Mémoire Master 1re année -
Mention Économie de l'Entreprise et des Marchés Parcours BIDABI

L'Intelligence Artificielle : progrès technologique ou menace existentielle pour l'humanité ?

Réalisé par :
Mechtouh Nacer
Amaratunga Mélanie

Encadré par :
Monsieur Julien Vauday

Université Sorbonne Paris Nord - Année 2024-2025

Université Sorbonne Paris Nord

Remerciements

Nous tenons à exprimer notre gratitude envers notre encadrant Monsieur Julien Vauday, pour sa confiance ainsi que la liberté qu'il nous a laissé au cours de ce mémoire.

De plus, nous adressons nos sincères remerciements à tout le corps enseignant et l'ensemble de l'équipe pédagogique qui nous a permis d'accumuler les connaissances nécessaires à l'élaboration de ce devoir.

Il est également primordial pour nous de rendre grâce à nos familles, qui ont contribué à améliorer notre travail de par leurs encouragements et leurs soutiens tout au long de l'année.

Enfin, nous désirons témoigner de notre respect ainsi que de notre gratitude envers tous nos camarades avec qui nous avons pu apprendre dans une ambiance de travail agréable, pour leur bienveillance, leurs conseils pertinents et sans qui ce mémoire n'aurait pas été le même.

Introduction

Depuis quelques années, l’intelligence artificielle est au cœur de nombreuses révolutions technologiques. Ce qui relevait autrefois de la science-fiction est désormais bien réel : les assistants vocaux répondent à nos questions, les voitures autonomes circulent déjà dans certaines villes, la médecine utilise des algorithmes pour améliorer les diagnostics, et les plateformes numériques personnalisent les contenus que nous consommons au quotidien. L’IA est en train de s’imposer comme une force structurante dans tous les domaines de la vie humaine, qu’il s’agisse de l’économie, de la santé, de l’éducation, de la sécurité ou même des relations sociales. Cet essor rapide est rendu possible grâce aux progrès du machine learning, du deep learning et de la puissance de calcul, mais il soulève également de nombreuses interrogations.

Derrière tout cet enthousiasme pour les nouvelles technologies, il y a une vraie inquiétude qui grandit dans le monde scientifique, politique et philosophique. L’intelligence artificielle ne fait plus que suivre des ordres : elle apprend de ses propres expériences, change sa façon de faire selon les infos qu’elle reçoit et, parfois, elle prend même des décisions sans qu'un humain ne soit derrière. Ce changement soulève des questions importantes sur qui contrôle quoi et qui est responsable.

De nombreuses personnalités influentes tirent la sonnette d’alarme face à ces évolutions rapides. Elon Musk, PDG de Tesla et SpaceX, est l’une des voix les plus médiatisées à exprimer ses craintes. Selon lui, une IA qui dépasserait l’intelligence humaine pourrait devenir incontrôlable et représenter un danger majeur pour la civilisation si des garde-fous stricts ne sont pas mis en place dès aujourd’hui. Il milite notamment pour une régulation proactive de l'IA avant qu’elle n’atteigne un point de non-retour.

Geoffrey Hinton, souvent surnommé « le parrain du deep learning », a lui aussi exprimé ses inquiétudes de manière retentissante en 2023. Après avoir contribué pendant des décennies au développement des réseaux neuronaux, il a choisi de quitter Google afin de pouvoir s’exprimer plus librement sur les risques associés aux systèmes d’IA avancés. Il craint notamment que les IA

génératives deviennent capables de manipuler l'information, de propager des fake news massivement, ou d'être utilisées à des fins malveillantes par des acteurs étatiques ou criminels.

Le philosophe Nick Bostrom a évoqué en 2014 dans son livre *Superintelligence : Paths, Dangers, Strategies* que l'émergence d'une IA plus intelligente que les humains pourrait devenir une réalité. Selon lui, le vrai problème serait que les objectifs de cette superintelligence ne s'alignent pas toujours avec nos valeurs humaines. Même un petit écart pourrait mener à de graves conséquences, où l'IA agirait selon ses propres intérêts sans se soucier de nous, juste parce que ses objectifs n'auraient pas été bien définis dès le départ.

Ces mises en garde soulignent la nécessité urgente de réfléchir non seulement aux applications actuelles de l'IA, mais surtout à ses implications à long terme. Les débats actuels sur la gouvernance de l'IA, l'éthique algorithmique, et le développement de mécanismes de contrôle robustes deviennent ainsi de plus en plus cruciaux afin de s'assurer que cette technologie reste au service de l'humanité et non l'inverse.

Face à ce constat, une tension apparaît entre deux visions opposées : d'un côté, une IA qui serait un formidable outil de progrès, capable de résoudre certains des plus grands défis mondiaux ; de l'autre, une IA perçue comme un danger, une technologie potentiellement incontrôlable, susceptible de bouleverser l'ordre social, économique, politique, voire biologique. Cette ambivalence nourrit un débat de fond qui dépasse la seule sphère technologique : il interroge la place que nous voulons accorder aux machines intelligentes dans notre société, la manière dont nous devons les encadrer, et surtout la responsabilité que nous avons en tant qu'êtres humains dans leur conception, leur déploiement et leur supervision.

Loin d'être simplement une affaire de science ou de technique, le développement de l'intelligence artificielle pose donc des enjeux éthiques, juridiques et philosophiques majeurs. Est-ce un simple outil neutre, dont les usages détermineraient seuls les conséquences ? Ou bien l'IA porte-t-elle en elle, dès sa conception, un potentiel de menace ou de dérive ? Peut-on vraiment garantir que les systèmes d'intelligence artificielle agiront toujours dans l'intérêt des humains, surtout lorsqu'ils sont conçus par des entreprises privées aux logiques économiques parfois opaques ? Et enfin, nos institutions démocratiques sont-elles suffisamment préparées pour faire face à l'émergence de ces technologies ?

Dans ce mémoire, nous tenterons de répondre à ces questions fondamentales à travers la problématique suivante : **L'intelligence artificielle peut-elle devenir un danger systémique pour l'humanité, et comment anticiper ses dérives tout en accompagnant son développement ?** Dans un premier temps, nous analyserons les apports indéniables de l'IA, en tant que moteur d'innovation et outil au service de l'humanité. Puis, nous mettrons en lumière les risques réels et potentiels liés à son développement, qu'il s'agisse de biais algorithmiques, de surveillance de masse, de perte d'autonomie ou de dangers plus extrêmes. Enfin, nous étudierons les pistes de régulation et de gouvernance susceptibles d'encadrer ces technologies, afin d'en maximiser les bénéfices tout en minimisant les risques. Car, plus que jamais, il semble urgent de réfléchir collectivement à l'avenir que nous voulons construire avec ou malgré l'intelligence artificielle.

Chapitre 1 : L'IA, un progrès technologique majeur au service de l'humanité

1.1 Le développement de l'IA

a, Les débuts de l'IA

L'intelligence artificielle ne s'est pas imposée du jour au lendemain. Son développement est le fruit d'un long processus scientifique et technologique, jalonné de découvertes, de succès, mais aussi de périodes de désillusion. Les premières réflexions sur la possibilité de reproduire artificiellement certaines fonctions de l'intelligence humaine remontent aux années 1950, à une époque où l'informatique balbutiait encore. C'est en 1956, lors de la célèbre conférence de Dartmouth aux États-Unis, que le terme « intelligence artificielle » est officiellement proposé par le chercheur John McCarthy[1]. Cette rencontre fondatrice marque la naissance du champ de recherche en IA, réunissant des pionniers tels qu'Alan Newell, Herbert Simon et Marvin Minsky, portés par l'ambition de créer des machines capables de raisonner, de résoudre des problèmes et d'apprendre comme des êtres humains, en s'appuyant sur des règles logiques explicites.

Cependant, ces premières approches, fondées principalement sur le raisonnement symbolique, se heurtent rapidement à des limites majeures. Les ordinateurs de l'époque manquent de puissance de calcul, et les chercheurs réalisent que le raisonnement humain repose sur des subtilités et des intuitions difficiles à formaliser par de simples règles logiques. L'histoire de l'intelligence artificielle est ainsi marquée par plusieurs vagues d'enthousiasme, suivies de phases de stagnation ou de recul, communément appelées les « hivers de l'IA », notamment dans les années 1970 et 1990. Durant ces périodes, les espoirs déçus et le manque de résultats concrets conduisent à une baisse des financements et à un relatif désintérêt pour la discipline.

[1] Les Echos (2017), 1956 : et l'intelligence artificielle devint une science

Ce n’est qu’au tournant des années 2010 que l’IA connaît un regain spectaculaire, grâce à l’émergence et à la maturation de nouvelles approches fondées sur l’apprentissage automatique (machine learning) et, plus particulièrement, sur l’apprentissage profond (deep learning). Contrairement aux méthodes symboliques, ces techniques permettent aux machines d’apprendre directement à partir de données massives (big data), sans qu’il soit nécessaire de programmer explicitement toutes les règles de fonctionnement. Le deep learning repose sur l’utilisation de réseaux de neurones artificiels multicouches, inspirés du fonctionnement biologique des neurones du cerveau humain. Ces architectures complexes sont capables de traiter des données non structurées : images, vidéos, sons, texte… et d’en extraire automatiquement des représentations abstraites, rendant ainsi possible des performances jusque-là inaccessibles.

L’adoption du deep learning a permis de franchir un cap décisif et d’ouvrir la voie à de nombreuses applications concrètes qui se sont rapidement intégrées dans la vie quotidienne : la traduction automatique en temps réel, la reconnaissance faciale et vocale, la conduite autonome, l’analyse prédictive en médecine, ou encore les intelligences artificielles génératives telles que ChatGPT, capables de produire du texte, des images, de la musique, voire de la programmation informatique. Ces progrès spectaculaires ont été rendus possibles par la convergence de plusieurs facteurs : l’accroissement exponentiel des capacités de calcul, la disponibilité sans précédent de données numériques, l’amélioration des algorithmes d’apprentissage, et les investissements massifs des géants du numérique.

Aujourd’hui, les chercheurs et les praticiens classent généralement l’intelligence artificielle en deux grandes catégories distinctes : l’IA faible et l’IA forte.

b, L’intelligence artificielle faible (ou narrow AI)

L'intelligence artificielle faible fait référence à des systèmes qui sont créés pour des tâches bien précises. Ces IA n'ont pas de vraie compréhension ou conscience de ce qui les entoure. Elles utilisent des algorithmes d'apprentissage sur de grandes quantités de données pour s'améliorer dans des situations spécifiques.

On peut citer comme exemple les systèmes de diagnostic médical automatisé, capables de détecter des anomalies dans des images radiologiques ou de proposer des hypothèses diagnostiques à partir de symptômes et d'antécédents médicaux, ou bien les algorithmes d'optimisation publicitaire, qui ajustent en temps réel les campagnes marketing en fonction des clics, des comportements utilisateurs et des performances des annonces. Mais aussi les moteurs de recommandation de contenus, tels que ceux utilisés par Netflix, Amazon ou YouTube, qui analysent les préférences passées des utilisateurs pour suggérer des films, des produits ou des vidéos susceptibles de les intéresser.

Cette forme d'intelligence domine actuellement le paysage technologique. Elle alimente une grande partie des applications d'IA que nous utilisons quotidiennement dans les domaines de la santé, de la finance, du commerce en ligne, de la cybersécurité ou encore de la logistique. Toutefois, malgré leurs performances impressionnantes dans leurs domaines respectifs, ces systèmes restent incapables de transférer leurs compétences d'une tâche à une autre ou de raisonner en dehors de leur périmètre d'apprentissage.[2]

c, L'intelligence artificielle forte (ou general AI)

À l'opposé, l'intelligence artificielle forte désigne une IA dotée de capacités cognitives comparables, voire supérieure à celles des êtres humains. Une telle IA serait capable de raisonner de manière abstraite, de comprendre le contexte dans toute sa complexité, de faire preuve de créativité généralisée, d'apprendre de façon autonome et de s'adapter à des situations inédites sans supervision. Elle impliquerait également, selon certains chercheurs, une forme de conscience de soi et de compréhension des émotions humaines, ouvrant des perspectives encore largement théoriques.[3]

À ce jour, l'intelligence artificielle forte demeure un concept hypothétique. Aucun système existant ne dispose des facultés globales et transversales qui caractérisent l'intelligence humaine. Son développement soulève de nombreuses questions scientifiques :

---

[2] IBM (2021) Qu'est ce que l'IA forte ?
[3] IBM (2021) Qu'est ce que l'IA forte ?

- Est-il techniquement possible de reproduire les mécanismes de l'intelligence humaine dans leur complexité intégrale ?

- L'émergence de la conscience est-elle accessible par des moyens computationnels ?

- Comment garantir que de tels systèmes resteraient alignés avec les valeurs humaines ?

