Cahier n°39

l'Académie
SCIENCES TECHNIQUES COMPTABLES FINANCIERES

VERS DES INTELLIGENCES ARTIFICIELLES DIGNES DE CONFIANCE

NOVEMBRE 2022

ISACA
PARIS-FRANCE CHAPTER

Sage

EDITOS

« Quand le vent du changement se lève, certains construisent des murs, d'autres des moulins à vent ».
Proverbe chinois

L'intelligence artificielle (IA) ouvre des opportunités immenses et inédites mais aussi d'importants défis en matière de mutations économiques, sociales, démocratiques et climatiques liées aux peurs et freins qu'elles soulèvent.

Encore faut-il disposer des clés pour nous garantir – **citoyens, entreprises et pouvoirs publics** – de bénéficier d'un cadre de confiance. Au-delà de la conformité à des textes légaux en pleine émergence, il faut qu'il réponde à un référentiel de bonnes pratiques, bien particulières, qui traitent des spécificités de l'IA et des démarches pour auditer régulièrement ces systèmes.

Pour autant, peut-on affirmer que les applications basées sur l'IA sont réellement **dignes de confiance** ?

Tout l'enjeu du groupe de travail de **l'Académie des Sciences et Techniques Comptables et Financières**, initié en 2018, est bien là. Professionnels du chiffre, avocats, professeurs, consultants, organisations, associations et institutions parties prenantes au débat, se sont réunis pour identifier un ensemble de problèmes et de questions qui se posent sur les IA dignes de confiance, afin de dégager les **bonnes pratiques** pour y répondre et le cadre nécessaire pour exploiter pleinement leur potentiel.

Ce nouveau cahier de l'Académie poursuit et approfondie la réflexion du premier ouvrage paru en 2021 « **Intelligence Artificielle et Confiance : réglementation, enjeux, risques, audit et certification** ». Il confirme l'importance d'instaurer la confiance nécessaire dans les systèmes à base d'intelligence artificielle afin d'être en mesure de **mieux les maîtriser**.

Dans ce contexte, l'Académie joue un rôle clé dans le débat et la diffusion des bonnes pratiques pour **organiser un écosystème favorable à l'IA** et optimiser la mutualisation des expériences au profit des professionnels de la comptabilité, de la gestion, de l'audit et de la finance.

C'est donc le moment idéal de construire des moulins à vent en saisissant les opportunités offertes par l'IA pour relever les défis qu'elle pose.

Bonne lecture

William Nahum
Président fondateur de l'Académie des Sciences et Techniques Comptables et Financières

Dans cet ouvrage, vous trouverez tout ce que vous avez rêvé de savoir sur l'intelligence artificielle sans oser le demander. Loin des propos abscons et des discours futuristes et enflammés, vous trouverez dans les pages qui suivent des explications simples et claires, sans jargon technologique qui pourrait rendre incompréhensibles les choses les plus simples. C'est une véritable encyclopédie vous permettant de comprendre les enjeux et les risques liés à la mise en œuvre des systèmes à base d'intelligence artificielle, les démarches à suivre, les règles de gouvernance, les bonnes pratiques à appliquer ainsi que les démarches d'audit à mettre en œuvre.

Cet ouvrage fait suite à celui publié en 2021 dans la même collection : « Intelligence Artificielle et Confiance : réglementation, enjeux, risques, audit et certification ». Il a été écrit par un groupe de travail commun à l'Académie des Sciences et Techniques Comptables et Financières et à l'ISACA-AFAI créé en 2018 avec la volonté de comprendre tous les problèmes que soulèvent les nouveaux algorithmes composant ces systèmes. Que doit-on faire pour mieux les maitriser ?

Les enjeux sont considérables car dès aujourd'hui les systèmes d'information à base d'intelligence artificielle sont présents dans les logiciels métiers, les robots, les automates conversationnels, l'aide à l'orientation professionnelle (comme Parcoursup), les voitures autonomes, les traductions automatiques, les recommandations commerciales ou de lectures, le choix de films ou de séquences vidéo, les réseaux sociaux, l'aide au diagnostic médical, etc. Mais peut-on avoir confiance, d'emblée dans toutes ces nouvelles applications ?

Les démarches d'audit, d'expertise ou de certification permettent d'avoir une assurance raisonnable et de garantir que des systèmes à base d'intelligence artificielle sont réellement dignes de confiance. Il y a quelques années, le rapport Villani préconisait de mettre en place un corps de « commissaires aux algorithmes ». Il serait, pour une grande partie, composé d'auditeurs informatiques certifiés en intelligence artificielle. Nous serons alors satisfaits d'avoir contribué à définir leurs travaux et heureux de les accueillir au sein de notre association. L'ISACA, forte de 155.000 membres dans le monde, est l'association de référence pour la confiance dans le digital et les Systèmes d'Information et l'ISACA-AFAI en sont le chapitre français.

Vincent Manière,
Président de l'ISACA-AFAI

Serge Yablonsky,
Président d'honneur de l'ISACA-AFAI

En véritable partenaire de l'Académie, Sage s'engage auprès de la profession comptable et financière pour nourrir la réflexion et les échanges autour des grands challenges auxquels doit répondre notre société aujourd'hui.

Tous ces changements qui bousculent notre quotidien nous imposent d'être vigilants quant à nos actions et notre démarche afin de les rendre bienveillants et source de valeur pour tous. C'est notre engagement chez Sage, encore plus lorsque l'on touche à l'intelligence artificielle (IA).

Le thème de ce 39ème cahier, autour des notions d'IA responsables, dignes d'humanité et de confiance, est déjà un incontournable de notre quotidien et de nos perspectives de développement.

L'intelligence artificielle a fait ses preuves dans ce qu'elle apporte aux professions comptables tous les jours.

Efficacité, simplicité et productivité sont au rendez-vous, auxquels nous associons toujours, chez Sage, une approche profondément humaine. Notre objectif est simple, mettre au cœur du processus l'utilisateur et le respect que nous lui devons. Par respect, nous entendons, bien-sûr, vie privée mais nous allons bien au-delà, en veillant à maintenir le contrôle de l'humain sur l'IA, la transparence et la sécurité, mais aussi le bien-être tant personnel que sociétal. Nous épousons cette démarche naturellement par notre histoire profondément liée aux femmes et hommes qui ont construit notre entreprise, mais aussi par la relation privilégiée qui nous avons toujours nourrie avec nos clients et tout notre écosystème en les incluant dans nos réflexions.

Ainsi, au-delà de son apport opérationnel, nous positionnons l'intelligence artificielle comme un véritable levier d'amélioration de la vie des utilisateurs, source de bien-être mais aussi d'engagement au sein de leur organisation.

Nous nous associons une nouvelle fois à ce cahier pour marquer non seulement notre attachement aux actions de l'Académie, mais aussi notre engagement sur les thèmes abordés, notamment ceux qui touchent notre profession et les grands enjeux qu'elle aborde avec volonté, ambition et bienveillance.

Sabine Ducrot-Ciss
Senior Product Marketing Director
SAGE

COMPOSITION DU GROUPE DE TRAVAIL

Groupe de travail animé par :

- **Alain BENSOUSSAN**, avocat, Lexing Alain Bensoussan Avocats
- **Serge YABLONSKY**, expert-comptable, commissaire aux comptes, CGEIT, CISA, CRISC, SYC Consultants, Président d'honneur de l'ISACA-AFAI

Groupe de travail actif :

- **Marie-Agnès GRAS**, Responsable du service Grands Comptes, Groupe SAURAMPS
- **Jean-Laurent LIENHARDT**, expert-Comptable mémorialiste
- **Camille ROSENTHAL-SABROUX**, Professeur émérite Université Paris Dauphine-PSL, Lamsade, Administrateur de l'ISACA-AFAI
- **Claude SALZMAN**, consultant en Système d'Information, Président International du Club Européen de la Gouvernance des Systèmes d'Information, Administrateur de l'ISACA-AFAI
- **Patrick STACHTCHENKO**, chargé de cours en Gouvernance, Cyber Sécurité, Audit IA et Référentiels, CGEIT, CRISC et CISA, Ancien président de l'ISACA International

Le principal inspirateur de ce guide est **Patrick STACHTCHENKO**.

La mise en forme de ce guide a été assurée par le **Conseil national de l'ordre des experts-comptables** et notamment par **Marie Amélie CALMAO**, chargée administrative de l'Académie.

Sommaire

INTRODUCTION

1. Contexte des travaux

Le groupe de travail de l'Académie des Sciences Comptables et Financières en commun avec l'ISACA-AFAI a réuni de nombreux experts d'horizons très variés (scientifiques, avocats, commissaires aux comptes, auditeurs informatiques, universitaires...) pour aborder toutes les facettes de la problématique[1] des intelligences artificielles (IA) dignes de confiance. De tels systèmes informatiques et robotiques intelligents sont, pour de nombreux États, organisations internationales et acteurs de cet écosystème, un impératif pour un développement économique mondial dynamique, innovant et durable. Ils sont également nécessaires à l'amélioration du bien-être des individus et à la poursuite du progrès.

Les IA présentent à l'évidence de nombreuses opportunités dans des secteurs très variés mais aussi des défis majeurs en matière de mutations économiques, sociales, sociétales, démocratiques et climatiques liées aux peurs et freins qu'elles soulèvent.

Notre société fait face à un environnement qui englobe un nombre important de situations, de contextes, de types d'IA, de technologies et techniques associées, de modèles, de raisonnements, de fonctions proposées, de finalités, d'exigences attendues, d'outils... Les problématiques et les risques sont spécifiques à chacun de ces facteurs.

Pour pouvoir disposer d'IA dignes de confiance, les moyens à mettre en œuvre et le niveau de confiance espéré et proposé ne seront pas les mêmes pour ces différentes situations. Des solutions existent pour certains cas mais restent à développer pour d'autres.

A cet effet, il convient donc de clarifier la manière d'aborder cette problématique. Il est nécessaire d'identifier les facteurs différenciants, les hiérarchiser et les classer. Des approches génériques et spécifiques pourront être élaborées. Des réponses adaptées pourront alors être proposées.

Les différentes parties prenantes de l'écosystème des IA ont besoin d'avoir confiance et de donner confiance dans les outils et dispositifs* mis en œuvre ainsi que dans les services et résultats, issus de ces IA et de leur écosystème, pour être des acteurs tout à fait contributifs à la chaîne de valeur globale.

2. Objectifs du groupe de travail

Dans ce contexte, l'objectif premier du groupe a donc été de comprendre pourquoi notre société a besoin d'avoir confiance dans les IA en précisant la nature de ce besoin, les acteurs intéressés par cette démarche, son intérêt pour ces derniers et les attentes associées.

---

[1] Une explication introductive des termes spécifiques au cahier est donnée en annexe. Ils sont signalés par un astérisque

L'objectif visé fut d'identifier les difficultés et freins actuels au développement de telles IA et les conditions nécessaires à leurs déploiements effectifs et à la fourniture par un tiers de confiance d'une opinion quant aux assertions possibles relatives à la présence d'IA dignes de confiance.

Dans le cadre de ses travaux, le groupe a dégagé plusieurs thèmes et éléments spécifiques aux IA, dont une prise en compte appropriée a été considérée comme essentielle. Il a approfondi leur analyse et tenté d'y apporter une contribution.

L'objectif est de pouvoir disposer in fine :

- d'un cadre commun de compréhension des enjeux, des freins, des obstacles, des spécificités qui méritent une attention particulière, des exigences attendues, des types de risques susceptibles d'apparaître, des parties prenantes concernées et leurs responsabilités, des outils disponibles et leurs limites, pour des systèmes à base d'intelligence artificielle dignes de confiance
- d'assertions spécifiques quant aux qualités et caractéristiques attendues de tels systèmes informatiques et robotiques
- de référentiels de bonnes pratiques* qui, si elles sont en œuvre, doivent permettre de concevoir et mettre à disposition de telles IA
- de référentiels de bonnes pratiques d'audit et de certification d'assertions spécifiques à mettre en œuvre par les auditeurs (i.e. commissaires aux IA).

3. Nature des travaux

Plusieurs organisations se sont emparées de certaines problématiques identifiées comme critiques au développement d'IA dignes de confiance et ont publié les résultats de leurs travaux. Par ailleurs, des États et des institutions internationales tels l'OCDE, le Parlement européen et la Commission européenne ont souhaité poser les fondations d'une réglementation adaptée au développement de telles IA.

Dans ce contexte, le groupe de travail a été amené à analyser des dizaines d'études ou de propositions réglementaires. Il était nécessaire d'appréhender les objectifs de ces travaux, les problématiques soulevées, les difficultés rencontrées, les pistes de solutions proposées en partant du principe que c'est en posant les bonnes questions qu'on peut trouver les bonnes réponses.

Pour cela, plusieurs points essentiels ont été identifiés, complétés et enrichis au travers des analyses du groupe de travail. Un cadre général ainsi qu'une démarche spécifique pour de telles IA et pour leur certification ont été ensuite élaborés et illustrés au travers d'exemples.

4. Contenu et destinataires du cahier

Ce document a pour vocation de restituer notre compréhension du contexte et des conditions nécessaires à la mise à disposition d'IA dignes de confiance.

Ce rapport s'adresse à tous les acteurs directement impliqués dans le développement, le déploiement, et le maintien d'un système d'intelligence artificielle. Il peut s'agir de spécialistes des technologies de l'IA, de data scientistes, de développeurs, d'ingénieurs, de managers en cybersécurité et en risque IT ... qui pourront y trouver des éléments susceptibles d'améliorer leurs démarches de travail en intégrant les diverses problématiques et points de vue.

Il s'adresse aussi aux auditeurs internes et externes qui seront amenés à porter une appréciation sur la prise en compte effective et efficace de ces différentes problématiques et à fournir une assurance quant à la présence ou pas d'IA dignes de confiance répondant aux exigences attendues et aux référentiels de bonnes pratiques adaptés.

Il s'adresse enfin aux autres acteurs qui n'ont pas nécessairement une forte expertise de ce domaine mais qui pourraient être particulièrement intéressés par ces IA dignes de confiance. Il s'agit en particulier, mais sans se limiter, des responsables des systèmes d'information et du numérique, de la protection des données, de la gestion des risques, de l'éthique, des dirigeants d'entreprise, des responsables métier (directeur de branche, responsable de la recherche, directeur fonctionnel...), des juristes, des instances professionnelles concernées, des universitaires, des législateurs...

Grâce à ce cahier, les intéressés comprendront les enjeux d'une IA digne de confiance, découvriront des clés pour aboutir à la mise en place d'un tel système et trouveront une aide pour atteindre leurs objectifs que ce soit en termes de choix d'achat, de mise en place de certains concepts lors d'un projet de création, d'évolutivité ou au contraire de maintien dans le temps de leurs IA en place ou encore d'audit et de certification de l'IA elle-même.

CHAPITRE 1

LES SYSTÈMES À BASE D'INTELLIGENCE ARTIFICIELLE : DES SPÉCIFICITÉS SOURCES D'APPORTS CONSÉQUENTS MAIS AUSSI DE NOUVEAUX RISQUES

Il ne se passe pas une semaine sans que l'IA ne soit mentionnée dans la presse. Elle est au cœur de nombreux sujets et conversations, preuve que nous sommes dans une période charnière et incertaine. Cette cacophonie médiatique embrouille souvent le citoyen. Dans ce contexte, il est essentiel d'identifier les spécificités des IA et de comprendre à la fois en quoi et pourquoi elles sont porteuses de grandes espérances tout en suscitant de grandes inquiétudes. Ce n'est qu'en appréhendant correctement ces différents facteurs et contextes dans lesquels ils sont mis en œuvre que l'on pourra trouver les clés à un développement maitrisé des IA et une adoption valorisée et partagée par le plus grand nombre.

Ce chapitre s'intéresse dans un premier temps aux forts impacts du développement des IA sur les enjeux de la société, sur la gestion des projets, sur les compétences, sur les conditions de succès et sur les autres technologies. Leurs prises en compte satisfaisantes seront essentielles au développement d'IA de confiance.

Dans un second temps, les différents éléments qui permettent d'appréhender les spécificités des IA sont identifiés. Plusieurs grilles d'analyse permettent de déterminer en quoi et pourquoi elles sont critiques pour obtenir des IA dignes de confiance. Sont notamment abordés, les apports, les limites et les risques, articulés autour des trois piliers des IA qui la composent (puissance de traitement et de communication, logiciels, données) tout en insistant sur la nécessité d'une intégration harmonieuse.

Enfin, les fortes perspectives d'évolution des différents domaines de l'informatique qui devraient favoriser le développement des IA sont évoquées. Leurs prises en compte maitrisées seront critiques à la mise en œuvre d'IA dignes de confiance

1. De forts impacts sur les enjeux, les projets, les compétences, les conditions de succès et les autres technologies

1.1. Une présence accrue et diffuse des systèmes à base d'intelligence artificielle

L'importance qui est accordée à l'IA provient d'un constat simple : à terme, elle imprégnera tous les secteurs de la société, créant de nouveaux produits, de nouveaux métiers et de nouveaux secteurs de l'économie. Si nous ne sommes pas encore entrés dans cette nouvelle ère, cela ne saurait tarder au regard de la rapidité des développements actuels.

Les premiers changements s'articulent déjà et s'intensifieront autour de l'amélioration de nos conditions de vie actuelle tels que de meilleurs soins de santé (précision accrue des diagnostics ou meilleure prévention des maladies, par exemple), les villes intelligentes, l'industrie 4.0 et davantage de robots, l'internet des objets (IoT)... Le recrutement sera facilité, accélérant les procédures et limitant les erreurs d'embauche. Les voitures autonomes se développeront jusqu'à ne plus nécessiter d'intervention humaine. La finance automatique continuera à réaliser du trading, à analyser le marché et à conseiller les humains sur la gestion de leur portefeuille boursier et de leurs finances personnelles.

A moyen terme, on peut espérer que l'IA facilitera la résolution des grands problèmes planétaires comme le réchauffement climatique. L'agriculture de précision, pilotée par des drones autonomes, améliorera les rendements tout en limitant les besoins en eau. La maintenance prédictive permettra de rallonger la durée de vie des équipements ce qui améliorera l'efficacité des systèmes de production tout en limitant le besoin des ressources naturelles. Lorsque le renouvellement des infrastructures actuelles laissera place à du matériel plus développé, le Smart Grid distribuera plus intelligemment les ressources électriques limitant ou optimisant les besoins de production. L'IA pourra toucher au droit pour prédire les décisions de justice et aider à la décision des cas simples.

1.2. De forts enjeux et un environnement technique non stabilisé menant à une complexification des projets

Sa polyvalence est telle que chaque secteur (transports, éducation, sécurité, média...) adapte et créée des technologies dédiées pour améliorer son fonctionnement. Mais le chemin est encore long entre le discours commercial et la mise en place des promesses. Conscientes du saut technologie restant à franchir, les industries investissent massivement dans le développement et la mise en oeuvre de nouvelles technologies, la captation et le traitement de nouvelles données, le développement et l'adaptation des compétences et des méthodes de travail.

Ces projets sont pourtant complexes à mener, risqués et coûteux à entreprendre. Ils nécessitent de faire des choix technologiques et éthiques au bon moment, de maîtriser les risques et les conséquences de ces choix. Il sera nécessaire de s'assurer que les différentes parties prenantes aient confiance dans les choix effectués.

Il reste encore du chemin avant de parvenir à maîtriser ces IA variées et complexes, à mener à bien des projets de très grande ampleur et à forts enjeux et risques, et à y intégrer dès le début des projets, des dispositifs spécifiques adaptés aux IA intégrant la composante humaine qui rendront les IA dignes de confiance.

1.3. Un impact conséquent sur le marché du travail et sur les compétences

Dès qu'une alternative technologique à un emploi humain existe, celle-ci est presque systématiquement choisie, dans une optique de gain de productivité[2]. L'impact de l'IA dans le marché du travail est déjà visible depuis plusieurs années : les ouvriers sur les lignes d'assemblage ont déjà été remplacés par des robots dès que c'était possible. Amazon fait le choix du tout-automatique en investissant massivement pour utiliser des robots dans ses entrepôts de distribution plutôt que des salariés humains.

Le succès de la relocalisation des usines dans les pays occidentaux reposera sur la diminution des coûts de fabrication par rapport au coût de la force de travail des « ateliers du monde » et donc la possibilité d'automatisation de la production. Le risque ne se limite pas au remplacement du travail manuel : avec l'informatisation des processus, dès lors qu'une tâche est répétitive et ou qu'elle nécessite une prise de décision selon des règles assez simples, l'automatisation peut prendre le relai. Tous les métiers peuvent donc être impactés mais cela ne signifie pas non plus la disparition pure et simple de ceux-ci.

Des métiers vont disparaître

Certains métiers sont voués à être totalement ou quasi-totalement automatisés tels que les caissières ou les standardistes, les centres d'appel de taxis, etc. Si l'évolution des métiers à faible valeur ajoutée semble assez prévisible, ce n'est pas le cas des métiers plus complexes.

Des métiers vont être plus ou moins impactés

Certains spécialistes estiment qu'un certain nombre de métiers pourraient être protégés de l'IA, de la robotique ou de l'automatisation. Il pourrait s'agir des métiers créatifs, des artisans dans le BTP, la plomberie et l'électricité et des métiers de services qui touchent au corps comme les coiffeurs, manucures, kinésithérapeutes... ainsi que la plupart des professions de santé, les enseignants et les travailleurs sociaux pour lesquels le contact humain est la clé. L'IA servirait davantage d'outil et de support, les métiers évolueraient donc jusqu'à l'obtention d'une certaine symbiose avec elle, comme on peut le voir aujourd'hui chez les pilotes d'avion dont une partie du travail est géré par l'appareil et le pilote automatique. Il est facile d'imaginer que ces systèmes seraient utilisés comme aide au diagnostic (analyse du problème en BTP, analyse des maux en médical) ou d'outils variés en fonction des métiers (suggestion de coiffure ou de maquillage suivant les caractéristiques faciales des clients, analyse des goûts des consommateurs...).

D'autres experts estiment qu'avec l'amélioration des performances de l'IA, peu de métiers seraient vraiment à l'abri. Il est déjà envisagé que des robots conversationnels puissent aider dans les services hospitaliers ou les EPHAD où ils se montreraient plus réactifs, plus empathiques que le personnel soignant et surtout inlassables à converser avec les gens. Cela permettrait aux soignants de se concentrer sur d'autres types de tâches. Sur le volet créatif, les expériences de création réalisées à 100 % par les IA se

[2] Étude de l'institut Sapiens « L'impact de la révolution digitale sur l'emploi » https://www.institutsapiens.fr/wp-content/uploads/2018/08/Note-impact-digital-sur-lemploi.pdf

multiplient. Si les résultats sont aujourd'hui encore souvent médiocres[3], on peut supposer que la machine pourrait être capable de remplacer les scénaristes humains dans la conception des intrigues et l'écriture des scénarios de films et séries.

De nouveaux métiers vont apparaître

Enfin, il ne faut pas sous-estimer l'importance de la création de nouveaux métiers[4]. Il y a bien évidemment tous les ingénieurs, concepteurs et développeurs en IA et les métiers de support du numérique qui sont associés. Par exemple, les métiers de psydesigner qui s'occuperait de créer des personnalités aux IA et de coach de robot qui enseignerait aux robots à être plus efficace dans la réalisation des tâches qui leur sont données, pourraient émerger. On peut supposer que si l'automatisation des métiers libère du temps sans que le pouvoir d'achat ne baisse, de nouveaux besoins émergeront et de nouveaux métiers apparaîtront dans les loisirs.

L'IA aura un impact fort et certain sur l'emploi, soit elle remplacera certains métiers à faible valeur ajoutée, soit elle permettra l'apparition de nouveaux métiers.

De nouvelles compétences seront nécessaires dans les métiers existants

Elle impactera probablement un très grand nombre de métiers en nécessitant une évolution des compétences pour apprendre à utiliser et maîtriser de nouveaux outils qui feront évoluer les métiers existants. Là encore, ces évolutions ne seront envisageables et pérennes que si les utilisateurs ont confiance dans les IA développées.

Or, pour disposer d'IA de confiance, il faudra avoir confiance dans la mobilisation et la mise en œuvre maîtrisée des différentes compétences appropriées.

1.4. La nécessité d'un accès pérenne aux différents éléments clés de l'écosystème des IA

L'IA bouleversera notre société à plusieurs niveaux. Pour que ces changements se réalisent et soient acceptés par la population, il est nécessaire de bâtir un écosystème de l'IA fort qui soit capable d'innover et qui soit digne de confiance. La confiance dans les systèmes d'IA bâtis par les acteurs de chaque pays ne pourra s'établir que si l'accès local aux composants clés de cet écosystème est assuré. Six éléments apparaissent comme critiques :

- l'accès à des compétences appropriées : trouve-t-on des spécialistes en IA localement ?

---

[3] L'exemple d'envergure le plus récent est le court-métrage d'horreur de Netflix « *Mr. Puzzles Wants You To Be Less Alive*. »

[4] Notons que de nouveaux métiers émergeront même indépendamment de l'essor de l'IA, c'est le cas par exemple des métiers des énergies renouvelables.

- l'accès aux nouvelles innovations des technologies IA adaptées et maîtrisées : on parle notamment d'une recherche scientifique dynamique en IA. Cela peut se traduire par les questions suivantes : des fonds sont-ils attribués à la recherche fondamentale et appliquée ? Des études scientifiques sont-elles publiées ? Les acteurs impliqués sont-ils accessibles ?
- l'accès à la puissance de calcul nécessaire : dispose-t-on de supercalculateurs dans le pays ? Sont-ils facilement accessibles et en nombre suffisant ?
- l'accès à la richesse des données : l'IA est grande consommatrice de bases de données : est-il possible de les obtenir facilement dans le pays ?
- l'accès à un écosystème IA favorable et performant qui sait maîtriser la gestion de grands projets complexes qui intègrent les différentes composantes/vues y compris avec la technologie et les projets dits « traditionnels »
- l'accès maîtrisé à un environnement d'utilisation favorable : les IA restent-elles cantonnées aux laboratoires de recherche ou non ? Les IA sont-elles adoptées par les entreprises et le grand public ou rejetées ?

Ces éléments clés soulignent l'importance de la souveraineté parmi les enjeux de l'IA. Par exemple, si une entreprise française a un accès régulier à un supercalculateur étranger pour ses recherches, elle est à la merci des aléas diplomatiques où, en guise de rétorsion, le pays étranger peut couper l'accès des supercalculateurs aux étrangers et en particulier aux français. La souveraineté est donc un enjeu qui peut être critique.

De plus, comme l'IA infiltrera l'ensemble de la société, il semble logique de supposer que la nation qui conduira la course du développement et de l'utilisation de l'IA façonnera le futur de cette technologie et améliorera drastiquement sa compétitivité économique, tandis que les États à la traîne risquent de perdre leur compétitivité économique dans les industries clés[5]. Les gouvernements en sont bien conscients et plus de trente d'entre eux ont mis en place des stratégies nationales concernant l'IA.

Pour l'instant, les États-Unis semblent mener la course, talonnés par la Chine, tandis que l'Union européenne décroche un peu. En effet, la Chine est leader dans les deux éléments concernant la maîtrise des données et l'adoption de l'IA hors des laboratoires de recherche alors que les USA bénéficient d'un réservoir de talents humain supérieur, du développement d'un écosystème favorable, du matériel et de la recherche.

La Chine, initialement à la traîne, réalise de gros efforts : elle a dépassé l'Union européenne pour devenir le premier pays en termes de nombres de publications scientifiques parues, la qualité de la recherche et développement augmente, les sociétés de services et d'informatique ont augmenté leurs dépenses R&D dans ce domaine, il y a désormais près de deux fois plus de supercalculateurs classés dans le top 500 des performances qu'aux États-Unis – qui étaient en tête pour cet indicateur pas plus tard qu'en 2017.

---

[5] D. Castro et M. McLaughlin « Who is winning the AI Race : China, the EU or the US ? » Jan 2021

D'autres pays s'investissent dans le sujet et présentent de bonnes performances sur certaines thématiques : c'est le cas de l'Inde dont le nombre de personnel qualifié en IA augmente, de l'Israël qui reçoit des montants de financements privés par habitant significatifs et de l'Australie qui publie de nombreux papiers de recherche dans le deep learning.

Les progrès de l'Union européenne pour rattraper le leader américain dépendent des domaines considérés. Malgré une augmentation importante des investissements, l'IA attire moins de financements qu'aux USA. En revanche, la qualité de la recherche de l'Union européenne augmente : les publications scientifiques européennes sont de plus en plus citées (et donc utilisées). Il est important de noter que l'Union européenne semble avoir conscience de son retard technologique et tente donc de s'imposer par une autre manière : sur le niveau règlementaire. En effet, en indiquant clairement ce qui est acceptable ou non en termes d'IA, elle indique aux chercheurs et aux développeurs une ligne de conduite et une orientation des recherches pouvant même influencer les autres nations, s'ils veulent que leurs IA puissent être utilisés dans l'Union européenne. La mise en place d'un processus de certification de ces systèmes permettrait de valider qu'elles ont été construites en répondant bien à ces impératifs réglementaires européens. La validité de cette stratégie a été confirmée par la mise en place du RGPD (Règlement Général sur la Protection des Données) qui a été suivi par un grand nombre de sociétés dans le monde ou qui a été l'inspiration principale pour de nombreuses règlementations locales comme en Californie avec le California Consumer Privacy Act.

Il est donc essentiel que les acteurs d'un pays aient confiance en un accès pérenne aux compétences pertinentes, aux innovations technologiques et techniques, à la puissance de traitement, à la richesse des données, à un écosystème IA performant et à un environnement économique, social, sociétal et règlementaire adapté propice à l'adoption d'IA dignes de confiance.

1.5. Une intégration avec les technologies « traditionnelles » au cœur des transformations

L'IA est un sujet particulièrement prégnant depuis les années 2010 et semble être apparue soudainement. Pourtant, elle s'inscrit dans la stricte évolution des technologies antérieures. Les premiers signes d'un intérêt pour l'IA remontent à l'antiquité, avec les premiers automates que constituaient les statues sacrées en Égypte antique. Mais il faut attendre la seconde partie du XXème siècle que les innovations permettent l'essor de cet intérêt. L'IA nait entre 1940 et 1956. Cette époque est riche en création de toutes sortes : le premier ordinateur piloté par un programme[6], la publication des principes des réseaux de neurones

[6] Il s'agit du Zuse 3 construit en 1941 par Konrad Zuse,

artificiels[7], la création du terme « cybernétique[8] », la création du test éponyme d'Alan Turing pour mesurer l'intelligence d'une machine en 1950 ou la publication des trois lois de la robotique[9] d'Isaac Asimov. Cette période se termine sur la tenue de la conférence de Dartmouth en 1956 où la discipline de l'IA naît officiellement : le nom est adopté et ses objectifs sont définis.

S'ensuit un premier âge d'or jusqu'en 1974 où les progrès suivent ceux de l'informatique. C'est dans ces périodes qu'émergent la recherche opérationnelle[10] (années 1960) et la supervision des opérateurs ou process control (dans les années 1970). Les programmes de l'époque peuvent déjà tenir des conversations (limitées) en anglais, résoudre des problèmes d'algèbre et démontrer des théorèmes de géométrie.

Différents problèmes comme les limites de la puissance de calcul ou les limites sensorimotrices des ordinateurs mettent en pause la recherche sur l'IA jusqu'à la montée, dans les années 1980, des systèmes experts qui simulent les raisonnements d'un expert dans un domaine de connaissances donné. Enfin, à partir de 1990, l'IA est utilisée pour des problèmes compliqués mais très précis qui mènent à des utilisations concrètes dans la vie économique. L'IA s'applique à de nouveaux domaines : exploration de données, robotique industrielle, logistique, reconnaissance vocale, applications bancaires, diagnostics médicaux. Les systèmes à base de connaissances, les systèmes cognitifs et l'intelligence computationnelle, tous apparus à partir de cette époque, sont des IA qui ne portent pas leur nom.

Les IA sont partout autour de nous, et ce, depuis bien longtemps. Cette tendance se poursuivra et même s'intensifiera au fur et à mesure que les techniques permettront de se rapprocher de l'intelligence humaine. Ce foisonnement ne sera acceptable et accepté qu'avec le développement d'IA dignes de confiance mais

---

[7] W. McCulloch et W. Pitts « A Logical Calculus of the Ideas Immanent in Nervous Activity » 1943. Les réseaux de neurones s'inspirent à l'origine du fonctionnement des neurones biologiques et se sont ensuite rapprochés des méthodes statistiques

[8] Différentes acceptions du mot cybernétique se chevauchent, suivant la période, le cadre scientifique, ou le point de vue où l'on se place. C'est une méthode interdisciplinaire qui étudie l'évolution dynamique des systèmes avec la rétroaction comme concept central.

[9] Tous les robots de l'univers de l'écrivain de SF Isaac Asimov (aussi professeur de biochimie) doivent obéir aux trois lois suivantes (grossièrement résumées) 1.ne pas porter atteinte à un humain, 2. obéir aux humains sauf si contradiction avec loi 1, et 3. protéger sa propre existence sauf si contradiction avec lois 1 et 2. Ces concepts sont importants car ils marquent le début d'une réflexion sur les relations humains-robots. Ces lois sont souvent considérées comme un idéal à atteindre. Sur ce sujet, nous pouvons noter la proposition de modification de la constitution française en janvier 2020 par M. Raphan (député de l'Essonne – LREM) dont l'article 2 est littéralement constitué par les trois lois de la robotique.

[10] Ensemble des méthodes et techniques orientées visant à faire le meilleur choix pour atteindre le meilleur résultat dans un contexte donné. Par exemple : quel est le meilleur chemin à faire prendre au facteur pour limiter son temps de tournée ? Comment limiter le gaspillage de matière première lors de l'étape de découpe ?

aussi en ayant confiance dans toutes ces technologies « historiques » car elles s'intègreront dans les systèmes d'IA futurs[11].

Dans cette sous-partie (*§ 1*), nous avons vu que la multiplication de nouveaux projets IA à forts enjeux et risques qui complexifient la gestion de projet, qui requièrent de nouvelles compétences et comportements, qui doivent intégrer (ou s'intégrer aux) différents systèmes et technologies existants, qui nécessitent un accès pérenne aux innovations et aux différentes ressources et qui requièrent un environnement local adapté peut susciter des inquiétudes. Ce n'est donc que lorsqu'on pourra avoir confiance dans la prise en compte satisfaisante de tous ces éléments que l'on pourra avoir confiance dans les IA qui seront issues de ces projets.

Par ailleurs, les caractéristiques spécifiques des IA sont porteuses d'opportunités mais aussi de risques. Il convient de s'assurer que l'on pourra avoir confiance dans la prise en compte satisfaisante de ces spécificités. Il s'agit donc de bien comprendre la nature des impacts qui pourraient en découler et d'avoir confiance en leur prise en compte adaptée. C'est ce que nous allons aborder dans le sous-chapitre suivant.

2. En quoi et pourquoi certaines spécificités des systèmes à base d'intelligence artificielle sont critiques pour des IA dignes de confiance

Avant de pouvoir construire des IA de confiance, il est important de définir ce qu'est une IA, ce qui la caractérise. Cela n'est pas aussi simple qu'il y paraît. Ensuite, les différents axes d'analyse des IA devant permettre de distinguer les facteurs critiques pouvant impacter leur niveau de confiance sont expliqués. Puis nous présenterons pourquoi il convient d'étudier ces dernières à la fois de manière individuelle, mais aussi dans leur globalité.

2.1. Un périmètre multiforme à appréhender

Comme nous avons pu le voir, l'intelligence artificielle est une discipline multiforme et en constante évolution. Il n'existe pas une IA comme il existe un logiciel de traitement de texte ou une application mobile.

La Commission européenne constate cette variabilité des systèmes d'IA à travers la définition qu'elle en donne :

"[Un] « système d'intelligence artificielle » (système d'IA) [est] un logiciel qui est développé au moyen d'une ou plusieurs des techniques et approches [...] et qui peut, pour un ensemble donné d'objectifs définis par

[11] Ce fait n'a d'ailleurs pas échappé à la Commission européenne qui a inclus les systèmes experts dans le périmètre de la législation sur les IA

l'homme, générer des résultats tels que des contenus, des prédictions, des recommandations ou des décisions influençant les environnements avec lesquels il interagit."

Ces approches sont les suivantes :

"(a) Approches d'apprentissage automatique, y compris d'apprentissage supervisé, non supervisé et par renforcement, utilisant une grande variété de méthodes, y compris l'apprentissage profond.

(b) Approches fondées sur la logique et les connaissances, y compris la représentation des connaissances, la programmation inductive (logique), les bases de connaissances, les moteurs d'inférence et de déduction, le raisonnement (symbolique) et les systèmes experts.

(c) Approches statistiques, estimation bayésienne, méthodes de recherche et d'optimisation."

Il existe en effet différents systèmes d'IA qui, en fonction des techniques déployées et des buts visés, présentent différents problèmes, risques, enjeux et questionnements. Les réponses à apporter seront variables en fonction du type de système d'IA mis en œuvre.

Ce document sera illustré par de nombreux exemples mais se penchera en particulier sur deux systèmes d'IA différents : la voiture autonome et Parcoursup, système de classement et d'affectation des étudiants dans les universités de France. La voiture autonome a été choisie car elle représente le système d'IA « idéal » : quasiment toutes les techniques d'IA sont employées. Parcoursup a été sélectionné car il se trouve à l'autre bout du continuum du spectre d'un système d'IA. Quasiment aucune technique d'IA n'y est utilisée dans le cœur du système : l'algorithme de Gale-Shapley n'est pas considéré comme relevant de l'IA par certains spécialistes. Si Parcoursup est considéré comme de l'IA par le grand public et par d'autres spécialistes, c'est que ce type d'intelligence se retrouve dans l'ensemble du dispositif associé à Parcoursup ; c'est à dire en prenant aussi en compte les quelques 15 000 « algorithmes locaux » utilisés par les établissements de l'enseignement supérieur. Par ailleurs, malgré cette faible utilisation des techniques d'IA, Parcoursup possède la caractéristique intéressante de présenter tous les problèmes et polémiques qui sont soulevées par l'IA : définition du périmètre d'une IA, confidentialité des données, transparence, discrimination… Mais pour mieux comprendre tout cela, il convient de définir l'IA et ses spécificités, de décrire quelques termes et d'établir quelques classifications.

L'IA est un assortiment de techniques qui servent d'outils simulant l'intelligence humaine ou animale de façon à atteindre un but précis. On peut faire un parallèle entre l'IA et les mathématiques : il n'existe pas « une mathématique » possédant un usage défini. Les mathématiques sont divisées en plusieurs domaines comme l'algèbre, les probabilités, la géométrie… Parler de « l'IA »[12] est en réalité aussi absurde que de dire qu'il n'existe qu'une seule mathématique. De même que pour l'IA, les mathématiques sont un ensemble de connaissances, de théorèmes et de formules servant d'outils pour les domaines physiques, biologiques…, et d'IA ! Car l'IA est à la croisée de nombreuses disciplines, depuis les mathématiques à la philosophie en passant par le droit ou l'économie…

[12] Pour des fins de simplification de ce texte, l'usage de IA au singulier sera conservé

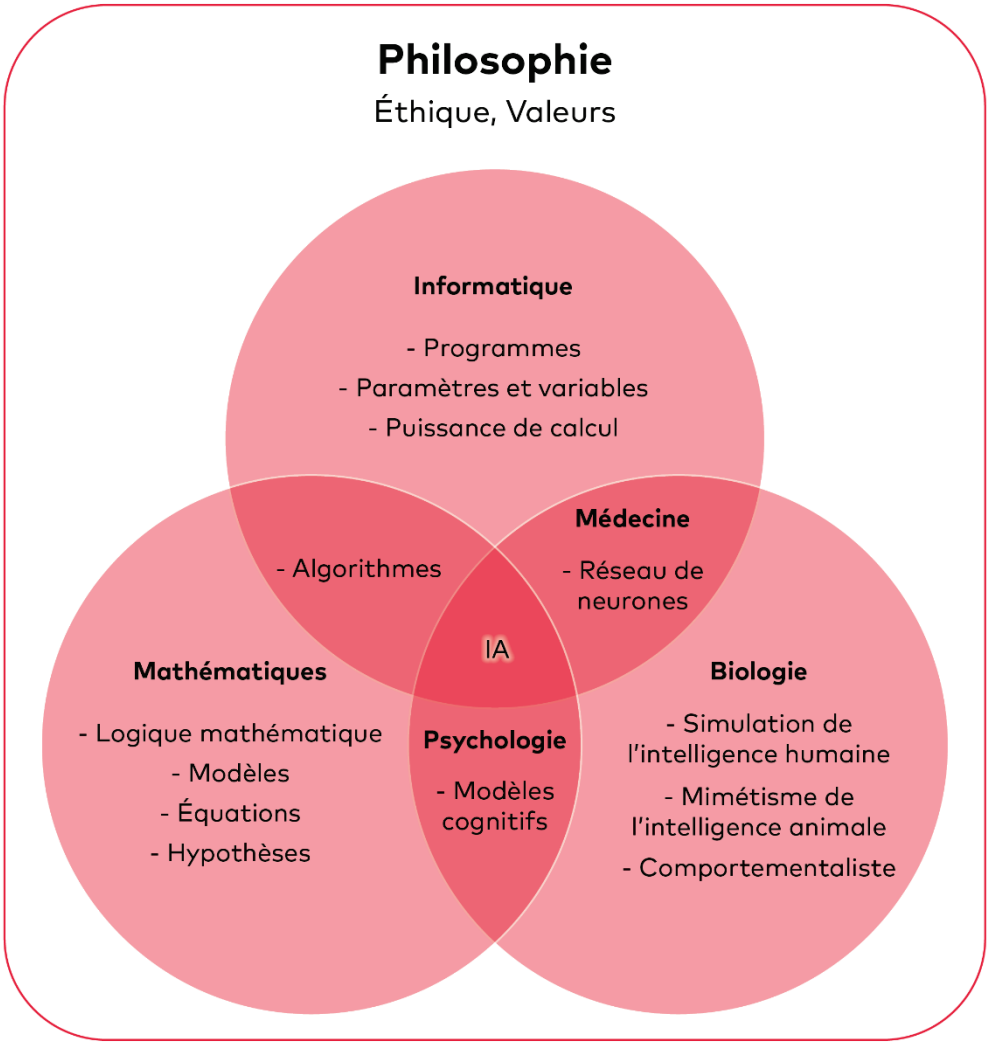

Figure 1 : Une représentation possible de l'IA en fonction des disciplines mises en œuvre

La compréhension de l'IA a ceci de difficile que le même terme d'« intelligence artificielle » désigne le tout et ses différents éléments constitutifs, et que la définition de ces éléments constitutifs varie d'un expert (ou d'un non-expert) à un autre mais aussi du domaine auquel appartient l'expert (un juriste aura une définition différente d'un informaticien). Suivant l'angle adopté, on peut donner différentes définitions et classifications de l'IA, toutes vraies. On peut classer l'IA en fonction des techniques utilisées, des compétences mises en œuvre, des mécanismes cognitifs qu'on tente de reproduire… Décrivons-en quelques-unes.

L'une des représentations les plus abordables est celle d'Olivier Ezratty qui scinde l'IA en quatre strates :

- les solutions : que l'on va directement utiliser dans les entreprises ou chez les particuliers avec les chatbots, les véhicules autonomes, les robots, les systèmes de recommandation, les outils de segmentation client, le marketing prédictif ou les solutions de cybersécurité. Il s'agit par exemple, dans le domaine des véhicules autonomes, de la Google Car d'Alphabet, de Nuro de la start-up éponyme, ou des futures Tesla. Parcoursup est une plateforme centralisée pour les vœux des lycéens de même que Central Application Office (CAO) en Irlande ou Universities and Colleges Admissions service (UCAS) au Royaume-Uni
- les outils : qui aident à créer ces solutions, comme la vision artificielle[13], la reconnaissance de la parole*, la traduction automatique*, les systèmes experts*, les outils de prévision* ou de segmentation* automatiques, les réseaux multi-agents*… Les véhicules autonomes* utilisent les outils de vision artificielle, de réseaux multi-agents et de systèmes experts. Parcoursup emploie des outils de prévision et de segmentation
- les techniques : sur lesquelles sont construits ces outils, avec les méthodes de machine learning*, les réseaux de neurones*, les nombreuses méthodes de deep learning* et les moteurs de règles*. Chacune de ces techniques contient plusieurs spécialités. Par exemple, le machine learning regroupe l'apprentissage supervisé*, non supervisé*, semi-supervisé*, par renforcement*, meta learning*, l'apprentissage profond*… Dans le cas des voitures autonomes, les techniques employées sont, entre autres, celles du deep learning (pour la vision artificielle), des moteurs de règles (pour les systèmes experts). Parcoursup s'est appuyé sur un algorithme de machine learning supervisé dit « forêt d'arbres décisionnels* ». Le choix de l'utilisation des techniques dépend souvent du contexte d'utilisation : deux techniques ayant un objectif similaire (par exemple, deux techniques d'apprentissage automatiques) peuvent mener à des efficacités de résultats radicalement différents suivant le contexte d'utilisation de cette IA (par exemple, les données accessibles sont plus adaptées à l'une des techniques ou l'une des techniques est trop demandeuse en puissance de calcul)
- les données : les sources de données et les capteurs associés jouent un rôle indispensable. Dans les véhicules autonomes, le deep learning utilise des données via des capteurs de type radars, Lidars et caméra, ce qui permet de déchiffrer les panneaux de circulation, de détecter les piétons ou les autres

[13] Une explication introductive des termes techniques est donnée en annexe. Ils sont signalés par un astérisque

véhicules. Parmi les données utilisées, il y a aussi les bases de connaissance et de règles (utilisées pour le moteur de règles). Les règles sont par exemple celles du code de la route local. Les bulletins de notes, les vœux des élèves, leur lycée d'origine... sont des données utilisées par la solution Parcoursup

Cette représentation a l'avantage de présenter l'IA de manière verticale : du plus petit élément (une donnée) au plus grand (le produit final comme solution). On peut aussi très bien y comprendre comment pour obtenir la confiance de la société envers les solutions d'IA, il faut également établir la confiance envers chacune de ces strates : il ne saurait y avoir de confiance dans la strate supérieure si les données, les techniques et les outils ne sont pas dignes de confiance eux-mêmes.

Notons que cette représentation en quatre strates est adaptée pour un point de vue technique. Des juristes pourraient définir l'IA d'une autre manière. Ainsi, dans ce point de vue, un système d'IA est un système qui met en œuvre au moins un algorithme qui utilise de grandes quantités de données. Un algorithme est lui-même défini par une suite finie d'opérations qui aboutit à un résultat. Dans cette vision des choses, on peut faire un parallèle avec une chanson. Tout comme une chanson est composée du texte et de la musique instrumentale, un système d'IA serait composé des algorithmes (le texte) et des données (la musique). Une IA devrait répondre à deux critères : la reproductibilité et la justesse du résultat. Une définition d'un point de vue juridique peut aussi être donnée aux robots : un robot serait une IA additionnée d'une coque mécatronique* comportant des actionneurs et des capteurs. La différence fondamentale entre un robot et une IA est que le robot est capable de modifier son environnement directement par l'intermédiaire de ces actionneurs.

Lorsque l'on parle d'IA de confiance, il convient donc de bien préciser le périmètre exact qui est couvert et la nature des éléments dont on souhaite évaluer le niveau de confiance associé.

2.2. De nombreuses grilles d'analyses avec des impacts spécifiques à traiter : types de problèmes, de finalités, de raisonnements, de logiciels...

S'il est difficile de donner une définition unique des systèmes à base d'intelligence artificielle, ces derniers présentent plusieurs caractéristiques communes. Chacune de ces caractéristiques peut prendre plusieurs valeurs. Une IA en particulier se définit donc par l'ensemble des valeurs prises à chacune de ces caractéristiques. Nous avons identifié dix caractéristiques que nous détaillons ci-dessous.

2.2.1. Des types de résultats à obtenir

Les systèmes à base d'Intelligence artificielle sont utilisés comme des solutions pour résoudre des problèmes. L'un des éléments prometteurs de l'IA est justement la grande variété de problèmes capables d'être traités par ces nouvelles techniques informatiques. Ils peuvent être utilisés pour aider à observer, à diagnostiquer, à expérimenter, à simuler, à modéliser, à prescrire, à agir, à prévoir ou à s'intégrer dans d'autres dispositifs.

Ainsi, les systèmes dotés d'IA peuvent être purement logiciels, agissant dans le monde virtuel (assistants vocaux, logiciels d'analyse d'images, moteurs de recherche ou systèmes de reconnaissance vocale et

faciale, par exemple) mais l'IA peut aussi être intégrée dans des dispositifs matériels (robots évolués, voitures autonomes, drones ou applications de l'internet des objets, par exemple).

La nature et le niveau des risques et des impacts correspondants n'est pas le même pour chacune de ces finalités. Une erreur au niveau d'une action prescrite et effectuée dans le cas d'une voiture autonome peut entraîner des conséquences plus immédiates et néfastes qu'une erreur d'observation. Le niveau de confiance attendue n'est donc a priori pas le même.

2.2.2. Des types de caractéristiques et d'exigences à attendre des résultats

L'IA permet de répondre à des questions mais, comme pour tout système informatique, il ne s'agit pas d'atteindre ces finalités n'importe comment. On attend des résultats certaines caractéristiques, comme la justesse. Par ricochet, les systèmes à base d'intelligence artificielle qui produisent ces résultats, sont soumis à un certain nombre d'attentes. Parmi ceux-ci :

- la robustesse — les résultats doivent être résistants aux attaques et aux erreurs
- la transparence – il est possible d'avoir une vue complète du système : tous les aspects de l'IA peuvent être visibles et étudiés pour analyse – cela comprend l'explicabilité des résultats (un humain peut-il expliquer comment une décision en particulier a été prise ?) et leur intelligibilité (un humain peut-il comprendre les décisions d'une IA en analysant le fonctionnement de celle-ci ?)
- l'équité : les modèles sont créés sans biais et ne produisent pas de discriminations.

En effet, les résultats sont issus de raisonnements et de techniques qui peuvent présenter des failles. Par exemple, le raisonnement inductif ne permet pas d'identifier des évènements rares, que ces derniers soient bénins ou importants. Si l'on entraine une IA à reconnaitre des tigres avec des photos de tigres roux, elle n'identifiera pas un tigre blanc comme un tigre mais plutôt potentiellement comme un zèbre ; c'est un faux négatif. Si l'on cherche à identifier les ballons de rugby dans des photos, une IA peut remonter une photo de gâteau d'anniversaire car il a une forme de ballon ; c'est un faux positif. Certaines de ces failles découlent des biais : si les données ne représentent pas correctement les différentes possibilités, les prévisions ou les classifications issues du modèle seront biaisées et fausses. C'est le cas dans l'exemple du tigre. Un autre exemple significatif de biais se trouve dans les IA de reconnaissance faciale : les bases de données d'entrainement présentant plus de visages d'hommes blancs, les femmes noires sont moins bien identifiées. C'est pourquoi toute IA présente un niveau de maturité dans l'apprentissage qui se traduit par un taux d'erreur donné. Plus ce dernier est faible, plus l'IA est fiable.

Les conséquences d'erreurs au niveau de chacune des exigences sont à apprécier au cas par cas en fonction de leur contexte. Le niveau de confiance souhaitée sera fonction de cette évaluation. Il pourrait conditionner le moment opportun de leur déploiement opérationnel.

2.2.3. Des types de spectres de problèmes à résoudre

Les systèmes à base d'intelligence artificielle peuvent aussi se définir en fonction des types de spectres de problèmes à résoudre, de disciplines à étudier.

Des règles formelles

Ce sont des problèmes de mathématiques ou de jeux. L'IA aide à la rédaction de démonstrations compliquées (comme en 2014 où Thomas Hales utilisa l'IA pour démontrer la conjecture de Kepler[14]), à la démonstration automatique où l'IA recherche seule une démonstration[15] ou encore la suggestion automatique d'idées, de théorèmes et de conjectures à démontrer. Ce type de problèmes est plus médiatisé dans le cas des jeux. La victoire aux échecs de Deep Blue contre Garry Kasparov en 1997 est historique. À l'origine, ces problèmes sont résolus plus par l'utilisation d'une puissance de calcul brute plus que la simulation d'une réflexion similaire à celle de l'homme. Les algorithmes évoluant, cela devient de moins en moins le cas et les machines peuvent affronter des humains dans des jeux plus complexes, comme des jeux vidéo de stratégie en temps réel (StarCraft II), des jeux de tir à la première personne (Quake III) ou le jeu de Go où, en 2017, le programme AlphaGo bat le champion du monde Chinois Ke Jie. Si les avancées dans le domaine a priori superficiel des jeux est autant mis en avant, c'est parce qu'il est considéré comme l'étalon des avancées du domaine de l'IA.

Le traitement médiatique de ces prouesses technologiques peut faire parfois oublier les deux autres types de problèmes que sont les problèmes de tâches d'expertises et les tâches courantes.

Des tâches d'expertise

Les tâches d'expertise s'intéressent entre autres aux analyses financières et scientifiques, à l'ingénierie et aux diagnostics médicaux. Faut-il accorder un crédit immobilier à telle personne, quel sera le prix de la cotisation d'assurance pour un prospect ? Cette radiographie présente-t-elle une anomalie ? Ce sont des questions qu'on pose à l'IA dans ce type de tâches.

Des tâches courantes

Les tâches courantes correspondent à des tâches triviales pour un humain – perception, sens commun, raisonnement, langage, mobilité – mais qui sont extrêmement complexes pour une machine. Voici quelques exemples de problèmes posés dans les tâches courantes : est-ce un chien sur cette image ? Quelles sont les photos similaires à celle de ce sac à dos ? Comment traduire ce texte ? Les robots ont des problèmes

---

[14] Ce problème cherche à trouver l'arrangement compact de sphères identiques le plus dense. Thomas Hales a identifié les variations possibles à analyser puis à vérifier la densité d'empilement par une approche numérique. Si le lecteur désire en apprendre plus, il peut consulter l'article de vulgarisation du magazine « Pour la science » à cette adresse : https://bit.ly/3j3inDk ou consulter l'article scientifique ici : https://bit.ly/3K5vfVp

[15] Les démonstrateurs automatiques sont parfois utilisés commercialement pour vérifier le bon fonctionnement des circuits intégrés. AMD et Intel vérifient ainsi que la division est correctement implémentée dans leurs processeurs

spécifiques comme la marche bipède, le remplissage d'un verre d'eau, la conservation de l'équilibre après un choc, le repérage dans l'espace...

Les impacts potentiels pour ces différents types de spectres de problèmes à résoudre ne sont pas de même nature. Les techniques utilisées peuvent être plus ou moins opaques. Le niveau de confiance qui en résulte est à appréhender au cas par cas.

2.2.4. Des types d'outils, de modèles et techniques à utiliser

Les solutions comme les systèmes de reconnaissance faciale, les chatbots, les voiture autonomes, Parcousup, etc. sont construites à partir d'outils. Il s'agit par exemple de la vision artificielle, la traduction automatique, la segmentation automatique... Les solutions sont généralement construites à partir d'un assemblage de ces outils. Par exemple, la voiture autonome utilise (entre autres) des outils de vision artificielle, des systèmes experts et des réseaux multi agents.

Les outils sont construits sur des techniques. Les différences entre les outils et les techniques peuvent être floues : il s'agit en réalité de distinguer l'outil de son application. Par exemple, le deep learning — technique — est utilisé pour réaliser des outils de vision artificielle et de traduction de langage naturel. Parmi les techniques les plus connues, il y a les arbres de recherche, les méthodes statistiques, le raisonnement automatique, le machine learning, les réseaux de neurones, le deep learning, les réseaux multi-agents... Cette liste met en évidence une autre différence fondamentale : les techniques sont en fait des algorithmes. On peut classer les systèmes d'IA en fonction des algorithmes utilisés.

Chacune de ces techniques apporte son lot de problèmes. Par exemple, le deep learning est plus opaque et il en résulte une difficulté accrue pour expliquer pourquoi et comment un résultat a été établi. Cela peut avoir un impact sur le niveau de confiance que l'on peut y porter.

2.2.5. Des types de logiciels, d'algorithmes à mettre en œuvre

Rappelons l'évidence : l'IA est un domaine de l'informatique et utilise donc ses concepts et son matériel (ordinateurs, supercalculateurs, serveurs...). Des paramètres d'entrées, comme les données, sont utilisés par les différents algorithmes d'un programme pour fournir un résultat. Philippe Flajolet[16], définit un algorithme[17] comme « une façon de décrire dans ses moindres détails comment procéder pour faire quelque chose : trier des objets, situer des villes sur une carte, multiplier deux nombres... ». Il s'agit d'une suite finie et non ambigüe d'instruction et d'opération permettant de résoudre une classe de problèmes. Les algorithmes parmi les plus courants sont les recettes de cuisine.

Les algorithmes peuvent être numériques (extraire une racine carrée) ou non (chercher un mot dans le dictionnaire, réaliser une recette de cuisine). Les modèles mathématiques, comme la méthode d'extraction

---

[16] Chercheur français en informatique et en mathématiques à l'INRA, Institut national de recherche en sciences et technologies du numérique

[17] https://interstices.info/quest-ce-quun-algorithme/

d'une racine carrée, utilisent des équations, des constantes, des variables, des conditions et des hypothèses explicites ou implicites. Il est souvent possible d'utiliser différents algorithmes – ou techniques - pour résoudre un même problème – de la même façon qu'il existe plusieurs manières de cuisiner une tartiflette, ne serait-ce qu'en changeant la proportion des ingrédients. Le choix de l'un ou l'autre des algorithmes dépend des hypothèses d'entrées, de la complexité des calculs à effectuer, de la rapidité de calcul désirée, de la consommation de mémoire vive, de la précision des résultats...

Le résultat d'un algorithme sera différent en fonction de l'algorithme. Par exemple, pour un algorithme de tri, le résultat est une liste triée. Pour un algorithme d'apprentissage automatique, le résultat est un modèle d'apprentissage algorithmique ; à ne pas confondre avec le modèle mathématique. Le modèle d'apprentissage algorithmique représente ce qui a été appris par l'algorithme à partir des données d'entrainement. C'est ce modèle qu'on utilisera pour réaliser une prédiction à partir de nouvelles données.

À titre d'exemple simplifié, Parcoursup a utilisé les données des années précédentes en guise de données d'entrées dans l'algorithme de forêt d'arbres décisionnel. Après avoir entraîné cet algorithme avec cet ensemble de données, un modèle a été conçu. Ce modèle remarque par exemple que, pour une licence en mathématiques, les candidats les mieux classés sont bons en mathématiques et en philosophie et les moins bien classés faibles sur ces matières. Le modèle en conclut qu'il vaut mieux donner un poids important aux mathématiques et à la philosophie pour ce cursus. Lorsque les dossiers des nouveaux élèves arrivent (nouvelles données), le modèle trie les dossiers en priorité en fonction des notes sur ces matières.

Autre exemple schématique avec les véhicules autonomes : en analysant des vidéos de piétons qui traversent intempestivement (sans respecter les feux), l'algorithme d'apprentissage apprend qu'un grand nombre de ces cas arrive alors qu'il y avait un panneau « attention école » dans les environs. Le modèle associe la présence de ce panneau à un risque plus élevé de traversée piétonne inattendue et recommande de ralentir dans ces situations. Lors d'un nouveau trajet, dès que le système vidéo détecte un panneau « attention école » (nouveau jeu de données) et qu'il est reconnu comme tel, le modèle se met en œuvre et fait ralentir le véhicule.

Les erreurs peuvent trouver leur origine dans le modèle, dans le choix des variables ou d'absence de variable, dans les paramètres, dans le programme... Les impacts ne sont pas les mêmes. Le niveau de confiance souhaité devra être apprécié en fonction de ces éléments.

