# Introduction

*Aux grandes promesses qu'apporte l'IA, doit répondre*
*une grande responsabilité humaine*

Le monde digital se caractérise par son instantanéité, sa densité d'informations, son omniprésence, en contraste avec le monde concret des choses. Désormais, avec la multiplication des moyens de connexion, la baisse des coûts des technologies, les nouvelles capacités de collecte et de traitement algorithmique de la donnée, on s'aperçoit que nous pouvons faire communiquer des éléments de notre environnement jusqu'à présent muets. On assiste au développement multiforme des Nouvelles technologies de l'information et la communication (NTIC), illustré par l'émergence des technologies associées aux *Big Data*, aux objets connectés, aux algorithmes, aux nanotechnologies, biotechnologies, informatique et sciences cognitives (NBIC), à la *blockchain*, à l'intelligence artificielle (IA), à la réalité virtuelle et augmentée, voire à l'informatique quantique. L'IA se développe à un rythme extrêmement rapide. Nous devrions nous attendre à voir des changements significatifs dans notre société à mesure que les systèmes d'IA s'intègrent à de nombreux aspects de nos vies.

Ce phénomène numérique multifacette est en train de réunir les différents univers en ajoutant aux objets associés à ces NTIC la vitesse, l'intelligence et l'ubiquité propres au numérique. Les développements majeurs relatifs à l'IA dans la santé, les véhicules autonomes, la cybersécurité, l'éducation, les robots domestiques et de services améliorent chaque jour la qualité et le confort de nos vies. Désormais, l'IA est fondamentale pour relever bon nombre des grands challenges auxquels l'humanité est confrontée, tels que le changement climatique, la santé et le bien-être dans le monde, la valorisation des ressources naturelles ou des dispositifs légaux et démocratiques fiables et pérennes. Cette technologie bouleverse alors nos modes de vie, de consommation, de fonctionnement et de travail. Cela s'illustre par une coupure avec le passé dans

la relation et le lien que chacun a avec son prochain. Dès lors, ces interactions obligent le système à repenser chaque activité humaine. C'est le commencement d'une révolution silencieuse, mais bien présente qui se passe bien sous nos yeux. Une nouvelle ère de changement et de disruption où la survie passe inéluctablement par de la réactivité, de l'adaptabilité, de la créativité et donc par de l'innovation.

Dès lors, ce contexte technoscientifique est propice au développement d'un mouvement culturel et intellectuel international de plus en plus important, à savoir le transhumanisme, dont l'objectif est d'améliorer les caractéristiques physiques et mentales de l'être humain en s'appuyant sur les biotechnologies et les autres technologies émergentes. Ce courant de pensée considère que certains états de la condition humaine comme la maladie, le handicap, la douleur, le vieillissement ou la mort ne sont pas une fatalité en soi et peuvent être corrigés, voire supprimer.

Ainsi, les révolutions technologiques ont permis un changement d'échelle dans l'exploitation des données numériques, notamment dans le domaine de la génétique. Elles peuvent être produites en grande quantité, de façon de plus en plus précise et conservées sur un temps indéfini. On observe que les progrès en matière informatique ont rendu possible, grâce à la création de programmes spécifiques, l'interopérabilité des bases de données permettant par la même la fusion de données provenant de sources diverses et multiples. À cela, on peut rajouter le développement des nouveaux modes d'accès aux *data*, en particulier à travers la multiplication des sources de données en tout genre. Le *crowdsourcing*[1] devient l'un des nouveaux dispositifs permettant un accès facilité, en temps réel, aux données numériques afin de développer des recherches (Khare *et al.* 2015).

> TRAITEMENT ALGORITHMIQUE.–
>
> Un traitement algorithme est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre un problème ou d'obtenir un résultat. On retrouve aujourd'hui des algorithmes dans de nombreuses applications telles que le fonctionnement des ordinateurs, la cryptographie, le routage d'informations, la planification et l'utilisation optimale des ressources, le traitement d'images, le traitement de texte, etc. Un algorithme est une méthode générale pour résoudre un ensemble de problèmes. Il est dit correct lorsque, pour chaque instance du problème, il se termine en produisant la bonne sortie, c'est-à-dire qu'il résout le problème posé.