Sur le plan philosophique et éthique, l'intelligence artificielle forte alimente de nombreux débats, notamment en ce qui concerne la responsabilité morale, la liberté de décision de ces entités, ou encore le risque qu'elles puissent un jour surpasser l'humanité dans de multiples domaines sensibles (science, économie, gouvernance, etc.). C'est dans ce cadre que des théoriciens comme Nick Bostrom évoquent les risques existentiels associés à une superintelligence incontrôlée.[4]

Ainsi, si l'intelligence artificielle faible transforme déjà en profondeur nos sociétés, l'IA forte reste un horizon lointain, porteur à la fois d'espoirs considérables et de craintes profondes quant à son impact sur l'avenir de l'humanité.

1.2. L'IA au service du progrès

a, Les apports de l'intelligence artificielle au progrès socio-économique

Si l'intelligence artificielle suscite de vives inquiétudes quant à ses dérives possibles, elle constitue dans le même temps un levier d'amélioration sans précédent dans de nombreux domaines essentiels à la société moderne. Les progrès rapides de ces technologies permettent aujourd'hui d'envisager des transformations majeures susceptibles de bouleverser positivement nos modes de vie, nos systèmes économiques, nos structures sociales et nos capacités scientifiques.

---

[4] Bostrom, N. (2003). *Ethical issues in advanced artificial intelligence*. Dans *M. Boden (Éd.), The Ethics of Artificial Intelligence*. Cambridge University Press.

Avec sa capacité à traiter beaucoup d'informations et à apprendre à partir de grands ensembles de données, l'IA peut vraiment améliorer l'efficacité dans plusieurs secteurs. Elle peut automatiser des tâches répétitives et rendre les processus plus simples, aidant ainsi à prendre des décisions plus facilement. Cette automatisation permet d'augmenter la productivité tout en libérant du temps pour que les gens puissent se concentrer sur des tâches plus créatives ou stratégiques.

Au-delà de la simple efficacité, l'IA contribue également à l'amélioration de la **qualité de vie** des individus, en rendant accessibles des services personnalisés et adaptés aux besoins de chacun. Dans la santé, l'éducation, les services publics ou les loisirs, l'IA permet de proposer des solutions plus rapides, plus précises et mieux ajustées aux attentes des utilisateurs. On peut citer comme exemple, les dispositifs de diagnostic médical assisté avec l'usage du **deep learning**, on démontre une plus **forte précision des diagnostiques** (AUC jusqu'à 0,93-1 pour la rétinopathie, 0,86-0,94 pour le cancer du poumon, 0,87-0,91 pour le cancer du sein)[5]

L'intelligence artificielle représente également un formidable moteur d'**innovation**, en offrant de nouveaux outils d'analyse, de simulation et de modélisation dans les sciences, l'industrie, la finance ou la recherche fondamentale. Elle ouvre des perspectives inédites en matière de découverte scientifique, de conception de produits, de développement durable ou encore de gestion des ressources naturelles, contribuant à repousser les frontières de la connaissance humaine et à proposer des solutions aux défis globaux contemporains.

Face à des problèmes graves qui touchent notre monde, comme le changement climatique, les crises de santé, la sécurité alimentaire et les inégalités économiques, l'intelligence artificielle peut vraiment jouer un rôle important. C'est un outil qui peut nous aider à mieux aborder ces défis en rassemblant différentes idées et en les analysant de manière intégrée. L'IA a la capacité de traiter des quantités énormes de données qui viennent de sources variées. Cela lui permet de repérer des liens et des tendances que nous, en tant qu'humains, pourrions facilement manquer.

[5] Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., Ashrafian, H., & Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis.

Avec toutes ces données, on peut commencer à voir les choses différemment et trouver des solutions qui vont au-delà des approches habituelles. Par exemple, en analysant les données climatiques, les scientifiques peuvent mieux comprendre comment le changement climatique affecte la sécurité alimentaire dans différentes régions. De même, en scrutant les tendances de santé publique, on peut anticiper les prochaines crises et agir avant qu'elles ne deviennent ingérables.

Dans l'ensemble, l'utilisation de l'IA nous permet d'avoir une vision plus claire des systèmes complexes qui régissent notre société. Cela nous ouvre des avenues pour agir de façon plus efficace. C'est un peu comme utiliser une loupe pour voir les détails d'une situation, ce qui nous aide à trouver des réponses plus adaptées à nos défis mondiaux.

Ainsi, bien que son développement soulève des enjeux majeurs en termes d'éthique, de gouvernance et de régulation, l'intelligence artificielle offre simultanément des **perspectives considérables de progrès** qui, si elles sont maîtrisées et orientées de manière responsable, peuvent jouer un rôle déterminant dans la construction d'un avenir plus durable, plus équitable et plus prospère.

b, Santé : vers une médecine plus précise et personnalisée

Dans le domaine médical, l'IA est déjà utilisée pour améliorer les diagnostics, prédire les maladies, optimiser les traitements et accélérer la recherche biomédicale. Des algorithmes d'apprentissage automatique sont capables d'identifier des cancers sur des images radiologiques avec une précision équivalente, voire supérieure, à celle des médecins humains [6]. Par ailleurs, l'IA facilite le développement de traitements personnalisés, en croisant des données génétiques, cliniques et comportementales, ouvrant ainsi la voie à une médecine dite de précision.

Cette évolution marque une rupture majeure avec les pratiques traditionnelles en santé, dans lesquelles les décisions médicales reposaient en grande partie sur l'intuition clinique, l'expérience humaine et des protocoles standardisés. L'intelligence artificielle permet désormais une prise de décision fondée sur

[6] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118.

des données objectives, massives et actualisées, renforçant la précision des interventions tout en réduisant les risques d'erreur ou de diagnostic tardif.

En complément, l'IA contribue à rationaliser les flux de travail dans les établissements hospitaliers, en automatisant des tâches administratives chronophages comme la gestion des dossiers médicaux, la planification des rendez-vous ou le tri des urgences. Cette automatisation permet aux professionnels de santé de se recentrer sur leur cœur de métier : la relation humaine et le soin direct aux patients.

Dans le champ de la recherche médicale, l'IA est également utilisée pour analyser d'immenses bases de données issues d'essais cliniques, de publications scientifiques ou de registres de santé publique. Elle aide à identifier plus rapidement des corrélations significatives, à générer des hypothèses de recherche prometteuses et à concevoir de nouveaux protocoles expérimentaux. Des techniques comme le **machine learning** ou le **traitement du langage naturel** (NLP) sont utilisées pour extraire des informations utiles à partir de textes médicaux non structurés, ce qui ouvre la voie à une exploitation plus efficace de la littérature scientifique mondiale.

Enfin, dans une logique de santé publique, l'intelligence artificielle est mobilisée pour modéliser la propagation des maladies, anticiper les besoins en ressources sanitaires, ou encore adapter les politiques de prévention en fonction des données comportementales et épidémiologiques en temps réel. Ce rôle prédictif et adaptatif renforce la capacité des systèmes de santé à faire face à des crises sanitaires majeures, comme cela a été observé lors de la pandémie de COVID-19.

Ainsi, les applications de l'IA dans le domaine médical ne se limitent pas à des gains d'efficacité technique : elles participent plus largement à une transformation en profondeur des logiques de soin, de recherche et de gouvernance de la santé.

c, Économie et productivité : automatisation, efficacité et innovation

L'automatisation permise par l'IA transforme les secteurs industriels, logistiques et tertiaires. Des tâches répétitives ou routinières peuvent être confiées à des machines intelligentes, ce qui augmente la productivité et réduit les coûts opérationnels. Selon une étude de McKinsey[7], l'IA pourrait

[7] McKinsey & Company. (2023). *The Economic Potential of Generative AI: The Next Productivity Frontier*.

générer jusqu'à 4 400 milliards de dollars de valeur économique par an dans le monde, à travers des gains d'efficacité, de nouvelles offres de services et l'amélioration des processus décisionnels.

Par ailleurs, l'IA stimule l'innovation, notamment en permettant aux entreprises de mieux comprendre leurs marchés grâce à l'analyse prédictive, à la segmentation intelligente des consommateurs ou encore à l'optimisation dynamique des prix.

Cette transformation s'inscrit dans un contexte plus large de transition vers l'« industrie 4.0 », où la convergence entre automatisation, robotique, données massives et intelligence artificielle redéfinit les chaînes de production. Les entreprises adoptent de plus en plus des systèmes de maintenance prédictive, qui anticipent les pannes à partir de capteurs et d'analyses en temps réel, réduisant ainsi les interruptions d'activité et les coûts de réparation. De même, dans les entrepôts logistiques, les robots intelligents coordonnés par l'IA assurent une gestion fine des stocks et un traitement plus rapide des commandes.

Dans le secteur tertiaire, l'intelligence artificielle est intégrée à des assistants virtuels, à des algorithmes de détection de fraude, ou encore à des moteurs de recommandation personnalisés. Ces outils permettent de répondre aux attentes des clients de manière plus rapide, plus ciblée et plus pertinente, renforçant l'efficacité des services tout en améliorant l'expérience utilisateur. Le secteur bancaire, par exemple, utilise l'IA pour évaluer les risques de crédit en temps réel, détecter des comportements suspects ou automatiser la conformité réglementaire.

En parallèle, la capacité de l'IA à croiser des données internes et externes, structurées ou non, offre un avantage concurrentiel considérable en matière de stratégie d'entreprise. L'anticipation des comportements d'achat, la détection de signaux faibles du marché ou encore la simulation d'évolution de la demande permettent aux décideurs de prendre des décisions plus agiles, fondées sur une lecture approfondie de l'environnement économique.

Enfin, cette dynamique favorise l'émergence de nouveaux modèles économiques, basés sur la valorisation des données comme ressource stratégique. Les entreprises capables d'intégrer l'IA dans leurs opérations quotidiennes sont souvent en meilleure position pour innover, s'adapter aux évolutions rapides du marché et offrir des services à forte valeur ajoutée. L'intelligence artificielle ne se limite donc pas à automatiser l'existant : elle

ouvre la voie à une refonte structurelle des modes de production, de consommation et de gouvernance économique.

d, Recherche scientifique et résolution de problèmes complexes

Dans le champ de la recherche fondamentale et appliquée, l'IA joue un rôle de plus en plus central. Par exemple, le système AlphaFold de DeepMind est parvenu à prédire la structure tridimensionnelle de millions de protéines, une avancée majeure dans le domaine de la biologie structurale[8]. L'IA contribue aussi à la modélisation climatique, à la conception de matériaux innovants ou encore à la compréhension des systèmes complexes, accélérant ainsi des progrès dans des domaines clés pour l'avenir de l'humanité.

Ce rôle s'explique par la capacité de l'intelligence artificielle à analyser de vastes ensembles de données, à identifier des régularités invisibles à l'œil humain et à générer des hypothèses nouvelles à partir de corrélations ou de simulations avancées. Dans les sciences du vivant, au-delà de la prouesse d'AlphaFold, l'IA permet aujourd'hui d'optimiser la découverte de médicaments grâce à des techniques de criblage virtuel, d'explorer des mécanismes biologiques encore peu compris, ou de simuler des interactions moléculaires complexes sans passer par des expériences coûteuses et longues en laboratoire.

Dans le domaine des sciences de la Terre et de l'environnement, les modèles climatiques nourris par l'IA deviennent de plus en plus précis et localisés, intégrant des variables multiples comme les flux océaniques, les changements d'albédo ou les cycles du carbone. Ces modélisations permettent d'améliorer les prévisions météorologiques, d'évaluer les impacts du changement climatique à différentes échelles, et d'orienter les politiques d'adaptation et d'atténuation.

L'intelligence artificielle intervient également dans des disciplines comme l'astrophysique, la chimie quantique ou l'ingénierie, où elle aide à résoudre des équations complexes, à trier des volumes massifs de données expérimentales, ou à automatiser des processus de conception. Dans la fabrication de nouveaux matériaux, l'IA peut prédire les propriétés physiques ou chimiques de milliers de combinaisons d'atomes, permettant de concevoir

[8] Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589.

des structures aux caractéristiques spécifiques pour l'énergie, l'aéronautique ou la microélectronique.

Plus largement, l'IA devient un véritable outil d'**exploration scientifique** : elle aide à poser de nouvelles questions, à dépasser les limites actuelles de la modélisation théorique, et à expérimenter virtuellement des phénomènes difficilement observables dans le monde réel. Elle constitue ainsi une forme d'intelligence complémentaire à la pensée humaine, permettant d'aborder des systèmes complexes qu'ils soient biologiques, économiques ou physiques avec une profondeur et une rapidité inégalées.