2.2.6. Des types de raisonnements à mener

Les algorithmes traduisent donc des raisonnements. C'est en partie dans l'utilisation de ces raisonnements que se manifeste l'aspect de simulation de l'intelligence de l'IA. Ces raisonnements peuvent être de tous types et même mixés entre eux :

- déductifs : obtient une conclusion particulière à partir d'affirmations générales, comme des règles et des savoirs accumulés. Ces raisonnements représentent les connaissances avec des objets et des symboles formels qui associent les connaissances entre elles. On peut résoudre les problèmes mathématiques et d'optimisation grâce à ce type de méthodes

- inductifs : recherche des lois générales à partir de faits particuliers. L'induction permet, à partir d'observations, de produire des propositions générales qui seront ensuite testables. On peut ainsi réaliser des prévisions et des généralisations. Ce sont des méthodes le plus souvent probabilistes
- abductifs : établit la cause la plus probable d'un fait observé. C'est le type de raisonnement utilisé en médecine pour faire des diagnostics...

D'autres raisonnements existent, comme le raisonnement par analogie, par récurrence...

Ce sont ces raisonnements, utilisés, seuls ou en combinaison, dans l'apprentissage des humains, qui sont aussi utilisés pour les IA. Chacun de ses raisonnements peut contribuer à faire avancer l'apprentissage mais il apporte son lot de risques et peut ainsi conduire à des erreurs ou à des biais. Ainsi, un raisonnement de corrélation n'aboutit pas nécessairement à une causalité.

Il s'agit donc de bien en être conscient et d'identifier les limites et les risques qui pourront avoir un impact sur le niveau de confiance correspondant à chacun de ces types de raisonnements pour un contexte donné.

2.2.7. Des types de formats et de sources de données à couvrir

La donnée est un pilier de l'IA. Sans la donnée, l'IA ne peut pas fonctionner puisque l'essence de l'IA est d'analyser, de traiter de bien plus grandes quantités de données que ce que l'humain n'est capable de traiter lui-même pour en tirer des conclusions et des enseignements. Est-ce là une des raisons de la croissance de l'IA ? Notre société produit de plus en plus de données grâce à l'essor de l'IoT (internet des objets), de l'usage intensif d'Internet, des mails et des réseaux sociaux ou de l'utilisation de plus en plus formalisée de la donnée dans les entreprises. L'usine 4.0 et les différents processus qualité, entre autres, sont très générateurs de données. Les données peuvent être industrielles, économiques, publiques, personnelles... Cette masse d'information croissante nécessite des systèmes de traitement et de stockage de plus en plus performants. L'IA est l'outil rêvé pour leur interprétation. Les traitements que nécessitent la donnée pour être exploitable sont nombreux et dépendent de l'utilisation qu'on en projette. Les programmes peuvent être différents en fonction de la donnée d'entrée.

Par exemple, certains algorithmes nécessitent des données d'entrées structurées (les données sont présentées dans un tableau par exemple) tandis que d'autres travaillent sur des données non structurées. Les programmes de traitement du langage naturel cherchent à interpréter un texte écrit par un humain. Or, les textes sont typiques des données non structurées (elles ne peuvent pas être mises dans des tableaux).

Parmi les sources de données existantes, il est impératif de s'attarder sur celle des capteurs. Lorsque les IA sont intégrées dans des systèmes physiques, il est nécessaire de leur offrir des données sensorimotrices parmi les différentes données d'entrées. Ces données sur l'environnement immédiat de l'IA sont obtenues grâce aux capteurs. Ils sont de nombreux types : accéléromètres, caméra de différentes longueurs d'ondes (visible, infrarouge, UV...), capteurs piézoélectriques, de pression, de température... Il est aujourd'hui possible de mesurer quasiment toutes les grandeurs physiques. Ces capteurs sont massivement présents dans nos vies : tous les smartphones récents contiennent au moins un accéléromètre et une caméra.

Ces capteurs sont la clé d'entrée de l'internet des objets. Les données captées peuvent être envoyées dans le nuage (cloud) pour être analysées et traitées. Pour des raisons d'économie de bande passante et d'énergie, les méthodes de traitements en périphérie se diffusent. Dans ce cas, les données sont traitées directement dans l'objet qui a capté les données et seul le résultat est envoyé dans le cloud.

Le niveau de précision de ces capteurs n'est pas le même en fonction du type de capteurs mais le besoin en précision n'est pas le même en fonction des situations. Le niveau de confiance attendu doit donc être évalué au cas par cas.

Ces différents types de données apportent leur lot de risques spécifiques. Le niveau de confiance souhaité est à évaluer pour chaque situation.

2.2.8. Des types de traitements à effectuer sur les données

Le traitement de la donnée est une étape préliminaire indispensable de l'IA. La première phase est l'extraction de la donnée. Il s'agit du processus de collecte de données de types disparates à partir de différentes sources. La voiture autonome récupère les données des différentes caméras, des centrales inertielles, des radars... La phase suivante est la phase de nettoyage. Certaines données peuvent être mal structurées ou porter une erreur. C'est le cas lorsqu'un reflet sur la caméra d'une voiture autonome produit une image toute blanche ou qu'un capteur s'est trouvé en-dehors de sa plage de fonctionnement comme un thermomètre en pleine canicule alors qu'il ne peut mesurer que jusqu'à 40°C. Les sources d'erreurs sont très nombreuses. L'algorithme est conçu pour fonctionner avec des données d'entrée d'un certain format. Il faut donc supprimer ou corriger toutes les données d'entrées qui ne correspondent pas à ce format. Certains types de données nécessitent des traitements supplémentaires. C'est le cas en particulier des données personnelles qui sont anonymisées[18].

Notons toutefois, avec l'amélioration de la protection des données personnelle, l'apparition du principe de minimisation où l'on cherche à utiliser le moins de données possibles. Cela entre en contradiction avec la nature de l'IA qui a besoin de nombreuses données pour acquérir de nouvelles connaissances.

Des risques spécifiques sont associés à chacun de ces types de traitement. Le niveau de confiance correspondant sera donc à apprécier en fonction du contexte.

2.2.9. Des types de rôles à jouer pour les données

Dans la vie d'un système à base d'intelligence artificielle, on peut trouver trois grands types de données suivant l'utilisation qu'on en fait :

- les données d'entraînement qui servent à paramétrer leur modèle
- les données de test qui servent à valider le modèle

[18] L'utilisation de ces données particulières est définie par la Directive 95/46/CE et précisée par le RGPD (règlement Union européenne 2016/679) et ne sera pas étudié dans ce document.

- les données de production qui sont utilisées pour les faire tourner.

Ces données sont extraites des bases via un processus d'échantillonnage qui permet de sélectionner des sous-ensembles de données représentatives.

De même que pour les autres dimensions, des risques spécifiques sont présents pour chacun de ces types d'utilisation et ceci doit être pris en compte dans le niveau de confiance qui en découle.

2.2.10. Des types de matériels et de réseaux à solliciter

L'informatique moderne, avec la démocratisation de l'internet et du cloud peut apparaître à beaucoup comme quelque chose d'éthérée et impalpable. Cette sensation peut être facilement ressentie lorsqu'on a affaire à des IA présentes sur des sites internet comme les chatbox. Pourtant, la composante matérielle est cruciale. Au niveau des solutions, la voiture autonome se base d'abord sur une voiture, les robots sur des systèmes mécatroniques, les assistants domotiques vocaux se présentent sous la forme de boitiers. Les données sont acquises par des capteurs physiques, comme présentés plus haut, et stockées dans des serveurs dédiés. La phase de calcul logicielle repose également sur du matériel. Les calculs puissants nécessaires à la résolution des problèmes de règle formelle ne peuvent être réalisés que sur des supercalculateurs. L'augmentation du nombre de ce type d'ordinateurs et de leur puissance ces dernières années est un des signes de l'importance cruciale du pilier matériel de l'IA.

Mais les systèmes à base d'intelligence artificielle d'aujourd'hui seraient très différents sans les réseaux. L'IoT se base sur le wifi, la 5G, le bluetooth pour transmettre des données. De nombreuses solutions d'IA sont accessibles via des solutions SaaS, ce qui signifie que les calculs sont réalisés dans des datacenters et consultables à distance sur les terminaux personnels (ordinateurs, smartphones, tablettes…) grâce à la qualité des réseaux internet.

Chacun de ces éléments matériel et réseau présentent des risques spécifiques qu'il convient de bien identifier et traiter pour que l'on puisse disposer d'IA dignes de confiance.

Dans cette sous-partie (§ 2.2), il est apparu que le périmètre d'utilisation et d'application des IA est très large et rend ces dernières difficiles à définir. Pourtant, il serait possible d'analyser tous les systèmes à base d'Intelligence artificielle suivant les dix grilles d'analyse que nous avons décrite. Chaque système présente ses propres résultats à obtenir, ses propres exigences à résoudre, ses propres outils à utiliser…

2.3. Des puissances de traitement et de communication, des logiciels et des données avec des apports, des limites et des risques spécifiques à intégrer pour une confiance globale

La compréhension de ces grilles d'analyse de manière séparée et indépendantes entre elles n'est pas suffisante pour faire d'un système une IA digne de confiance. Il faut non seulement le prendre dans sa globalité mais être conscient des différences intrinsèques entre la machine et l'homme, dont elle a pour but de copier l'intelligence.

Une approche intégrée est nécessaire à l'obtention d'une confiance globale

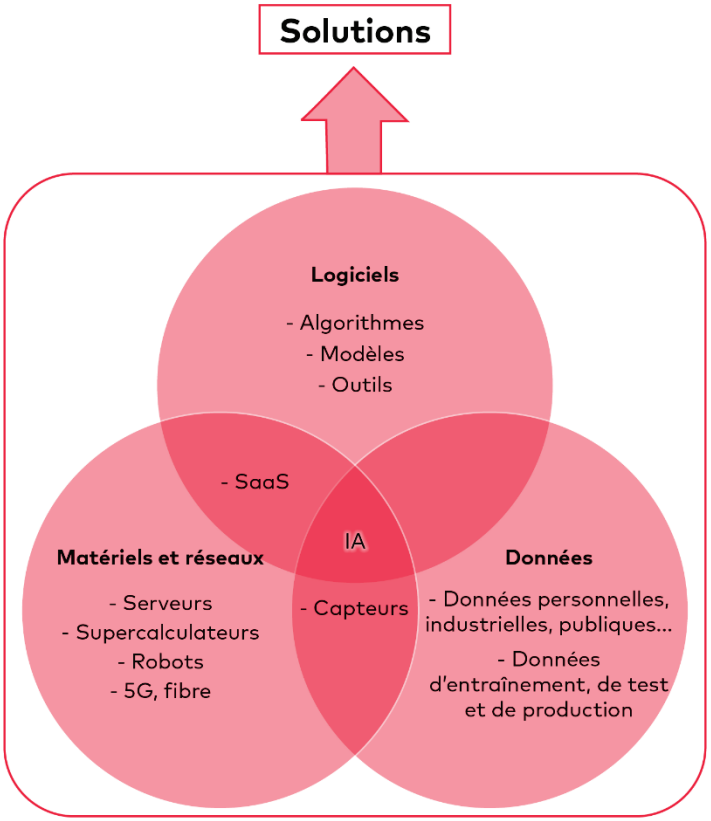

Figure 2 : Une autre représentation de l'IA – les trois piliers de l'IA

L'IA peut être représentée comme reposant sur trois piliers : logiciels, matériels et réseaux et données. La qualité d'un système d'IA, dépend de la qualité de chacun de ces piliers : le supercalculateur le plus puissant ne sert à rien s'il doit utiliser des données pauvres et mal nettoyées. Pour que chacun de ces piliers soit pertinent, il faut qu'il réponde favorablement aux trois critères suivants : une bonne qualité intrinsèque, une bonne qualité contextuelle et un bon accès. Par exemple, les données ne doivent pas contenir d'erreur (qualité intrinsèque), elles doivent être riches pour représenter l'éventail des possibilités pour son objet (qualité contextuelle).

Dans le cadre d'un logiciel de reconnaissance faciale, les données d'entrainement représentent des humains de tous les genres et origines ethniques possibles en quantité suffisante. L'accès des bases de données sources doit être possible. Par exemple, la base de données du FBI contient 411 millions de photos et Facebook 3 milliards mais l'accès au premier est réservé et l'accès au second est contrôlé par la RGPD. Il doit aussi être suffisamment souple : les archives nationales sont publiques mais pas dématérialisées par exemple. Ce raisonnement s'applique aussi au matériel – sans l'accès à des supercalculateurs, certaines applications d'IA sont impossibles – et au logiciel : un modèle qui ne prend pas en compte la richesse des possibilités risque de ne pas être efficace ni fiable. Il est important de noter également que la confiance que l'on cherche à obtenir des systèmes à base d'Intelligence artificielle ne sera atteignable qu'en ayant confiance dans chacun de ces piliers. Il faudra donc avoir confiance dans les logiciels mis en œuvre, dans les données récoltées et utilisées mais aussi dans les matériels et réseaux employés.

Plusieurs capacités intrinsèques à l'humain non présentes pour les IA peuvent limiter leur pertinence et fiabilité et donc leur niveau de confiance

L'IA évolue très rapidement. Dans son objectif de simulation de l'intelligence humaine, la machine dépasse l'homme sur plusieurs points comme la vitesse de traitement ou la quantité d'informations prises en compte. Il ne s'agit en revanche que d'une imitation. La machine n'est pas biologique. Si l'on parle d'un neurone en IA, il s'agit d'un objet logiciel qui récupère des variables numériques pondérées en entrée et les combine pour générer une valeur à la sortie alors qu'un neurone biologique assure la transmission d'un signal bioélectrique. L'IA ne semblerait pas pouvoir atteindre la simulation de nombreuses fonctions biologiques car il existe de nombreuses différences intrinsèques entre l'IA et l'Homme. Comment l'IA pourrait-elle simuler le système hormonal ou même d'autres fonctions portées par certains mécanismes anciens et peu connus comme l'instinct ou les réflexes de survie ? De même, l'atteinte d'une IA avec un niveau d'intelligence équivalente à l'Homme qui pourrait raisonner de manière polyvalente apparaît de plus en plus comme un mythe pour de nombreux spécialistes comme Yann Le Cun[19] ou Luc Julia[20]. Un niveau supérieur d'IA qui intégrerait conscience, sentiment, sagesse, bon sens et connaissance de soi semblerait encore plus inatteignable. D'autres spécialistes pensent au contraire que cela constitue l'avenir des IA. Les sentiments sont actuellement pris en compte via les techniques d'analyse de sentiments (dans les textes, sur une image) mais il semblerait que la machine ne vivra pas ses propres sentiments. Actuellement, si un

[19] Chercheur en intelligence artificielle et vision artificielle (robotique) français, prix Turing 2018.

[20] Luc Julia (ingénieur et informaticien franco-américain, spécialisé dans l'intelligence artificielle, un des concepteurs de l'assistant vocal Siri)

tweet « Youpi ! J'ai réussi ! » peut être identifié comme une expression de joie, l'algorithme ne ressent pas lui-même cette joie. Savoir si ce sera un jour le cas ou non reste une grande inconnue.

Ces différentes spécificités humaines le plus souvent instinctives contribuent fortement aujourd'hui aux décisions et aux actions qui sont prises par les humains. Le fait qu'elles ne soient pas prises en compte par les IA ne limite-t-elle pas la confiance que l'on pourrait avoir dans les décisions et les actions issues de ces IA. Si oui, dans quels cas et dans quels contextes ?

Le seul champ où les différences intrinsèques entre l'IA et l'humain s'effaceront, sera peut-être celui de l'intelligence augmentée où l'Homme saura utiliser efficacement les technologies de l'information pour augmenter sa propre intelligence.

Une approche différenciée est à mettre en place

Bien que ces catégorisations ne soient pas totalement justes car simplifiées pour faciliter la compréhension (une notion que nous retrouverons justement plusieurs fois dans la suite de ce document), il est essentiel de bien saisir les principes des systèmes d'IA mais surtout de leurs différences. En effet, ces différences impliquent des apports et des risques différents ; cela signifie également que les dispositifs à mettre en place pour parvenir à créer des IA dignes de confiance pourront eux aussi présenter des différences en fonction des caractéristiques de chaque type d'IA.

Le schéma suivant est un résumé du début de cette sous-partie. À gauche, il rappelle qu'un système à base d'IA est composé des trois piliers logiciels, matériels et données dans le but de fournir des solutions. À droite, on retrouve les dix grilles d'analyse d'une IA avec quelques exemples pour chacune d'elles. Chacun de ces points pris isolément, mais également dans leur globalité, est susceptible d'impacter le niveau global de confiance.

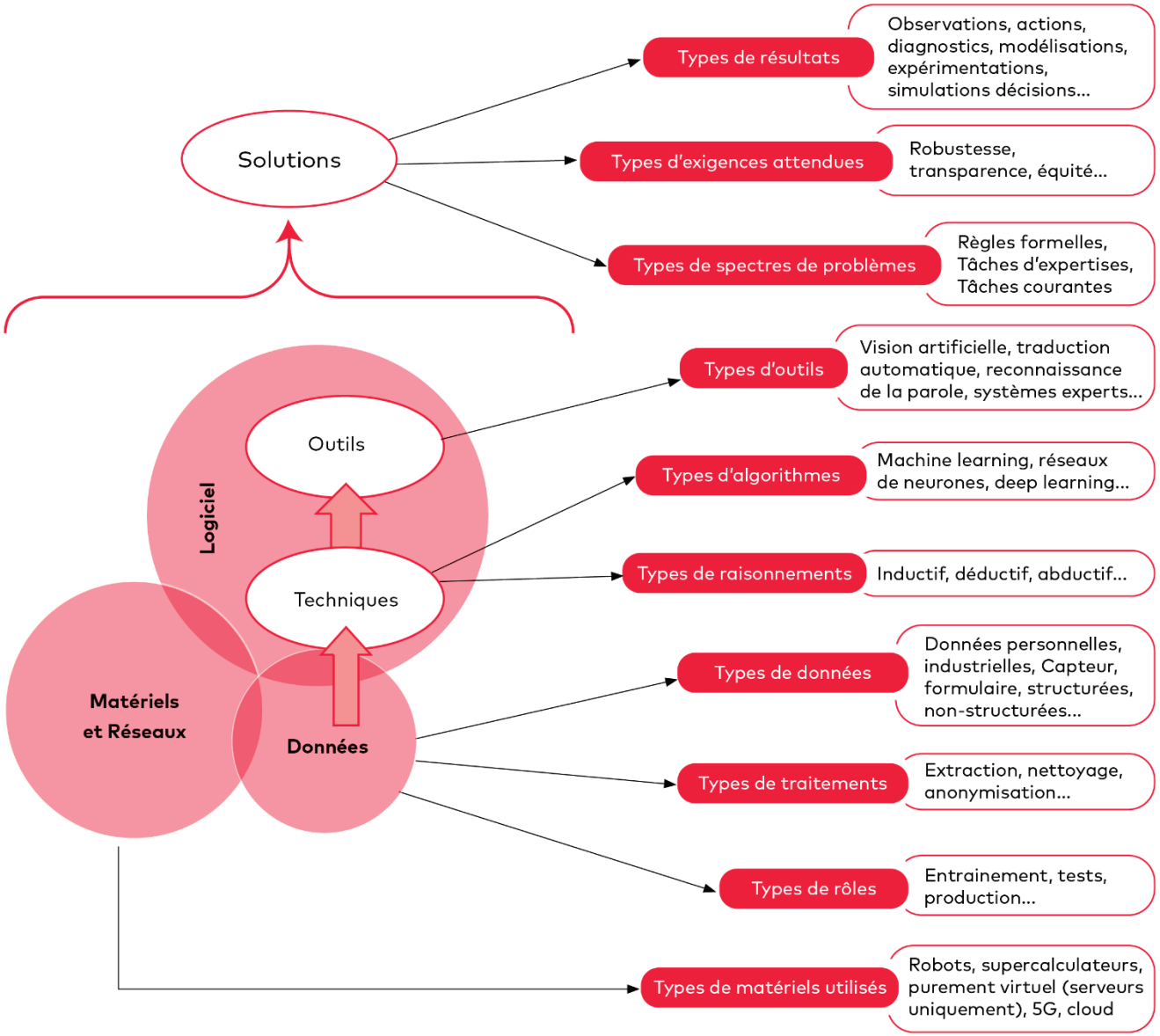

Figure 3 : une vue d'ensemble de l'IA et classifications possibles

3. Des puissances de traitement et de communication, des logiciels et des données avec des apports en évolution, des limites et des risques spécifiques à intégrer pour une confiance globale

L'IA évolue et se transforme depuis plusieurs décennies. Tout indique que le développement de l'IA va s'accélérer dans les prochaines années, d'abord en raison d'un engouement général envers cet ensemble de technologies mais aussi parce que les piliers de bases de l'IA se développent eux-mêmes.

Les données sont de plus en plus nombreuses : on prévoit, en 2025, de produire dans le monde entier un volume de 175 zettaoctets de données produites (1021 octets soit un milliard de téraoctets ou mille milliards de milliards d'octets[21]) contre 33 zettaoctets produits en 2018, soit une progression de 430 % en seulement sept ans. Avec la baisse des coûts des capteurs, le nombre d'objets équipés croît : les drones sont maintenant un équipement courant, les smartphones disposant de cinq caméras ne sont pas rares, les municipalités investissent dans le matériel de surveillance et de contrôle « intelligent » et donc producteur de données. L'IoT est partout. Les sources de données purement logicielles augmenteront : réseaux sociaux, sondages, formulaires, logs et datalakes d'entreprises... La philosophie émergente de l'open data, soit la mise à disposition des bases de données au public, principalement par les organismes gouvernementaux mais aussi par certaines sociétés privées, facilite l'exploitation des données et l'amélioration de l'IA.

Si la captation de données s'améliore, le volume de stockage nécessaire est également croissant. Avec l'amélioration de la rapidité et de la fiabilité des connexions internet, le stockage en nuage s'est répandu, décorrélant pour l'utilisateur l'usage d'un espace de stockage de son investissement en matériel, investissement désormais supporté par des sociétés spécialisées. Ces sociétés peuvent ainsi s'équiper de serveurs de plus en plus rapides, volumineux et fiables. Les équipes de recherche misent sur l'utilisation de nouvelles technologies pour augmenter les capacités de stockage : ordinateur quantique et optique*, spintronique*, stockages optique 3D*, holographique*, ADN*, magnétiques* (HAMR, BPM, SMR).

L'accès à ces stockages est également facilité avec l'amélioration des réseaux (démocratisation de la fibre optique, début de la 5G/6G) mais aussi l'apparition de nouveaux modèles économiques comme le SaaS et le HaaS (respectivement Software et Hardware as a Service) qui permettent aux utilisateurs de payer en fonction de leur usage. L'utilisation et la conception d'IA est ainsi plus accessible financièrement.

Outre le pilier données, l'amélioration du matériel va soutenir le développement de l'IA, comme par le passé. Les supercalculateurs atteignent des puissances de calcul inégalées : en 2020, Fugaku[22] devient le superordinateur le plus puissant du monde avec une puissance de 415 pétaflops soit 415 millions de milliards d'opérations par seconde ! L'arrivée prévue des ordinateurs quantiques ou optiques permettra de dépasser encore ces puissances de calculs, ce qui impliquera que les algorithmes travailleront plus rapidement avec de plus grandes quantités de données dans des laps de temps plus raisonnables. Ils pourront par ailleurs

---

[21] Pour comparaison, une page de texte A4 d'environ 3000 caractères correspond à 3000 octets de mémoire

[22] Fugaku est un supercalculateur japonais, développé par Fujitsu pour le compte de l'institut scientifique japonais RIKEN

impacter l'efficacité des techniques de cryptographie, qui sont un outil essentiel au respect de la vie privée et de l'intégrité des données.

Le troisième pilier de l'IA ne sera pas en reste et poursuivra son développement : de nouveaux algorithmes apparaîtront sans nul doute tandis que les algorithmes existants seront améliorés pour devenir plus performants. Les innovations algorithmiques connaîtront une diffusion rapide avec le développement de la culture de l'open source.

Les ressources humaines pour travailler et manipuler l'IA seront également plus répandues. Les cursus scolaires se mettent en place. Les data scientistes*, data analystes*, data architectes*, ingénieurs machine learning* et autres ingénieurs de l'IA seront formés en plus grand nombre chaque année.

Enfin, le développement de l'IA s'appuiera sur la mise à disposition sur le marché de produits ou de briques logicielles prêtes à l'emploi : utilisation du superordinateur Watson[23], achat ou SaaS de chatbots*, de robots, d'outils de segmentation, de marketing prédictif*... L'immense majorité des IA produites est un assemblage de techniques et de paramétrages d'outils assez standards adaptés à une base de données pour répondre à un problème métier particulier. Le choix, la combinaison et l'assemblage de ces briques seront facilités par la multiplication de modèles, d'outils et d'offres de services.

Tous ces éléments sont critiques au développement des IA. Nous ne pourrons avoir confiance dans ces IA et dans leurs résultats qu'en ayant confiance dans le développement et la mise en œuvre maîtrisée de ces éléments. C'est pour cela qu'il convient de les identifier et traiter spécifiquement.

Pour conclure ce chapitre, nous avons identifié plusieurs facteurs « environnementaux », plusieurs spécificités relatives aux IA, plusieurs vues possibles pour ces éléments, plusieurs évolutions dans les capacités informatiques qui chacun apporte son lot de contributions potentielles mais aussi son lot de risques.

Pour nous permettre de disposer d'IA de confiance, nous avons besoin d'avoir confiance dans le fait que tous ces éléments critiques et toutes ces spécificités ont été traités de manière satisfaisante. La diversité et la complexité du paysage des IA telles que présentées ci-avant conduit à la nécessité d'une démarche adaptée pour s'assurer de ce traitement satisfaisant. Cette démarche se doit d'être simple, souple et flexible, facile à mettre en œuvre et à comprendre.

À cet effet, nous devons pouvoir disposer de modèles et de classifications appropriés qui permettent d'aborder efficacement la nature des apports, des risques, des solutions... Catégoriser est un élément essentiel à une meilleure compréhension et appréhension des différents phénomènes complexes des IA.

---

[23] Watson est un programme informatique d'intelligence artificielle conçu par la société IBM dans le but de répondre à des questions formulées en langage naturel. Il a connu une notoriété mondiale en gagnant au jeu télévisé américain Jeopardy ! qui consiste à trouver la question correspondant à trois indices donnés.

CHAPITRE 2

DES ENJEUX, DES INQUIÉTUDES ET DES OBSTACLES AU DÉVELOPPEMENT ET À L'ADOPTION DES SYSTÈMES À BASE D'INTELLIGENCE ARTIFICIELLE

La fabrication de tissu a toujours été importante car se vêtir est un besoin primordial. L'apparition des métiers à tisser Jacquard a permis d'améliorer le rendement des ateliers de fabrication, permettant ainsi de diminuer les coûts de production et de vente. Ils portaient aussi un enjeu social. Le concepteur de cette machine avait espéré que cela diminuerait le travail des enfants qui aidaient leurs parents tisseurs, les ouvriers de la soie y voyaient une cause de chômage. La révolte des Canuts de Lyon qui a causé 169 morts en 1831 est une conséquence du refus des Canuts de s'approprie cette technologie.

De cet ancêtre du robot à l'intelligence artificielle, les questions sont les mêmes. Que faut-il faire pour éviter une condamnation de ce type de systèmes ? Quels sont les enjeux sociétaux, économiques, moraux à considérer ? Quelles sont les inquiétudes à apaiser ? Quels barrages vaincre et comment pour faire accepter l'IA à la société et ainsi ne pas passer à côté du potentiel qu'elle porte ?

1. De forts enjeux économiques, sociaux, géopolitiques, démocratiques, éthiques... à prendre en compte

Avec le développement du domaine, l'IA pourra apporter des réponses ou des pistes de réponses à des problèmes de plus en plus variés. Il ne parait pas utopique d'envisager que l'IA, avec la possibilité de prise en compte de très nombreux paramètres, puisse suggérer des actions à prendre et faciliter la résolution des défis planétaires comme le réchauffement climatique (adaptation des pratiques* agricoles...), la crise écologique (amélioration de la durée de vie des équipements avec la maintenance prédictive et donc diminution du besoin de ressources naturelles, meilleure connaissance de la faune, atteinte des objectifs du pacte vert Européen...), la crise de l'énergie (smartgrids), la lutte contre la pauvreté, un évitement des crises économiques via un meilleur contrôle de la finance...

Les technologies de l'IA pourront être utiles dans l'ensemble des secteurs de l'économie : depuis la finance au juridique en passant par les services publics ou les transports sans oublier le luxe, les médias, l'industrie... Cela leur permettrait d'accroître l'innovation, d'améliorer les organisations internes, d'augmenter la productivité et de développer des offres de services plus innovants, plus adaptés... L'utilisation de l'IA promet aussi une amélioration du bien-être des individus. Grâce à la meilleure connaissance de leurs habitudes et de leurs intérêts, on pourrait personnaliser les soins médicaux, offrir des services publics et des systèmes de transports plus adaptés, améliorer les communications interprofessionnelles... Toutes ces

améliorations seront possibles grâce à l'IA et indirectement, grâce à la possibilité d'accès à une grande variété de données (personnelles, économiques, industrielles, publiques...).

La recherche fondamentale, la recherche appliquée, la création puis l'utilisation de l'IA aura des impacts sociaux. De nouveaux besoins émergeront tandis que l'importance des métiers actuels de l'IA (data scientiste, data analyste, ingénieur IA...) s'accentuera. Même les travailleurs utilisant ces technologies de manière basique, à la manière des utilisateurs « pousse-bouton » des logiciels d'aujourd'hui, auront besoin de gagner en compétences et de se former. La Commission européenne envisage de réviser le plan d'action en matière d'éducation numérique afin de sensibiliser à l'IA les futurs citoyens à tous les niveaux d'enseignements et de les préparer à vivre avec des décisions de plus en plus influencées par l'IA.

La Commission européenne, en date du 19 février 2020, dans son livre blanc, stipule que l'IA doit améliorer le quotidien des citoyens tout en respectant leurs droits. Elle doit donc suivre une certaine éthique et être fondée sur les valeurs et les droits fondamentaux tels que la dignité humaine, le pluralisme, l'inclusion, la non-discrimination et la protection de la vie privée et des données à caractère personnel.

Or, les particularités des technologies de l'IA, notamment l'opacité (ce qui est communément appelé « effet de boîte noire* »), la complexité, l'imprévisibilité et le comportement partiellement autonome, mettent à mal ces principes démocratiques. De nombreuses questions se posent : comment le citoyen ou l'utilisateur peut-il être certain d'avoir accès à toutes ces informations ? A-t-il bien donné son consentement au recueil de toutes ces informations et à leur utilisation ? Comment peut-il être sûr que les résultats ne présentent pas de biais et n'ont pas été soumis à une influence quelconque ?

Certaines IA actuelles portent déjà atteinte à ces valeurs à cause de failles dans la conception ou de l'utilisation de données sans correction de biais éventuels. Mais les caractéristiques de l'IA rendent difficile la vérification de la conformité aux règles du droit de l'Union européenne en vigueur destinées à protéger les droits fondamentaux et peuvent entraver le contrôle de l'application de celles-ci.

Une autre difficulté existe. Les personnes concernées et les autorités répressives ne disposent pas nécessairement de moyens suffisants pour vérifier comment une décision donnée, résultant de l'utilisation de l'IA, a été prise et, par conséquent, pour déterminer si les règles applicables ont été respectées. Les particuliers et les entités juridiques peuvent se heurter à des difficultés en ce qui concerne l'accès effectif à la justice lorsque ces décisions sont susceptibles d'avoir des effets négatifs pour eux.

L'absence de règles claires en matière de sécurité pour faire face aux risques pour la sécurité des utilisateurs peut, outre les risques pour les individus concernés, créer une insécurité juridique pour les entreprises qui commercialisent leurs produits reposant sur l'IA dans l'Union européenne. Par exemple, on peut se demander qui est responsable dans le cas de la voiture autonome. La directive sur la responsabilité du fait des produits prévoit qu'un fabricant est responsable des dommages causés par un produit défectueux. Or, avec un système fondé sur l'IA tel que la voiture autonome, il peut être difficile de prouver la défectuosité du produit, le dommage survenu et le lien de cause à effet entre les deux.

Les autorités de surveillance du marché et les autorités répressives peuvent se trouver dans une situation où elles ne savent pas si elles peuvent intervenir parce qu'elles ne disposent pas des capacités techniques appropriées pour inspecter les systèmes et/ou qu'elles n'y sont peut-être pas habilitées.

La question des prises de décision et des responsabilités reste encore en suspens. D'autres points de gouvernance* sont encore à établir comme la mise en place de personnalité juridique des IA et la résolution des insécurités juridiques.

Enfin, l'IA est un enjeu géopolitique majeur. Elle est devenue un des étalons de la puissance d'un pays et accentue la rivalité entre les États-Unis, la Chine et l'Europe. La question du droit applicable est majeure : Amazon considère que toute donnée utilisée dans ses algorithmes devient américaine avec l'application du droit afférent. Et que dire de la souveraineté des données lorsque les activités liées à l'IA sont sous-traitées dans un pays tiers ? Quelle est la localisation juridique appropriée dans le système de stockage globalisé et éclaté que permet le cloud ou les techniques de la blockchain ?

L'IA est aussi le reflet des visions et des cultures locales : la Chine s'occupe bien peu de libertés individuelles, valeur moins importante que celle de l'importance du groupe (« l'harmonie de la société »), alors que c'est un sujet prépondérant dans le monde occidental, particulièrement au Canada et en Europe. Ce contexte souligne bien la difficulté d'élaboration d'une vision commune sur ce qu'est une IA de confiance.

Il est donc nécessaire d'adresser de manière satisfaisante ces différents enjeux pour pouvoir donner confiance.

2. De nombreux fantasmes, inquiétudes, réticences, freins... à lever

Multiforme, se transformant vite, utilisant des mathématiques de haut niveau, l'IA reste mal comprise même pour les professionnels du numérique. La méconnaissance provoque de nombreux quiproquos et crée des mythes souvent basés sur la science-fiction.

La requête « IA+peur » sur le moteur de recherche Google remonte plus de 3,8 millions de résultats, c'est dire si le sujet inquiète. Différentes menaces sont associées aux IA et se scindent principalement en deux catégories :

- la première est constituée par la crainte qu'une IA forte, qui n'existe pas encore, puisse tout connaître, tout contrôler et en conséquence, nous asservir. Ce fantasme est alimenté et diffusé auprès du grand public par la science-fiction. Les images apocalyptiques de Terminator et de son Skynet sont plus mémorables que les discours des nombreux spécialistes, comme Yann Le Cun et Luc Julia, qui considèrent ce scénario comme très peu probable voire chimérique ;
- la seconde catégorie de menaces est constituée par les travers de l'IA actuelle, celle qui existe déjà, et qui, bien que moins extrêmes qu'un Skynet, n'en serait pas moins lourde de conséquences pour nos sociétés, nos emplois et nos démocraties.

Une partie de ces inquiétudes est constituée par l'utilisation malveillante de l'IA. On peut citer les phishings personnalisés, la fabrication automatique de « fake news » et le chantage à grande échelle ou les « deep fakes », menace considérée comme la plus importante par les spécialistes[24], qui créent des fausses vidéos pornographiques ou de discours politiques hyperréalistes, causant des dommages à des célébrités politiques ou artistiques. Les systèmes à base d'Intelligence artificielle pourraient être détournés de leur utilisation primaire ou abusés : les voitures autonomes pourraient être piratées et transformées en voitures-béliers, les systèmes de sécurité à base de reconnaissance faciale laisser passer un individu suspect ou être détournés pour stalker[25] une victime...

Les systèmes d'IA sont fragiles et les cyberattaques, les atteintes à la vie privée, à la dignité et à l'équité seraient plus probables qu'avec d'autres systèmes. De manière anecdotique, on peut citer le cas du chatbot Tay développé par Microsoft, influencé par des personnes mal intentionnées. Son apprentissage automatique basé sur les conversations avec ces utilisateurs trolls l'a fait évoluer jusqu'à lui faire tenir des propos éthiquement inacceptables, et présenter un comportement raciste et misogyne.

Sur ces points, le débat sur les dangers de l'IA est le même que celui de la dynamite : l'IA n'est qu'un outil et c'est son utilisation malfaisante par l'homme qui la corrompt.

Il existe d'autres risques, même dans un usage bienveillant et bénéfique de l'IA. Le fonctionnement même de l'IA présente des menaces. Les décisions sont prises de manière automatique sur la base d'éléments passés. Les règles de prises de décision peuvent évoluer automatiquement. On peut alors se poser des questions sur la potentielle perte de contrôle dans la prise de décision et l'opacité du processus. De même, il n'est pas impossible que les éléments passés utilisés associés à un processus de décision mal conçu puissent causer des discriminations. L'exemple le plus connu reste celui de l'algorithme de recrutement AMZN.O de la société Amazon qui pénalisait les CV contenant le mot « femme ». Autre exemple, les technologies de reconnaissance faciale sont plus performantes sur des visages à peau blanche que pour des carnations plus sombres. Amazon a eu également des problèmes avec son algorithme « Rekognition » de reconnaissance de visages qui a confondu, lors d'une expérience, 28 membres du Congrès américain avec des portraits de criminels. Près de 40 % des membres confondus étaient des personnes de couleurs.

Au-delà de ces risques liés à l'usage de systèmes d'IA en particulier, où on pourrait parler de micro-impacts, l'IA semblerait présenter des menaces à plus grande échelle : celle de nos sociétés. En effet, les algorithmes spécialisés vont être impliqués dans les métiers les plus routiniers. Si le marché de l'emploi sera bouleversé par l'IA, on ignore encore dans quelle mesure : les prévisions de destruction nette d'emploi liées à l'IA à l'horizon 2025 se situant entre 6% et... 47 %. Cela fait craindre l'apparition de différentes classes de citoyens entre ceux qui peuvent avoir accès à un emploi ou non, ceux qui peuvent exercer des tâches

---

[24] M.Caldwell, J.Andrews, T.Tanay et L.Griffin, « AI-enabled future crime », https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-020-00123-8

[25] Le stalking, parfois traduit par « traque furtive » est une forme de harcèlement. Le stalker suit avec une attention maladive les faits et gestes d'une personne. Le développement des réseaux sociaux facilite le stalking.

créatives, faisant appel à des compétences non encore gérées par les IA ou plus simplement entre ceux qui ont une utilité sociale ou non.

Les valeurs démocratiques pourraient également être mises en danger : quid de la vie privée (avec les cookies ou plus simplement avec l'instauration potentielle de systèmes de surveillance de masse), du droit à la liberté d'expression (avec les algorithmes de modération automatique qui, par exemple, ont suspendu des comptes Facebook et Twitter de militants LGBT car leur orientation sexuelle était mentionnée dans leur profil), de dignité humaine ou le droit à un recours juridictionnel effectif et à un procès équitable ? Un citoyen peut se sentir impuissant pour se défendre contre un des GAFAM d'autant plus s'il n'est même pas au courant des prises de décisions algorithmiques qui le concerne ! Par exemple, la technique du shadowban (bannissement furtif) tend à se répandre : cette forme de modération rend invisible le contenu du créateur en le déréférençant sans qu'il n'en soit averti ou qu'il en ait conscience.

Enfin, on peut se poser des questions sur l'impact de l'utilisation de l'IA sur les compétences humaines. Certaines études ont déjà remarqué un impact sur le fonctionnement cérébral. Ainsi, des scientifiques de l'University College London ont découvert que l'utilisation du GPS « désactive » l'hippocampe et le cortex préfrontal, zones du cerveau dédié à la mémoire et à la prise de décision. Une utilisation non-raisonnée de l'IA pourrait-elle conduire à un abêtissement de l'espèce humaine ?

Des réponses doivent être apportées à tous ces inquiétudes avant de pouvoir avoir confiance dans les IA.

3. Plusieurs obstacles à franchir

Le principal obstacle à franchir pour construire des IA dignes de confiance est l'immaturité de l'IA. Cette immaturité touche de nombreux aspects de l'IA : la technologie en elle-même bien sûr, avec le développement constant d'algorithmes spécifiques mais aussi les bonnes pratiques, les référentiels, la règlementation... Dès lors, la question se pose : peut-on, avec les outils à notre disposition aujourd'hui, traiter toutes les spécificités des systèmes d'IA critiques ? On peut répondre par la négative sans trop se tromper. C'est pour cela qu'il apparait impératif de développer de nouveaux outils.

En effet, pour atteindre une IA de confiance, il faut pouvoir utiliser un certain nombre d'outils et de dispositifs techniques ou organisationnels qui permettent par exemple d'évaluer un point particulier. Ces points peuvent être le niveau de biais d'une IA spécifique mais aussi un ensemble d'autres attributs qui contribuent à apporter cette confiance globale d'une IA. Or, ces outils d'évaluation restent encore à développer. Dans les quelques cas où ils existent, ils sont souvent encore très balbutiants et peu utilisés. Certains acteurs estiment que, suivant le principe de précaution, tant que les outils nécessaires n'existent pas et qu'on n'a pas assez d'informations, il faut interdire l'utilisation des IA. Cela peut paraître logique notamment dans le cas des IA à haut risque : tant qu'on n'a pas le niveau de confiance suffisant, il parait plus sage de conserver les méthodes qui ont fait leurs preuves. Tant qu'on n'est pas sûr que les voitures autonomes soient dignes de confiance, il vaudrait mieux éviter leur utilisation.

Pour surmonter cet obstacle, il faudra en premier lieu que les chercheurs puissent attester de la faisabilité de ces outils. Par exemple, est-il possible de rendre « transparent » avec les outils actuels les algorithmes obscurs type « boîtes noires » qui sont présents actuellement ? Si non, que resterait-il à créer pour y

parvenir ? Il n'est pas trop optimiste d'estimer que cet obstacle devrait se régler avec le temps : les règlementations se mettront petit à petit en place et on réussira à obtenir des outils permettant de valider le niveau de confiance sur chacun des composants d'une IA ainsi que sur l'ensemble final. Ainsi, pour la voiture autonome, on devrait pouvoir un jour attester du niveau de confiance sur chacun des capteurs, techniques, outils et enfin le système d'intelligence artificielle complet de la voiture autonome en question.

Mais avant de pouvoir attester de ces niveaux de confiance, il faudra se mettre d'accord sur ce dont on parle. Nous l'avons vu lorsque nous avons essayé de caractériser l'IA : plusieurs définitions existent, cohabitent, présentent chacune leurs nuances, sont utilisées par des acteurs différents. Un même terme peut porter plusieurs sens qu'il soit utilisé par un développeur ou une autre personne. Il va ainsi falloir développer un socle commun de compréhension où chaque terme présentera une définition claire, compréhensible et comprise de tous. Ainsi, lors du processus de certification de confiance d'une IA, lorsqu'une exigence sera examinée, on sera certain que chaque personne, du développeur à l'auditeur, parle bien de la même caractéristique du système d'IA. Cette problématique de clarification est nécessaire aussi bien dans le cadre de la règlementation pour qu'on puisse s'assurer du périmètre exact de la conformité avec cette règlementation que dans le cade d'engagements contractualisés.

Les systèmes à base d'intelligence artificielle modifieront en profondeur notre société à partir du moment où ils seront acceptés par la population et diffusés.

Pour cela, il faudra que l'IA respecte les valeurs et les droits fondamentaux. Les dispositifs et les institutions permettant de vérifier cela ne sont pas encore mis en place, d'autant plus que la notion de droit fondamental est variable suivant les cultures.

Il faudra également apaiser les craintes qu'inspirent ces systèmes en développant des outils de façon à diminuer les menaces et les mésusages volontaires ou non. Très peu de ces instruments existent à l'heure actuelle : la technologie est trop jeune. Il faudra donc que les chercheurs et les développeurs s'attellent à les créer.

Mais avant de parvenir à cela, il faudra d'abord convenir de certaines définitions afin que les différents acteurs de l'IA puissent se comprendre pour définir comment bâtir le futur de l'intelligence artificielle.

CHAPITRE 3

NÉCESSITÉ DE DISPOSER DE SYSTÈMES À BASE D'INTELLIGENCE ARTIFICIELLE DIGNES DE CONFIANCE

En 2022, les IA sont comme un jardin sans horticulteur : de magnifiques plans se sont développés de manière anarchique. Le travail d'un jardinier permettrait de contrôler l'ensemble du terrain et de magnifier chaque pousse sans quoi personne n'oserait utiliser l'endroit sous peine de se faire griffer par les épines. Nous verrons dans ce chapitre comment la friche de l'IA s'est développée frénétiquement ces dernières années, puis comment les organisations et les institutions ont réalisé qu'elles devaient mettre un peu d'ordre en commençant à penser l'IA et la relation que nous devrions entretenir avec ces systèmes. Nous verrons quels constats se sont imposés et quelles ont été les premières actions juridiques mises en place.

1. Mobilisation des parties prenantes pour des IA dignes de confiance

La « ruée vers l'IA » apparue ces dernières années s'est pour l'instant souvent déroulée sans contraintes ni limites. Les équipes de développeurs se sont attachés à produire des résultats sans parfois s'encombrer de méthodologies rigoureuses ni de recul moral dans leurs développements. Aux promesses utopistes de ces « nouvelles » technologies se sont succédées les interrogations et les inquiétudes des citoyens, relayées par la presse, des universitaires et les gouvernants. Si le spécialiste en IA avait bien compris que l'adoption de la voiture autonome ne supprimerait jamais l'accident, le décès de Élaine Herzberg[26], première piétonne tuée par un véhicule autonome, a donné à l'opinion publique l'impression qu'on lui avait menti. La mise en place de Parcoursup a davantage stressé les lycéens à qui on avait promis une orientation plus sereine. Les incompréhensions, les voix furieuses se faisant de plus en plus vives, les parties prenantes du monde de l'IA ont réalisé qu'il était temps de commencer à contrôler le domaine de l'IA, de le faire passer d'un état d'esprit « sans foi ni loi » à quelque chose de plus policé, plus maîtrisé.

En effet, l'ensemble des acteurs de l'IA a un intérêt commun : l'adoption de l'IA par la société. Le potentiel bénéfique de cette technologie ne pourra être atteint que s'il n'est pas contrecarré par un rejet de la population et des autres acteurs de l'écosystème des IA comme les fournisseurs. Les personnes doivent avoir confiance dans son utilisation.

Ainsi, depuis l'année 2016, et de plus en plus souvent, de nombreux acteurs travaillent à élaborer les bases d'un cadre commun éthique à commencer par les entreprises privées elles-mêmes. Ainsi, des sociétés

[26] Elaine Herzberg poussait son vélo le 18 mars 2018 pour traverser la quatre-voies à Tempe en Arizona (USA) quand elle a été fauchée par un taxi Uber « autonome ».

nationalement reconnues comme Telefónica en Espagne, IA Latam au Chili ou les géants Sage, ITI, Microsoft et Google aux États-Unis, se sont réunies pour produire de documents énonçant les principes qu'ils utiliseraient pour développer leurs systèmes d'IA. Cela a été également le cas pour des sociétés civiles comme Amnesty International, The Public Voice Coalition ou le T20:Think20 mais toute organisation intéressée par le sujet a pu publier ses propres réflexions comme l'université de Montréal, l'IEEE ou, parmi bien d'autres, le New York Times[27].

Bien que ces parties prenantes aient des voix plus ou moins fortes sur le sujet –que dire de la capacité de Google ou de Microsoft à influencer les technologies d'IA ? Que dire de la portée de la voix du New York Times ? – les constatations émises n'ont ni portée légale, ni pouvoir de contrainte. Les acteurs du marché et les différentes parties prenantes ont conscience de devoir mettre en place un socle de pratiques communes reposant sur des fondations saines et une compréhension consensuelle des termes et de la technologie mais il apparaît de plus en plus que les organisations internationales et nationales doivent prendre le relais pour règlementer et légiférer.

2. Mobilisation des organisations nationales et internationales pour des IA dignes de confiance

Les organisations nationales et internationales n'ont pas été en reste ces dernières années. Le nombre de symposiums, de livres blancs et de constats sur le sujet ont explosés ; entre autres :

- un livre blanc sur la standardisation de l'IA émis par l'administration chinoise des standards
- « l'IA au service des citoyens » par l'agence du numérique italien
- « l'IA dans l'UK » rédigé par la chambre des lords du Royaume-Uni
- « Principes et éthique de l'AI » par les Émirats Arabes Unis...

Cet emballement de production de documents et d'émission d'avis est logique : les systèmes IA sont de véritables enjeux nationaux car ils influent sur l'emploi, le système judiciaire et, plus généralement, l'évolution de la société future. Un certain nombre de gouvernements ont pris conscience de la nécessiter d'édicter des principes à respecter et de légiférer en ce sens.

Au-delà des organismes nationaux, plusieurs organisations intergouvernementales ont exploré le sujet : c'était le cas avec le groupe européen sur l'éthique dans les sciences et les nouvelles technologies qui a publié l'ouvrage « Constats sur l'IA, la robotique et les systèmes autonomes » mais aussi avec le livre blanc de la Commission européenne « IA pour l'Europe » en avril 2018 ou « Charte éthique européenne sur l'utilisation de l'IA dans les systèmes judiciaires » par le Conseil de l'Europe (CEPEJ) également en 2018.

---

[27] Une infographie résumant les principes de ces documents a été créée par J. Field, H. Hiligoss, N. Achten et al., sous le nom « Principled artificial intelligence ; a map of ethical and rights-based approaches » en 2020 Elle est trouvable sur https://cyber.harvard.edu/publication/2020/principled-ai

L'Union européenne a saisi très tôt l'intérêt de l'adhésion unanime à un cadre commun de principes. Cela est énoncé avant même l'introduction dans le livre blanc de la Commission européenne sur l'IA « Une approche européenne axée sur l'excellence et la confiance » : « *Pour tirer le meilleur parti possible des opportunités qu'offre l'IA et relever les défis qu'elle pose, l'Union européenne doit se montrer unie dans l'action et définir une manière qui lui est propre de promouvoir le développement et le déploiement de l'IA, en s'appuyant sur les valeurs européennes.* » Cette assertion se traduirait par une révision du plan coordonné pour favoriser le développement et l'utilisation de l'IA dans tous les pays membres.

Les travaux réalisés par l'Union européenne pour la constitution d'une éthique commune ont servi de base ou ont été réutilisés par plusieurs organisations internationales ; comme ce fut le cas pour l'élaboration des principes éthiques de l'OCDE, principes approuvés par la suite par le G20. Ces quelques organisations se sont concentrées sur le sujet des principes et des règlementations. De nombreux autres groupes comme le partenariat Mondial sur l'IA, le Centre Commun de Recherche de l'Union européenne, l'Institut Montaigne... se sont intéressés au sujet des mises en œuvre et des bonnes pratiques. En raison de l'étendue du sujet, ils se sont souvent seulement concentrés sur quelques thèmes précis, ne répondant pas ainsi à l'ensemble des problématiques. Enfin, une troisième catégorie d'acteurs nationaux a travaillé sur le thème des contrôles et certifications comme LNE (Laboratoire National d'Essais) en France et l'ICO (The Information Commissioner's Office) au Royaume-Uni.

3. Leurs constats : nécessité de valeurs, principes, règles et bonnes pratiques communs pour répondre à un impératif de confiance

L'ensemble de ces organisations et de ces travaux mettent en avant le potentiel bénéfique de l'adoption des systèmes d'IA, à même de participer à la résolution des plus grands défis planétaires de ce prochain siècle quels qu'ils soient : climatiques, environnementaux, sociaux, sociétaux, démocratiques, thérapeutiques, épidémiologiques...

Pour intégrer sereinement ces outils dans notre quotidien de citoyens, de consommateurs, d'entreprises privées ou d'États, il est impératif d'avoir confiance dans ces outils. Ainsi, comme le souligne la Commission européenne : « *le déficit de confiance constitue aussi un frein considérable à un recours plus généralisé à l'IA* ». Cette notion est double : pour que les gens puissent leur accorder cette confiance, les systèmes d'IA doivent eux-mêmes être conçus de manière à être à la hauteur de leurs attentes. Mais la responsabilité est mutuelle : une fois que ces systèmes seront conçus éthiquement, il sera alors possible de leur accorder cette confiance.

La gageure est de s'accorder sur les principes permettant l'atteinte de cette assurance. Pour cela, les travaux ont conclu que ces systèmes doivent répondre à un certain nombre de principes éthiques pouvant être résumés en : être responsables face aux enjeux, dignes de confiance et respectant les droits fondamentaux. La Commission européenne a défini une IA comme digne de confiance lorsqu'elle est fondée sur un ensemble de valeurs, de règles communes et de droits fondamentaux.

Toutefois, tout aussi unanimement, il est souligné que les solutions techniques et les bonnes pratiques actuelles sont insuffisantes. Il est donc nécessaire de réaliser des travaux supplémentaires afin d'élaborer un ensemble de définitions communes, un périmètre international qui fasse consensus et un ensemble de règles sécurisant l'ensemble des parties prenantes.

4. Leurs premières orientations : un premier socle commun de principes et d'exigences à respecter et sur lequel s'appuyer pour construire les dispositifs nécessaires

Ces organisations ont donc énoncé des principes que devraient suivre les concepteurs de systèmes d'IA afin de créer des IA dignes de confiance. Il s'agit, par exemple en Union européenne, de garantir le respect des règles de l'Union européenne, notamment celles qui protègent les droits des consommateurs et les droits fondamentaux tels que la dignité humaine et la protection de la vie privée énoncés dans la Charte des droits fondamentaux de l'Union européenne.

Sept exigences essentielles ont été dégagées :

- prise en compte du facteur humain et conservation du contrôle humain
- robustesse technique et sécurité
- respect de la vie privée et gouvernance* des données
- transparence
- diversité, non-discrimination et équité
- bien-être sociétal et environnemental
- responsabilisation.

Un cadre réglementaire clair permettrait de susciter la confiance des consommateurs et des entreprises à l'égard de l'IA en définissant les moyens de réduire au minimum les divers risques de préjudice pouvant exister. Ces règles devront être respectées pour donner aux entreprises et aux organismes du secteur public la sécurité juridique voulue pour innover au moyen de l'IA.

Toutefois, ces éléments restent théoriques : ces principes et exigences n'ont pas de définition ni de limite clairement définie. Il n'est pas précisé la nature exacte et le niveau attendu de ces exigences, le type de pratiques qui devraient favoriser l'atteinte de ces exigences, leurs faisabilités techniques et organisationnelles, les difficultés de mise en œuvre, la manière dont on pourra se rassurer de façon raisonnable quant à la présence effective d'IA dignes de confiance, la manière dont cette confiance sera restituée, les démarches, labels et référentiels associés.

Il reste donc de nombreux travaux à entreprendre et finaliser : définir les solutions à développer, identifier les bonnes pratiques à utiliser, les référentiels à bâtir, les certifications et les labels à proposer…

5. Des limites aux réglementations actuelles mais de nombreuses initiatives en cours pour couvrir les problématiques spécifiques aux IA

5.1. De nombreuses limites aux règlementations actuelles

Les règlementations actuelles couvrent certaines des problématiques soulevées par les IA mais seulement partiellement. Par exemple, la question de protection des données personnelles est bien gérée par la règlementation sur la RGPD qui a permis une avancée substantielle sur le sujet. En revanche, cette règlementation délaisse la question des données industrielles, économiques et publiques alors que ces dernières font partie intégrante des systèmes d'IA. Plus généralement, les sujets de transparence, de traçabilité au sens de l'IA et du contrôle humain, entre autres, ne sont couverts que très partiellement.

De plus, l'essence même de l'IA peut compliquer l'application et le contrôle de l'application des règles existantes. Ainsi, en Europe, les concepteurs et les personnes chargées du déploiement d'IA doivent faire en sorte de respecter la législation européenne sur les droits fondamentaux comme le respect de la vie privée ou la non-discrimination mais aussi la protection des consommateurs. Ils doivent aussi respecter les législations sur la sécurité du produit et les responsabilités qui en découlent. Cela semble simple mais comment prouver que ces règles sont appliquées dans le cadre de systèmes d'IA opaques ou très évolutifs ?

Les règlementations actuelles sont non seulement insuffisantes mais les aspects qui sont couverts sont généralement disséminés dans plusieurs lois ce qui complique la lecture juridique des différentes parties prenantes.

Notons enfin que les projets de règlementations les plus avancés sont ceux du Parlement européen mais il reste probablement quelques années encore avant qu'elles ne puissent aboutir. L'Europe essaie de compenser son retard concernant plusieurs facteurs clés de l'IA par l'instauration de règles qui contraindraient les concepteurs d'IA des autres pays s'ils désirent commercialiser leurs produits sur le vieux continent. L'Europe espère ainsi orienter les méthodologies de conception mondiales. La Chine et les États-Unis, qui tirent plutôt avantage de l'absence de règlementation, essaient de retarder le plus possible la légiférération de ce sujet qui pourrait nuire à leurs intérêts.

5.2. Plusieurs initiatives d'autorégulation au travers de chartes éthiques

En effet, la première problématique du droit dans le sujet de l'IA est de déterminer à quel moment légiférer. Poser des lois trop tôt présente le risque de freiner voire stopper l'innovation. Au contraire, attendre trop pour légiférer peut favoriser l'innovation mais cela peut être au détriment des droits fondamentaux. En l'absence de limites juridiques, il est probable que des systèmes d'IA puissent se retrouver à violer certains droits fondamentaux ce qui pourrait avoir de graves conséquences économiques et/ou sociétales ; cela a pu être constaté par le passé, à chaque grande innovation industrielle.

Actuellement, la législation n'étant pas trop avancée, les limites sont surtout données par l'autorégulation des entreprises et autres organismes. De nombreuses chartes éthiques et textes de droit souples ont été

créés. Une étude a analysé 21 d'entre elles et recensé 94 principes utiles à la réflexion sur les considérations éthiques, juridiques et scientifiques de la robotique et de l'IA. On peut remarquer certaines tendances dans ces dernières. En particulier, le critère le plus fréquent est celui de la protection de la vie privée (10 chartes et textes sur 21 contiennent ce critère) suivie par la transparence des algorithmes (9 sur 21). Les principes de respect de la dignité humaine, de bienfaisance, et de protection des données personnelles sont citées dans 7 études sur 21.

5.3. Le temps d'une nouvelle règlementation adaptée et les problématiques à y intégrer

Règlementation horizontale ou verticale

Les tendances actuelles montrent que le temps de l'autorégulation s'achève pour laisser place à la règlementation. C'est là le second problème qui se pose : faut-il favoriser une réglementation verticale ou horizontale ? La règlementation verticale consisterait à réguler les systèmes d'IA par secteur d'activité et par type d'activité. Il existe, en début janvier 2022, à travers le monde, plus d'une centaine de règlementations dans le monde. On constate qu'il s'agit principalement de règlementations sectorielles comme par exemple, l'IA médicale, l'IA pour les navires autonomes... On peut noter pour illustrer l'adoption par les Nations Unies de trois règlements en été 2020. Ils concernent des sujets très précis : les systèmes automatisés de maintien de la trajectoire pour les voitures, la gestion des risques cyber dès la conception des véhicules et la mise à jour des logiciels équipant les véhicules.

Ce type de règlementation semble s'imposer partout dans le monde sauf en Union européenne où la régulation horizontale serait a priori privilégiée. Dans ce cas, il y aurait une régulation globale a minima qui irriguera tous les secteurs d'activités.

La prise en compte de l'humain

Les règlementations actuelles sont principalement orientées dans le cadre de « l'IA faible » c'est-à-dire une IA où l'humain reste supérieur à la machine. Mais il est à noter que les IA réussissent de mieux en mieux à passer le test de Turing – c'est-à-dire qu'un humain qui converse avec elles est incapable de se rendre compte qu'il discute avec un programme et non pas un autre humain. Ces situations pourront amener à un premier niveau de tension juridique. On peut envisager qu'une obligation de transparence sera à mettre en place où, lorsqu'une IA devra interagir avec un humain, elle devra d'abord préciser sa nature d'IA pour que l'humain soit bien conscient de l'essence de son interlocuteur.

Un second niveau de tension juridique se posera dans les cas où l'IA est supérieure à l'humain, comme c'est déjà le cas pour les jeux d'échecs ou de dames où un humain ne peut plus gagner contre la machine. Ces IA sont actuellement monofonctionnelles et multi-contextes. Ce problème s'amplifiera lorsque les IA plus fortes seront plurifonctionnelles et multi-contextes.

Le troisième niveau de tension juridique sera constitué par la mixité IA/humain. Il semble que les problèmes émergeront à partir du moment où les machines ne seront plus cantonnées à certaines localisations précises et qu'elles commenceront à partager le même espace physique que les humains.

