Thèse PRO 2023
Nathalie MOHEISSEN

TEASING

L' IA Responsable

et ses applications dans la banque

MBA MCI
LET'S GO DIGITAL

MBA Marketing Digital, Communication, Innovation

Code NSF 312-326, certification professionnelle de niveau I (Fr) et de niveau 7 (Eu) enregistrée au RNCP

Promotion 2022-2023

Thèse professionnelle de

Nathalie MOHEISSEN

L'IA responsable et ses applications dans la banque

Tuteur de thèse : David GUIRADO

Date de remise : 03 septembre 2023

DEVINCI EXECUTIVE EDUCATION

47, boulevard de Pesaro

92000 Nanterre

Remerciements

Je suis très reconnaissante des apports et des échanges que j'ai pu avoir avec les 14 personnes que j'ai pu interviewer :

Patrick BEZOMBES
Président du comité de normalisation de l'IA – **AFNOR**, vice-président du comité de normalisation JTC 21 au niveau européen (Normalisation IA)

Benoit ROTTEMBOURG
Responsable **REGALIA** chez **INRIA**
Un scientifique de formation, Mathématicien

Mohamed BEJI
NLP Engineer &
Data Scientist chez CACIB

Ludovic GIBERT
Chief Data Officer & Innovation Leader
Vice-Président de l'**IMA** (Innovation Makers Alliance)

David CORTES
Président,founder **AIvidence**
Co-Pilote de la French Tech Corporate Community

Matthieu CAPRON
Chef de projet IA
Responsable Design authority IA
DataLab Groupe Crédit Agricole

Grégoire COLOMBET
PDG d'une start-up AI decision
Coordination au sein d'**IBM** &
Auteur du livre « Humain ou IA ? Qui décidera le futur ? »

Alexandra COURTAUX
Infolinguiste &
Chef de projet IA Factory CACIB

Ghassen BETTAIEB
Data Scientist
DataLab Groupe Crédit Agricole

Christophe MOULIS
Account Manager for IA &
Innovation Architect at CACIB

Michel BONVIN
expert en intelligence artificielle chez CACIB

Céline LUNET VAUTRIN
Manager of the Compliance IA team CACIB

Gilbert HANNOUCHE
Leader innovation chez CACIB

Pierre CASTELLANI
AI and ML model validation at CACIB
Head of Market Risk IT, Solution manager

Je les remercie vivement pour le temps qu'ils ont bien voulu me consacrer.

Mes remerciements vont également à mon directeur de thèse David GUIRADO, pour ses nombreux conseils.

Je remercie également Gilbert HANNOUCHE pour son aide au choix du sujet et Yoanne DRAUCOURT pour ses conseils.

Je remercie sincèrement mon époux Salem MOHEISSEN mon conseiller et relecteur pour son soutien ainsi que nos enfants qui se souviendront de ces longues heures d'isolement de leur maman pour réaliser cette thèse et tout l'investissement que cela a impliqué pour la réalisation de ce travail.

Je remercie également l'Institut Léonard de Vinci, et plus particulièrement nos directeurs pédagogiques du MBA MCI, Alexandre STOPNICKI et Louis DUROULLE à #MBAMCI (élu meilleur Master Marketing Digital et e-business par Eduniversal pour la 12ème année consécutive).

Enfin, merci à tous mes collègues du MBAMCI pour cette belle aventure, leur soutien, leur bonne humeur, et leurs conseils.

Un dernier merci à ChatGPT pour son aide à la rédaction pour les peu de fois que je l'ai sollicité.

Ce travail fut pour moi l'opportunité d'une véritable exploration, riche de découvertes et d'enseignements, sur un sujet très vaste qui est l'Intelligence Artificielle ainsi que les domaines associés tel que les algorithmes, la data, l'application de l'intelligence artificielle dans les banques, l'innovation responsable.

Résumé

L'utilisation de l'Intelligence Artificielle se généralise de plus en plus de nos jours et se propage à tous les aspects de nos vies, qu'ils soient personnels ou professionnels.

Je vous propose dans un premier temps de comprendre cette technologie appelée « Intelligence Artificielle », souvent discutée mais rarement comprise étant donné sa complexité, via des explications simples, nécessaires à la bonne compréhension de ce sujet très vaste.

Si l'IA a été conçue pour aider l'humain à prendre des décisions en déléguant cette responsabilité à une machine, ses avancées sont fulgurantes ces dernières années et avec elles, apparaissent des limites et dérapages. En effet, l'Intelligence Artificielle n'est pas infaillible et différents risques existent quant à son utilisation.

Ces derniers soulèvent de nombreux débats et de nombreuses voix convergent vers un appel à faire une pause en attendant la mise en place d'un cadre réglementaire pour une IA responsable. Celle-ci vise à inspirer confiance, être perçue comme juste, non biaisée, explicable et à augmenter les capacités humaines sans s'y substituer.

Cette thèse vous permettra de découvrir plus en détail ce qu'est l'IA responsable ainsi que les différentes initiatives en cours pour définir sa mise en œuvre, notamment à travers un focus sur la futur réglementation européenne appelée « AI Act ».

Mais si cette dernière donne des principes, il ne s'agit pas d'une démarche. Je vous propose donc d'aborder ensemble une démarche pour aller vers une IA responsable, à travers principalement deux cas de figure :

- Vous avez déjà un système d'IA :
  - Une grille d'auto-évaluation selon les sept principes de l'AI Act vous permettra de savoir où vous en êtes et les efforts à faire pour atteindre votre objectif.
  - La thèse vous proposera également des systèmes d'évaluations dit de « tierce partie » (externe) avec les organismes recommandés.
- Vous envisagez ou vous avez démarré récemment la mise en place d'un système d'IA :
  - Une démarche vous est proposée pour une application par phase pour la construction d'un projet d'IA responsable, de la conception jusqu'à sa mise en service et son suivi.

- **Mots clés :**

Lien du Teasing de la thèse : https://youtu.be/IgllwYlJp28?feature=shared

Le QR CODE est sur la page de garde.

Bonne lecture !

Summary

The use of Artificial Intelligence is becoming more and more widespread nowadays and is spreading to all aspects of our lives, whether personal or professional.

First, I suggest that you understand this technology called "Artificial Intelligence", often discussed but rarely understood given its complexity, via simple explanations, necessary for a good understanding of this very vast subject.

If AI was designed to help humans make decisions by delegating this responsibility to a machine, its advances have been dazzling in recent years and with them, limits and slippages appear. Indeed, Artificial Intelligence is not infallible and there are various risks associated with its use.

The latter raise many debates and many voices converge on a call to take a break while waiting for the establishment of a regulatory framework for responsible AI. This aims to inspire confidence, to be perceived as fair, unbiased, explainable and to increase human capacities without replacing them.

This thesis will allow you to discover in more detail what responsible AI is as well as the various initiatives underway to define its implementation, in particular through a focus on the future European regulation called "AI Act".

But if the latter gives principles, it is not a question of an approach. I therefore suggest that you approach together an approach to move towards a responsible AI, mainly through two scenarios:

- Already have an AI system ?
  - A self-assessment grid according to the seven principles of the AI. Act will allow you to know where you are and the efforts to be made to achieve your goal.
  - The thesis will also offer you so-called "third party" (external) evaluation systems with the recommended organizations.
- Are you considering or have you recently started setting up an AI system ?
  - An approach is proposed to you for a phased application for the construction of a responsible AI project, from design to its commissioning and monitoring.

- **Keywords :**


Ethics
Transparency
Skid
Big Data
Artificial intelligence
Environment
Security
Bias
Trust
Data
Limit
Risk
Regulation
Responsible AI
Bank

Table des matières

Introduction

L'Intelligence Artificielle est devenue une force omniprésente qui façonne notre quotidien. Au cœur des révolutions technologiques actuelles, elle offre à la fois aux machines et aux humains des capacités autrefois inimaginables. Cette quatrième révolution industrielle, principalement axée sur le numérique, est en réalité une révolution humaine. Elle promet de transformer profondément notre façon de vivre, de travailler et d'interagir.

Pour appréhender l'IA, il est essentiel de comprendre d'abord ce qu'est l'intelligence humaine. L'intelligence, qu'elle soit naturelle ou artificielle, est souvent associée à la capacité d'adaptation. Plusieurs définitions ont été proposées au fil des ans, mais l'adaptabilité revient constamment comme un trait central de l'intelligence.

En effet, l'humain a toujours été reconnu pour sa remarquable capacité d'adaptation, qui repose sur une série de choix et de décisions. Si nous considérons l'intelligence comme la capacité d'apprendre, de déduire des informations, de reconnaître des motifs répétitifs, de résoudre des problèmes et de prendre des décisions, alors l'humain est véritablement une créature intelligente. Cependant, la prise de décision est un processus complexe, souvent entravé par l'incertitude du futur.

L'histoire montre que l'humanité a toujours cherché des moyens de prédire l'avenir pour faciliter sa prise de décision. Des méthodes mystiques, comme consulter des devins ou lire les astres, ont été utilisées. Cependant, avec l'avènement des probabilités modernes au 17ème siècle, une approche plus scientifique de la prédiction est née.

Dans ce contexte, l'IA a été conçue pour aider l'homme à prendre des décisions éclairées en déléguant certaines responsabilités à une machine. L'IA, en tant qu'outil, peut traiter d'énormes quantités de données, fournir des informations et, finalement, aider à acquérir le savoir nécessaire pour prendre une décision.

Cependant, chaque révolution apporte son lot de craintes, en particulier lorsqu'elle est mal comprise. L'IA, malgré ses promesses, n'est pas exempte de ces inquiétudes. Ces craintes ont été alimentées au fil des années notamment par les films et romans de science-fiction mais des craintes bien plus concrètes voient récemment le jour, au fur et à mesure que l'utilisation de l'IA se généralise. Quand certains voient en elle une menace pour nos emplois, d'autres voient également des dangers liés à la sécurité des nations de la démocratie.

En effet, l'IA est partout : une réalité omniprésente dans nos vies mais aussi dans nos emplois. De la santé à l'automobile, de l'éducation à l'agriculture, elle a infiltré presque tous les secteurs. Cependant, parmi ces domaines, le secteur bancaire se distingue par son adoption précoce et sa dépendance croissante à l'égard de l'IA.

La banque, en tant qu'institution, a toujours été un pilier de nos sociétés. Depuis l'Antiquité, elle a joué un rôle central dans l'économie, facilitant les transactions, accordant des crédits, investissant dans des entreprises et offrant une multitude de services financiers. Ces services, essentiels à la fois pour les individus et les entreprises, ont façonné le développement économique, social et culturel de nos sociétés. Dans ce contexte, toute décision prise par une institution bancaire peut entraîner des répercussions profondes. Les implications ne sont pas seulement financières, mais aussi sociales, culturelles et, dans certains cas, politiques.