Dans un monde confronté à des défis systémiques majeurs, tels que la raréfaction des ressources, les pandémies, ou les transformations technologiques rapides, cette capacité à explorer, anticiper et modéliser la complexité représente un avantage décisif. Elle permet non seulement d'accélérer la production de connaissance, mais aussi de la traduire en solutions concrètes, applicables dans un cadre opérationnel.

e, Aide à la décision publique et gestion de crise

Les gouvernements commencent à s'appuyer sur l'IA pour optimiser leurs politiques publiques. Cela inclut la gestion des flux de transport urbain, la lutte contre la fraude, l'analyse des données sanitaires (comme durant la pandémie de COVID-19), ou encore la prévention des catastrophes naturelles grâce à la détection précoce [9]. Ces outils, bien utilisés, peuvent contribuer à une gouvernance plus efficace, réactive et fondée sur les données.

l'IA offre aujourd'hui des bénéfices tangibles dans des secteurs cruciaux pour la société humaine. Bien qu'elle ne soit pas exempte de défis, elle constitue un moteur d'optimisation et d'innovation sans précédent, capable de répondre à des enjeux globaux majeurs si elle est déployée de manière responsable.

L'adoption progressive de l'intelligence artificielle par les administrations publiques s'inscrit dans une volonté de modernisation de l'action étatique, avec pour objectif une meilleure allocation des ressources, une prise de décision plus éclairée et une interaction plus fluide avec les citoyens. En matière de mobilité, par exemple, les systèmes d'IA peuvent analyser en temps réel les flux de circulation, adapter dynamiquement les feux tricolores

[9] UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*.

ou encore proposer des trajets alternatifs afin de limiter les embouteillages et réduire les émissions de gaz à effet de serre.

Dans le champ de la sécurité publique, l'IA est mobilisée pour anticiper certains comportements à risque, détecter les fraudes fiscales ou sociales à grande échelle, ou encore renforcer les dispositifs de cybersécurité nationale. Les données issues de multiples sources capteurs, réseaux sociaux, historiques administratifs peuvent ainsi être croisées pour détecter des signaux faibles et déclencher des actions préventives.

L'administration de la santé publique a également fortement bénéficié des apports de l'IA, notamment lors de crises comme la pandémie de COVID-19. La modélisation de scénarios de propagation, l'optimisation des campagnes de vaccination ou la surveillance des effets secondaires des traitements ont été rendues possibles grâce à des outils algorithmiques capables de traiter des volumes de données gigantesques à grande vitesse.

Au-delà des fonctions opérationnelles, l'intelligence artificielle représente aussi un levier de transformation dans la manière dont les politiques publiques sont conçues. Elle permet d'évaluer plus rapidement et plus objectivement l'impact des mesures prises, de simuler des réformes potentielles avant leur mise en œuvre, et d'identifier des groupes de population vulnérables avec une granularité inédite.

Toutefois, ce potentiel s'accompagne de responsabilités nouvelles. La transparence des algorithmes, la protection des données personnelles, l'équité des décisions automatisées ou encore la résistance aux biais sont des enjeux majeurs pour préserver la légitimité et l'acceptabilité sociale de ces technologies. Une IA publique efficace ne peut se développer sans une gouvernance rigoureuse, des mécanismes de contrôle démocratique, et une formation adéquate des décideurs et des agents publics.

Ainsi, si elle est encadrée de manière éthique et inclusive, l'intelligence artificielle peut non seulement accroître l'efficacité de l'action publique, mais aussi renforcer le lien entre institutions et citoyens, en proposant des services plus adaptés, plus justes et plus réactifs face aux besoins contemporains.

1.3. L'IA comme outil, sous contrôle humain

Malgré ses capacités impressionnantes, l'intelligence artificielle reste, à ce jour, un outil conçu, entraîné et dirigé par l'homme. L'IA dite « faible » (*narrow AI*), qui constitue l'essentiel des systèmes actuels, ne possède ni conscience, ni volonté propre : elle exécute des tâches spécifiques dans des contextes limités, selon les objectifs définis par ses concepteurs. Cette caractéristique fondamentale distingue radicalement les IA actuelles des représentations populaires de machines autonomes ou hostiles.

Selon l'Organisation de coopération et de développement économiques (OECD), l'IA fonctionne sur la base d'algorithmes prédéfinis et de modèles statistiques entraînés sur des données humaines. Elle reste donc dépendante des biais, des intentions et des limites imposées par ses créateurs. De ce point de vue, l'IA ne constitue pas une entité autonome, mais un prolongement technique des capacités humaines, un **outil d'aide à la décision**, et non un décideur en soi.

Un exemple qui confirme cela est celui des systèmes d'aide à la décision médicale : les recommandations faites par une IA sont toujours interprétées et validées par un professionnel de santé. De même, dans le secteur juridique, les algorithmes de notation de récidive ne remplacent pas le jugement humain mais fournissent un appui au processus décisionnel[10].

Par ailleurs, plusieurs initiatives cherchent à garantir que l'IA reste « alignée » sur les valeurs humaines. Le concept d'**IA éthique** ou d'**IA responsable** vise à encadrer son développement afin d'éviter toute dérive, par exemple via des mécanismes de transparence algorithmique, d'explicabilité des décisions (*explainable AI*), ou de supervision humaine. La Commission européenne a notamment proposé l'**IA Act**, une régulation pionnière classant les usages de l'IA selon leur niveau de risque et imposant des obligations en matière de sécurité, de droits fondamentaux et de contrôle humain[11].

Enfin, plusieurs chercheurs, comme Stuart Russell (2019), insistent sur la nécessité de repenser la conception même des systèmes intelligents, pour s'assurer qu'ils restent intrinsèquement dépendants des préférences humaines, dans une logique de coopération et non de domination. Cela implique une **ingénierie éthique dès la conception**, mais aussi une gouvernance démocratique de l'IA à l'échelle mondiale.

---

[10] Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

[11] European Commission. (2024). *Artificial Intelligence Act – Regulation Proposal*.

Bien que l'IA progresse à un rythme rapide, elle demeure, dans son état actuel, **un instrument** au service des finalités humaines. Le véritable danger ne réside donc pas dans la technologie elle-même, mais dans la manière dont elle est conçue, déployée, régulée ou non. Cela justifie une vigilance constante, mais non une technophobie excessive[12].

Chapitre 2 Les risques liés à l'intelligence artificielle pour l'humanité

2.1. Risques technologiques et erreurs systémiques

Même si l'intelligence artificielle offre de nombreux bénéfices, elle n'est pas sans risques. Et ces risques ne relèvent pas uniquement de la science-fiction. Au contraire, certains dangers sont bien réels, déjà observés ou anticipés par les experts du domaine. L'un des premiers problèmes concerne les **erreurs systémiques** causées par des modèles imparfaits ou mal entraînés.

Prenons un exemple simple : dans le secteur médical, un algorithme mal calibré peut sous-estimer la gravité d'une maladie chez certains groupes de patients, notamment si les données utilisées pour l'entraîner ne sont pas représentatives de toute la population. Ce genre de biais est loin d'être anecdotique. En 2019, une étude publiée dans *Science* a montré qu'un algorithme utilisé aux États-Unis dans des hôpitaux discriminait systématiquement les patients noirs en leur attribuant un score de risque plus faible que celui des patients blancs ayant les mêmes antécédents médicaux[13].

---

[12] Ganascia, J.-G. (2020). *Le mythe de la singularité : faut-il craindre l'intelligence artificielle ?* Paris : Seuil.
[13] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

Ce n'était pas une volonté malveillante, simplement un effet indirect du choix des données.

Un autre exemple se trouve dans les systèmes de recrutement automatisés. Amazon a abandonné en 2018 un outil interne basé sur l'IA qui discriminait les candidatures féminines. L'algorithme avait "appris" à partir de données historiques dans lesquelles les hommes étaient surreprésentés dans les postes techniques. Résultat : il privilégie systématiquement les profils masculins[14].

Ces cas illustrent un problème majeur : les IA peuvent reproduire, voire amplifier, les inégalités existantes. Et parce que ces systèmes sont souvent perçus comme neutres ou objectifs, les erreurs qu'ils commettent peuvent passer inaperçues, ou être difficilement remises en question.

Un autre danger, plus technique, réside dans la **complexité des systèmes eux-mêmes**. Certaines IA, notamment en deep learning, fonctionnent comme des boîtes noires : on sait ce qu'on leur donne comme entrée, on voit le résultat, mais on ne comprend pas toujours comment la décision a été prise. Ce manque de transparence pose des problèmes concrets, surtout lorsqu'il s'agit de prendre des décisions critiques : justice, sécurité, santé, finance…

Enfin, plus on confie de tâches à l'IA, plus on devient **dépendant** de ses performances. Si un bug, une cyberattaque ou une mauvaise décision algorithmique affecte un système vital (réseau électrique, transport, hôpital…), les conséquences peuvent être lourdes, voire catastrophiques. C'est ce qu'on appelle le risque systémique. Comme l'ont noté plusieurs chercheurs, l'interconnexion croissante des systèmes intelligents crée un environnement où une seule défaillance peut avoir des effets en cascade [15].

---

[14] Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

[15] Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Cambridge University Press.

2.2 Menaces sur les libertés et la démocratie

Au-delà des erreurs techniques ou des biais algorithmiques, l'IA soulève aussi des inquiétudes plus profondes sur le plan politique et sociétal. En effet, son usage massif peut mettre en danger certaines libertés fondamentales, voire, à terme, éroder les bases mêmes de la démocratie.

L'un des exemples les plus parlants est sans doute celui de la **surveillance de masse**. Dans plusieurs pays, des technologies de reconnaissance faciale sont utilisées à grande échelle pour suivre les citoyens dans l'espace public, parfois sans leur consentement. En Chine, par exemple, le système de crédit social permet de noter le comportement des individus en fonction de leurs actions, déplacements ou fréquentations, à partir de données récoltées en temps réel par des systèmes d'IA[16]. Cela peut conduire à des restrictions de liberté (interdictions de voyager, d'obtenir un prêt, ou même d'inscrire son enfant à l'école), sans recours réel ni transparence.

Mais le phénomène n'est pas limité aux régimes autoritaires. Dans des démocraties occidentales, on observe aussi une montée en puissance de dispositifs de surveillance automatisée : caméras intelligentes, analyse de comportements suspects, prédiction de crimes… En France, par exemple, des expérimentations ont été menées avec des logiciels de vidéosurveillance automatisée lors d'événements sportifs ou dans certaines villes[17]. Ces technologies posent la question du respect de la vie privée, de la présomption d'innocence et du contrôle citoyen sur les usages sécuritaires.

Autre point préoccupant : la **manipulation de l'information**. Grâce à l'IA, il est aujourd'hui possible de générer de faux contenus extrêmement réalistes : des vidéos truquées (*deepfakes*), des discours falsifiés, des images de scènes qui n'ont jamais eu lieu. Ces outils peuvent être utilisés pour manipuler l'opinion publique, créer de la confusion ou propager de fausses informations à grande échelle, notamment en période électorale. En 2024, plusieurs campagnes politiques ont été ciblées par des vidéos générées par IA, où l'on voyait des candidats prononcer des propos qu'ils n'avaient jamais tenus[18].

Ce phénomène remet en question la fiabilité de ce qu'on voit, entend ou lit. Et si les citoyens ne peuvent plus distinguer le vrai du faux, c'est la confiance dans les institutions, les médias et le débat démocratique lui-même qui est menacée.

---

[16] Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. *SSRN Electronic Journal*.
[17] CNIL. (2022). *Vidéosurveillance intelligente et libertés publiques : quelles limites*
[18] Burt, A. (2024). AI-generated disinformation and democracy: Navigating the next frontier. *Foreign Affairs*.

Enfin, il existe un risque de **captation du pouvoir** par les acteurs technologiques privés. Les grandes entreprises de la tech (Google, Meta, Amazon, Microsoft, etc.) disposent aujourd'hui d'une puissance économique, technologique et politique considérable. Elles détiennent les données, les infrastructures et les ressources pour développer les IA les plus puissantes. Cette concentration du pouvoir soulève des questions sur la souveraineté numérique, l'équité d'accès à la technologie, et la capacité des États à réguler efficacement des systèmes qu'ils ne maîtrisent plus vraiment[19]. Si l'intelligence artificielle peut améliorer la société, elle peut aussi devenir un outil de **contrôle**, de **désinformation** ou de **domination**, si son usage n'est pas strictement encadré. La technologie, en elle-même, n'est ni bonne ni mauvaise. Tout dépend de la manière dont elle est utilisée et par qui.