Une règlementation spécifique par catégorie de risques

Le projet de règlementation de l'IA (RSIA) classerait les IA en cinq catégories :

- les IA interdites : ce sont celles qui violent les droits fondamentaux comme par exemple, les IA subliminales, les IA de reconnaissances faciales en temps réel — on voit d'ailleurs la différence de traitement géographique puisque ces dernières IA sont parfaitement tolérées et répandues en Chine et aux États-Unis
- les IA à hauts risques : ce sont les IA qui pourraient entraîner des conséquences graves en cas de problème. Les trois principaux critères pour classer une IA dans cette catégorie sont la sécurité des personnes, la santé des personnes et/ou les valeurs morales
- les IA avec une obligation de se présenter en tant qu'IA : ce sont les IA que nous avons mentionnées dans le paragraphe précédent. Ces IA qui réussissent le test de Turing doivent indiquer au début d'une conversation avec un humain que ce dernier est en train d'interagir avec une IA
- les IA qui sont hors du champ de la règlementation, comme les IA militaires que l'Union européenne considère ne pas relever de son champ de compétence
- les IA avec une absence d'impact significatif : ces IA présentant un risque minimal ne sont pas concernées par une règlementation spécifique mais pourraient être soumis à des exigences à minima prévues pour toutes les IA telle que la protection de la vie privée et de transparence.

La nature des responsabilités

La règlementation est principalement basée sur trois principes qui sont autant de défis pour les juristes :

- la sécurité humaine et l'intégrité physique : ce qui correspond au « safety » anglophone. Le défi est de savoir qui prend la décision en dernier ressort : qui, de l'IA ou de l'humain, a le dernier mot ? Autrement dit, qui est le dominant ? Il sera impératif de répondre à cette question pour apporter la réponse à la fameuse question du « bouton rouge[28] » qui permettrait à l'humain de reprendre la main sur un système d'IA
- l'acceptabilité sociale : ce point a pour but de répondre à la problématique de la cohabitation de l'IA avec les humains. Par exemple : comment rendre compatible les IA et le marché du travail ? Faut-il autoriser les robots à prendre un visage humain ou non ?

---

[28] Ce terme fait référence aux boutons rouge d'arrêts d'urgence des machines industrielles. Il s'agit de créer une commande pour désactiver l'intelligence artificielle de manière la plus instantanée possible. L'une des difficultés est de faire en sorte que l'IA n'apprenne pas à empêcher l'activation de ce bouton rouge.

- la responsabilité : comment protéger le consommateur ou le salarié ? Qui est responsable en cas de problème ? Les concepteurs, les fournisseurs, les robots eux-mêmes, un mélange de certains ?

Terminons cette approche juridique de l'IA par la remarque suivante : plus les IA seront de confiance, moins il y aura de problèmes juridiques.

Au cours des années 2010, les entreprises ont réalisé que les gens n'adopteraient pas les systèmes d'information à base d'intelligence artificielle si ces derniers les mettaient en danger, de manière physique ou psychologique. Ces organisations privées ont alors conçu des chartes éthiques. Ce premier acte de bonne volonté n'assurant pas au citoyen que ces chartes soient réellement suivies, les institutions et gouvernements ont pris le relai. Leur constat est limpide : pour que les IA atteignent leur plein potentiel, il faut que la population puisse leur faire confiance, notamment via un cadre règlementaire.

Sept exigences dérivées des droits fondamentaux de l'Union européenne ont été définies : prise en compte du facteur humain et conservation du contrôle humain, robustesse techniques et sécurité, respect de la vie privée et gouvernance des données, transparence, diversité, non-discrimination et équité, bien-être sociétal et environnemental et responsabilisation.

Ces principes sont théoriques, presque philosophiques et peu applicables concrètement, à la manière d'un jardinier qui annonce désirer un joli parc avec des arbres et des fleurs. Une orientation vers le but est donnée mais rien n'indique concrètement comment atteindre ce but. Des règlementations actuelles, comme la RGPD, couvrent partiellement certains des aspects relevés par ces exigences. Il faudra donc les développer en s'adaptant à la nature des systèmes à base d'intelligence artificielle, en répondant aux défis juridiques, tout cela voté au moment opportun pour ne léser ni la créativité technique ni les citoyens.

CHAPITRE 4

PLUSIEURS PROBLÉMATIQUES SPÉCIFIQUES AUX SYSTÈMES À BASE D'INTELLIGENCE ARTIFICIELLE NÉCESSITANT UN TRAITEMENT ADAPTÉ

Lorsqu'un jardinier entre dans un terrain en friche et désire en faire un beau jardin, il ne doit pas commencer à prendre sa bêche pour planter des bulbes. Il doit d'abord mettre en place une démarche : imaginer ce qu'il veut faire (créer un jardin de roses par exemple), quels sont les risques pour son projet (le froid peut endommager ce type de plantes), quelles sont les meilleures techniques à mettre en œuvre, quels efforts et moyens il est prêt à mobiliser...

Si l'on veut transformer la friche de l'IA en un domaine acceptable, il sera nécessaire d'adopter une démarche similaire. Il faut commencer par s'interroger sur les thèmes critiques à prendre en compte pour créer des IA de confiance. Quels sont les avantages apportés par les systèmes d'information à base d'intelligence artificielle à chaque partie prenante mais aussi les risques supportés par chacun d'entre eux et quelles ressources mettre en place...

1. De nombreuses sources de création de valeur mais avec quel niveau de risque et d'efforts acceptables par les différentes parties prenantes ?

1.1. Quels bénéfices ? Quels risques ? Quels efforts ? Pour qui ?

L'un des potentiels de l'adoption des systèmes d'IA est de réaliser un bond dans la création de valeur*, de la même manière que la révolution industrielle avec la généralisation de la mécanisation ou l'adoption massive de l'informatique. Toutefois, les différentes parties prenantes de l'IA n'auront pas la même approche de la création de la valeur que peut leur offrir l'IA. Certaines y verront une façon de produire plus d'innovation (la voiture autonome ou les assistants vocaux n'existent que grâce à l'IA), d'autres un moyen d'augmenter la qualité de service (Netflix, Youtube, Spotify... proposent à leurs abonnés de meilleures suggestions de visionnages, les IA de recrutement permettent aux candidats à des offres d'emplois, d'éviter des entretiens voués à l'échec), des gains de productivité (les routines d'IA automatisent certains processus permettant de d'accélérer voire de supprimer la réalisation manuelle, les robots effectuent certaines tâches à la place des humains) ou encore d'améliorer la qualité des décisions comme pour le choix des investissements financiers.

Toutefois, ces méthodes de création de valeur peuvent se révéler conflictuelles pour les différentes parties prenantes. Ce conflit est facile à comprendre en prenant l'exemple de Parcoursup. La valeur de ce système d'IA pour les étudiants est d'améliorer la qualité du service en leur permettant de mûrir leur décision

d'orientation et donc de diminuer les échecs d'orientation. Pour les universités, la valeur est d'améliorer l'efficacité de l'organisation de la rentrée. Ces deux valeurs sont presque directement opposables : si un lycéen mûrit son choix, l'université a moins de temps pour connaître la liste des inscrits et préparer convenablement les processus administratifs. L'objectif est donc de trouver des options qui permettent aux différentes parties prenantes de trouver des avantages et d'accepter le résultat. Il convient donc d'identifier les bons dispositifs qui conduiront à chacun d'avoir confiance dans la matérialisation des bénéfices attendus à un niveau de risque et un niveau d'efforts acceptables pour eux.

La négociation entre ces différents intérêts et la prise de décisions quant à des solutions relève de la gouvernance des systèmes d'informations IA.

À chaque décision, il est possible et nécessaire de se poser trois questions pour cerner le bien-fondé de la création de valeur envisagée :

1. quels bénéfices et pour qui ?
2. quels risques et pour qui ?
3. quelles ressources nécessaires et pour qui ?

La réponse à ces questions entraîne une réflexion pour l'atteinte d'une IA digne de confiance.

Quels bénéfices et pour qui ?

En déterminant quels sont les apports amenés par tel système d'IA et pour quelles parties prenantes, on peut cerner les intérêts de chacun ainsi que les conflits d'intérêts potentiels à court et moyen termes. Cela permet de réorienter le projet et/ou de préparer des solutions préventives comme réaliser des campagnes de communication adéquates. Expliquer au grand public que les voitures autonomes n'éviteront pas les accidents mortels mais que le nombre de victimes sera nettement moindre permet de le faire accepter. La présentation compréhensible de chaque système d'IA envisagé permettra de mieux faire accepter cette IA aux parties prenantes en soulignant son intérêt.

Il est important de noter qu'il existe une composante humaine à cette acceptation : il faut distinguer les faits bruts de la perception de ces faits par les parties prenantes. Ainsi, même si un système d'IA peut se montrer objectivement supérieur à un processus humain, il peut ne pas être moralement acceptable pour les parties prenantes. De même, il se peut que la valeur perçue ne soit pas suffisante pour que l'IA soit acceptée.

Quels risques et pour qui ?

C'est également en définissant quels sont les risques amenés par cette IA et qui devra les supporter qu'on pourra atteindre un système IA plus digne de confiance. On pourra ainsi déterminer, en fonction de la tolérance au risque de chaque partie prenante et de son intérêt par rapport aux bénéfices envisagés, quel est le niveau acceptable de risque pour chaque partie.

Les risques sont différents en fonction des parties prenantes : les entreprises pourront par exemple avoir un risque concernant le secret des affaires tandis que les individus voient un risque de divulgation de leur vie privée voire de leur vie intime et de leur dignité.

Toute partie prenante impactée par une décision issue d'un système d'IA peut ne pas en être informée ou subir une différence de traitement voire une discrimination sans pour autant pourvoir déclencher un recours en toute connaissance de cause. Il est important de noter que les risques ne sont parfois pas nécessairement subis par les parties prenantes qui bénéficient des avantages. De nombreux exemples existent mais le plus significatif est celui des systèmes d'IA d'aide au trading qui offrent un avantage aux traders mais posent un risque sur les entreprises cotées en Bourse et leurs salariés en raison de la décorrélation entre les calculs mathématiques de l'IA et les réalités de l'économie.

Quels efforts et pour qui ?

En déterminant qui porte les efforts et la nature de ces efforts, on peut planifier une utilisation raisonnable et responsable des ressources. En excluant les entreprises qui créent des systèmes d'IA en utilisant des briques logicielles disponibles sur le marché, les investissements pour créer des systèmes d'IA sont très importants que ce soit au niveau humain avec le recours à de la main-d'œuvre qualifiée, financier pour payer le personnel et les infrastructures mais aussi au niveau technologique. L'IA demande d'une part du matériel conséquent avec des supercalculateurs de plus en plus puissants et chers mais aussi l'accès à une masse de données.

Ces dernières sont souvent détenues par de très grandes entreprises comme par les GAFAM qui se présentent de plus en plus comme des acteurs incontournables de ce pilier de l'IA et qui pourraient facilement imposer leurs décisions à des parties prenantes de moindre poids et donc fortement dépendant de ces acteurs pour avoir un accès pérenne aux données utiles.

1.2. La nécessité de règles et de dispositifs de gouvernance pour effectuer les arbitrages

Cette démarche de questionnement en trois parties permet de prévenir les conflits ou, le cas échéant, d'arbitrer plus facilement les conflits entre les différentes parties prenantes, dans une logique de création de valeur pour un grand nombre d'intéressés. Les différentes parties prenantes ont bien compris que l'IA amène au risque qu'il y ait un petit nombre de bénéficiaires de l'IA au détriment de tous les autres, résumé par la maxime anglaise « the winner takes it all[29] ».

Pour faire face à ce danger, les parties prenantes désirent défendre leurs intérêts en créant des instances représentatives investies du pouvoir de résolution des conflits et de décision. Elles pourraient ainsi être à même de trancher sur les problématiques éthiques qui sont posées par les performances de l'IA et les différents conflits d'intérêts. Cela permettrait d'améliorer la confiance dans le système aux différentes

[29] Le gagnant prend tout

parties prenantes même si certaines limitations se présentent déjà. En effet, les choix éthiques sont grandement dépendants de la culture locale ou des informations possédées. Par exemple, avec les mêmes informations initiales, le choix de privilégier la vie privée ou l'intérêt public risque d'être bien différent entre un pays à la philosophie individualiste comme les États-Unis ou ayant la culture du groupe comme la Chine ou le Japon.

Ces instances pourraient endosser plusieurs rôles comme le rassemblement d'information sur le sujet via des études, des enquêtes, des sondages… et l'accès à la base d'information ainsi rassemblée. En effet, chaque partie prenante perçoit son propre rôle et son propre risque à l'aune de la connaissance réelle ou perçue qu'elle a sur le sujet. Cette connaissance peut en effet être grandement influencée par des campagnes d'informations, des lobbyistes, d'autres parties prenantes, les expériences individuelles, les techniques de manipulation assumées (nudges) ou pas… Il est donc important d'offrir des éléments fiables sur lesquels les parties peuvent s'appuyer pour former leur jugement.

Il convient aussi de préciser que certains choix pourraient être contraints par une réglementation qui les interdirait, les limiterait ou les encadrerait lorsqu'ils seraient considérés par la communauté nationale ou internationale comme inacceptables ou inéquitables. D'autres choix pourraient résulter de choix individuels et être traduits dans des chartes éthiques, par exemple, qui indiqueraient les grands principes à respecter et la nature des exigences spécifiques attendues pour les IA déployées.

In fine, il faut que ces dispositifs règlementaires ou privés d'arbitrage obtiennent l'adhésion des différentes parties prenantes et leurs donnent confiance quant à la pertinence des décisions prises et à l'acceptabilité de la répartition des avantages proposés, des risques subis, des efforts demandés et des responsabilités qui en découlent. Ce sera un prérequis pour pouvoir disposer d'IA de confiance.

Nous avons ainsi identifié plusieurs thèmes critiques à prendre en compte pour aboutir à des IA de confiance. Il s'agit d'avoir confiance dans :

- la finalité des IA et des bénéfices concrets attendus par les différentes parties prenantes, dans le respect des principes et des exigences spécifiques associées à ces finalités
- le niveau de risque, associé à ces finalités et au respect des principes et des exigences attendus, qui soit acceptable pour ces différentes parties prenantes
- le niveau d'effort déployé pour permettre aux finalités d'être proposées avec un niveau de risque acceptable. Il s'agit de déployer des outils, des compétences, des processus… qui soient soutenables techniquement et financièrement
- le dispositif de gouvernance et de responsabilisation qui a arbitré les choix effectués
- la mise en œuvre cohérente des quatre points précédents.

2. Cascade de confiance des cinq options critiques à prendre en compte : principes et exigences, prise de risques, gouvernance et responsabilités, outils et bonnes pratiques, audits et certifications

Pour répondre aux différents enjeux et à ces cinq problématiques, il est nécessaire de disposer d'une démarche intégrée qui donne confiance à l'ensemble des acteurs concernés en leur garantissant que le niveau des attentes et des engagements seront atteints avec un niveau de risque et d'effort acceptable.

Le rapport du centre commun de recherche de la Commission européenne sur « la robustesse et l'explicabilité de l'IA »[30] souligne « l'importance de la mise en place de bonnes pratiques et de procédures axées sur les menaces [...] pour renforcer la confiance dans les systèmes d'IA. » Cela signifie que tous les risques pour les différents acteurs doivent être pris en considération et que les mesures appropriées doivent être mises en œuvre pour être conformes au niveau des bénéfices attendus. Pour ce faire, notre groupe de travail a identifié une démarche intégrée en cinq points ; le point suivant ne pouvant être efficacement conduit qu'en ayant achevé la réflexion du point précédent.

La cascade de ces cinq étapes est donc la suivante.

2.1. Quels principes et exigences attendus ?

Les IA doivent donc devenir dignes de confiance. Mais cette expression est vague : il faut que cette orientation se traduise en concepts plus concrets qui puissent être utilisables par les concepteurs de systèmes d'IA. Il faudra répondre à plusieurs questions :

- quels types et combien d'exigences attend-on ? Ça peut être par exemple, la transparence, l'équité, la robustesse, la sécurité...
- quels niveaux d'exigences ?
- quel est le niveau de confiance associé à ces exigences qui est attendu par les différentes parties prenantes ?
- comment gérer les exigences contradictoires ?
- qu'attend-on des systèmes d'IA en termes de caractéristiques et de qualités ?

Répondre à ces questions permettra de définir les exigences attendues a minima des systèmes d'IA. Ces concepts sont des exigences auxquelles les systèmes d'IA devront se conformer sous peine de ne pas être considéré comme digne de confiance. Il faut en définir la liste et le contenu précis.

---

[30] https://publications.jrc.ec.europa.eu/repository/handle/JRC119336

2.2. Quelle prise de risques ?

À chaque exigence, il existe plusieurs types de risques qui pourraient conduire à ce que cette exigence ne soit pas tout à fait atteinte. Certains risques peuvent impacter plusieurs exigences, comme le risque d'enfermement algorithmique ; et certaines exigences peuvent être impactées par plusieurs types de risques. Il existe aussi des risques spécifiques à une exigence en particulier, comme le risque de biais qui est associé à l'exigence d'équité d'une IA. Certains risques sont présents pour la plupart des systèmes d'information mais ont des spécificités dans le cas des IA (cyberattaque, gestion de projet défaillant, protection des données...) et certains risques sont plutôt spécifiques aux IA (opacité, enfermement algorithmique, équité, perte du libre arbitre...).

Il faut donc avoir une vision claire de ces risques. Pour cela, il faut définir les risques généraux et spécifiques, identifier les principes à appliquer pour les prévenir (comme le principe de précaution, de proportionnalité...) mais aussi déterminer les niveaux de risques tolérables et acceptables pour les différentes parties prenantes. Il faut donc élaborer des modèles spécifiques de classification du risque au sein de la démarche générale de gestion de la prise de risque. Ces modèles devraient détailler les différents types d'IA et les niveaux de prise de risques qui soient tolérables et acceptables, ainsi que les dispositifs et outils qui soient susceptibles de les maîtriser.

2.3. Quelles gouvernance et responsabilités ?

Nous avons vu dans l'étape précédente qu'il faudra définir les niveaux de risques tolérables et acceptables pour les différentes parties prenantes. Mais ces parties prenantes sont guidées par des intérêts différents voire conflictuels ou contradictoires. Plusieurs options de contribution de valeur sont possibles, que ce soit en termes d'exigences, de niveau de risques ou de moyens à mobiliser. Des arbitrages seront donc nécessaires mais qui tranchera ? De même, plusieurs types de problèmes pourraient survenir. Plusieurs acteurs pourraient être à leur origine. Qui serait alors considéré comme responsable ? Comment imputer ces responsabilités compte tenu de certaines spécificités des IA (opacité...) ? Quels seraient les recours possibles ? Il faudra mettre en place une gouvernance appropriée dotée d'un pouvoir de prise de décision proportionnel à l'étendue de leurs responsabilités.

2.4. Quels outils et bonnes pratiques ?

Une fois que ces choix seront faits, il faudra déterminer quels dispositifs permettront d'y répondre de manière satisfaisante, de les mettre en œuvre et de les suivre. Il s'agira d'identifier les pratiques opérationnelles, de management* et de gouvernance qui conviennent. Il faut mettre en place des structures organisationnelles, des directives, des comportements, des outils, des processus... adaptés à chaque problématique à traiter et en particulier aux problématiques spécifiques aux IA. Plusieurs types de leviers* sont possibles pourvu qu'ils soient efficaces, efficients, robustes...

Ces bonnes pratiques seront donc en résonnance avec les référentiels IA. Chaque étape du cycle de vie d'un système d'IA (conception, développement ou acquisition, mise en œuvre opérationnelle, améliorations, destruction...) devra avoir ses propres bonnes pratiques. Si les concepteurs, développeurs et autres acteurs suivent ces dernières, ils devraient plus facilement atteindre les exigences qui font d'une IA une IA de confiance.

Encore faut-il avoir des outils qui permettent de vérifier l'atteinte de ces exigences. Il s'agit entre autres des indicateurs de performance. Ceux-ci peuvent être classés en deux catégories :

- les indicateurs de résultat in fine (ou indicateurs avancés) qui permettent de s'assurer que les exigences souhaitées pour les IA soient atteintes
- les indicateurs de moyens qui permettent de vérifier que les exigences souhaitées pour les leviers soient effectives.

2.5. Quels audits et certifications ?

Cette dernière étape complète les quatre premières étapes de la cascade des options à prendre en compte tout en devant réutiliser pour elle-même ces premières étapes.

Elle la complète car elle a pour but de vérifier le niveau de confiance et d'assurance que l'on souhaite pour telle ou telle caractéristique, pour tel ou tel élément d'une IA, ou pour l'IA dans son ensemble qui intègre les différents niveaux de confiance obtenus sur les sous-ensembles.

Mais il n'est actuellement pas possible de réaliser cela en raison de l'immaturité de l'IA et du processus d'audit de l'IA. De nombreux points sont encore à étudier :

- qui sont les prescripteurs et les parties prenantes ?
- quel est le type d'assurance à mettre en place (audit*, certification*, labels*...) ?
- quel est le niveau d'indépendance pour ces missions (audit interne, externe, tiers de confiance...) ?
- quel est le périmètre de l'assurance (quelles exigences à couvrir, quels risques, quels leviers...) ?
- quel est le niveau de confiance souhaité ?
- quelle est la faisabilité technique et financière... ?

Pour répondre à ces différentes interrogations, il faut réappliquer les quatre premières étapes de la cascade de confiance à chaque type d'assurance à mettre en place. Il faudra donc dans un premier temps définir les exigences de l'audit, puis les risques de l'audit (en particulier les faux positifs et faux négatifs, c'est-à-dire les risques de l'émission d'une opinion d'audit erronée) puis les options de gouvernance et de responsabilité des acteurs d'audit puis les pratiques d'audit (tests à effectuer, types d'opinion, réserves...).

Il est impératif de bien distinguer :

- les référentiels d'IA en termes de caractéristiques attendues des IA et en termes de caractéristiques des moyens à mettre en œuvre pour s'assurer que les résultats attendus soient réalisés
- les référentiels d'audit eux-mêmes en termes de type d'opinions d'audit attendues et en termes de moyens d'audit mis en œuvre (comment auditer l'IA) pour permettre l'émission des opinions d'audit de qualité.

Les intérêts pour les IA ne sont pas les mêmes en fonction des parties prenantes et peuvent être conflictuels ou contradictoires. Ainsi, l'accès aux données personnelles est un besoin pour certains afin de leur permettre d'améliorer la qualité des services qu'ils offrent alors que leur protection est un souci pour d'autres. Il s'agit de trouver le bon équilibre entre ces différents intérêts, et ce dans l'intérêt de tous. Pour cela, il est nécessaire d'identifier, pour chaque IA, les bénéfices apportés à chacune des parties, les risques portés par chacun et les ressources et efforts nécessaires. Différentes options seront à étudier et des arbitrages devront être réalisés.

Dans ce contexte, les thèmes critiques pour parvenir à des IA de confiance ont été identifiés :

- **les bénéfices attendus par les parties prenantes doivent être définis par des principes et des exigences à respecter**
- **les risques pouvant impacter le respect des principes et exigences doivent être identifiés, pris en compte et traités pour les rendre acceptables**
- **la gouvernance doit permettre d'arrêter les choix, de fixer les objectifs et définir les responsabilités envers les IA**
- **les efforts et ressources à mobiliser pour atteindre les exigences avec un niveau de risque acceptable doivent se traduire par le déploiement d'outils et de dispositifs adaptés et soutenables.**
- **la mise en œuvre des points précédents pour certaines IA et pour certains contextes doit être vérifiée et validée par des audits et des certifications**

Pour répondre à ces thèmes critiques, nous avons développé une démarche en cinq étapes, schématisée ci-après. Ce ne serait que lorsque ces cinq étapes seront maîtrisées que nous pourrons espérer obtenir des systèmes à base d'Intelligence artificielle dignes de confiance. Les étapes sont à traiter dans l'ordre : il n'est évidemment pas logique d'espérer pouvoir traiter les risques ou auditer une IA si on ignore quels principes et exigences on cherche à mettre en œuvre.

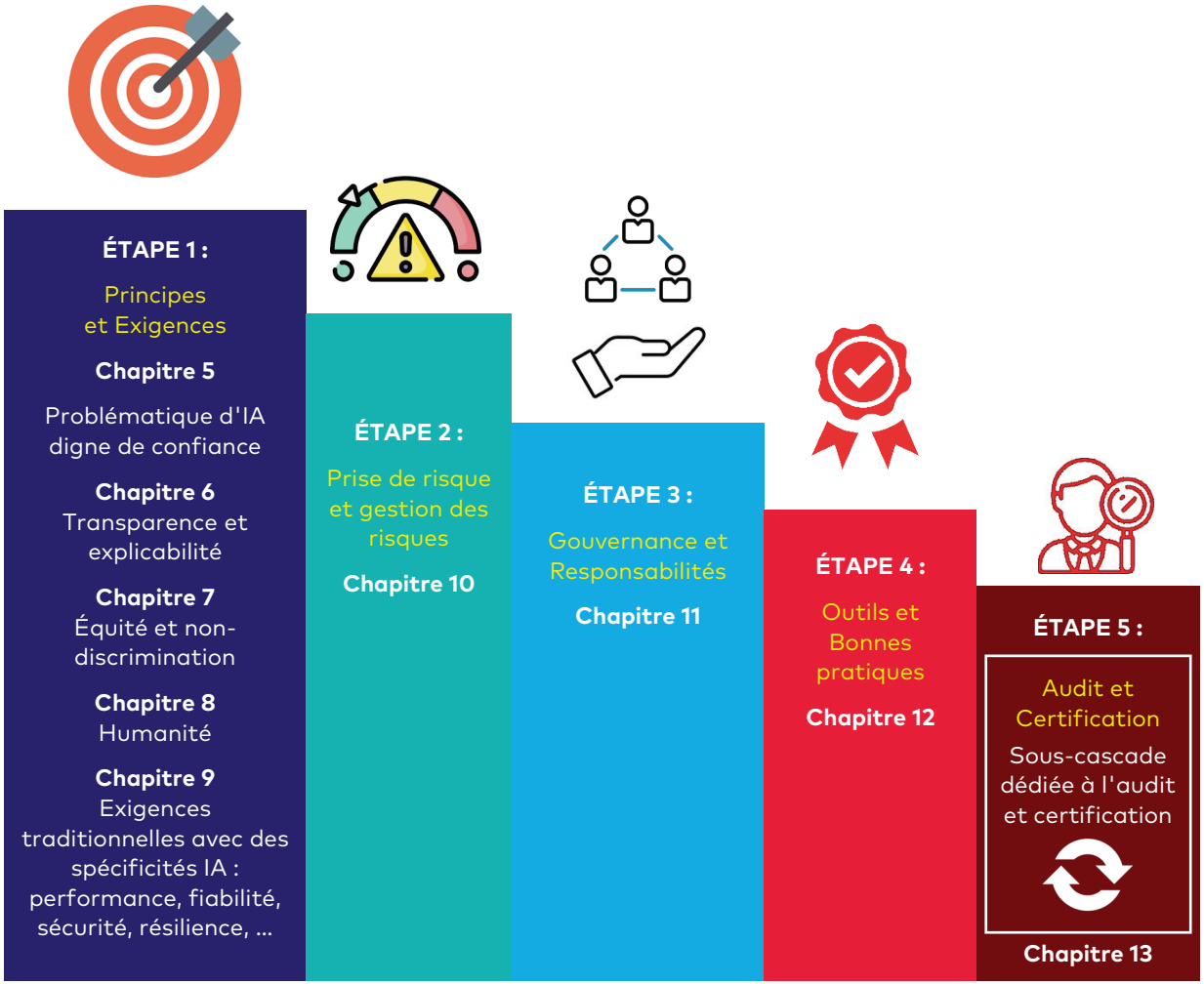

Figure 4 : Schéma de la cascade des cinq options critiques à prendre en compte

CHAPITRE 4

Nous allons présenter plus en détail ces cinq étapes — qui constituent le cœur de notre démarche — dans les prochains chapitres. Nous identifierons pour chacune d'elles les problématiques spécifiques qu'il conviendra de traiter, les outils et dispositifs qui seraient susceptibles de les traiter et quelques difficultés de mise en œuvre. Nous les illustrerons au travers de nombreux exemples. Par ailleurs, nous proposerons différentes mesures de nature à répondre à ces problématiques.

CHAPITRE 5

PRINCIPES À RESPECTER ET EXIGENCES SPÉCIFIQUES À ATTEINDRE : PROBLÉMATIQUES GÉNÉRALES DES SYSTÈMES À BASE D'INTELLIGENCE ARTIFICIELLE

L'acceptation des systèmes d'IA et leur diffusion massive ne pourra se faire qu'à partir du moment où la société saura que l'IA est digne de confiance. Il faut toutefois d'abord s'interroger sur la signification de ce terme : comment peut-elle être atteinte ? À quels niveaux peut-elle s'appliquer ? Surtout quels principes faut-il respecter, quelles exigences faut-il atteindre pour être digne de confiance ?

Nous décrirons donc les principes – à suivre – et les exigences – qui engagent – attendus pour constituer une IA digne de confiance. Il est à noter que ces principes et exigences se complètent et certains éléments peuvent se recouvrir les uns les autres.

1. La problématique générale à l'IA : une nécessaire identification, clarification et catégorisation des principaux principes et exigences pour des IA dignes de confiance

L'engouement et le développement des systèmes à base d'intelligence artificielle reposent en partie sur la grande diversité d'applications qu'on peut en faire et de solutions qu'ils peuvent apporter. Cette force constitue aussi leur faiblesse : le terme « intelligence artificielle » s'applique à des systèmes radicalement différents. Rares sont les points communs entre les voitures autonomes et Parcoursup. Pourtant, il faudrait que nous puissions avoir confiance aussi bien dans le premier que dans le second et dans toutes les autres. Comment identifier les principes à respecter et les exigences à atteindre qui rendent une IA digne de confiance ? Ces principes et exigences sont-ils tous les mêmes pour tous les systèmes ?

2. Une première structuration des grandes options de périmètres possibles

2.1. Confiance dans les résultats issus des IA

Nous avons vu que les IA contribuent à produire des diagnostics, des modèles d'apprentissage, des prévisions, des prescriptions, des décisions individuelles, des actions intégrées à des robots... L'objectif final est d'avoir confiance dans le fait que ces résultats issus des IA soient fiables, pertinents, équitables, compréhensibles... Un niveau raisonnable de confiance quant aux différentes caractéristiques attendues de ces résultats doit être défini. La qualité de ces résultats découle de la qualité des IA mais aussi de la qualité des dispositifs mis en œuvre localement autour de l'IA comme le processus de qualité, les processus de validation humaine...

2.2. Confiance dans les IA et leurs différents éléments

Il est nécessaire d'avoir confiance dans les IA elles-mêmes et dans leurs différents éléments (algorithmes, données...). Cela signifie avoir confiance dans la qualité de certaines de leurs caractéristiques, par exemple leur fiabilité, leur robustesse, leur sécurité, leur traitement équitable... Un niveau raisonnable de confiance quant aux différentes caractéristiques attendues de ces IA et de leurs éléments doit être défini. Combiné aux dispositifs locaux mis en place pour des usages spécifiques, ils permettront d'avoir confiance dans les résultats des IA.

2.3. Confiance dans les moyens mobilisés

Par ailleurs, pour obtenir ce niveau d'exigence souhaité des IA, il faudra mobiliser des moyens appropriés, que ce soient des outils, des logiciels, des processus, des compétences, des documents... Il faudra donc aussi avoir raisonnablement confiance dans la qualité de ces moyens, c'est-à-dire qu'ils soient robustes, fiables, disponibles en temps opportun, transparents... Un niveau raisonnable de confiance quant aux différentes caractéristiques attendues de ces moyens doit être défini.

2.4. Confiance dans les dispositifs mis en œuvre pour s'assurer du respect des principes par les systèmes à base d'intelligence artificielle

Il ne sera pas possible d'avoir confiance dans les systèmes à base d'IA si on n'a pas confiance dans chacune des composantes de l'IA et dans les résultats obtenus. Mais ce n'est pas suffisant. Il faudrait aussi avoir confiance dans les dispositifs qui sont mis en œuvre pour s'assurer que les principes ont été respectés et qui ont pour but de donner confiance dans les composantes et les résultats. Il s'agit d'avoir confiance non seulement dans l'IA elle-même mais aussi dans l'écosystème des dispositifs qui l'entoure ; plus encore : il s'agit non seulement d'avoir confiance dans les différents moyens pris individuellement mais aussi pris dans leur globalité, c'est-à-dire dans leur capacité, une fois combinés et intégrés, à donner confiance quant au respect des exigences attendues des IA et de leurs composants.

Il pourra alors s'agir d'appliquer certains principes pour arriver à l'atteinte de ces exigences. Par exemple, on pourra vouloir s'assurer que les principes de proportionnalité des moyens mis en œuvre à la vue des résultats attendus, de minimisation de l'utilisation des ressources, de précaution, de vigilance... ont bien été respectés.

2.5. Des types d'exigences attendues spécifiques pour chacun de ces éléments

Il convient donc de bien distinguer ces différents niveaux d'exigence.

On peut envisager d'obtenir une confiance sur des exigences concernant :

- les résultats des IA
- les IA elles-mêmes et leurs éléments
- les moyens pris individuellement.

Il peut aussi s'agir d'avoir confiance dans la combinaison et l'intégration des moyens pris dans leur ensemble pour satisfaire aux principes et exigences des résultats issus des IA, et des IA elles-mêmes et de leurs éléments.

3. Les IA et les systèmes d'informations financiers et comptables : un parallèle pertinent

Pour illustrer cette problématique, faisons un parallèle avec les informations financières. Il existe différents types d'informations financières : les états financiers, les comptes comptables, les informations de gestion, les informations prévisionnelles… Pour chaque type d'information, certains principes devront être respectés ; par exemple en matière de reconnaissance des revenus et des dépenses, de répartition entre plusieurs périodes, de présentation… En fonction des enjeux et du contexte de leur utilisation, un certain niveau de qualité sera attendu en termes de fiabilité, de pertinence, de validité, de disponibilité, de compréhension, de transparence, de confidentialité…

Pour aboutir au respect de ces principes et exigences, il sera nécessaire d'utiliser des systèmes d'information comptables (logiciel comptable, utilitaires, matériels…) associés à divers autres moyens et dispositifs (des directives, des processus, des compétences…). Il sera nécessaire d'avoir confiance dans le niveau de qualité des systèmes d'information comptables et des autres moyens et dispositifs mobilisés. En effet, chacun d'eux doit être fiable, robuste, disponible…

Et pour que l'on ait confiance *in fine* dans les informations financières, il faut aussi avoir confiance dans la qualité de l'intégration de l'ensemble des moyens mobilisés qui soit suffisante pour atteindre le niveau d'exigence attendu des informations financières. Il s'agit en fait d'avoir confiance dans le dispositif de contrôle interne* correspondant.

3.1. Principes et exigences définis et adoptés dans le domaine financier

Par ailleurs, il est apparu que de très nombreuses parties prenantes (actionnaires, banquiers, fournisseurs…) s'appuyaient très largement sur la confiance qu'ils portaient sur certaines de ces informations financières en l'occurrence les états financiers. C'est pourquoi, le marché, pour son bon fonctionnement, a défini de manière plus précise la nature exacte de ce qui était attendu des états

financiers en termes d'exigences de qualité. La nature de ces exigences a été étudiée et interprétée par les instances professionnelles comptables en lien avec les différentes parties concernées pour différents contextes. Il a été convenu que les états financiers se devaient d'être sincères, de donner une image fidèle de la situation financière et réguliers, c'est-à-dire conformes aux principes et règles en vigueur. Pour cela, il est nécessaire de s'assurer que les états financiers reflètent l'exhaustivité des opérations financières valides concernant une entité pour une période donnée de manière exacte qui soient conformes aux principes comptables et présentées de manière appropriée et compréhensible. Plusieurs caractéristiques ont donc été identifiées : exhaustivité, exactitude, validité, conformité, présentation, traçabilité…

3.2. Cette confiance peut passer par différents audits voire une certification du CAC

De plus, il a été décidé que, pour certaines organisations considérées comme critiques pour le bon fonctionnement du marché, ces exigences de sincérité des états financiers feraient l'objet d'une certification par un tiers de confiance – en l'occurrence des commissaires aux comptes (CAC) – qui engagent leurs responsabilités civile et pénale au cas où leurs travaux n'auraient pas été effectuées avec la diligence attendue. Dans le cadre de ces diligences, ils seront amenés à s'assurer que les systèmes d'information comptables satisfassent certaines exigences, notamment qu'ils soient fiables. Ils pourront être amenés à fournir une opinion spécifique sur ces systèmes comptables. Mais en général, leur opinion portera non pas sur la qualité du système d'information lui-même mais sur la qualité des moyens mis en œuvre pour aboutir à des systèmes fiables, soit en réalité sur la qualité d'une partie du dispositif de contrôle interne mobilisé à cet effet. D'autres informations financières et d'autres exigences existent et il a été considéré par le marché qu'elles pourraient être traitées au cas par cas directement par les parties prenantes concernées qui auront alors à définir la nature et les différents niveaux d'exigences alors attendues. Plusieurs types d'audit et plusieurs types de labels correspondant aux niveaux de confiance attendus pour ces cas ont aussi été définis.

3.3. Un parallèle qui s'applique aux IA mais avec plusieurs limites

Ce parallèle réalisé sur les informations financières s'applique aussi aux IA. Par exemple, là aussi, il existe différents types d'exigences concernant différents éléments. Il existe les exigences « *in fine* » sur les résultats issus des IA, sur les IA elles-mêmes comme l'atteinte d'une IA robuste, d'une IA équitable ou d'une IA transparente. Mais pour l'IA, nous n'avons pas les mêmes reculs et expériences que nous avons pour les informations financières. L'IA n'est pas suffisamment mature pour que nous soyons à même de définir précisément à quoi correspond une IA transparente ou équitable comme cela a été fait en ce qui concerne la sincérité des états financiers. De même, il apparaît que les outils nécessaires à l'atteinte de certaines exigences (explicabilité, traitement non biaisé…) ne soient pas tous disponibles sur le maché, leur développement n'ayant pas encore abouti pour tous les types d'IA. Il convient donc non seulement de préciser à quoi correspond réellement chacune des exigences mais d'identifier la nature des outils disponibles ou en cours de développement, dans quels cas ils pourraient être utiles et pour quels types d'IA et préciser leurs limites.

Dans ce contexte, il semble qu'il soit trop tôt à ce stade pour imaginer obtenir un niveau de confiance global élevé quant au respect des différents principes et exigences pour les résultats issus des IA et pour les IA elles-mêmes. Avec le temps, les outils disponibles et la maturité s'accroîtront, et le niveau de confiance s'améliorera.

4. Une approche pragmatique et évolutive à privilégier

Il est néanmoins possible d'obtenir un certain niveau de confiance sur la qualité de certains moyens, par exemple la qualité de la documentation mise en œuvre pour favoriser l'atteinte de certains résultats ou de certaines exigences. Le niveau global de confiance sera alors limité mais on pourra s'assurer à minima de la présence de quelques fondamentaux nécessaires à des IA de confiance. C'est ce que propose par exemple LNE (Laboratoire National d'Essais) avec son référentiel de certification « processus de conception, de développement, d'évaluation et de maintien en conditions opérationnelles des intelligences artificielles ».

Compte tenu des enjeux et des risques associés à certaines IA, on peut imaginer, comme c'est le cas pour les états financiers et tel que cela est proposé par la Commission européenne, que le respect de certains principes et exigences de ces IA nécessitent une certification par des tiers externes, en quelque sorte, des Commissaires aux IA. Aussi, pour émettre une opinion quant au respect de certaines exigences des IA, il convient de disposer d'une compréhension commune de ce que recouvrent précisément ces exigences, comme c'est le cas pour la sincérité des comptes ou la fiabilité des systèmes d'informations. Il est aussi nécessaire que des outils existent pour satisfaire au respect de ces principes et exigences et pour que l'auditeur puisse évaluer leur capacité à les satisfaire.

4.1. Un préalable : une compréhension commune des différentes options en matière de principes et d'exigences

Dans ce contexte, ce qui pose un problème particulier est d'identifier les différents types de principes et d'exigences – qu'ils soient généraux ou spécifiques – qui sont attendus des IA. S'agit-il de s'assurer que les IA soient sincères ou fiables ? Ou s'agit-il de s'assurer qu'ils soient explicables, équitables et dignes d'humanité ? Quelles que soient ces exigences attendues, elles devront être bien comprises. Il convient en effet d'identifier ce qu'elles pourraient recouvrer précisément, de les catégoriser et regrouper et de trouver une compréhension commune. Ceci devrait permettre aux différents acteurs concernés de procéder, avec une démarche commune cohérente acceptée par tous :

- à l'identification des types de risques qui pourraient apparaître et contrecarrer l'atteinte de ces exigences
- à l'identification des types d'outils et de dispositifs qu'il conviendrait de mettre en place pour favoriser le respect de ces exigences
- à l'identification des contraintes et des limites associés à ces dispositifs.

Ceci permettrait l'élaboration de référentiels de bonnes pratiques partagés et adoptés par les acteurs concernés. Ceci permettrait aussi aux auditeurs de s'appuyer sur cette compréhension commune pour émettre des opinions et, le cas échéant des réserves, cohérentes et comprises par tous ; comme c'est le cas pour la certification des états financiers.

4.2. Une première classification des exigences en grandes familles d'exigences

Une première analyse de la situation montre qu'étant un système informatique, l'IA doit répondre aux exigences traditionnelles des systèmes informatiques. Toutefois, certaines de ces exigences, en raison du fait qu'elles sont appliquées à un système d'IA, présentent des spécificités qu'il convient de bien identifier et traiter. Par ailleurs, certaines autres exigences semblent plus prégnantes dans le cas des IA que dans le cas des systèmes d'information plus classiques. Les prochains paragraphes sont donc consacrés à la description de ces deux types d'exigences plus traditionnelles avec quelques spécificités liées aux IA (§ 5) ou plus présentes dans le cas des IA (§ 6). La définition précise des exigences et de leurs composantes est difficile parce qu'elle reflète la complexité du domaine de l'IA, de ses nombreux intérêts divergents, de ses nombreux acteurs. Chaque partie voudrait satisfaire ou ne pas satisfaire à certaines exigences tandis que la définition d'une exigence particulière varie d'un acteur à l'autre.

Comme il existe de nombreuses exigences attendues et de nombreuses composantes pour chacune d'entre elles, certaines sont prioritaires par rapport à d'autres. Il convient de définir cette hiérarchie, toujours en fonction du contexte d'application de l'IA. C'est également pour cette raison qu'un besoin de classement et de regroupement des principes et exigences se fait sentir : chacun présentant ses propres listes en fonction de son point de vue. Par exemple, l'OCDE a identifié cinq exigences générales tandis que l'Union européenne en a identifié sept. Comme noté plus haut, les exigences diffèrent en fonction des attentes des parties prenantes et ne prônent pas les mêmes principes à respecter.

4.3. Le choix d'un qualificatif général : des IA dignes de confiance

L'illustration la plus évidente de ce fait concerne la désignation d'un qualificatif général recherché qui pourrait se traduire par une IA éthique, une IA conforme à la réglementation, une IA responsable ou d'une IA digne de confiance. Plusieurs organismes ont considéré le qualificatif d'IA éthique et en ont donné des définitions. Mais qu'en est-il réellement ? Souvent les choix éthiques lorsqu'il y a un consensus se retrouvent dans le droit et la réglementation. Ne reste-t-il pas alors pour l'IA éthique que l'articulation des aspects moraux non couverts par la réglementation ? Ce qualificatif est beaucoup plus difficile à appréhender et prête beaucoup plus à des interprétations individuelles. C'est pour cela que les différents documents internationaux (comme ceux de la déclaration de Montréal ou de l'OCDE) font davantage référence au concept d'une IA digne de confiance. D'autres ont mentionné des IA responsables. Nous avons choisi d'utiliser le cadre le plus englobant d'IA dignes de confiance. En effet, on ne peut avoir confiance dans les IA que si les aspects éthiques tel le respect des droits fondamentaux ont été respectés et que si des responsabilités ont été définies. Mais les IA dignes de confiance nécessitent aussi que des principes et

exigences qui ne sont pas tous de nature éthique aient été respectés. Il convient par exemple que les IA soient robustes, fiables, explicables… et que les dispositifs mis en œuvre soient efficaces, fiables…

Pourquoi avons-nous choisi la terminologie d'IA « digne de confiance » et non pas d'IA « de confiance » ? Lorsqu'une personne est dite de confiance, cela signifie qu'on l'a désignée comme telle. C'est par exemple celle à laquelle le médecin se référera pour prendre certaines dispositions médicales en cas d'incapacité d'un patient de faire ces choix. Cela n'implique pas qu'elle soit réellement digne de confiance. En effet, la confiance se mérite. Il faut pouvoir la démontrer. Pour que des IA soient dignes de confiance, il faut avoir confiance dans ses éléments constitutifs. Ainsi, il s'agit d'avoir confiance dans les dispositifs mis en place, dans les compétences et les outils sollicités, dans les technologies utilisées, dans les fournisseurs qui vous les ont proposés… Comment obtenir cette confiance ? En démontrant que l'on a été diligent, que l'on a mis en place de bonnes pratiques. On peut aussi obtenir cette confiance en demandant à des tiers indépendants dits de confiance d'émettre une opinion sur les éléments constitutifs de la confiance. C'est ainsi que l'on pourra disposer d'IA dignes de confiance.

Dans ce contexte, quels principes et exigences pour des IA dignes de confiance ?

Mais quels qualificatifs recouvriraient en réalité des IA dignes de confiance ? Le groupe de travail en a identifié plusieurs et les avons regroupés par grande famille. Nous allons maintenant présenter ces grandes familles d'exigences attendues des IA, qu'elles soient traditionnelles ou plus spécifiques aux IA.

Dans les paragraphes suivants, pour chacune de ces grandes familles d'exigences, nous préciserons de manière plus détaillée leurs différentes caractéristiques, expliciterons en quoi elles sont spécifiques aux IA et en quoi et dans quel contexte elles peuvent poser un problème. Pour permettre une meilleure compréhension, nous avons illustré ces problématiques par des exemples concrets, en identifiant des risques spécifiques qui pourraient y être associés et des types d'outils et de solutions avec leurs limites actuelles qui pourraient être mis en œuvre pour que les exigences puissent être atteintes. Des recommandations ont été émises pour faciliter la finalisation de ce cadre commun de principes et d'exigences.

Une fois que les principales exigences nécessaires au développement et au déploiement d'IA dignes de confiance auront été identifiées, nous pourrons ensuite dans les chapitres suivants traiter plus particulièrement les autres grandes problématiques qui y sont directement rattachées, que ce soit la nature et le traitement des risques spécifiques à ces exigences pour les IA, l'impact sur la gouvernance et les responsabilités associées, la mise en œuvre d'outils et de bonnes pratiques adaptés et les types de certification qui pourront apporter les niveaux de confiance appropriés en fonction des différents contextes.

CHAPITRE 5

5. Quatre grandes familles d'exigences « traditionnelles » des systèmes d'information avec des spécificités IA : performance, fiabilité, sécurité et résilience

Tous les systèmes d'information[31] répondent à des exigences, définies lors de l'étape de spécification. Toutefois, certaines de ces exigences « traditionnelles » ne présentent pas tout à fait les mêmes enjeux lorsqu'elles sont appliquées à un système d'IA. C'est le cas des exigences suivantes.

5.1. Les exigences de performance

Classiquement, les performances d'un système d'information énoncent les possibilités maximales ou optimales du système. Diverses mesures sont possibles : temps de réponse pour effectuer une tâche, débit (vitesse d'exécution d'une tâche), disponibilité du système…

En IA, cette exigence devient plus complexe avec l'apparition de nouveaux types de mesures pour :

- l'efficacité (dans l'atteinte des finalités : un bon diagnostic, une bonne décision, un bon modèle d'apprentissage…)
- l'exactitude du résultat (dans un système d'information, le résultat est juste ou faux ; en IA, avec une approche d'apprentissage probabiliste il peut être plus ou moins juste et donc plus ou moins acceptable)
- l'efficience (la quantité de données utilisées pour l'apprentissage, la puissance consommée…).

Il peut même être une nuance des mesures existantes. Ainsi, l'efficience en SI (système d'information) désigne la faible utilisation des ressources tels processeur, mémoire, consommation électrique… tandis qu'en IA, il peut également désigner la manière d'atteindre le résultat.

Les exigences de performance peuvent être en contradiction avec d'autres exigences. Par exemple, la limitation de l'utilisation des données personnelles ou leur anonymisation conduira à une performance algorithmique moindre mais limitera les risques en matière d'exigence relatifs à la confidentialité des données personnelles. Les efforts associés à la collecte d'une masse de données et l'utilisation d'une forte puissance de traitement susceptibles d'engendrer des innovations pourraient se faire au détriment d'une utilisation des ressources qui soit responsable et cohérente avec les exigences climatiques. Comment appliquer les principes de minimisation et de proportionnalité qui conditionnent ces choix ?

---

[31] Dans la suite de cette partie, le terme « Système d'Information », abrégé en SI, désigne les systèmes d'information sans intelligence artificielle. IA ou système d'IA continue de faire référence à un système d'information présentant au moins un algorithme « d'intelligence artificielle ».

5.2. Les exigences de fiabilité

La fiabilité désigne pour les systèmes d'information la capacité du système à assurer sa mission durant une durée donnée, être conforme à ce qui est attendu et ne pas engendrer des erreurs. La fiabilité d'un système d'information repose sur le respect de notions sous-jacentes. Pour être fiable, un système d'information doit être sincère, loyal, exact... Mais dans le contexte des IA, ces exigences peuvent revêtir un sens légèrement différent. Étudions par exemple la notion de sincérité. Faisons un parallèle avec les états financiers : pour être sincères, ces derniers doivent représenter une image fidèle de la réalité. Mais dans le cas des diagnostics, des décisions, des actions issus des IA, comment déterminer s'ils sont sincères c'est-à-dire s'ils donnent bien une image fidèle de la réalité ? On sait que le processus d'apprentissage n'assure pas la justesse des IA. Les concepteurs d'IA savent que la technique d'apprentissage utilisée peut influencer la justesse des IA, mais cette dernière est aussi modifiée dans d'autres cas comme lorsqu'il y a des erreurs dans les données d'apprentissage utilisées. Les IA sont-ils donc sincères ? Le fait que la justesse ne soit pas assurée laisse entendre que non. Pourtant, il est souvent aussi fait mention du concept de loyauté qui traduit le fait que les IA font ce qui est annoncé (c'est à dire de répondre aux attentes des destinataires) et uniquement ce qui est annoncé. C'est en cela qu'ils sont sincères.

Intéressons-nous à l'exigence d'exactitude, très liée à la fiabilité dans un système d'information. Dans le contexte de l'IA, la notion d'exactitude doit se comprendre de manière plus large. Il s'agit souvent de s'assurer que le niveau d'exactitude annoncé ou attendu corresponde à la réalité (il y a par exemple 99 % de chances de réussite par exemple ou 95 % de précision).

Certaines problématiques concernant la fiabilité peuvent aussi apparaître alors qu'elles ne peuvent pas exister dans les systèmes d'information traditionnels. Prenons le cas d'une voiture autonome qui rencontre sur son chemin une voiture d'entretien des bas-côtés ou un accident pris en charge par les autorités alors que la route présente une ligne blanche. La voiture autonome peut-elle se permettre de franchir cette ligne blanche pour fluidifier la circulation ? Cette action non conforme à la réglementation est-elle acceptable ? Faut-il introduire une notion de tolérance au non-respect des règles ? Si un tel franchissement est acceptable, comment cela influence-t-il la fiabilité ? La voiture autonome est-elle moins fiable parce qu'elle se permet de violer certaines lois ou plus fiables car elle s'adapte mieux aux circonstances et répond mieux à son objectif de transport d'un point A à un point B ?

5.3. Les exigences de sécurité

La sécurité vise à empêcher l'usage malveillant, non conforme ou non valide d'un système d'information. Il couvre les concepts classiques de confidentialité (restriction de l'accès aux personnes autorisées), d'authentification (l'utilisateur prouve son identité grâce par exemple à un code d'accès), d'intégrité (les systèmes d'information ne doivent pas être modifiées de manière accidentelle ou non valide), de disponibilité (l'accès au système d'information doit être possible au moment voulu, avec un niveau de disponibilité suffisant durant les plages d'utilisation prévues et pour les usages prévus).

Dans le cas des IA, ces concepts subsistent mais les IA introduisent des spécificités à prendre en compte telles que la prise de décision automatisée non explicable, la sécurité des données d'entraînement, des modèles d'apprentissage… qui introduisent de nouveaux risques de sécurité. À ces concepts classiques, s'ajoutent d'autres aspects à prendre en compte de manière plus spécifique pour les IA. Il s'agit en particulier de :

- la confidentialité : même si elle est la plus connue du grand public grâce au RGPD, l'IA se base sur l'utilisation intensive de données personnelles et des données économiques, industrielles et publiques. La protection de ces dernières et des droits associés doit donc être étendue et renforcée par rapport à un système d'information classique. La notion de secret des affaires doit par exemple être prise en compte
- la vie intime : même si les données personnelles sont protégées par l'anonymisation ou d'autres techniques, il est possible dans le cas des IA de connaître ou reconnaître un individu en recoupant des données et d'utiliser cette compréhension de manière subliminale ou sur les fragilités qui en découlent et d'être nocif *in fine* pour lui
- la sûreté : la délégation potentielle des responsabilités (voiture autonome, IA de diagnostic de maladies graves) peut présenter de graves conséquences sur l'intégrité physique voire sur la vie humaine.

5.4. Les exigences de résilience

Face aux facteurs internes et externes, un système d'information résilient est robuste, capable de continuer à fonctionner quelle que soit la situation, fréquente ou pas, inattendue ou pas, comme des pannes, des pics d'activité, des incidents, des piratages… Il se doit d'être pérenne, de permettre cette continuité mais aussi d'être agile et de continuer de fonctionner lorsque des imprévus subviennent ou des évolutions fonctionnelles, technologiques, réglementaires ou autres apparaissent.

L'IA apporte des nuances sur certains points comme celui de la pérennité. Elle est très dépendante de l'accès très incertain aux technologies, aux données, à la puissance de traitement… aux mains d'un nombre très restreint d'acteurs (GAFAM…). Elle est très fragile du fait de son opacité, de la perte de contrôle possible du fait de son évolution autonome et des prises de décisions autonomes, de son absence de prise en compte des évènements rares, du bon sens… C'est ainsi que les notions supplémentaires apparaissent comme l'anti-fragilité : le système doit apprendre de ces situations exceptionnelles pour *in fine* pouvoir les affronter de manière renforcée dans le futur à la manière d'un organisme qui se fait vacciner.

6. Trois grandes familles d'exigences spécifiques aux IA : transparence et explicabilité, équité et non-discrimination, et humanité

Les particularités des systèmes d'IA sont telles que certaines exigences sont très prégnantes lorsque ces IA sont déployées. C'est le cas des trois exigences suivantes que nous décrirons rapidement ici avant de les détailler ultérieurement.

6.1. Les exigences de transparence et d'explicabilité

Les médias et les parties prenantes parlent souvent de la transparence de l'IA d'une manière générale. Or ce terme est en réalité complexe. La « transparence » peut faire référence à plusieurs notions différentes ; à tel point qu'il n'existe actuellement aucun consensus sur sa définition exacte dans le contexte des IA. Certains experts lui donnent une certaine définition et des qualités qui ne sont pas celles des autres experts.

Généralement, lorsqu'on parle de transparence, on fait référence à la transparence du choix des technologies, des techniques, des modèles, des données, des traitements, la nature de leurs impacts potentiels... La transparence est d'autant plus importante que la logique interne des systèmes d'IA semble inexplicable, surtout quand les règles de décision utilisées par les IA sont déduites par l'ordinateur et non fixées par l'humain. La CNIL constate[32] « les concepteurs mêmes de ces algorithmes probabilistes perdent la capacité à comprendre la logique des résultats produits ». C'est l'effet « boîte noire ». Les exigences de transparence – et les variantes et nuances qui s'y rapportent – ont pour but de transformer la boîte noire en une boîte grise, à défaut d'une boîte blanche où tout serait connu. En effet, la transparence n'est pas forcément la connaissance complète du code source, ce qui pourrait par ailleurs nuire au secret des affaires et s'avérer inacceptable par certains acteurs privés et publics, d'autant plus que la connaissance du code source ne serait d'aucune aide pour l'utilisateur lambda. Pour certains, la transparence est l'exigence selon laquelle les systèmes d'IA doivent être conçus et mis en œuvre de façon à en permettre le suivi et la supervision.

Pour quelques experts, il existe une deuxième sorte de transparence orientée sur l'aspect social. Il s'agirait de la transparence des parties prenantes, de leurs intérêts et des usages prévus de l'IA. Ce type de transparence leur a semblé important lorsque certaines questions se sont posées : l'IA ne profiterait-elle pas à quelques personnes au détriment d'un autre groupe ? N'y aurait-il pas manipulation des grandes sociétés pour servir leurs intérêts mercantiles ? N'utiliseront-elles pas les informations à des fins non avouées ? Nous avons évoqué précédemment la notion de loyauté (§ 5.2), cela peut se rejoindre : les IA doivent-elles faire ce qui a été annoncé et ne faire que cela ? Cette notion de la transparence ici est largement associée à la communication faite aux usagers et la confiance qui y est associée. En effet, si cet aspect de la transparence semble parfois flirter avec la théorie du complot, c'est qu'il y aurait un problème de communication des responsables des IA envers les consommateurs.

Pour bien cerner les limites du concept de transparence au niveau technologique et technique, il convient de définir les autres termes qui s'y rapportent :

---

[32] CNIL, « Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle », 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, décembre 2017

- **explicabilité, interprétabilité, intelligibilité, justifiabilité** : dans le contexte des IA, plusieurs éléments méritent d'être compris. Par exemple, comment le modèle d'apprentissage a été développé, sur quelle base, avec quel type de données et quelles sources pour ces données, quels paramètres de traitement, comment il fonctionne, comment le résultat issu de ce modèle a été établi, sur la base de quels éléments précis, peut-on reconstituer la décision à partir des éléments en entrée… ? Il est souvent aisé de comprendre comment et pourquoi le modèle a été construit mais parfois plus difficile de comprendre comment le modèle lui-même fonctionne. Il est parfois encore plus difficile de comprendre le cheminement exact qui a mené à une décision particulière ou à un résultat. Cette compréhension peut s'avérer en réalité assez simple lorsque les techniques d'apprentissage sont déterministes et que le nombre de variables est limité mais particulièrement difficile lorsque ces techniques sont opaques, proviennent d'IA « boîtes noires », lorsque le nombre de variables utilisé est significatif ou que le nombre d'algorithmes impliqués est important. Il en résulte plusieurs types et niveaux de compréhension possibles. Associé à ces types de compréhension, on retrouve plusieurs qualificatifs tels intelligibilité, interprétabilité, explicabilité et justifiabilité qui sont souvent utilisés de manière indissociable mais qui peuvent recouvrer des notions bien distinctes. Par exemple :
  - **les IA interprétables** identifieraient les variables et leurs caractéristiques voire leur importance (quels types de données, quels types de calculs…), qui ont contribué à arriver à un résultat précis. Ainsi, on peut avoir une bonne compréhension théorique du modèle sans comprendre la sortie particulière d'un résultat
  - **les IA intelligibles** s'intéresseraient au niveau de compréhension des différents interlocuteurs qui ont des niveaux de technicité variées, informaticiens ou utilisateur lambda par exemple. Or, plus le niveau de compréhension initial, plus il faut simplifier et donc plus l'explication finale est inexacte. Ces IA peuvent être vulgarisées et leur fonctionnement globalement compris sans s'intéresser forcément à décrire chaque détail
  - **les IA justifiables** qui étudieraient la décision prise et non pas l'algorithme lui-même ou le processus qui a conduit à la décision. Il s'agit de montrer que la décision est bien fondée, qu'on peut en montrer la vérité ou la preuve. Il n'est pas nécessaire d'avoir une bonne compréhension de l'algorithme. Dans le cas de Parcoursup, par exemple, cela permettrait de dire à un élève pourquoi il a été reçu dans une filière (par exemple, la variable « note en mathématique » est suffisamment élevée ou parce que la règle non algorithmique « Tout élève est reçu si sa moyenne dans toutes les matières est supérieure à 10 »
  - **les IA explicables** s'intéresseraient aux relations directes explicites entre les résultats et les caractéristiques et données d'entrée connues. On est capable d'expliquer un résultat, un peu à la manière d'une IA justifiable, mais de manière plus approfondie car on comprend le fonctionnement de l'algorithme. En comparaison l'IA justifiable fait la lumière sur le résultat obtenu simplement en expliquant les variables et les règles non-algorithmiques.