Avec l'avènement de l'IA, les enjeux associés à ces décisions sont montés d'un cran. Les systèmes d'IA, avec leur capacité à traiter d'énormes quantités de données à une vitesse vertigineuse, ont le potentiel d'améliorer considérablement l'efficacité et la précision des services bancaires. Ils peuvent automatiser des tâches autrefois laborieuses, offrir des analyses plus approfondies et personnaliser les services pour répondre aux besoins spécifiques des clients. Cependant, ces avantages ne viennent pas sans défis.

Bien qu'elle soit une technologie puissante, l'IA n'est donc pas sans limites. Ces limites, lorsqu'elles ne sont pas correctement adressées ou comprises, peuvent rapidement se transformer en dérapages majeurs. Plusieurs incidents notables dans le passé ont mis en lumière ces dérapages et notamment dans le milieu bancaire où des systèmes d'IA ont refusé des prêts à des individus sur la base de critères discriminatoires, ou des algorithmes ont mal évalué les risques, entraînant des pertes financières massives.

Les algorithmes utilisés par les systèmes d'IA, aussi avancés soient-ils, ne sont donc pas infaillibles. Ils peuvent être biaisés, opaques et, dans certains cas, prendre des décisions qui peuvent sembler arbitraires ou injustes. De plus, étant donné la nature complexe de ces algorithmes, même les experts peuvent avoir du mal à comprendre pleinement comment ils fonctionnent, rendant la transparence et la responsabilité juridique encore plus difficiles.

C'est ici que la notion d'IA responsable entre en jeu. Au cœur de cette idée se trouve un engagement envers l'éthique. Il s'agit de garantir que les systèmes d'IA sont développés et déployés de manière à respecter les principes fondamentaux de justice, de transparence, de respect de la vie privée et d'équité. Cela signifie reconnaître et aborder les biais potentiels, garantir la transparence des processus décisionnels et assurer que les utilisateurs, qu'il s'agisse d'individus ou d'entreprises, peuvent faire confiance à ces systèmes et comprendre comment ils fonctionnent.

L'IA responsable n'est pas seulement une question de programmation ou de technologie. Elle nécessite une collaboration étroite entre les experts en éthique, les régulateurs, les développeurs d'IA, les parties prenantes du secteur bancaire et, bien sûr, les clients eux-mêmes. Elle exige une compréhension profonde des implications éthiques de l'IA et une volonté de mettre en place des mécanismes pour surveiller, évaluer et, si nécessaire, corriger les décisions prises par ces systèmes.

L'IA responsable est aussi une réponse à des préoccupations majeures quant à l'utilisation et à la gouvernance de ces technologies et elle met en exergue la nécessité d'un cadre réglementaire solide qui permettra d'instaurer un écosystème où l'IA est à la fois responsable, éthique et digne de confiance. Cependant, la mise en place d'un tel cadre n'est pas une mince affaire. De nombreuses initiatives existent et des organisations internationales, des gouvernements et des groupes de réflexion travaillent ensemble pour élaborer des directives, des normes et des réglementations pour l'IA. Néanmoins, un flou subsiste sur les réponses données par ces cadres et la façon d'atteindre une IA responsable.

Ainsi, alors que nous nous aventurons plus profondément dans cette ère de l'IA, il est impératif de se poser les bonnes questions pour tenter d'y voir plus clair. Comment pouvons-nous garantir que l'IA est utilisée de manière éthique et responsable, en particulier dans le secteur bancaire ? Comment pouvons-nous nous assurer que les décisions prises par ces systèmes sont justes et équitables ? Et surtout, comment pouvons-nous maintenir la confiance des clients dans un monde où les machines jouent un rôle de plus en plus important dans des décisions aussi cruciales que l'octroi d'un prêt ou la gestion des investissements ?

Cette thèse vise donc à explorer ces questions, à mettre en lumière les défis associés à l'utilisation de l'IA, son application dans le secteur bancaire et à proposer des solutions pour une utilisation responsable de cette technologie révolutionnaire. En plongeant dans les pages suivantes, nous espérons offrir une perspective équilibrée sur les promesses et les pièges de l'IA, et souligner l'importance d'une approche éthique dans son déploiement et son utilisation, en particulier dans le milieu bancaire.

1 L'Intelligence Artificielle

L'Intelligence Artificielle est aujourd'hui omniprésente dans notre quotidien. En effet, elle est au centre des révolutions technologiques qui marquent notre époque. Elle donne aux machines et aux humains des capacités qui étaient inimaginables il y a encore quelques années. L'Intelligence Artificielle n'a pas fini d'évoluer et permettra d'arriver à des innovations plus surprenantes les unes après les autres.

En effet, face à une quatrième révolution industrielle qui n'a d'industrielle que le nom tant elle concerne le numérique, l'Intelligence Artificielle (ou IA) va bouleverser nos vies, tout changer! Mais surtout, tous nous changer ! En cela, elle pourrait être requalifier de révolution humaine.

Mais l'humain a une particularité : toute révolution qui bouleverse notre quotidien et qui est mal-comprise lui amène de la crainte, voir le braque jusqu'à en être réfractaire. Il est donc plus qu'important de comprendre ce qu'est l'Intelligence Artificielle ou pour les « intimes », l'IA.

Et si la question se pose certainement très souvent, il faut à mon sens bien comprendre de ce qu'est l'intelligence humaine avant de définir l'Intelligence Artificielle.

1.1 Qu'est-ce que l'Intelligence ?

1.1.1 Vers une définition

L'intelligence non-artificielle, celle humaine, est le plus souvent appelé « intelligence naturelle ». Et s'il est difficile de définir avec exactitude ce qu'est l'Intelligence Artificielle c'est parce qu'il est aussi très difficile de définir ce qu'est l'intelligence naturelle.

Voici quelques définitions :

> *L'intelligence a été décrite comme une faculté d'adaptation (apprentissage pour s'adapter à l'environnement) ou au contraire, faculté de modifier l'environnement pour l'adapter à ses propres besoins.*

Wikipédia – 2023

> *L'intelligence consiste à savoir donner à des problèmes de tous ordres, une réponse adaptée. Les solutions intelligentes se distinguent radicalement des solutions obtenues par tâtonnement aveugle, par instinct ou par habitude.*

Dictionnaire Hachette – 1977

> *L'intelligence : faculté de connaître, de comprendre ; qualité de l'esprit qui comprend et s'adapte facilement.*

Dictionnaire leRobert- 2023

> *L'intelligence : aptitude d'un être humain à s'adapter à une situation, à choisir des moyens d'action en fonction des circonstances.*

Dictionnaire Larousse - 2022

> *Les espèces qui survivent ne sont pas les espèces les plus fortes, […], mais celles qui s'adaptent le mieux aux changements*

*Attribué à **Darwin***

Vous l'aurez compris : ce qui ressort ou revient le plus souvent est **l'adaptabilité**, qui finalement pourrait se caractériser comme une forme de force.

Il est souvent dit, particulièrement entre humains, que l'intelligence est le propre de l'Humanité. Cette particularité a été très largement démontré comme fausse notamment vis-à-vis des animaux mais nous devrions retenir que le propre de l'Humanité est sa capacité à s'adapter et évoluer : penser, stocker son savoir en mémoire et utiliser ces données pour construire de nouvelles choses, inventer des outils nécessaires à sa subsistance qui lui permettrait ainsi de réduire sa fatigue, pour anticiper le futur en imaginant demain à partir de ce qu'il connaît d'hier.

L'Humain a donc une capacité d'adaptation incroyable : l'adaptation ne peut se faire que par une série de choix et donc de décision. Ainsi, s'il représente l'espèce la plus adaptable, c'est que l'Humain est capable de prendre les meilleures décisions.

Et si nous devions retenir une définition : l'intelligence serait la capacité intellectuelle (cognitive) d'apprendre, de déduire des informations, de reconnaître des motifs répétitifs, de **résoudre des problème et de prendre des décisions**.

1.1.1.1 Les différentes formes de l'Intelligence naturelle

Howard GARDNER, psychologue américain, indique dans son livre « *Les intelligences multiples* » (édition de 1983) qu'il existe plusieurs formes d'intelligences. Il décrit alors huit formes possibles d'intelligence qu'il complète d'une huitième en 1993 dans la réédition de son livre.

Voici les huit formes d'intelligence qu'il recense :

- **L'intelligence linguistique** : Aptitude à manier et à saisir les mots ainsi que leurs subtilités de signification. A donc une affinité pour les mots, écrits et parlés. Elle donne du sens aux mots.
- **L'intelligence logico-mathématique** : habileté logique et mathématique, capacité à manipuler les nombres et de résoudre des problèmes logiques (et mathématique). Elle donne du sens aux nombres.
- **L'intelligence spatiale** : capacité de créer le modèle mental d'un espace, d'y évoluer et d'y opérer, de trouver son chemin dans un environnement donné et d'établir des relations entre les objets dans l'espace. Elle nous permet de savoir où nous sommes, le sens de l'orientation.
- **L'intelligence musicale** : aptitude à percevoir et créer des rythmes et mélodie, reconnaitre des sons et des harmonies. Elle nous permet de jouer d'un instrument, de chantier, de composer et d'apprécier la musique écoutée.
- **L'intelligence corpo-cinétique ou corporelle kinesthésique** : habilité à utiliser son corps dans l'espace, capacité d'utiliser le contrôle fin des mouvements du corps. Elle permet d'avoir du sens de notre corps dans l'espace, l'intelligence des sportifs. Elle est celle qui se développe à force d'intense pratique et d'expertise.
- **L'intelligence interpersonnelle** : capacité de comprendre les autres, de communiquer avec eux et d'anticiper l'apparition d'un comportement, de ressentir, de se projeter sur les réaction et d'avoir de l'empathie. Elle permet de reconnaitre les motifs comportementaux entre êtres humains.
- **L'intelligence intrapersonnelle** : capacité à créer un modèle de compréhension de soi-même de manière précis et véridique, et de mettre à profit cette compréhension dans la vie quotidienne. Elle est plus complexe que les autres car permettant de se connaitre soi-même, savoir lire ses propres comportements et d'en déduire les motifs qui guide nos choix et nos décisions. C'est celle qui guide nos **prises de décision**.
- **L'intelligence naturaliste** : faculté de percevoir et de comprendre l'environnement vivant qui nous entoure. Elle se manifeste par la capacité d'apprécier, de distinguer et de catégoriser la faune, la flore et les minéraux. Par extension, cette compétence s'applique également à la compréhension et à l'interprétation de l'univers culturel.