1. En France, on parle de production participative définie selon la Commission générale de terminologie et de néologie (2014) comme le « Mode de réalisation d'un projet ou d'un produit faisant appel aux contributions d'un grand nombre de personnes, généralement des internautes ». *JORF*, 0179(91), 12995.

*BIG DATA.–*

Les Big Data, littéralement en français « grosses données », ou mégadonnées, parfois appelées données massives, désignent des ensembles de données qui deviennent tellement volumineux qu'ils en deviennent difficiles à exploiter avec des outils classiques de gestion de base de données ou de gestion de l'information. Le terme de Big Data désigne une nouvelle discipline qui se situe au croisement de plusieurs secteurs tels que les technologies, les statistiques, les bases de données et les métiers (marketing, finance, santé, ressources humaines, etc.). Ce phénomène peut être défini selon 7 caractéristiques, les 7V (volume, variété, vitesse, véracité, visualisation, volatilité, valeur).

*BLOCKCHAIN.–*

En français « chaîne de blocs » informatique, protégée contre toute modification, dont chacun contient l'identificateur de son prédécesseur. La *blockchain* enregistre un ensemble de données comme une date, une signature cryptographique associée à l'expéditeur et tout un ensemble d'autres éléments spécifiques. Tous ces échanges sont traçables, consultables et téléchargeables gratuitement sur Internet, par toute personne qui souhaite vérifier la validité et la non-falsification de la base de données en temps réel. L'intérêt majeur de ce dispositif est de pouvoir stocker avec chaque transaction une preuve d'information, afin de pouvoir prouver ultérieurement et à chaque instant l'existence et le contenu de cette information originale à un moment donné. Sa mission est donc de créer de la confiance en protocolisant un actif numérique ou une base de données en le rendant auditable.

*CROWDSOURCING.–*

Pratique qui correspond à faire appel au grand public ou aux consommateurs pour proposer et créer des éléments de la politique marketing (choix de marque, création du slogan, création de vidéo, idéation/cocréation produit, etc.) ou même pour réaliser des prestations marketing. Dans le cadre du *crowdsourcing*, les prestataires professionnels ou amateurs peuvent alors être récompensés, rémunérés ou parfois uniquement valorisés lorsque leurs créations sont choisies par l'annonceur ou parfois simplement pour leur effort de participation. Le *crowdsourcing* s'est surtout développé avec Internet qui favorise la sollicitation des consommateurs ou freelances par le biais de plateformes spécialisées.

Dès lors, l'IA apparait comme une évolution essentielle dans le traitement de l'information numérique. Elle représente la partie de l'informatique consacrée à l'automatisation de comportements intelligents. Cette approche constitue la recherche de moyens susceptibles de doter les systèmes informatiques de capacités intellectuelles

comparables à celles des êtres humains. L'IA doit être capable : d'apprendre, de s'adapter et de modifier son comportement.

L'idée d'élaborer des machines autonomes remonte vraisemblablement à l'Antiquité grecque avec les automates construits par Héphaïstos relatés notamment dans l'Iliade (Marcinkowski et Wilgaux 2004). Pour Brian Krzanich, président-directeur général d'Intel (premier fabricant mondial de microprocesseurs), l'IA n'est pas seulement le prochain raz-de-marée de l'informatique, mais c'est également le prochain virage majeur dans l'histoire de l'humanité. Elle ne représente pas un programme informatique classique : elle s'éduque plus qu'elle ne se programme. Force est de constater que le procès intenté à l'IA a mêlé fantasmes, science-fiction, et futurologie à long terme, en oubliant même les définitions de base de cette dernière.

Ainsi, le concept de l'IA[2] est d'élaborer des programmes informatiques capables d'effectuer des tâches accomplies par des humains demandant un apprentissage, une organisation de la mémoire et un raisonnement. L'objectif est de donner des notions de rationalité, des fonctions de raisonnement et de perception (par exemple, visuelle) pour commander un robot dans un milieu qui lui est inconnu. Son engouement est associé à de nouvelles techniques, comme le *Deep Learning*, qui donnent la possibilité à un programme d'apprendre à représenter le monde grâce à un réseau de neurones virtuels qui réalisent chacun des calculs élémentaires, de manière similaire à notre cerveau.