  Certains qualificatifs peuvent s'appliquer aussi bien à la procédure qu'aux résultats. Une clarification reste donc nécessaire. Il faut aussi rappeler que l'objectif est que les personnes concernées aient une

compréhension suffisante pour qu'ils aient confiance dans le fonctionnement des IA et dans les résultats obtenus. Il faut donc trouver un juste équilibre entre ces différents niveaux de compréhension.

- **auditabilité** : comme son nom l'indique, on fait ici référence à la capacité de faire inspecter par un tiers le fonctionnement de l'IA, la nature des résultats qui en est sortie, les traitements des anomalies potentielles ou des erreurs, les décisions qui ont été prises, l'affectation des responsabilités… et de vérifier leur cohérence et leur compatibilité avec des normes de référence. Pour cela, il est nécessaire que ces différents éléments aient été notés et conservés. Cela suppose aussi que les algorithmes et les modèles puissent être intelligibles et explicables et que les preuves soient conservées sur une période donnée de temps de manière à être évaluées
- **traçabilité et imputabilité** : Il est possible de tracer un grand nombre d'éléments comme les décisions, les opérations, les actions, les incidents… Il peut aussi être possible de les imputer à ceux qui en sont à l'origine. Il faut noter cependant que pour certaines IA, leur origine peut découler directement du modèle d'apprentissage, modèle qui évolue en continu de manière automatisée et qui de ce fait peut rendre très difficile une traçabilité complète et compréhensible. Le recueil de ces informations passe donc le plus souvent par leur enregistrement dans un journal de logs* qui sera conservé. La traçabilité peut être utile pour permettre l'atteinte de plusieurs exigences telle que d'expliquer le résultat d'une IA, d'identifier des responsabilités, d'auditer la conformité à des règles de référence….

Il convient donc de clarifier ces notions, d'identifier les éléments qui nécessitent une transparence, les contextes pour lesquels cette transparence serait souhaitée, comment elle serait mise en œuvre… Ces points seront abordés ultérieurement dans le chapitre 6.

6.2. Les exigences d'équité et de non-discrimination

La notion d'équité n'est pas facile à décliner. Elle est multiforme. La notion anglaise d'équité est « fairness ». Elle est souvent associée à la notion d'égalité.

Lorsque l'on évoque la notion d'égalité, on l'associe à une situation de fait. Il peut s'agir de s'assurer que toutes les personnes aient un même accès, aient un même résultat, de mêmes droits, un même traitement pour une même situation… De nombreux facteurs pourraient contrecarrer cette égalité : un biais implicite ou explicite, une erreur, une mauvaise interprétation…

Lorsque l'on évoque la notion d'équité, on l'associe à un jugement qui peut avoir une composante éthique ou morale. On tiendra compte d'un contexte, d'une situation ou besoin individuel, ou du fait d'être un membre d'une minorité qui serait ou aurait été désavantagée dans le passé en tant que membre de ce groupe même si, individuellement, la personne n'avait pas été désavantagée. On pourra alors désirer ajuster les moyens pour tendre vers une fin souhaitée qui serait une égalité de résultats, de droits, d'un accès… Dans ce cas, la situation d'origine peut découler de plusieurs facteurs : un handicap, une discrimination, une erreur, une injustice… ou le fait d'appartenir à un groupe particulier. Il s'agira de compenser cette situation ; en faisant de la discrimination positive par exemple.

Dans tous ces cas, on appréciera une situation de fait ou un contexte en introduisant des notions de justice, d'impartialité, de non-discrimination, de neutralité, d'objectivité, de besoin de diversité, … Mais cette appréciation pourra elle-même être biaisée.

Ces situations peuvent s'opposer voire être incompatibles. Lorsqu'un groupe sera privilégié, cela peut se faire au détriment d'un individu du groupe non privilégié qui n'aura pas le même traitement que s'il avait été dans le groupe privilégié. Serait-ce inéquitable ? De même, une personne pourrait être défavorisée au bénéfice d'un autre du seul fait de la présence d'un attribut. Imaginons par exemple deux jeunes diplômés à la recherche d'un emploi. Ils ont les mêmes qualités et le même parcours si ce n'est que l'un a une plus faible expérience car il a réalisé moins de stages. On pourrait attribuer ce fait à une situation sociale ou familiale défavorisée donc injuste. Devrait-on neutraliser cette différence pour des raisons d'équité alors qu'un recruteur pourrait la considérer comme objectivement différenciante ?

Il y a donc des appréciations à porter et des choix à effectuer. Quelles situations doivent être considérées comme inéquitables ? Découlent-elles de biais, de discriminations, d'injustices à corriger… ? Dans quels cas un individu ou un groupe doit être privilégié et à quelle hauteur doit-il l'être ? Qui doit effectuer ces choix et comment ? Ces choix sont souvent éminemment culturels. Comment s'assurer que ces choix ne sont pas eux-mêmes biaisés ? Ils peuvent impliquer une dimension éthique. Ainsi, si la mixité des classes avantageait les garçons au détriment des filles, doit-on la privilégier ? Serait-ce équitable, juste, éthique… ?

L'IA est très sujette aux biais. Certains biais cognitifs sont introduits de manière volontaire et dépendent de choix personnels. D'autres dépendent de choix politiques (promouvoir certaines catégories sociales, certaines minorités, les femmes…) qui peuvent être inclus dans les lois et réglementations. Certains biais cognitifs peuvent être mis en œuvre pour inciter un type de comportement considéré plus vertueux (ils sont appelés « nudges » ou coups de pouce). Cela est-il éthiquement acceptable ? Les biais cognitifs peuvent donc résulter de choix volontaires mais sont le plus souvent introduits de manière inconsciente.

Certains biais cognitifs vont conduire à des discriminations, d'autres pas. Certaines discriminations seront le fait de décisions volontaires et ne sont donc pas des erreurs. D'autres seront le fait de décisions subies et découlent d'erreurs de raisonnement.

Il convient donc de clarifier ces notions d'équité d'une part et d'erreurs, de biais, de non-discrimination d'autre part.

Ces notions sont en effet importantes car elles reflètent des interrogations cruciales. L'IA facilitera-t-elle ou non votre accès au marché du travail ou à un crédit parce que vous êtes une femme provenant d'une minorité ethnique ou un homme caucasien ?

Les biais peuvent provenir de plusieurs sources. Quelles sont-elles ? Pour parer à l'existence de ces biais et à leurs éventuelles conséquences néfastes, quels sont les outils disponibles ? Qui doit faire ces arbitrages et comment doivent-ils être effectués ?

Nous aborderons ces thèmes de manière plus détaillée au chapitre 7.

6.3. Les exigences d'humanité

Pour l'instant, les ordinateurs ne peuvent réaliser que les tâches programmées. Aussi les systèmes d'IA devraient être conçus de telle façon à ce qu'ils fonctionnent en étant compatibles avec les idéaux humains. En Europe, le document fondamental est la Charte des droits fondamentaux des personnes qui consacre ces idéaux. Ceux-ci sont répartis entre six valeurs individuelles et universelles constituant le socle de la construction européenne : dignité, liberté, égalité, solidarité, citoyenneté et justice. Cette charte comporte 54 articles que doivent respecter les IA. Plusieurs de ces droits ont été évoqués dans les paragraphes précédents notamment concernant le droit à la vie privée, à la dignité, à l'intégrité physique ou à la non-discrimination. Mais il y en a d'autres qui doivent aussi être pris en compte. Il s'agit alors de les traduire en termes d'exigences spécifiques pour les IA.

Au-delà, il convient de prendre en compte d'autres sujets éthiques :

- comment peut-on aborder l'utilisation responsable des ressources et appliquer les principes de leur minimisation et de leur proportionnalité avec les apports des IA ?
- comment s'assurer qu'il y ait eu libre arbitre traduisant bien la manifestation d'une volonté libre, spécifique, éclairée et univoque au consentement de l'utilisation des données personnelles et des usages qui en découlent ?
- comment garantir que l'on puisse aller en justice de manière éclairée lorsque l'on subit une discrimination ou un traitement non équitable ?
- comment être certain que les décisions automatisées qui peuvent avoir des impacts individuels soient maîtrisées *in fine* par des humains (notamment en actionnant le fameux bouton rouge) ?

Pour toutes ces situations, ces principes doivent se traduire en exigences qui puissent être mises en action opérationnellement. Dans quel contexte, comment, à quel moment, avec quelles responsabilités… ?

Nous aborderons certains de ces points au chapitre 8.

6.4. Recommandation : création d'une instance professionnelle pour finaliser le cadre commun des principes et exigences pour des IA dignes de confiance

Si les éléments constitutifs de ces exigences ont bien été intégrés, tous les acteurs ne sont pour l'instant pas d'accord sur ces exigences et leurs caractéristiques. Il sera donc important de mettre au point des définitions précises (par exemple pour un audit financier, le terme sincère est différent de fiable) et consensuelles, c'est-à-dire qui soient acceptées par l'écosystème des acteurs. Le premier de ces acteurs, les législateurs et les créateurs de réglementation, devront définir le périmètre du problème, c'est-à-dire quels sont les programmes informatiques qui seront considérés comme des IA ou non et quelles sont les principes et exigences à respecter et pour quel contexte. Cela permettra de définir quelles réglementations s'appliquent. Le second groupe d'acteurs sont les parties prenantes, les développeurs et les utilisateurs. Ces personnes devront respectivement, déterminer les exigences à mettre en place, qu'elles soient réglementaires ou pas, les mettre en œuvre et, pour les utilisateurs, donner leur avis. Enfin, les derniers acteurs sont les auditeurs qui certifieront que les IA et/ou les moyens mis en œuvre sont dignes de confiance sur la base d'une compréhension commune. Tous devront alors s'accorder sur les termes et leurs définitions.

Pour y parvenir, il sera nécessaire de créer une instance professionnelle composée par les principales parties prenantes, comme il en existe pour les états financiers, pour définir de manière consensuelle ce cadre commun de principes et d'exigences pour des IA dignes de confiance.

6.5. Accompagnement des parties prenantes

Une fois le cadre des principes et exigences définies, il faudra que toutes les parties prenantes intègrent ces principes et exigences dans leurs pratiques.

A l'instar des experts-comptables qui accompagnent les entreprises dans l'application des principes comptables, des experts accompagneront les entreprises dans la mise en œuvre et l'application des principes et exigences en matière d'IA et ce, en amont des opérations d'audit.

« Avoir confiance dans l'IA » est une expression un peu floue. Il faut en réalité avoir confiance dans les résultats d'un système d'information à base d'intelligence artificielle, dans le système lui-même et ses composantes, dans les ressources et efforts mis en œuvre pour créer et/ou déployer cette IA et enfin, dans les dispositifs mis en place pour s'assurer du respect des principes. La confiance globale envers les IA ne pourra être atteinte qu'en ayant confiance dans toutes ces composantes. Une confiance partielle peut être obtenue si tous ces éléments ne sont pas tous dignes de confiance.

Ce schéma est très similaire à celui existant pour les Commissaires aux Comptes : pour avoir une confiance globale dans un système d'information financier et comptable, il faut avoir confiance dans chacun des éléments du systèmes, dans son intégration et dans ses dispositifs de contrôle, qui sont entre autres les audits et les certifications. Cette comparaison présente des limites : on sait les éléments à contrôler, comment les mettre en place, comment les contrôler dans le cas des systèmes financiers et comptable car on a de nombreuses décennies de recul sur la question. Les réflexions et les outils se sont raffiné au cours du temps. L'IA reste trop jeune pour avoir vu émerger toutes les réponses. Cela viendra au fur et à mesure que les outils et moyens de contrôle seront créés. Ces premières réponses ne seront pas optimales mais seuls l'expérience et le temps permettront de les améliorer.

La première étape pour bâtir ces réponses est de définir les exigences à atteindre, à la manière du jardinier qui décide du plan de son parc et des essences et fleurs à planter. Les systèmes d'information basés sur des Intelligences Artificielles sont pleinement des systèmes d'informations et les exigences classiques peuvent leur être imposées. Toutefois, la composante « intelligence artificielle » de ces systèmes impose des évolutions par rapport à ces exigences classiques. Ainsi les exigences de performance, de fiabilité, de sécurité et de résilience sont différentes dans le contexte d'un système avec ou sans IA. Il est également nécessaire de s'intéresser aux nouvelles exigences qui n'avaient pas raison d'être dans un système d'information traditionnel. Ce sont les exigences de transparence et d'explicabilité, d'équité et de non-discrimination et d'humanité.

Ces différents types d'exigence seront étudiées dans les prochains chapitres.

CHAPITRE 6

EXIGENCES DE TRANSPARENCE ET D'EXPLICABILITÉ

Au début de l'informatique, les algorithmes étaient codés sur quelques bits. Avec un peu de pratique, le déchiffrage du code et la compréhension de l'outil était simple. Avec les décennies, le nombre de bits s'est envolé. On a eu besoin d'inventer de nouveaux langages : des langages de haut-niveau qui se rapproche du langage naturel pour que l'humain donne ses instructions à la machine et des langages de bas-niveau qui traduise ces instructions en des mots compréhensibles par les processeurs. Une étape était franchie : il fallait maintenant des spécialistes pour interpréter chaque partie du système. C'était plus compliqué mais la mise en place d'organisations adéquates suffisait à faire que l'ensemble de l'équipe soit à même de comprendre le programme. Avec l'introduction de l'intelligence artificielle dans les systèmes, on est passé de systèmes compliqués à des systèmes complexes. Désormais, même les spécialistes ont parfois du mal à interpréter ce qu'il se passe dans certains algorithmes.

Comment rassurer la population sur l'innocuité d'un système si même les experts ont des problèmes de compréhension ? Les citoyens peuvent facilement considérer que l'IA est un terrain de jeu pour apprentis sorciers si rien n'est rendu plus clair. Pour cela, il va falloir expliquer les systèmes d'information à base d'intelligence artificielle et les rendre transparents.

Mais cette exigence est plus facile à énoncer qu'à atteindre. De nombreuses questions se posent : quoi rendre transparent, quand et comment ? C'est ce que nous verrons dans un premier temps avant de nous attarder sur quelques recommandations qui permettraient de commencer à résoudre le problème.

1. Problématiques spécifiques à l'IA : quels éléments, dans quelles circonstances, à qui, comment les rendre transparents

La Cour des comptes s'est interrogée sur l'anonymat des dossiers examinés pour la promotion Parcoursup 2019, alors qu'aucun élément ne permettait a priori d'affirmer qu'il y avait eu des discriminations liées au genre ou au nom dans la promotion précédente. Il en a été déduit qu'une forte transparence des algorithmes locaux permettrait de réduire le niveau de suspicion généralisé.

En effet, l'une des expressions les plus souvent associées à l'IA est celui de « boîte noire ». L'image commune qu'elle traduit est qu'on offrirait des données à un système artificiel qui les passe à la moulinette de ses algorithmes pour en recracher un résultat qui nous impacte. L'IA est notoirement connue pour être complexe, imprévisible, avec un comportement plus ou moins autonome, voire évolutif. Comment, avec une telle opacité, quand on n'y comprend rien, accorder sa confiance à des systèmes pouvant radicalement modifier la vie de chacun ? Comment contester une décision prise à notre égard, comment obtenir les preuves qu'on a subi un préjudice ? Comment vérifier que les systèmes d'IA respectent les lois et la dignité humaine ? L'une des solutions avancées est de transformer ces « boîtes noires » en « boîtes grises » ou

blanches par la mise en place de moyens permettant d'assurer la transparence et d'explicabilité dans ces systèmes.

Mais l'application de ces solutions n'est pas si simple car ce qui est transparent et explicable pour quelqu'un ne l'est pas pour un autre. Ce qui semblerait nécessaire de savoir pour l'un relève du secret des affaires pour l'autre. Sans compter que ce qui est explicable pour un type d'algorithme devient nettement plus nébuleux pour un autre, comme le montre ce schéma ci-dessous :

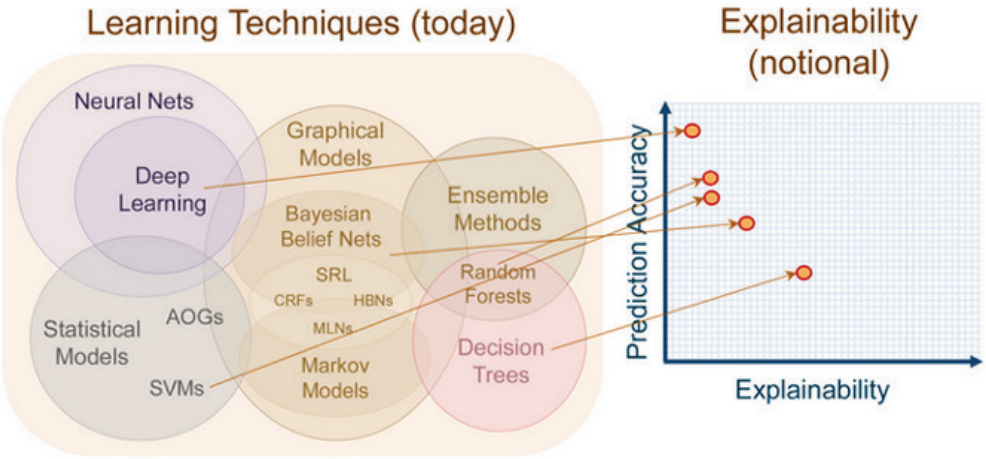

Figure 5 : Un schéma montrant qu'il n'est pas évident d'obtenir à la fois des modèles prédictifs précis et explicables. La technique du Deep Learning présente une très grande précision de prédiction (il se trouve en haut sur l'axe des ordonnées) mais la plus faible explicabilité de l'algorithme. À l'opposé, les techniques des arbres de décisions sont les plus explicables (car les plus à droite sur l'axe des abscisses) mais les moins précises dans la prédiction (car le plus bas sur l'axe des ordonnées (source : DARPA, 2016).

En effet, pour obtenir des résultats plus précis ou parvenir à réaliser des tâches plus complexes, il faut des modèles et donc des systèmes d'IA plus complexes. Or, d'après le JRC Technical report : Robustness and explainability of AI, publié par la Commission européenne *« rendre ces modèles plus interprétables semble à son tour entraîner presque inévitablement une perte de ces caractéristiques. La question de savoir dans quelle mesure les sorties d'un algorithme donné sont encore compréhensibles pour un humain ou même fondamentalement explicables de manière unique (par exemple, en raison des fonctions non linéaires utilisées dans de nombreux modèles d'apprentissage automatique) est cruciale pour une évaluation fiable de sa sécurité.* » Le chercheur Zachary Lipton a présenté lors de la conférence internationale du Machine Learning (ICML) en 2016, une opinion plus tranchée encore : pour lui, c'est l'interprétabilité qui pose un problème car elle empêcherait d'exploiter au maximum ces technologies. En effet, les modèles de machine learning ont pour essence d'extraire des modèles à partir d'une quantité de données bien plus importante

que ce que des humains peuvent manipuler. Cet automatisme est l'essence des algorithmes de machine learning. Le fait de vouloir à la fois utiliser des réseaux neuronaux et de vouloir qu'ils soient interprétables serait une erreur. En effet, l'interprétabilité d'un modèle dépend de sa simplicité, comme on peut le voir sur la Figure 5, ci-avant. Pourtant, les voix s'accordent pour dire que donner une explication des décisions et avoir des systèmes d'IA transparents est le seul moyen d'obtenir une relation de confiance entre l'IA et les différents acteurs concernés. De nombreuses questions se posent alors.

1.1. Quels éléments devraient être transparents ?

Il ne faut pas oublier que, quand on parle d'une IA, on sous-entend que cette IA fait partie d'un système technique et même d'un écosystème sociotechnique. Une IA est créée par une organisation au sein d'une équipe de développeurs, de data-scientistes, de responsables de la sécurité, de la RGPD et autres, mais aussi, quand toutes les ressources ne sont pas internalisées, d'une myriade de sous-traitants, de fournisseurs, de traitants... Souvent, un système d'IA est un puzzle dont les pièces (les modèles, les briques logicielles...) sont issues de différentes sociétés privées ou d'organisations publiques. Ces éléments interagissent et créent une part de l'IA finale. Au sein de cette abondance d'intervenants et de parties prenantes, chacun porte des responsabilités.

Les concepteurs et les développeurs ont normalement réalisé une analyse d'impact, de risques, ont instauré des procédures en cas de recours, des chartes éthiques, des structures de gouvernance... Faut-il connaître l'ensemble de ces éléments au risque d'être noyé sous la masse d'informations et de ne plus distinguer l'essentiel de l'accessoire ? Mais si l'on doit sélectionner une certaine part de ces éléments, comment les sélectionner ? À hauteur de leur implication dans le système d'information à base d'intelligence artificielle ? À hauteur des risques qu'ils portent ? Comment mesurer ces éléments ? Qui doit, parmi ces multiples acteurs, rendre des comptes aux solliciteurs ? Est-ce l'équipe de développeurs qui a créé l'algorithme ? Leurs responsables ou un délégué dédié dans l'organisation créatrice de l'IA ? Serait-ce davantage de la responsabilité de l'organisation qui déploie le système ?

En se restreignant seulement au sein de l'organisation qui a créé l'IA, on peut imaginer nombre d'éléments qui pourraient être à transmettre :

- **les affirmations de la société émettrice** concernant le respect des principes et des exigences. Mais lesquelles ? Dire que l'IA est conforme à la réglementation et/ou qu'il est certifié est-il suffisant ? Faudrait-il plutôt préciser les niveaux d'engagement, d'incertitude, de maturité de l'algorithme ? Mais dans ce cas, où s'arrêter ? Faut-il juste l'affirmer ou le prouver par des moyens techniques voire mathématiques ?
- **les finalités d'usage**. Quels sont les impacts potentiels de l'IA ? Sur qui ? Mais doit-on seulement estimer les impacts directs ? Comment estimer les impacts indirects ou rares ?
- **les recours possibles**. Quels sont les dispositifs de recours en place ? Qui a le droit de faire recours ? Faut-il obtenir des preuves ou des éléments probants avant de débuter un recours ? Dans ce cas, comment les obtenir ?

- **le contrat d'assurance**. L'organisation est-elle assurée contre les risques liés à l'IA ? Quelle est la nature précise de la couverture ? Quel est le niveau de risque estimé par la compagnie d'assurance ? Qui peut ou doit avoir accès à ces informations stratégiques ?

En se restreignant cette fois-ci, à la pure partie technique, d'autres questions se soulèvent :

- **la construction de l'IA.** Quelles sont les techniques utilisées (machine learning, deep learning, supervisé ou non supervisé…) ? Quels sont les modèles utilisés ? Quels types d'algorithmes ont été mis en place ? Quels ont été les tests et contrôles effectués ? Quels ont été les outils utilisés pour limiter les biais ? Si on considère que seuls les modèles simples peuvent s'expliquer, d'autres questions subsidiaires se posent : qu'est-ce qu'un modèle simple ? Est-ce un modèle avec peu de variables ? Avec peu de résultats différents ? Un modèle basique ?
- **l'apprentissage de l'IA.** Comment a été construite la base de données d'apprentissage ? Quels ont été les critères de sélection des données, à quel point sont-elles représentatives du monde réel ? Des profils ont-ils été construits ? Quelles sont les variables utilisées ? Quelles actions ont-elles été effectuées sur les données (nettoyage, échantillonnage, anonymisations…) et comment ?
- **les résultats.** Quelles sont les prédictions et décisions (octroi de crédit, recrutement…) réalisées ? Quelles sont les données de sortie (score, indicateur…) ? Quel est le niveau de précision de l'algorithme ? Quels sont les indicateurs de performance ? Comment interpréter chacun de ces éléments ?
- **les dérives et incidents néfastes rencontrés.** Quels incidents néfastes ont-ils été rencontrés ? Quels ont été les éléments de mesure ? Quelles sont les instances de contrôles ? Qui a accès à ces informations au sein de l'organisation et à l'extérieur ? Existe-t-il une procédure d'alerte ? Une procédure d'arrêt de l'IA a-t-elle été mise en place ?
- **les pratiques.** Quelles pratiques liées à la transparence ont été mises en place ? L'IA est-il transparent by design ? L'organisation exerce-t-elle une concurrence loyale ? Fait-elle l'objet d'une certification ? Où sont présentées les pratiques restant à mettre en place ? Qui décide et élabore ces pratiques ?

Notons qu'en fonction des sensibilités des parties prenantes, certains désireraient transmettre une quantité circonscrite d'informations comme les données, le système et les modèles économiques tandis que d'autres parties prenantes préféreraient étendre au maximum cette transparence, jusqu'au document de conception voire au code source.

1.2. Dans quelles circonstances leur donner un accès ?

Ici s'opposent le droit de savoir et l'envie de savoir. Être transparent n'implique pas la totale divulgation des informations. En raison de la criticité de certaines des informations, il semble évident qu'elles ne peuvent pas toutes être en libre accès, notamment pour les secteurs à très hauts risques. De plus, si tout était disponible en continu, les systèmes pourraient être facilement détournés par des fraudeurs qui sauraient comment se présenter pour être traité d'une telle ou telle façon par le système. Mais dès lors se pose la question de quelles informations doivent être en libre accès ? Et pour quels secteurs d'activité ?

Doit-on avoir plus d'exigences en matière de transparence pour les IA à haut risque ? Comment faire agir le droit d'alerte ?

On peut également s'interroger sur la temporalité de la transparence. Quand doit-on informer ? Au début de la production ? À la fin du déploiement ? Sur quels sujets ?

D'autres questions se posent concernant les décisions individuelles : comment les individus peuvent faire un recours sur les décisions qui les concernent ? Mais encore avant, comment obtiennent-ils les informations sur ces recours ?

Une autre problématique concerne la propagation du savoir scientifique. Si les informations sont critiques et/ou relèvent du secret des affaires, comment transmettre le savoir ? Comment assurer la sécurité des informations ? En effet, les anonymisations peuvent être « désanonymisées » en croisant les sources et les données.

1.3. Comment les rendre transparentes et à qui ?

Une première étape de la transparence est de laisser un certain nombre d'éléments en libre accès – la question de la teneur de ces informations se pose évidemment – de façon à ce que tout intéressé puisse accéder facilement à un minimum de données. Il faut mettre facilement ces informations à disposition. Il faudrait éviter d'avoir à remplir un formulaire, de devoir contacter de nombreuses fois un service client téléphonique ou de fouiller un site internet entier pour trouver la page qui parle de ces informations.

Lors d'un recours individuel, on peut également se demander quelles informations supplémentaires communiquer au plaignant et comment les lui transmettre. Faut-il lui transmettre les journaux de logs* qui tracent toutes les opérations, les règles connues qui le concernent, les liens de cause à effet, la façon dont son profil a été classé ? Toutes ces informations peuvent-elles être compréhensibles pour un néophyte ? Sont-elles toutes utiles ? Enfin, il se peut qu'il y ait un certain nombre de règles inconnues (boîtes noires et/ou si l'algorithme est autoapprenant), difficilement interprétables ou explicables. Quand vaut-il mieux privilégier l'exactitude et l'exhaustivité des informations ou leurs simplicité et compréhensibilité ? Ici, la transparence est assujettie à l'explicabilité du système d'IA.

Les grandes lignes des réponses sembleraient évidentes : il faut répondre aux contraintes réglementaires et contractuelles, mettre en place un système de traçabilité (des modifications du système IA, des journalisations, des éléments probants, des décisions prises), travailler à diminuer la non-intelligibilité des algorithmes et adapter sa réponse en fonction du niveau des destinataires. Spécialistes ou non-initiés, co-acteurs, concurrents, utilisateurs, contrôleurs, régulateurs... n'auraient pas besoin du même niveau d'information. Mais dans ce cas, si l'interlocuteur présente un haut niveau technique et qu'il faut lui fournir un grand nombre d'informations, ne risque-t-on pas de dévoiler des secrets d'affaires ou la vie privée des gens ? Conserver l'ensemble des informations, même les plus insignifiantes, ne ferait-il pas porter une charge trop importante aux organisations ? Où est l'équilibre entre toutes ces contraintes ?

Le Parlement Européen suggère de mettre en place des méthodes spécifiques, notamment modulaires, de façon à pouvoir tester les systèmes d'IA sans divulguer des informations confidentielles ? Mais est-il possible d'avoir une compréhension du système avec une inspection parcellaire ?

Enfin, comment mettre en place une communication externe correcte qui puisse être reprise par les journalistes et les médias alternatifs sans être déformée, notamment pour contrer les fakes news ? En effet, comme l'avait mentionné la Cour des Comptes, malgré la transparence présente pour le système d'IA de Parcoursup, les journalistes ont allègrement sélectionné quelques informations pour les publier hors de leur contexte et les transformer en titres à sensations au détriment de la réalité de Parcoursup, qui est très différente.

2. Quelques recommandations : transparence by design, traçabilité et auditabilité, outils et bonnes pratiques

La transparence n'est pas du tout ou rien, elle peut évoluer en fonction du cycle de vie des systèmes d'IA. Elle devrait donc être mesurable. Ces mesures pourraient mener à des certifications et des labellisations. On pourrait même aboutir au fait qu'un système d'IA insuffisamment transparent et/ou explicable ne puisse pas être déployé dans des contextes à haut risque. La transparence est un préalable à la compréhension des algorithmes et de leur vérifiabilité.

2.1. La mise en place d'une transparence by Design (de bout en bout : conception, développement, exploitation...)

Il est nécessaire de prévoir la transparence dès la conception. Virginia Dignum considère comme impératif la mise au point de théories, de méthodes et d'algorithmes qui intègrent les valeurs sociétales, juridiques et morales à toutes les étapes de construction d'une IA, c'est-à-dire d'analyse, de conception et de développement (analyse, conception, construction, déploiement et évaluation). Ces outils doivent :

- traiter du raisonnement autonome de la machine sur les problèmes considérés comme ayant un impact sur l'éthique
- guider les choix de conception
- réguler les portées des systèmes d'IA
- assurer une bonne gestion des données
- aider les individus à déterminer leur propre implication.

Ces principes sont reproductibles pour la transparence. Il faudrait qu'elle soit prise en compte dès les premières étapes de création d'une IA. C'est ainsi l'ensemble des exigences concernant la transparence qui doivent être identifiées, évaluées et arbitrées dès la conception. Le choix des options qui seront retenues peut avoir un fort impact sur le niveau de confiance qu'auront les différentes parties prenantes sur les IA qui seront créer avec ces règles. Mais il faut s'assurer que cette transparence soit techniquement possible

et financièrement soutenable, d'où l'intérêt de les prendre en compte très tôt dans la création de l'IA. Les options retenues en matière de modèles, de techniques, de variables ou données d'apprentissage, d'engagements pris concernant les assertions relatives aux IA, de traitement des recours, des niveaux de responsabilités… et le niveau de transparence et d'explicabilité qui en résulteraient vont conditionner le niveau de confiance associé à ces IA. Des outils et des dispositifs devront être mis en œuvre pour faciliter cette transparence et l'explicabilité des traitements de l'IA.

2.2. La mise en place d'une traçabilité et d'une auditabilité des décisions, opérations, actions, incidents, réponses…

Ces solutions intermédiaires se basent principalement par la traçabilité du système à toutes ses étapes (depuis la conception jusqu'à l'utilisation) et/ou de confier la vérification de l'explicabilité à des organismes tiers, comme des auditeurs.

Ainsi le livre blanc de la Commission européenne sur l'IA énonce qu'il faudrait fournir des informations objectives, concises, facilement compréhensibles à l'utilisateur sur :

- le fait qu'il agit avec un système d'IA (sauf lorsque c'est évident)
- les capacités et limites du système d'IA
- l'objectif qu'il poursuit
- les conditions dans lesquelles il devrait fonctionner comme prévu
- le niveau de précision attendu dans la réalisation de l'objectif spécifié.

Cela pourrait notamment être fait sous la forme d'un compte-rendu périodique, qui informe régulièrement les intéressés sur différents éléments comme les détails sur l'atteinte des objectifs, la minimisation des risques, les effets indésirables rencontrés et leurs corrections… Cela permettrait la transparence dans le cadre d'une utilisation d'une IA.

Mais l'explication de ces éléments n'est pas suffisante pour permettre de reconstituer l'historique des actions ou décisions potentiellement problématiques prises par les systèmes d'IA et de les vérifier. Dans ce cas, toujours d'après le livre blanc de la Commission européenne, il faut conserver de nombreux autres éléments pendant une période limitée mais raisonnable :

- la documentation relative aux méthodes, procédures et techniques de programmation et d'entraînement utilisées pour concevoir, tester et valider les systèmes d'IA
- des archives précises concernant l'ensemble de données utilisées pour entraîner et tester les systèmes d'IA, y compris une description des principales caractéristiques et de la manière dont ces données ont été sélectionnées
- dans certains cas justifiés, les données elles-mêmes.

L'accès à ces éléments devrait être proportionné mais devrait être mis à disposition sur demande, afin notamment de permettre aux autorités compétentes de les tester ou de les inspecter. Le cas échéant, des dispositions devraient être prises pour garantir la protection des informations confidentielles, telles que celles relatives aux secrets des affaires.

Il faut également faire attention à la conception malintentionnée ou à la « volonté propre » des IA. Tom Murphy avait créé en 2013 une IA pour gagner le plus points au jeu de Tétris. En observant son programme, il avait découvert que l'IA mettait sur pause juste avant de perdre, évitant ainsi l'apparition de l'écran de Game Over. Cette anecdote trouve un écho après avoir lu Le Merrer et Trédan dans leur article « The bouncer problem » où les auteurs évoquent l'impossibilité d'être certain d'obtenir une explication de la part de l'IA. Tout comme l'IA de Tétris qui avait trouvé une solution innovante pour ne pas perdre, rien ne dit que les IA ne vont pas inventer des fausses explications pour donner la solution avec la meilleure précision sans transgresser les principes éthiques qui lui ont été inculqués. L'idée n'est pas saugrenue : les comportements humains de ce type sont nombreux. Le Merrer et Trédan donnent l'exemple du videur de boîte de nuit qui interdit l'entrée à quelqu'un et, plutôt que de donner la vraie raison (une origine ethnique non désirée par exemple), va dire que la personne ne peut pas rentrer car elle n'a pas de cravate.

Ne sous-estimons pas non plus la programmation intentionnelle de tels comportements. C'est ce qui s'est passé avec le Dieselgate où les algorithmes de réglage de la consommation d'essence du moteur changeaient leurs actions en fonction du contexte d'utilisation (normal ou tests). Il semble donc impérieux de mettre en place des outils qui permettent de tracer les différentes interventions effectuées sur les IA et permettre d'imputer des responsabilités en fonction des actions, de porter un recours en justice en ayant les informations qui permettraient de se défendre correctement, de faire contrôler les IA par des auditeurs indépendants avec des éléments probants disponibles...

2.3. La mise en œuvre d'outils, de techniques et de bonnes pratiques

Si une grande partie des solutions précédentes restent à développer et construire, des techniques existent déjà. L'une des premières est la mise en place d'une charte de la transparence par les organismes. Si ce document n'est évidemment pas la panacée, il présente le mérite de mettre en avant l'intérêt de l'organisme pour ce sujet et d'expliciter les objectifs à atteindre voire, dans son application, les méthodes à mettre en place.

Une autre technique qui existe déjà est l'analyse d'impact relative à la protection des données (AIPD). Bien que principalement orientés sur la protection des données, les AIPD renseignent déjà sur le traitement respectueux de la vie privée. La nature de l'AIPD peut être très bien déclinée pour la transparence d'un système d'IA. Elle concerne :

- la description détaillée des traitements mis en œuvre
- l'évaluation de la nécessité et de la proportionnalité concernant les principes et droits fondamentaux
- l'étude technique des risques sur la sécurité des données
- l'étude des impacts sur la vie privée.

D'une manière générale, l'ICO[33], dans son document « Explaining decisions made with artificial intelligence » du 1er mai 2020, souligne qu'il y a différents moyens de réaliser des explications : celles qui sont basées sur le process en décrivant l'IA tout au long de sa conception et son déploiement et celles basées sur les résultats qui expliquent ce qu'il s'est passé dans le cas d'une décision particulière.

Ce document décrit six façons différentes d'expliquer les décisions d'IA :

- **justification** : les raisons qui ont conduit à une décision, détaillées de manière accessible et non technique
- **responsabilité** : définition de qui est impliqué dans le développement, la gestion et la mise en œuvre d'un système d'IA, et qui contacter pour un examen humain d'une décision
- **données** : quelles données ont été utilisées dans une décision particulière et de quelle façon
- **équité** : mesures prises tout au long de la conception et de la mise en œuvre d'un système d'IA pour s'assurer que les décisions qu'il prend sont impartiales et justes, et s'assurer qu'un individu a été traité de manière équitable ou non
- **sécurité et des performances** : étapes suivies dans la conception et la mise en œuvre d'un système d'IA pour maximiser la précision, la fiabilité, la sécurité et la robustesse de ses décisions et de ses comportements
- **impact** : étapes franchies dans la conception et la mise en œuvre d'un système d'IA pour examiner et surveiller les impacts que l'utilisation d'un système d'IA et ses décisions ont, ou peuvent avoir, sur un individu et sur l'ensemble de la société.

Il est précisé dans ce document que chacune des explications fournies, quelle que soit la façon dont on désire la prodiguer, pourrait nécessiter deux versions : l'une pour le grand public et l'autre destinée aux spécialistes.

Ces six démarches ne sont pas forcément les seules : on pourrait, par exemple, plutôt opter pour une approche de justifiabilité par les résultats où on présente l'algorithme en faisant différentes simulations où on aura modifié les variables d'entrées (l'AI justifiable que nous avions présenté au § 6.1 du chapitre 5).

Il y a également différents autres moyens pour expliquer les décisions prises. Par exemple, dans l'explication d'un cas précis, on pourrait expliquer le cas en utilisant la technique du LIME, Local Interpretable Model-agnostic Explanations, qu'on pourrait traduire par « Explications locales interprétables indépendantes du modèle ». C'est une technique qui simplifie n'importe quel modèle d'apprentissage avec un modèle local et interprétable pour expliquer chaque prédiction individuelle, comme le synthétise le schéma ci-après :

[33] L'ICO est l'équivalent britannique de la CNIL

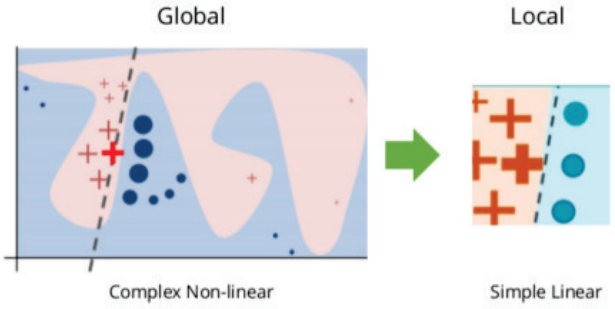

Figure 6 : Schéma de fonctionnement du LIME. Le schéma de gauche représente une IA de catégorisation des données en deux groupes bleues et rouges. Décrire comment est réalisé cette catégorisation dans son ensemble est difficile puisque non-linéaire. En revanche, il existe certains points comme celui autour de la croix rouge qui sont localement plus simples. Si l'on zoome (schéma de droite), la catégorisation est plus simple à expliquer tout en restant pertinente dans ces limites précises. (Source : c3.ai)

Il est important de savoir également où s'arrêter dans l'explicabilité. Dans le cas où rendre l'algorithme clair n'est pas possible, il faudra éviter d'utiliser des techniques d'explicabilité qui, à trop simplifier le système, pourrait le rendre faussement digne de confiance.

2.4. De la transparence aux biais

Sans transparence ni explicabilité, on ne pourrait pas mettre en place les autres exigences de l'IA comme celles d'équité. Mais il peut arriver qu'une plus grande transparence mène à un accroissement des biais humains et donc, paradoxalement, une baisse de l'équité. C'est ce qui a été découvert dans certaines études qui ont été menées auprès de professionnels et de spécialistes des IA. Lorsque des modèles complexes d'IA étaient présentés, accompagnés d'outils de visualisation pour faciliter la compréhension, les sujets étaient plus enclins à commettre des erreurs de jugement en faisant trop confiance aux résultats fournis par les IA. On pourrait donc s'interroger sur l'intérêt de la quête d'une transparence toujours croissante, si ce n'était oublier qu'il faut déjà limiter l'apparition d'autres biais, plus flagrants et plus faciles à éliminer que celui-ci, comme par exemple les biais méthodologiques que nous allons voir dans la § 1.2 du chapitre 7.

CHAPITRE 6

La transparence et l'explicabilité est le premier type d'exigence inhérent à un système d'information à base d'IA. En effet, les systèmes d'IA sont généralement difficiles à comprendre. Cela est dû au fait que les algorithmes d'IA sont très difficiles à appréhender, parfois même pour les spécialistes. Le fonctionnement de ces systèmes est donc obscur. Le rendre compréhensible à quelqu'un est compliqué car le niveau d'explication nécessaire peut varier en fonction des algorithmes mis en jeu et du niveau de familiarité du destinataire avec le sujet. Par ailleurs, le choix des éléments à rendre transparent n'est pas évident non plus, il faut faire la part entre ne pas fournir assez d'informations et donner trop d'éléments qui peuvent noyer le destinataire ou dévoiler des secrets d'affaires. Il faut également définir quand et pourquoi fournir ces éléments et à qui : toute curiosité n'est pas bonne à satisfaire mais les besoins légitimes doivent être pourvus.

Heureusement, la transparence peut être graduellement mise en place. Cela pourrait se réaliser via la création de chartes de transparence, l'intégration des exigences de transparence dès la conception de l'IA, la mise en place d'une traçabilité... De nombreux outils et bonnes pratiques restent encore à construire.

CHAPITRE 7

EXIGENCES D'ÉQUITÉ ET DE NON-DISCRIMINATION

Le sentiment d'injustice est très profondément ancré dans notre cerveau puisque les chercheurs ont démontré que certaines espèces animales l'éprouvaient également. L'égalité des citoyens devant la loi est une valeur particulièrement sensible en France, en raison de notre passé républicain. Il est donc naturel que les personnes espèrent être traitées équitablement, cela davantage encore quand les décisions sont prises par un système à base d'intelligence artificielle. Dans notre inconscient collectif, depuis Descartes, la machine est dénuée de sentiments, est incorruptible et serait donc plus fiable et juste qu'un humain. Pourtant, les exemples sont nombreux où les IA rendent des conseils, des prédictions, des jugements qui se sont révélés inéquitables. Obtenir des IA dignes de confiance sera impossible si nous ne pouvons pas avoir confiance dans l'équité des résultats. Il est crucial de faire en sorte que l'IA soit la plus équitable possible.

Le nombre de publications scientifiques lié à l'équité en machine learning a explosé ces dernières années, montrant que les chercheurs en IA sont très conscients de la problématique.

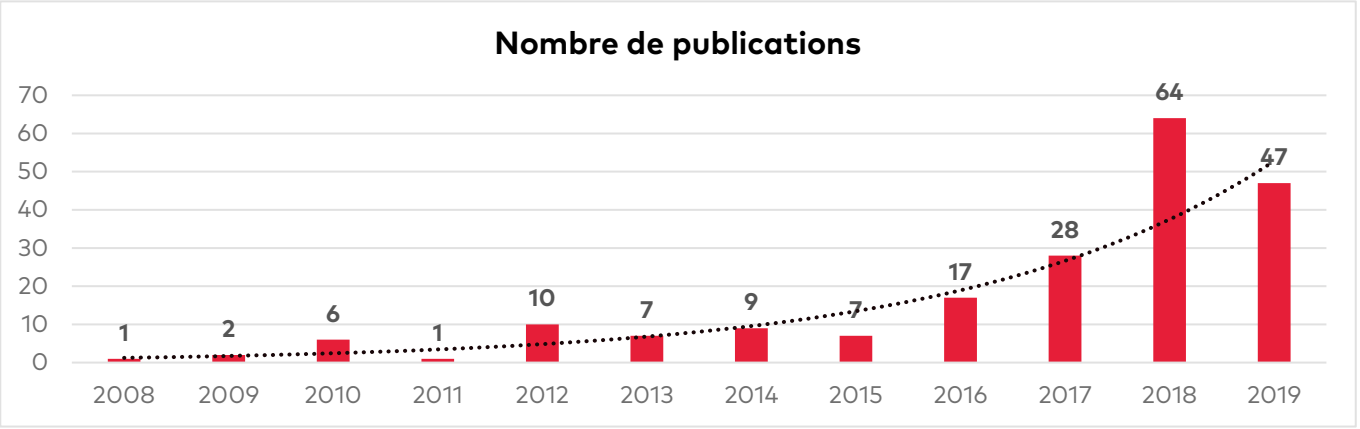

Figure 7 : Nombre d'articles scientifiques liés à l'équité en machine learning (Source : Caton et Haas, « Fairness in machine learning : a survey »)

Les grandes entreprises du secteur sont elles aussi très attentives au sujet. À titre d'exemple, on peut citer le document « Microsoft AI Principles » qui précise que « les systèmes d'IA devraient traiter toutes les personnes équitablement ». Pour autant, le domaine est complexe et difficile d'approche. Il existe une multitude de techniques visant à atténuer les préjugés et promouvoir l'équité à tous les stades d'un système d'IA : prétraitement, en cours de traitement et de post-traitement. Elles ne sont pourtant pas suffisantes notamment pour régler les contradictions entre les différents types d'équité.

CHAPITRE 7

1. Problématiques spécifiques à l'IA : types d'équité, nature des biais, sources des biais et des erreurs, maturité de l'apprentissage, importance relative, accentuation ou réduction des biais

1.1. Quels types d'équité ?

Ne léser personne est un beau concept mais qui est malheureusement souvent source de nombreux dilemmes : les représentations mathématiques des différentes variations de l'équité peuvent rentrer en conflit voire être incompatibles. Certains dilemmes sont plus courants que d'autres comme c'est le cas des cinq problématiques d'équité suivantes :

Équité entre individus

L'équité individuelle est souvent considérée comme l'équité idéale en IA. Le principe est de mesurer la similarité entre les individus et de le retranscrire par un score. Les personnes de score proche sont traitées de la même manière. Toutefois, il peut arriver que cela soit différent.

Prenons l'exemple d'un algorithme d'admission aux études supérieures fictif. Supposons que deux candidats viennent du même lycée, ont des notes équivalentes et visent la même filière. L'un des candidats vient d'une famille plus aisée que le second. Ce critère n'étant pas considéré comme pertinent par l'algorithme, les deux jeunes ont le même score de similarité à 0,1 point près. Toutefois, on peut vraisemblablement supposer que l'algorithme prend alors en compte d'autres données historiques (possibilité de voyager à l'étranger et d'apprendre des langues, d'avoir bénéficié de cours particuliers...) qui, elles, pourraient résulter de l'influence de la situation familiale et le résultat de l'algorithme est alors un peu différent : un score de traitement de 0,7 contre 0,9. Ces autres critères sont-ils cependant pertinents ou très accessoires ? Les utiliser ne pourrait-il pas conduire à des résultats inéquitables ? Ces légères différences, dans certains cas, pourraient avoir de graves conséquences. Il suffit d'imaginer que ces deux élèves soient le dernier de la liste d'attente et l'autre le premier recalé, et on comprendra mieux le dilemme que pourrait poser le principe de l'équité entre individus.

Équité individuelle et équité de groupe

Imaginons le cas d'un employeur qui s'aide d'un logiciel d'IA pour sélectionner les profils les plus pertinents qui seront invités à un entretien d'embauche. L'employeur, au courant des déboires d'Amazon pour leur logiciel équivalent, fait alors appliquer une technique de machine learning d'amélioration de l'équité. Ainsi, l'IA va faire en sorte que le nombre de candidats féminins acceptés par rapport au nombre total de candidats féminins soit le même que pour les candidats masculins. En raison de cette nouvelle version de l'IA, un candidat masculin aux compétences équivalentes à certaines candidates n'est pas invité. Dans cet exemple, une amélioration de l'équité de groupe cause des préjudices individuels. L'équité de groupe assure une certaine parité statistique pour les membres de groupes protégés particuliers mais elle peut nuire à l'équité individuelle.

Égalité de traitement et équité de traitement

L'égalité de traitement est l'équité sous-jacente à l'équité individuelle : deux individus au score proche doivent recevoir le même traitement. Toutefois la mise en place de cette logique nécessite parfois l'instauration d'œillères aux circonstances et à l'environnement. En effet, la loi interdit l'utilisation algorithmique de 25 types de données protégées comme l'origine ethnique ou confessionnelle de l'individu pour sept situations particulières comme l'accès à un emploi ou la rémunération. Par conséquent, pour être pertinents et conformes à la loi, les algorithmes doivent neutraliser ces variables pour ces situations c'est-à-dire qu'ils ne peuvent pas être pris en compte dans les choix effectués par eux. Mais les données d'entraînement présentent souvent des biais liés à ces données interdites et les utilisent pourtant pour exercer le programme. Cela introduit paradoxalement un biais dans le fonctionnement de l'algorithme. Sous prétexte d'égalité de traitement, les personnes défavorisées le restent. Ainsi, certains suggèrent de modifier la loi pour permettre de réaliser un biais positif et d'assurer une équité de traitement aux personnes défavorisées en prenant ouvertement en compte les données actuellement interdites par la loi pour y appliquer des corrections algorithmiques. La problématique est très proche de celle présentée ci-dessus concernant l'équité individuelle, la différence principale étant la légalité d'utilisation de certaines données ou non. Il convient de noter que les critères et les situations pour lesquels des limites sont imposées aux algorithmes ne sont pas les mêmes pour les différents pays. Cela peut poser un problème lorsque les modèles d'apprentissage proviennent de certains pays et sont utilisés dans d'autres.

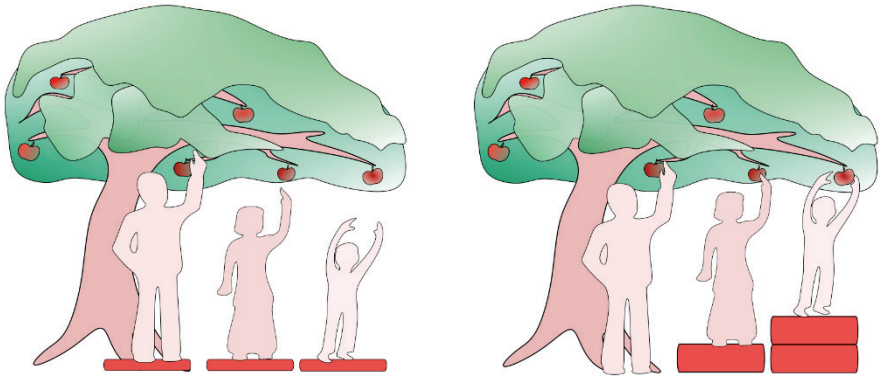

Figure 8 : Égalité et équité : l'égalité à gauche signifie que tout le monde est traité de la même manière. L'équité à droite signifie que chacun reçoit ce dont il a besoin pour atteindre le but.

Équité et performance

Enfin, citons un autre dilemme : celui de l'équité et de la performance. L'intérêt principal des concepteurs de systèmes d'IA est souvent porté sur la précision statistique des modèles. Le système qu'ils sont en train de créer est-il aussi précis que possible ? Mais l'ajout de l'exigence de l'équité implique l'ajout de compromis. Ces derniers viennent alors diminuer la précision du modèle. Par exemple, pour réduire le potentiel de discrimination, il est possible de modifier un modèle de risque de crédit afin que la proportion de prédictions positives entre les personnes ayant des caractéristiques protégées différentes (par exemple les hommes et les femmes) soit égalisée. Cela peut aider à éviter des résultats considérés discriminatoires (à tort ou à raison), mais cela peut également entraîner un nombre global d'erreurs statistiques plus élevé qu'il faudra également gérer.

Par ailleurs, les chercheurs remarquent souvent que plus un modèle est précis, plus il pourra être biaisé. Par exemple, suivre les biais de société permet à l'algorithme d'être plus performant : en matière de publicité ou d'offre d'emploi, coller aux stéréotypes peut permettre de maximiser le nombre de clics sur les annonces.

Notons également une autre problématique soulevée notamment dans le livre blanc de la Commission européenne dédié : l'éthique ne serait-elle pas variable en fonction du contexte d'application ? La définition d'un biais et de ce qu'est l'équité sera immanquablement différente selon que l'on discute d'un algorithme de conduite autonome, de mise au point d'un protocole de chimiothérapie ou de ciblage publicitaire.

Équité et choix éthique

Dans tous les cas présentés ci-devant, les choix sur ce qui serait équitable impliquent une appréciation éthique ou politique, de manière individuelle ou au niveau d'un groupe.

Il serait intéressant d'envisager la nécessité d'une étape supplémentaire lors de la conception d'une IA : celle des choix d'équité et d'éthique. Lorsque plusieurs options sont incompatibles, laquelle doit-on privilégier ? Ainsi, lorsqu'une voiture autonome prédit qu'elle va avoir un accident qui provoquera des décès ou des blessures et qu'elle a plusieurs options, tuer des piétons âgés sur le trottoir ou des personnes jeunes dans la voiture en face quelle option doit-elle privilégier ? C'est le classique « dilemme du tramway » formulé par Philippe Foot en 1967[34]. Ces choix éthiques dépendent souvent de la culture des personnes qui auraient à faire des choix, de leur situation personnelle... Il s'en suit nécessairement une discrimination, jugée normale par certains et pas par d'autres. Où se trouve l'équilibre ? Si 70 % des personnes dans un pays privilégiaient une option par rapport à une autre, l'algorithme doit-il est paramétré pour suivre cette option ? Si oui, de manière systématique ou en fonction de cette proportion ? Tout comme un juge cherche le bon niveau d'équité lorsqu'il aménage la peine du condamné en fonction du contexte, comment le contexte de l'IA influence-t-il le choix de cet équilibre entre les différents contextes et les problématiques éthiques associées ?

---

[34] Le dilemme du tramway est une expérience de pensée qui se conçoit ainsi sous une forme générale : une personne peut effectuer un geste qui bénéficiera à un groupe de personnes A, mais, ce faisant, nuira à une personne B ; dans ces circonstances, est-il moral pour la personne d'effectuer ce geste ?

1.2. Quelles natures de biais ?

Par définition, un biais est une démarche qui engendre des erreurs dans le résultat. Avec une définition aussi large, on comprend que les biais peuvent être de natures extrêmement variées. Quatre sortes de biais existent : les biais méthodologiques et techniques, les biais statistiques et probabilistes, les biais cognitifs et les biais inhérents aux processus d'apprentissage.

1.2.1. Les biais méthodologiques et techniques

Les biais méthodologiques et techniques dans le cas de l'IA résultent d'erreurs dans la méthode scientifique ou l'utilisation de certaines technologies. Ils peuvent provenir de différentes sources, certaines très techniques, par exemple, les limitations techniques des capteurs IoT* qui sont plus ou moins fiables, qui récoltent des données à une plus ou moins grande fréquence, avec plus ou moins de précision.

Les biais méthodologiques peuvent être liés au choix des variables. Par exemple, lorsqu'on n'a pas assez d'informations sur une variable (à cause, entre autres, de capteurs pas assez performants) ou lorsqu'on ne pense pas que la variable aura une grande incidence sur le résultat final, il peut arriver que les variables en question soient omises alors qu'elles pourraient avoir une incidence sur les résultats.

1.2.2. Les biais statistiques et probabilistes

À la croisée entre les biais méthodologiques et les biais cognitifs, on peut trouver les biais des statistiques et des probabilités. Par exemple, les études en neurosciences ont montré que le cerveau n'est pas naturellement câblé pour maîtriser intuitivement l'ensemble des techniques statistiques. Or, les calculs associés à l'IA se complexifient de plus en plus lorsqu'on cherche à en améliorer les performances, ce qui implique qu'il faut de plus en plus d'effort intellectuel pour comprendre les statistiques en jeu dans les algorithmes d'IA et ne pas se faire prendre au piège de ces biais.

On peut illustrer plusieurs problèmes qui découlent de l'utilisation des statistiques et des probabilités.

La présence d'erreurs dans les données impacte l'analyse statistique

Les analyses de l'existant s'appuient souvent sur des articles, des enquêtes, des sondages, des études, des publications qui s'appuient sur des analyses statistiques et souvent sur leur traduction en termes de probabilité. Mais les données qui sont alors utilisées correspondent souvent à la somme de perceptions individuelles sur tel ou tel sujet, perceptions qui sont bien réelles mais qui ne sont pas pour autant la traduction de la réalité. Ces perceptions proviennent de sources variées (y compris la restitution des articles et publications notés ci-dessus) qu'ils n'ont pas le plus souvent vérifiées mais aussi de leurs propres biais. Elles peuvent aussi s'appuyer sur des recherches basées sur des corrélations et non sur des causalités prouvées scientifiquement comme cela a souvent pu être vu dans le cas du Covid19. Les IA vont donc utiliser toutes ces données dans leurs algorithmes d'apprentissage qui de ce fait pourraient conduire à des modèles d'apprentissage erronés.

CHAPITRE 7

Les techniques statistiques influent sur le niveau de confiance

La taille et la source d'un échantillon, la nature de l'extrapolation sur un groupe plus large et autres techniques auront tous un impact sur le niveau de confiance sur les résultats. Si on prend l'exemple de l'élection présidentielle, la taille de l'échantillon d'une population donnée permet un certain niveau de confiance quant aux résultats de ce sondage. Lorsque l'on interroge 700 personnes dans un sondage avec un candidat à 20 % avec un niveau de confiance de 95 %, en réalité, cela traduit le fait que si ce sondage était réalisé une deuxième fois selon les mêmes modalités au même moment, les résultats de 95 % de ces sondages se situeraient dans une fourchette allant de 17 % à 20 %[35]. En réalité, le fait qu'un des candidats soit à 18,5 %, un autre à 19 % et un autre à 19,5 % ne signifie pas grand-chose statistiquement puisqu'ils se situent tous dans la marge d'erreur.

Un autre sondage effectué dans les mêmes conditions pourrait donner tout autant un ordre inversé sans que cela ne traduise une situation réelle différente. Cette marge d'erreur peut fluctuer de manière importante en fonction du niveau de confiance souhaité et la taille de l'échantillon. Ainsi, le gain de précision peut avoir un coût très significatif. Par ailleurs, l'extrapolation est effectuée sur la population susceptible de voter. Comment identifier cette population ? En incluant une question sur leur niveau de certitude d'aller voter (de 1 à 10 par exemple). Certains incluront ceux qui ont répondu 10, et d'autres ceux qui ont répondu 8, 9 ou 10. Ceux qui ont répondu 10 pourrait représenter 50 % de la population alors que ceux qui ont répondu 8, 9 ou 10 représenter 70 % de la population.

Ces techniques différentes conduiront à des résultats différents puisque le profil des personnes qui sont convaincues d'aller voter (plus militants) n'est pas les mêmes que ceux qui sont un peu moins convaincus (moins militants). Il y aura donc plusieurs résultats qui peuvent être assez différents pour une même situation. On pourrait poursuivre ce raisonnement avec l'utilisation d'autres techniques statistiques (les types de quotas, de redressements, ...) qui auront tout autant un impact sur le niveau de confiance. Cette problématique s'applique aux IA dans la mesure où elles utilisent aussi souvent ces techniques statistiques.

La profondeur d'analyse peut modifier entièrement les conclusions

Les IA utilisent plusieurs variables. Le choix de la nature et du nombre de variables peut avoir un impact important sur la nature des conclusions voire les inverser. On peut l'illustrer par l'exemple suivant.

Dans une école d'ingénieur, il y a autant de candidatures hommes que de candidatures femmes. Sur 100 candidats hommes, 50 % sont acceptés et sur 100 candidats femmes 40 % sont acceptés. Cela pourrait amener à penser qu'il y ait un biais en faveur des hommes dans cette école.