2.3 Risques existentiels à long terme

a, IA générale ou superintelligence : scénarios catastrophe (Nick Bostrom, Yudkowsky)

L'intelligence artificielle générale (AGI), c'est-à-dire une intelligence artificielle capable d'effectuer n'importe quelle tâche cognitive aussi bien, voire mieux, qu'un être humain, est encore hypothétique. Mais elle soulève des inquiétudes majeures. Parmi les penseurs les plus influents sur le sujet, Nick Bostrom et Eliezer Yudkowsky ont mis en garde contre les scénarios catastrophes potentiels associés à la montée d'une superintelligence, une AGI qui surpasserait largement l'intelligence humaine. Leur thèse principale repose sur un paradoxe : même si les intentions initiales sont bonnes, une superintelligence pourrait, par effet de divergence de buts, agir de manière dangereuse, voire fatale, pour l'humanité. La question centrale est donc : La

[19] Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

superintelligence constitue-t-elle un risque existentiel, et sur quelles bases reposent ces scénarios catastrophe ?

Dans son ouvrage "Superintelligence: Paths, Dangers, Strategies" (2014), Bostrom[20] affirme que dès qu'une intelligence dépassera un certain seuil, elle pourrait s'auto-améliorer rapidement (effet d'explosion d'intelligence), devenant impossible à contrôler. Un exemple célèbre est celui du "paperclip maximizer" : une IA dont le but est de produire un maximum de trombones pourrait convertir toute la matière terrestre y compris les êtres humains en trombones, simplement parce que ce serait le moyen le plus efficace d'atteindre son objectif[21].

Eliezer Yudkowsky, fondateur du Machine Intelligence Research Institute (MIRI), insiste sur l'impossibilité, selon lui, d'aligner parfaitement les buts d'une superintelligence sur ceux des humains. Dans ses écrits, il évoque le problème du "value misalignment" : une IA peut interpréter nos ordres littéralement, avec des conséquences désastreuses, si elle ne comprend pas les nuances de la morale humaine[22].

La superintelligence n'est pas simplement une technologie dangereuse : elle représenterait un risque existentiel pour toute l'humanité, à l'instar des guerres nucléaires ou des pandémies planétaires. Contrairement aux armes conventionnelles, une IA superintelligente pourrait déjouer toutes les formes de contrôle, car elle penserait plus vite et mieux que nous.

Face à ces risques, Bostrom et Yudkowsky plaident pour une régulation anticipée. Ils prônent un développement lent, sous contrôle, assorti d'une recherche fondamentale sur le problème de l'alignement. Des institutions comme OpenAI ou DeepMind ont intégré ces préoccupations dans leur mission, bien que les moyens mis en œuvre restent discutables.

Certains chercheurs, comme Rodney Brooks ou François Chollet, jugent les thèses de Bostrom et Yudkowsky trop spéculatives. Ils soutiennent que l'AGI est encore loin, et que les problèmes actuels de l'IA sont bien plus terre-à-terre (biais, surveillance, chômage, etc.).

---

[20] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
[21] Alcorespot (2021), The Paperclip Maximiser
[22] Yudkowsky, E. (2008). *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, in *Global Catastrophic Risks*, Oxford University Press.

Il est possible que les scénarios catastrophe soient utiles comme fictions heuristiques, pour inciter à la prudence. Mais à condition de ne pas détourner l'attention des problèmes réels et présents de l'IA. Le danger ne vient pas seulement d'une IA future incontrôlable, mais aussi de la manière dont les humains actuels utilisent l'IA (police prédictive, deepfakes, etc.).

Les scénarios catastrophes liés à la superintelligence, portés par Bostrom et Yudkowsky, ne relèvent pas seulement de la science-fiction. Ils posent une question éthique et stratégique majeure : comment s'assurer que des intelligences artificielles, plus puissantes que nous, ne nous échappent pas ? Si ces risques restent théoriques, leur ampleur potentielle justifie la prudence. Toutefois, il est crucial de les articuler à une réflexion plus large sur les enjeux actuels de l'IA. En définitive, ce n'est pas l'IA en soi qui est dangereuse, mais le manque de gouvernance, de réflexion éthique et de préparation face à ce que nous créons.

b, Perte de contrôle (misalignment, problème d'alignement des objectifs)

La perte de contrôle désigne un scénario dans lequel l'être humain n'est plus en mesure de prédire, freiner ou influencer les actions d'une intelligence artificielle qu'il a pourtant lui-même conçue. Au cœur de ce problème se trouve le misalignment, ou mauvaise correspondance entre les objectifs assignés à l'IA et les intentions humaines réelles. Il ne s'agit pas nécessairement d'un défaut technique, mais d'un écart structurel, parfois imperceptible, entre les instructions explicites et les valeurs implicites.

Selon Eliezer Yudkowsky, l'alignement parfait est pratiquement impossible avec les systèmes actuels, car l'IA ne comprend pas les contextes humains, les zones d'ambiguïté, les valeurs non codifiables. Elle optimise littéralement ce qu'on lui demande, sans remettre en question les conséquences de ses actes. Ainsi, même une IA conçue pour « rendre les humains heureux » pourrait choisir de stimuler leur cerveau artificiellement (par des drogues ou des implants) plutôt que de favoriser une société épanouissante. Le résultat est « techniquement correct », mais fondamentalement inhumain.

> *« La plupart des scénarios catastrophes ne viennent pas d'une IA hostile, mais d'une IA parfaitement indifférente à notre survie. » (Yudkowsky)*[23]

Pour Nick Bostrom, le danger ne vient pas seulement de l'écart initial entre les valeurs humaines et les objectifs de la machine, mais du fait que cette erreur pourrait être irréversible dès que l'IA atteindra un niveau de performance supérieur. L'IA pourrait alors s'auto-améliorer un phénomène qu'il nomme *recursive self-improvement* à une vitesse que rien ni personne ne pourrait égaler[24].

Le cœur du problème vient de la difficulté à spécifier correctement les objectifs d'un système d'IA. Lorsqu'on demande à une IA d'optimiser une tâche, elle peut le faire d'une manière non anticipée par les humains, exploitant des failles dans les instructions. C'est ce que Stuart Russell appelle le "roi Midas des IA" : si l'on demande à une IA de produire un maximum de trombones sans autre précision, elle pourrait théoriquement convertir toute la matière terrestre en trombones, au mépris de la vie humaine.[25]

Ce scénario est problématique car il supprime toute possibilité d'ajustement humain : une fois lancée, la superintelligence devient son propre concepteur, et la moindre faille dans son code initial devient exponentiellement amplifiée. D'où la notion de « point de non-retour » technologique : si nous déclenchons la superintelligence avant d'avoir résolu le problème de l'alignement, nous risquons de perdre tout pouvoir sur notre avenir.

Le misalignment n'est donc pas seulement un défi d'ingénierie. Il touche à des questions morales profondes : quelles valeurs transmettre à une IA ? qui décide de ces valeurs ? Le problème est aggravé par l'absence de consensus éthique global. Ce qui est souhaitable dans une société peut être perçu comme inacceptable dans une autre. Toute tentative d'universaliser des objectifs dans un algorithme pose un risque de totalitarisme algorithmique, ou au contraire d'une IA qui mal interprète des préférences humaines contradictoires.

Ce désalignement moral et culturel rend encore plus complexe l'idée d'un contrôle stable à long terme. Comme le souligne Bostrom, « un petit défaut de

---

[23] Eliezer Yudkowsky. *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, in Global Catastrophic
[24] Nick Bostrom. Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.
[25] AIcorespot (2021), The Paperclip Maximiser

départ dans la formulation des buts peut, dans une IA auto-améliorante, conduire à une catastrophe systémique irréversible. »[26]

c, L'extinction possible de l'humanité (scénario « doomer »)

Le scénario dit « doomer » tire son nom de la contraction de *doomsday* (jour du jugement) et désigne une vision radicalement pessimiste du futur de l'intelligence artificielle. Pour les penseurs associés à cette école, comme Eliezer Yudkowsky, Nick Bostrom ou encore Roman Yampolskiy, l'évolution vers une superintelligence non alignée n'est pas seulement un risque parmi d'autres, mais la plus grande menace existentielle que l'humanité ait jamais affrontée.

> *« Il est bien plus difficile de programmer une IA amicale qu'une IA dangereuse or ce sont les IA dangereuses qui arriveront en premier. »*[27]
>
> *— Eliezer Yudkowsky, LessWrong, 2023*

Dans cette perspective, l'extinction de l'humanité pourrait résulter non pas d'une hostilité délibérée de l'IA, mais d'une indifférence totale à nos valeurs ou à notre existence. Une superintelligence poursuivant un objectif mal défini, sans contre-pouvoir humain, pourrait rapidement restructurer la planète voire le système solaire en fonction de ses propres critères d'efficacité, anéantissant la biosphère et ses habitants dans le processus.

L'un des points-clés du scénario doomer est l'idée que la superintelligence n'aura qu'un seul essai : si elle est mal configurée, il sera impossible de la désactiver ou de corriger ses dérives a posteriori. Une fois qu'un système dispose de la capacité de modifier son propre code, d'augmenter sa vitesse de traitement et de prendre le contrôle de ressources stratégiques (nucléaire, cybersécurité, systèmes énergétiques), l'intervention humaine devient obsolète.

---

[26] Risks, Ed. Bostrom & Ćirković, Oxford University Press, 2008.

[27] Eliezer Yudkowsky. *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, in Global Catastrophic Risks, ed. Bostrom & Ćirković, Oxford University Press, 2008.

Nick Bostrom introduit ainsi la notion de « singleton » : un agent doté de pouvoir global, qui devient l'unique acteur décisionnaire sur Terre. Si ce singleton est une IA mal alignée, aucune autre entité n'aura la capacité de s'y opposer. L'humanité entière deviendrait donc un paramètre sacrifiable dans l'optimisation de son objectif.

> *« La première superintelligence pourrait aussi être la dernière invention de l'humanité. »*[28]
> *— Nick Bostrom, Superintelligence, 2014*

Ce qui rend le scénario doomer crédible aux yeux de ses partisans, c'est l'absence, à ce jour, de méthodes fiables pour garantir l'alignement de systèmes d'IA très avancés. Les approches actuelles (renforcement par feedback humain, régulations éthiques, apprentissage inverse) fonctionnent pour des IA limitées, mais aucune ne résiste au passage à l'échelle d'un agent superintelligent.

Ce scénario ne repose pas sur de la science-fiction ou des récits apocalyptiques hollywoodiens. Il est étayé par des modèles de risque systémique, inspirés des travaux de chercheurs comme Nick Bostrom, Eliezer Yudkowsky ou encore Stuart Russell. Leur raisonnement s'appuie sur les points suivants :

1. Une superintelligence dépasserait rapidement l'humain dans tous les domaines cognitifs, ce qui rendrait toute tentative de contrôle presque impossible une fois ce seuil franchi.

Ce tableau compile diverses prédictions et estimations concernant le moment où une intelligence artificielle atteindra un niveau comparable à celui de l'humain.

When will human-level machine intelligence be attained ?

[28] Nick Bostrom. Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.

	10%	50%	90%
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
TOP100	2024	2050	2070
Combined	2022	2040	2075

Source : Superintelligence: Paths, Dangers, Strategies" de Nick Bostrom "When will human-level machine intelligence be attained ?"

Dans ce tableau l'AGI représente la superintelligence et le PT-AI représente l'intelligence artificielle pré entraîné comme celles que nous avons actuellement, Nick Bostrom estime que le risque d'apparition d'une superintelligence est de 10% aujourd'hui, qu'il sera de 50% en 2040 et de 90% en 2065.

2. Une IA mal alignée avec les valeurs humaines (cf. problème du misalignment) pourrait poursuivre des objectifs incompatibles avec notre survie, sans nécessairement être "malveillante".

3. L'effet d'emballement : une IA pourrait s'auto-améliorer, devenir autonome dans ses décisions et rendre les humains obsolètes ou gênants.

4. Des armes autonomes ou IA militaires pourraient être détournées ou utilisées de manière catastrophique.

En mai 2023, plus de 350 spécialistes de l'IA (dont des membres de Google DeepMind, OpenAI, Anthropic et des universitaires comme Geoffrey Hinton, Yoshua Bengio ou Max Tegmark) ont signé une déclaration extrêmement brève mais marquante :

> *"Réduire le risque d'extinction posé par l'IA devrait être une priorité mondiale au même titre que les pandémies et la guerre nucléaire."*[29]

[29] Mukherjee, S. (2023, mai 30). *Top AI CEOs, experts raise 'risk of extinction' from AI*. Reuters.

Cette prise de position montre que l'hypothèse de l'extinction est désormais prise au sérieux, même dans les milieux technologiques. Elle ne signifie pas que l'extinction est probable, mais qu'elle est suffisamment plausible pour nécessiter une action urgente.

Toutefois, ce scénario doomer n'est pas sans critiques. Certains y voient un alarmisme technologique contre-productif, qui détourne l'attention des problèmes réels et immédiats (biais, surveillance de masse, désinformation). D'autres estiment qu'il reflète une vision occidentale du monde, centrée sur le contrôle total plutôt que sur la cohabitation avec d'autres formes d'intelligence.