> *Deep Learning*.–
>
> Cela fait plus de vingt années que l'on utilise ce système algorithmique pour différentes actions sous la forme de réseaux de neurones, en particulier pour faire de l'« apprentissage ». Un neurone représente une fonction simple qui prend différentes entrées et calcule son résultat qu'elle envoie à différentes sorties. Ces neurones sont principalement structurés et organisés en couches. La première couche emploie des données quasi brutes et la dernière va générer un résultat. Plus le nombre de couches est important et plus la capacité d'apprentissage et de performance sera accrue. On peut prendre l'exemple de la reconnaissance de caractères à partir de l'écriture manuscrite. La première couche va prendre en compte la totalité des pixels constituant un caractère écrit, par exemple une lettre ou un chiffre, et chaque neurone aura quelques pixels à analyser. La dernière couche indiquera « c'est un T avec une probabilité de 0,8 » ou « c'est un I avec une probabilité de 0,3 ». On effectue une opération de rétropropagation à partir du résultat final pour remodifier les paramètres de chaque neurone.

2. La norme ISO 2382-28:1995 définit l'intelligence artificielle comme la « capacité d'une unité fonctionnelle à exécuter des fonctions généralement associées à l'intelligence humaine, telles que le raisonnement et l'apprentissage ».

La machine est programmée pour « apprendre à apprendre ». L'IA n'existe pas pour remplacer l'homme, mais pour compléter, assister, optimiser et étendre les capacités humaines. On distingue deux types d'IA :

– IA faible : dont l'objectif est de débarrasser l'homme de tâches fastidieuses, à l'aide d'un programme informatique reproduisant un comportement spécifique. Cette IA est rapide à programmer, très performante, mais sans aucune possibilité d'évolution. C'est l'IA actuelle ;

– IA forte : dont l'objectif est de constituer des systèmes de plus en plus autonomes, ou des algorithmes capables de résoudre des problèmes. C'est l'approche la plus similaire du comportement humain. Cette IA apprend ou s'adapte très facilement. Grâce à des boucles de rétroactions algorithmiques, la machine peut modifier ses paramètres internes employés pour gérer la représentation de chaque strate à partir de la représentation de la strate précédente. Ces strates de fonctionnalités sont apprises par la machine elle-même et non par l'homme. À partir de ce postulat, on peut dire que la machine devient autonome et intelligente, en construisant ses propres structures de « computérisation » et en s'appuyant sur des décisions axiomatiques. C'est l'IA à venir qui devrait se développer dans une dizaine d'années.

IA FAIBLE.–

L'IA faible (*Weak Artificial Intelligence*) ou étroite (*Narrow Artificial Intelligence*) simule des facultés cognitives spécifiques comme la compréhension du langage naturel, la reconnaissance de la parole, ou la conduite automobile. Elle n'effectue que des tâches pour lesquelles elle est programmée. Elle est donc très spécialisée. C'est une machine pour laquelle le monde physique revêt un caractère un peu énigmatique, voire fantomatique, si tant est qu'elle le perçoive. Elle n'a même aucune conscience du temps passé. Cette IA est inintelligente et fonctionne uniquement à partir de scénarii préétablis par les concepteurs et les développeurs.

IA FORTE.–

L'IA générale (*Artificial General Intelligence*) ou forte (*Strong Artificial Intelligence*) a des capacités de raisonnement analogues – et même supérieures – à celles des êtres humains. La machine est dotée de capacité non limitée à certains domaines ou à certaines tâches. Elle reproduit ou vise à restituer un esprit, voire une conscience, sur une machine. C'est-à-dire une machine évolutive disposant de son propre raisonnement et d'une conscience, capable notamment d'élaborer en toute indépendance des stratégies et/ou des décisions qui dépassent l'humain afin de le comprendre pour l'aider (dans le meilleur des cas) ou de le tromper voire le détruire (dans le pire des cas).