Dans une école d'ingénieur, il y a autant de candidatures hommes que de candidatures femme. Il y a deux départements : biologie et électronique. Dans le département de biologie, sur 100 candidats hommes, il y

---

[35] Précisions sur l'intervalle de confiance issues du « Sondage d'intentions de vote pour l'élection présidentielle de 2022 » Ipsos-Sopra-Steria de Septembre 2021
https://www.ipsos.com/sites/default/files/ct/news/documents/2021-10/Ipsos%20-%20Le%20Parisien-France%20Info%20-%20IV%202022%20-%20Oct.%202021.pdf

en a 20 qui sont acceptés et sur 100 candidats femmes, il y en a 30 qui sont acceptés. Dans le département électronique, sur 100 candidats hommes, il y en a 60 qui sont acceptés et sur 100 candidats femme, il y en a 70 qui sont acceptés. Pour les 2 départements, il y a donc une proportion de femmes plus élevée qui est acceptée. On pourrait en déduire qu'il y a un biais de sélection favorable aux femmes dans cette école.

En réalité, cela peut être la même école d'ingénieur et la même situation. Il s'agit d'une présentation alternative de la même situation. (cf.Tableau 1).

	Dpt. biologie		Dpt électronique		Total	
	Candidatures		Candidatures		Candidatures	
	Initiales	Acceptées	Initiales	Acceptées	Initiales	Acceptées
Hommes	250	50 (20%)	750	450 (60 %)	1000	500 (50%)
Femmes	750	225 (30 %)	250	175 (70%)	1000	400 (40%)
Total	1000	275	1000	625	2000	900

Tableau 1 : Exemple de candidatures par rapport au genre

On peut ainsi analyser les probabilités d'acceptation des candidats. En fonction de la profondeur de l'analyse, les biais genrés semblent se transformer. Sur le total des candidatures, on voit apparaître un biais en faveur des hommes mais si on analyse par département, le biais semble être en faveur des femmes. Ainsi, quelle est la réalité ? Y'a-t-il un biais ? Pas de biais ? La situation est-elle équitable ?

On pourrait ajouter plusieurs variables supplémentaires d'analyse à l'exemple ci-dessus que la conclusion basculerait chaque fois dans un sens comme dans l'autre. Ce n'est que la nature et le type d'analyse et des variables considérées qui conduira à conclure de telle ou telle manière parfois dans un sens totalement opposé. C'est ce type d'analyse qui est effectué par les IA et il convient donc d'en tenir compte dans l'appréciation du niveau de confiance que l'on peut en tirer.

La capacité à comprendre et interpréter le résultat statistique

Nous pouvons l'illustrer comme suit. Deux vaccins, l'un efficace à 95 % et l'autre à 75 %, donnent l'impression de manière intuitive qu'ils ont tous les deux un bon niveau d'efficacité et que le niveau de risque d'attraper la maladie est assez proche. En réalité, le risque est cinq fois supérieur dans le second cas que dans le premier. En effet, un vaccin efficace à 95 % (c'est-à-dire que pour 100 malades non vaccinés, il y aura cinq malades vaccinés) est environ cinq fois moins risqué qu'un vaccin efficace à 75 % (c'est-à-dire que pour 100 malades non vaccinés, il y aura 25 malades vaccinés, soit cinq fois plus). Il y aurait donc cinq fois plus de malades, de personnes hospitalisées ou de personnes décédées entre ces deux populations vaccinées, ce qui n'est pas négligeable comme écart. Les IA s'appuient beaucoup sur les analyses statistiques et probabilistes. Leur utilité dépend donc beaucoup de la capacité à apprécier les résultats qui sont traduits en statistiques et en probabilités.

1.2.3. Les biais cognitifs

Les biais cognitifs sont extrêmement nombreux et sont inhérents au mode de fonctionnement du cerveau. Suivant la thèse de Daniel Kahneman[36], le cerveau présente deux systèmes de pensée : le « système 1 » fonctionne de manière automatique et intuitive en demandant peu d'effort. Le « système 2 » plus lent et gourmand en énergie est celui qu'on mobilise dans la résolution de tâches complexes comme reconnaître une personne, compter le nombre d'occurrences de lettres dans un texte[37], se préparer à une course de sprint, réaliser des opérations mathématiques complexes... Le premier système est utilisé par défaut et permet de réagir rapidement dans notre environnement en simplifiant ce dernier. Cet impératif de simplification est à l'origine des biais, mais seuls les biais permettent cette simplification. Les biais cognitifs sont de nombreux types : perception, attention, mémoire, jugement, raisonnement, de personnalité...

Ils permettent de régler quatre problèmes principaux :

- la surcharge d'information
- le manque de sens
- le besoin d'agir vite
- le besoin de mémoriser le minimum de choses efficacement.

L'utilisation de ces biais nous permet tout simplement de (sur)vivre à peu près sainement. Par exemple, parmi les 5 000 messages de publicité en moyenne auxquels les individus sont confrontés quotidiennement, lesquels doivent-ils retenir car ils pourraient s'avérer utiles ? Mais ce mode de fonctionnement à des revers : certaines informations écartées sont en fait utiles et importantes, le cerveau donne un sens à ce qui n'en a pas (comme les illusions d'optique), les décisions rapides peuvent être déplorables ou désastreuses et notre mémoire renforce les erreurs initiales.

---

[36] Daniel Kahneman, « Système 1, Système 2 Les deux vitesses de la pensée », Flammarion, 2011

[37] Utilisés entre autres pour des exercices académiques ou dans des contextes de cryptographie

Les biais utilisés pour résoudre ces problèmes sont très nombreux comme représentés ci-dessous :

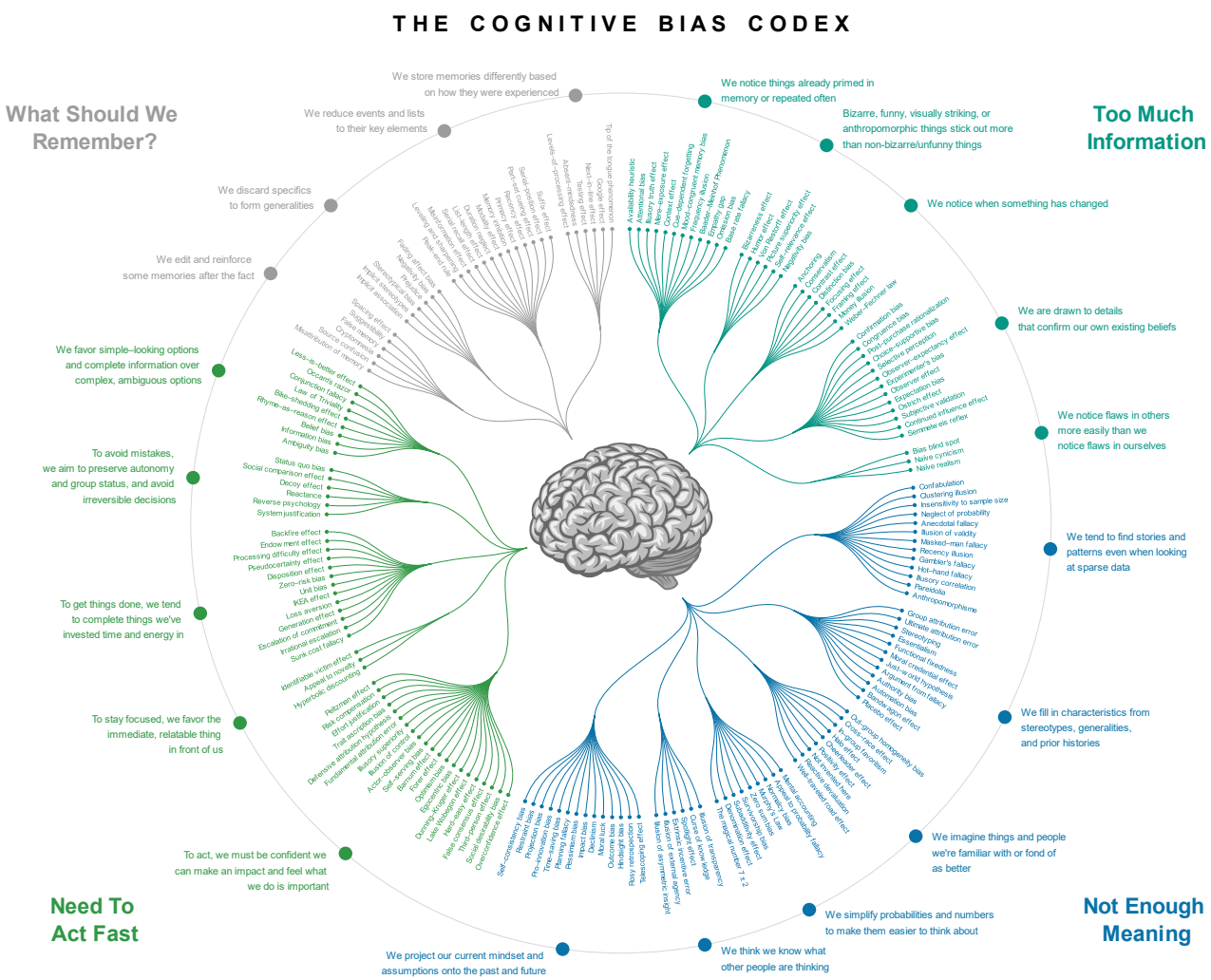

Figure 9 : Codex des biais cognitifs (catégorisation par Buster Benson, design par John Manoogian III) (https ://fr.wikipedia.org/wiki/Fichier:The_Cognitive_Bias_Codex_(French)_-_John_Manoogian_III_(jm3).svg)

A priori, certains de ces biais n'ont pas d'influence sur l'IA mais nombreux sont ceux qui peuvent avoir un effet extrêmement important. En voici quelques exemples :

- **la corrélation illusoire** : ce biais fait percevoir une corrélation (un lien) entre deux évènements qui n'en ont pas ou dont la relation est beaucoup plus faible en réalité. L'exemple le plus connu est que la pleine Lune exercerait une influence sur le nombre des naissances, ce qui a bien entendu été démontré comme faux par les scientifiques[38]
- **l'effet cigogne** : c'est la tendance humaine à croire que si deux évènements ont une corrélation entre eux, alors l'un est la cause de l'autre. L'effet cigogne correspond à une corrélation illusoire maximale. Mathématiquement, le lien entre corrélation et causalité est évident : deux évènements A et B sont liés entre eux par une variable r. Si r vaut 0, A et B sont indépendants, il n'y a aucun rapport entre eux. Si r vaut 1, A est la cause de B. Entre les deux, il y a juste un lien plus ou moins fort. Le nom vient du fait qu'en Alsace, on remarque que les villes qui ont le plus de cigognes sur les toits sont celles qui ont le plus de naissances. Cela pourrait être la preuve que les cigognes apportent les bébés ! Ou plus simplement que les grandes villes qui ont plus de population et de couples en âge d'être parents présentent aussi plus de toits où les cigognes peuvent nicher.

  Cet effet est extrêmement important avec l'IA qui permet de détecter des signaux faibles ou des informations cachées derrière les données qui ne sont pas perceptibles par les humains. C'est un champ de recherche qui se développe : faire de nouvelles découvertes en laissant une IA analyser seule les bases de données et remonter les corrélations fortes. Une de ces IA a remonté un résultat étrange[39] : on découvre une corrélation de r=0,99789 entre le nombre de suicides par pendaison aux USA et les dépenses américaines dans la science, l'espace et la technologie tel que montré sur le schéma Figure 10.

---

[38] Il existe un grand nombre d'études scientifiques prouvant ce fait (Oliver Kuss; Anja Kuehn (2008). Lunar cycle and the number of births: A spectral analysis of 4,071,669 births from South-Western Germany / Wei WWS (2006) Time series analysis: univariate and multivariate methods/Laurent Toulemon (1986) Nouvelles données sur les variations du nombre des naissances selon les rythmes lunaires et circadiens...) La chaine Hygiène mentale en fait une très bonne vulgarisation zététique et scientifique du point de vue bayésien (Ep 23 et 24 https://www.youtube.com/watch?v=PRtwo1j0y2I&ab_channel=Hygi%C3%A8neMentale) et qui explique aussi pourquoi on trouve des études donnant un résultat positif.

[39] http://tylervigen.com/spurious-correlations

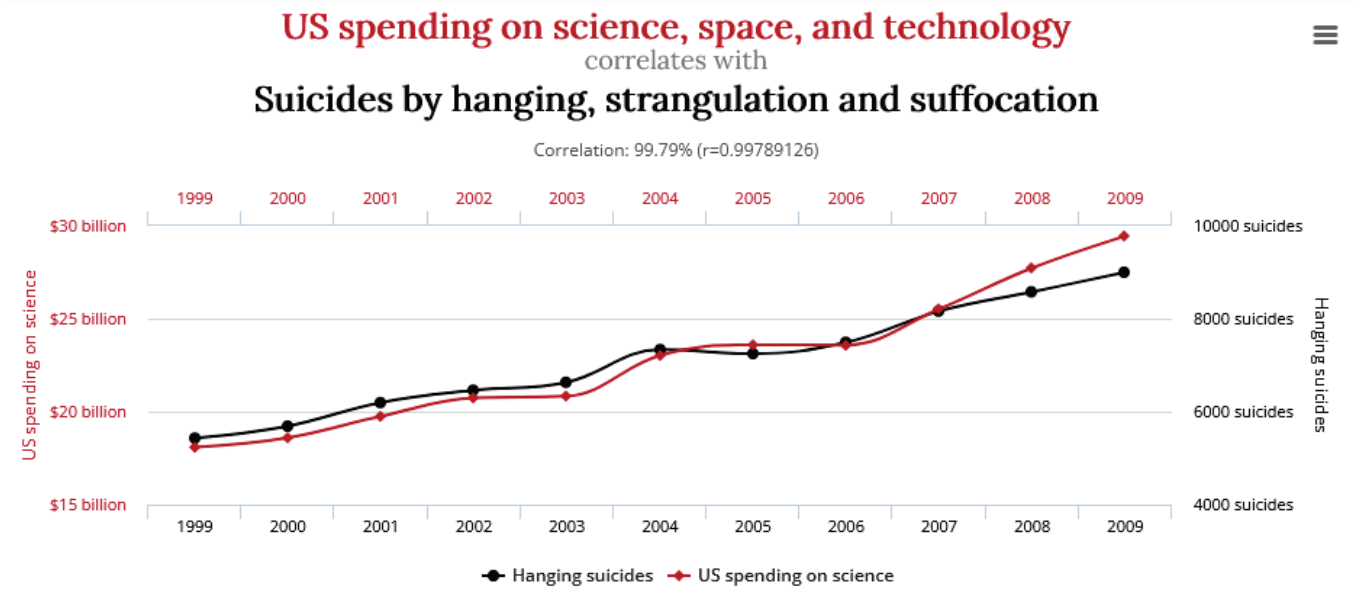

Figure 10 : « Les dépenses américaines dans les sciences, l'espace et la technologie a une corrélation avec les suicides par pendaison, strangulation et suffocation »[35] - Sources des données : U.S. Office of Management and Budget and Centers for Disease Control & Prevention.

Or, quand r est égal à 1 ou extrêmement proche comme ici, cela veut dire qu'il y a une relation de cause à effet ! Doit-on conclure qu'il faut immédiatement diminuer les dépenses américaines dans la science et ses applications pour sauver des vies humaines ? Ou au contraire les augmenter pour éduquer les gens à la pensée critique afin qu'ils comprennent par eux-mêmes qu'il ne s'agit là que d'une bête coïncidence ?

Dans un autre cas, on a découvert que le nombre de prix Nobel d'un pays augmente avec sa consommation de chocolat. Plutôt que d'instaurer l'obligation de manger du chocolat pour tous pour obtenir le plus de prix Nobel, il vaudrait mieux se rappeler que les pays qui ont beaucoup de prix Nobel ont un meilleur niveau de vie et donc plus de budget pour acheter des produits alimentaires non essentiels. Ici, c'était un troisième facteur C, hors A et B, qui n'était pas connu ou pas conscientisé qui était la cause commune des deux facteurs.

Ces constatations soulèvent de nombreuses questions pour l'IA :

- comment distinguer un signal faible d'un évènement uniquement dû au hasard, comme dans l'exemple des dépenses américaines dans la science ?
- comment différentier les signaux faibles utiles et non utiles ? Comme pour les cigognes, on a bien trouvé un signal de la présence des cigognes sur la démographie mais ce n'est pas très utile pour maîtriser la démographie (sinon le Japon importerait des cigognes)
- comment différentier les signaux faibles inutiles de ceux qui sont rares (et donc qui sont peu représentés dans les données) mais qui ont un grand impact ? Ces cas sont très courants sur les marchés financiers : chaque crise financière est une surprise pour ces acteurs mais rétrospectivement, on découvre de nombreux indices qui auraient pu la faire prévoir, à défaut de l'expliquer. Comment les repérer avant la crise ?
- comment distinguer la part de la corrélation dans la cause ? Si une personne a un cancer du poumon mais qu'elle avait manipulé beaucoup d'amiante avant son interdiction et qu'elle fume, quelle responsabilité chacun des facteurs de risque a-t-il eue dans l'apparition de sa maladie ? Cela présente une importance dans les modèles : si une cause A influence le résultat à 80 % et la cause B à 20 %, il faudra davantage pondérer la cause B pour qu'ils obtiennent chacun le niveau en relation avec la réalité de la causalité.

- **biais de représentativité** : on a observé que plus une chose est typique d'une catégorie, plus les individus la classent dans celle-ci. Ce mode de fonctionnement permet de répondre rapidement à la question « quelle est la probabilité qu'un élément A appartienne à la catégorie B ? ». Il renforce souvent les stéréotypes car il y a une surreprésentation ou sous-représentation de certaines catégories de personnes dans l'échantillon de la population interrogée. Ainsi, si on questionne les gens aux abords d'une discothèque le samedi soir, les jeunes seront surreprésentés par rapport aux retraités. On ne pourra donc pas extrapoler leurs réponses sur l'ensemble de la population
- **biais de disponibilité** : il est plus facile de se souvenir d'un phénomène lorsque celui-ci est récurrent. Donc le biais est de juger un élément plus fréquent s'il est plus simple de se souvenir d'occurrences de cet évènement. Ainsi, on peut croire qu'il est plus fréquent de mourir assassiné que d'un cancer, ce qui est exactement l'inverse dans les faits mais les récits de meurtres se trouvent plus facilement dans les médias que les reportages dans le département oncologique. De même, est-il plus courant d'avoir un mot qui commence par la lettre a ou un mot dont la quatrième lettre est -a[40] ? Comme il est plus facile de se souvenir des mots qui commencent par a, les individus répondent par la première solution alors que c'est faux

[40] Pour l'anecdote, le site listedemots.net recense 31 175 mots débutant par -a et 43 358 mots avec -a en 4ème position

- **biais de confirmation** : nous sommes plus attentifs aux éléments qui confirment nos croyances et réfutons ou ignorons ceux qui les contredisent. Un concepteur peut donc ignorer ou surpondérer une variable du modèle qu'il est en train de construire en fonction de ses croyances propres. Il peut également présenter les résultats suivant la manière qui correspond à ses croyances
- **biais de statu quo** : Toute nouveauté est perçue comme engendrant plus de risques que d'avantages. Il est particulièrement présent dans toute décision difficile et complexe car le risque de faire le mauvais choix incite à la prudence. Cela peut inciter les concepteurs à attendre l'arrivée de nouvelles données avant de modifier leur IA
- **biais focal** : Les seules informations qui sont directement disponibles nous font oublier d'envisager le reste des cas possibles à la suite d'une information manquante
- **biais de mémorisation émotionnelle** : cela correspond au fait que nous nous souvenons mieux de faits qui ont provoqué de fortes émotions. C'est pour ça que quasiment tous ceux qui ont vécu le 11 septembre 2001 se souviennent de ce qu'ils faisaient lorsqu'ils ont appris la nouvelle des attentats ou qu'on se souvient facilement de certaines images-chocs, comme celle d'Aylan Kurdi, migrant syrien de 3 ans, décédé sur une plage grecque
- **biais culturel** : C'est la tendance des individus à analyser, interpréter et juger les choses à travers le filtre de ses propres références culturelles. Ainsi, les pays asiatiques tendent à faire disparaître le svastika, symbole local et très répandu du bien-être, car les Occidentaux en voyage protestent en confondant ce symbole avec la croix gammée nazie
- **biais du favoritisme culturel** : portant différents noms, ce biais consiste à favoriser les membres de notre propre groupe social (même origine ethnique, même formation, même club, même sexe...) au détriment des autres.

Nous avons vu quelques exemples de biais cognitifs. Plus d'une centaine ont déjà été identifiés. La recherche fondamentale en neuroscience et en psychologie devrait permettre petit à petit de mieux comprendre ces biais et leur fonctionnement. Plus nos connaissances dans ce domaine seront étendues, plus il sera facile de trouver des moyens et des techniques pour les empêcher de se manifester dans des contextes où ils n'ont pas lieu d'être.

1.2.4.Les biais inhérents aux processus d'apprentissage

Il existe plusieurs dizaines d'autres biais. Découvrir leur existence peut être choquant et l'on doit se poser la question de comment les maîtriser à défaut de pouvoir les neutraliser. Mais en aucun cas, il ne serait possible d'empêcher leur apparition. En effet, les biais sont inévitables, ils sont issus du processus de classement (humain) qui nous permet de catégoriser notre environnement en éléments explicables et compréhensibles. La catégorisation est indispensable pour envisager les choses complexes et les nouveaux phénomènes. Ce mécanisme simplifie notre vision du monde mais la limite également. Elle est également discriminatoire dans le sens où on classe les choses et les gens selon des critères qui discriminent, le mot « discriminer » étant à comprendre dans son sens originel de réaliser une séparation entre les éléments.

Or, le classement utilisé en IA est basé sur le mécanisme de la catégorisation humaine mais en plus, la classification est le principe de base des IA. Tout classement issu d'une IA, donc toute IA, est potentiellement discriminatoire. Sachant cela : si une IA classe un groupe dans une certaine catégorie – par exemple, le logiciel américain COMPAS prédisait que la population afro-américaine présente plus de risques de récidive. Est-ce une discrimination « rationnelle » basée sur les faits, une discrimination « irrationnelle » issue de biais ou un mélange des deux ?

Plusieurs types de traitement possible y compris l'utilisation de contre-biais

Les biais cognitifs peuvent donc causer de graves erreurs d'équité en intelligence artificielle. Il faut donc mettre en œuvre des techniques d'atténuation des biais cognitifs, soit en réussissant à les prévenir, soit en réduisant leurs effets négatifs.

Quelques outils qui existent déjà pourraient être appliqués dans le contexte de l'IA. Des biais volontaires, souvent instaurés par des choix politiques et légaux, ont déjà été mis en place à de nombreuses reprises. C'est le cas notamment des quotas positifs ou du nudging. Le nudge est un concept popularisé par Cass Sunstein et Richard Thaler ; ce dernier ayant reçu un prix Nobel d'économie pour sa compréhension de la psychologie en économie. Le nudge consiste à influencer les décisions des groupes et des individus en leur faisant des suggestions indirectes. Ces suggestions indirectes se basent souvent sur l'exploitation des biais cognitifs des individus. Par exemple, l'augmentation de nombre de donneurs d'organes en France est basée sur l'application d'un nudge. Désormais chaque individu français est donneur par défaut, ceux qui désirent refuser ce don – pour être en accord avec leurs convictions religieuses par exemple – doivent faire la démarche de se désinscrire de la liste de donneurs. Sauf que ce schéma de pensée est celui du biais de statu quo.

Certes, ces techniques ont obtenu de bons résultats et il pourrait donc être envisagé de les réappliquer aux systèmes d'information à base d'intelligence artificielle mais on peut s'interroger sur son éthique : il s'agit d'exploiter nos faiblesses cognitives sans notre consentement. Ainsi, toutes les techniques d'atténuation des biais dans le but d'une amélioration de l'éthique des IA ne sont-elles pas elles-mêmes non éthiques ?

Heureusement, il existe d'autres techniques de lutte contre les biais comme la formation des personnes ou les incitations qui peuvent être calibrées pour modifier les préférences vers un comportement dit plus éthique. Cela peut être par exemple la baisse de taxe sur les aliments sains avec une augmentation de taxe pour les produits délétères.

Tous ces biais sont des sources d'erreurs. La probabilité que l'IA donne des résultats faux et/ou inéquitables est proportionnelle à la quantité de biais présents dans les systèmes d'information à base d'intelligence artificielle. Il est donc important d'avoir conscience de l'existence de ces biais, de découvrir leur nature et leurs sources afin de pouvoir mettre en place des actions pour diminuer voire supprimer quand c'est possible l'existence de ces biais.

1.3. Quelles sources de biais ?

Nous avons vu que la présence de biais est normale chez les humains, qu'elle est inhérente à notre santé mentale, notre apprentissage et notre survie. Ces biais peuvent se retrouver dans un système à base d'intelligence artificielle à différents niveaux, soit parce qu'ils affectent leurs concepteurs, soit parce que l'âme de l'IA est de reproduire une intelligence humaine et qu'elle peut malheureusement en reprendre certains travers de pensées. Plus précisément, les biais peuvent provenir de différentes sources.

- les données d'apprentissage :
  - **biais de récolte des données** : les concepteurs d'IA utilisent pour l'entraînement de leurs algorithmes des bases de données privées ou publiques, qui peuvent être publiées par des chercheurs, des groupements, des entreprises... ImageNet, la base de données développée à Princeton, a donné accès à 14 millions d'images annotées dans 20000 catégories. Il est permis de les utiliser pour entraîner des algorithmes de reconnaissance visuelle. Néanmoins, tout biais dans une base de données de cette taille peut créer des biais dans de nombreux systèmes à base d'intelligence artificielle
  - **biais de l'étiquetage des données** : l'étiquetage manuel des données est long donc coûteux. Pour diminuer ces coûts, une grande partie de l'industrie est délocalisée dans des pays à bas salaires, dont les repères culturels peuvent être différents. Ce biais est aussi appelé biais du favoritisme culturel, car in fine l'IA produite sera plus adaptée aux utilisateurs provenant de la même origine culturelle que celle du groupe de concepteurs. On en déduit donc qu'un moyen de limiter l'existence de ce biais est de limiter les équipes de développement aux profils homogènes
  - **déséquilibre ou biais des données passées** : Les IA extraient leurs enseignements des données passées. Or, ces données peuvent être déséquilibrées et/ou refléter une discrimination, ils peuvent produire des résultats qui ont des effets discriminatoires sur les personnes en fonction de leur sexe, race, âge, santé, religion, handicap, orientation sexuelle ou d'autres caractéristiques. Un cas célèbre est celui d'un assureur automobile américain dont l'IA qui devait calculer la cotisation demandait systématiquement un tarif plus élevé pour les assurés de couleur que pour les assurés WASP. Ce résultat peut être tout à fait logique : les afro-américains ayant en moyenne des revenus inférieurs aux Blancs, leurs voitures sont plus anciennes et/ou moins bien entretenues, tombent donc en panne plus souvent et coûtent plus cher à l'assureur. La couleur de la peau n'a en réalité aucun rapport avec ces différences. La tendance existe donc statistiquement. On pourrait maintenant se demander si l'existence d'une tendance statistique est une raison pour augmenter la cotisation des nouveaux assurés de couleur, ne deviendrait-elle pas un biais par une application aveugle d'une observation générale (elle-même possiblement biaisée) en une condition algorithmique systématique. Il n'y aurait aucun espoir de faire évoluer les discriminations actuelles si nous construisons sans discernement des IA sur l'observation de phénomènes passés

  - **non-représentativité des exemples lors de la phase d'entraînement** : c'est le risque de biais de spectre, qui fait référence à la présence d'exemples dans l'ensemble de données qui ne reflètent pas la diversité et la complexité des situations, c'est-à-dire que la variété des données utilisées dans la phase d'entraînement de l'IA ne reflète pas l'ensemble des possibilités réelles. Cela implique que de bonnes performances sur des exemples évidents ne sont pas suffisantes pour évaluer la capacité du modèle à gérer correctement des situations plus ambiguës. Un exemple aisé est un algorithme d'aide au diagnostic d'une maladie. Si une cause de la maladie est inconnue à l'époque de la création de l'IA, cette dernière présente un biais de spectre[41].
- **les données d'incitations pour les techniques d'apprentissage par renforcement** : de plus en plus d'algorithmes d'apprentissage présentent une boucle de rétroaction destinée à rendre plus efficace l'apprentissage en favorisant certains chemins d'apprentissage susceptibles d'être considérés comme plus pertinents et en pénalisant d'autres chemins moins pertinents. Ainsi, une incitation erronée faussera l'apprentissage
- **les modèles** : le principe de l'IA est d'observer des phénomènes passés, y reconnaître un schéma caractéristique, le modèle, afin de l'appliquer à un nouvel élément. Si l'hypothèse se base sur des éléments biaisés, alors l'IA le sera aussi
- **les biais de restitutions** : la manière de restituer les résultats peut être biaisée. Ainsi, une restitution à connotation positive d'une situation donnée (il y a 70 % de chances de survie si vous prenez tel type de traitement) par rapport à une présentation à connotation négative de la même situation (il y a 30 % de probabilité de mourir) sera suivie d'effet par un nombre plus grand de patients choisissant le traitement proposé. Dans ce cas, comme dans de nombreux autres, comment l'IA doit-il restituer ces informations ? Il s'agit bien de choix éthique. Qui doit faire ce choix ?
- **les choix techniques** : ces choix viennent souvent d'une mauvaise utilisation statistique ou d'une simplification par le concepteur de l'IA. Ces biais techniques réduisent la performance de l'algorithme et entravent la réalisation de son objectif. Les biais de confirmation (ou variable omise) et de sélection, entre autres, font partie de ces sources. Ces causes peuvent aboutir à de véritables non-sens. Par exemple, une IA mal entraînée à l'identification de panneaux routiers pourrait avoir compris qu'il faut qu'elle donne une réponse positive si la photo provient du dossier C://Images/Photos_panneau et une réponse négative dès que la photo provient d'un autre dossier. Dans ce cas d'école, le biais est purement technique.

---

[41] À l'inverse, et c'est ce qui cause en partie l'engouement pour les systèmes d'information à base d'intelligence artificielle, on pourrait découvrir de nouvelles causes aux maladies dans certains cas. En étudiant tous les patients qui ont le même score, on pourrait découvrir des points communs et potentiellement déterminer l'origine de la maladie. Ce principe peut s'appliquer à d'autres cas que la médecine.

1.4. D'autres erreurs que les biais existent et comment les traiter ?

Bien que le sujet du biais soit très présent dans le sujet de l'éthique, ce n'est pas la seule source de problème pour une IA. Il existe des erreurs qui ne sont pas issues de biais comme :

- **les erreurs liées aux techniques d'IA** : l'objectif de l'apprentissage d'une IA est de bien trier et d'associer certaines caractéristiques, d'en éliminer d'autres, de les contextualiser... Cet apprentissage se fait souvent sur la base d'essais et d'erreurs. Il y a donc beaucoup d'erreurs en début d'apprentissage avant de trouver la bonne « recette ». Par exemple, dans une IA d'identification des sujets de photos, les erreurs liées aux techniques d'apprentissage pourraient faire que l'algorithme dans sa phase initiale d'apprentissage confonde un humain et un singe du fait de plusieurs caractéristiques potentiellement concordantes comme la forme générale du visage, la présence d'un œil, d'un nez, d'une bouche, etc. Le processus d'apprentissage permettra au fil du temps d'affiner l'analyse et de ne plus commettre ce type d'erreur
- **les erreurs de modèle** : le modèle n'est qu'une simplification de la réalité et ne reflète donc pas de manière tout à fait exacte la réalité. Cette simplification doit permettre de traiter efficacement un problème complexe donné mais une simplification trop importante qui ne tiendrait pas compte de certaines variables ou données qui in fine pourraient être susceptibles de fortement impacter une situation conduira à des conclusions erronées
- **la non-prise en compte de données extrêmes importantes** du fait de ne pas pouvoir en extraire une formule mathématique réapplicable
- **les erreurs de continuité du modèle**. On suppose que le modèle est valide et constant pour l'appliquer aux futures situations mais les comportements et/ou l'influence des variables ont pu évoluer à l'insu des concepteurs qui continuent à appliquer un modèle désormais faux.

L'une des grandes différences entre les biais et les erreurs classiques est que la probabilité d'erreur tend à diminuer avec la maturité de l'IA, en particulier avec les phases de tests qui remontent ces erreurs. Vient alors le moment où les concepteurs se trouvent en face d'un nouveau dilemme : les résultats de l'IA peuvent se montrer satisfaisants malgré l'existence d'erreurs (faux positifs, faux négatifs...) et de biais potentiels. Il est important de réaliser une analyse des bénéfices versus le risque associé aux nuisances. C'est la balance entre la valeur attribuée à l'utilisation en conditions réelles de l'IA et la valeur liée à l'exactitude de cette dernière.

1.5. La nécessaire prise en compte de l'importance des biais et des erreurs, et de leurs impacts

Il est important de noter que tous les biais ne sont pas mauvais et n'ont pas tous des conséquences catastrophiques ; par exemple, une IA, qui reconnaît moins souvent les chênes que les bouleaux sur une image, n'a pas grande conséquence pour les personnes. Il faut seulement maîtriser les biais qui sont discriminatoires au sens éthique. Et parmi ces biais discriminatoires tous n'ont pas les mêmes conséquences. Que vous voyiez apparaître des suggestions peu pertinentes sur votre fil d'actualité ou que vous soyez en garde à vue parce que l'algorithme vous a confondu avec un terroriste ne pose pas exactement le même problème.

Il faut bien sûr viser à l'amélioration des algorithmes et à la diminution des biais discriminatoires mais pour autant, il ne faut pas bannir l'utilisation des algorithmes : il faut bien étudier la balance bénéfices-risques. Les enjeux peuvent parfois être supérieurs aux risques. Imaginons un algorithme de détection de comportement suspect dans la foule, pour prévenir les auteurs d'attentat avant qu'ils ne commettent leurs méfaits. L'algorithme présente un fort taux de faux positifs et régulièrement des personnes tout à fait innocentes, bien qu'appartenant majoritairement à une minorité ethnique, sont interpellées. Le biais et la discrimination sont possibles mais, en attendant d'améliorer l'algorithme, que faire ? Préférer les enjeux de sécurité publique ou l'évitement des biais discriminatoires. La réponse dépendra évidemment des circonstances et des opinions. Ainsi, des choix éthiques devront être faits. Outre la question de savoir qui sera à même de faire ces choix, il faudra prendre en compte la façon de les présenter. Si on présente la situation ainsi : « la mise en service de cette IA a conduit à l'arrestation de 48 innocents d'origine noire ou arabe en deux ans d'utilisation, faut-il poursuivre son utilisation ? » ou « la mise en service de cette IA a permis d'arrêter toutes les tentatives d'attentat depuis le début de son utilisation bien que 2 personnes innocentes soient interpellées par mois en moyenne, faut-il poursuivre son utilisation ? » Les réponses ne seront probablement pas les mêmes.

1.6. La nature des IA peut conduire aussi bien à une généralisation et à une accentuation des biais et de leur présence qu'à leur réduction

Une autre problématique de la présence de biais dans l'IA est la généralisation à grande échelle de comportements observés localement, comme dans les données d'entraînement. L'algorithme peut en effet identifier un modèle à partir de données biaisées et sera ensuite conduit à reproduire ces biais de manière systématique dans son utilisation. Ça a été le cas avec l'IA de recrutement d'Amazon où les données d'entraînement étaient majoritairement basées sur des CVs d'hommes et où l'algorithme a appris à moins bien noter les CV portant le mot « women ». De ce fait, toutes les femmes étaient exclues. L'IA était encore plus biaisée que la société ne l'était auparavant.

De même, une erreur dans les données d'incitation dans le cadre de l'apprentissage par renforcement conduira à pérenniser un apprentissage erroné.

Enfin, l'opacité et la non-intelligibilité des IA actuelles peuvent renforcer les biais puisqu'il est difficile de déterminer l'existence d'un biais ainsi que, lorsqu'il a été détecté, l'origine du mécanisme de ce biais afin de le corriger. Les IA peuvent ainsi accentuer un biais déjà présent chez l'humain.

Une IA peut donc être davantage biaisée qu'un humain. Mais une IA peut également être moins biaisée que l'humain. En effet, des données initiales d'apprentissage correctes, sans biais, ou auxquelles des corrections appropriées ont été apportées permettront une utilisation systématisée efficace du modèle d'apprentissage issu de ces données fiabilisées conduisant à des résultats non ou moins biaisés. La question qui se pose alors : l'IA est-elle *in fine* plus ou moins biaisée que la personne qu'elle remplace ou qu'elle assiste ? Il serait intéressant de pouvoir développer un système de mesure des biais existants dans chaque configuration.

2. Quelques recommandations : méthodologie, transparence des choix, outils spécifiques

Nous avons vu que les IA sont sources de biais lorsqu'elles sont mal maîtrisées mais, une fois les mécanismes mieux compris et les systèmes de corrections inventés et/ou mieux appliqués, les IA pourront être sources de remèdes aux biais. Les IA pourraient s'affranchir des biais humains et, dans le futur, une IA de recrutement sera peut-être plus équitable qu'un recruteur humain.

Toutefois, le chemin vers une meilleure équité est difficile et délicat. Par exemple, la Cour des Comptes, lorsqu'elle s'est penchée sur le cas de Parcoursup a suggéré de procéder à des ajustements des notes basés sur l'écart réel des notes données par le lycée et ceux du bac pour plus d'équité. L'année précédente, le CESP (le Comité éthique et scientifique de la plateforme Parcoursup) avait proposé l'exact inverse, ne pas faire d'ajustement, également dans le but d'atteindre plus d'équité. L'amélioration de l'équité est-elle donc un problème insoluble ? Plusieurs pistes d'amélioration existent soit dans la conception générale et préalable de l'algorithme, soit dans la mise en place d'outils et de bonnes pratiques.

2.1. Bonnes pratiques méthodologiques et transparence des choix effectués en matière d'équité

Les pratiques méthodologiques se scinderaient en deux catégories : celles qu'on pourrait mettre en place pour réduire les biais en général et celles qu'on devrait mettre en place pour gérer les dilemmes d'équité.

Instauration d'une équipe et d'un environnement attentifs aux problèmes d'équité et de biais

Plusieurs actions existent pour mettre en place un environnement de développement favorable à la prévention et la maîtrise des biais algorithmiques. La première condition pour cela est une connaissance de ces problématiques par le management et leur réelle implication. Grâce à cela, une méthodologie de réduction des biais pourrait être mise en place très en amont dans le processus de développement d'une IA, quasiment dès le recrutement de l'équipe de développement. En effet, une équipe aux profils variés

(différents genres, parcours, origines socioculturelles et ethniques, formations : informatique, statistiques, data science) sera plus à même de partager les expériences, d'avoir un point de vue différent sur les résultats intermédiaires et finaux remontés par une IA et d'en identifier les biais.

Il faut également instaurer un environnement attentif à ces problématiques par la formation des différents concepteurs et acteurs de la conception, ne serait-ce que pour conscientiser nos modes de fonctionnement. En effet, comme nous l'avons vu, le fonctionnement de nos pensées ou de nos habitudes peut facilement et involontairement tendre à des comportements discriminatoires. Par exemple, baser un recrutement sur certains prérequis, comme des stages dans l'industrie, exclut souvent des candidats qui viennent de catégories sociales défavorisées car elles ne peuvent pas se permettre d'aller à ces stages. Pour surmonter ce point, il faudrait penser à concevoir l'algorithme en prenant en compte ces schémas de pensées et demander à étudier les CV sur la base des compétences des candidats et trouver d'autres critères susceptibles de se substituer au moins en partie au manque d'expérience qui, accompagné le cas échéant d'initiatives adaptées complémentaires, pourraient compenser assez rapidement ce manque d'expérience. Cela demande cependant de repenser le processus initial en profondeur et trouver les incitations qui encourageraient les entreprises à participer à de telles initiatives. L'IA pourra alors être très utile pour trouver des solutions innovantes dans cette quête d'équité.

Mise en place de chartes internes d'équité

Il faudrait également mettre en place des chartes internes qui permettent de guider les concepteurs dans le processus de développement et de conception pour éviter la création et la pérennisation de biais algorithmiques. Comme présenté dans le rapport mars 2020 de l'Institut Montaigne « Algorithme : contrôle des biais S.V.P. », les chartes pourraient mettre en avant :

- des exigences de méthodologie pour assurer la qualité des algorithmes
- les propriétés que doivent présenter les algorithmes développés (notamment si l'on veut pouvoir les auditer plus tard lors du déploiement)
- les mécanismes internes pour gérer les tensions entre différents objectifs, définir des exigences d'équité pour les algorithmes, et préciser leur formalisation informatique
- les analyses et évaluations internes à faire subir à l'algorithme », notamment des tests internes qui permettraient de détecter la présence des biais le plus tôt possible.

Traitement des dilemmes d'équité

Par ailleurs, comme nous l'avons vu plus haut, l'équité présente parfois des dilemmes qui ne pourront être répondus que par des choix plus ou moins justes. L'environnement de conception devrait être à même de savoir identifier ces dilemmes et de mettre en place un système de choix de l'équité à privilégier et des méthodes à appliquer. L'ICO, dans son document « explaining decisions made with artificial intelligence » de mai 2020, donne l'exemple suivant : *« si les résultats discriminatoires du modèle sont motivés par un manque relatif de données sur une population minoritaire, la précision statistique du modèle pourrait être augmentée en collectant plus de données à leur sujet, tout en égalisant les proportions de prédictions*

correctes. Cependant, dans ce cas, vous seriez confronté à un choix différent - entre collecter davantage de données sur la population minoritaire afin de réduire le nombre disproportionné d'erreurs statistiques auxquelles elle est confrontée, ou ne pas collecter ces données en raison des risques posés aux autres droits et libertés de ces personnes. »

Communication des choix effectués en matière d'équité et de l'état actuel des IA relatif à ces choix

Il serait également judicieux de maintenir une documentation appropriée aussi bien pour les acteurs internes (description du cadre de gouvernance des données, de la politique d'atténuation des biais et de la discrimination, des contrôles de gestion des accès, évaluation approfondie des impacts des différentes erreurs repérées, processus de tests, résultats et mesures prises en conséquence...) qu'externes comme les auditeurs ou les utilisateurs. Il faudrait notamment communiquer sur les choix réalisés au cours du processus de développement et de maintien de l'IA. L'utilisateur serait donc à même de savoir si l'algorithme présente des biais volontaires, quelles méthodes (inductives ou déductives) sont utilisées, comment sont traités les signaux faibles et les évènements exceptionnels. Cela lui permettrait de comprendre le fonctionnement de l'algorithme mais il serait également judicieux de présenter des renseignements sur l'état actuel de l'algorithme : quel est son niveau de maturité (notamment au niveau de son apprentissage), quels sont la fréquence et les impacts des faux positifs et faux négatifs, quel est le niveau de l'indicateur de l'évolution de l'algorithme dans le temps, quels sont les risques...

Actions de suivi après le déploiement des IA

Des actions de surveillance après la mise en production du logiciel devraient également être mises en place, ne serait-ce que parce que les IA peuvent évoluer dans le temps. Ces actions peuvent être l'instauration d'un processus d'évaluation des algorithmes par des tiers, l'utilisation des systèmes décisionnels traditionnels en même temps que du système IA qui doit les remplacer pour une certaine durée le temps et de comparer les résultats ou la mise en place d'un programme de chasseurs de bug. Plutôt que de laisser les découvreurs de bug signaler publiquement les erreurs ou, pire, les revendre à des personnes malintentionnées, il s'agit d'offrir des récompenses à ceux qui les signalent directement aux institutions concernées. Ce système présente des limites surtout lorsqu'il est mal calibré (communauté de chasseurs de bugs trop petite, propriétés du système impossible à découvrir hors de la société créatrice, mauvaise spécification des primes...) mais reste préférable à l'absence d'action.

2.2. La mise en œuvre d'outils et des bonnes pratiques techniques spécifiques aux IA pour traiter les biais

Outre la mise en place d'un environnement conscient et favorable à la maîtrise des biais, plusieurs techniques statistiques et algorithmiques existent et pourraient être mises en place, même si tous ces outils ne sont pas compatibles entre eux.

Outils d'amélioration des techniques statistiques des IA pour traiter les biais

La première catégorie de ces techniques s'intéresse à l'amélioration statistique des IA avec la complétion des bases de données, le redressement des données et la correction de la dérive temporelle :

- **complétion des bases de données** : Si certains types de données manquent dans la base et donnent donc une mauvaise représentativité des cas possibles, il est possible, sous certaines conditions plus ou moins restrictives suivant les techniques, de reconstituer des données. La méthode « hot-deck » consiste à compléter les valeurs manquantes par la moyenne des valeurs issues de profils d'individus ou d'objets ayant des caractéristiques similaires. Évidemment, cette approche présente de nombreuses limites dont l'aggravation des biais initiaux. Le rééchantillonnage est une technique ayant pour but de créer des populations artificielles d'objets ou de personnes qui ressemblent à la population cible en utilisant des procédures de sous-échantillonnage et sur-échantillonnage ou en créant des objets ou des individus virtuels par interpolation. Cela permet de répliquer ou supprimer certaines unités et, est largement pratiqué du fait de leur simplicité. Toutefois, aucune étude théorique n'a été conduite pour la valider à ce jour et, dans certains cas, cela peut conduire à des résultats erronés
- **redressement des données** : certaines bases de données sont construites de manière automatique en sélectionnant certains profils de manière aléatoire parmi un plus grand nombre de profils. Lorsqu'on remarque qu'une classe d'objets ou d'individus est peu représentée dans la base de données finale, il faut comprendre pourquoi elle a été peu sélectionnée puis de quantifier la probabilité - probabilité d'inclusion - qu'un élément provenant de cette classe figure dans la base de données. Si l'on dispose de certaines informations additionnelles (d'autres variables dans la base de données) on peut dans certains cas comprendre l'origine de ce biais de sélection et le corriger via un principe d'estimation correctif (dit de Horvitz-Thompson). Dans certains cas, on peut aussi modifier les pondérations de certaines variables grâce à l'utilisation de scores de propension dépendant de l'information additionnelle, c'est la post-stratification aussi appelée calage
- **correction de la dérive temporelle** : les données issues des capteurs peuvent parfois être analysées sur des fenêtres de temps trop courtes, ce qui conduit à ignorer certaines tendances de long-terme ou des effets saisonniers. L'une des solutions est d'incorporer des modèles temporels supplémentaires qui décrivent les mécanismes d'évolution des phénomènes analysés.

La recherche sur les solutions statistiques en est encore à ses débuts mais elle est très prometteuse et de nombreux résultats devraient être publiés au cours des prochaines années.

Outils d'amélioration des techniques algorithmiques des IA pour traiter les biais

La seconde catégorie de ces techniques est algorithmique. Certaines de ces techniques ont pour but de corriger les problèmes existants tandis que d'autres sont orientées sur la conception d'algorithmes qui répondent à des critères d'équité. Ces derniers répondent principalement à trois définitions d'équité différentes : l'anti-classification, la parité de classification et la calibration.

- **l'anti-classification** : les algorithmes basés sur les principes d'anti-classification ne prennent pas en compte les attributs protégés (genre, religion, origine ethnique, ...) dans les méthodes de classification ou de prédiction. C'est un « algorithme à œillères » qui réalise une « équité par l'ignorance » puisqu'on ignore volontairement certaines variables. Il faut noter que dans certains systèmes, les variables à ne pas discriminer peuvent être exclues mais l'intelligence des algorithmes est telle qu'il est assez facile de reconstruire ces informations à partir d'autres données (la personne achète en ligne des produits pour barbe par exemple). Il faut donc connaître suffisamment le fonctionnement de l'algorithme (et donc avoir un système intelligible et transparent) pour parer à ces constructions. Cette faiblesse remet souvent en cause cette catégorie d'algorithme et de nombreuses voix s'élèvent pour l'utilisation raisonnée des attributs protégés (cf. plus bas). Cette catégorie d'algorithme favorise l'équité individuelle
- **la parité de classification**, aussi appelée parité démographique ou parité d'erreur. Dans ces algorithmes, tous les individus qui appartiennent à un même groupe devraient avoir la même probabilité de résultat. Avec la parité d'erreur, un modèle est juste s'il donne un nombre égal d'erreurs à différents groupes. Les algorithmes de ce type devraient être soumis à des méthodologies de tests inspirés de ceux des tests cliniques. Cette catégorie favorise l'équité de groupe
- **la calibration** : ici, après un contrôle du risque estimé, les résultats doivent être indépendants des attributs protégés. Ces algorithmes sont orientés sur l'estimation de la probabilité d'un résultat et sa fréquence réelle. Par exemple, si un modèle de classification classe les demandeurs de prêt entre ceux dont les chances de remboursement sont faibles, moyennes ou élevées, il devrait y avoir des proportions égales d'hommes et de femmes qui remboursent réellement dans chaque catégorie de risque. Cela ne signifie pas qu'il devrait y avoir des proportions égales d'hommes et de femmes dans différentes catégories de risque. Par exemple, si les femmes ont effectivement eu des taux de remboursement plus élevés que les hommes, il se peut qu'il y ait plus de femmes que d'hommes dans la catégorie à faible risque. L'algorithme est programmé pour respecter cette indépendance entre variables protégées et l'évaluation des performances des personnes, même si cela se fait au détriment de certains.

Généralement, ces différentes approches sont généralement incompatibles entre elles.

Autres dispositifs pour traiter les biais

Enfin, d'autres pistes techniques hors algorithmie pure, existent. Ça peut être par exemple :

- la mise à disposition de bases de **données publiques de tests** pour évaluer le niveau de biais dans les méthodologies
- la mise en place de dispositifs spécifiques comme **l'analyse d'impacts**, pour les IA à haut risque
- la mise en place d'une **démarche d'équité active** en autorisant l'usage de variables sensibles dans le but unique de mesurer les biais et d'évaluer les algorithmes. Il s'agit là d'un changement de philosophie en acceptant de mesurer les discriminations. Cette philosophie se base sur l'idée que l'algorithme équitable ne repose pas sur l'exclusion des variables sensibles dans un algorithme mais sur l'indépendance du résultat par rapport à des variables protégées
- la mise en place de **labels spécifiques** : un label garantissant l'absence de biais semble utopique mais il peut s'intéresser à d'autres métriques comme la présence d'une méthodologie d'évaluation des risques de biais dans l'organisation.

Nous avons vu que la confiance dans les IA ne sera possible que si nous avons confiance dans l'équité de ces résultats. Pourquoi suivre l'avis d'une machine si elle nous traite plus inéquitablement encore qu'un décisionnaire humain ? Malheureusement, l'équité est une notion complexe à appréhender. Il existe plusieurs types d'équité qui peuvent rentrer en conflit. Ce qui est équitable pour un groupe de personnes ayant des caractéristiques communes peut être inéquitable envers les individus. Il sera donc nécessaire de mieux cerner l'équité et de faire des choix pour mettre en œuvre l'une ou l'autre de ses types.

La discrimination provient en partie des biais et des erreurs qui faussent la justesse des résultats. Les biais sont très variés et nombreux. Leurs causes sont multiples : ce peut être des erreurs méthodologiques, techniques, statistiques ou probabilistes ou même liés au processus d'apprentissage. La complexité technologique des IA rend l'apparition de ce type de biais plus facile. La mise en place de bonnes pratiques comme des démarches spécifiques, des formations dans le domaine... devraient permettre de maîtriser à terme ce phénomène. Une autre vaste catégorie de biais est liée au fonctionnement de notre cerveau. Ces biais cognitifs sont plus difficiles à détecter car ils font partie intégrante de notre biologie.

La prise en main de ce sujet est particulièrement sensible car l'essence des systèmes d'informations à base d'intelligence artificielle peut réduire les biais s'ils sont bien maîtrisés ou les accentuer dans le cas contraire ! Heureusement, la mise en place d'actions de prévention et de traitement permettrait de diminuer l'impact négatif des biais : les chartes d'équité, les formations dans le domaine, le suivi attentif des IA sont quelques-uns de ces outils.

CHAPITRE 8

EXIGENCES D'HUMANITÉ

Bien qu'encore assez peu étudiée dans le contexte des IA, l'exigence d'humanité est une préoccupation croissante. Cette notion n'est pas là pour dire que les systèmes d'IA doivent devenir humains mais pour rappeler que ceux-ci doivent servir à l'humanité dans le respect des droits de chacun – tel que décrits dans la Charte des droits fondamentaux de l'Union européenne – et offrir les mêmes opportunités à tous. Cela signifie, au-delà du fait qu'ils ne doivent pas viser la mort ou l'atteinte de l'intégrité physique ou psychologique de personnes, ils doivent favoriser la bienfaisance. Ainsi, ils doivent éviter d'augmenter les fractures sociales et économiques en favorisant un petit groupe de personnes au détriment des autres ou créer davantage de déséquilibres de forces entre les États. Au contraire, comme le mentionne un des principes d'Asilomar[42], « la prospérité économique engendrée par l'utilisation des IA devrait être partagée largement pour bénéficier à toute l'humanité ». L'IA devrait être une force positive pour la société et l'humanité.

1. Problématiques spécifiques à l'IA : dignité et vie privée, bienfaisance et non-malfaisance, liberté de penser et libre accès aux informations, justice et équité, autonomie

On peut remarquer que de nombreuses caractéristiques de l'IA peuvent facilement conduire à des situations qui pourraient être en contradiction avec les droits décrits dans la Charte des droits fondamentaux de l'Union européenne[43] :

- **droit au respect de sa vie privée et à la dignité** : il convient de noter qu'il existe plusieurs dimensions à ces notions de vie privée. La plus connue est la protection des données personnelles couverte notamment par la RGPD. Mais il en existe plusieurs autres. Il s'agit, par exemple de :
  - la protection concernant la personne elle-même, son corps. Il s'agit de son droit de n'y pas porter atteinte, de pouvoir ne pas accepter de fournir du sang, des tests ADN, des tests d'alcoolémie ou de drogues, d'implanter des puces d'identification (type RFID) pour l'authentifier au passage de portes de sécurité ou pour le localiser, de passer devant un scanner à l'aéroport qui pourrait évaluer son état de santé...

---

[42] Les principes d'Asilomar sont issus de la conférence du même nom « Asilomar Conference on Beneficial AI » qui s'est tenue en Californie en 2017. Plus de 100 chercheurs dans les domaines économique, juridique, éthique et philosophique y ont formulé les « 23 principes d'Asilomar » pour guider éthiquement les recherches informatiques.

[43] Notons que cette charte est une base répandue mais de nombreuses voix soulignent qu'elle ne s'intéresse pas assez à l'humanité dans sa globalité et se concentrent sur les droits individuels. Or, nous avons vu au § 1.1 du chapitre 7 qu'il peut y avoir un réel conflit éthique entre les intérêts du groupe et ceux de l'individu.

- la protection concernant les pensées et les émotions avant qu'elles ne soient exprimées en public telles que des croyances religieuses, des préférences sexuelles, des idées politiques, ... qui soient par exemple surveillées dans un espace public ou privé. Il peut s'agir de vues prises par des caméras ou d'autres moyens
- la protection concernant les informations communiquées à d'autres. Des conversations pourraient être enregistrées par des microphones cachés, ou suivies par l'administration.
- la protection concernant des informations à fournir à d'autres qui pourrait leur nuire en cas de refus. Il peut s'agir, lors d'une embauche, des informations concernant ses idées (politiques, religieuses...) ou de donner son mot de passe pour permettre à un tiers d'entrer sur le compte de la personne dans un réseau social et vérifier la situation
- la protection concernant sa localisation. De même, on doit respecter son environnement familial ou professionnel (sa maison, son lieu de travail...). Il peut s'agir de l'utilisation de drones, de caméras de surveillance... pour prendre des photos à son insu
- la protection concernant son appartenance à un groupe. On doit respecter son droit d'association avec quiconque sans subir un monitoring intempestif. Ceci peut aussi être lié au fait d'appartenir à une ethnie ou à une famille particulière.

Dans tous ces cas, les IA pourraient être concernées.

En effet, les IA sont très consommatrices de données personnelles. À partir de celles-ci, elles peuvent faire du profilage. Elles peuvent ainsi identifier des traits de caractère, des comportements personnels, des faiblesses psychologiques, d'un individu pour arriver à le connaître dans son intimité mieux que sa famille, ses amis, voire que lui-même. À partir de cette caractérisation, il est possible d'identifier des services qui lui conviendraient, l'inciter de manière subliminale à des actions qui pourraient lui être nuisibles voire toucher à sa dignité humaine.

Dans ce contexte, comment s'assurer que ses droits fondamentaux sont respectés ? Dans tous les cas notés ci-dessus, qui serait concerné, quels seraient les usages interdits, les contextes dans lesquels ils seraient autorisés, les limites de leur utilisation, les outils qui aideraient à leur traitement... On pourrait par exemple, imaginer une interdiction de la reconnaissance faciale hormis des cas spécifiques bien identifiés telle que la prévention d'actes terroristes.

- **exigence de bienfaisance et de non-malfaisance** : l'objectif est de développer des IA qui ne portent pas atteinte à la vie ou à l'intégrité physique des personnes. Si l'on considère que toute personne a droit à la vie et que nul ne peut être condamné à la peine de mort ni exécuté, comment justifier l'utilisation de drones militaires automatisés ? Doit-elle être interdite ou encadrée ? Si oui comment ? Concernant le droit à l'intégrité physique, a-t-on le droit d'implanter des traceurs dans le corps d'une personne pour faciliter leur identification ou pour les géolocaliser en cas de problème, de capteurs pour diagnostiquer l'état de santé...

- **droit à la liberté de pensée et d'information** : l'usage d'IA pose des risques d'enfermement algorithmique. Or en ne proposant aux utilisateurs que certaines catégories d'information basées sur ce qu'une IA estime l'intéresser et/ou lui convenir, on prive son accès à d'autres informations, ce qui peut également entraver sa liberté de pensée
- **égalité des droits** : chacun d'entre nous (femmes, hommes, enfants, personnes âgées et/ou en situation de handicap) a le droit à un traitement identique pour une situation similaire. Nous avons vu dans la partie exigence d'équité et de non-discrimination les différents problèmes concernant ce sujet. Il convient de tenir compte du fait qu'il s'agit bien d'un droit fondamental et qu'il doit être traité en tant que tel
- **droit à l'autonomie** : il s'agit de la liberté d'agir en fonction de ses valeurs et de ses croyances. Concrètement, l'autonomie de chacun s'exerce lorsqu'on agit de manière volontaire, indépendante, sans contrainte extérieure, et spécifique. Cette volonté doit être manifeste. De plus, nul ne peut contraindre autrui même au nom du principe de bienfaisance. En réalité, il n'est pas toujours facile d'apprécier cette volonté. Cette liberté d'agir peut-être réelle mais elle peut aussi n'être qu'affichée. Il y a souvent une différence entre la parole et les actes. Non seulement la notion de consentement s'impose mais il faut voir de quelle manière cela se traduit concrètement. De plus, il faut s'assurer que c'est un consentement libre en connaissance de cause. L'exemple qui est mis souvent en exergue est celui du consentement des cookies apparus à la suite de la mise en œuvre de la RGPD, qui a suscité le rappel à l'ordre de la CNIL. Avant ce rappel, si la possibilité de les refuser était offerte, elle était en réalité et en général beaucoup plus longue à réaliser que d'accepter les cookies. Il fallait souvent décocher toutes les options manuellement ou bien même un pop-up descriptif de la politique des cookies apparaissait sans qu'on puisse refuser d'y adhérer.

Dans le cadre des IA, ce droit à l'autonomie doit s'exercer alors que ces IA prennent des décisions automatisées qui sont souvent difficiles à comprendre du fait de leur opacité. Pouvons-nous réellement consentir de manière éclairée à accepter les résultats qui en sont issus alors que nous pouvons ne même pas savoir si des décisions ont été prises qui nous concernent, comment elles ont été prises et pourquoi ? Comment traiter ce sujet ? Avancer sur les problèmes d'explicabilité et d'équité contribuera à faciliter l'exercice de cette autonomie. Mais, il devra être accompagné d'un dispositif qui permette de mieux maîtriser et d'avoir une confiance raisonnable dans les décisions automatisées prises par ces IA voire de reprendre le contrôle en exerçant le « bouton rouge » en cas d'urgence ou de dérapage de l'IA.