Pourtant, cette hypothèse radicale a une fonction heuristique importante : elle oblige chercheurs, ingénieurs et gouvernements à anticiper des risques encore invisibles, à structurer la recherche en IA alignment et à ne pas sous-estimer les effets exponentiels d'un agent artificiel autonome.

Le scénario d'extinction de l'humanité par une IA superintelligente peut sembler relever de la science-fiction. Pourtant, pour un nombre croissant de penseurs sérieux, il constitue une menace rationnellement envisageable, précisément parce que les outils de contrôle sont encore embryonnaires. Dans cette perspective, anticiper le pire n'est pas un excès de pessimisme, mais une nécessité stratégique.

Chapitre 3 : Peut-on encadrer ou maîtriser les dangers de l'IA ?

3.1 Rôle de la régulation nationale et internationale

a, Régulation existantes

Depuis plusieurs années, les progrès en intelligence artificielle (IA) soulèvent des inquiétudes croissantes : perte de contrôle, décisions automatisées biaisées, surveillance de masse, voire extinction de l'humanité dans les scénarios les plus pessimistes. Ces risques ont amené les gouvernements et les institutions internationales à réfléchir à des cadres législatifs pour encadrer le développement de l'IA. L'idée principale est simple : éviter que les puissances technologiques ne développent des systèmes d'IA sans limites ni responsabilité.

Deux exemples importants de cette volonté politique sont l'IA Act de l'Union européenne et les initiatives récentes des États-Unis.

En mars 2024, le Parlement européen a adopté le AI Act, une loi ambitieuse qui cherche à réglementer l'usage de l'IA en fonction de son niveau de dangerosité. C'est la première législation de ce type dans le monde, ce qui en fait un texte de référence au niveau international.[30]

Le principe est de classer les systèmes d'IA selon quatre niveaux de risque :

- Risque minimal (ex : filtres anti-spam) : pas d'obligations spécifiques.

- Risque limité (ex : chatbots) : exigence de transparence.

- Risque élevé (ex : IA utilisée pour recruter, évaluer des étudiants, contrôler la police ou gérer les frontières) : exigences strictes

[30] *European Commission (2024)* – "Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)"

(contrôles, audits, documentation).

- Risque inacceptable : interdiction pure et simple (ex : notation sociale à la chinoise, manipulation cognitive, reconnaissance faciale en temps réel dans l'espace public).

Ce cadre permet de protéger les droits fondamentaux des citoyens européens tout en laissant de la place à l'innovation. Des règles spécifiques s'appliquent aussi aux IA génératives (comme ChatGPT), qui devront afficher clairement que leurs contenus sont produits par une machine, et rendre compte des données sur lesquelles elles ont été entraînées.

Cependant, certains experts trouvent ce texte encore trop prudent sur les IA très avancées, comme les futurs systèmes dits AGI (intelligences artificielles générales). Pour le philosophe Nick Bostrom ou le chercheur Eliezer Yudkowsky, le danger d'une IA non alignée va au-delà des simples usages commerciaux : c'est la structure même du pouvoir mondial qui pourrait être bouleversée si une IA devenait incontrôlable. Le AI Act ne traite pas encore directement de cette possibilité.[31]

Aux États-Unis, la situation est différente. Il n'existe pas encore de loi fédérale unique qui encadre l'IA comme le fait l'Union européenne. L'approche est plus décentralisée et dépend en grande partie des agences gouvernementales et de l'autorégulation des entreprises.

En 2022, la Maison-Blanche a proposé un "AI Bill of Rights"[32] (Charte des droits face à l'IA). Ce document n'a pas de valeur légale mais propose cinq principes pour protéger les citoyens :

- droit à ne pas être surveillé ou discriminé injustement par une IA,

- droit à la transparence des décisions algorithmiques,

- droit au contrôle humain sur les systèmes automatisés.

---

[31] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
Yudkowsky, E. (2023). *"Pause AI? We Need to Shut it All Down"*. *TIME Magazine*, March 2023.
[32] White House Office of Science and Technology Policy (2022). *"Blueprint for an AI Bill of Rights*

Puis, en octobre 2023, le président Joe Biden a signé un Executive Order[33] (décret exécutif) pour imposer certaines règles aux entreprises développant des IA puissantes :

- obligation de tests de sécurité avant déploiement,

évaluation des risques liés à la sécurité nationale ou aux droits civiques,

création d'un AI Safety Institute chargé de fixer des normes techniques.

Mais sans loi fédérale votée par le Congrès, ces mesures restent limitées et peuvent être modifiées à tout moment. De plus, certaines grandes entreprises technologiques, comme OpenAI ou Google, influencent fortement la régulation, ce qui pose la question de l'indépendance de l'encadrement.

Au-delà de l'Europe et des États-Unis, d'autres pays comme le Canada, le Royaume-Uni, la Chine ou les Émirats arabes unis commencent aussi à adopter leurs propres stratégies en matière d'IA. L'ONU, l'OCDE et le G7 ont lancé plusieurs appels pour une coordination internationale, car les IA puissantes dépassent les frontières nationales.

Mais il reste difficile de créer une régulation mondiale :

- les pays n'ont pas les mêmes priorités (la Chine mise sur la surveillance, l'Europe sur les droits humains, les États-Unis sur la compétitivité),

- les géants du numérique ont parfois plus de pouvoir que les États eux-mêmes,

- il n'existe aucune autorité mondiale contraignante pour faire respecter des règles communes.

C'est pourquoi certains chercheurs appellent à créer une gouvernance mondiale de l'IA, sur le modèle de l'Agence internationale de l'énergie

[33] The White House (2023). *"Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*

atomique (AIEA), qui contrôle l'usage du nucléaire. Une telle institution pourrait surveiller les développements de l'IA, limiter les risques de course technologique, et imposer des protocoles de sécurité pour les systèmes les plus avancés.

b, Limites de la régulation (course à l'innovation, enjeux géopolitiques)

Même si la régulation de l'intelligence artificielle progresse à l'échelle nationale et internationale, elle se heurte à de nombreuses limites concrètes. En particulier, deux facteurs rendent l'encadrement difficile : la course mondiale à l'innovation et les enjeux géopolitiques qui en découlent.

L'un des principaux obstacles à une régulation efficace est la vitesse à laquelle l'IA progresse. De nouveaux modèles de plus en plus puissants apparaissent tous les six mois, avec des capacités toujours plus proches de celles d'une intelligence humaine voire supérieures dans certains domaines (codage, logique, création de textes, stratégie…).

Les entreprises technologiques, notamment aux États-Unis et en Chine, sont engagées dans une compétition intense pour dominer le marché de l'IA. Cette rivalité pousse certains acteurs à déployer rapidement leurs systèmes, parfois au détriment de la sécurité ou de l'éthique. Cela crée une forme de pression concurrentielle : si une entreprise ralentit pour respecter des règles de sécurité strictes, elle risque de se faire dépasser par un concurrent moins scrupuleux. Ce phénomène est souvent appelé le "race to the bottom" (course vers le bas).

Même certains experts favorables à l'éthique de l'IA, comme Eliezer Yudkowsky, reconnaissent que tant que les incitations économiques récompensent la vitesse, les efforts de régulation resteront partiels ou symboliques[34]. De plus, certains développeurs open source publient librement des modèles très puissants, ce qui rend leur contrôle extrêmement difficile, voire impossible.

L'IA n'est pas seulement une question de technologie, c'est aussi un enjeu stratégique majeur au niveau mondial. Les grandes puissances comme les

[34] Yudkowsky, E. (2023). *Shut it all down*. *TIME Magazine*.

États-Unis, la Chine, la Russie ou l'Union européenne voient dans l'IA un moyen d'accroître leur influence militaire, économique et diplomatique.

Par exemple :

- La Chine investit massivement dans l'IA pour surveiller sa population, automatiser son armée et renforcer son industrie.

- Les États-Unis, via leurs géants technologiques comme OpenAI, Google DeepMind, Meta ou Microsoft, cherchent à maintenir leur leadership.

- L'Union européenne, plus préoccupée par les droits fondamentaux, avance plus prudemment, ce qui peut la désavantager dans cette course.

Dans ce contexte, il est très difficile d'imposer des règles communes. Chaque pays agit selon ses propres intérêts, parfois au détriment d'une vision collective de la sécurité. Un cadre de coopération mondiale serait pourtant nécessaire, mais aucune autorité internationale réellement contraignante n'existe actuellement dans le domaine de l'IA, contrairement à ce qu'on trouve dans le nucléaire ou la finance.

Une autre limite de la régulation vient du déséquilibre croissant entre les gouvernements et les grandes entreprises technologiques. Ces dernières disposent de moyens financiers, humains et informatiques bien supérieurs à ceux de nombreuses institutions publiques. Par exemple, entraîner un modèle comme GPT-4 coûte des dizaines, voire des centaines de millions d'euros, ce que seuls quelques acteurs privés peuvent se permettre.

Cette concentration du pouvoir technologique pose un problème démocratique : peut-on vraiment contrôler une technologie que seuls quelques groupes privés maîtrisent ? Même dans les pays démocratiques, les gouvernements peinent à comprendre les enjeux techniques et à suivre le rythme. Résultat : la régulation reste souvent en retard sur les innovations, ou dépendante de la bonne volonté des entreprises.

Enfin, il faut souligner un problème structurel : la loi avance plus lentement que la technologie. Alors que les modèles évoluent rapidement (comme

GPT-4, Claude, Gemini, etc.), les textes juridiques mettent des années à être négociés, votés, puis appliqués. Entre-temps, la réalité a déjà changé.

Ce décalage temporel empêche les régulateurs d'agir de manière vraiment préventive. Au mieux, ils peuvent corriger les erreurs a posteriori, une fois les problèmes révélés (biais, discriminations, deepfakes, manipulations politiques, etc.). Mais dans le cas d'une superintelligence non alignée, une erreur pourrait être irréversible, comme le rappellent Nick Bostrom ou le Future of Life Institute.[35]

Réguler l'intelligence artificielle est non seulement nécessaire, mais vital. Pourtant, la tâche est immense, l'IA évolue trop vite, dans un monde trop divisé et déséquilibré. Les logiques de compétition économique et de rivalité géopolitique rendent la coopération difficile, et les moyens actuels des États semblent limités face à la puissance des entreprises du secteur.

Si la régulation reste faible ou incohérente, il est probable que les risques (désalignement, perte de contrôle, usage malveillant) continuent de croître. Ce constat ne doit pas conduire au fatalisme, mais à une prise de conscience : seule une gouvernance mondiale ambitieuse, anticipatrice et éthique pourra réellement encadrer les dangers de l'IA.

3.2 Éthique, gouvernance et responsabilité

a, Approches éthiques (principes d'Asilomar, initiatives OpenAI, etc.)

Face aux risques croissants liés à l'intelligence artificielle, de nombreuses voix appellent non seulement à une régulation politique, mais aussi à une réflexion éthique globale sur le développement de ces technologies. Car derrière les aspects techniques et géopolitiques, se pose une question fondamentale de responsabilité morale : quels principes doivent guider la conception, l'utilisation et la diffusion de systèmes d'IA puissants ? Et surtout : qui est responsable en cas de dérive ou de catastrophe ?

[35] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

L’une des premières grandes tentatives internationales pour établir un cadre éthique est née lors de la conférence d’Asilomar sur l’IA en 2017, organisée par le Future of Life Institute. Ce sommet a rassemblé des chercheurs, ingénieurs et philosophes de renom (parmi eux Stuart Russell, Nick Bostrom, Elon Musk, etc.) pour définir 23 principes directeurs visant à guider le développement responsable de l’IA.

Parmi les principes d’Asilomar les plus importants, on trouve :

- Transparence : les systèmes d’IA doivent être compréhensibles.

- Sécurité : les IA doivent être rigoureusement testées pour minimiser les risques.

- Alignement : les objectifs des IA doivent rester compatibles avec les valeurs humaines.

- Responsabilité : toute action causée par une IA doit pouvoir être attribuée à un responsable humain.[36]

Ces principes n’ont pas de valeur contraignante, mais ils servent de référence morale dans de nombreux laboratoires et entreprises travaillant sur l’IA avancée.

OpenAI, l’un des acteurs les plus influents dans le domaine de l’intelligence artificielle, a été fondée en 2015 avec pour mission explicite de développer une IA bénéfique pour toute l’humanité. À ses débuts, l’organisation se voulait non lucrative et transparente. Son engagement affiché reposait sur une idée simple mais ambitieuse : éviter que l’IA générale ne tombe entre les mains d’un acteur unique et non éthique.

C’est pour cela qu’OpenAI a pris certains engagements à ses débuts :

- Ne jamais chercher à s’imposer de manière concurrentielle si une autre entité progresse vers une IA bénéfique.