1.1.1.2 La prise de décision chez l'Humain

Si nous revenons à la notion d'adaptation, celle-ci est fondée sur des choix. L'être humain, que ce soit dans sa vie de tous les jours pour des éléments "banals" ou bien pour des choix cruciaux, se doit de prendre des décisions. Cependant, prendre une décision n'est pas toujours aisé : cela est même toujours un processus complexe et difficile.

La raison de cette difficulté pour l'Humain de prendre une décision est unique : **l'incertitude du futur**. L'être humain n'a aucune idée de ce qui pourrait se passer dans le futur. Cela a des conséquences sur sa prise de décision :

- L'Humain ne peut qu'ignorer la ou les conséquence d'une décision
- L'Humain ne peut que décider que dans l'incertitude
- L'Humain ne peut jamais savoir à l'avance s'il prend une bonne décision

Mais alors, qu'est-ce qu'une bonne décision ? Le caractère bon ou mauvais d'une décision est souvent jugé à posteriori, par son résultat.

En réalité, une bonne décision est une décision mesurée : néanmoins, les probabilités d'occurrence d'une conséquence (ou plusieurs) n'étant pas forcément connu lors de la prise de décision, elle peut donc avoir parfois des conséquences négatives. Mais cela ne remet pas en cause le fait que la décision était initialement une bonne décision. Donc, nous pouvons en déduire que **prendre la meilleure décision sera celle qui sera mesurée en connaissance probable des conséquences**.

Cela nous amène donc à la question suivante : **l'être Humain est-il naturellement équipé pour raisonner de manière probabiliste** ?

Daniel KAHNEMAN, psychologue et économiste américain ayant obtenu le prix Nobel d'économie en 2022, a tenté de comprendre durant sa carrière les mécanismes de prise de décision humaine. A la question précédente, **la réponse qu'il obtient est un NON formel**. **L'être Humain n'est clairement pas naturellement équipé pour raisonner de manière probabiliste**.

C'est dans son principal ouvrage « *Thinking, Fast & Slow* » résumant ses travaux qu'une explication peut être trouvée (celle qui lui aura permis d'obtenir le prix Nobel) :

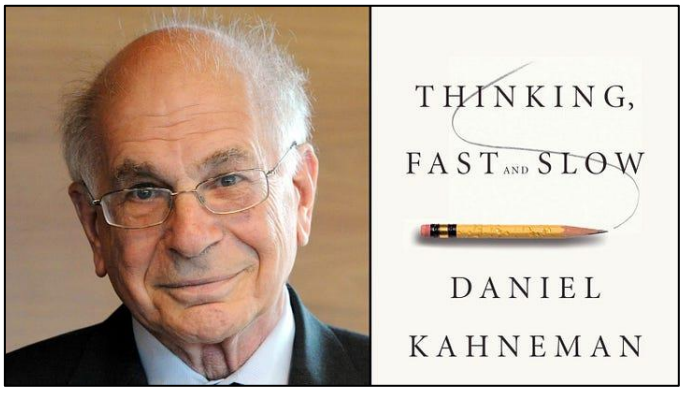


Figure 1 : Daniel KAHNEMAN et son ouvrage "Thinking, Fast & Slow"

En effet, dans ce livre, il nous explique que les choix et décisions de l'être humain repose sur deux systèmes auxquels il a préféré ne pas attribuer de noms :

- **Système 1** : celui « Fast », qui pense vite. Il est le système « instinctif », proche du « réflexe » qui est en charge de la survie et de la préservation de notre espèce. C'est le système qui **prend des décisions rapides**, souvent en cas de danger.
- **Système 2** : celui « Slow », qui pense lentement. C'est celui qui va analyser, calculer lesquelles sont des étapes qui prennent plus de temps. Celui-ci a des faiblesses car pouvant être soumis à des **biais**, appelés « **heuristiques de décisions** » par Daniel KAHNEMAN

Il développe alors l'idée que **les individus peuvent être soumis à des biais**, des raccourcis mentaux que nous utilisons tous les jours, **et donc commettre des erreurs**. Par conséquent, **les individus ne prendraient pas toujours leurs décisions de manière rationnelle**.

Sur la base de ses travaux, de nombreux chercheurs ont trouvé des heuristiques, considérées comme des « règles approximatives » ou « raccourcis de la pensée », dont le nombre serait de plusieurs dizaines.

Daniel KAHNEMAN en a identifié initialement quatre :

- **La représentativité** : pour identifier le lien qui unit deux éléments
- **La disponibilité** : pour évaluer la fréquence d'un évènement, ou la probabilité qu'il se produise
- **L'ancrage** : pour faire des estimations sans avoir de connaissances certaines à propos de l'information à estimer, en comparant à une action ou un événement similaire
- **L'ajustement** : souvent liée à l'ancrage, l'estimation est ajustée selon la référence (point d'ancrage)

L'heuristique de disponibilité fonctionnerait comme suit : au moment de prendre une décision, l'Humain utilise les informations immédiatement disponibles, qu'il a de plus fraiches. Notre mémoire fonctionnant comme une « stack » ou une « pile » (*LIFO : Last In First Out*), **la dernière information entrée est la première sortie dans l'urgence d'une prise de décision rapide**.

Ainsi, par exemple, c'est une des raisons pour lesquelles nous sommes submergés de messages marketing à longueur de journée : l'idée est pour les marketeurs de s'assurer qu'au moment de prendre une décision d'achat, le message concernant leur produit sera disponible car dernier arrivé.

Parmi ces quatre heuristiques, l'une des plus utilisées pour les choix est **l'heuristique d'ancrage** car permettant d'estimer la grandeur (fréquence, probabilité…) d'un nouvel évènement en se basant sur les informations connues (pertinentes ou non) par un individu.

Effectivement, l'effet d'ancrage se manifeste dans de nombreux phénomènes, y compris dans la prévision de la probabilité d'un événement futur, ce qui a une influence sur la prise de décision. Néanmoins, si cet ancrage a été introduit par une partie tierce, l'estimation est biaisée car influencée.

Pour résumer, **l'Humain n'est donc capable de décider que dans l'incertitude du futur** : source de craintes, d'anxiété, d'angoisse. **Il a manifesté le désir de transférer ou d'externaliser cette charge de prise de décisions relevant du « système 2 » vers des systèmes informatiques ou des machines**.

C'est dans ce but que l'Intelligence Artificielle, appelée aussi « IA », a été imaginée : aider l'être humain à prendre la meilleure décision en déléguant la responsabilité de la décision vers une machine.

Figure 2 : L'humain et la machine, main dans la main - source : theregister.com

1.1.1.3 Comment un être humain peut être aidé pour prendre la meilleure décision possible ?

L'être humain doit posséder le savoir s'il veut prendre les meilleures décisions possibles. Mais qu'est-ce que le savoir ?

Si la définition du savoir pouvait se résumer à l'ensemble des connaissances d'un individu acquises par l'étude, par l'observation, par l'apprentissage ou même par l'expérience, **le savoir serait alors tourné « vers le passé » et l'acquis mais ne permettrait de se faire une idée du futur possible**.

Personne ne peut connaitre le futur et particulièrement les conséquences réelles d'une prise de décision. Néanmoins, ce qui est recherché est l'idée de ce qui pourrait se passer : à défaut de pouvoir avoir une certitude car impossible, c'est avoir une **probabilité**.

Cette interrogation au sujet du futur a toujours existé : pour tenter de palier à cette incertitude du futur qui a pour conséquence de rendre la prise de décision si difficile, l'Humain a eu recours dans l'histoire a de nombreuses techniques « mystiques » tels que :

- Interrogation de « devins »
- Lecture d'entrailles d'animaux sacrifiés
- Lecture d'une boule de cristal, de cartes, de marcs à café, etc.
- Lecture des astres

Figure 3 : Techniques "mystiques" pour la prédiction du futur – Source : autrice

Si depuis des temps immémoriaux, l'être humain a tenté de connaitre son futur, ce qui allait se dérouler ou tout simplement d'avoir une indication, c'est au 17ème siècle, plus exactement en 1654 qu'il a tenté une autre approche.

En effet, en France cette année-là, **Antoine GOMBAUD**, Chevalier de Méré, noble, écrivain français et fervent amateur de jeux de hasard demande à un de ses amis, **Blaise PASCAL**, célèbre mathématicien, de l'aider à résoudre le problème de la répartition des gains lors d'une partie interrompue avant sa conclusion.

Accompagné de **Pierre DE FERMAT**, un célèbre mathématicien tout comme lui, ils débutent ensemble une étude sur la quantification du hasard. Ils finissent par inventer les probabilités modernes, tel que nous les connaissons aujourd'hui.

Figure 4 : Antoine GOMBAUD, Blaise PASCAL et Pierre DE FERMAT, inventeurs des probabilités modernes

Il s'agit alors d'une méthode de quantification de l'occurrence d'un événement dans le futur à l'aide d'outils mathématiques, initialement simples mais devenus complexes au fil du temps.

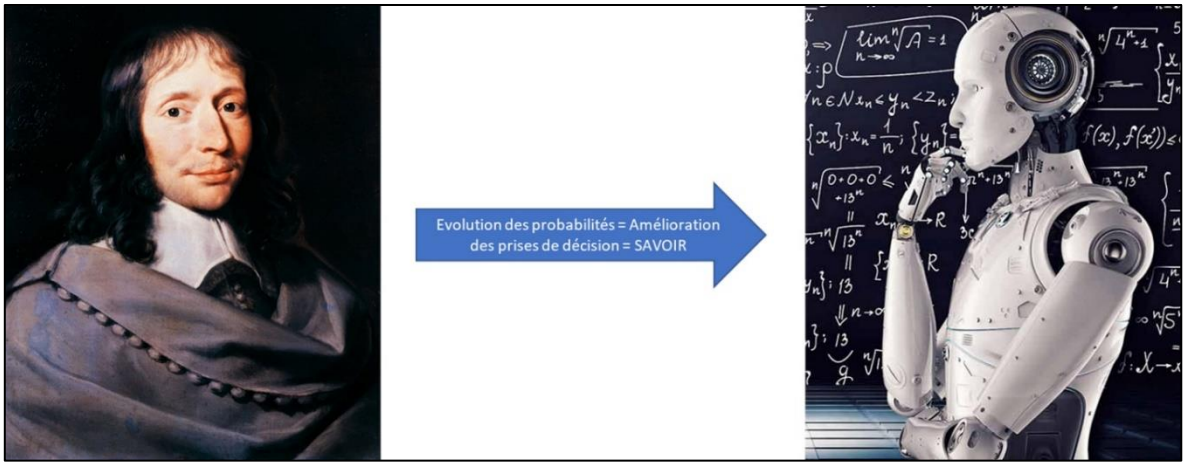


Figure 5 : Evolution des probabilités = savoir – Source : autrice

A travers les âges, la prise de décision a été rendue plus sereine, plus simple, via l'évolution de la probabilité d'occurrence d'un évènement futur : **le fait de pouvoir s'en faire une idée est en quelque sorte équivalent à posséder le savoir**.