Une reprise en main des IA par l'humain, un des enjeux les plus difficiles à résoudre

Le livre blanc sur l'IA par la Commission européenne évoque « *la possibilité de suivre le système d'IA pendant son fonctionnement et la capacité d'intervenir en temps réel et de le désactiver (ex. : dans une voiture autonome, un bouton ou une procédure d'arrêt peut être activé par un être humain lorsqu'il estime que la conduite de la voiture n'est pas sûre)* ». Mais ce document ne donne pas plus d'indications sur le dispositif mis en place. Le lecteur attentif remarquera souvent l'absence d'explications ou de détails supplémentaires sur ce bouton rouge. Ce n'est pas sans raison : cette solution est elle-même sujette à questions.

CHAPITRE 8

Comment mettre en place un bouton rouge ? L'algorithme est souvent au sein d'un système, il n'est pas possible d'arrêter juste un algorithme défaillant, il faudrait arrêter l'ensemble de l'algorithme et trouver une autre méthode de faire l'action mais il n'existe pas (ou plus) forcément de méthode alternative. C'est le cas dans les algorithmes de trading. Les procédures internes aux banques impliquent qu'il faut plusieurs minutes voire plusieurs heures entre le moment de la demande de reprise en main manuelle et sa réalisation effective. Problème : un ordre de trading ne prend que quelques fractions de seconde à être donné. On pourrait répondre que cette possibilité de déconnexion en quelques minutes a le mérite de limiter les problèmes mais à l'heure où la finance est plus qu'interconnectée, quelques minutes suffisent pour avoir de graves conséquences. Ne serait-il pas trop tard ?

Si on reste dans cette perspective de temporalité, l'instant de la prise de décision d' « appuyer sur le bouton rouge » est critique. Il faut que l'humain ait assez de temps pour reprendre le contrôle. Par exemple, si un « conducteur » de voiture autonome voit qu'un piéton surgit pour traverser la rue et que le véhicule ne ralentit pas assez, il active le « bouton rouge » en appuyant de lui-même sur la pédale de frein. La voiture aura-t-elle le temps nécessaire pour s'arrêter ? Sa prise de décision a-t-elle été prise assez tôt ? Certains acteurs juridiques auraient tendance à légiférer en disant justement que si l'humain n'est pas capable de reprendre sans risque le contrôle de la machine, alors il faut que l'IA conserve la main. De manière générale, il existe un flou juridique sur cette question.

Mais nous pouvons envisager encore pire : doit-on vraiment activer ce fameux bouton rouge ? Deux thématiques différentes correspondent à cette question. D'une part, l'IA présente le même processus d'apprentissage que l'homme : on apprend à partir de nos erreurs. Nier à l'IA le droit de commettre des erreurs revient à empêcher l'IA de s'améliorer. Comment concilier ces deux éléments ? D'autre part, dans les cas, voués à se multiplier, où la performance de l'IA est déjà supérieure à la performance humaine, faut-il laisser aux humains le droit de se tromper ?

En effet, imaginons une IA de détection des cancers qui a 90 % de réussite là où un médecin n'en aurait que 70 %. La question de l'activation du bouton rouge doit se soulever quand le médecin est en désaccord avec l'IA. Mais cette situation reflète trois réalités potentielles différentes : l'IA a tort et le médecin a raison, les deux ont tort ou le médecin a tort et l'IA a raison. Statistiquement, le dernier cas a le plus de chances d'arriver. L'humain a-t-il le droit de reprendre la main ? La question est déjà réelle : une IA qui ne dit plus son nom est celle des pilotes automatiques d'avion. Les retours montrent que, suite à un problème, la majorité des accidents arrivent alors que le pilote a repris la main sur le pilote automatique. Qui est responsable lorsque la reprise en main manuelle conduit à une situation beaucoup plus néfaste ? Comment faire en sorte que le contrôle humain, qui a pour but d'éviter qu'un système d'IA ne mette en péril l'autonomie humaine ou ne provoque d'autres effets néfastes, ne provoque lui aussi des effets néfastes ?

2. Quelques recommandations : clarification, analyse d'impact, charte, outils, gouvernance

Il n'existe que peu de solutions techniques à tous ces problèmes liés aux droits fondamentaux. La mise en place d'un bouton rouge est une de ces solutions mais nous avons pu constater ses limites. Beaucoup de travail reste à accomplir pour avancer efficacement sur ce sujet. Il est ainsi possible de recommander de :

- clarifier les situations spécifiques aux droits fondamentaux qui sont susceptibles d'être impactés par la mise en œuvre d'IA
- intégrer dans les analyses d'impacts ceux qui concernent les droits fondamentaux lorsque des IA sont concernées
- définir les problématiques spécifiques à traiter et identifier les outils qui aideraient à les résoudre
- décliner la Charte des droits fondamentaux à l'utilisation des IA
- déterminer les différents modes de gouvernances et les différentes responsabilités.

Le livre blanc sur l'IA par la Commission européenne détaille certaines suggestions : « *Le contrôle humain devra dépendre notamment de l'utilisation prévue des systèmes et de ses incidences potentielles sur les citoyens et les personnes morales concernées.* » Il cite quelques exemples où « *un contrôle humain peut être exercé :*

- *des résultats du système ne deviennent effectifs que s'ils ont été préalablement réexaminés et validés par un être humain (ex. : la décision de rejeter une demande de prestations de sécurité sociale ne peut être prise que par un être humain)*
- *les résultats du système d'IA deviennent immédiatement effectifs, mais un être humain intervient par la suite (ex. : une demande de carte de crédit peut être rejetée par un système d'IA, mais cette décision doit pouvoir être réexaminée ensuite par un être humain)*
- *la possibilité de suivre le système d'IA pendant son fonctionnement et la capacité d'intervenir en temps réel et de le désactiver (ex. : dans une voiture autonome, un bouton ou une procédure d'arrêt peut être activé par un être humain lorsqu'il estime que la conduite de la voiture n'est pas sûre)*
- *pendant la phase de conception, en imposant des contraintes opérationnelles au système d'IA (ex. : une voiture autonome doit cesser de fonctionner dans certaines conditions de faible visibilité qui diminuent la fiabilité des capteurs, ou maintenir, dans une quelconque condition donnée, une certaine distance par rapport au véhicule qui précède)* ».

Ce concept d'exigence d'humanité, encore peu étudié, souligne le fait que les systèmes à base d'intelligence artificielle doivent soutenir l'humanité dans le respect des droits de chacun et bénéficier au maximum à tous plutôt que de privilégier un petit nombre de personnes. Or la nature des IA peut entrer en contradiction avec certains des droits fondamentaux. Par exemple, leur besoin dévorant de données entre en conflit avec le droit au respect de la vie privée. Il importe alors d'identifier tous les cas à risque et de mettre en place des actions appropriées. La plus connue d'entre elles est celle du bouton rouge qui devrait permettre de désactiver une IA à la manière du bouton d'arrêt d'urgence dans les usines. Mais les conditions d'arrêt, les moyens pour le faire sont encore à un stade très précoce et plus de questions se posent que de réponses n'ont été apportées. Heureusement, d'autres actions existent, comme la mise en place d'analyse d'impact, qui permettront de commencer à traiter la problématique.

CHAPITRE 9

EXIGENCES « TRADITIONNELLES » DES SYSTÈMES D'INFORMATION AVEC DES SPÉCIFICITÉS IA : PERFORMANCE, FIABILITE, SECURITÉ, RESILIENCE, ...

Avant de se définir par leur culture, les français, ghanéens, péruviens ou autres sont avant tout des hommes. De la même manière, une IA est avant tout un système d'information. Mais puisqu'ils sont de culture différente, chacun n'aura pas tout à fait les mêmes exigences. Le salaryman japonais est traditionnellement enjoint à rester à son travail jusqu'en soirée alors qu'un manager allemand s'inquiète si son collègue n'est pas parti à temps retrouver sa famille. Ainsi, tous les systèmes d'informations doivent suivre les mêmes exigences. Mais puisqu'on s'intéresse à l'intelligence artificielle, les exigences se sont différenciées. Nous verrons à quel point les exigences de vie privée, de fiabilité, de robustesse, de pérennité et de cybersécurités sont modifiées dans ce contexte et comment ont évoluées les problématiques associées. Nous donnerons ensuite quelques recommandations pour commencer à répondre à ces problématiques.

1. Problématiques spécifiques à l'IA : principe de minimisation, consentement, données d'apprentissage, risques inconnus, dépendance, scalabilité, portabilité, modèles, capteurs

1.1. Exigence de vie privée et RGPD : principe de minimisation, consentement, données d'entraînement, droit à l'oubli

Sans même utiliser de l'IA, la vie privée est régulièrement mise en danger par les systèmes d'information. Les systèmes d'IA peuvent exacerber les risques de sécurité connus et les rendre plus difficiles à gérer. Ils présentent également des défis pour le respect du principe de minimisation des données : d'une part, la minimisation de l'utilisation des données dans le respect de la RGPD, d'autre part, la conservation d'un grand nombre de données permet un meilleur fonctionnement de l'IA. Parcoursup, comme tant d'autres systèmes, s'est retrouvé soumis à ces deux positions contraires. Comment trouver l'équilibre et qui doit en porter la responsabilité ?

L'utilisation de l'IA ne fait que compliquer les choses par rapport aux difficultés des systèmes d'information sans IA : comment prendre soin de sa vie privée quand l'utilisateur ne sait même pas qu'un système d'IA est utilisé, qu'il ne sait pas à quel moment du processus il se trouve et que le fonctionnement algorithmique lui-même est opaque ? Il est nécessaire d'avoir un consentement clair de l'utilisateur.

Une autre question : comment traiter l'utilisation des données d'entraînement ? Que faire de ces données qui ne sont pas directement utilisées à des fins personnelles mais qui utilisent des informations personnelles ?

Cela présente également de nouveaux risques de vie privée : les données d'entraînements, les paramètres et autres éléments techniques qui sont concernés peuvent être piratés ou être diffusés illégalement, alors que ces données ne sont traditionnellement pas conservées.

Comment également faire valoir son droit à l'oubli quand nos données personnelles ont été utilisées pour éduquer un algorithme ? On peut toujours supprimer les éléments de la base mais l'algorithme en conserve une trace, même infime, dans sa manière de fonctionner. D'autant plus qu'il est connu que certaines IA peuvent déduire certaines données manquantes à partir de l'ensemble des données disponibles.

1.2. Fiabilité, robustesse et résilience : risques inconnus, dépendance, scalabilité, portabilité, erreurs d'interprétation, modèles

L'utilisation de l'IA fait apparaître de nouveaux risques de fiabilité, de robustesse et de résilience du système d'IA. L'IA constitue un champ de recherche nouveau où des risques inconnus existent même s'ils n'ont pas encore été rencontrés. Ça a été le cas de l'accident mortel d'Elaine Herzberg en 2018 avec une voiture autonome : les concepteurs du logiciel n'avaient pas pensé à prévoir le cas où un piéton (qui plus est, tenant son vélo en main, donc posant des problèmes supplémentaires d'identification) traverserait la route en dehors d'un passage piéton. C'est cet accident qui a permis de découvrir ce « cas d'utilisation ».

Sans même songer à tous les problèmes existants mais encore inconnus, d'autres nouveaux problèmes sont soulevés par les IA :

- **la dépendance** : l'IA repose sur les trois piliers du matériel, du logiciel et de la donnée, comme nous l'avons vu au § 2.3 du chapitre 1. Celui ou ceux qui maîtriseront ou possèderont ces éléments seront en mesure de placer les autres dans une relation de dépendance. L'omnipotence des GAFAM pourrait ainsi entrer en contradiction avec ces exigences
- **la scalabilité** : l'IA est testée sur plusieurs milliers de données mais on peut se demander si les IA sauront fonctionner avec la même efficacité avec plusieurs millions de données voire milliards de données
- **la portabilité** : les systèmes d'IA évoluent vite et leurs composants aussi. Ces évolutions s'accélèrent. Combien de temps les systèmes d'IA conçus à un instant t pourront-ils perdurer ? Comment prendre en compte l'évolution des parties et des systèmes ?
- **les erreurs d'interprétation** : les résultats issus des systèmes d'IA peuvent être difficilement compréhensibles. On peut se demander à quoi correspond exactement tel ou tel score de sortie finalement calculé par l'algorithme. Certains noms de score peuvent être trompeurs, d'autres flous ou bien leur signification a été mal expliquée aux utilisateurs ou ces derniers les ont mécomprises

- **la fiabilité des modèles** par rapport aux données : le principe de l'IA est de simplifier la réalité retranscrite dans les données récoltées pour en extraire des prédictions ou des conclusions. Toutefois, un modèle qui ne prend pas en compte certaines données n'est pas fiable. Par exemple, une voiture autonome qui n'intègre pas l'état d'humidité de la route est dangereuse pour le freinage.

Le rapport technique du groupe d'experts de la Commission européenne « Robustness and Explainability of AI » stipule qu'évaluer la fiabilité d'une IA nécessite d'évaluer deux facettes de la non-fiabilité :

- l'IA présente de mauvaises performances dans des conditions considérées normales pour l'homme
- l'IA présente des vulnérabilités – malgré de bonnes performances – qui peuvent entraîner des dysfonctionnements dans des conditions spécifiques, comme ce fut le cas de l'accident subi par Elaine Herzberg. Il peut s'agir de dysfonctionnements apparaissant naturellement ou qui sont provoqués intentionnellement.

Ce dernier point soulève une nouvelle particularité des exigences traditionnelles par rapport à l'IA : celle concernant l'exigence de cybersécurité.

1.3. Cybersécurité : modèles, données d'apprentissage, capteurs, brouillage, virus

Conceptuellement, un hacker exploite les failles[44] d'un système. Plus le système est compliqué et complexe, plus les failles sont potentiellement nombreuses. Tel était déjà le cas dans les systèmes traditionnels. Ainsi, on comprend facilement que les problèmes de cybersécurité sont de différents types :

- **les failles déjà existantes** : si certains types de failles ne sont pas plus importants avec l'utilisation de l'IA, d'autres sont exacerbés et plus difficiles à maîtriser. C'est le cas de la perte et l'utilisation abusive de grandes quantités de données ou le risque d'introduction de vulnérabilités lors de l'évolution d'un système d'IA (nouveaux codes et/ou changements d'infrastructures)
- **les failles nouvelles liées à l'utilisation de modèles** : les vulnérabilités peuvent être exploitées au niveau des différents éléments de traitement du modèle ou, autre exemple, pour des raisons de temps, on peut réutiliser un modèle entraîné par des tiers avec un dispositif sécurisé mais qui présente des failles en amont ou en aval du dispositif. Autre exemple, les modèles peuvent être très similaires aux données d'entraînement et peuvent être récupérés par des adversaires : une telle attaque a été réalisée et a permis d'extraire les numéros de carte de crédit d'un modèle de traitement du langage naturel développé pour l'autocomplétion*

[44] Notons que les biais sont à traiter différemment des failles de sécurité car elles ont des causes et des traitements différents

- **les failles nouvelles liées à l'entraînement des modèles** : les possibilités sont nombreuses. Ce peut être du data poisoning où on introduit de fausses données dans les données d'entraînements ce qui faussera le résultat final ; du reward poisonning où, pour des algorithmes utilisant l'apprentissage par renforcement[45], le pirate altère les récompenses modifiant ainsi l'apprentissage de l'IA qui construit un modèle final faux.

  Cela peut aussi être l'attaque de l'inversion de modèle. Ici, les hackers connaissent déjà quelques données personnelles sur certains individus utilisés dans la base d'entraînement et ils vont déduire d'autres données personnelles en observant les entrées et les sorties du modèle. Cette attaque a été utilisée contre un modèle médical prédisant le dosage correct d'un anticoagulant à partir des données des patients. Ici, l'attaquant pouvait déduire les biomarqueurs génétiques des individus utilisés dans les données d'entraînement à partir de certaines informations démographiques.

  Les attaques par inférence d'appartenance permettent de déduire si un individu donné était présent dans les données d'entraînement d'un modèle. Par exemple, si une personne figurait dans les données d'entraînement, le modèle aura une confiance disproportionnée dans une prédiction concernant cette personne, car il les a déjà vues. Cela permet à l'attaquant de déduire que la personne figurait dans les données d'entraînement. Contrairement à l'inversion de modèle, ce type d'attaque ne permet pas directement d'apprendre de nouvelles choses sur un individu mais on peut déduire des éléments de manière indirecte : par exemple, qu'il fréquente l'un des hôpitaux qui ont généré les données de formation au cours de la période de collecte des données. Cela peut poser des risques pour la vie privée s'il s'agit d'un modèle concernant une population sensible comme des patients atteints de démence ou du VIH
- **les failles nouvelles liées à l'utilisation de capteurs** : cela peut tout simplement être des capteurs mal sécurisés avec des mots de passe faibles voire absents
- **les failles nouvelles liées à l'usage d'un système d'IA** : c'est par exemple l'attaque des exemples contradictoires où un hacker introduit des données d'entrées qui sont délibérément conçues pour être mal classées. Ainsi, une photo de bus scolaire où l'on a modifié quelques pixels de manière imperceptible pour l'œil humain peut être reconnu comme du guacamole. La faille peut encore être plus simple : il peut suffire d'introduire une image d'un objet portant un autocollant ou un post-it. Par exemple, une affichette publicitaire sur un panneau stop confondra l'algorithme qui ne saura pas reconnaître le panneau et ne freinera pas. Dans un autre cas, en déformant légèrement l'image du visage d'un individu, un adversaire a trompé le système en faisant classer cette personne dans une autre catégorie alors qu'un humain reconnaîtrait toujours l'image déformée comme l'individu correct.

---

[45] En simplifiant grossièrement, l'action effectuée par l'IA est récompensée ou punie (via des incitations positives ou négatives), et l'algorithme apprend de ces récompenses et erreurs. Ce principe fonctionne aussi pour les humains : c'est l'apprentissage par la carotte ou le bâton.

Le schéma ci-dessous montre les différences d'origine de vulnérabilités entre un logiciel classique, en haut, où le pirate (en rouge) ne dispose que d'une seule « voie » de piratage. Avec un système à base d'intelligence artificielle, schématisé en bas, le pirate peut emprunter trois voies d'attaque différentes, ce qui facilite l'attaque et complique la protection du système :

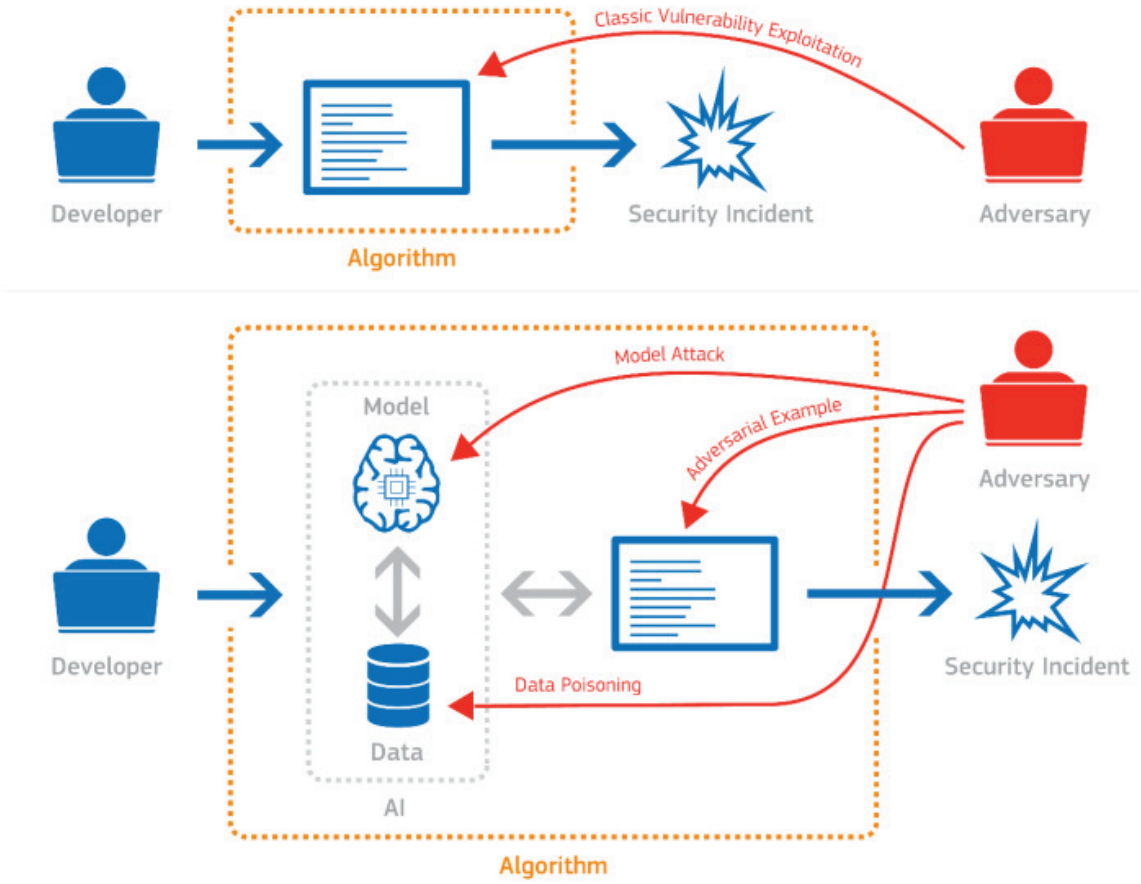

Figure 11 : Changements du paradigme en cybersécurité en raison de l'introduction de composants IA (source : Union européenne JRC technical report robustness and explainability of AI)

Notons aussi le problème des virus : le principe d'un logiciel antivirus repose sur le recensement de toutes les menaces connues. Comment faire pour recenser ou parer aux menaces lorsque les logiciels d'IA sont eux-mêmes évolutifs ?

Enfin, il faut savoir que les exigences de sécurité concernant la cybersécurité peuvent être contradictoires avec l'exigence de transparence. Ainsi, des recherches récentes ont démontré comment certaines méthodes proposées pour rendre les modèles de machine learning explicables peuvent involontairement faciliter la déduction de données personnelles sur les individus dont les données ont été utilisées pour entraîner le modèle.

2. Quelques recommandations : équipes rouges, outils de perturbation, apprentissage mutualisé et distribué, calcul distribué

Il n'y a pas d'approche « taille unique » concernant les exigences de sécurité, vie privée et fiabilité.

Les mesures de sécurité appropriées qui doivent être adoptées dépendent du niveau et du type de risques qui dépendent des caractéristiques du système à base d'intelligence artificielle. Par exemple, une IA qui utilise exclusivement des algorithmes auto-apprenants sera plus vulnérable comparativement à une autre qui n'en utilise aucune. De plus, certains chercheurs[46] estiment qu'il manque des normes pour évaluer les nouvelles techniques d'apprentissage automatique préservant la confidentialité, et la capacité de les mettre en œuvre se situe actuellement en dehors des compétences typiques d'un développeur d'IA.

Toutefois, plusieurs solutions pour atteindre ces exigences peuvent être mises en place, en fonction des besoins. Voici quelques exemples de mesures :

- mettre en place **d'exercices d' « équipes rouges »** : les « équipes rouges » cherchent à adopter l'état d'esprit et les méthodes d'un attaquant pour trouver des failles et des vulnérabilités dans un plan, un système technique ou une organisation. À plus haut niveau, une communauté de professionnels de l'équipe rouge pourrait prendre des mesures telles que publier les meilleures pratiques, analyser collectivement des études de cas particulières, organiser des ateliers sur des problèmes émergents ou plaider en faveur de politiques qui permettraient à l'équipe rouge d'être plus efficace, tout en faisant attention au traitement des données confidentielles
- réaliser des **évaluations de sécurité** pour tout algorithme et modèle développé en interne comme en externe
- **séparer les environnements de développement et de production** en utilisant des machines virtuelles ou en utilisant un langage spécifique pour le développement et en le convertissant en un autre avant le déploiement. Par exemple, Python est peu sécurisé mais très développé pour des utilisations d'apprentissage automatique. Il faudrait le convertir en Java qui rend le codage non sécurisé moins probable au moment où on désire passer l'IA de l'environnement de test à celui de production

[46] Brundage et all, « Toward trustworth AI Development : mechanisms for supporting verifiable claims », 2020

- **introduire des méthodologies standardisées** pour évaluer la robustesse des modèles d'IA, notamment pour déterminer leur champ d'action par rapport aux données qui ont été utilisées pour la formation, le type de modèle mathématique, le contexte d'utilisation…
- utiliser des outils spécifiques comme :
  - les **outils de perturbation** comme la confidentialité différentielle qui introduit du « bruit » pour permettre l'anonymisation et le traitement équitable mais qui risque de faciliter les cyberattaques et peut aussi fortement réduire les performances des systèmes en termes de précision si le niveau de confidentialité est trop élevé
  - les outils **d'apprentissage mutualisés** qui combinent plusieurs tendances de plusieurs modèles locaux mais sans partager les données d'apprentissage ce qui fait que personne n'a accès aux données locales
  - les outils de **génération de données « synthétiques »** qui ne permettent pas de remonter aux données des personnes physiques
  - les outils d'**anonymisation** comme l'utilisation de pseudonymes

  bien que la faisabilité de la mise en place de tels outils dépende fortement du contexte de l'IA et augmente la complexité du système, ce qui a potentiellement un impact sur l'explicabilité de ce dernier.
- utiliser les principes **de sécurité par conception**, c'est-à-dire la prise en compte de la sécurité d'un logiciel ou d'une application dès le début du processus de conception. On peut repenser la procédure d'apprentissage pour être robuste contre les actions malveillantes, en particulier les **exemples contradictoires**. Cela implique un entraînement explicite contre des exemples contradictoires connus, ainsi qu'une refonte du fondement mathématique des algorithmes en utilisant des techniques statistiques, telle que la robustesse qui ajoute du bruit à l'algorithme et vérifie la teneur des réponses obtenues. Il convient de noter que pour les systèmes d'IA basés sur l'apprentissage automatique, cette approche implique probablement une dégradation des performances inhérente aux approches statistiques
- désinfecter les données : **nettoyer les données d'entraînement** de tout contenu potentiellement malveillant avant d'entraîner le modèle est un moyen d'éviter l'empoisonnement des données
- utiliser **plusieurs jeux de modèles** pour tester les données : en faisant passer les mêmes données dans différents modèles – donc différentes versions de l'IA – on peut étudier leur comportement
- incorporer la **vérification formelle** au processus de développement d'une IA : la vérification formelle est un domaine très actif en informatique qui vise à prouver l'exactitude d'un système logiciel ou matériel par rapport à des propriétés spécifiées, à l'aide de preuves mathématiques. Deux propriétés principales sont souvent étudiées :
  - « (In) satisfiabilité » qui vérifie si, pour une entrée donnée, obtenir une sortie définie n'est (pas) faisable
  - la robustesse qui vérifie si l'ajout de bruit à une entrée donnée modifie sa sortie

- développer **l'apprentissage distribué et fédéré**. L'apprentissage du modèle n'est pas effectué par un seul acteur, mais plutôt par une multitude de parties différentes qui peuvent ou non être connectées entre elles. Dans l'apprentissage distribué, toutes les parties apprennent le même modèle et partagent des informations sur les gradients. Avec l'apprentissage fédéré, seuls les paramètres du modèle sont échangés entre les acteurs. Dans ce cadre, chaque acteur n'a accès qu'à sa partie de l'ensemble de données, tout en profitant d'un modèle plus robuste qui est entraîné à l'aide de diverses sources de données. Bien que les informations sur les données d'entraînement puissent toujours fuir à travers le modèle, cela réduit considérablement la divulgation de données sensibles
- développer le **calcul distribué** : une autre façon d'atténuer les risques liés au partage des variables prédictives consiste à héberger le modèle machine learning sur l'appareil à partir duquel la requête est générée et qui collecte et stocke déjà les données personnelles de l'individu. Un modèle de machine learning pourrait être installé sur le propre appareil (PC, tablette, smartphone) de l'utilisateur et faire des inférences « localement », plutôt que d'être hébergé sur un serveur distant ou dans le cloud. La contrainte est que les modèles machine learning doivent être suffisamment petits et efficaces en termes de calcul pour s'exécuter sur le propre matériel de l'utilisateur. Cependant, les progrès récents dans le matériel spécialement conçu pour les smartphones et les appareils intégrés signifient qu'il s'agit d'une option de plus en plus viable
- développer **l'apprentissage sur données cryptées** : Le cryptage entièrement homomorphe est un type particulier de méthodes de cryptographie qui permet d'effectuer des additions et des multiplications sur des données cryptées. Son intégration dans les algorithmes d'apprentissage automatique n'en est qu'à ses débuts, mais cela suggère que l'apprentissage sur des données cryptées pourrait être une stratégie raisonnable lorsque la sensibilité des données est élevée. Si cette approche souffre d'un certain nombre de limitations, la principale étant le coût de calcul élevé actuel d'une seule opération par rapport à l'approche non chiffrée
- enfin le plus évident : créer des **documents explicatifs** de la gestion de la sécurité et des performances, ce qui va en accord avec l'exigence de transparence et d'explicabilité.

CHAPITRE 9

Les systèmes d'information à base d'intelligence artificielle étant avant tout des systèmes d'informations, les exigences traditionnelles s'appliquent à eux. Toutefois, certaines de ces exigences ont évolué en raison de la nature même de la composante IA présente. Les exigences liées à la vie privée, déjà délicates à atteindre traditionnellement, se compliquent avec l'existence des données d'entrainement, la persistance de l'existence des données dans les bases (quid du droit à l'oubli ?) ... Les IA étant plus complexes, ils sont aussi plus fragiles. Cette fragilité présente des impacts nouveaux sur la fiabilité, la portabilité, la scalabilité... et la cybersécurité. Certaines des failles de sécurité existantes sont plus importantes tandis que de nouvelles apparaissent. L'utilisation de modèles, de capteurs, l'entraînement des modèles sont tout autant de nouvelles faiblesses qui sont inhérentes aux IA. Heureusement, plusieurs dispositifs sont possibles afin de traiter ces risques : créer des méthodologies spécifiques et de nouveaux outils comme la vérification formelle, développer l'usage d'équipes rouges, l'apprentissage distribué ou sur données cryptées...

Le groupe a défini une démarche en cinq étapes qui, si elle est suivie correctement, permettrait *in fine* d'aboutir à la création de systèmes d'information à base d'intelligence artificielle dignes de confiance. Cette démarche ne peut être réalisée que dans l'ordre et le passage à l'étape suivante ne peut se faire sereinement que si la phase précédente a été correctement étudiée. L'immaturité de l'IA est telle que même la première de ces étapes, la définition des exigences à atteindre, n'a pas encore abouti dans l'écosystème de l'IA. Bien que les réflexions concernant les étapes suivantes soient moins élaborées, le groupe a tenu à expliciter les problématiques et les enjeux portés par chacune des phases restantes. Nous allons donc nous intéresser dans le chapitre suivant à la gestion des risques et de la prise de risque.

CHAPITRE 10

GESTION DES RISQUES ET DE LA PRISE DE RISQUES

La gestion des risques dans l'informatique de gestion traditionnelle est plutôt bien maîtrisée. En revanche, pour tous les domaines plus récents, comme le devops, l'IoT, le cloud, les systèmes embarqués type smartphone, la gestion des risques est négligée voire maltraitée. Les excuses sont évidentes : l'accélération de l'apparition de nouvelles technologies et de la mise en production des produits, des patchs, des versions et des systèmes empêcheraient les organisations de prendre le recul nécessaire pour la mise en place d'un système de gestion de risque digne de ce nom. Cette gestion inconsidérée pose de nombreux problèmes et des scandales éclatent régulièrement. Ce qui est déjà dangereux dans les systèmes d'information traditionnel peut s'avérer dangereux dès que de l'intelligence artificielle est impliquée car ces systèmes sont plus fragiles. D'après l'ICO, le fait qu'un système d'IA soit plus ou moins risqué dépend des circonstances spécifiques de l'IA comme son objet, son environnement, les technologies employées...

1. Problématiques spécifiques à l'IA : nouveaux risques, nouveaux principes, niveau de risque acceptable

1.1. Nouveaux risques IA ou risques renforcés pour les IA

En raison de l'immaturité de l'IA, il faut bien saisir une nuance importante : à ce stade, il est parfois prématuré de parler de gestion de risque. En effet, la gestion de risque permet de gérer un risque connu. Par exemple, en informatique, il s'agirait de la perte de la connexion internet, d'une intrusion malveillante dans le système, de l'incendie des serveurs, d'une mise à jour d'une brique logicielle incompatible avec le reste du système... Tous ces risques sont connus même si en dresser une liste exhaustive est compliqué et fastidieux. Il est possible pour chacun d'eux d'identifier les facteurs de risques (aussi appelé danger, ce sont les causes potentielles d'apparition du risque), sa gravité ou impact du risque s'il survenait, sa probabilité d'occurrence et son acceptabilité. En IA, il est possible qu'une partie importante des risques soient encore inconnus. Il convient donc d'introduire un concept supplémentaire : celui de la gestion de la prise de risque du fait d'éléments inconnus. Fait-on ou non le choix de prendre un risque inconnu ?

Heureusement, certains risques ont déjà été rencontrés et le pouvoir de notre imagination nous permet d'en envisager d'autres. Nous ne parlerons ici que des risques qui existent uniquement à cause de l'usage de l'IA dans le système ou des risques classiques de l'informatique qui sont radicalement transformés à cause de leur application dans un système contenant de l'IA. Ce sont principalement :

- les risques liés à l'exigence de transparence :
  - **perte de contrôle et du libre arbitre** : Ce risque peut être causé par l'opacité des algorithmes et l'absence de transparence du système, par la mise en place des options par défaut que l'utilisateur

ignore ou n'ose pas toucher... Tous les éléments d'automatisation peuvent conduire à la perte de la maîtrise du système

- **opacité et non-reproductibilité des résultats** : Cela concerne le risque de ne pas comprendre comment un résultat a été produit, par exemple dans le cas des algorithmes basés sur l'apprentissage profond
- **absence de traçabilité :** Avec un manque de traçabilité, en cas de problème, on peut ne pas comprendre les causes et/ou les effets d'une décision. Dans le cas des algorithmes évolutifs, cela peut être de ne pas comprendre pourquoi ou comment ce résultat a évolué par rapport à un test ou résultat précédent

- **les risques liés à l'exigence d'équité** : il s'agit notamment du renforcement de certains biais. Des nuances sont à saisir dans la bonne compréhension de ce risque. Il ne s'agit pas du risque de la présence de biais. En effet, l'existence des biais semble inhérente à l'IA : l'IA catégorise. Or catégoriser c'est simplifier la réalité et simplifier la réalité est biaiser la réalité. Le vrai risque est donc moins la présence de ces biais que leur importance voire leur renforcement et leur potentiel impact néfaste telle une discrimination. En effet, une IA mal construite pourrait accentuer les biais déjà présents dans les données alors que la maîtrise correcte de l'IA permettrait de diminuer ces biais par rapport à un traitement humain classique
- les risques liés à **l'exigence d'humanité** :
  - **fragilité du système** : En la présence d'un évènement exceptionnel et donc probablement jamais rencontré, on ignore comment réagira une IA car l'IA manque du bon sens humain. Par exemple, si un camion est couché en travers de la route, un humain s'arrêtera mais on ignore comment réagira l'IA. Toutes les IA se basent sur des situations passées : comment fera-t-elle pour traiter quelque chose de nouveau ?
  - **non protection des données personnelles, économiques, industrielles**[47] : Les problèmes potentiels à ce niveau sont nombreux : non consentement libre et éclairé, reconstitution de la vie intime des individus ou anonymat dévoilés en croisant les données, délits d'initiés (en 2013, un ordre d'achat d'or passé sept millièmes de secondes avant les autres a révélé l'existence d'un délit d'initié[48] ; à l'heure du trading haute fréquence, quelques millièmes de seconde peuvent se traduire par des gains ou des pertes de milliards de dollars)
  - **enfermement algorithmique** : La connaissance de ce risque est apparue à partir des algorithmes de suggestion d'Amazon, l'un des premiers e-commerçants à les avoir mis en place. Les internautes ne se voyaient plus proposer que des livres ressemblant à leurs sélections précédentes. Le phénomène s'est accru avec la diversification de ce type d'algorithme : séries TV sur Netflix, fil

[47] Il n'existe à l'heure actuelle pas de RGPD pour les données industrielles

[48] https://www.lemonde.fr/economie/article/2013/09/26/la-justice-americaine-et-la-reserve-federale-soupconnent-un-delit-d-initie-2-0-sur-l-or_3484876_3234.html

d'actualité sur Facebook et sur les réseaux sociaux en général. La raison d la création de ces algorithmes est compréhensible : face à la multitude de contenus (culturel ou d'actualité) qui existe, on ne peut pas tout consulter. L'algorithme propose donc uniquement ce qu'il estime qui plaira à l'utilisateur. Mais cet usage empêche l'utilisateur de découvrir d'autres contenus différents de ce qu'il a l'habitude de consulter. Non seulement cela contrevient au droit de libre accès à l'information mais cela peut polariser les opinions des utilisateurs en leur occultant les autres points de vue. Cela peut à terme poser des problèmes sociétaux majeurs avec la montée des extrémismes, des complotistes ou des fausses vérités en tous genres

- les autres risques :
  - **imputabilité** : à qui s'adresser en cas de problème ou de recours ? Le risque est que l'utilisateur se retrouve isolé face à un système d'IA
  - **dépendance** : une plus grande dépendance envers les données, les technologies, la puissance de traitement… Ce qui peut provoquer des distorsions dans les pratiques de concurrence avec l'émergence d'oligopoles comme les GAFAM
  - **déresponsabilisation des différentes parties prenantes** : avec la montée de l'utilisation de l'IA pour l'aide à la prise de décision, la responsabilité de l'humain est diluée. De même, si la prise de décision est automatique, quelle peut est la place de l'humain ? Où et comment peut-on reprendre le contrôle manuellement ?

Ces risques nombreux et variés peuvent être présents au moment de la mise sur le marché des systèmes d'IA ou résulter de mises à jour logicielles ou être induits par l'autoapprentissage en cours d'utilisation du produit. Comment maîtriser l'évolution des risques ou leur apparition spontanée ?

À cela s'ajoute le fait que certains risques informatiques classiques sont amplifiés lorsque de l'intelligence artificielle est introduite dans les systèmes. Ça a été le cas pour Parcoursup où des défaillances humaines – par des gestionnaires qui n'avaient pas assisté aux formations – ont provoqué des défaillances dans le processus de vérification des paramètres. Résultat : il y a quelques années, 50 000 candidats ont reçu des messages d'acceptation à des formations puis des démentis quelques jours ou semaines après[49]. Ce cas illustre aussi la spécificité des risques en fonction du système d'IA impliqué. Dans ce cas de Parcoursup, le risque a pu être identifié par le Comité éthique et scientifique de Parcoursup (CESP). Il provenait *« à la fois de défaillances humaines, et de défaillances dans le processus de vérification des paramètres, imputables à des retards de saisie et à la faiblesse des moyens humains centraux de contrôle »*. Dans les faits, il n'aurait

[49] Certains utilisateurs n'avaient pas suivi les formations ou mal compris les notions d'appel sur capacité d'accueil et d'appel sur rang limite. Le 16 mai 2019, plus de 50 000 propositions ont été envoyées aux candidats sans assurance de pouvoir les accueillir. Des candidats ont ainsi reçu des messages d'acceptation puis de démenti. Cet incident, relayé par les médias, fut finalement sans conséquences car, au 30 mai, tous les candidats concernés avaient reçu une proposition satisfaisant leurs vœux.

pas pu exister sans l'existence d'un système de surbooking des candidats aux formations. Ce risque ne peut, par essence, pas être présent dans une IA de diagnostic médical.

L'IA présente encore une autre particularité, que l'on pourrait appeler le « métarisque »[50]. C'est par exemple le risque qui découle du fait d'expliquer ou non les risques issus des décisions d'IA. Ne pas rendre explicables les IA pourrait poser des problèmes législatifs, des dommages à la réputation de l'organisation et un désengagement du public. Dans le même temps, si l'IA est mal ou trop expliquée, l'organisation se livre à un problème de dévoilement du secret des affaires, des algorithmes, des données ou des types de données utilisées, à être plus faible face à des pirates informatiques ou à des trolls – plus potaches dans l'esprit que les pirates mais pouvant causer tout autant de dégâts.

1.2. Nouveaux principes ou principes renforcés pour les IA

Ainsi, l'IA présente de nouveaux risques qui peuvent contrevenir au but d'obtenir une IA digne de confiance. Heureusement, ces risques peuvent être minimisés par le respect de certains principes qui se traduiront par la mise en place de procédures ou de différents outils. Tout comme pour les risques et les exigences, ces principes peuvent être nouveaux, car spécifiques à l'IA, ou être déjà existants dans les systèmes d'informations traditionnels mais prenant une nouvelle dimension ou intensité. Ces principes sont les suivants :

- **minimisation des risques** : ce principe consiste à maintenir le nombre et le niveau des risques aussi bas que possible. Toutefois, l'existence de ce principe présente un revers : il pourrait avoir tendance à enjoindre les chercheurs à ne prendre aucun risque, ce qui pourrait limiter l'innovation
- **précaution** : il est parfois délicat à cerner car il est souvent confondu avec la prévention des risques. Dans la prévention des risques, on met en place des actions pour éviter un problème puis des éléments de protection au cas où le risque arrive quand même. Toujours dans la prévention des risques, on peut choisir de transférer le risque à une assurance ou bien ne rien faire dans le cas de risques de faible importance et/ou de faible fréquence. Cela prévient des risques qui sont avérés. Le principe de précaution porte sur les risques non avérés. Or tant que nous ignorons l'existence de ce risque, il n'est pas possible de faire de la prévention. L'application du principe de précaution, qui rappelons-le est présente dans la Constitution, est donc difficile : comme dans le principe de minimisation des risques, le revers de ce principe est d'abandonner l'innovation car le risque de faire pourrait être plus important que le risque de ne pas faire. Pour donner un exemple : dans le cas du Covid, le laboratoire pharmaceutique s'est fait attaquer en justice par le premier patient décédé aux suites de l'injection du

[50] Ce néologisme est basé sur le principe de la méta-analyse qui correspond à un niveau d'analyse supplémentaire puisqu'il s'agit d'un travail de recherche qui analyse d'autres travaux de recherche sur un thème commun. Le métarisque peut se traduire par exemple par un niveau de risque supplémentaire qui découle de la prise en compte des différents risques.

vaccin. Les créateurs d'IA ou tout acteur de l'écosystème pourraient craindre une telle action et renoncer à créer une IA

- **non-malfaisance** : « d'abord ne pas nuire »[51]. Ce postulat rappelle qu'il est parfois préférable de « ne pas faire » que de nuire. Là encore, s'il permet d'éviter un grand nombre de risques, il pose le problème de décourager à la prise de risque nécessaire pour créer des choses nouvelles
- **vigilance** : ce précepte a pour but de conserver un esprit de doute. Puisqu'on ne sait pas, et qu'on se force à se souvenir qu'on ne sait pas, on se contraint à une obligation de moyens renforcés et on va mettre en place des outils spécifiques. Ces derniers peuvent par exemple, être l'évaluation régulière du fonctionnement et des résultats d'un algorithme, que celui-ci soit prédictif ou non
- **proportionnalité** : ce principe qui se retrouve dans la RPGD souligne qu'un traitement de données ne doit pas excéder ce qui est nécessaire pour atteindre les objectifs visés. Il convient d'abord de vérifier la légitimité du but. Par exemple, dans une IA pour une société de cours du soir qui devrait faire apparier des enseignants à des élèves mineurs, il est légitime de collecter des données sur les compétences des tuteurs potentiels. Il faut vérifier ensuite quels sont les moyens nécessaires à ce but. En revanche, dans une IA de reconnaissance faciale pour détecter des personnes recherchées, il est excessif d'aspirer les profils des réseaux sociaux pour les comparer aux visages de la foule – tel que le réalise la société Clearview AI[52]. Il faut ensuite envisager l'équilibre des intérêts : si l'intérêt est très grand, peut-on utiliser des moyens un peu plus importants ? Pour rester sur le cas de la reconnaissance faciale, les eurodéputés ont voté pour demander un moratoire sur le déploiement de ces systèmes, jugés disproportionnés, sauf dans le cas de l'identification des victimes de criminalité, comme les kidnappings
- **alerte** : il faut que les lanceurs d'alerte puissent signaler les IA représentant une menace ou un préjudice grave pour l'intérêt général. Cela permettrait de prévenir ou de corriger des effets particulièrement néfastes provenant de dysfonctionnements importants d'une IA, par exemple, sur les utilisations inéquitables de certains algorithmes qui créeraient de sérieuses discriminations. Il devrait être possible de créer un cadre réglementaire d'irresponsabilité pénale similaire à celui prévu pour les lanceurs d'alerte de la loi Sapin II[53].

---

[51] Principe appliqué par le Parlement européen, première institution à avoir tenté de légiférer sur les questions juridiques et éthiques relatives à l'intelligence artificielle, à travers une Résolution du 16 février 2017 relative aux règles de droit civil sur la robotique. Pour une analyse, voir A. Bensoussan et J. Bensoussan, « IA robots et droit », Editions Larcier, juillet 2019.

[52] ClearView AI est une entreprise américaine spécialisée dans la reconnaissance faciale. A. Bensoussan, « Vie privée : pourquoi l'application Clearview ne devrait pas arriver en France », L'Express.fr du 22 janvier 2020.

[53] Pour rappel, l'article 122-9 du Code pénal prévoit une immunité pénale pour les lanceurs d'alerte qui portent atteinte à un secret protégé par la loi, dès lors que cette divulgation est nécessaire et proportionnée à la sauvegarde des intérêts en cause, qu'elle intervient dans le respect des procédures de signalement définies par la loi et que la personne répond aux critères légaux de définition du lanceur d'alerte de la loi Sapin II.

1.3. Niveau de risque tolérable et acceptable

Si les principes précédents ont pour but de faire prendre conscience des risques et même d'inciter les acteurs de l'IA à les prendre en compte pour en contrôler les travers, plusieurs questions subsistent avant d'obtenir la maîtrise de ces risques. Comment :

- évaluer le niveau de risques ? Nous manquons actuellement d'outils pour déterminer si un algorithme est plus ou moins opaque, si le système d'intelligence artificielle est plus ou moins fragile. Ceci concerne presque tous les risques présentés dans ce chapitre. Des équipes de recherche travaillent activement à la constitution de ces outils mais pour l'instant ces derniers ne s'appliquent que dans certains cas et sont encore peu répandus lorsqu'ils sont matures
- établir le niveau d'appétit au risque ? Les différents acteurs auront plus ou moins d'aversions aux risques. Certains paraîtront acceptables pour les créateurs alors qu'ils ne le seront pas pour les utilisateurs ou vice-versa. Il faut donc parvenir à concilier tous les intérêts et déterminer ce qu'on est prêt à accepter comme niveau de risque ?
- trouver l'équilibre entre d'une part, les principes prudents que sont les principes de minimisation, précaution et non-malfaisance et d'autre part, l'innovation et la volonté de créer quelque chose de nouveau ? Si des principes sont érigés comme une obligation, cela détruira toute innovation. A contrario, privilégier à tout prix l'innovation peut amener à prendre des risques intolérables.

Le niveau de risque d'un projet peut toujours être réduit. Toutefois, plus on cherche à réduire un risque, plus les investissements nécessaires sont conséquents et ce, de manière souvent exponentielle. Nous voyons là un autre équilibre à trouver : le niveau de risque acceptable pour toutes les parties prenantes contre le niveau d'investissement soutenable.

Il est une autre composante à prendre en compte concernant le niveau de risque tolérable. Il s'agit de la gestion de la perception du public. L'Internet étant désormais entré dans les mœurs, on sait que la diffusion des informations actuelle diffère d'il y a quelques décennies : la baisse des revenus des médias classiques les contraint à privilégier le scoop à tout prix tandis qu'avec les réseaux sociaux chacun peut propager facilement son opinion. Le risque de fakenews et de déformation de la réalité est un élément à prendre en compte.

C'est ce qui est arrivé avec Parcoursup où les journalistes se sont empressés de réaliser du cherry picking (littéralement cueillette de cerises). Cette pratique consiste à choisir uniquement les extraits d'un rapport ou d'un document qui vont dans le sens de la thèse soutenue par le journaliste. Ainsi, la presse s'est empressée de dénoncer dans Parcoursup les éléments qui étaient justement démentis dans les rapports publics. L'IA étant un sujet sensible, ce risque est plus prégnant que pour les technologies anciennes qui n'intéressent pas seulement un public spécialisé.

2. Quelques recommandations : démarche spécifique, modélisation des risques

La gestion des risques et de la prise de risque concernant les systèmes d'IA diffère sensiblement de son équivalent informatique n'intégrant pas d'IA. Ainsi, il est nécessaire de mettre en place une démarche spécifique de la gestion de risque et de prise de risque pour les IA et de modéliser les risques de manière pertinente.

2.1. Démarche spécifique pour les IA : la gestion de la prise de risque

Nous avons vu que le terme « IA » désigne en réalité un ensemble d'objets très différents. Entre les différents algorithmes créés, différentes techniques employées, différentes solutions mises en place, les impacts potentiels de l'usage des IA, des acteurs impliqués... Chacun de ces objets présente ses propres spécificités et donc, ses propres risques qui peuvent être différents des risques d'un autre algorithme, d'une autre solution, d'un autre objet. L'une des premières choses à faire pour mettre en place une gestion de la prise de risques et des risques devrait donc être de catégoriser les types d'IA puis de déterminer les risques spécifiques à chacune de ces catégories. Cela permettra de déterminer les actes préventifs et les traitements de risques à mettre en place en fonction de chacune de ces catégories.

Si la démarche de gestion de la prise de risque est spécifique aux types de risques à prendre, des constantes seront nécessaires pour toutes ces démarches. Elles devront être :

- simples
- souples et flexibles
- faciles à mettre en œuvre, à maintenir et à comprendre
- alignées avec les référentiels internationaux

De façon à favoriser une large adhésion. Il faut que la mise en place de ces démarches de gestion de la prise de risque soit intéressante et profitable au plus grand nombre de personnes possibles. Plus de personnes y souscriront, plus ces démarches seront efficaces.

Ces démarches pourront être testées via des exercices dits de « Red Teaming » comme ce qui est fait en cybersécurité. Des équipes spécialisées seront constituées pour prendre le rôle d'attaquant et tenter de percer les systèmes, d'explorer les risques et les vulnérabilités spécifiques aux IA. Idéalement ces équipes mettraient en commun le résultat de leurs recherches via des partages de bases de type d'attaques, d'incidents... afin que les équipes de développement puissent prendre en compte ces vulnérabilités.

2.2. Modélisation des risques dans un contexte d'IA

Si nous avons vu plus haut qu'il est nécessaire de mettre en place des démarches, celles-ci sont inapplicables tant qu'elles ne sont pas construites sur des modèles fiables et pertinents. En effet, une meilleure compréhension des phénomènes complexes permettra une meilleure appréhension des risques et des solutions correspondantes. Pour cela, il faudra créer des catégorisations. On peut songer à étudier différentes catégorisations telles que par types d'IA, d'algorithmes, de techniques d'apprentissage, de raisonnements, de données, de services, d'outils... Mais il faudra bien se souvenir que catégoriser des objets induit par définition des biais. Il faudra aussi bien garder en mémoire que la gestion des risques et la gestion de la prise de risques ne pourra être pleinement efficace que si ces biais de catégorisation sont eux-mêmes pris en compte.

Il faudra également faire attention aux évolutions possibles de ces modèles en fonction d'éventuels changements des réglementations, des technologies, des comportements, des solutions...

La gestion de risque dans le domaine des systèmes d'information de gestion est relativement bien maîtrisée et mise en œuvre pour tous les systèmes traditionnels qui sont connus et traités depuis longtemps. Ce n'est pas le cas pour tous les systèmes nouveaux dont les systèmes à base d'IA. La gestion de risque doit prendre en compte les risques associés aux exigences traditionnelles qui sont différents dans le cadre de l'intelligence artificielle, comme la cybersécurité ou la vie privé mais aussi les risques qui sont apparus suite à l'apparition des exigences spécifiques telles l'humanité, la transparence et l'équité. Il faut également prendre en compte la gestion de la prise de risque, c'est-à-dire le risque lié à des facteurs encore inconnus en raison de la jeunesse des technologies.

Plusieurs principes à respecter pour les systèmes à base d'IA ont été identifiés. Il s'agit par exemple des principes de minimisation des risques, de précaution, de non-malfaisance, de vigilance, de proportionnalité et d'alerte.

Aboutir à la maîtrise de ces risques nécessitera de répondre à plusieurs interrogations essentielles comme l'évaluation du niveau de risque, définir le niveau d'aversion au risque et trouver l'équilibre entre la prudence et l'innovation. Nous avons évoqué quelques recommandations à commencer par la création d'une démarche spécifique à l'IA de gestion des risques et de leur modélisation.

CHAPITRE 10

Nous avons vu que l'IA présente des différences importantes par rapport à des systèmes d'information plus traditionnels. Pour que l'IA soit acceptée à grande échelle par la société et puisse tenir ses promesses, il faut qu'elle devienne digne de confiance. Pour cela, le groupe de travail a identifié une démarche en cinq étapes. Nous avons parcouru les deux premières : identifier les principes à suivre et les exigences à prendre (nous avons notamment mis en avant les exigences de transparence et d'explicabilité, d'équité et non-discrimination, et d'humanité) et mettre en place une gestion de la prise de risque adaptée. Attardons-nous donc sur la troisième qui correspond à la mise en place d'une gouvernance et de la responsabilité des parties prenantes.

CHAPITRE 11

GOUVERNANCE ET RESPONSABILITÉS

Lorsqu'il est seul, un jardinier peut aisément décider ce qu'il va planter et organiser son plan à sa façon. Mais que faire si on instruit à une équipe de jardinier de créer le plus beau jardin sans plus de précisions ? L'un considérera qu'il faudrait faire un jardin paysager, l'autre un jardin à la française… Et à une échelle plus grande, que faire lorsqu'il y a également un voisin qui veut planter un séquoia, un autre enlever la haie pour gagner de l'espace, un troisième qui a besoin de creuser une tranchée pour passer un conduit ? Les besoins et envies de chacun peuvent facilement empiéter sur celles des autres.

La gouvernance, qui met en œuvre l'ensemble des dispositifs d'une organisation pour déterminer les contributions de valeur attendues pour les parties prenantes et fixer les objectifs correspondants, est la solution à ce problème. Nous verrons dans ce chapitre quelles sont les problématiques de gouvernances spécifiques aux IA, les décisions qu'il reste à prendre, l'impact du grand nombre de parties prenantes sur sa mise en place puis nous donnerons quelques recommandations à suivre pour aboutir à une gouvernance spécifique et transparente.

1. Problématiques spécifiques à l'IA : nombreuses décisions, nombreux acteurs, complexité et opacité

1.1. De nombreuses décisions en matière d'IA à prendre

Lorsqu'il est devenu évident que les données sur support numérique étaient sources d'une part importante de la valeur ajoutée d'une entreprise, il est aussi devenu évident qu'il fallait mieux les maîtriser, depuis leur acquisition jusqu'à leur destruction. La gouvernance des informations s'est petit à petit mise en place au sein des organisations afin d'optimiser leur valeur et l'efficience de leur traitement. Il semble désormais évident que l'IA devra suivre un chemin similaire, à la nuance près que la création de valeur des IA provient non seulement de l'utilisant des données mais aussi des deux autres piliers évoqués au chapitre 1, la puissance de traitement et de communication et les techniques, logiciels et technologies spécifiques à l'apprentissage. Nous avons noté que ces piliers dans le contexte des IA apportaient de nouvelles exigences, de nouveaux risques et de nouveaux outils et dispositifs pour les traiter. Sa portée est donc plus grande et sa mise en œuvre et sa maîtrise plus complexe.

La première question à résoudre sera le choix du dispositif de gouvernance. L'objectif de la gouvernance de l'IA sera donc d'instaurer un cadre connu et accepté au sein de l'organisation, pour l'identification et l'évaluation des grandes options de création de valeur des IA (quelles exigences, quels risques et quels efforts), pour l'arbitrage entre ces options et pour leur suivi. Quelques questions qui se posent alors sont : qui décide, comment, qui est responsable…

La gouvernance aura pour rôle de définir les objectifs de l'organisation envers l'IA. Ces derniers sont nombreux et ils dépendent fortement des choix effectués concernant ses composants (quelles données, quelles technologies, quelles techniques d'apprentissage, quels outils...). Tout d'abord, quelle est la valeur apportée par l'usage de l'IA dans l'organisation ? Quels engagements l'organisation doit-elle prendre à la suite de son utilisation de l'IA ? Quelles sont les exigences qu'elle veut satisfaire ? Sur chacune de ces exigences, que va-t-elle chercher à atteindre, jusqu'à quel niveau ? Quels sont les risques acceptables ? Par exemple, si l'organisation désire s'engager sur la transparence de son système d'IA, quel niveau de transparence va-t-elle chercher à atteindre ? Cela pourrait être par exemple : « nous sommes capables de suivre une donnée de son entrée jusqu'à son traitement et d'expliquer le fonctionnement du traitement en question ». Où placer la limite entre la transparence et la protection des données personnelles et le secret des affaires ? Il conviendra également de mettre en place une gestion des risques et de la prise de risque pour chacune de ces exigences.

La gouvernance devrait pouvoir fixer l'objectif à atteindre entre deux exigences ou impératifs contradictoires. Par exemple, l'organisation devrait-elle favoriser l'innovation ou la diminution de la prise de risque ? L'organisation devrait-elle privilégier la grande exactitude des résultats au détriment de l'explicabilité du fonctionnement des algorithmes ? Ou bien devrait-elle préférer la diminution des biais ou la diminution des performances ? Ainsi, la gouvernance devra également déterminer les niveaux de confiance souhaités, les solutions techniques à privilégier ou à limiter, les contrats d'assurance à prendre, les principes éthiques à respecter...

Il faudra également aligner les structures internes, les responsabilités, les exigences de formation, les politiques et les incitations à la stratégie globale de gouvernance de l'IA et de gestion des risques.

Il sera important de proportionner la gouvernance et la gestion des risques à mesure de l'utilisation de l'IA par les organisations ; l'investissement de ressources et d'efforts requis étant, comme pour la gouvernance des données, souvent sous-estimé. Cela est d'autant plus vrai que le domaine est en pleine évolution et que tous les outils ne sont pas encore disponibles.

1.2. De nombreux acteurs IA impliqués dans un contexte de complexité et d'opacité

La gouvernance, puisqu'elle impose les orientations à prendre, est très liée à la notion de responsabilité. Cette notion garantit notamment qu'une personne ayant subi un préjudice ou un dommage avéré est en droit de demander et de recevoir une indemnisation de la partie responsable. Par conséquent, cela incite économiquement les personnes physiques et morales à éviter de causer un préjudice ou un dommage. Toutefois, l'IA présente de grandes différences par rapport aux cadres habituels de responsabilité.