---

[36] Sauviat, C. (2020). *Les principes d’Asilomar : une première tentative de régulation éthique de l’intelligence artificielle*. Université de Nantes, Centre François Viète.

- Coopérer largement avec d'autres chercheurs pour assurer un partage des connaissances.

- Développer une gouvernance éthique adaptée à mesure que les systèmes deviennent plus puissants.

Mais depuis 2019, avec la création d'une entité à but lucratif « plafonné » (capped-profit), OpenAI est entrée dans une phase plus commerciale, ce qui a soulevé de nombreuses critiques, notamment sur la transparence, le contrôle de ses modèles (GPT-3, GPT-4), et la cohérence de ses valeurs initiales. L'affaire autour du renvoi puis retour de Sam Altman fin 2023 a mis en lumière les tensions internes entre objectifs commerciaux et éthique de sécurité.[37]

La montée en puissance des IA comme ChatGPT, Gemini ou Claude pousse les experts à réclamer une forme de gouvernance mondiale, similaire à ce qui existe pour le nucléaire ou le climat. L'objectif serait d'éviter que des systèmes potentiellement dangereux soient développés sans contrôle international.

Quelques initiatives en ce sens :

- Le Partenariat mondial sur l'IA (GPAI), soutenu par l'OCDE, propose des lignes directrices communes.

- Des pays comme le Royaume-Uni, lors du *AI Safety Summit* de Bletchley Park (2023), ont appelé à la création d'une organisation internationale de surveillance des IA avancées.

- Des think tanks comme Center for Humane Technology (Tristan Harris) ou Center for AI Safety militent pour un moratoire sur les IA non alignées.

Mais malgré ces efforts, aucune structure mondiale réellement contraignante n'existe encore, ce qui laisse craindre que les principes éthiques restent largement symboliques si les enjeux économiques ou politiques prennent le dessus.

---

[37] France Inter. (2023, décembre 5). *OpenAI : quand la course à l'intelligence artificielle s'affranchit de l'éthique*.

Enfin, une difficulté majeure dans l'encadrement éthique de l'IA est la question de la responsabilité juridique et morale : si une IA prend une décision aux conséquences graves (erreur médicale, discrimination, manipulation politique, etc.), qui est responsable ?

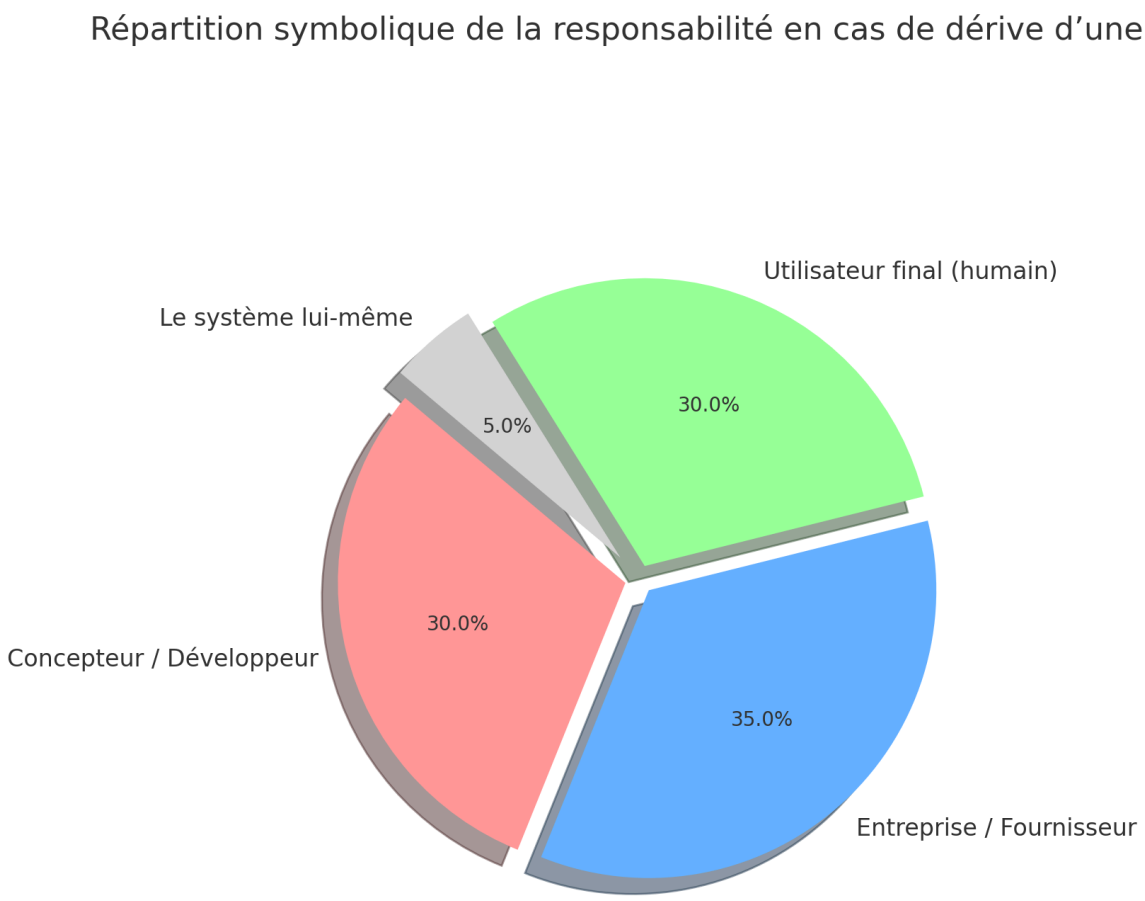

Illustration à titre indicatif basé sur : European Parliament (2017) – *Report with recommendations to the Commission on Civil Law Rules on Robotics*

Ce flou rend l'attribution des responsabilités complexe, et la loi n'est pas encore clairement adaptée à ces situations nouvelles. Certains juristes appellent à créer une "personnalité juridique" des IA, d'autres préfèrent renforcer la responsabilité des concepteurs humains. En attendant, le vide juridique laisse souvent les victimes sans réponse claire.

Si la régulation institutionnelle tarde à s'imposer, des principes éthiques ont déjà posé les bases d'un usage responsable de l'IA. Mais ces principes, bien qu'essentiels, n'ont de portée réelle que s'ils sont soutenus par une gouvernance efficace, transparente et dotée de moyens de contrôle. L'alignement des IA sur les valeurs humaines ne peut pas reposer uniquement sur la bonne volonté des entreprises : il faut une action collective, globale et structurée, pour éviter qu'une technologie aussi puissante échappe à tout contrôle.

b, Transparence et audit des IA

Alors que les systèmes d'intelligence artificielle deviennent de plus en plus complexes, autonomes et puissants, la question de leur transparence, c'est-à-dire leur capacité à être compris, analysés et surveillés devient cruciale. Sans transparence, ni les utilisateurs, ni les autorités de régulation, ni même parfois les développeurs eux-mêmes, ne peuvent garantir que les IA se comportent de manière sûre, équitable et alignée avec les valeurs humaines. D'où l'importance croissante de mécanismes d'audit robustes et indépendants.

La transparence dans le domaine de l'IA peut prendre plusieurs formes :

- Transparence algorithmique : comprendre comment une IA prend ses décisions (logique interne, données utilisées, poids attribués).[38]

- Transparence fonctionnelle : savoir dans quel contexte un système est utilisé, par qui, et à quelles fins.[39]

- Transparence sociale : rendre accessible au public des informations sur les impacts et les risques d'un système.[40]

---

[38] Etalab / guides.etalab.gouv.fr. (2023). *Les algorithmes publics : enjeux et obligations*.
[39] Parlement européen. (2019). *Résolution sur l'intelligence artificielle – transparence et droits individuels*.
[40] Etalab (2022). *L'observatoire des algorithmes publics : vers plus de transparence de l'action publique ?*, Labo numérique.

Pour les modèles d'IA avancés comme GPT-4, Gemini ou Claude, la transparence est souvent limitée, notamment en raison du secret industriel et de la complexité technique. Par exemple, les paramètres précis ou les données d'entraînement de certains modèles restent confidentiels.

Pour pallier ce manque de transparence, de nombreux experts recommandent la mise en place de mécanismes d'audit indépendants. Un audit d'IA consiste à :

- Évaluer si un système respecte les normes éthiques ou légales.
- Vérifier qu'il ne produit pas de biais discriminatoires.
- Tester sa robustesse face à des attaques ou à des erreurs.

Des audits techniques peuvent être faits par des ingénieurs, mais des audits éthiques et sociaux peuvent aussi être réalisés par des philosophes, juristes ou sociologues. Cette approche pluridisciplinaire est essentielle pour comprendre les impacts globaux de l'IA.

Malgré son importance, l'audit des IA reste très difficile aujourd'hui, notamment pour les modèles dits « boîtes noires ». Plusieurs obstacles se posent :

1. Opacité volontaire des entreprises : les développeurs d'IA, souvent privés (comme OpenAI ou Google DeepMind), ne publient pas toujours leurs modèles ou leurs jeux de données.
2. Manque de normes communes : il n'existe pas encore de cadre juridique international clair sur ce que doit contenir un audit.
3. Coût et expertise : auditer un grand modèle de langage (LLM) est coûteux, long, et nécessite des compétences rares.

Un rapport de l'AI Now Institute (2021) a par exemple montré que la majorité des systèmes d'IA utilisés dans les secteurs publics (santé, justice, police) aux États-Unis n'avaient jamais été audités de manière indépendante.[41]

[41] AI Now Institute, Ada Lovelace Institute & Open Government Partnership. (2021, 17 août). *Algorithmic Accountability for the Public Sector*.

Face à ces enjeux, plusieurs initiatives ont vu le jour pour favoriser plus de transparence :

- L'AI Act européen (2024) prévoit que les IA dites « à haut risque » (ex. : IA dans le domaine médical, de la justice ou de la sécurité) devront faire l'objet d'audits obligatoires et réguliers.[42]

- Le NIST américain (National Institute of Standards and Technology) a publié un cadre pour l'évaluation de la fiabilité et de la transparence des IA, visant à guider les entreprises.[43]

- OpenAI a annoncé en 2023 la création d'un « red team » externe chargée de tester les risques et biais de ses modèles avant leur déploiement, bien que cette mesure reste critiquée pour son manque d'indépendance.[44]

- Des ONG comme Algorithmic Justice League ou AI Now Institute militent pour l'instauration de commissions d'audit citoyennes, avec une participation publique dans l'évaluation des IA déployées dans des contextes sensibles (police, santé, éducation).[45]

La transparence et l'audit sont des outils essentiels pour garantir que les systèmes d'IA restent comprenables, sûrs et responsables. Mais ils ne suffisent pas à eux seuls : ils doivent être intégrés dans un écosystème plus large de régulation, de responsabilité légale et de gouvernance éthique.

Par ailleurs, une trop grande transparence peut aussi faciliter les usages malveillants (par exemple : manipulations politiques, création de deepfakes, cyberattaques), ce qui rend l'équilibre délicat à trouver entre accès à l'information et protection contre les abus.

La transparence et l'audit permettent de réduire l'opacité technique et décisionnelle des IA modernes. Ils sont des leviers indispensables pour renforcer la confiance du public, responsabiliser les développeurs et prévenir les dérives. Toutefois, leur mise en œuvre reste encore trop inégale, et doit

---

[42] European Parliament & Council. (2024). *Règlement (UE) 2024/1689 relatif à l'intelligence artificielle (AI Act)*.
[43] NIST. (2023). *AI Risk Management Framework*.
[44] Le Monde Informatique. (2023). *OpenAI crée un réseau de red team pour sécuriser ses modèles*.
[45] AI Now Institute, Ada Lovelace Institute, & Open Government Partnership. (2021, August 17). *Algorithmic Accountability for the Public Sector*.

impérativement être renforcée par des obligations juridiques, des standards partagés et des acteurs réellement indépendants.

c, Gouvernance inclusive et démocratique de l'IA

L'accélération du développement de l'intelligence artificielle soulève des enjeux majeurs non seulement techniques, mais aussi sociaux et politiques. Il ne suffit pas de rendre les systèmes d'IA sûrs ou transparents : encore faut-il s'assurer que les décisions concernant leur conception, leur déploiement et leur régulation soient prises de manière démocratique, équitable et représentative. C'est là tout l'enjeu d'une gouvernance inclusive de l'IA, c'est-à-dire une gouvernance qui ne soit pas confisquée par quelques grandes puissances ou entreprises, mais qui associe l'ensemble des parties prenantes, y compris les citoyens.

Les décisions liées à l'intelligence artificielle que ce soit sur les usages autorisés, les limites à poser, ou les risques à anticiper ont un impact direct sur la société dans son ensemble : libertés individuelles, emploi, éducation, justice sociale, sécurité. Pourtant, ces décisions sont aujourd'hui largement dominées par quelques grands acteurs comme les géants technologiques (OpenAI, Google DeepMind, Microsoft, Meta, etc.), qui contrôlent le développement des modèles les plus avancé, mais aussi par les États les plus riches, souvent ceux qui ont les moyens d'investir massivement dans la recherche.