1.1.2 Mais qu'est-ce que le savoir ?

Le savoir est une des étapes de l'évolution de l'information. Mais alors qu'est qu'une information ? c'est elle-même une évolution de la donnée. La donnée est un « élément qui sert de base à un raisonnement, de point de départ pour une recherche » selon le dictionnaire « **Le Robert** ».

Pour bien comprendre ce qu'est le savoir, il est préférable d'illustrer par un exemple : John doit prendre la décision d'achat d'un scooter en se basant sur le meilleur rapport qualité/prix…

Figure 6 : John veut acheter un scooter

John reçoit un fichier Excel très particulier de la part de son cousin un peu brouillon : dans le premier onglet, il ne voit que des montants et le premier qu'il voit est 1000… **Ceci est une donnée mais elle ne représente rien… John ne peut pas prendre de décision !**

John en parle à son cousin qui corrige le fichier et lui renvoie. John voit donc maintenant différents montants et sur la première ligne, il voit 1000 € pour un scooter d'une célèbre marque japonaise…

La **donnée est alors mise en contexte** : la donnée 1000 est alors exprimée avec une autre donnée qu'est la devise et est reliée à un modèle de scooter dont c'est le prix. **La donnée se transforme alors en une information**.

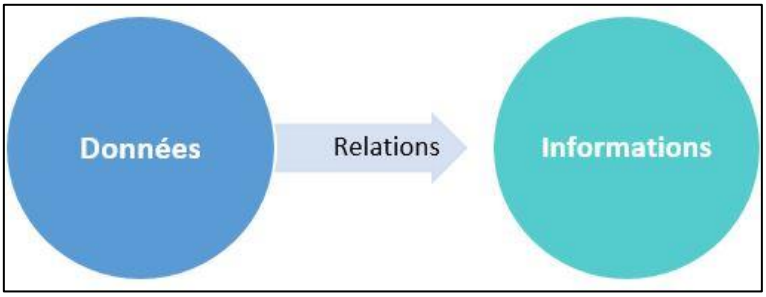


Figure 7 : Les données mises en relation sont des informations– Source : autrice

Hélas, John ne peut toujours pas prendre de décision : en effet, il veut acheter un scooter en se basant sur sa propre idée du meilleur rapport qualité/prix. Il lui manque alors la comparaison par rapport à une « population » d'autres scooters… **Il lui manque le savoir**.

John, en utilisant le fichier le plus récent de son cousin, a maintenant dans l'onglet le prix de plusieurs modèles ainsi que des avis, mais surtout **il sait maintenant que le fameux scooter japonais est dans le premier quartile dans la rangée des prix des scooter ayant les mêmes caractéristiques et les mêmes notes**. En gros, 75% des scooters équivalents sont plus chers… John est maintenant très content et il va pouvoir se décider en sachant qu'il fait probablement une bonne affaire.

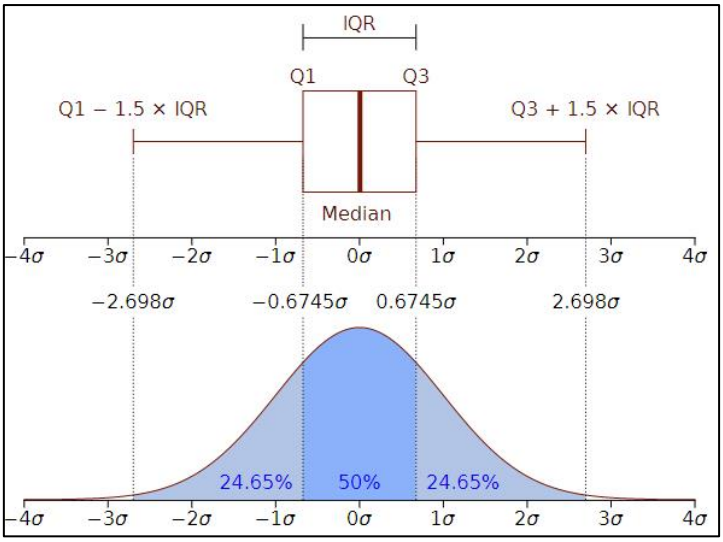


Figure 8 : Découpage par quartile - source : wikipedia.org

Les informations que John possède lui ont permis de réaliser son étude comparative et d'obtenir le graal : **le savoir ! Les informations se sont transformées en savoir !**

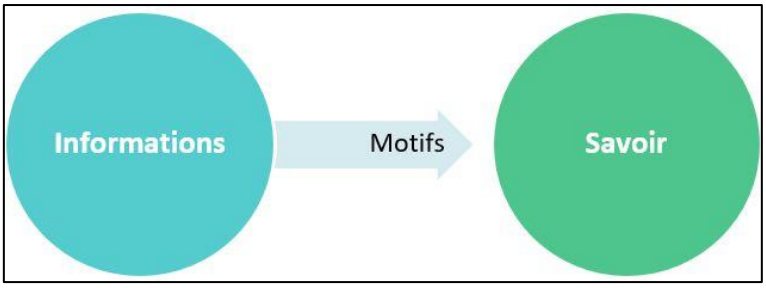


Figure 9 : Les informations mises en contexte donnent le savoir– Source : autrice

Et le savoir permet une prise de décision, ce que ne permettent pas les données « brutes » ou les informations.

Supposons maintenant que John aille plus loin et qu'il comprenne qu'il existe des évènements ou autres éléments qui influencent les prix des scooters, la bonne période pour acheter. Il a compris tout ce qui rentre en compte dans une prise de décision dans un contexte similaire. **Le savoir devient alors une compréhension de la décision !**

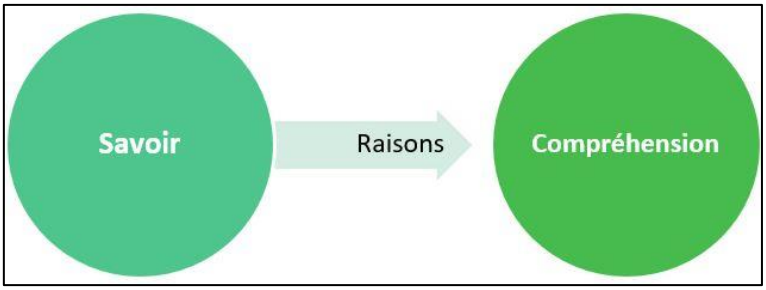


Figure 10 : Le savoir devient une compréhension de la décision– Source : autrice

Cette compréhension va permettre à John de reproduire son schéma décisionnel et même de faire des simulations si ensuite ce processus décisionnel compris est répété si souvent qu'il en devient automatique : **la compréhension s'est alors transformée en sagesse**.

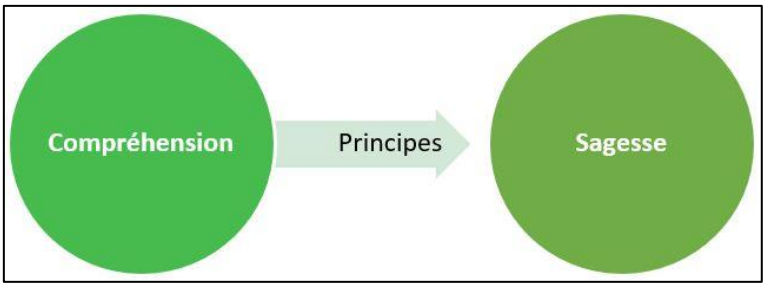


Figure 11 : La compréhension se transforme en sagesse si processus automatique de décision– Source : autrice

Ainsi, il peut être retenu qu'avec le savoir, issu d'une compréhension des informations basées sur des constats passés, la prise de décision est possible. Et s'il en sort une compréhension d'un modèle, **il en devient même possible d'exprimer un savoir sous forme de probabilité d'occurrence ce qui est nécessaire pour prendre la meilleure décision** !

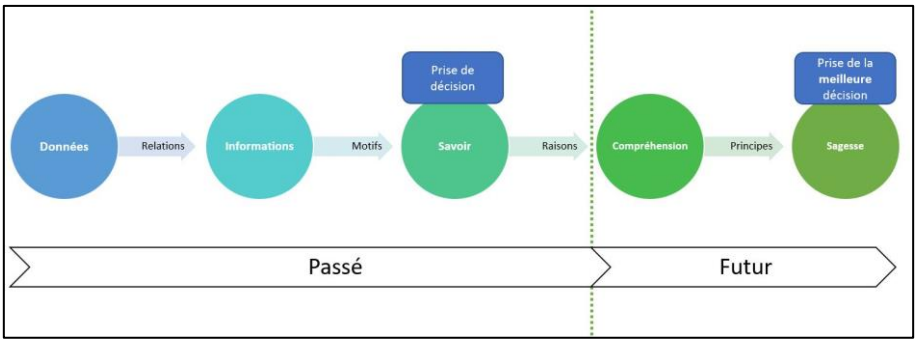


Figure 12 : processus de prise de la meilleure décision– Source : autrice

Pour résumer, nous avons commencé par définir ce qu'est l'intelligence naturelle : cette aptitude de l'être humain à s'adapter à une situation, à choisir des moyens d'action en fonction des circonstances pour évoluer.

Mais l'Humain n'est pas naturellement équipé pour prendre les meilleures décisions ! En effet, il possède une faible capacité à déterminer des probabilités avec certitude et reste donc dans l'incertitude du futur. **Cela provoque le besoin de se faire aider notamment par la technologie informatique.**

1.1.3 Quels outils utiliser aujourd'hui pour prendre des décisions ?

L'évolution de la donnée en savoir, appuyé par l'informatique, a donné naissance à plusieurs domaines ou outils de l'industrie de l'analyse des données.