D'un point de vue général, l'IA peut s'illustrer comme étant une matrice algorithmique qui a pour objectif d'optimiser les décisions « justement ou froidement ». Naturellement, la morale ou l'équité de ce jugement n'est pas prédéfinie, mais dépend, d'une part, de la manière avec laquelle sont apprises les règles (le critère objectif qui a été choisi), et d'autre part, de la façon avec laquelle a été construit l'échantillon d'apprentissage. Le choix des règles mathématiques permettant de créer le modèle est primordial. À l'instar du fonctionnement humain qui analyse une situation avant de changer son comportement, l'IA permet à la machine d'apprendre de ses propres résultats pour modifier sa programmation. Cette technologie existe déjà dans de multiples applications comme sur nos smartphones, elle devrait prochainement s'étendre à tous les secteurs de la vie quotidienne : de la médecine à la voiture autonome, en passant par la création artistique, la grande distribution ou la lutte contre la criminalité et le terrorisme. Le *Machine Learning* n'offre pas uniquement l'opportunité d'exploiter automatiquement de grandes quantités de données et d'identifier des habitudes dans le comportement des consommateurs. Désormais, nous pouvons également actionner ces données.

> *Machine Learning*.–
>
> L'apprentissage automatique ou apprentissage statistique (en anglais *Machine Learning*) concerne la conception, l'analyse, le développement et l'implémentation de méthodes permettant à une machine (au sens large) d'évoluer par un processus systématique, et ainsi de remplir des tâches difficiles ou impossibles à remplir par des moyens algorithmiques plus classiques. Les algorithmes utilisés permettent, dans une certaine mesure, à un système piloté par ordinateur (un robot éventuellement), ou assisté par ordinateur, d'adapter ses analyses et ses comportements en réponse, en se fondant sur l'analyse de données empiriques provenant d'une base de données ou de capteurs.

À notre sens, adopter la méthode *Machine Learning* n'est plus seulement une utilité, mais plutôt une nécessité. Ainsi, à l'aune de la transition numérique et cette « guerre des intelligences » (Alexandre 2017), les entreprises vont être la cible d'une importante transformation et investir dans des applications de l'IA afin de pouvoir, en particulier :

– augmenter l'expertise humaine *via* les programmes d'assistance virtuels ;

– optimiser certains produits et services ;

– apporter de nouvelles perspectives dans la R&D *via* l'évolution des systèmes autoapprenants.

Dès lors, l'IA est porteuse de grandes promesses, mais également de fortes angoisses, d'aléas et de dangers qu'il convient de corriger, voire de supprimer, afin d'en

garantir un déploiement conforme au cadre légal, à des valeurs morales et principes éthiques, et au bien commun. Les conflits en question peuvent être très variés. En effet, les machines telles que les assistants robotiques ignorent finalement les concepts du bien et du mal. Il faut tout leur apprendre. Les voitures autonomes sont susceptibles de nous impliquer dans un accident ou des situations dangereuses. Certains agents conversationnels peuvent insulter ou donner de mauvais conseils aux individus, et ne pas être bienveillants envers eux.

Ainsi, même si aujourd'hui, des préconisations éthiques impactent faiblement le périmètre fonctionnel d'une IA et qu'elles introduisent un niveau de complexité supplémentaire dans la conception des systèmes autoapprenants, il devient primordial, à l'avenir, de concevoir et d'intégrer des critères éthiques autour des projets digitaux relatifs à l'IA.

Plusieurs normes portant sur les systèmes algorithmiques, la transparence, la vie privée, la confidentialité, l'impartialité et plus généralement sur l'élaboration de systèmes éthiques ont été élaborées par des associations professionnelles comme : l'IEEE (Institut des ingénieurs électriciens et électroniciens) et par l'IETF (*Internet Engineering Task Force*)[3].

À cela, on peut rajouter les documents centrés sur les principes éthiques relatifs à l'IA, tels que :

– les principes Asilomar relatifs à l'IA, élaborés lors du *Future of Life Institute*, en collaboration avec les participants à la conférence de haut niveau d'Asilomar, de janvier 2017 ;

– les principes éthiques proposés dans la Déclaration sur l'intelligence artificielle, la robotique et les systèmes autonomes, publiée par le Groupe européen d'éthique des sciences et des nouvelles technologies de la Commission européenne, en mars 2018 ;

– les principes énoncés par le groupe d'experts de haut niveau sur l'IA, *via* un rapport intitulé : « Ethics Guidelines for Trustworthy AI », pour la Commission européenne, le 18 décembre 2018 ;

– la déclaration de Montréal pour une IA, élaborée à l'Université de Montréal, à la suite du forum sur le développement socialement responsable des IA, de novembre 2017 ;

3. IEEE P7000 : *Model Process for Addressing Ethical Concerns During System Design* ; IEEE P7001 : *Transparency of Autonomous Systems* ; IEEE P7002 : *Data Privacy Process* ; IEEE P7003 : *Algorithmic Bias Considerations* ; IETF *Research into Human Rights Protocol Considerations draft*.