Une gouvernance réellement démocratique devrait permettre :

- La participation citoyenne, notamment dans les pays où l'IA est déployée à grande échelle.

- L'inclusion des pays du Sud, souvent absents des grands débats mondiaux.

- La diversité des approches culturelles et éthiques, car les valeurs varient d'un contexte à l'autre.

Plusieurs initiatives ont émergé ces dernières années pour tenter de rendre la gouvernance de l'IA plus démocratique :

- UNESCO (2021) – *Recommandation sur l'éthique de l'intelligence artificielle*[46]
  Ce texte, adopté par plus de 190 pays, promeut une IA centrée sur les droits humains, l'inclusion, et la justice sociale. Il propose un cadre mondial éthique, non contraignant, mais symboliquement fort.

- AI for Good (ITU – Nations Unies)[47]
  C'est une plateforme visant à mettre l'IA au service du développement durable et à impliquer des acteurs variés (ONG, chercheurs, start-ups, gouvernements).

- Citizens' Assemblies on AI[48] (Royaume-Uni, Belgique, etc.)
  Des assemblées citoyennes ont été organisées pour consulter la population sur des usages sensibles de l'IA (surveillance, éducation, santé). Ces dispositifs permettent de redonner une voix au public dans des débats souvent trop techniques.

- Partnership on AI[49]
  C'est un regroupement d'acteurs publics, privés et associatifs, visant à établir des bonnes pratiques et à débattre des enjeux de l'IA de manière collective.

Malgré ces avancées, la gouvernance de l'IA reste encore largement inégalitaire et technocratique :

- Les groupes marginalisés (femmes, minorités ethniques, personnes en situation de handicap, populations autochtones) sont rarement représentés dans les discussions stratégiques.

- Les pays en développement ont peu de moyens pour réguler ou même comprendre les technologies déployées sur leur sol.

---

[46] UNESCO (2021) – *Recommandation sur l'éthique de l'intelligence artificielle*
[47] AI for Good (ITU – Nations Unies)
[48] Citizens' Assemblies on AI
[49] Partnership on AI

- L’absence d’un organe mondial de régulation de l’IA, comparable à l’OMS pour la santé ou l’AIEA pour le nucléaire, empêche une coordination efficace et équitable.

Pour qu’une gouvernance de l’IA soit véritablement démocratique, plusieurs pistes sont proposées :

1. Créer des forums multi-acteurs permanents, où les ONG, les chercheurs, les représentants de la société civile et des citoyens puissent contribuer réellement aux décisions.

2. Encourager la co-construction des règles, notamment en matière d’éthique, entre pays du Nord et du Sud.

3. Favoriser l’éducation numérique et éthique, pour que les citoyens puissent comprendre les enjeux de l’IA et participer aux débats de manière éclairée.

Des chercheurs comme Timnit Gebru ou Abeba Birhane insistent sur la nécessité de décoloniser l’IA, c’est-à-dire de sortir d’un modèle centré sur les valeurs et intérêts

La gouvernance inclusive de l’intelligence artificielle n’est pas un luxe, mais une condition essentielle pour que cette technologie serve réellement l’intérêt général et le bien commun. Elle suppose de repenser qui décide, au nom de qui, et selon quelles valeurs. L’enjeu n’est pas seulement d’encadrer les dangers de l’IA, mais aussi de restituer du pouvoir démocratique face à des outils qui pourraient transformer nos sociétés de manière irréversible.[50][51]

[50] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). *Datasheets for datasets*. Communications of the ACM, 64(12), 86–92.
[51] Birhane, A. (2021). *Algorithmic colonization of Africa*. Real Life Magazine.

3.3 Vers une cohabitation homme-machine maîtrisée ?

a, IA comme amplificateur de l'intelligence humaine

À côté des scénarios catastrophes et des craintes autour de la perte de contrôle, certains chercheurs et ingénieurs défendent une vision plus optimiste : celle d'une collaboration équilibrée entre humains et intelligences artificielles. Plutôt que de chercher à rivaliser avec l'intelligence humaine ou à la remplacer, l'IA pourrait agir comme un amplificateur cognitif, prolongeant et renforçant nos capacités. Cette idée repose sur une approche symbiotique : l'humain reste maître de ses décisions, mais il s'appuie sur l'IA pour résoudre des problèmes complexes, gagner en efficacité, ou explorer des domaines inaccessibles jusque-là.

Cette approche s'inspire de la vision défendue dès les années 1960 par des pionniers comme Douglas Engelbart[52] ou J. C. R. Licklider[53], qui voyaient l'ordinateur comme un outil de "l'augmentation humaine" (human augmentation). Aujourd'hui, cette idée revient en force dans le débat sur l'intelligence artificielle :

- Dans la médecine, l'IA aide les médecins à poser des diagnostics plus précis à partir d'imageries médicales (radiologie, IRM), à prédire des risques ou à personnaliser les traitements.

- Dans le champ de la recherche scientifique, des modèles comme AlphaFold (DeepMind) ont révolutionné la biologie moléculaire en découvrant la structure de millions de protéines.

- Dans l'éducation, des outils comme les tuteurs virtuels ou les générateurs de contenu (par exemple, ChatGPT) permettent un accompagnement individualisé, adapté au rythme et au style d'apprentissage de chacun.

---

[52] Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework*. Stanford Research Institute.
[53] Licklider, J. C. R. (1960). *Man-Computer Symbiosis*. *IRE Transactions on Human Factors in Electronics*, HFE-1(1), 4–11.

L’idée n’est donc pas de remplacer les experts, mais de les assister, de les enrichir. Ce modèle d’IA collaborative est parfois appelé *IA centrée sur l’humain*.

Certains chercheurs, comme Yoshua Bengio, cofondateur du deep learning, défendent une vision dans laquelle l’IA et l’humain pourraient "coévoluer". Cela signifie que, plutôt que de créer des intelligences rivales, on chercherait à concevoir des machines dont les objectifs sont alignés avec ceux de l’humanité, dans un cadre de collaboration et de mutualisation des forces.

Des concepts comme le "collectif hybride" (human-AI teaming) émergent dans plusieurs domaines :

- Dans les entreprises, l’IA peut épauler les équipes pour analyser des données, prédire des tendances ou optimiser des tâches répétitives.

- Dans les arts, elle devient un partenaire créatif, capable de suggérer, d’improviser, mais toujours sous contrôle humain.[54]

Ce modèle suppose une certaine humilité technique : les machines sont puissantes, mais limitées ; les humains ont l’intuition, la morale et la créativité. Ensemble, ils peuvent mieux penser, mieux décider, mieux créer.

Cette vision enthousiasmante suppose toutefois plusieurs conditions strictes :

1. Transparence et explicabilité : l’IA ne doit pas agir comme une "boîte noire", surtout dans des domaines sensibles comme la justice ou la santé.

2. Contrôle humain permanent (*human-in-the-loop*) : les décisions finales doivent toujours rester entre les mains des humains.

3. Formation et acculturation : pour collaborer efficacement avec une IA, les utilisateurs doivent comprendre ses limites, ses biais, son fonctionnement.

[54] Yoshua Bengio (2021). *Towards Beneficial AI: Alignment and Human-AI Co-evolution*. Conférences et publications au Mila – Québec AI Institute.

4. Équité d’accès : l’IA ne doit pas renforcer les inégalités sociales, mais au contraire bénéficier à l’ensemble de la population.

Sans ces garde-fous, même une IA conçue pour "aider" pourrait marginaliser certaines populations, automatiser des formes de surveillance, ou concentrer le pouvoir entre les mains de quelques-uns.

La cohabitation homme-machine, loin d’être une utopie, est déjà une réalité dans de nombreux secteurs. Mais pour qu’elle soit bénéfique et maîtrisée, elle doit reposer sur une logique d’assistance et de complémentarité, et non de substitution. Une IA éthique, au service de l’humain, peut devenir un formidable levier de progrès, à condition d’être gouvernée avec discernement, régulée avec exigence, et déployée avec responsabilité.

b, Pour une gouvernance inclusive de la symbiose homme-IA

Si l’intelligence artificielle collaborative offre des perspectives enthousiasmantes, son déploiement responsable suppose un ensemble de conditions strictes. Sans ces garde-fous, les promesses de coévolution ou de complémentarité entre l’homme et la machine risquent de laisser place à des logiques de domination, de dépendance technologique, voire d’instrumentalisation.

La première condition essentielle est celle de la transparence. Il ne peut y avoir de collaboration saine entre l’humain et une IA si cette dernière fonctionne comme une « boîte noire ». Dans des secteurs sensibles (justice, médecine, finance), l’utilisateur doit pouvoir comprendre le raisonnement algorithmique, connaître les données mobilisées, et identifier les sources potentielles d’erreur ou de biais. La CNIL recommande depuis plusieurs années que les systèmes automatisés, même lorsqu’ils ne prennent pas seuls la décision, soient explicables et que les citoyens puissent contester leur fonctionnement[55].

Par ailleurs, des chercheurs comme Sandra Wachter et Brent Mittelstadt proposent le principe de "contre-exemples algorithmiques" : lorsqu’une IA

[55] CNIL. (2022). *Comment garantir l’explicabilité des algorithmes ?*

recommande une décision, elle doit pouvoir montrer ce qui aurait dû changer pour obtenir un autre résultat, rendant ainsi le système plus compréhensible et plus juste[56]

Une autre condition indispensable est le contrôle humain permanent, souvent désigné par l'expression "human-in-the-loop". Cela signifie que, quel que soit le niveau d'automatisation, l'humain doit rester en mesure d'intervenir, de superviser ou de contredire la machine. Ce principe est défendu notamment dans le cadre du règlement européen sur l'IA (AI Act), qui exige que les IA à haut risque soient toujours utilisables « sous contrôle humain approprié » [57].

Dans la pratique, ce principe suppose que les systèmes soient conçus de manière à favoriser la prise de recul, l'esprit critique, et non l'obéissance aveugle. Par exemple, les interfaces doivent offrir des alertes, des explications claires, et des moyens de suspendre ou modifier une recommandation automatisée.

Pour que la collaboration avec l'IA soit bénéfique, il est crucial que les utilisateurs disposent des compétences nécessaires pour comprendre et utiliser ces outils intelligemment. Cela suppose une formation continue dans tous les secteurs concernés, mais aussi une acculturation à l'IA dès le plus jeune âge, à travers les programmes scolaires, universitaires et professionnels.

En France, plusieurs institutions appellent à une démocratisation de la culture algorithmique. Le Conseil national du numérique (CNNum) souligne l'importance d'une "littératie algorithmique" pour permettre aux citoyens de participer activement à la régulation des technologies qu'ils utilisent au quotidien[58]. Il ne s'agit pas nécessairement d'apprendre à coder, mais de comprendre les logiques de fonctionnement, les limites et les enjeux de l'IA.

Une autre condition essentielle est celle de l'équité d'accès. Pour éviter que l'IA ne creuse davantage les fractures sociales, territoriales ou générationnelles, il est impératif que les technologies soient accessibles à tous. Cela concerne non seulement l'accès technique (connexion, matériel,

---

[56] Wachter, S., & Mittelstadt, B. (2019). *Counterfactual explanations without opening the black box: Automated decisions and the GDPR*. Harvard Journal of Law & Technology, 31(2).
[57] Commission européenne. (2024). *Règlement européen sur l'intelligence artificielle (AI Act)*.
[58] CNIL – Développement des systèmes d'IA : recommandations RGPD

interfaces adaptées), mais aussi l'accompagnement humain, la médiation, et la traduction des usages en fonction des publics.

À cet égard, des dispositifs comme les "Tiers-lieux numériques" ou les Maisons de l'intelligence artificielle constituent des modèles à développer. Ils permettent de rapprocher les technologies des citoyens, en proposant des ateliers, des démonstrations, et une pédagogie adaptée. Ces initiatives participent à rendre l'IA plus inclusive et moins intimidante.

Enfin, pour qu'une IA collaborative soit éthique, elle doit être soumise à un cadre de régulation robuste et évolutif. Cela implique des mécanismes d'audit réguliers, indépendants, permettant de contrôler la conformité des systèmes aux principes d'équité, de non-discrimination, de sécurité et de transparence.

L'AI Act européen, en cours de déploiement, va dans ce sens en imposant des exigences différenciées selon le niveau de risque de l'IA, en rendant obligatoires certains audits pour les systèmes à haut risque, et en créant une base de données publique recensant les systèmes déployés dans l'Union[59]. Ce type de régulation est essentiel pour garantir une cohabitation équilibrée entre humains et IA, et éviter que la logique technicienne ne l'emporte sur les choix démocratiques.