Voici un schéma qui liste ces domaines et leurs utilités :

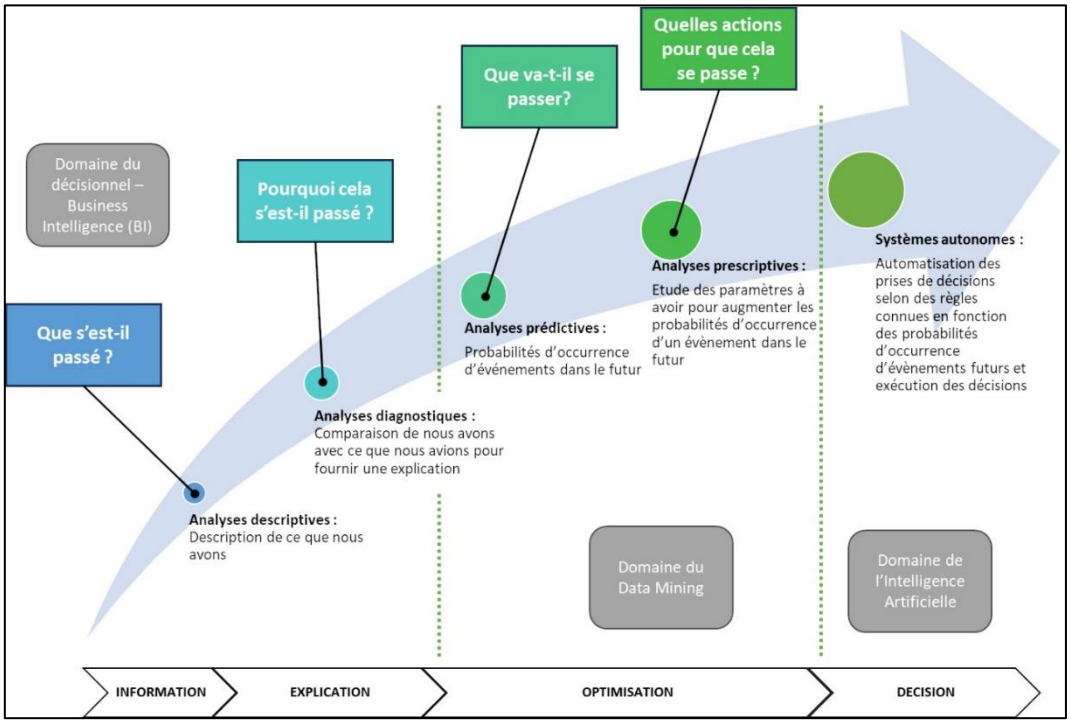


Figure 13 : Domaines de l'analyse des données : du décisionnel (BI) à l'Intelligence Artificielle – Source : autrice

Rentrons dans les détails de chacun de ces domaines :

- **L'analyse descriptive :**

Comme son nom l'indique, l'analyse descriptive est la description en termes statistiques de ce que nous avons.

Cela va comprendre des mesures de centrage, comme la moyenne ou la médiane où des mesures de dispersion comme l'écart-type ou la variance. Cela comprend aussi des pourcentages, des répartitions ou des distributions souvent représentées graphiquement.

Donc les statistiques descriptives sont la représentation de ce que nous avons aujourd'hui devant nous, une description du présent.

- **L’analyse diagnostique :**

L'étape suivante est celle de l’analyse diagnostique. Cela permet de comparer ce que nous avons aujourd'hui avec ce que nous avions hier. A partir de cela, nous essayons alors de trouver des explications.

Pourquoi avons-nous cela aujourd'hui ? y-a-t-il des éléments dans le passé qui aurait une relation avec ce que nous pouvons observer aujourd'hui ? Est-ce que nous pouvons l'expliquer dans une certaine mesure ?

C'est le domaine du **décisionnel** ou **business intelligence** (appelée aussi « BI »). La « BI » nous donne des informations sur ce que nous avions hier, ce que nous avons aujourd’hui et les possibles interactions entre les deux via des comparaisons.

- **L’analyse prédictive :**

L'étape suivante est celle des analyses prédictives. D’une complexité d’un niveau supérieur aux analyses vues précédemment, c’est une analyse qui a pour but de définir les probabilités d’occurrence d’évènements dans le futur. Sur la base de ces probabilités, il sera possible de prendre des décisions et potentiellement les meilleures parmi elles.

Avant cela, le décisionnaire (l’être humain) est contraint de trouver seul les possibles motifs et corrélations, donc d'évaluer une probabilité non quantifié avec un faible niveau de fiabilité et finalement de continuer de décider dans l'incertitude la plus absolue.

Cela s’appelle également la modélisation prédictive.

- **L’analyse prescriptive :**

Les analyses prescriptives constituent l’étape suivant et sont proches des analyses prédictives mais ont un but inversé.

En effet, si cette dernière consiste à déterminer la probabilité d’occurrence d’un événement futur pour un univers présent, l’analyse prescriptive consiste à déterminer ce qu’il faut changer dans mon univers présent pour augmenter la probabilité d’occurrence d’un événement bien déterminé dans le futur.

- **Le système autonome :**

Le système autonome est la dernière étape de toute cette évolution d’outils. Une fois les données acquises, leurs interactions comprises, leur comportement modélisé et grâce au savoir obtenu sous la forme de probabilité, il est alors possible de décider et d’exécuter. La prise de décision est souvent dépendante d'un seuil qui varie suivant les individus.

Voici un exemple de seuil par une illustration : un premier individu prendra un parapluie avec 50 % de probabilités de pluie lorsqu'un deuxième individu ne le fera qu'avec une probabilité de 70 ou 80 %.

Si des règles de gestion sont présentes et connues pour prendre des décisions selon les probabilités déterminées précédemment et des seuils, alors il est possible d'automatiser la prise de décision et donc de créer un système autonome : **l'Humain sera totalement retiré du processus mis en place, depuis l'acquisition des données jusqu'à l'exécution de la décision**.

Ce système autonome peut alors être considéré comme une Intelligence Artificielle**. Voyons plus précisément ce qu'est une Intelligence Artificielle.**

1.2 Qu'est-ce que l'Intelligence Artificielle ?

Nous avons pu voir ensemble dans la partie précédente, que l'origine de l'Intelligence Artificielle est la volonté de l'être Humain d'utiliser les probabilités comme moyens mathématiques d'accès au savoir, en l'occurrence pour obtenir l'ouverture d'une petite fenêtre vers le futur, basée sur la probabilité d'occurrence d'événement.

Et ces probabilités, si elles sont utilisées dans un cadre de règles connues et bien définies, permettent alors d'aller jusqu'à la prise de décision automatisée, à travers l'utilisation d'un système autonome tel que l'Intelligence Artificielle.

Nous comprenons donc que l'IA est en lui-même un système autonome et automatisé de prise de décision. Mais cela ne nous permet pas de définir ce qu'elle est vraiment, mais plutôt son but final. L'IA reste toujours à ce stade un concept mystérieux pour une majeure partie de personnes et nous allons tenter d'éclaircir ce concept au maximum dans cette partie.

1.2.1 Vers une définition

Plusieurs définitions existent. En voici quelques-unes :

> *Intelligence Artificielle : Ensemble des théories et techniques complexes de développement informatique capables de simuler certains traits de l'intelligence humaine.*

Dictionnaire leRobert – 2023

> *Intelligence Artificielle : Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine.*

Dictionnaire Larousse – 2023

> *It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.*

Article "What is AI ?" - John McCarthy, Un des pionniers de l'IA - 2004

Une traduction serait : « *Il s'agit de l'étude et de la conception de machines intelligentes, notamment de logiciels informatiques intelligents. Cette discipline est connectée à l'effort d'employer des ordinateurs pour comprendre l'intelligence humaine, cependant l'IA n'est pas contrainte de se restreindre uniquement aux méthodes perceptibles biologiquement* ».

> *The building of computer programs which perform tasks which are, for the moment, performed in a more satisfactory way by humans because they require high level mental processes such as: perception learning, memory organization and critical reasoning.*

***Marvin LEE MINSKY** - 1956*

Une traduction serait : « *Il s'agit de l'élaboration de logiciels capables d'exécuter des tâches qui, à l'heure actuelle, sont mieux réalisées par les humains car elles nécessitent des processus cognitifs supérieurs tels que la perception, l'apprentissage, la structuration de la mémoire et le jugement critique* ».

> *The study and design of intelligent agents, where an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success.*

Dictionnaire Larousse – 2023

Une traduction serait : « *Il s'agit de l'étude et de la création d'agents intelligents, un agent intelligent étant un système capable de percevoir son environnement et d'agir de manière à optimiser ses chances de succès* ».

En d'autres termes, **l'Intelligence Artificielle (ou « IA ») est une science qui imagine et conçoit des outils informatiques capables de mimer les processus cognitifs humains**, processus mentaux impliqués dans l'acquisition, la compréhension des connaissances et la prise de décision : des machines intelligentes capables de raisonner, d'analyser et d'interagir, mais surtout de prendre des décisions, tout comme nous.

Et si nous devions retenir peut-être une seule définition, nous retiendrons sans doute la définition donnée par **Stuart RUSSEL et Peter NORVIG**, de célèbres chercheurs américains en IA, dans leur manuel révolutionnaire intitulé « Intelligence Artificielle : une approche moderne », véritable

référence en la matière. Les auteurs RUSSEL et NORVIG abordent la question de l'IA en unifiant leurs travaux pour couvrir de façon complète et moderne aussi bien le volet théorique et pratique de l'Intelligence Artificielle. Ils centrent alors leur approche autour du thème des « agents intelligents » pour proposer une vision unifiée de l'IA.

Ainsi, RUSSEL et NORVIG expliquent qu'il y a 4 écoles de « pensée » ou **4 objectifs potentiels de l'IA**, en se basant sur la rationalité et la réflexion, selon l'action de l'IA. Ils ajoutent à cela également quelques définitions supplémentaires dont voici le tableau résumé :

Approche humaine	**Approche rationnelle**
Systèmes qui pensent comme des êtres humains	**Systèmes qui pensent rationnellement**
« La tentative nouvelle et passionnante d'amener les ordinateurs à penser. . . [d'en faire] des machines dotées d'un esprit au sens le plus littéral. » (**Haugeland**, 1985) « [L'automatisation d'activités que nous associons à la pensée humaine, des activités telles que la prise de décision, la résolution de problèmes, l'apprentissage. . . » (**Bellman**, 1978)	« L'étude des facultés mentales grâce à des modèles informatiques. » (**Charniak et McDermott**, 1985) « L'étude des moyens informatiques qui rendent possibles la perception, le raisonnement et l'action. » (**Winston**, 1992)
Systèmes qui agissent comme des êtres humains	**Systèmes qui agissent rationnellement**
« L'art de créer des machines capables de prendre en charge des fonctions exigeant de l'intelligence quand elles sont réalisées par des gens. » (**Kurzweil**, 1990) « L'étude des moyens à mettre en œuvre pour faire en sorte que des ordinateurs accomplissent des choses pour lesquelles il est préférable de recourir à des personnes pour le moment. » (**Rich et Knight**, 1991)	« L'Intelligence Artificielle (computational intelligence) est l'étude de la conception d'agents intelligents. » (**Poole et al.**, 1998) « L'IA. . . étudie le comportement intelligent dans des artefacts. » (**Nilsson**, 1998)

Figure 14 : Tableau des 4 catégories de RUSSEL & NORVIG – Source : Cours sur l'IA de Pearson France

Les système d'IA, selon leur nature, pourraient donc se classer dans une de ces catégories, classées selon deux dimensions :

- En ligne :
  - Ligne 1 : Approche basée sur le raisonnement
  - Ligne 2 : Approche basée sur le comportement
- En colonne :
  - Colonne 1 : se compare aux performances humaines
  - Colonne 2 : s'évalue selon la rationalité, si le système d'IA opère de manière appropriée en conséquence des éléments qu'il connait

Pendant que l'approche humaine se base sur des éléments issus d'observation et d'hypothèses sur le comportement humain, l'approche rationnelle (idéale) fait appel aux mathématiques et à l'ingénierie.