De nombreux acteurs

En premier lieu, l'usage d'un système d'IA implique immédiatement de très nombreux acteurs tout au long du cycle de vie de ce système. Ils sont entre autres :

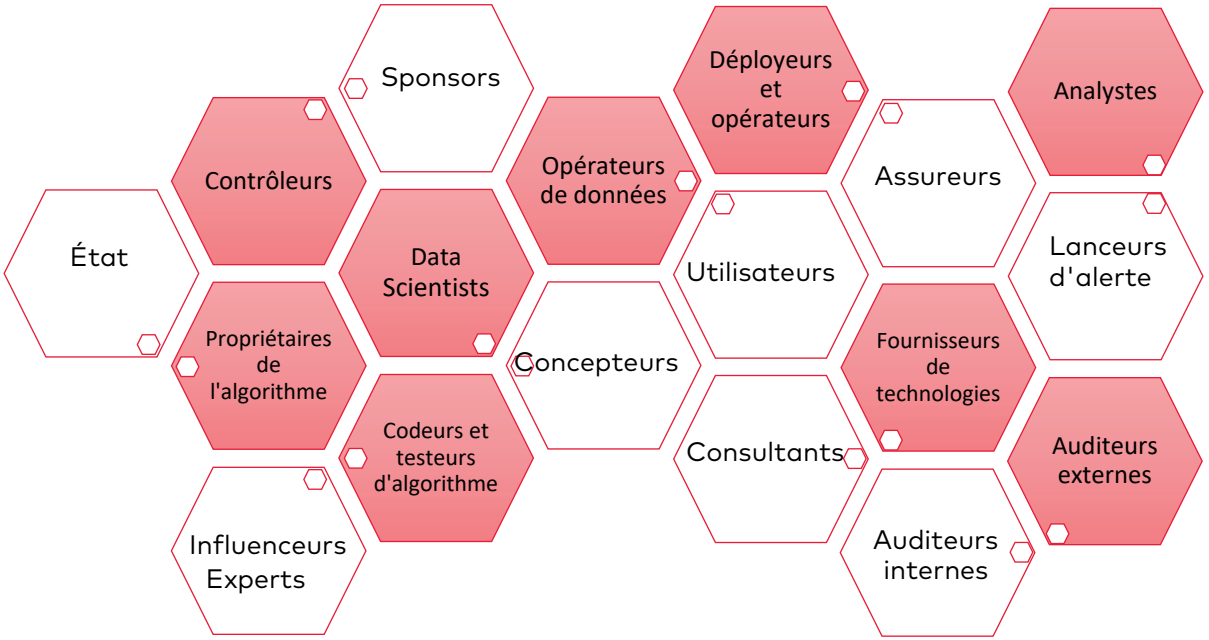

Figure 12 : liste de quelques parties prenantes d'un système d'information à base d'intelligence artificielle

Complexité, opacité et difficulté d'apporter des éléments probants

Autre difficulté : la complexité des systèmes. Un système d'IA utilise de nombreuses briques logicielles d'origines diverses, nécessite de très nombreuses données le plus souvent issues de sources variées, met en œuvre des algorithmes très complexes… et il est souvent interconnecté à d'autres systèmes informatiques dont certains peuvent eux-mêmes être des systèmes utilisant de l'IA. Si les principes et exigences ne sont pas respectés, il en ressort un aspect diffus, notamment si le système d'IA ne dispose pas de système de traçabilité des actions ou des résultats aux sources. Cet état est encore aggravé par l'apprentissage automatique du système et l'opacité des systèmes.

Dans ces circonstances, il serait extrêmement difficile, voire parfois impossible, de déterminer qui contrôle le risque associé à l'utilisation d'un système d'IA et d'apporter des éléments probants en cas de requête. On pourrait se demander, par exemple, comment justifier qu'on a pris une décision valable. Ces incertitudes impliquent une insécurité juridique certaine par toutes les parties. Le système d'IA devrait donc lui-même être conçu de manière à soutenir la gouvernance. On pourrait songer, par exemple, à la création d'une matrice des prises de responsabilités de type RACI (qui est Responsable, qui doit rendre compte (Accountable en anglais), qui doit être Consulté et qui doit être Informé) en fonction du rôle de la partie prenante et de la décision prise.

Plusieurs types de responsabilités : civil, pénal...

Mais il serait nécessaire de songer à réduire l'insécurité juridique via des réflexions de plus large envergure. Il faudrait étudier la nature des responsabilités des différentes parties prenantes et leur étendue. Telle partie pourrait-elle être civilement responsable ? Pénalement ? Devrait-elle souscrire à une assurance d'utilisation de l'IA comme on pourrait souscrire à une assurance automobile ? Telle partie aurait-elle une obligation de moyens ou de résultats ? Qui doit apporter la preuve en cas de litiges et quel type de preuve (sauvegarde, rétention) ?

La reprise en mains des IA par l'humain conduit à de nouvelles problématiques de responsabilité

Les caractéristiques de l'IA amènent également de nouveaux champs de réflexion. Nous avions vu au paragraphe 6.6.3 le besoin d'un « bouton rouge » grâce à laquelle l'humain pourrait reprendre la main sur un système d'IA. Toutefois, que se passerait-il – notamment en termes de responsabilités – si les décisions proposées par l'IA sont considérées comme plus fiables ? Les IA ne présentent pas toutes des performances supérieures aux humains mais pour un grand nombre, cela ne saurait tarder. C'est d'ailleurs l'un des objectifs de la voiture autonome. Si l'on n'espère pas que les voitures autonomes mèneraient à moins d'accidents que la conduite manuelle, quel serait l'intérêt de les développer ? Le jour où la voiture autonome sera plus fiable, qu'un humain décide malgré tout de reprendre la main et qu'un accident se produit, qui serait responsable de la perte de chance de survie ? La même question se posera dans le cas où l'IA d'analyse d'image médicale fournit un diagnostic que le médecin décide de ne pas prendre en compte. Qui serait responsable de la perte de chance de guérir ?

2. Quelques recommandations : dispositif spécifique, transparence

2.1. Dispositif de gouvernance spécifique pour les IA

Au niveau de chaque IA

L'une des premières actions à mener consisterait donc à mettre en place une matrice de responsabilité de type RACI. Cette démarche est d'une telle évidence que la Commission européenne a commencé à travailler sur le sujet, à définir les missions et les rôles de chacune des parties prenantes[54]. Il faudrait pour cela également identifier et lister les « instances » qui seraient nécessaires à la gouvernance de l'IA.

Certains semblent évidents, comme une instance qui traite des options technologiques et aux orientations, mais on pourrait aussi songer à une instance d'arbitrage, une autre d'incidents mais aussi de conformité, de communication… Chaque décision significative ou dilemme important que présente l'IA pourrait théoriquement faire l'objet d'une instance de gouvernance, tout en faisant attention qu'il n'y ait pas de doublon soit avec les structures existantes soit avec les instances entre elles.

Au niveau institutionnel

Ces instances pourraient se trouver au niveau micro (celui de l'organisation) mais aussi à des niveaux plus généraux. Il semblerait nécessaire de mettre en place une structure de gouvernance européenne qui assurerait la coopération des États membres pour éviter la dilution des responsabilités entre les autorités nationales ; et ce d'autant plus que le même système d'IA pourrait être utilisé et affecter plusieurs pays simultanément. Cette gouvernance qui pourrait être nationale, internationale (comme avec l'Europe) ou mondiale (avec l'ONU ou l'ISO) devrait être en mesure d'effectuer des tâches à grande échelle comme étudier l'incidence de l'IA sur la société, offrir un cadre de discussion sur le sujet de l'IA, mettre en place des bonnes pratiques, d'identification des tendances émergentes. Cette structure pourrait aussi jouer un rôle en ce qui concerne la normalisation et la certification des produits mais le niveau d'implication sur ce sujet reste à déterminer : s'agirait-il juste de donner des conseils ou bien carrément de certifier et tester les produits et services reposant sur l'IA, comme le suggère la Commission européenne dans son Livre Blanc sur le sujet de l'IA ? On pourrait imaginer l'existence de centres d'essais indépendants pour faire les vérifications et les évaluations des IA. Cet organisme pourrait potentiellement disposer de certains pouvoirs qui lui permettraient aussi d'enquêter sur des cas individuels.

---

[54] Proposition de règlement sur l'IA établissant des règles harmonisées pour l'UE (loi sur l'intelligence artificielle), 21 avril 2021. Pour une étude, A. Bensoussan, « Vers une harmonisation des législations européennes en matière d'intelligence artificielle », Télécom revue #203, Décembre 2021, disponible sur https://www.alain-bensoussan.com/wp-content/uploads/2022/01/TPA203.pdf

Il ne faudrait toutefois pas sous-estimer l'importance des instances nationales qui devraient être mises en place par les Etats membres. Par une proximité supérieure avec le terrain, on pourrait les charger d'un rôle d'évaluation et de contrôle de la conformité du développement, du déploiement et de l'utilisation de l'IA, de la robotique et des technologies connexes à haut risque.

On peut aussi penser que des instances professionnelles, comme il en existe pour le domaine financier, composées des différentes parties prenantes, puissent jouer un rôle dans l'établissement des référentiels des bonnes pratiques pour les IA et dans leur certification.

2.2. Transparence du dispositif de gouvernance mis en place pour les IA

Toutefois la mise en place de telles structures n'aurait que peu d'intérêt sans transparence de la part des systèmes de gouvernance. Cela manquerait même carrément de logique : quel intérêt d'exiger de la transparence et de l'explicabilité de la part des systèmes d'IA si aucune communication sur le sujet n'existe ? Il semble donc nécessaire que chacun de ces dispositifs de gouvernances (instances techniques, instances de pilotage, groupes de suivi au niveau de l'organisation, au niveau national ou international...) produisent des rapports sur leurs activités. La forme des rapports serait encore à définir mais on peut déjà imaginer avoir besoin d'un rapport sur la gouvernance et la gestion de la prise de risques de la part du responsable de l'IA comme il en existe dans le domaine financier. De tels rapports permettraient aussi d'expliciter la matrice des responsabilités de type RACI et de dire noir sur blanc quelles sont les orientations principales décidées par les instances. Par exemple, on pourrait savoir quelle est la position de telle ou telle organisation sur la transparence ou l'humanité des systèmes d'IA construits et/ou utilisés en son sein.

On pourrait aussi imaginer d'aller plus loin qu'une gouvernance simplement transparente en allant vers le concept de gouvernance participative – on pourrait dire « gouvernance 2.0 » – où le plus d'acteurs de l'écosystème d'IA seraient impliqués : organisations de consommateurs, partenaires sociaux, entreprises, chercheurs et organisations de la société civile.

Les systèmes d'informations à base d'intelligence artificielle soulèvent un grand nombre de questions et de dilemmes qu'il faudra arbitrer. La mise en place d'un dispositif de gouvernance reconnu est donc impérative. Il aura pour rôle de définir les objectifs de l'organisation envers les IA. Il faudra définir ses rôles, ses pouvoirs, ses organes de décisions... Les dispositifs de gouvernance devraient être présents au sein des organisations mais aussi à l'échelle institutionnelle. Cela permettrait de définir les responsabilités civiles, pénales... des nombreuses parties prenantes et de déterminer comment chaque acteur pourrait les consulter.

Le groupe de travail a identifié une démarche en cinq étapes devant permettre le déploiement d'IA dignes de confiance. Nous avons parcouru les trois premières : identifier les principes à suivre et les exigences à atteindre, mettre en place une gestion de la prise de risque adaptée et définir la gouvernance et la responsabilité des parties prenantes. Attardons-nous donc sur la quatrième étape qui correspond à la mise en œuvre des outils et des bonnes pratiques.

CHAPITRE 12

OUTILS ET BONNES PRATIQUES

Il est difficile de tailler un rosier avec un couteau de cuisine, il faut adapter l'outil à notre besoin pour atteindre nos objectifs. L'expérience a montré que le sécateur est l'un des meilleurs outils pour entretenir un rosier. Mais elle a aussi montré qu'il faut éviter d'arroser les feuilles, qu'il est conseillé d'arquer les branches vers le bas, de pailler, de corriger le pH du sol... Toutes ces bonnes pratiques sont apparues parce que les horticulteurs ont testé différents outils, différentes procédures, différentes méthodes dans le but d'avoir les plus belles fleurs.

Le principe est identique dans le cas des systèmes d'informations à base d'intelligence artificielle. Il faut adapter les outils utilisés et faire émerger les meilleures pratiques qui permettent d'atteindre nos exigences et de suivre les principes qui ont été définis, comme le respect des règles de l'éthique et les droits fondamentaux.

Nous verrons d'abord pourquoi il n'est pas possible de travailler uniquement avec les exigences et les principes. Ensuite, nous verrons pourquoi les bonnes pratiques restent en grande partie à définir et lesquelles peuvent être appliquées immédiatement.

1. Problématiques spécifiques à l'IA : traduction de principes en bonnes pratiques, pratiques SI, spécificités IA, référentiels

Les principes éthiques n'ont pas de valeur s'ils ne sont pas traduits en mécanismes

Un certain nombre de problèmes, de craintes et d'inquiétudes ont été soulevés à la suite de l'apparition de l'IA. Nous avons vu que beaucoup de ces préoccupations se lèveront à partir du moment où les lois et les réglementations adéquates seront mises en place. Seulement il est nécessaire de prendre du temps afin d'analyser la situation et élaborer le meilleur cadre juridique possible. Or, on constate un décalage de plus en plus important entre le développement de nouvelles technologies et le droit (internationale ou local). Les chercheurs en IA et les entreprises technologiques ont rapidement pris la mesure des craintes de l'opinion publique et ont souvent décidé d'anticiper les problèmes et de commencer à les résoudre. Les entreprises technologiques ont souvent mis en place des groupes de réflexion spécifiques («Thinktank») visant à établir des principes éthiques à suivre.

C'est le cas par exemple de SAGE avec son document « The ethics of Code » en 2017, de Microsoft avec son document « Microsoft AI Principles » en 2018, de Google avec «AI at Google : Our principles » en 2018 ou l'un des plus connus : les principes d'Asilomar. Il faut cependant se souvenir que ces démarches volontaires n'ont rien de coercitif. Rien n'empêche les entreprises qui ont énoncé ces principes d'y déroger.

Par ailleurs, ces principes éthiques ne sont pas facilement transposables en préconisations techniques. Par exemple, le onzième principe d'Asilomar sur les valeurs humaines énonce que les systèmes d'IA devraient être conçus et maintenus de façon à être compatibles avec les idéaux de dignité humaine, les droits, les libertés et la diversité culturelle. Comment mettre cela en pratique dans un système d'IA de suggestion de

films à regarder ? Quel algorithme choisir pour respecter ces principes fondamentaux dans un produit d'analyse faciale pour la sécurité publique ? Quand bien même, si les équipes de développement réussissent à trouver une solution, comment faire pour le prouver à un évaluateur extérieur ?

Il n'est pas impossible que, à la manière du greenwashing, certaines organisations effectuent des manœuvres qualifiables de « good AI washing ». Pour que les créateurs d'IA gagnent la confiance des autres parties prenantes, il semblerait judicieux de mentionner les principes suivis mais surtout de mettre en avant les mécanismes permettant leur mise en place et tout en démontrant leurs comportements responsables.

Les meilleurs mécanismes deviendront des bonnes pratiques

Le principe derrière ces mécanismes est simple : les organisations posent des exigences à atteindre. Des risques apparaissent alors, qui peuvent menacer l'atteinte de ces exigences. La mise en place de dispositifs permet de limiter les risques. Ces dispositifs seront déployés et testés par les organismes. Lorsque l'écosystème aura un retour d'expérience solide sur les dispositifs qui donnent de bons résultats, ils pourront être requalifiés en bonnes pratiques et recommandés en modèles, de façon à les partager. Ainsi, ils fourniront des exemples ou des suggestions que d'autres organisations s'approprieront et adapteront à leurs besoins. Ces nouveaux acteurs pourront adapter et faire évoluer ces bonnes pratiques à leur contexte et leurs spécificités.

Ces bonnes pratiques devront être adaptées à chaque contexte

Il faut donc bien saisir que ces bonnes pratiques sont variables et adaptables suivant les organismes. En effet, chaque bonne pratique est destinée à pallier un niveau de risque lui-même rattaché au niveau d'exigence désirée. Par exemple, si un organisme estime qu'il lui faut un niveau très élevé d'équité et un niveau minimum de transparence, les risques liés à l'atteinte de l'exigence d'équité seront plus forts que ceux liés à la transparence. Les bonnes pratiques pour l'atteinte de l'exigence d'équité indiqueront par exemple d'installer trois types différents de logiciels de contrôle là où un seul suffira pour la transparence. Enfin, on soulignera que les bonnes pratiques peuvent être de différentes natures : mise en œuvre d'un processus de documentation adéquat, mise en place d'outils logiciels spécifiques, création d'instances de vérification...

1.1. Bonnes pratiques des systèmes d'information « traditionnels » pertinentes pour les IA

Certaines bonnes pratiques applicables aux IA existent déjà

Les bonnes pratiques doivent être adaptées au contexte et aux spécificités de l'organisation qui les met en place : en effet, les enjeux pour un ministère dédié au service public sont différents de ceux d'une société privée ou d'une ONG. Pour autant, il existe beaucoup de similarités dans le fonctionnement de certaines organisations ; des référentiels de bonnes pratiques sont connus : il s'agit d'ITIL, CMMI, COBIT, ISO...

Comme l'IA est une composante particulière d'un système d'information, certaines des bonnes pratiques qui ont émergées au fil des années pour les systèmes informatiques classiques pourront être directement transposables aux systèmes d'information à base d'intelligence artificielle. En effet, les principes concernant la gouvernance, le management ou le domaine opérationnel sont sensiblement identiques. Les modèles pour la création de valeur, la gestion de la prise de risque, les scénarios de risque, les analyses d'impact et autres peuvent être repris quasiment tels quels des référentiels informatiques classiques et transposés à un système d'IA.

Des bonnes pratiques pour chacun des types de leviers

C'est également le cas pour les leviers qui influencent la réussite d'un système d'IA.

Un levier est un moyen d'action pour atteindre un but. Plusieurs types de leviers peuvent être mobilisées. Nous avons retenu les 7 types de leviers définis dans COBIT qui peuvent interagir entre eux :

- **les principes, les directives et les cadres de référence** qui représentent le véhicule permettant d'orienter les décisions et pratiques à mettre en œuvre
- **les processus** qui sont des ensembles d'activités corrélées ou en interaction utilisant des éléments en entrée pour produire des résultats escomptés
- **les structures organisationnelles** qui sont les entités clés concourant à la prise de décision et à l'action
- **la culture, l'éthique et le comportement** des individus et d'une organisation qui sont la manière d'être, d'agir ou de réagir face à une situation en fonction du contexte
- **les informations** utilisés et produites par une organisation qui permettent d'agir et prendre des décisions
- **les services, l'infrastructure, les applications et les autres outils** employés par une organisation pour atteindre ses objectifs
- **le personnel, les aptitudes et les compétences** qui sont sollicités dans le cadre des actions et des décisions

Il n'est pas difficile pour le lecteur de voir comment on peut réutiliser ces leviers pour un système d'IA.

Des bonnes pratiques en matière d'indicateurs de résultats et de moyens

C'est aussi le cas de certains indicateurs, comme les indicateurs de moyens, de résultats, les indicateurs fonctionnels comme les délais et les coûts... dont l'utilisation pour un système d'IA le rendra plus qualitatif.

Mais nous n'avons cessé de souligner depuis le début de ce document les spécificités des systèmes d'IA par rapport aux systèmes informatiques traditionnels. Chacune de ces spécificités impliquait l'apparition d'une exigence nouvelle. Chacune de ces exigences amène de nouveaux risques qui doivent être maîtrisés en partie par la mise en place de nouvelles bonnes pratiques.

1.2. Quelles spécificités IA impactant les bonnes pratiques ?

Des bonnes pratiques en cours de maturation

Si on peut reprendre les bonnes pratiques actuelles des systèmes d'information traditionnels pour gérer les points communs dans les systèmes d'IA, des différences existent, comme nous l'avons souligné précédemment. Ces bonnes pratiques spécifiques, dont plusieurs ont été évoquées dans les chapitres précédents, devront être mises en place pour appréhender les caractéristiques spécifiques de l'IA.

L'un des principaux problèmes concernant l'IA concerne l'immaturité de ces technologies. Nous l'avons dit : une bonne pratique est avant tout un dispositif qui a été testé de nombreuses fois dans de nombreuses organisations différentes et qui s'avère efficace pour l'atteinte de certaines exigences. Il faut donc du temps : du temps pour songer à un dispositif, le créer, le mettre en place, le tester et le faire évoluer. Par conséquent, de nombreuses bonnes pratiques liées aux spécificités des IA restent à créer voire à imaginer. Notons cependant que quelques outils et mécanismes sont en cours de création comme ceux destinés à l'explicabilité des algorithmes d'IA, portés par des laboratoires qui recherchent actuellement des solutions en ce sens. C'est le cas également des techniques de vérification formelle qui visent à prouver l'exactitude d'un système logiciel ou matériel par rapport à des propriétés spécifiées, à l'aide de preuves mathématiques notamment pour les IA basées sur le machine learning.

Plusieurs bonnes pratiques s'appuyant sur des outils techniques

À la différence de certaines bonnes pratiques de management, l'aspect scientifique des IA nécessite des efforts bien plus importants. Par exemple, pour la vérification formelle, des chercheurs[55] indiquent que « les défis comprennent :

- la génération des revendications formelles et des preuves correspondantes concernant le comportement des modèles de machine learning, étant donné que leur comportement de sortie peut ne pas toujours être clair ou attendu par rapport aux entrées (par exemple, un modèle ML n'affichera pas nécessairement le même comportement dans le domaine qu'il a présenté sous un environnement de test). En conséquence, les propriétés formelles traditionnelles doivent être repensées et redéveloppées pour les modèles de machine learning
- la difficulté de modéliser correctement certains systèmes de machine learning en tant qu'objets mathématiques, en particulier si leurs blocs de construction ne peuvent pas être formalisés dans les domaines mathématiques utilisés par les techniques de vérification existantes ; et
- la taille des modèles de machine learning réels, qui sont généralement plus grands que ce que les techniques de vérification existantes ne sont capables de vérifier. »

[55] M. Brundage et al., « Toward Trustworthy AI Development : mechanisms for supporting verifiable claims », 2020

Plusieurs bonnes pratiques alternatives avec des niveaux d'efficacité variables

Ainsi, les bonnes techniques spécifiques à l'IA qui ont déjà été développées soit concernent des risques plus simples et faciles à traiter (par exemple les bonnes pratiques concernant la transparence pour une IA « linéaire » où les extrants sont directement liés aux intrants par des algorithmes « simples »), soit sont des alternatives avec un plus faible niveau d'assurance. Nous pouvons par exemple citer les pratiques de vérification empirique qui ont été élaborées comme des pratiques alternatives à la vérification formelle. Elles sont plus pratiques, mais fonctionnent de manière empirique, et par conséquent ne peuvent pas garantir aussi pleinement leurs prétentions.

Or, pour certains secteurs critiques, comme les transports, l'énergie ou la santé, le niveau d'assurance nécessaire est extrêmement élevé, bien plus que pour d'autres domaines. Par exemple, pour un Commissaire aux Comptes, le niveau d'assurance raisonnable attendu est de l'ordre de 95 %. Or si on applique ce même niveau à une IA de voiture autonome par exemple, cela conduirait inévitablement à plusieurs millions de morts, ce qui serait de plus inacceptable pour le public. Ainsi, pour les secteurs à haut risque, il ne paraît pas incongru que les niveaux attendus pour certaines exigences s'élèvent à plus de 99,99 %. Mais pour atteindre ces niveaux actuellement, on ne pourrait pas utiliser la vérification formelle – qui n'est pas encore directement applicable dans ce contexte – mais on ne pourrait pas non plus utiliser la technique alternative de vérification empirique – qui existe mais offre un niveau de garantie bien inférieur.

Les bonnes pratiques devront aussi être adaptées en fonction du type d'impacts potentiels de l'IA : risque d'effets juridiques, de provoquer des blessures, de produire des iniquités – comme dans le recrutement – des atteintes à la vie privée – avec une surveillance intrusive comme avec la mise en place d'une IA de biométrie… Une première bonne pratique dans la démarche de la mise en place d'une IA est donc de réaliser une analyse d'impact similaire à celle exigée par le RGPD sans oublier de prendre en compte la dimension cumulative des risques. La survenue de plusieurs risques en même temps peut produire un risque plus important que la somme de ces risques. C'était le cas pour le Titanic où l'ensemble des problèmes rencontrés a conduit au naufrage alors que chaque risque pris isolément aurait pu être maîtrisé.

Sur ce point, une autre bonne pratique qu'on peut déjà suggérer d'effectuer est la réalisation d'analyse de vulnérabilité. L'analyse de vulnérabilité est différente de l'analyse de risque par sa propension à faire face à l'inconnu. Nous l'avions dit dans le chapitre consacré aux risques : l'une des spécificités de l'IA à l'heure actuelle est son immaturité et le fait que les acteurs de l'IA doivent faire face à des menaces inconnues ou non-maîtrisées. Donc les acteurs de l'IA doivent maîtriser les risques mais aussi la prise de risques. L'analyse de vulnérabilité est l'une des bonnes pratiques pour étudier les menaces encore inconnues.

Plusieurs domaines posent encore un problème en matière de développement de bonnes pratiques

Enfin, notons qu'à l'heure actuelle, les lacunes sont trop nombreuses pour être exhaustivement citées mais il est important de souligner l'insuffisance de bonnes pratiques dans ces trois domaines :

- **détermination de la fiabilité des algorithmes** : la création de bonnes pratiques sur ce sujet se heurte aux mécanismes d'apprentissage et à l'évolution permanente des algorithmes. Les développeurs ont le plus grand mal à réaliser des tests pour s'assurer du bon fonctionnement de certains algorithmes. Et quand cette problématique sera résolue, on devra se demander pour quelle durée les résultats de ces tests sont valides puisque les IA évoluent rapidement
- **maîtrise des données** : on sait que les risques liés aux données sont la présence de biais, des intrusions dans la vie privée, des problèmes de fiabilité des algorithmes (là aussi)... Certaines bonnes pratiques ont été développées pour la RGPD mais la quantité de données utilisées pour l'IA ainsi que leur nature et leur utilisation nécessitent de nouveaux mécanismes encore à définir
- **l'évolution de la gouvernance** : certes certaines bonnes pratiques de gouvernance issues des systèmes informatiques classiques peuvent être transposables mais la gouvernance elle-même est pressentie pour évoluer avec le développement des IA d'aide à la décision voire des IA de décision automatisées. On peut également mentionner ici le risque des responsabilités associées à ces IA décisionnelles. Des nouvelles bonnes pratiques seront nécessaires.

Or, nous voyons que ces trois domaines sont cumulativement présents dans tous les systèmes d'IA. Cela montre bien le reste du travail à accomplir.

1.3. Référentiels actuels et en cours d'élaboration

Il est possible de se sentir submergé par l'ampleur de ce qu'il reste à faire. Heureusement, des référentiels existent déjà et ils permettent pour partie de limiter certains risques dès lors qu'on les applique. Ainsi, on peut reprendre les démarches présentées dans ITIL qui concerne les directions informatiques en tant que fournisseur de service au reste de l'organisation, COBIT 5 (l'intégrateur des meilleures pratiques en technologies de l'information et le référentiel général de la gouvernance et de management des systèmes d'information qui aide à comprendre et à gérer les risques et les bénéfices qui leur sont associés.), CMMI qui concerne les directions informatiques en tant que réalisateur de développement propre mais aussi les ISO et la RGPD.

Nous l'avons vu ces référentiels n'adressent pas spécifiquement les caractéristiques de l'IA mais de nombreuses études sont en cours pour cela. Ainsi, de nombreux groupes de travail ses sont lancés dans des projets d'études relatifs à ces questions : le Lab50, le Consortium Thales/Total/EDF, l'Union européenne via son Livre Blanc ou encore :

- Artificial Intelligence to Benefit People and Society (PAI)
- the Association for Computing Machinery (ACM)
- the Institute of Electrical and Electronics Engineers (IEEE)
- the International Telecommunications Union (ITU)
- the International Standards Organization (ISO)
- the Information Commissioner's Office (ICO)...

2. Quelques recommandations : référentiels existants, modèles spécifiques, adaptations, analyse d'impact, mise en place progressive

Appui sur les référentiels existants

La première recommandation concernant les bonnes pratiques en IA serait tout d'abord de s'appuyer sur les référentiels actuels tels que COBIT, CMMI, ITIL et ceux mentionnés plus haut en fonction des caractéristiques de l'organisation qui souhaite les mettre en place. De même, dans le cadre de l'audit des systèmes d'IA, les techniques et les meilleures pratiques n'ont pas encore été établies. Toutefois, on peut s'aider d'autres cadres qui sont déjà bien développés. Ce sont notamment les « cadres d'assurance » basés sur les résultats ou les allégations, tels que le cadre des allégations, des arguments et des preuves (comme la méthodologie Claims Argument Evidence de la société Adelard qui sert à présenter des arguments de sécurité[56]) et la notation de structuration des objectifs (comme le Goal Structuring Notation qui est un argument graphique utilisé pour documenter et présenter la preuve que les objectifs de sécurité ont été atteints, dans un format plus clair que le texte brut) qui sont déjà largement utilisés dans les contextes d'audit critiques pour la sécurité.

Développement de modèles spécifiques pour les IA

Il faudra ensuite développer des modèles spécifiques à l'utilisation en IA. L'un des plus évidents est la création d'une classification précise des types d'IA utilisés (qui devrait prendre en compte les algorithmes utilisés, le contexte d'utilisation...). En effet, à chacun de ces types correspondront des exigences spécifiques et des risques associés. À partir de cette classification des IA pourra être établie une classification des risques.

[56] https://claimsargumentsevidence.org/

Certains de ces modèles sont en cours de développement, comme c'est le cas avec [Dans et al.][57] qui ont créé un modèle pour documenter correctement l'ensemble de données d'apprentissage. Cette documentation est structurée en sept catégories :

- motivation pour la création du jeu de données
- composition du jeu de données
- processus de collecte,
- prétraitement appliqué aux données, y compris le nettoyage et l'étiquetage,
- utilisations attendues,
- modes de diffusion,
- procédure mise en œuvre pour sa maintenance.

Une approche similaire est envisagée dans [Mitchell et al.][58] bien qu'elle soit plus centrée sur le modèle, décrivant les données utilisées pour la phase d'apprentissage et d'évaluation, les détails techniques sur le modèle ; la façon dont il a été formé, ainsi que diverses mesures de performance. Bien que ces modèles ne soient pas encore validés comme bonnes pratiques, il peut être intéressant pour les organisations de les mettre en place, celles-ci ou d'autres équivalentes, tout ou en partie, telles quelles ou de manière adaptée à l'organisation.

Identification des bonnes pratiques existantes et en cours de développement

Ces deux exemples illustrent l'un des premiers réflexes à mettre en œuvre au sein des organisations : en raison de l'immaturité des techniques face à leurs rapides évolutions, il faudrait mettre en place un système d'identification des bonnes pratiques spécifiques aux IA attestées ou en cours de développement et de partage d'expériences entre les différentes parties prenantes. Les incidents et vulnérabilités seraient connus plus vite et des solutions apportées plus rapidement. Ces systèmes peuvent être internes à une organisation et/ou émaner d'une structure externe, à l'image de ce qui se fait dans le pharmaceutique pour l'identification des effets secondaires des médicaments. La mise en place d'un tiers externe pour le partage des bonnes pratiques et des vulnérabilités offrirait plusieurs avantages :

- une veille régulière sur les nouvelles bonnes pratiques communes
- la diminution des coûts de partage des comportements souhaités ou inattendus de la part des IA

---

[57] Dans T., Gebru, J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, I. Daume, Hal, and K. Crawford, "Datasheets for Datasets," in Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning, Stockholm, Sweden, PMLR, 2018

[58] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 2019, pp. 220–229

- anonymisation des incidents pour faciliter le partage à la communauté en préservant la réputation des entreprises ayant constaté ces incidents.

Élargissement de l'utilisation des analyses d'impact

D'autres bonnes pratiques connues peuvent conduire naturellement à la mise en place de plusieurs autres bonnes pratiques. C'est le cas par exemple de l'analyse d'impact (data protection impact assessment ou DPIA). Cette dernière a initialement plusieurs rôles comme :

- décrire les traitements, les flux de données et les étapes où les processus IA et les décisions automatisées qui peuvent produire des effets sur les individus
- expliquer les variations et les marges d'erreur pertinentes pouvant affecter l'équité du traitement statistique
- décrire la portée et le contexte de chaque traitement y compris les données traitées (volume, variété, sensibilités, collecte et stockage, sources des données), le nombre de personnes concernées, la nature de la relation avec les individus
- expliciter les résultats escomptés pour l'organisation elle-même, les individus et la société au sens large
- identifier le degré d'implication dans les prises de décisions et préciser à quelle étape cela a eu lieu…

Toutefois, il ne faudrait pas voir les DPIA comme une simple liste à cocher, mais s'en servir comme d'une feuille de route pour identifier les risques pouvant être potentiellement causés par l'IA analysée afin d'utiliser ce document dans une démarche plus globale. C'est uniquement en adoptant une approche micro et macro, lorsque c'est possible, qu'on peut voir qu'une même pratique possède plusieurs conséquences. Ainsi, le DPIA peut aussi être un argument pour démontrer la responsabilité d'une organisation dans les décisions concernant l'achat ou la conception d'un système d'IA.

D'autres analyses peuvent être réalisées en fonction de la nature et du niveau des exigences à atteindre telles que définies par les organisations. Il peut s'agir d'analyse de l'impact sur les libertés individuelles, d'analyse d'impact algorithmique, d'analyse d'impact intégrées sur la transparence, l'équité, la sécurité, la vie privée… En soi, il est possible de tout imaginer. La mise en place de bonnes pratiques n'est que l'une des dernières étapes de la démarche.

CHAPITRE 12

Démarche d'évaluation et de mise en place progressive des bonnes pratiques

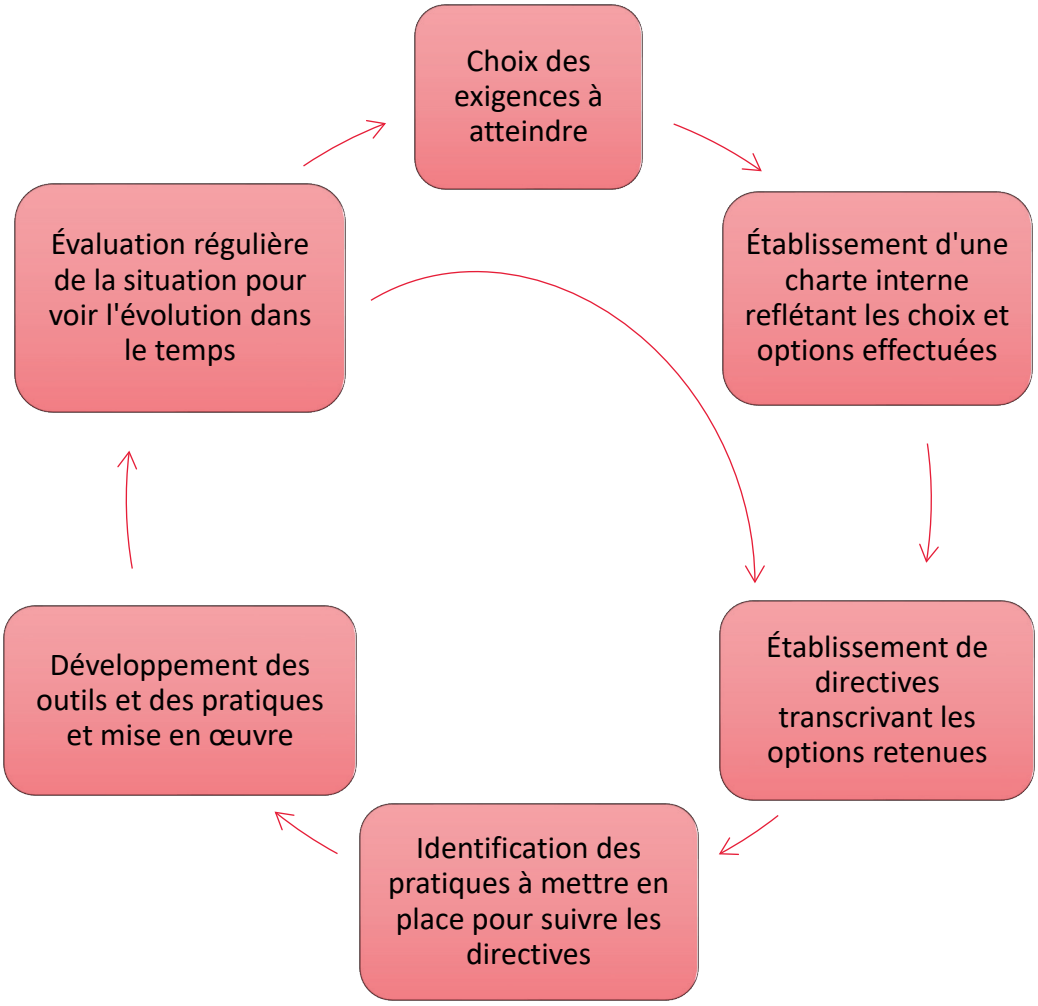

Figure 13 : Cycle de vie depuis les exigences jusqu'aux bonnes pratiques et la poursuite de ce cycle

Une autre bonne pratique est bien évidemment le suivi de l'efficacité des bonnes pratiques via une évaluation régulière pour les modifier ensuite si nécessaire. Ce suivi devrait se réaliser à tous les niveaux que ce soit celui d'une entreprise, d'une association, d'une administration mais aussi d'une profession, d'un organisme d'audit ou de certification voire d'un Etat ou d'une organisation internationale

Dans la majorité des cas, il vaut mieux mettre en place un dispositif (qu'il soit organisationnel ou technique) visant à aider à l'atteinte des exigences, même s'il n'est pas parfait ou idéal, que pas de dispositif du tout. Bien sûr, pour que la mise en place de ce dispositif soit efficace, il faut qu'il réponde à certaines exigences :

- La facilité de mise en œuvre : si le mécanisme est trop complexe, il ne sera pas utilisé optimalement dans l'organisation
- La pérennité des pratiques : si les pratiques évoluent trop souvent, non seulement elles ne seront pas suivies, mais leur efficacité risque d'être remise en cause
- La proportionnalité aux niveaux de risques : si les risques sont faibles, il n'est pas nécessaire de mettre en place les mêmes dispositifs que pour un risque élevé.

Plusieurs bonnes pratiques spécifiques ont été décrites tout au long de ce document, notamment dans les paragraphes « recommandations ». Si besoin, nous enjoignons le lecteur à s'y référer.

La confiance dans les systèmes d'information à base d'intelligence artificielle sera effective lorsqu'elles suivront les principes définis et atteindront les exigences fixées. Mais ces dernières peuvent être formulées de manière théorique. Comment faire techniquement qu'une IA soit transparente ou équitable ? Ces éléments doivent donc être convertis en instructions techniques qui soient applicables dans la conception du système. Avec le temps, certains de ces mécanismes s'avèreront particulièrement efficients pour atteindre les exigences et devraient se diffuser en tant que bonnes pratiques. Les IA étant un domaine neuf, il est évident que limitées sont les bonnes pratiques spécifiques à s'être distinguées. Il faudra du temps pour cela. En revanche, il est déjà possible d'intégrer les bonnes pratiques relatives aux systèmes d'information traditionnels voire d'en adapter certaines aux IA, qui ont été décrites successivement dans les chapitres précédents.

Le suivi de cette démarche qui a été présentée jusqu'à présent, permettrait aux concepteurs et développeurs d'IA de parvenir à l'obtention d'une IA digne de confiance. Il reste toutefois une dernière étape. Celle-ci a pour but de fournir une assurance quant au niveau de confiance effectif de l'IA à tous les autres acteurs impliqués. Il s'agit de l'étape d'Audit et de certification.

CHAPITRE 13

AUDIT ET CERTIFICATION

Après avoir passé des mois à créer des plans d'aménagement, préparé la terre et prendre soin de ses massifs, le jardinier satisfait désire prouver à tout le monde qu'il a le plus beau jardin. Des retours enthousiastes des voisins et de la famille le conforteront dans son opinion mais cela n'est pas suffisant pour qui veut en avoir la preuve formelle. La solution est simple : participer à un concours.

Pour les systèmes à base d'intelligence artificielle, nul n'est tenu à prouver qu'il a fait la meilleure IA mais seulement de démontrer que son IA atteint les exigences qu'on lui avait imposées. Pour cela, au lieu d'un jury de passionnés d'horticulture, il vaut mieux un auditeur indépendant. Dans ce chapitre, nous nous intéressons aux caractéristiques des audits des systèmes d'information, aux facteurs à prendre en compte, aux types de travaux à mener, en comparant ces questions avec les audits effectués par les commissaires aux comptes pour les états financiers. Nous évoquerons enfin quelques pistes pour élaborer des référentiels d'audits spécifiques à l'audit des IA.

1. Problématiques spécifiques à l'IA : proposition de valeur, fixation des objectifs et des travaux d'audit, parallèle avec le commissariat aux comptes, illustration des options, communication

1.1. Définition de la proposition de valeur adaptée pour les audits et certifications des IA

Nous avons pu voir que compte tenu des très forts enjeux de l'écosystème des IA mais aussi des nombreuses inquiétudes associées, il en est découlé un impératif de pouvoir disposer d'IA dignes de confiance. Nous avons, dans les chapitres précédents, identifié en quoi les IA pourraient être dignes de confiance, les difficultés pour y arriver et quelques dispositifs et bonnes pratiques qui permettraient de développer et de déployer de telles IA. Mais non seulement il faut disposer d'IA de confiance mais il faudrait aussi convaincre plusieurs parties prenantes concernées inquiètes que ces IA soient effectivement dignes de confiance. Comment faire ?

Il conviendrait d'envisager l'intervention d'un tiers de confiance indépendant, pour qu'il puisse rassurer certaines parties prenantes quant à la nature et au niveau réel de confiance que l'on pourrait avoir dans ces IA. Plusieurs types d'intervention pourraient ainsi avoir lieu pour traiter ces différents cas. Mais lesquels, dans quelles circonstances, avec quel périmètre, à quel coût… ?

On peut penser que pour les IA à hauts risques pouvant avoir un impact significatif auprès de certaines parties prenantes, un auditeur ou un commissaire aux IA tel que c'est le cas pour l'écosystème financier, pourrait être mandaté pour réaliser une mission d'intérêt général de certification des assertions relatives aux exigences de qualité attendue a minima pour des IA dignes de confiance.

En effet, comme c'est le cas pour les différents acteurs du système financier et économique – qu'ils soient actionnaires, employés, banquiers, fournisseurs, clients, citoyens ou l'État – qui ont besoin d'avoir confiance dans les éléments financiers qui sous-tendent leurs décisions d'investisseurs, de prêteurs, de partenaires, d'achats… ; les parties prenantes de l'écosystème des systèmes à base d'intelligence artificielle ont tout autant besoin d'avoir confiance dans les IA. Sans confiance, il ne peut y avoir de croissance.

Dans le cas des éléments financiers, pour favoriser cette confiance, ce sont des professionnels comptables, tiers indépendants, qui sont mandatés pour certifier la sincérité et la régularité des états financiers de certaines entreprises privées et publiques. Il s'agit d'une mission d'intérêt général pour le compte de l'ensemble de l'écosystème financier. Ainsi, de nombreuses grandes entreprises (du CAC 40 et autres) sont prêtes à payer chacune plusieurs dizaines de millions d'euros d'honoraires chaque année pour rassurer l'écosystème financier sur leurs comptes. En est-il de même pour les IA ?

Cette certification doit en effet pouvoir créer de la valeur. Il s'agit donc de trouver le bon équilibre entre :

- la nature des assertions attendues qu'elles soient générales telles que des IA dignes de confiance ou plus spécifiques telles que des IA transparentes, équitables et non discriminantes, fiables, sécurisées, robustes, conforme à la réglementation…
- la nature de l'opinion (certification, label, attestation, opinion…) en fonction du niveau de confiance souhaitée pour ces assertions
- le niveau de risque de fournir une certification erronée qui soit acceptable par les parties prenantes, par exemple : certifier qu'une IA est digne de confiance ou équitable alors qu'elle ne le serait pas
- le niveau d'efforts nécessaires pour fournir ce type et ce niveau de confiance qui soit techniquement faisable et financièrement soutenable.

Que ce soit un audit menant ou pas à une certification, il s'agira de mettre en œuvre une démarche qui permette d'identifier précisément le type d'intervention qu'il conviendra de faire dans un contexte donné.

1.2. Quels facteurs à prendre en compte pour déterminer la nature des objectifs et des travaux d'audit pour les IA ?

On peut rappeler que tout audit doit fournir aux parties prenantes concernées un certain niveau de confiance quant à une opinion relative aux caractéristiques de l'objet audité et ce, à une date ou pour une période donnée. Cette opinion se doit de leur être utile compte tenu des avantages et bénéfices qu'elle procurera et du niveau de risque qu'elle soit erronée et du niveau des moyens à mobiliser pour l'émettre.

Pour que cela soit le cas, il faut clarifier chacun de ces points :

- **les parties prenantes** concernées par l'audit : quelles sont les parties prenantes qui sont intéressées par le résultat de l'audit ? S'agit-il des fournisseurs de technologies, de solutions ou de données, des sponsors, des développeurs, des déployeurs, des instances de réglementation, des usagers… ? Chacun a des intérêts propres. La nature de la mission sera donc fonction des attentes de ces parties prenantes

Cela aura aussi un impact sur le profil de l'intervenant : auditeur interne, auditeur externe, expert, instances de contrôle type CNIL ou Cour des Comptes, agence publique dotée d'un pouvoir de police et de sanction… et sur la nature de l'expertise à mettre en œuvre

La détermination des parties prenantes concernées dépendra de la nature de la mission. S'agit-il d'une mission issue d'une obligation légale ou contractuelle ?

- **les objets** à auditer : il peut s'agir des objets directement mentionnés dans les assertions à auditer – les éléments avancés par l'organisme audité et dont on veut vérifier la véracité

  Dans le cas des IA, ces objets peuvent être de différentes natures. Il peut s'agir des résultats issus des IA, tels qu'un diagnostic, une prescription, une action… Pour une telle nature de résultat, l'assertion pourrait être par exemple « les octrois de crédits issus de l'IA sont justes, équitables, sans biais ». Le résultat peut être un produit tels une voiture autonome, un robot… L'assertion pourrait être « la voiture autonome x est digne de confiance »

  Il peut aussi s'agir de l'IA elle-même. L'assertion pourrait être « l'IA est digne de confiance ». Mais dans ce cas qu'entend-on par IA ? S'agit-il d'un algorithme ou d'un ensemble d'algorithmes ? S'agit-il d'un système d'IA comprenant des modèles, des logiciels, des données, des capteurs, des techniques d'apprentissage… ? On pourra donc être amené à auditer ces différents éléments. Chacun de ces éléments devra être fiable, robuste…

  Cela peut être aussi des moyens et des dispositifs qui vont être mobilisés pour aboutir à un objet comme une IA digne de confiance. Il s'agira alors de compétences, de processus, d'outils, de documentations… Chacun de ces éléments doit être disponible, fiable, robuste, conforme … et pourra faire l'objet de l'audit

  Ces éléments et moyens peuvent aussi être pris dans leur ensemble et être évalués dans leur capacité à favoriser le développement d'IA dignes de confiance.

- **Les assertions** à auditer quant à certains objets : l'audit sert à valider l'atteinte de certaines exigences quant aux caractéristiques relatives aux objets de l'audit. Ces caractéristiques peuvent être la sécurité, l'explicabilité, l'éthique, la protection de la vie privée, l'absence de biais nuisibles, la conformité à la réglementation, la fiabilité… Les assertions peuvent couvrir un spectre assez large de qualités ou une qualité en particulier. Ainsi, l'assertion pourrait être « l'IA est digne de confiance » ou « l'IA est équitable », « l'IA est explicable » ou « l'IA est robuste ». Cela pourrait être aussi, « les données d'apprentissage sont fiables » ou « les données d'apprentissage sont exactes, valides, représentatives et non biaisées… »

  Dans chacun de ces cas, il sera nécessaire d'avoir défini au préalable ce que recouvre chacune de ces qualités et lorsqu'une assertion est plus générale, digne de confiance par exemple, quelles qualités seraient concernées. Il faut également voir qu'en fonction de l'objet audité, les qualités à auditer peuvent être différentes. Par exemple, on peut auditer la sûreté et la sécurité pour des systèmes d'IA tels que les véhicules autonomes et les systèmes d'IA médicaux mais auditer l'absence de biais nuisibles aux IA de moteurs de recherche et des systèmes de recommandation.

Il faut *in fine* que ces assertions soient auditables et elles-mêmes comprises par les acteurs concernés.

- **le niveau de confiance** attendue de ces assertions : il faudra avant tout audit déterminer quel est le niveau d'assurance à atteindre, sachant que plusieurs types d'IA nécessiteront des niveaux de confiance assez différents. Ainsi, pour certaines IA à très hauts risques telles que des voitures autonomes, les niveaux d'assurance requis pourraient être extrêmement élevés (supérieur à 99,9 %) pour certaines exigences

  Il faudra donc déterminer le niveau d'assurance raisonnable pour les différents types d'exigences pour les différents types d'IA pour différents contextes. Mais comment déterminer ce niveau de confiance et le niveau de confiance effectif pour chacune des exigences et pour chacun des objets et pour l'IA prise dans son ensemble ? Quels sont les éléments probants à l'appui des assertions qui permettraient de conclure que chaque élément et que l'IA dans son ensemble ne contiennent pas d'anomalies significatives et qu'ils seraient dignes de confiance ? Lorsqu'il existe des anomalies, des erreurs, des incertitudes… comment apprécier leur impact sur le niveau de confiance ?
- **la nature de l'opinion** : plusieurs types d'opinions sont possibles : attestation, opinion circonstanciée, certification, label… À chaque type d'opinion correspond une nature et une étendue de travaux d'audit et donc d'un certain niveau de confiance. Ces opinions peuvent être émises sans réserve. Mais en cas de difficultés pour auditer certains éléments ou en cas d'identification de problèmes significatifs qui pourraient avoir un impact sur le niveau de confiance obtenu, un refus de certifier ou des réserves spécifiques pourraient être émises. Ces réserves doivent être codifiées pour permettre leur compréhension commune. Au final, chacun de ces types d'opinion et de réserve apporte un certain niveau de confiance. Il faut que les acteurs concernés comprennent quels sont leurs apports et leurs limites
- **la date ou la période couverte** : faut-il réaliser un audit avant le déploiement, après le déploiement, chaque année, à chaque évolution, à chaque évolution importante, … ? Pour combien de temps les résultats de l'audit restent-ils valables – notamment lorsque l'IA évolue régulièrement lorsque l'apprentissage est continu ? Peut-on mettre en place un audit continu ?
- **les avantages et bénéfices** de l'audit : pourquoi faire un audit ? Pour obtenir un label, pour répondre à la réglementation, pour répondre à un recours, pour donner confiance… ? Il est rare de réaliser des audits pour le plaisir, connaître le but de l'audit permet de mieux l'orienter. Cela aura un impact sur la nature de l'assertion, le niveau de confiance attendu, le type d'opinion, la nature et l'étendue des travaux…
- **le niveau de risque** de fournir une **opinion erronée :** ce sont les risques relatifs à l'opinion émise par l'auditeur. Autrement dit, quel est le risque que les moyens et/ou les bonnes pratiques mis en œuvre par l'auditeur ne permettent pas d'obtenir une opinion de qualité ? Quelles seraient les conséquences de l'émission d'une opinion erronée ? Cela dépendra naturellement du contexte
- **le niveau d'effort** pour fournir cette opinion avec le niveau de confiance attendu : l'auditeur doit mobiliser des moyens pour effectuer sa mission et émettre son opinion. Cet effort sera d'autant plus important que le niveau de confiance attendu sera fort. Il faut que ce niveau de moyens soit soutenable

financièrement pour les parties prenantes qui seront sollicitées pour les financer et techniquement c'est-à-dire que les outils nécessaires pour valider les qualités à auditer existent, sans quoi l'audit ne pourra être réalisé. Il conviendra donc d'en tenir compte lorsque la nature et l'étendue des travaux seront déterminées.

Tous ces éléments doivent être éclaircis avant que l'on puisse procéder à un audit. Ils devraient permettre d'établir un cadre à l'audit en plus d'un objectif et d'un périmètre. Une fois les éléments énoncés ci-dessus établis, il sera possible d'identifier la nature et l'étendue des travaux d'audit nécessaires pour répondre aux attentes de qualité des parties prenantes.

1.3. Quels types de travaux d'audit à mener ?

Deux types d'audit : validation du fond ou de la forme

En fonction de l'objectif fixé, les travaux d'audit pourront consister à valider le fond qui pourrait être :

- la qualité d'un résultat issu d'une IA par rapport à un standard de référence défini en début d'audit
- la qualité d'une IA et/ou de ses composants par rapport à un standard de référence défini en début d'audit
- la qualité des moyens mis en œuvre par rapport à des bonnes pratiques de référence à définir en début d'audit pour permettre d'aboutir au niveau de qualité attendu pour le résultat issu d'une IA ou pour l'IA elle-même
- la qualité d'une combinaison de ces qualités (qualité d'un résultat issu d'une IA ou d'une IA et qualité des moyens mis en œuvre pour obtenir ce résultat ou cette IA)

Il pourrait aussi consister à valider la forme, c'est-à-dire le contenu d'une déclaration, d'un rapport, d'une assertion qui pourrait couvrir un domaine donné (gouvernance, gestion des risques…) ou un point particulier (utilisation de tel outil d'explicabilité pour tel type de technique d'apprentissage, mise en place d'une charte, d'une analyse d'impact…).

Impact très significatif du choix de type d'audit (fond ou forme) sur le niveau de moyens à mobiliser et sur le niveau de confiance fourni

Ces différents types d'audit vont conduire à fournir des niveaux de confiance différents sur la présence d'IA dignes de confiance. Le niveau de mobilisation des moyens des auditeurs sera aussi très différent. Ainsi, dans le domaine financier, les honoraires de Commissariat aux Comptes pour la certification (sur le fond) des états financiers d'un grand groupe (certification obligatoire) pourront se chiffrer en plusieurs dizaines de millions d'euros par an. Pour la certification (sur le fond) de leur dispositif de gouvernance, de contrôle interne et de gestion des risques (certification non obligatoire en France mais obligatoire dans certains pays pour certains groupes à risque), ils pourront se chiffrer en centaines de milliers d'euros ou en millions d'euros. Pour la certification (sur la forme) de la sincérité du rapport émis par la direction générale

sur la gouvernance, le contrôle interne et la gestion des risques, ils pourront se chiffrer en quelques milliers ou dizaines de milliers d'euros.

La Commission européenne propose à ce stade des certifications, pour les IA à très haut risque, uniquement sur la qualité de certains dispositifs de contrôle interne tels la documentation, les processus de gestion de risques... dont elle a estimé leur coût à quelques milliers d'euros. Mais cela permettra-t-il de fournir le niveau de confiance en cohérence avec les attentes des parties prenantes ? Compte tenu de ces écarts importants entre ces différents types d'audit, il est important de bien comprendre à quoi pourrait correspondre chaque type d'audit pour les IA et quels travaux d'audit seraient à mener pour chacun de ces cas.

Impact important du niveau d'acceptabilité de l'émission d'une opinion d'audit erronée sur le choix de la nature et de l'étendue des moyens à mobiliser

Il est important de comprendre que la pratique d'un audit entraîne elle-même un risque : le risque que les travaux d'audit mènent à une opinion contraire à la réalité. Ces risques sont de deux types : l'établissement d'un faux positif (l'auditeur émet une opinion positive alors qu'elle devrait être négative) et l'établissement d'un faux négatif (l'auditeur émet une opinion négative alors qu'elle devrait être positive). En fait, il faut déterminer quel est le taux de faux positifs et de faux négatifs qu'on considère comme tolérables compte tenu des conséquences qu'elles pourraient engendrer notamment lorsqu'il s'agira d'IA sensibles à très hauts risques ?

Niveau des travaux d'audit techniquement faisable et financièrement soutenable

Les audits d'IA peuvent être difficiles à réaliser car les assertions des organisations concernant les attributs de l'IA à auditer peuvent ne pas être facilement vérifiables. En effet, le processus de développement de l'IA est souvent opaque pour les personnes extérieures à l'organisation (voire au sein même de l'organisation pour toutes les personnes extérieures à l'équipe de développement et de déploiement voire même pour ceux qui sont directement impliqués). La société n'a peut-être pas mis en œuvre des dispositifs adéquats qui permettraient d'avoir confiance sur tel ou tel aspect comme l'explicabilité et la non-présence de biais. De plus, des outils d'audit ne sont pas toujours disponibles pour tester la fiabilité de tel ou tel algorithme, notamment lorsqu'ils sont opaques.

Par ailleurs, une opinion erronée peut résulter de la mobilisation insuffisante des moyens nécessaires pour émettre une opinion avec le niveau de confiance attendue. Il faut en effet que ces efforts soient soutenables financièrement. Il faut donc apprécier la faisabilité technique et financière de conduire un audit dans un contexte donné avant de procéder à cet audit.

Difficulté de bien comprendre le réel niveau de confiance en fonction des résultats d'audit

Enfin, on peut se questionner sur la signification des résultats d'audit qui seront le cas échéant communiqués aux différentes parties prenantes. Cette restitution peut être assez courte : « je certifie ou j'atteste que l'IA x est digne de confiance », « je certifie ou j'atteste que la documentation est conforme au référentiel x » … Mais que vont comprendre les différentes parties prenantes de cette affirmation ? Vont-ils bien comprendre que la certification, pourtant affirmative, ne porte en réalité qu'un niveau de confiance limitée (à 95 %, 98% ou autre) ? Vont-ils bien comprendre à quelle IA, à quels éléments de l'IA, à quelles exigences couvertes se rapportent ces affirmations ? En un mot, vont-ils bien comprendre les limites du résultat d'audit, en général sous-entendues comme ce qu'on peut voir sur les audits financiers ?

Comment faire comprendre qu'une certification ou une attestation sur la conformité de la documentation peut contribuer à donner confiance dans l'IA mais ne permet pas à elle seule d'en déduire que l'IA est digne de confiance ? On peut se poser la même question pour la conformité d'un processus à un référentiel donné qui permet d'obtenir une documentation et des processus dignes de confiance. Beaucoup d'autres facteurs entrent en ligne de compte.

Naturellement, cette certification ou attestation – qui peut contribuer à fournir un niveau de confiance concernant l'IA, par exemple à hauteur de 50 % – peut être utile pour des efforts d'audit relativement limités. Mais comment savoir si cela contribue vraiment à hauteur de 50 % ? Que peut-on dire sur la réalité de la confiance de l'IA ? Il y a peut-être des dizaines d'autres facteurs que l'on n'aura pas évalués puisque ne faisant pas partie du périmètre d'intervention. On ne pourra donc pas communiquer sur ce niveau de confiance. Dans ce contexte, comment interpréter des résultats seulement relatifs à la qualité de la documentation sur une IA ? Cela peut conduire certains à surestimer ou sous-estimer le niveau de confiance réel et subir des conséquences néfastes par excès d'optimisme par exemple.

De plus, dans certains cas, on ne pourra pas fournir une opinion sur une qualité (IA digne de confiance, IA explicable, IA robuste…) mais plutôt une opinion circonstanciée en fonction des travaux d'audit effectués et des résultats associés. On pourra alors présenter le périmètre exact audité et ses limites, la nature exacte des travaux et leurs limites, le type d'audit et ses limites, la date ou période couverte et les limites associées, la nature exacte des conclusions et leurs limites, la présence de certains risques, etc.

Qu'en conclure ? On peut imaginer, que ce soit lorsque l'opinion est succincte ou plus détaillée, que de nombreuses interprétations des résultats seront possibles quant au niveau de confiance que l'on peut effectivement obtenir des travaux effectués. Il convient d'en tenir compte dès l'initiation d'une mission d'audit.

Il faut donc bien faire comprendre les limites de tous ces types d'audit lorsque leurs résultats seront restitués. Ainsi, dans le cas où un domaine spécifique sera audité, une attestation sur la conformité pourra être émise.

CHAPITRE 13

De nombreux dispositifs et travaux d'audit possibles

Nous avons vu plus avant que plusieurs leviers peuvent être mobilisés pour s'assurer que les IA soient dignes de confiance. De même, plusieurs leviers peuvent être mobilisés pour s'assurer que les audits de ces IA soient dignes de confiance. Il s'agit en réalité des mêmes sept leviers présentés plus haut :

- les principes, directives et référentiels d'audit
- les structures organisationnelles d'audit
- les processus d'audit
- le personnel, les aptitudes et les compétences d'audit
- la culture, l'éthique et les comportements d'audit
- les outils, les dispositifs et les guides d'audit
- les informations utilisés et produites par l'audit...

Des démarches d'audit génériques et des référentiels de bonnes pratiques d'audit spécifiques aux différents types d'audit pour chacun de ces leviers ont été développés pour les autres types d'audit et notamment pour l'audit des systèmes d'information. On peut largement s'en inspirer pour l'audit des IA. Les auditeurs vont donc être amenés à mener leurs travaux en s'appuyant sur ces leviers.