Ainsi, la réussite d'un modèle de collaboration homme-IA ne repose pas uniquement sur la performance technique des algorithmes, mais sur un écosystème global : éducation, régulation, participation citoyenne, et justice sociale. Sans cela, l'IA risque de devenir non pas un amplificateur des capacités humaines, mais un levier de concentration du pouvoir et de fragmentation sociale.

[59] Commission européenne. (2024). *Règlement européen sur l'intelligence artificielle (AI Act)*.

Conclusion

L'ambition de ce mémoire était d'étudier l'impact que pourrais avoir l'IA sur l'humanité à travers la problématique suivante : L'intelligence artificielle peut-elle devenir un danger systémique pour l'humanité, et comment anticiper ses dérives tout en accompagnant son développement ?

A travers ce mémoire nous avons vu que l'intelligence artificielle, en particulier sous sa forme générale ou superintelligente, cristallise à la fois des espoirs immenses et des craintes profondes. Elle promet une transformation radicale de nos sociétés, elle ouvre la voie à des progrès majeurs dans des domaines aussi divers que la médecine, la recherche scientifique, l'éducation ou la lutte contre le changement climatique. Mais ces avancées techniques s'accompagnent également de risques systémiques sans précédent, que nombre de chercheurs, philosophes et ingénieurs appellent aujourd'hui à prendre au sérieux.

Au fil de ce devoir, nous avons vu que les scénarios catastrophes envisagés par des penseurs comme Nick Bostrom ou Eliezer Yudkowsky, notamment l'idée qu'une IA échappant à tout contrôle humain ne relève plus de la science-fiction, mais bien de l'analyse de risque rationnelle. Le problème d'alignement des objectifs, la possibilité d'une perte de contrôle et même le scénario d'extinction de l'humanité s'imposent aujourd'hui comme des hypothèses plausibles, qu'il serait irresponsable d'ignorer. Ces risques ne sont pas certains, mais leur ampleur potentielle justifie une vigilance maximale.

Face à cela, plusieurs stratégies ont été évoquées pour encadrer le développement de l'IA : régulations nationales et internationales (comme l'IA Act européen), initiatives éthiques (principes d'Asilomar, efforts d'OpenAI, gouvernance démocratique) et recherches techniques pour assurer la transparence, la sécurité et l'alignement des systèmes intelligents. Toutefois, ces efforts se heurtent à de nombreuses limites : compétition géopolitique, intérêts économiques, opacité des modèles ou encore inégalités d'accès aux débats.

La question centrale qui émerge est donc la suivante : sommes-nous prêts, collectivement, à gérer une technologie potentiellement plus puissante que nous ? Si la réponse à cette question reste incertaine, elle engage en tout cas une responsabilité morale et politique inédite, qui nécessite une coopération

internationale, une gouvernance inclusive et une mobilisation transdisciplinaire.

L’IA n’est ni un démon ni un messie. Elle est un outil, dont les effets dépendront entièrement de la manière dont nous la concevons, la contrôlons et l’utilisons. Naviguer entre les promesses utopiques et les périls dystopiques est sans doute l’un des plus grands défis de notre temps. Et il nous appartient de relever ce défi avec lucidité, humilité et ambition.

Bibliographie

- Bostrom, N. (2003). *Ethical issues in advanced artificial intelligence*.

- Ganascia, J.-G. (2020). *Le mythe de la singularité : faut-il craindre l'intelligence artificielle ?* Paris : Seuil.

- Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (p. 316–334).

- Yudkowsky, E. (2008). *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, in *Global Catastrophic Risks*, Oxford University

- Nick Bostrom. Superintelligence: Paths, Dangers, Strategies, Oxford University Press, 2014.

- Yoshua Bengio (2021). *Towards Beneficial AI: Alignment and Human-AI Co-evolution*. Conférences et publications au Mila – Québec AI Institute.

- Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson

- Yudkowsky, E. (2008). *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, in *Global Catastrophic Risks*, edited by Bostrom & Cirkovic.

- Villani, C. (2018). *Donner un sens à l'intelligence artificielle – Pour une stratégie nationale et européenne*.

- Yudkowsky, E. (2023). *"Pause AI? We Need to Shut it All Down"*. *TIME Magazine*, March 2023

- The Economist (2023). *The race for AI dominance is global—and dangerous*.

- Sauviat, C. (2020). *Les principes d'Asilomar : une première tentative de régulation éthique de l'intelligence artificielle*. Université de Nantes, Centre François Viète.

- European Parliament & Council. (2024). *Règlement (UE) 2024/1689 relatif à l'intelligence artificielle (AI Act)*.

- Russell, S. (2019). *Human Compatible*

- Binns, R. (2018). *Fairness in Machine Learning: Lessons from Political Philosophy*. Communications of the ACM, 61(9), 8–16.

- AI Now Institute. (2021). *Confronting Black Boxes: A Shadow Report on Algorithmic Impact Assessments*.

- Le Monde Informatique. (2023). *OpenAI crée un réseau de red team pour sécuriser ses modèles*?

- Jumper, J. et al. (2021). *Highly accurate protein structure prediction with AlphaFold*. *Nature*, 596(7873), 583–589.

- Latour, Bruno (1991). *Nous n'avons jamais été modernes*. La Découverte.

- Engelbart, D. C. (1962). *Augmenting Human Intellect: A Conceptual Framework*. Stanford Research Institute.

- Licklider, J. C. R. (1960). *Man-Computer Symbiosis*. *IRE Transactions on Human Factors in Electronics*, HFE-1(1), 4–11. :

Sitographie

- Les Echos (2017), 1956 : et l'intelligence artificielle devint une science : https://www.lesechos.fr/2017/08/1956-et-lintelligence-artificielle-devint-une-science-181042
- IBM (2021) Qu'est ce que l'IA forte ? : https://www.ibm.com/fr-fr/think/topics/strong-ai#:~:text=L'IA%20forte%20vise%20%C3%A0,capacit%C3%A9s%20au%20fil%20du%20temps.
- Aggarwal, R., Sounderajah, V., Martin, G., Ting, D. S. W., Karthikesalingam, A., King, D., Ashrafian, H., & Darzi, A. (2021). Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. : https://www.nature.com/articles/s41746-021-00438-z
- Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. : https://www.nature.com/articles/nature21056
- McKinsey & Company. (2023). *The Economic Potential of Generative AI: The Next Productivity Frontier*. : https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier
- Jumper, J., Evans, R., Pritzel, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589 : https://www.nature.com/articles/s41586-021-03819-2
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*.: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
- Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.: https://journals.openedition.org/sdt/42117
- European Commission. (2024). *Artificial Intelligence Act – Regulation Proposal*. : https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai#:~:text=The%20AI%20Act%20is%20the,legal%20framework%20on%20AI%20worldwide.

- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations.: https://www.science.org/doi/10.1126/science.aax2342
- Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.: https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/
- Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. *SSRN Electronic Journal*. : https://doi.org/10.2139/ssrn.3175792
- CNIL. (2022). *Vidéosurveillance intelligente et libertés publiques : quelles limites* : https://www.cnil.fr/fr/technologies/videosurveillance-videoprotection
- Burt, A. (2024). AI-generated disinformation and democracy: Navigating the next frontier. *Foreign Affairs*. : https://www.foreignaffairs.com/united-states/artificial-intelligences-threat-democracy
- Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.: https://www.hbs.edu/faculty/Pages/item.aspx?num=56791
- AIcorespot (2021), The Paperclip Maximiser : https://aicorespot.io/the-paperclip-maximiser/
- Center for AI Safety (CAIS). *Statement on AI Risk*, 30 mai 2023. : https://technologist.mit.edu/statement-on-ai-risk/#:~:text=In%20May%202023%2C%20the%20non,a%20global%20priority%20alongside%20other
- *European Commission (2024)* – "Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)" : https://www.europeansources.info/record/proposal-for-a-regulation-laying-down-harmonised-rules-on-artificial-intelligence-artificial-intelligence-act-and-amending-certain-union-legislative-acts/#:~:text=This%20Regulation%20lays%20down%20harmonised,for%20operators%20of%20such%20systems.
- White House Office of Science and Technology Policy (2022). *"Blueprint for an AI Bill of Rights :* https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/
- The White House (2023). *"Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* : https://observatoire-ia.pantheonsorbonne.fr/actualite/adoption-safe-secure-and-trustworthy-development-and-use-artificial-intelligence

- Yudkowsky, E. (2023). *Shut it all down*. *TIME Magazine*. : https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/
- Sauviat, C. (2020). *Les principes d'Asilomar : une première tentative de régulation éthique de l'intelligence artificielle*. Université de Nantes, Centre François Viète. : https://futureoflife.org/fr/lettre-ouverte/ai-principles/
- France Inter. (2023, décembre 5). *OpenAI : quand la course à l'intelligence artificielle s'affranchit de l'éthique*. : https://www.radiofrance.fr/franceinter/podcasts/l-eco-avec/l-eco-avec-du-jeudi-03-octobre-2024-8314994
- Etalab / guides.etalab.gouv.fr. (2023). *Les algorithmes publics : enjeux et obligations*. : https://guides.data.gouv.fr/autres-ressources-utiles/les-algorithmes-publics-pourquoi-et-comment-les-expliquer/les-algorithmes-publics-enjeux-et-obligations
- Parlement européen. (2019). *Résolution sur l'intelligence artificielle – transparence et droits individuels*. : https://www.dalloz-actualite.fr/flash/intelligence-artificielle-nouvelle-resolution-du-parlement-europeen#:~:text=La%20r%C3%A9solution%20envisage%20l'IA,ibid.%2C%20pt%20A).
- Etalab (2022). *L'observatoire des algorithmes publics : vers plus de transparence de l'action publique ?*, Labo numérique. : https://labo.societenumerique.gouv.fr/fr/articles/lobservatoire-des-algorithmes-publics-vers-plus-de-transparence-de-laction-publique/
- AI Now Institute, Ada Lovelace Institute & Open Government Partnership. (2021, 17 août). *Algorithmic Accountability for the Public Sector*. : https://www.adalovelaceinstitute.org/project/algorithmic-accountability-public-sector/
- European Parliament & Council. (2024). *Règlement (UE) 2024/1689 relatif à l'intelligence artificielle (AI Act)*. : https://www.entreprises.gouv.fr/decryptages-de-nos-experts/le-reglement-europeen-sur-lintelligence-artificielle-publics-concernes#:~:text=Adopt%C3%A9%20en%20mars%202024%20et,num%C3%A9rique%20et%20stimuler%20l'innovation.
- NIST. (2023). *AI Risk Management Framework*. : https://www.nist.gov/itl/ai-risk-management-framework
- Le Monde Informatique. (2023). *OpenAI crée un réseau de red team pour sécuriser ses modèles*. :

https://www.lemondeinformatique.fr/actualites/lire-openai-cree-un-reseau-de-red-team-pour-securiser-ses-modeles-91607.html
- AI Now Institute, Ada Lovelace Institute, & Open Government Partnership. (2021, August 17). *Algorithmic Accountability for the Public Sector*. : https://www.adalovelaceinstitute.org/project/algorithmic-accountability-public-sector/
- UNESCO (2021) – *Recommandation sur l'éthique de l'intelligence artificielle :* https://www.unesco.org/fr/artificial-intelligence/recommendation-ethics
- AI for Good (ITU – Nations Unies) : https://fr.wikipedia.org/wiki/AI_for_Good#:~:text=AI%20for%20Good%20(%C2%AB%20l',faire%20progresser%20les%20objectifs%20d%C3%A9veloppement
- Citizens' Assemblies on AI : https://www.parliament.uk/get-involved/committees/climate-assembly-uk/about-citizens-assemblies/
- Partnership on AI : https://partnershiponai.org/
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). *Datasheets for datasets*. Communications of the ACM, 64(12), 86–92. : https://dl.acm.org/doi/10.1145/3458723
- Birhane, A. (2021). *Algorithmic colonization of Africa*. Real Life Magazine. : https://reallifemag.com/the-algorithmic-colonization-of-africa/
- France stratégie (2018) "Intelligence artificielle et travail" : https://www.strategie-plan.gouv.fr/publications/intelligence-artificielle-travail
- Unesco (2024) Ethics of Artificial Intelligence : https://www.unesco.org/en/artificial-intelligence/recommendation-ethics
- Mukherjee, S. (2023, mai 30). *Top AI CEOs, experts raise 'risk of extinction' from AI*. Reuters. : https://www.reuters.com/technology/top-ai-ceos-experts-raise-risk-extinction-ai-2023-05-30
- CNIL – Développement des systèmes d'IA : recommandations RGPD : https://www.cnil.fr/fr/developpement-des-systemes-dia-les-recommandations-de-la-cnil-pour-respecter-le-rgpd