– les bonnes pratiques en matière d'IA du *Partnership on AI*, en 2018, cette organisation multipartite – composée des universitaires, des chercheurs, des organisations de la société civile et des entreprises qui construisent et utilisent l'IA – a étudié et formulé les meilleures pratiques en matière de technologies d'intelligence artificielle. L'objectif était d'améliorer la compréhension du public sur l'IA et pour servir de plateforme ouverte de discussion et d'engagement sur l'intelligence artificielle et ses influences sur les individus et la société ;

– les « cinq principes fondamentaux pour un code sur l'IA », proposés au paragraphe 417 de la UK House du rapport du Lords Artificial Intelligence Committee, « AI au Royaume-Uni : prêt, disposé et capable ? », publié en avril 2018 ;

– la charte éthique rédigée par la Commission européenne, pour l'efficacité de la justice (CEPEJ) sur l'utilisation de l'IA dans les systèmes judiciaires et leur environnement. Il s'agit du premier texte européen énonçant des principes éthiques relatifs à l'usage de l'IA dans les systèmes judiciaires (voir annexe 1) ;

– les principes éthiques de Luciano Floridi *et al*. dans leur article intitulé : « Taken together, they yield 47 principles. An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommandations », paru en décembre 2018 dans le *Minds and Machines* ;

– le rapport de l'OPECST (Office parlementaire d'évaluation des choix scientifiques et technologiques) (De Ganay et Gillot 2017) ;

– les six recommandations pratiques du rapport de la CNIL (Commission nationale de l'information et des libertés)[4] sur les enjeux éthiques des algorithmes et de l'intelligence artificielle, rédigé en 2017 (voir annexe 2) ;

– le rapport publié par le député Cédric Villani (2018) sur l'intelligence artificielle ;

– la déclaration sur l'éthique et la protection des données dans le secteur de l'intelligence artificielle, lors de la *40e Conférence internationale des commissaires à la protection des données et de la vie privée* (ICDPPC), mardi 23 octobre 2018, à Bruxelles ;

– les sept lignes directrices[5] élaborées par le groupe d'experts européens de haut niveau sur l'IA, publiées le 8 avril 2019 par la Commission européenne ;

4. CNIL (2017). Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle. Rapport de synthèse du débat public animé par la CNIL dans le cadre de la mission de réflexion éthique confiée par la loi pour une République numérique.

5. Ces sept exigences essentielles comprennent : le facteur humain et contrôle humain, la robustesse technique et sécurité, le respect de la vie privée et gouvernance des données, la transparence, la diversité, non-discrimination et équité, le bien-être sociétal et environnemental, et la responsabilisation.

– les cinq principes énoncés par la recommandation de l'OCDE sur le développement, la mise en œuvre et l'utilisation de l'IA, adoptés le 22 mai 2019 par le Conseil aux ministres de l'OCDE[6] ;

– etc.

Quelle est la meilleure pratique des cadres éthiques, des réglementations, des normes techniques et de meilleures pratiques durables sur le plan environnemental et socialement acceptables ? Il est clair que ces encadrements partagés ne garantissent pas le succès. Les erreurs et les comportements illégaux continuent à se générer. Mais leur disponibilité nécessite d'avoir une idée claire et précise de ce qui doit être fait et de la manière d'évaluer des solutions concurrentes.

Cette diversité d'approches et d'initiatives sur le sujet traduit l'enjeu majeur d'établir un cadre commun autour d'une gouvernance éthique relative à l'IA. Se pose alors une question délicate et décisive : comment définir, ou par quelles caractéristiques « mesurables », traduire des notions de loyauté, de responsabilité, de confiance, et donc d'éthique appliquées aux décisions algorithmiques lorsqu'elles sont la conséquence ou le résultat d'une prévision ?