C'est cette approche qui est la plus répandue, celle basée sur des probabilités.

1.2.2 Les fondations de l'IA

L'Intelligence Artificielle, qui est un sujet complexe par excellence, est fondée sur une base issue de la convergence de plusieurs domaines :

- **Les mathématiques** : ce sont la vraie fondation de l'Intelligence Artificielle car elle repose sur la modélisation prédictive qui est un calcul mathématique de probabilités.
- **L'économie** : c'est un domaine d'application de l'Intelligence Artificielle dans la mesure où l'on recherche toujours une forme d'optimisation.
- **Les neurosciences et la psychologie** : l'Intelligence Artificielle repose sur les principes cognitifs de choix et de prise de décision et tout le domaine des réseaux neuronaux tente de reproduire le mode de fonctionnement du cerveau humain.
- **La philosophie** : en relation avec les choix et décisions, la philosophie est un fondement notamment vis-à-vis des dimensions morales et éthiques de l'IA.
- **L'informatique et la cybernétique** : Ce sont les moyens technologiques de calcul, de mise en œuvre et d'exécution.

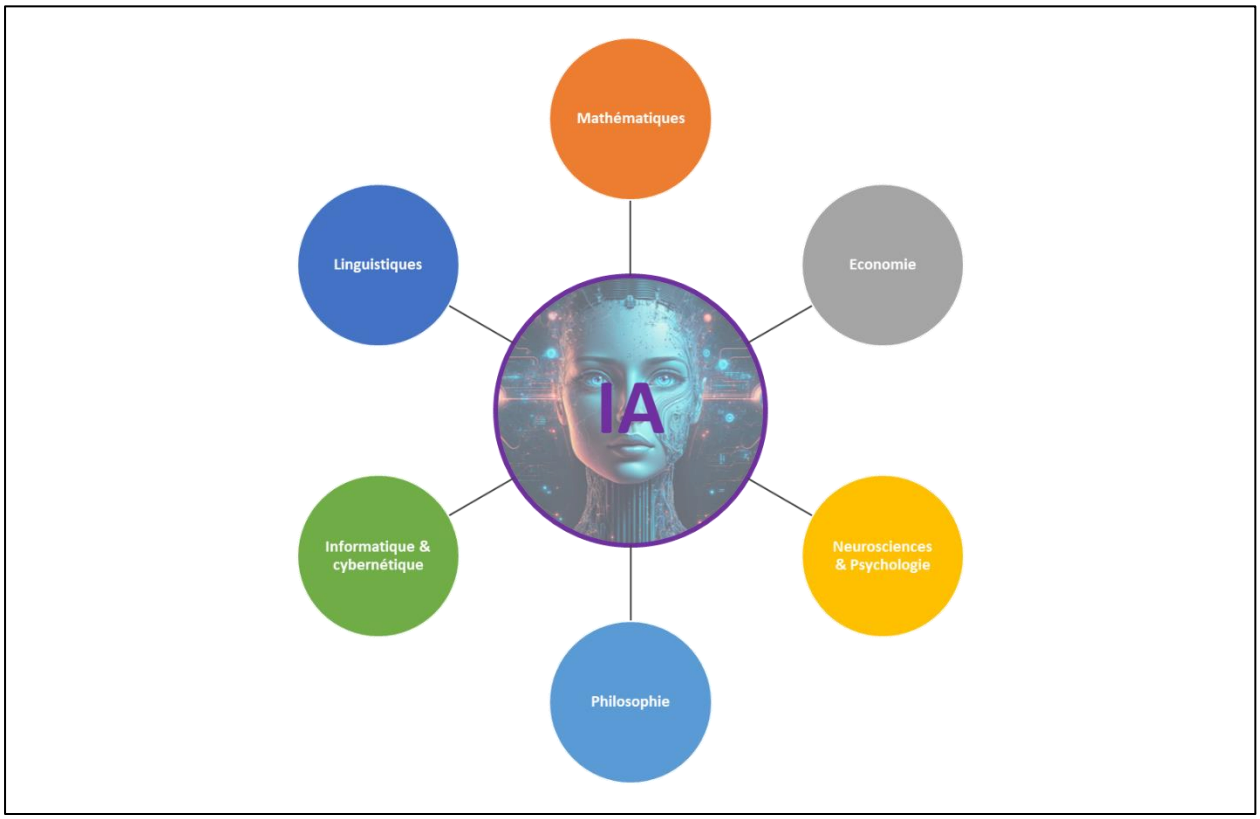


Figure 15 : Convergence de plusieurs domaines pour constituer l'IA – Source : autrice

1.2.3 L'histoire de l'IA

L'intérêt pour l'être humain pour une Intelligence Artificielle ne date pas d'hier. En effet, l'Humain a depuis longtemps imaginé des **« systèmes autonomes » qu'on pourrait appeler d'automate**.

1.2.3.1 L'histoire lointaine : des automates…

Dès la mythologie grecque, on retrouve une référence au géant **Talos, gardien de la Crète**. Sa tâche consiste à faire trois fois le tour de l'île chaque jour et à chasser les intrus au nom du roi légendaire de Crète, **Minos**. Selon la légende, il aurait été donné par le dieu **Héphaïstos** à **Minos** et est un automate forgé par Héphaïstos lui-même.

Figure 16 : géant Talos, gardien de la Crète – Source : wikipedia.org

Héron d'Alexandrie (fin 1er siècle après J.-C), géomètre et ingénieur grec et originaire d'Alexandrie, rédigea des manuels de construction d'un grand nombre d'automates dans deux ouvrages :

- Dans **Pneumatica**, où il présente des automates utilisant l'air, la vapeur ou l'eau,
- Dans **Automata**, où il présente des machines conçues pour le théâtre, dont les mécanismes sont à base de poids et de contrepoids.

Il n'a pas été prouvé que ces automates aient été un jour construits. Néanmoins, ces ouvrages présentent des conceptions de mécanismes très avancés ne nécessitant pas d'intervention humaine.

Figure 17 : Reproduction de l'éolipyle de Héron par John R. Bentley

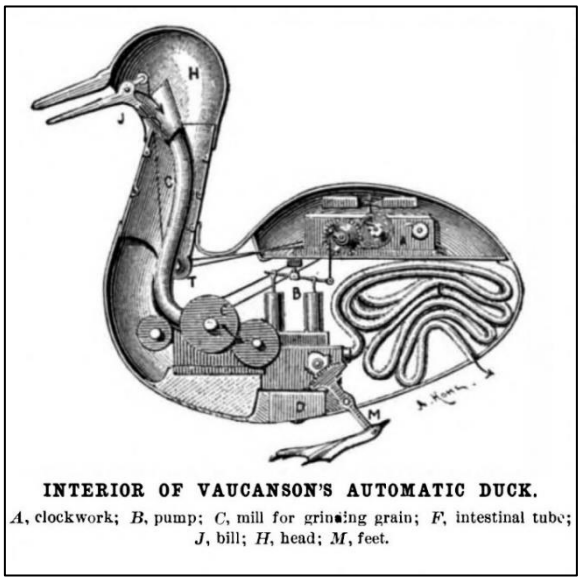
Le moyen âge voit la création de plusieurs automates célèbres dont un lion créé par **Léonard DE VINCI** en 1515. Ce lion mécanique extraordinaire a été créé initialement pour amuser et impressionner **François Ier**, roi de France de l'époque. Il lui a été présenté lors du passage de ce dernier dans la ville de Lyon.

Grâce à un mécanisme interne, la poitrine de l'animal mécanique s'ouvrait pour faire apparaître une fleur de lys. Symboliquement, le lys représente le roi de France et la ville de Florence, tandis que le lion est associé à la ville de Lyon, au blason de Florence, au pape **Léon X**, ainsi qu'à Léonard lui-même.

Figure 18 : Reconstitution du lion de Léonard DE VINCI réalisée en 2009 – Source : lyonhistorique.fr

En 1738, **Jacques DE VAUCANSON**, inventeur et mécanicien français, invente plusieurs automates dont le flûteur et le canard artificiel de VAUCANSON.

Concernant le flûteur automate, il recrée fidèlement le jeu d'un flûtiste sur un instrument identique à ceux en usage à l'époque. Le canard de Vaucanson fut très célèbre pour la diversité de ses mouvements naturels ainsi que pour le réalisme avec lequel il simule la digestion jusqu'à la défécation.



INTERIOR OF VAUCANSON'S AUTOMATIC DUCK.
A, clockwork; *B*, pump; *C*, mill for grinding grain; *F*, intestinal tube; *J*, bill; *H*, head; *M*, feet.

Figure 19 : Anatomie du canard artificiel de VAUCANSON – Source : wikipedia.org

1.2.3.2 L'histoire récente : 73 ans d'évolution

Nous avons pu voir précédemment que l'Humain, par son ingéniosité, étaient tout à fait capable de réaliser des systèmes « mécaniques » ne nécessitant pas d'intervention humaine.

Ce n'est qu'au début des années 50 qu'apparaissent les premières questions autour de l'Intelligence Artificielle, notamment avec le développement des premiers ordinateurs. Considérés comme les deux fondateurs de l'informatique, **Alan TURING** et **John VON NEUMANN** s'y intéresse d'ailleurs tous deux rapidement et débute ainsi l'histoire récente de l'Intelligence Artificielle moderne :

- **1950** : Dans son célèbre article « *Computing Machinery and Intelligence* », **Alan TURING** démarre une réflexion sur la fameuse question « les machines peuvent-elles penser ? ». Il a également développé le test qui porte son nom (**le test de Turing**), qui vise à établir si un ordinateur peut manifester le même niveau d'intelligence qu'un être humain (ou capable de produire des résultats ou des réponses de manière autonome, comme le ferait un humain). Ce test fait objet de nombreux débats encore aujourd'hui et nous aurons l'occasion d'y revenir plus en détails plus tard dans cette thèse.

- **1956** : Lors d'une conférence au **Dartmouth College**, **John MCCARTHY** invente le terme « Intelligence Artificielle ». Cette même année, **Allen NEWELL**, **Herbert SIMON** et **Cliff SHAW** créaient ce qui s'apparent au tout premier progiciel d'Intelligence Artificielle, le « **Logic Theorist** ». Celui-ci a été conçu pour reproduire les compétences d'un être humain en termes de résolution de problèmes. Commence alors une période florissante pour la recherche, durant laquelle sont réalisées de nombreuses avancées notamment autour des premiers réseaux de neurones.