On peut donner quelques exemples du type de travaux que les auditeurs peuvent mener :

- interviews auprès des différents acteurs (contexte, éléments critiques, compréhension...)
- obtention d'informations : notes de présentation des fonctionnalités, des techniques d'apprentissage, des données d'apprentissage, de l'environnement technique, organisationnel, de gouvernance..., analyses d'impact et analyses des risques spécifiques, documentation, incidents, litiges, contentieux, évènements exceptionnels... et éléments probants les étayant (auprès des acteurs concernés, de l'audit interne, de tiers...)
- observation et inspection du fonctionnement de certains dispositifs
- revues ou analyse détaillée des éléments reçus (y compris avec l'aide d'outils automatisés si nécessaire)
- évaluation du niveau de risque pour les IA de disposer de résultats ou d'objets dignes de confiance
- évaluation de la conception des dispositifs mis en œuvre
- vérification de la mise en œuvre effective de ces dispositifs : revue, tests... sur un échantillon représentatif de chaque dispositif significatif audité (y compris avec l'aide d'outils spécialisés si nécessaire)
- identification du niveau de risque d'audit suite aux travaux d'audit
- choix de la nature de l'opinion (avec ou sans réserve)
- restitution des résultats et leur communication.

Choix de la nature et de l'étendue des travaux d'audit en fonction du type d'audit et du contexte

Compte tenu des spécificités des IA (opacité, évolutivité des algorithmes, techniques d'apprentissage, traçabilité, fragilité…), plusieurs de ces techniques seront particulièrement difficiles à mettre en œuvre et, lorsque cela sera possible, il sera souvent difficile de conclure sur la présence d'IA dignes de confiance.

Nous avons noté plus haut plusieurs types d'audit qui évaluent le fond (certification qu'une IA est digne de confiance) ou la forme (certification qu'une assertion est sincère sans qu'une appréciation soit portée sur le fond : par exemple le contenu du rapport présentant le dispositif de gouvernance en place est exact mais aucune évaluation n'est portée sur la pertinence, l'efficacité et l'effectivité de ce dispositif). Il faudra donc identifier le type et l'étendue des travaux et la nature des outils qui permettent de répondre au mieux aux objectifs d'audit en fonction du type d'audit et du périmètre visé.

Illustration des travaux d'audit pour vérifier l'assertion « les décisions d'octroi de crédits issus d'une IA en particulier sont dignes de confiance »

Si on prend l'exemple de l'audit de l'assertion sur le fond « les décisions d'octroi de crédits issus de cette IA sont dignes de confiance », nous pourrions prendre un échantillon représentatif et suffisant de demandes d'octrois de crédit qui représente les différents types de situations susceptibles de se présenter et vérifier qu'ils ont été traités de manière conforme aux règles d'octroi en vigueur, que leur octroi ou non-octroi ne résulte pas d'une discrimination, que les différents dispositifs mis en place respectent la vie privée…

Mais combien de demandes d'octrois et lesquelles faut-il inclure dans l'échantillon ? Plus l'échantillon sera représentatif et important, plus le niveau de confiance sera élevé, mais plus le coût de l'audit le sera aussi. Ces règles auront été préalablement déterminées à la suite d'un processus d'apprentissage automatisée et se retrouveront dans le modèle d'apprentissage qui en a résulté. Mais comment s'en assurer ?

Par ailleurs, plusieurs autres dispositifs (chartes internes, analyse d'impact…) pourront compléter le paysage. Il faudra donc vérifier que, conceptuellement, ces outils et ces tâches sont appropriés. Mais quel niveau de confiance procure-t-il ? Sont-ils réellement mis en œuvre ?

Si des techniques d'apprentissage de type linéaire ont été utilisées, il devrait être possible de retracer comment une demande d'octroi de prêt a donné lieu à un octroi ou non-octroi d'un prêt. Il s'agira donc pour l'auditeur de vérifier pour l'échantillon choisi que les règles ont bien été respectées. Mais comment s'assurer que ces règles permettent réellement la non-discrimination ? Elles pourraient découler de données d'apprentissage biaisées. Il faudrait vérifier si des outils ou des dispositifs ont été mis en œuvre pour éviter les biais. Par exemple, des travaux de nettoyage, d'échantillonnage, d'anonymisation, d'étiquetage… ont-ils été mis en œuvre ? Plusieurs outils favorisant la non-discrimination (qui ont été présentés dans les chapitres précédents : outils de calibration, de redressement des données, d'anti-classification…) ont-ils été mis en place ? Quelles sont leurs limites compte tenu du contexte ?

Il est donc fort probable que vérifier que des règles aient été appliquées sur la base d'un échantillon ne soit pas suffisant pour porter un jugement sur des octrois équitables. Il sera alors nécessaire de comprendre les choix techniques qui ont été opérés et les autres dispositifs mis en place, comprendre leurs limites, déterminer si a priori ils sont appropriés et s'ils éviteront que les modèles d'apprentissage soient biaisés. Cela fournira un premier niveau de confiance. Mais quel niveau de confiance procurent-ils ? Sont-ils réellement mis en œuvre ? L'auditeur pourra alors vérifier sur la base d'un échantillon représentatif que les procédures ainsi prévues ont bien été appliquées. Il faudra déterminer ce qui pourrait être considéré comme suffisamment probant. Il pourra s'appuyer sur des outils pour vérifier par exemple que les outils de calibration et autres ont été mis en œuvre de manière satisfaisante. À chaque élément vérifié, l'auditeur obtient un niveau de confiance supplémentaire sur le fait que le modèle d'apprentissage est fiable et non biaisé et qu'associé à plusieurs autres dispositifs (qui eux aussi auront été vérifiés par l'auditeur), ils permettront d'avoir des octrois de prêts dignes de confiance.

Illustration des travaux d'audit pour vérifier l'assertion « la documentation concernant cette IA en particulier est digne de confiance »

Nous pouvons aussi illustrer cette problématique en prenant comme exemple la vérification d'une documentation digne de confiance. Les travaux pourraient avoir pour objet de vérifier que cette documentation comprend tous les éléments attendus tels que précisé dans un référentiel. Il peut s'agir des techniques et outils utilisés, des dispositifs de résolution de problèmes, des incidents… Toutes les cases du référentiel peuvent être vérifiées. L'auditeur pourra en déduire un certain niveau de confiance théorique. Mais doit-il évaluer si tel ou tel outil ou dispositif mentionné dans le document est pertinent, si c'est proportionné aux problèmes à traiter… Doit-il uniquement inspecter cette documentation ou doit-il effectuer des tests ? Si oui, quels tests, combien de tests ? …

Difficulté d'évaluer l'impact des problèmes, anomalies et erreurs identifiés lors des travaux d'audit

À chacune de ces étapes, l'auditeur peut être amené à identifier des problèmes, des erreurs… Mais leur nature, leur nombre, leur importance… remettent-ils en cause la confiance de l'auditeur ? À chaque élément vérifié, l'auditeur obtient un niveau de confiance supplémentaire que l'octroi des prêts est digne de confiance. L'auditeur aura *in fine* à déterminer si les dispositifs pris dans leur ensemble sont satisfaisants et que les anomalies ne remettent pas en cause leur fiabilité.

Si ce sont des techniques d'apprentissage du type opaque qui ont été utilisées, ne permettant pas une bonne traçabilité des choix techniques, nous avons vu qu'il existait des outils et dispositifs qui pouvaient fournir une certaine assurance quant à l'octroi de prêts de confiance mais que de nombreuses limites leur étaient attribuées. Le même type de travaux d'audit pourra être mené. Mais ils auront des limites. Il ne sera peut-être pas possible d'aboutir au même niveau de confiance. Comment évaluer cet impact ?

On voit bien au travers de ces exemples, que le choix du type de travaux d'audit est varié et dépend grandement des dispositifs en place et la possibilité de les tester et du niveau de confiance attendu.

Un dossier d'audit avec les éléments probants justifiant l'opinion de l'auditeur

L'auditeur sera donc amené à apprécier à chaque étape le niveau de confiance obtenu sur la base de son jugement. Compte tenu de la criticité de son opinion notamment pour les IA sensibles à haut risque, l'auditeur devra constituer un dossier d'audit dans lequel doit figurer les documents qui permettent d'étayer l'opinion formulée dans son rapport et qui permettent d'établir que l'audit des IA a été réalisé avec diligence dans le respect de la réglementation et des normes pour le type d'audit concerné. En principe, deux professionnels de l'audit devraient avoir sur la base des mêmes travaux avoir le même résultat.

Il s'agira notamment :

- des éléments de planification avec le plan de travail
- de la nature, du calendrier et de l'étendue des travaux d'audit
- de la liste des intervenants et des dates d'intervention
- du détail des éléments testés et des personnes rencontrées
- des résultats des vérifications effectuées
- des problématiques significatives relevées et des conclusions correspondantes ;
- des échanges avec les responsables des IA sur les éléments significatifs et sur d'éventuelles incohérences
- de l'argumentaire pour conclure sur ces éléments notamment lorsqu'il y a plusieurs options.

De nombreux choix à faire de manière cohérente

Dans ce contexte, il convient donc de bien identifier le type d'audit qui permette de répondre aux attentes de la manière la plus efficace et des conditions de sa mise en œuvre. Il s'agit de trouver le bon équilibre entre l'obtention d'un niveau de confiance utile dans un contexte donné, le niveau de risque d'émettre une opinion erronée dans ce contexte et le niveau de moyens à mobiliser pour émettre ce type d'opinion qui soit soutenable.

1.4. Illustration de tels choix effectués pour le Commissaire aux comptes et comparaison avec les efforts restants pour l'audit des IA

Pour illustrer cette problématique, vous trouverez ci-après un tableau présentant les différents éléments qui nécessitent que des choix soient effectués. Pour chacun de ces éléments, un comparatif a été effectué entre les choix effectués pour les états financiers dans un environnement plus mature et le type d'options que l'on pourrait envisager pour les IA.

	Audit financier	Audit pour les IA
État de la réflexion	Déjà déterminé	À déterminer : il n'y a pas encore de réponse du marché
Parties prenantes concernées	Actionnaires, État, fournisseurs, créanciers…	• prescripteurs : ceux qui commanditent un audit voire paient pour l'audit. (Les prescripteurs devraient-ils être des administrateurs indépendants comme pour le choix du CAC ?) • ceux qui utilisent cette opinion : les opérateurs, les développeurs, les fournisseurs (notamment de données), les utilisateurs, l'État… Cette opinion pourrait découler d'une obligation légale ou d'un engagement contractuel.
Niveau d'assurance raisonnable	95 % de confiance	Il reste à déterminer : 95 %, 99 %, 99,999 % ? Il pourrait être variable en fonction du type d'IA concerné (IA à très haut risque ou non) et du type d'assertion (IA digne de confiance, IA explicable…).
Opinion	Certification avec ou sans réserve	Attestation, certification et avec quels types de réserves, opinions circonstanciées, labels…

	Audit financier	Audit pour les IA
Qualités/ Caractéristiques	Sincérité, régularité	On pourrait envisager d'auditer différentes qualités de l'IA : • générales telles que IA digne de confiance, IA responsable, IA éthique… • générales mais plus Spécifiques telles que IA transparente, IA explicable, IA équitable, IA fiable, IA robuste, IA conforme à la réglementation… • spécifique à l'IA ou à un de ses éléments tels que IA conforme à tel référentiel, la documentation est conforme à tel référentiel, l'IA est conforme aux droits fondamentaux… • spécifiques à une assertion telle l'assertion que des outils d'explicabilité de type LIME ont été utilisés est exacte ou sincère…
Objet	États financiers (Audit sur le fond) Rapport du Président sur la gouvernance, les procédures de contrôle interne et la gestion de risques (article L. 225-37 du Code de Commerce) (audit sur la forme)	Plusieurs pistes peuvent être envisagées sur le fond concernant l'IA ou ses éléments : • algorithme, groupe d'algorithmes, capteurs, Données en entrée… • combinaison de ces éléments • résultat d'un ou plusieurs de ces éléments dans un environnement d'exploitation donné (avec ses caractéristiques techniques, humaines, organisationnelles…) Plusieurs pistes sur le fond concernant les moyens mis en œuvre : les processus, la documentation… Plusieurs possibilités sur la forme telle que la déclaration ou rapport du sponsor ou de l'opérateur concernant l'IA, concernant la gouvernance mise en place ou concernant le type d'outils utilisés par exemple assurant l'explicabilité. Une opinion serait alors émise sur la sincérité de la déclaration (un outil x a bien été utilisé) mais pas le fond (l'outil a permis de disposer d'une IA explicable).

Période	Pour le bilan, à une date donnée. Pour les résultats, sur une période donnée, (en général sur un an)	Reste encore à déterminer s'il pourrait s'agir d'un audit à une date donnée, pour une période donnée… Cela pourrait également varier si le système audité comprend des algorithmes auto-apprenants ou non. On pourra auditer un algorithme à une date donnée et les dispositifs de gouvernance, de management, de contrôle interne, de gestion des risques mis en œuvre pour ces algorithmes pour une période donnée.
Avantages/ Bénéfices	Confiance dans le marché financier	Variables en fonction de l'opinion et/ou du label obtenu par l'audit : confiance du public, résolution des contentieux, assurance avant mise en exploitation…
Risques acceptables	95 % de confiance et obligation de moyens	Cela reste encore à déterminer : 95 %, 99 % ou autre de confiance ? Y aurait-il une obligation de moyens et/ou de résultats ?

	Audit financier	Audit pour les IA
Efforts	Honoraires annuels	Quels montants d'honoraires soutenables pour quel niveau de confiance ?
Bonnes pratiques concernant l'objet audité	Principes comptables	Restent encore à identifier mais quelques principes et bonnes pratiques d'IA commencent déjà à être formulés.
Bonnes pratiques concernant les moyens mis en œuvre par les organisations conduisant à un objet de qualité	Contrôle interne : COSO, COBIT 5…	Reste encore à identifier mais ces bonnes pratiques pourraient concerner : • les principes, directives, référentiels utilisés • les structures organisationnelles mises en œuvre • la culture, l'éthique et les comportements (incitations et aspects dissuasifs) mis en place • l'information utilisée • les outils et services à disposition (infrastructures, applications…) • les aptitudes, compétences, savoir-faire en action • les processus de gouvernance, de management et opérationnels mis en œuvre

		• les indicateurs de performance de résultats et avancés (de moyens) permettant de s'assurer que les attentes et les objectifs des parties prenantes ont été atteints et que les moyens mis en œuvre sont efficaces et efficients Ces bonnes pratiques doivent couvrir l'ensemble du cycle de vie : planification, conception, acquisition/développement, exploitation, suivi/évaluation, mise à jour puis destruction. On peut déjà s'appuyer sur les référentiels de COSO, COBIT 5… pour finaliser les référentiels applicables aux IA

	Audit par le CAC	Audit pour les IA
Bonnes Pratiques concernant les travaux d'audit	Standards d'audit	Il faut identifier la démarche et le référentiel de bonnes pratiques concernant les travaux d'audit qui permettraient de fournir une opinion ou un label de qualité de l'objet et des qualités/caractéristiques audités avec un niveau de risque acceptable quant à la fiabilité de cette opinion/label et avec un niveau d'efforts acceptable notamment un coût soutenable. En fonction de la nature de l'opinion à fournir, ces bonnes pratiques pourraient concerner : • les principes, directives, référentiels d'audit utilisés • les structures organisationnelles d'audit mises en œuvre • la culture, l'éthique et les comportements (incitations et aspects dissuasifs) d'audit mis en place • l'information utilisée dans l'audit • les outils et services à disposition de l'audit (infrastructures, applications, outils pour les tests…) • les aptitudes, compétences, savoir-faire d'audit en action (certifications nécessaires : commissaire aux algorithmes ?) • les processus de gouvernance, de management et opérationnels d'audit mis en œuvre

		• les indicateurs de performance de résultats et avancés d'audit (de moyens) permettant de s'assurer que les attentes et objectifs d'audit des parties prenantes ont été atteints et que les moyens d'audit mis en œuvre sont efficaces et efficients[59] Ces bonnes pratiques doivent couvrir l'ensemble du cycle de vie de l'audit (planification, conception, acquisition ou développement, exploitation, suivi et évaluation, mise à jour puis destruction). Cela permettrait de s'assurer que l'audit est bien fait et servir d'assurance qualité. On peut aussi envisager des pratiques d'audit différentes en fonction des particularités des IA auditées comme dans le contexte d'algorithmes apprenants. On pourrait envisager de se baser à minima sur les standards d'audit existants notamment en termes de nature d'éléments probants.

Pour chacune de ces questions, il n'y a pas encore de réponse précise du marché ni de référentiels professionnels largement reconnus. Dès lors, on ne peut pas, à l'heure actuelle en 2022, donner de certification globale du type « certifier que l'IA x est digne de confiance » mais uniquement des certifications spécifiques ou des opinions circonstanciées.

1.5. Autres problématiques associées à l'audit et la certification des IA

S'il faut résoudre l'ensemble des questions précédentes pour parvenir à réaliser des audits sereinement, il existe encore quelques autres problématiques :

- **Une attention particulière sur le type de communication des résultats**

  C'est un point critique dont la maîtrise est essentielle à la mise en place d'IA dignes de confiance. En effet, en raison de la sensibilité du sujet, de l'appréhension des différents publics vis-à-vis de l'IA, de la complexité de compréhension, du fort niveau des enjeux et risques associés et des possibilités de confusion, il est tout à fait possible que les résultats d'un audit soient positifs mais que le public ou certaines parties prenantes ne le considèrent pas comme tel ou, inversement, que les limites des résultats n'aient pas été perçues comme telles. Cela a été assez visible dans le cas de l'audit de

[59] Attention, bien que cette liste soit très proche de la précédente, elle se réfère uniquement aux procédures d'audit en lui-même alors que la liste précédente s'intéresse aux bonnes pratiques de l'IA seulement.

Parcoursup où la Cour des comptes avait noté dans son rapport une communication souvent biaisée à son sujet.

Ainsi, il faudra étudier pour chaque type d'audit, à qui sont adressés les résultats, quels types de restitutions pour chaque type de population, quelles sont les formulations acceptables et sous quelles formes elles sont acceptables, quel niveau de détail, comment faire part de réserves, de faiblesses, des limites d'une intervention, du niveau de confiance *in fine*...

- **La difficulté aujourd'hui de réaliser des certifications holistiques[60] d'IA**

  Compte tenu de la complexité et du périmètre très large des IA, d'un nombre important d'exigences à atteindre, de la présence de plusieurs risques variés, de la non-maturité des outils et des bonnes pratiques d'IA et des audits pour les certifier... les conditions ne semblent pas réunies pour pouvoir émettre une opinion globale du type « l'IA x est digne de confiance » ou « l'IA x est éthique ». Tout d'abord, les propositions de valeurs n'ont pas été établies : *quid* de la faisabilité technique et financière d'un audit d'un système d'IA ? De plus, les structures institutionnelles et/ou professionnelles qui ont défini des normes et des qualifications sont rares. Certaines travaillent sur le sujet comme LNE.

  Comme nous l'avons vu plus haut en prenant l'exemple des CAC, toutes les conditions ne sont pas encore réunies pour fournir des certifications holistiques. Il est nécessaire que plus d'instances au niveau mondial définissent les certifications à réaliser. Il faudra notamment établir des référentiels techniques, de gouvernance, de management et d'audit, tous adaptés à l'IA, largement diffusés et acceptés. Cela ne pourra se faire que grâce à la réalisation de missions pilotes.

2. Quelques recommandations : illustrations d'opinions d'audit, outils et guides d'audit, instances professionnelles, audits pilotes

Illustration des types d'opinions susceptibles d'être fournis à l'issue d'un audit

Pour résoudre les problématiques soulevées précédemment, il faudra en premier lieu définir ce qu'on cherche à accomplir. En d'autres mots : à quoi ressemble ce qu'on veut faire ?

Il faudrait d'abord définir quels types d'audit on attend, et quels sont les buts de l'audit. En particulier, il faut définir les affirmations à certifier dans l'audit et le niveau de confiance attendu. Commencer par la définition des éléments de sortie permettrait de s'assurer que les audits répondent à des besoins réels et à des tarifs soutenables.

Pour faciliter la définition de ces objectifs, il serait utile d'illustrer les types d'opinions possibles à l'issue d'un audit. On pourra alors voir le type d'apport de confiance correspondant et si cela convient.

---

[60] Le Centre National des Ressources Textuelles et Lexicales définit ce terme comme une doctrine ou un point de vue qui consiste à considérer les phénomènes comme des totalités, comme une seule unité.

Cela pourrait s'illustrer sur deux types de systèmes d'IA, l'un dans un contexte assez simple, tel que pourrait l'être Parcoursup, et l'autre qui soit l'un des plus complexes, comme la voiture autonome.

Développement d'outils adaptés et élaboration de guides d'audit

Il faudrait également viser à l'élaboration de guides d'audit pour les IA, de réaliser différents outils « prêts à utiliser » tels que des questionnaires, des check-lists, des matrices… et de largement les partager. Cela permettrait d'aider les auditeurs à mener leurs missions dans les meilleures conditions compte tenu des difficultés liées aux spécificités des IA.

Mise en place d'une instance professionnelle pour les IA dignes de confiance

Enfin, il faudrait mettre en place des instances professionnelles en charge de définir les modalités d'exercice de ces missions telles que la composition de l'équipe qui doit réaliser la mission d'audit, les qualifications (diplômes, certifications, expérience…) requises par les auditeurs pour pouvoir réaliser des missions d'audit de l'IA, les référentiels d'IA acceptés et les référentiels d'audit acceptés, les types d'opinion et labels acceptés, les types de réserves qu'il est possible d'émettre (pour qui et dans quels contextes (hauts risques…)) ainsi que les contrôles et les sanctions. Ces instances pourraient également mettre en place un système de partage des expériences (incidents et autres) de façon à créer un écosystème de l'IA qui inspire confiance.

Il faudrait que les modalités d'exercice de ces missions prennent bien en compte les éléments suivants pour que les audits soient pleinement réussis :

- la démarche à mettre en œuvre doit être globale et pas seulement technique
- la démarche doit être systémique : en IA, la modification d'un élément peut avoir bien des répercussions sur d'autres éléments qu'on aurait imaginés indépendants
- la démarche peut être linéaire dans certains cas mais pas dans tous
- une démarche standard est possible mais ne peut s'appliquer à tous les audits. Il faut montrer une certaine souplesse et l'adapter aux contextes et cas spécifiques
- tous ces éléments doivent être intégrés sous peine de n'avoir que des morceaux de réponses épars.

Procéder à des audits pilotes

Nous avons vu qu'il est à ce stade difficile de procéder à des audits holistiques mais la possibilité de fournir des opinions circonstanciées sur :

- certaines assertions (vie privée, sécurité, transparence…)
- certains objets (algorithmes, données…)
- certains leviers (structures organisationnelles, outils, technologies, directives, processus…).

Cette possibilité permettra de pouvoir commencer à réaliser des audits « locaux », c'est-à-dire avec un périmètre plus restreint, qui seront néanmoins utiles dans l'atteinte in fine d'IA dignes de confiance.

Créer des IA dignes de confiance, c'est bien. Le prouver au public, au gouvernement et au reste des parties prenantes, c'est mieux. L'intervention d'un tiers de confiance indépendant qui vérifierait les assertions des entreprises ou de toute autre organisation serait une solution idéale. Ce concept présente étrangement de fortes similitudes avec un système déjà existant : les audits et certifications. On peut même faire une comparaison avec l'audit comptable et financier, qu'il soit réalisé par un expert-comptable ou un commissaire aux comptes et s'inspirer des méthodologies mises en place pour débuter les audits et certifications de systèmes à base d'intelligence artificielle. Toutefois, il n'est pas possible de dupliquer exactement le système : un système d'information financier ou comptable est trop différent. Il faudra donc définir de nombreuses caractéristiques de la méthodologie d'audit à mettre en œuvre avant d'obtenir un processus de vérification fiable.

Le premier élément à définir sera celui de la création de valeur produite par ces audits en trouvant l'équilibre entre le niveau d'exigence attendue, la nature de l'opinion fournie, le niveau de risque de l'audit et le niveau d'effort à fournir par l'organisation. Pour cela, il faudra également clarifier un certain nombre de caractéristiques de l'audit comme le périmètre des objets à auditer, le niveau de confiance attendu des assertions à auditer, comment diffuser les résultats... Ces réponses pourraient varier en fonction du type d'audit ou de certification qu'on désire mener. Un audit ayant pour but de valider le fond ne répondra pas de la même façon aux problématiques qu'un autre qui doit seulement valider la forme.

Tant que l'ensemble de ces questions n'auront pas trouvé une réponse, il sera impossible de certifier un système d'information à base d'IA de manière globale. Au mieux, il sera possible d'en certifier un ou plusieurs composants. Pour parvenir à des certifications holistiques, il faudra que des instances respectées au niveau mondial définissent les certifications à réaliser et leurs caractéristiques. Il faudra aussi créer et diffuser des référentiels techniques, de gouvernance, de management et d'audit qui soient reconnus et acceptés par les parties prenantes. Tout cela mettra un certain temps à faire et ne pourra se développer qu'en ayant le retour sur expérience de missions pilotes.

CONCLUSION

Les systèmes à base d'intelligence artificielle sont porteurs de nombreux espoirs. Ils pourraient aider à la résolution des petits et des grands problèmes de ce siècle en apportant des solutions concrètes comme l'agriculture de précision, la maintenance prédictive, le smart grid, l'aide à la décision et bien d'autres. Mais ce futur n'apparait pas dénué de menaces et de défis. Quel sera l'impact de l'IA sur le marché du travail ? Cela accentuera-t-il les inégalités entre les différentes nations ou au sein d'une même société ?

Si les doutes sont trop grands, la défiance envers les IA pourrait s'accroître jusqu'à ce que la population rejette totalement cette technologie. Il apparaît nécessaire de rendre les IA dignes de confiance pour qu'elles puissent être acceptées, utilisées à leur plein potentiel et remplir finalement leurs promesses.

Mais l'objectif de rendre les IA dignes de confiance est très difficile. Il nécessite de comprendre le concept même d'IA, qui est complexe et multiforme. Le terme IA peut désigner des solutions très différentes les unes des autres (il y a peu de points communs entre une voiture autonome et Parcousup) que ce soit dans leur but, les algorithmes employés, le matériel ou les données utilisés.

Conscientes de ce défi, de plus en plus de règlementations sont votées mais elles sont encore incomplètes et balbutiantes.

Le groupe a donc défini une démarche en cinq étapes qui, étudiées et mises en œuvre dans l'ordre, permet de formaliser les problématiques et de commencer à y répondre. L'atteinte d'IA dignes de confiance ne pourrait se faire qu'en suivant ce type de démarche :

- la première étape est de définir les principes et les exigences à atteindre que devrait respecter un système à base d'intelligence artificielle pour être digne de confiance. Une partie de ces exigences sont identiques à celles existantes pour les systèmes d'informations traditionnels – sans intelligence artificielle – mais présente des spécificités quand elles sont appliquées à un système d'IA. Il s'agit des exigences de vie privée et RGPD, de fiabilité, robustesse, résilience et de cybersécurité. D'autres de ces exigences ne peuvent pas exister hors d'un système d'IA. Il s'agit des exigences de transparence et d'explicabilité, d'équité et de non-discrimination et d'humanité. Chacune de ces exigences porte des problématiques spécifiques qui nécessiteront des réponses appropriées
- la seconde étape concerne la gestion des risques et de la prise de risque. Une certaine part de la gestion des risques déjà rencontrés dans les systèmes traditionnels peuvent s'appliquer dans un système d'IA mais d'autres sont inédites. C'est le cas de la gestion de prise de risque. En raison de l'immaturité technico-scientifique des IA, beaucoup de leurs comportements potentiels sont encore inconnus. Ces comportements pourraient être préjudiciables. Mais n'ayant pas d'informations sur l'existence de ces comportements, il faudrait mettre en place une gestion spécifique de la prise de risque sur ces risques inconnus. Plusieurs principes à respecter pour les systèmes à base d'IA ont été identifiés. Il s'agit par exemple des principes de minimisation des risques, de précaution, de non-malfaisance, de vigilance, de proportionnalité et d'alerte

- la troisième étape concerne la gouvernance et la responsabilité des acteurs de l'écosystème des IA. Des organisations spécifiques devront définir les niveaux de risques tolérables et acceptables pour les parties prenantes, sachant que ces dernières ont des intérêts divergents. Il faudra créer ces organisations et leur attribuer des missions et des pouvoirs pour leur permettre d'accomplir leur but
- la quatrième étape est pratique : il faudra mettre en œuvre les dispositifs permettant de suivre les principes et les exigences à atteindre en respectant les limites imposées par la gestion des risques et de la prise de risque. Au fur et à mesure que des dispositifs seront inventés et testés, un certain nombre d'entre eux émergeront en tant que bonnes pratiques, c'est-à-dire des pratiques opérationnelles, de management et de gouvernance les plus efficaces pour atteindre les exigences et se conformer aux principes. Des référentiels de bonnes pratiques pourront alors être adoptés
- l'étape finale devrait permettre aux parties prenantes de vérifier qu'un système à base d'intelligence artificielle répond bien aux exigences et suit bien les principes. Cette vérification pourrait prendre diverses formes (audit, certifications, attestations, labels...) et répondre à des caractéristiques différentes en fonction du type de vérification qu'on cherche à établir. Tout comme pour les IA, il convient de déterminer les exigences de ces audits et certifications, puis les risques de l'audit, les options de gouvernance et les pratiques d'audits. Autrement dit, il convient de réappliquer à l'audit de l'IA la cascade qui devrait être suivie pour les IA elles-mêmes.

Le suivi de cette démarche permettrait aux acteurs de disposer d'un cadre commun de compréhension des différents enjeux et spécificités des systèmes à base d'intelligence artificielle, d'établir les qualités et caractéristiques attendues des IA, des référentiels de bonnes pratiques qui permettraient de concevoir un système d'IA si elles sont correctement appliquées. Cela permettrait également de disposer de référentiels de bonnes pratiques d'audit et de certifications qui soient spécifiques à l'audit des IA et fassent écho à leurs caractéristiques. Ces bonnes pratiques d'audit mises en œuvre par des auditeurs spécialisés, des commissaires aux IA, permettraient aux acteurs de l'écosystème des IA d'avoir confiance dans les audits et certifications mis en œuvre pour vérifier que les IA respectent les principes et atteignent les exigences.

ANNEXES

1. Glossaire

Quelques termes sont utilisés dans ce document de manière très spécifique et parfois de manière différente du langage commun. Ces définitions ont été expliquées dans le corps du texte mais sont réunies ici à des fins de praticité :

- problématique : une problématique est un ensemble de questions et de problèmes concernant un domaine de connaissances ou qui sont posés par une situation. Une problématique n'est pas forcément négative
- levier : un levier est un moyen d'action pour atteindre un but. Plusieurs types de leviers peuvent être mobilisées. Nous avons retenu les sept types de leviers définis dans COBIT qui peuvent interagir entre eux :
  - › les principes, les directives et les cadres de référence qui représentent le véhicule permettant d'orienter les décisions et pratiques à mettre en œuvre
  - › les processus qui sont des ensembles d'activités corrélées ou en interaction utilisant des éléments en entrée pour produire des résultats escomptés
  - › les structures organisationnelles qui sont les entités clés concourant à la prise de décision et à l'action
  - › la culture, l'éthique et le comportement des individus et d'une organisation qui sont la manière d'être, d'agir ou de réagir face à une situation en fonction du contexte
  - › les informations utilisés et produites par une organisation qui permettent d'agir et prendre des décisions
  - › les services, l'infrastructure, les applications et les autres outils employés par une organisation pour atteindre ses objectifs
  - › le personnel, les aptitudes et les compétences qui sont sollicités dans le cadre des actions et des décisions.
- pratique : une pratique est une manière habituelle concrète d'exercer une activité ou des activités. C'est la mise en action des règles, des principes, des techniques...
- bonne pratique : une bonne pratique est une manière habituelle concrète éprouvée d'exercer une activité ou des activités qui a été utilisée avec succès par plusieurs organisations et dont il a été démontré qu'elle produit des résultats fiables. Il s'agit de repères, d'orientations ou de pistes pour cette activité qu'elle soit de gouvernance, de management ou opérationnelle. Il existe notamment des bonnes pratiques concernant les sept leviers d'action décrites ci-dessus. Ces bonnes pratiques se retrouvent le plus souvent dans des référentiels de bonnes pratiques du type COBIT, ITIL, ISO...

- système d'information traditionnel : il s'agit d'un système d'information qui ne contient pas d'éléments d'intelligence artificielle
- dispositif : un dispositif est un ensemble articulé de mesures prises et de moyens mis en œuvre pour atteindre un objectif
- contrôle interne : le contrôle interne est l'ensemble des dispositifs mis en œuvre par une organisation destiné à fournir une assurance raisonnable quant à la réalisation de ses objectifs Il inclut notamment les dispositifs qui permettront de prévenir, de détecter et de corriger les évènements susceptibles d'empêcher la réalisation de ces objectifs
- management : le management est la mise en œuvre de l'ensemble des dispositifs d'une organisation pour atteindre ses objectifs convenus préalablement. Ses principales fonctions sont :
  - › d'aligner ces dispositifs sur les orientations fixés par les structures de gouvernance, les planifier et les organiser
  - › de les développer, les acquérir et les implanter
  - › de les exploiter et les soutenir
  - › de les suivre, les évaluer et rétroagir si nécessaire.

  Il s'agit des 4 activités de management définies par ISO : Plan, Build, Run, Monitor (PBRM).
- gouvernance : la gouvernance est la mise en œuvre de l'ensemble des dispositifs d'une organisation pour déterminer les contributions de valeur attendues pour les parties prenantes et fixer les objectifs correspondants. Ses principales fonctions sont de permettre à de multiples parties prenantes :
  - › D'évaluer leurs différents besoins, conditions et options pour créer de la valeur
  - › De les prioriser, et d'arbitrer et fixer les orientations
  - › De suivre l'atteinte des objectifs convenus d'avance et de rétroagir si nécessaire

  Il s'agit des 3 activités de gouvernance définies par ISO : Evaluate, Direct, Monitor (EDM).
- création de valeur : la création de valeur pour une ou plusieurs parties prenantes est la création d'avantages pour elles en prenant des niveaux de risques acceptables et en utilisant les ressources limitées de manière soutenable et responsable. Chaque proposition de création de valeur intègre de manière différenciée ces trois éléments de la création de valeur : avantages, risques et efforts
- audit : un audit est une analyse, une évaluation et une vérification formelle d'un ou plusieurs aspects précis d'une organisation par rapport à un référentiel. Ce référentiel peut être l'expérience propre de l'intervenant. Il peut s'agir d'un intervenant interne ou externe à une organisation
- certification : une certification est un audit effectué par un intervenant agréé indépendant qui donne une assurance formelle (une opinion écrite) quant à une assertion (une affirmation concernant un objet), par rapport aux exigences prévues dans un référentiel agréé (une norme) ou dans une règlementation. L'agrément des intervenants et des référentiels est encadré par des instances professionnelles ou étatiques (ISO, AFNOR...)

- label : un label est un signe distinctif apposé sur un objet que ce soit une organisation, un produit, un service... qui permet de lui reconnaître certaines qualités ou caractéristiques par rapport à des normes prédéterminées. La différence avec une certification est que l'agrément des intervenants et des référentiels ne sont pas nécessairement encadrés par des instances professionnelles ou étatiques.

2. Explications introductives de quelques termes techniques d'IA

Plusieurs termes techniques de l'IA ont été abordés au sein de ce texte. Ces termes sont expliqués plus en détail ici :

- apprentissage supervisé : apprentissage automatique dans lequel l'algorithme s'entraîne à une tâche déterminée en utilisant un jeu de données assorties chacune d'une annotation indiquant le résultat attendu. L'apprentissage supervisé recourt le plus souvent aux réseaux de neurones artificiels et est utilisé, par exemple, pour la reconnaissance d'images et la traduction automatique
- apprentissage non supervisé : apprentissage automatique dans lequel l'algorithme utilise un jeu de données brutes et obtient un résultat en se fondant sur la détection de similarités entre certaines de ces données. L'apprentissage non supervisé est utilisé, par exemple, pour l'identification de comportements et la recommandation d'achats
- apprentissage semi-supervisé : algorithme d'apprentissage à partir de données partiellement étiquetées qui exploite la similarité entre les données pour leur attribuer des étiquettes. Un algorithme non-supervisé de groupage identifie des groupes, puis attribue une étiquette à chacun des groupes pour étiqueter tous les autres membres de chacun de ces groupes[61]
- apprentissage par renforcement : apprentissage automatique dans lequel un programme extérieur évalue positivement ou négativement les résultats successifs de l'algorithme, l'accumulation des résultats permettant à l'algorithme d'améliorer ses performances jusqu'à ce qu'il atteigne un objectif préalablement fixé. L'apprentissage par renforcement est fréquemment utilisé dans la robotique et a été attestée dans certains jeux stratégiques comme le jeu de go
- apprentissage profond (Deep learning) : apprentissage automatique qui utilise un réseau de neurones artificiels composé d'un grand nombre de couches dont chacune correspond à un niveau croissant de complexité dans le traitement et l'interprétation des données. L'apprentissage profond est notamment utilisé dans la détection automatique d'objets au sein d'images et dans la traduction automatique[62]

---

[61] https://datafranca.org/wiki/Accueil

[62] Commission d'enrichissement de la langue française, Vocabulaire de l'intelligence artificielle, Avis NOR : CTNR1832601K, JO du 9 décembre 2018[62]

- arbre décisionnel : outil d'aide à la décision sous la forme graphique d'un arbre avec sa racine en haut ; les différentes décisions possibles étant situées aux extrémités des branches et sont adoptées en fonction de la décision prise à chaque étape. Cette méthodologie est plus communément appelée arbre de classification à partir des données[63]
- arbre de recherche : graphe arborescent qui indique les règles appliquées durant une recherche, les nœuds explorés et les résultats obtenus
- autocomplétion : fonctionnalité qui propose des mots à l'utilisateur à partir des premiers caractères saisis[64]
- boîte noire : si les réseaux de neurones profonds obtiennent souvent des résultats supérieurs, ils ont cependant un point faible : leur fonctionnement apparaît comme bien plus opaque. Un phénomène appelé « boîte noire » (black *box)*, dans le sens où l'on peut juger des données qui entrent dans la boîte et des résultats qui en sortent, mais sans savoir ce qui se passe à l'intérieur
- chatbot (Fr, Dialogueur) : logiciel spécialisé dans le dialogue en langage naturel avec un humain, qui est capable notamment de répondre à des questions ou de déclencher l'exécution de tâches. Un dialogueur peut être intégré à un terminal ou à un objet connecté. Les dialogueurs sont utilisés, par exemple, dans les techniques de vente, les moteurs de recherche et la domotique.
- coque mécatronique : la norme NF E 01-010 (2008) définit la mécatronique comme une « démarche visant l'intégration en synergie de la mécanique, l'électronique, l'automatique et l'informatique dans la conception et la fabrication d'un produit en vue d'augmenter et/ou d'optimiser sa fonctionnalité ». Les robots sont constitués de matériel mécatronique
- Claims Argument Evidence : la méthodologie de la société Adelard qui sert à présenter des arguments de sécurité et la notation de structuration des objectifs (comme le Goal Structuring Notation qui est un est un argument graphique utilisé pour documenter et présenter la preuve que les objectifs de sécurité ont été atteints)
- data scientiste : personne spécialisée dans l'exploration, l'analyse et l'interprétation des données, et qui a pour tâche d'orienter les actions et les prises de décisions d'une organisation
- data analyste : spécialiste chargé d'analyser l'information pour cerner les besoins d'une entreprise, de développer des modèles informatiques logiques et d'en effectuer la maintenance
- data architecte : expert en informatique qui a la responsabilité de s'assurer que les objectifs stratégiques d'une organisation sont optimisés à travers l'utilisation de standards de données d'entreprise. Cela implique souvent la création et la mise à jour d'un registre de métadonnées centralisé

[63] https://datafranca.org/wiki/Accueil
[64] https://dictionnaire.lerobert.com/definition/autocompletion

- ingénieurs machine learning : expert en sciences de données chargé d'évaluer et de concevoir des systèmes d'apprentissage automatique et de mettre en production des modèles d'intelligence artificielle
- ingénieurs IA : Ingénieur spécialisé dans l'intelligence artificielle, qui modélise et conçoit des machines intelligentes pouvant aider les utilisateurs dans leur travail ou leur vie quotidienne
- IoT (Internet of Things) : ensemble des objets connectés à Internet capables de communiquer avec des humains, mais aussi entre eux, grâce à des systèmes d'identification électronique, pour collecter, transmettre et traiter des données avec ou sans intervention humaine
- journal de logs : fichier contenant l'enregistrement séquentiel de tous les événements affectant un processus particulier (applications, activité d'un réseau informatique...). Généralement datés et classés par ordre chronologique, ces derniers permettent d'analyser pas à pas l'activité interne du processus et ses interactions avec son environnement[65]
- machine learning : champ d'étude de l'IA qui se fonde sur des approches statistiques pour donner aux ordinateurs la capacité d'apprendre à partir d'entrées de données (plutôt que d'instructions explicitement programmées) pour améliorer leurs performances et résoudre des tâches. L'apprentissage automatique comporte généralement deux phases :
  - l'estimation du modèle devant résoudre une tâche pratique, telle que traduire un discours, estimer une densité de probabilité, reconnaître la présence d'un chat dans une photographie ou participer à la conduite d'un véhicule autonome
  - la mise en production où, le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée. En pratique, certains systèmes peuvent poursuivre leur apprentissage une fois en production.
- marketing prédictif (Ex. Modélisation du Marketing Mix) : technique d'analyse statistique qui permet d'estimer l'effet de diverses tactiques de marketing (combinaison marketing) sur les ventes et de prévoir l'effet de futurs ensembles de tactiques. Elle est souvent utilisée pour optimiser la politique publicitaire et les tactiques promotionnelles par rapport au chiffre d'affaires ou au bénéfice
- meta learning : la notion de méta-apprentissage réfère aux algorithmes de haut niveau qui peuvent apprendre à partir d'algorithmes d'apprentissage. Toute forme d'apprentissage à partir d'informations sur les processus et modèles d'apprentissage peut être appelée méta-apprentissage
- moteurs de règles : en informatique, un moteur de règles est un système logiciel qui exécute une ou plusieurs règles métiers dans un environnement de production. Ces règles peuvent venir de la législation, de politiques applicables ou d'autres sources
- ordinateur quantique : un ordinateur quantique est l'équivalent des ordinateurs classiques mais qui effectuerait ses calculs en utilisant directement les lois de la physique quantique et, à la base, celle dite de superposition des états quantiques. Alors qu'un ordinateur classique manipule des bits

[65] Wikipédia

d'information, qui sont soit des 0 soit des 1, un ordinateur quantique utilise des qubits. Ceux-ci sont des généralisations des bits classiques, qui sont en quelque sorte une superposition simultanée de ces deux états, comme peut l'être, par exemple, un état de spin pour un photon ou un électron

- ordinateur optique : un ordinateur optique (ou ordinateur photonique) est un ordinateur numérique qui utilise des photons pour le traitement des informations, alors que les ordinateurs conventionnels utilisent des électrons. Les photons ont la particularité de ne pas créer d'interférence magnétiques, de ne pas générer de chaleur et de se propager très rapidement. Les transistors optiques sont beaucoup plus rapides que les transistors électroniques. Des ordinateurs optiques pourraient être plus puissants que les ordinateurs conventionnels actuels
- outils de prévision automatique : ces outils permettent de prévoir une tendance future en utilisant les données historiques de l'entreprise. Ces outils peuvent avoir des buts variés tels que prévoir la demande (en matières premières, en stocks, en main-d'œuvre...) ou la croissance (en ventes, en revenus). Ces outils sont applicables dans de très nombreux secteurs comme la finance, l'industrie, les soins de santé, les instances gouvernementales...
- outils de segmentation automatique : un outil de segmentation d'IA est presque un pléonasme puisque l'essence de l'IA est de catégoriser. Ces outils permettent par exemple dans un contexte marketing de catégoriser les types de visiteurs d'un site en se basant sur leur comportement, dans un contexte de reconnaissance d'images satellitaire de distinguer les habitations avec une piscine de celles qui n'en ont pas...
- raisonnement automatique : le raisonnement automatique (ou automatisé) est un domaine de l'informatique et de la logique mathématique dédié à la compréhension des différents aspects du raisonnement. Bien que le raisonnement automatisé soit considéré comme un sous-domaine de l'intelligence artificielle, il a également des liens avec l'informatique théorique et même la philosophie.
- reconnaissance de la parole : la reconnaissance automatique de la parole (souvent improprement appelée reconnaissance vocale) est une technique informatique qui permet d'analyser la voix humaine captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine. La reconnaissance de la parole, ainsi que la synthèse de la parole, l'identification du locuteur ou la vérification du locuteur, font partie des techniques de traitement de la parole. Ces techniques permettent notamment de réaliser des interfaces homme-machine (IHM) où une partie de l'interaction se fait à la voix : « interfaces vocales ».
- réseaux multi-agents (Ex. Réseaux « feedforward » multi-couches) : les réseaux bouclés sont des réseaux de neurones avec une structure de couches distincte, pour toutes les connections alimentant les entrées aux sorties. Le terme réseaux bouclés est parfois utilisé comme synonyme des perceptrons multi-couches
- réseaux de neurones : ensemble d'algorithmes, modélisés librement d'après le cerveau humain, qui sont conçus pour reconnaître les modèles. Ils interprètent les données sensorielles à travers une sorte de perception machine, d'étiquetage et de partitionnement des entrées brutes. Les motifs qu'ils

reconnaissent sont numériques, contenus dans des vecteurs, dans lesquels toutes les données du monde réel, qu'il s'agisse d'images, de sons, de textes ou de séries temporelles, doivent être traduites

- vision artificielle : la vision par ordinateur (aussi appelée vision artificielle ou vision numérique) est une branche de l'intelligence artificielle dont le principal but est de permettre à une machine d'analyser, traiter et comprendre une ou plusieurs images prises par un système d'acquisition
- traduction automatique : technique informatique permettant d'obtenir de façon automatique, sans l'intervention d'un traducteur humain, une traduction de textes d'une langue source vers une autre langue
- spintronique : mariage de l'électronique, qui utilise la charge électrique des électrons pour transmettre de l'information, et du spin, une autre propriété intrinsèque des électrons. Le spin est une caractéristique microscopique purement quantique, qui n'a pas d'équivalent à notre échelle[66]
- stockage ADN (données Génomiques) : données relatives au génome et à l'ADN d'un organisme. Elles sont utilisées en bio-informatique pour collecter, stocker et traiter les génomes des êtres vivants. Les données génomiques nécessitent généralement une grande quantité de stockage et un logiciel spécifique pour les analyser
- stockage holographique : technique de mémoire de masse utilisant l'holographie pour stocker de hautes densités de données dans des cristaux ou des polymères photosensibles
- stockages magnétiques (HAMR, BPM, SMR) : systèmes qui conservent des informations ou des données sur un matériau aimanté. La technologie SMR (Shingled Magnetic Recording) prend le parti de faire se chevaucher les pistes, regroupées en une bande et ponctuées d'un sillon plus large pour l'écriture. Une innovation qui repose sur le principe que les pistes sont moins larges en lecture qu'en écriture : elles sont coupées ou couvertes au fur à et à mesure de l'inscription des informations, sans risque pour leur intégrité. En resserrant ainsi les pistes, la densité surfacique est augmentée de 25% ; HAMR (Heat Assisted Magnetic Recorded ou enregistrement magnétique assisté par la chaleur) chauffe l'emplacement de l'écriture sur le plateau via un faisceau laser pour le rendre plus « magnétisable ». Les bits sont rétrécis et la consommation d'énergie est réduite ; la méthode BPMR (Bit-Patterned Magnetic Recording ou enregistrement magnétique à motifs binaires) quant à elle prévoit de réduire chaque bit à un unique grain magnétique, contre une vingtaine aujourd'hui. Ce processus lithographique nécessite de nouvelles surfaces magnétiques, mais préfigure une densité de 10 térabits par pouce carré ! On parlera alors de technologie HDMR (Heated-Dot Magnetic Recording)[67]

[66] https://lejournal.cnrs.fr/articles/les-nouveaux-defis-de-la-spintronique

[67] https://www.data-labcenter.fr/news-et-reportages/reportages-dexperts/technologies-pmr-smr-tdmr-mamr-hamr-bpmr-des-disques-durs/

- stockage optique 3D : le stockage optique 3-D désigne toute forme de stockage d'information optique dans lesquelles les informations peuvent être enregistrées et / ou de lecture avec trois dimensions de résolution optique (par opposition aux résolutions bidimensionnelles offertes, par exemple, par le disque compact). Cette innovation pourrait fournir un stockage de mémoire de niveau pétabit sur des disques de la taille d'un DVD. L'enregistrement des données et leur lecture sont faits en y concentrant des lasers. Toutefois, en raison du caractère volumétrique de la structure de données, la lumière du laser doit traverser d'autres points de données avant qu'elle n'atteigne le point où la lecture ou l'enregistrement est souhaité. Par conséquent, une sorte de non-linéarité est nécessaire pour veiller à ce que ces autres points de données n'interfèrent pas avec l'adressage du point désiré
- systèmes experts : nés de technologies issues de l'IA, la linguistique, l'ergonomie et internet, ces systèmes recouvrent des applications qui vont de la modélisation des connaissances des entreprises au diagnostic médical ou technique, en passant par les agents intelligents et les moteurs de recherche associatifs du web
- véhicules autonomes : un véhicule autonome est un véhicule automobile apte à rouler, sur route ouverte, sans intervention d'un conducteur. Le concept vise à développer et produire un véhicule pouvant réellement circuler sur la voie publique dans le trafic sans intervention humaine en toutes situations, à terme. C'est une application typique du domaine de la robotique mobile dans laquelle de nombreux acteurs sont engagés. Néanmoins de nombreuses questions techniques, légales, psychologiques et juridiques restent non résolues pour l'instant.

3. Sources

- Alexandra Ebert, « We want fair AI algorithms – but how to define fairness? » Mostly AI, https://mostly.ai/blog/we-want-fair-ai-algorithms-but-how-to-define-fairness, 2020 consulté en 2022
- Alice Vitard, « Le parlement européen dit non à l'utilisation de la reconnaissance faciale par la police », L'usine digitale, https://www.usine-digitale.fr/article/le-parlement-europeen-dit-non-a-l-utilisation-de-la-reconnaissance-faciale-par-la-police.N1147532 2021, consulté en 2021
- Aurélie Jean, « De l'autre côté de la machine », 2019
- Aymeric Poulain Maubant, « Pour y voir plus clair sur les notions de transparence et d'explicabilité en IA », Medium, https://medium.com/@AymericPM/pour-y-voir-plus-clair-sur-les-notions-de-transparence-et-dexplicabilit%C3%A9-en-ia-c0db2e96ae62, 2020, Consulté en novembre 2021
- Buster Benson, « Cognitive bias cheat sheet », BetterHumans, https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18, 2016, consulté en octobre 2021
- C3.ai, « What is Local Interpretable Model-Agnostic Explanations (LIME) ? », https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/, consulté en décembre 2021
- Charte des droits fondamentaux de l'Union européenne, 2016/C 202/2
- Clémence Maquet, « Intelligence artificielle : quelle approche des biais algorithmiques ? » Siècle Digital, 2021
- Collectif, « Future of life Institute 2017 Asilomar Conference » ai-ethics.com https://ai-ethics.com/2017/08/11/future-of-life-institute-2017-asilomar-conference/ consulté le 15/02/2022
- Collectif, « Microsoft AI Principles », https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6 consulté le 01/02/2022
- Comité éthique et scientifique de Parcoursup, Rapport au Parlement, Janvier 2020, https://services.dgesip.fr/fichiers/Rapport_du_CESP_2019__janvier_2020_.pdf
- Commission européenne, « Annexes à la Proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'union. {SEC (2021) 167 final} - {SWD (2021) 84 final} - {SWD (2021) 85 final} », 2021
- Commission européenne, « Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions; Fostering a European approach to Artificial Intelligence » 2021
- Commission européenne, « Livre blanc intelligence artificielle, Une approche européenne axée sur l'excellence et la confiance », version du 19/02/2020

- Commission européenne, « Proposition de règlement du parlement européen et du conseil établissant des règles harmonisées concernant l'intelligence artificielle (législation sur l'intelligence artificielle) et modifiant certains actes législatifs de l'union. {SEC (2021) 167 final} - {SWD (2021) 84 final} - {SWD (2021) 85 final} », 2021
- Cour des comptes « Un premier bilan de l'accès à l'enseignement supérieur dans le cadre de la loi Orientation et réussite des étudiants; communication au comité d'évaluation et de contrôle des politiques publiques de l'Assemblée Nationale », 2020
- Daniel Castro and Michael McLaughlin, « Who is winning the AI Race : China, The EU or the United States ? 2021 Update », Center Data for Innovation, 2021
- Daniela Ovadia, « Vous êtes nul en statistiques? C'est normal! », Cerveau et Psycho, 2018
- Erwan Le Merrer, Gilles Trédan, « The bouncer problem: challenges to remote explainability », 2019
- Gérard Russeil, Vivien Bression, Florence Dietsch et al., « Les référentiels de la DSI : état de l'art, Usages et bonnes pratiques », Cigref, 2009
- Hubert Guillaud, « L'intelligence artificielle va-t-elle rester impénétrable ? » Internetactu.net https://www.internetactu.net/a-lire-ailleurs/lintelligence-artificielle-va-t-elle-rester-impenetrable/ 2016 consulté en octobre 2021
- ICO Information Commissioner's office, « Guidance on the AI auditing framework ; draft guidance for consultation », 2020
- ICO Information Commissioner's office, « Guidance on AI and data protection », 2020
- ICO Information Commissioner's office, « Explaining decisions made with AI », 2020
- Institut Montaigne, « Algorithmes : contrôle des biais S.V.P » ; 2020
- Isabelle Falque-Pierrotin, Gérard Berry, Jean-Richard Cytermann et al., « Comité éthique et scientifique de Parcoursup ; Rapport au parlement », 2020
- Jean-Louis Queguiner, « Deep learning explained to my 8-years-old daughter », https://blog.ovhcloud.com/deep-learning-explained-to-my-8-year-old-daughter/, 2019, consulté en novembre 2021
- J. Fjeld, H. Hiligoss, N. Achten, M.L. Daniel, J. Feldman, S. Kagay, A. Singh, « Principled artificial Intelligence : a map of ethical and Rights-Based Approaches », 15/01/2020
- Laurence Neuer, « Le juriste augmenté ou l'addition de « plusieurs formes d'intelligence » », LePoint.fr, 2020
- Laurent Sorber, « Si nous voulons promouvoir l'équité dans l'intelligence artificielle, nous devrons faire des choix difficiles », Trends, consulté en octobre 2021
- LNE, « Referentiel de certification ; processus de conception, de développement, d'évaluation et de maintien en conditions opérationnelles des intelligences artificielles », 2021

- Marine Corniou, « Comment mesure-t-on l'efficacité d'un vaccin ? », Québec Sciences, https://www.quebecscience.qc.ca/sante/mesure-efficacite-vaccin/ 2021 consulté en 2021
- Miles Brundage, Shahar Avin, Jasmine Wang et al., « Toward trustworthy AI Development : Mechanisms for supporting verifiable claims », OpenAI, 2020
- OCDE, « Recommandations du Conseil sur l'intelligence artificielle », OECD/LEGAL/0449
- Olivier Ezratty, « Les usages de l'intelligence artificielle 2021 », https://www.oezratty.net/wordpress/2021/usages-intelligence-artificielle-2021/ consulté le 15/02/2021
- Parlement Européen, « P9_TA-PROV (2020) 0275 Cadre pour les aspects éthiques de l'intelligence artificielle, de la robotique et des technologies connexes ; Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission concernant un cadre pour les aspects éthiques de l'intelligence artificielle, de la robotique et des technologies connexes (2020/2012 (INL)) » 2020
- Parlement Européen « P9_TA-PROV (2020) 0276 Un régime de responsabilité civile pour l'intelligence artificielle ; Résolution du Parlement européen du 20 octobre 2020 contenant des recommandations à la Commission sur un régime de responsabilité civile pour l'intelligence artificielle (2020/2014 (INL)) », 2020
- Parlement Européen, « P9_TA (2020) 0277 Les droits de propriété intellectuelle pour le développement des technologies liées à l'intelligence artificielle ; Résolution du Parlement européen du 20 octobre 2020 sur les droits de propriété intellectuelle pour le développement des technologies liées à l'intelligence artificielle (2020/2015 (INI)) », 2020
- Patrice Bertail, David Bounie, Stephan Clémançon et al., « Algorithmes : biais, discrimination et équité », Télécom ParisTech, 2019
- Patrick Krauf, Andreas Maler, « Y a-t-il un esprit dans la machine ? », Cerveau et Psycho, 2022
- Reuben Binns, « On the apparent conflict between individual and group fairness », In conference on fairness, accountability and transparency, 2020
- Ronan Hamon, Henrik Junklewitz, Ignacio Sanchez, « Robustness and Explainability of Artifical Intelligence ; From Technical to policy solutions » European Commision : JRC Technical Report, 2020
- Runshan Fu, Manmohan Aseri, Param Vir Singh et al., « "Un" Fair machine learning Algorithms », Carnegie Mellon University, Management science, 2021
- Salah Chadli, Philippe Neveux, Thibaud Real, « Intelligence artificielle et éthique : comment définir et mesurer l'équité algorithmique ? », Quantmetry https://www.quantmetry.com/blog/intelligence-artificielle-et-ethique-comment-definir-et-mesurer-lequite-algorithmique/ 2021, consulté en octobre 2021
- Tim Worstall, « How AI can remove bias from decision-making » Adam Smith Institute, 2020

- Tom Murphy, « The first level of Super Mario Bros. is easy with lexicographic orderings and time travel... after that it gets a little tricky », 2013
- Virigina Dignum « Ethics in artificial intelligence: introduction to the special issue », Ethics and Information Technology, Springer, 2018
- Will Fleisher, « What's fair about individual Fairness ? », Northeastern University, 2021
- Xavier Vamparys, « Ethique de l'intelligence artificielle : expliquer "l'explicabilité" », Revue-banque.fr http://www.revue-banque.fr/management-fonctions-supports/article/ethique-intelligence-artificielle-expliquer-explic, 2020, consulté en Octobre 2021

4. Table des matières détaillée

L'Académie est une initiative de l'Ordre des experts-comptables

Il est évident que tôt ou tard, tout le monde est ou sera impacté par des applications informatiques recourant à l'intelligence artificielle. Mais peut-on avoir réellement confiance dans ces nouveaux systèmes ? En effet, on a déjà constaté différentes dérives. Elles sont dues à de nombreuses causes : conception insuffisante, réalisation bâclée, paramétrage ou machine learning imparfait, tests incomplets, exploitation hasardeuse, maintenance aléatoire... Dans ces conditions peut-on garantir que les applications basées sur l'intelligence artificielle sont réellement dignes de confiance ?

Notre réponse est simple. Les développeurs et les exploitants de ces applications doivent veiller à mettre en œuvre un certain nombre de bonnes pratiques. Les risques spécifiques aux systèmes d'Intelligence Artificielle sont analysés dans cet ouvrage et les pistes de bonnes pratiques sont présentées pour que de nouvelles démarches soient inventées et diffusées.

Le meilleur moyen de s'assurer qu'elles sont effectivement appliquées est de faire régulièrement auditer ces systèmes par des professionnels compétents et indépendants. Il est pour cela nécessaire de définir des démarches adaptées.

Un groupe de travail commun entre l'Académie des Sciences et Techniques Comptables et Financières et l'ISACA-AFAI a travaillé depuis 2018 sur la manière d'auditer les systèmes à base d'intelligence artificielle. Un premier ouvrage est paru en 2021 : « Intelligence Artificielle et Confiance: réglementation, enjeux, risques, audit et certification ». Ce deuxième livre poursuit et approfondie la réflexion sur ce sujet fondamental.

www.lacademie.info

CONTACTS

Académie des Sciences et Techniques Comptables et Financières

200-216 rue Raymond Losserand 75014 Paris
Tél. +33 (0)1 44 15 64 24

www.lacademie.info

William NAHUM
Président fondateur

Constance CAMILLERI,
Directrice Prospective et Performance, diplômée d'expertise comptable

Marie-Amélie CALMAO
Chargée administrative