C'est à partir de cette vision d'universalisation que nous avons ressenti la nécessité de rédiger ce *vade-mecum* autour de l'encadrement de l'IA applicable à tous. De ce fait, nous avons développé un cadre moral accompagnant les projets numériques de l'IA en observant un certain nombre d'exigences, de préconisations et de règles, élaborées, vérifiées et discutées à chaque étape de conception, de mise en place et d'usage. Cela nous a permis d'en concevoir des critères éthiques, selon nos déterminants, essentiels et universels, basés sur le principe de l'*Ethics by Design*[7] (éthique dès la conception) ou de l'*Human Rights by Design* (droits de l'homme dès la conception) pour tendre vers un principe totalement novateur de l'*Ethics by Evolution* (éthique durant/par l'évolution) que nous développerons tout le long de ce livre. L'objectif est d'aboutir à une IA plus sûre, sécurisée, adaptée aux besoins, éthiques, et humaine dans le temps. Cela contribuera à optimiser notre faculté à suivre le progrès par rapport à des critères

6. Le 22 mai 2019, le Conseil aux ministres de l'OCDE, 42 pays (les 36 pays de l'OCDE et l'Argentine, le Brésil, la Colombie, le Costa Rica, le Pérou et la Roumanie) ont adopté les principes énoncés par la recommandation de l'OCDE sur l'IA, en faisant ainsi le premier accord intergouvernemental visant à stimuler l'innovation et renforcer la confiance dans l'IA en promouvant une approche responsable au service d'une IA digne de confiance, tout en garantissant le respect des droits de l'homme et des valeurs démocratiques.

7. Cela consiste à intégrer des règles et exigences éthiques dès la conception et l'apprentissage de ces NTIC, en interdisant de porter atteinte directement ou indirectement aux valeurs fondamentales protégées par les conventions.

de durabilité et de cohésion sociale. L'IA n'est donc pas une fin en soi, mais plutôt un moyen d'augmenter le bien-être individuel et sociétal.

> *Ethics by Design.–*
>
> Approche qui intègre des exigences et des préconisations éthiques dès la conception des NTIC.

> *Ethics by Evolution.–*
>
> Approche qui incorpore des recommandations et des règles éthiques, de manière évolutive dans le temps, tout le long du cycle de vie des NTIC, c'est-à-dire jusqu'à sa mise en place, son utilisation évolutive.

Cet ouvrage est destiné à catégoriser les questions d'éthique relatives au digital, à la fois du point de vue de l'utilisateur et du concepteur de solutions et/ou de services numériques. Il invite à la réflexion (quelles questions les entreprises peuvent-elles se poser sur l'éthique du numérique) et suggère des pistes d'action. C'est une démarche qui vise à fournir des balises, à faire émerger les valeurs que l'on veut collectivement mettre en avant pour aider les législateurs à formuler des lois qui vont encadrer l'IA. Ce référentiel n'est pas exhaustif. Il se veut généraliste, ouvert à toutes contributions, et évolutif. Il doit être mis régulièrement à jour afin de garantir sa cohérence et sa pertinence constante au fur et à mesure de l'évolution de l'environnement digital et de nos connaissances technologiques. Il n'est pas non plus destiné à rappeler les devoirs de l'entreprise en matière de réglementation, laquelle définit avec précision ce qui est permis ou interdit, et les sanctions qui s'appliquent. L'entreprise a l'obligation de se mettre en conformité, et cela ne concerne pas le domaine de l'éthique. Les moyens par lesquels elle se met en conformité peuvent en revanche faire l'objet d'une réflexion éthique.

Enfin, ce livre s'adresse à toutes les parties prenantes impliquées dans le développement, le déploiement ou l'usage d'une IA, comprenant des organisations, des entreprises, des services publics, des chercheurs, des personnes ou d'autres structures. Ce document doit donc être considéré comme étant la première brique de base d'une discussion entre ces différents acteurs tournés vers une IA éthique, responsable, digne de confiance visant à protéger et à servir de manière bénéfique les individus et le bien commun pour une meilleure adoption au niveau mondial.