- **Début des années 70** : Bien que les premiers ordinateurs montraient de grandes capacités machines, ils furent très vite limités. **La puissance et la mémoire de l'époque sont alors un frein au développement de l'IA**, les programmes les plus performants peinant à manipuler des versions simplistes des problèmes qu'ils sont censés résoudre. **L'IA connait alors une période moins florissante et les financements sont suspendus**. En raison d'un pessimisme quant aux véritables capacités de l'IA, de nombreux fonds et subventions ont été réduits et plusieurs projets ont été laissés de côté. Cela mène à un ralentissement de la recherche en IA dans les années 70-80, période appelée « **1er hiver de l'IA** ».

- **Début des années 80** : En dépit de la baisse des financement, les travaux ne s'interrompent pas totalement. Ainsi, **les systèmes experts apparaissent, permettant de résoudre des problèmes**. Ces systèmes imitent la capacité de l'homme à prendre des décisions et **utilisent un raisonnement basé sur des « règles »** en recourant principalement à une structure « si-alors » mise en œuvre pour répondre à des questions. Par ailleurs, de grandes avancées permettent l'évolution des réseaux neuronaux jusque-là limités. Enfin, durant cette décennie, **la puissance de calcul augmente de manière fulgurante** et l'informatique est de plus en plus accessible et utilisée, dans de plus en plus nombreux secteurs. Cela permet alors l'accélération du développement de l'IA.

- **Fin des années 80 et début des années 90** : c'est le « **2ème hiver de l'IA** ». Basés sur des règles et des connaissances, les **systèmes experts** montrent alors des limites et provoquent l'abandon de nombreux projets ainsi qu'une forte réduction des financements. Malgré cela, les recherches en IA se poursuivent particulièrement dans le sous-domaine de l'apprentissage du **Machine Learning, basé sur l'analyse statistique de grandes quantités de données**, permettant alors quelques avancées comme **la reconnaissance d'image ou l'identification et la compréhension de mots dans une langue humaine**.

- **1997** : L'Intelligence Artificielle n'avait pas vraiment marqué les esprits jusque-là. Il aura donc fallu un match d'échecs (6 parties) en plein « **Hiver de l'IA** », et que **Deep Blue d'IBM** batte le champion du monde de l'époque, **Garry KASPAROV**.

Figure 20 : Deep Blue d'IBM bat Garry KASPAROV aux échecs – Source : theworld.org

Grâce à cet événement, le monde se rendit compte véritablement du potentiel de l'Intelligence Artificielle.

- **2011** : **Watson**, programme informatique d'IA conçu par IBM, devient en battant ses concurrents humains (**Ken JENNINGS** et **Brad RUTTER**) le champion du jeu télévisé américain « **Jeopardy!** ». Supercalculateur, il réussit cette prouesse notamment grâce à sa superpuissance (90 serveurs & 2880 processeurs) et la quantité phénoménale de données qu'il a ingurgité : 15 téraoctets de données (pour l'époque !!!) soit l'équivalent de 1 million de livres.

Figure 21 : Watson d'IBM devient champion du jeu télévisé américain « Jeopardy! » - Source : nytimes.com

- **2015** : **Minwa**, **système d'IA et moteur de recherche chinois de la société Baidu**, utilise un type particulier de réseau de neurones profonds (sera détaillé plus loin dans cette partie). Celui-ci est appelé **réseau de neurones convolutifs** et permet d'identifier et classer des images avec un taux de précision très élevé. En mai 2015, il relève le défi **ImageNet** consistant à analyser 1,2 millions d'images répartis en 1000 catégories puis à tenter de reconnaitre 15000 images. A l'issue du test, il obtient 4,58% comme pourcentage d'erreurs, ce qui est meilleur que les résultats de ses concurrents (**Google** et **Microsoft** entre autres). Cela fait de lui **le système d'IA ayant le taux de précision supérieur à celui de l'être humain moyen le plus élevé** dans cet exercice de reconnaissance d'image (5% comme pourcentage d'erreurs pour l'humain).

- **2017** : **AlphaGo**, système d'IA capable de jouer au **jeu de go** et développé par l'entreprise **Deepmind** rachetée en 2014 par Google, **bat le champion du monde** de jeu de go **Ke JIE** (9ème dan professionnel). Sa victoire la plus célèbre reste sa prouesse en 2016 en battant Lee SEDOL (9ème dan professionnel). AlphaGo avait également battu auparavant le champion d'Europe **Fan HUI** (2ème dan professionnel).

Figure 22 : AlphaGo bat Lee Sedol au jeu de go en 2016 – Source : leparisien.fr

- **Années 2020** : Les années 2020 consacrent le **développement des enjeux de confiance et d'explicabilité dans l'IA**. Ils sont également synonymes de grand bonds en avant notamment autour des modèles utilisés pour la génération autour du langage (**« Natural Language Processing »** ou **NLP**). Ainsi, les deux grands acteurs que sont **Microsoft** et **OpenAI** annonce dès 2020 respectivement **Turing NLG** (**T-NLG**) et **GPT-3**, permettant un bond considérable en avant.

- **2022** : La startup française **NukkAI bat les meilleurs joueurs de bridge du monde** au moyen d'un programme d'IA hybride expliquant ses raisonnements. Par ailleurs, l'événement le plus célèbre est la mise à disposition mondiale de l'agent conversationnel **ChatGPT**, développé par **OpenAI** et utilisant le modèle **GPT-3** et près de 175 milliards de données avec une fraicheur allant jusqu'à 2021.

L'histoire récente de l'IA est donc riche et est résumé à travers le schéma suivant :

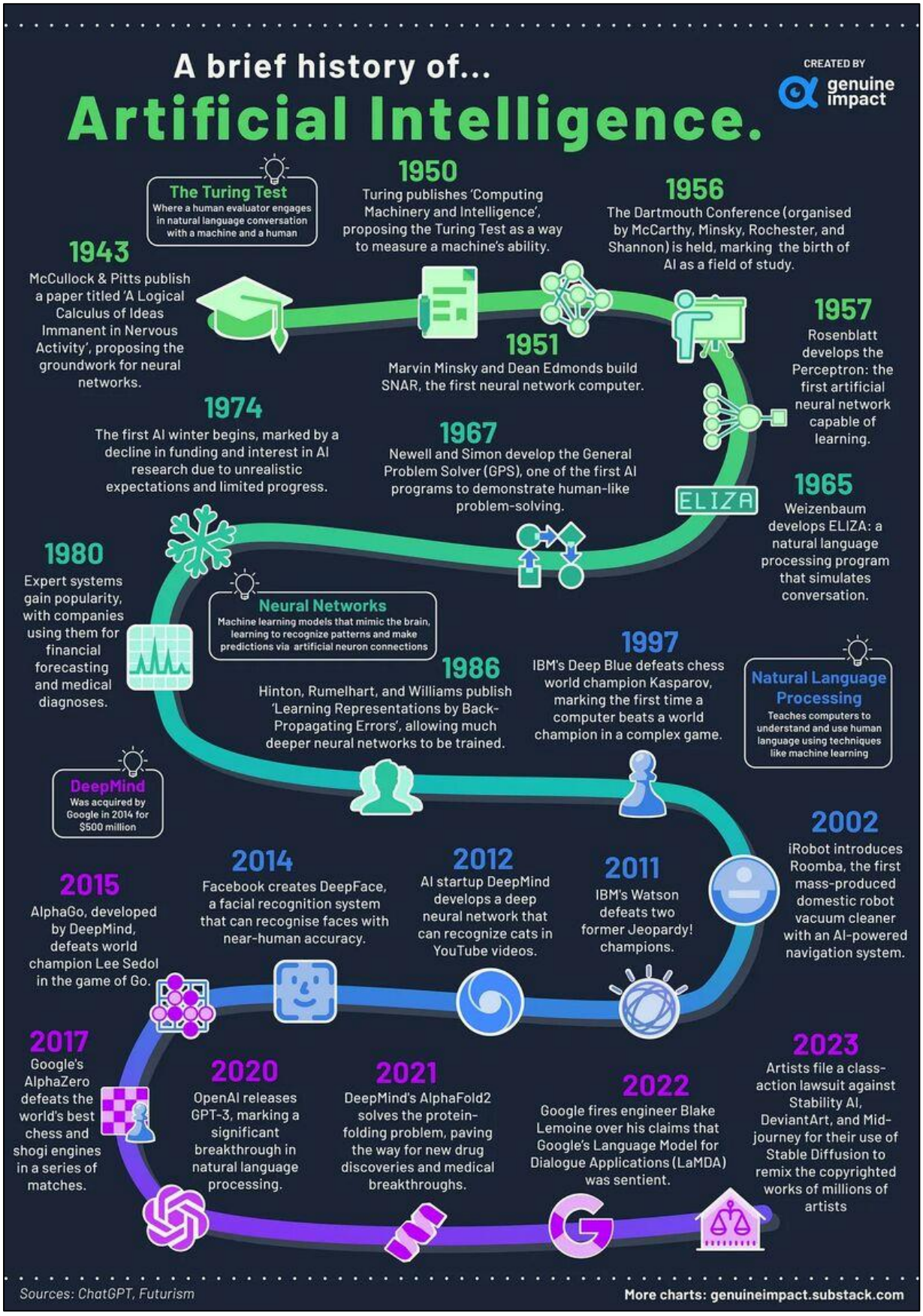


Figure 23 : L'historique de l'Intelligence Artificiel selon Genuine Impact - Source : [substack.com]

1.2.4 L'accélération de ces dernières années

Nous l'avons vu précédemment : l'histoire de l'Intelligence Artificielle n'a rien de nouveau. Et pourtant, en à peine quelques années, son évolution a bouleversé nos vies.

Mais il est important de bien distinguer deux éléments :

- La prise de conscience du potentiel de l'IA par le public laquelle a été fulgurante et a généré une incroyable frénésie autour de l'IA depuis 2022 et notamment avec l'arrivée de ChatGPT : c'est le point de vue « utilisateur »
- Les capacités de l'IA, lesquels ont réalisé des progrès rapides, depuis cette dernière décennie : c'est le point de vue « scientifique », de celui de l'IT.

Pourquoi ces progrès si rapides en termes de capacité de l'IA dans la dernière décennie ?

Aujourd'hui, l'IA est sur une voie express en raison de quatre axes ou secteurs d'améliorations majeurs :

- La puissance de calcul
- Les capacités de stockage
- Les capacités logicielles (algorithmes, autres)
- Les investissements

1.2.4.1 L'augmentation de la puissance de calcul : les lois de MOORE et « loi de HUANG »

L'IA a besoin d'une très grande capacité de calcul. D'ailleurs, les systèmes d'IA étaient initialement basés sur des supercalculateurs. Or, qui dit calcul, dit puissance de calcul et donc processeur (ou CPU en anglais).

Un CPU est un réseau de milliards de transistors reliés entre eux et qui prennent chacun la valeur de 0 ou 1. De nos jours, les processeurs les plus puissants comptent plus de 50 milliards de transistors sur une surface de quelques cm² (6 cm² exactement).

Gordon E. MOORE, est le cofondateur avec **Robert NOYCE** et **Andrew GROVE** de la société **Intel** en 1968, premier fabricant mondial de microprocesseurs.

En 1958, il fit l'observation que **le nombre de transistors par circuit intégré de même taille doublait tous les ans pour un même coût de fabrication**. Le 19 avril 1965, il publie alors dans le magazine **Electronics** cette observation comme la célèbre loi empirique qui porte son nom aujourd'hui. Plus exactement, elle fut appelée « **première loi de Moore** ».

En 1975, Gordon MOORE réévalua sa prédiction : le nombre de transistors des microprocesseurs sur une puce de silicium double selon lui tous les deux ans. C'est ce qui est appelé la « deuxième loi de Moore ».

En effet, depuis l'invention du 1er microprocesseur par Intel en 1971, qui comptait 2 300 transistors, **la densité de transistors par puce se multipliait par 2 environ tous les 2 ans**, au rythme de la prédiction de Gordon MOORE de 1975.

MOORE avait également anticipé que l'industrie parviendrait à miniaturiser les transistors sans augmenter le coût. Cela explique pourquoi chaque nouvelle génération d'ordinateur coûte aussi cher que la précédente mais avec un doublement de leur puissance de calcul.

En particulier chez Intel, le leader du secteur entre 1990 et 2010, cette règle a fonctionné plus de 40 ans. Mais depuis les années 2010, la loi de MOORE n'est plus vérifiée et la production de microprocesseur est de plus en plus coûteuse, notamment chez Intel.

De plus, elle ne colle plus avec les besoins en termes de puissance de l'époque moderne, notamment ceux des systèmes d'IA. Depuis 2013, **il a été observé que la puissance de calcul d'un système d'IA double en réalité tous les 4 mois, soit 6x plus rapidement que la loi de MOORE**.

De son côté, lorsqu'il fonda **Nvidia** au début des années 1990, **Jensen HUANG** avait pour objectif de **concevoir des puces graphiques nommées GPU** (**Graphic Processing Unit**) pour obtenir un meilleur rendu graphique que les processeurs d'Intel (**CPU** - **Central Processing Unit**).

En 2007, après s'être spécialisé dans un premier temps à faire des puces graphiques pour les jeux vidéo et être devenu le leader dans ce domaine, Nvidia décide alors de construire des puces permettant d'accélérer les calculs scientifiques et notamment ceux des systèmes d'IA. **Les GPUs de Nvidia deviennent alors incontournables à partir de 2016** dans les serveurs et les datacenters des entreprises. Aujourd'hui, pour certaines types d'usage, **les GPUs de Nvidia sont 1000x plus puissants que les CPUs d'Intel**.

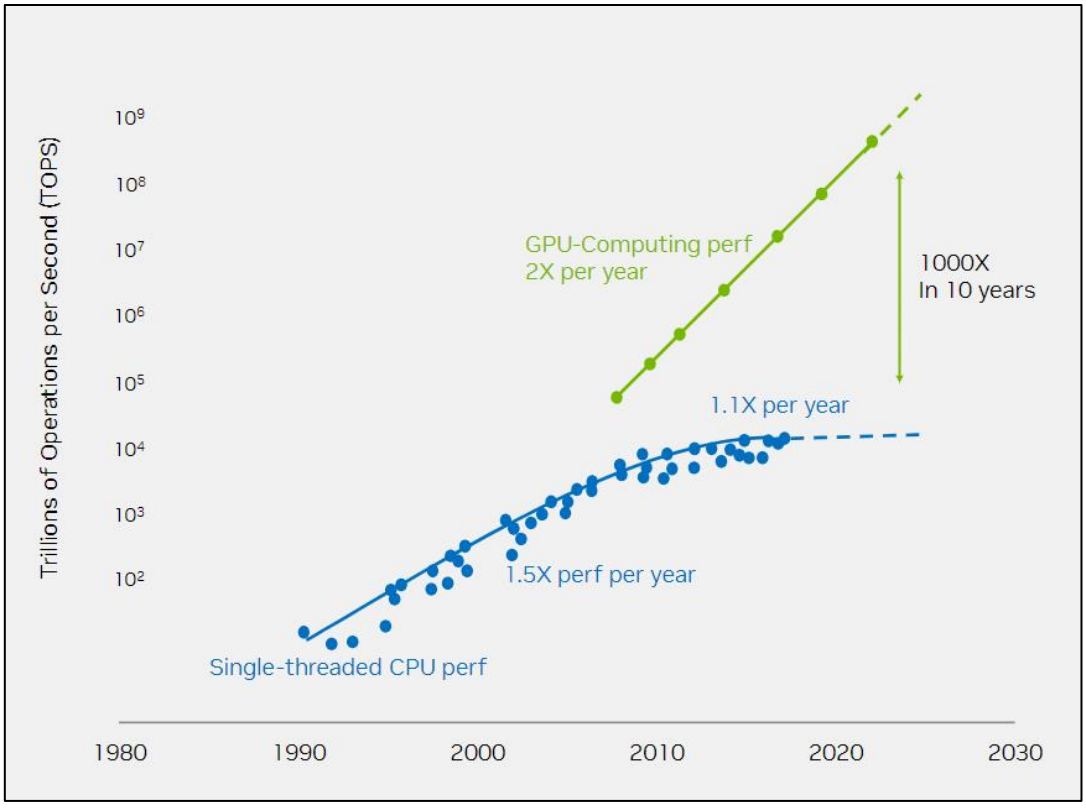


Figure 24 : La loi de Huang est applicable aux GPUs - Source : banquetransatlantique.com

Alors qu'Intel se concentra ces dernières années sur la miniaturisation de ses transistors, Nvidia s'est concentrée sur la conception de **meilleures architectures** beaucoup plus adaptées à l'IA que les processeurs d'Intel, en faisant le leader incontestée dans ce domaine et surtout permettant à l'IA de faire un bond considérable. Ce choix fait effet : sa côté boursière a augmenté en une journée seulement de 180 milliards de dollars fin mai 2023 avec la sortie de son dernier GPU (le H 100) dédié à l'IA.(source rtl.fr).

Pour HUANG, PDG de Nvidia : « une nouvelle loi est en cours ». Selon lui, les GPU actuels de Nvidia seraient aujourd'hui 25 fois plus rapides qu'il y a 5 ans.

1.2.4.2 L'augmentation des capacités de stockages : le Big Data

La vitesse de création des données numériques, leur variété et leur quantité augmentent sans cesse année après année.

Selon les dernières estimations du site **statista**, **le volume de données** numériques au niveau mondiale **est passé de 2 zettaoctets (1000 milliards de gigaoctets) à 64 zettaoctets en 2020**, soit une multiplication par plus de 30 au cours de cette décennie.

Et cela ne va pas cesser de croître dans les prochaines années : le volume de données à l'échellement mondiale pourrait **dépasser 180 zettaoctets en 2025**, **soit une croissance moyenne annuelle d'environ 40 % par an sur cinq ans**. Devant une telle volumétrie, il y a un défi technologique majeur.

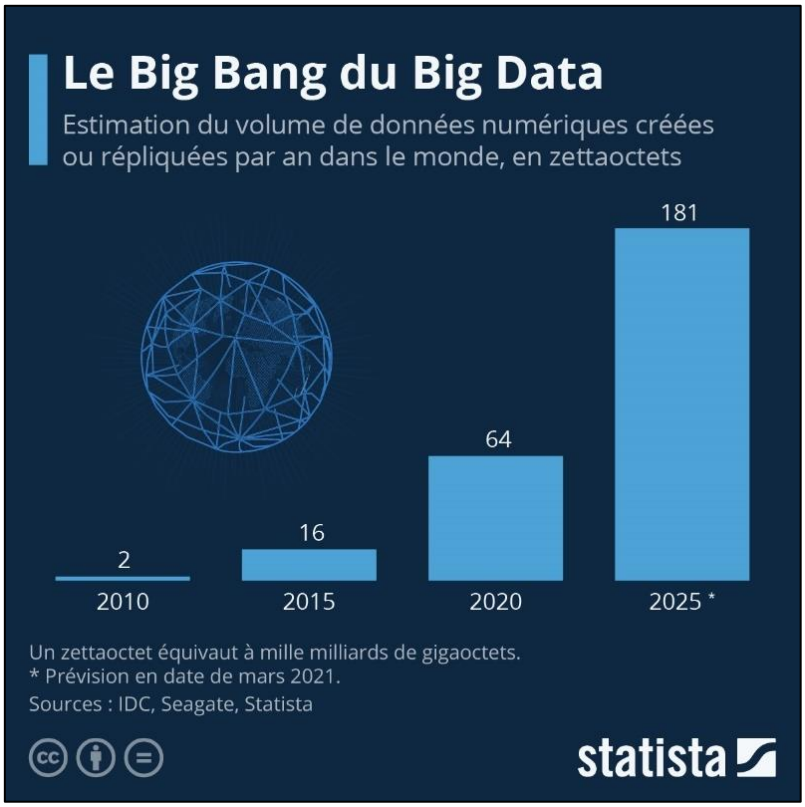


Figure 25 : Augmentation du volume de données numériques par an dans le monde - Source : statistica.com

Le « **Big Data** » fait justement référence à ces données volumineuses, ou données « **massives** ». C'est un terme utilisé pour qualifier un **très grand volume de données**, lesquelles ne peuvent être traitées par des outils classiques/traditionnels.

Le Big Data correspond à des données complexes, provenant essentiellement de nouvelles sources, tel que les objets connectés (« Internet des ObjeTs » ou « Internet Of Things » en anglais - **IOT**), appareils et services comme la géolocalisation, les réseaux sociaux, la gestion des historiques de navigation web, des applications mobiles, etc.

Ces données ont principalement **3 caractéristiques majeures** (les fameux 3 « V » du Big Data) :

- Le **Volume** : une quantité très importante de données est à traiter,
- La **Vitesse** : la vitesse à laquelle les données sont reçues et éventuellement traitées potentiellement en quasi-temps réel,
- La **Variété** : de nombreux types de données peuvent être traités (structurées, non-structurées, semi-structurées, etc.).

Deux autres « V » ont vu le jour récemment pour les amener à 5 « V » :

- La **Valeur** : les données ont de la valeur existant à l'intérieur qui est à découvrir,
- La **Véracité** : la fiabilité des données collectées est à vérifier et doivent être corrigées si nécessaire.

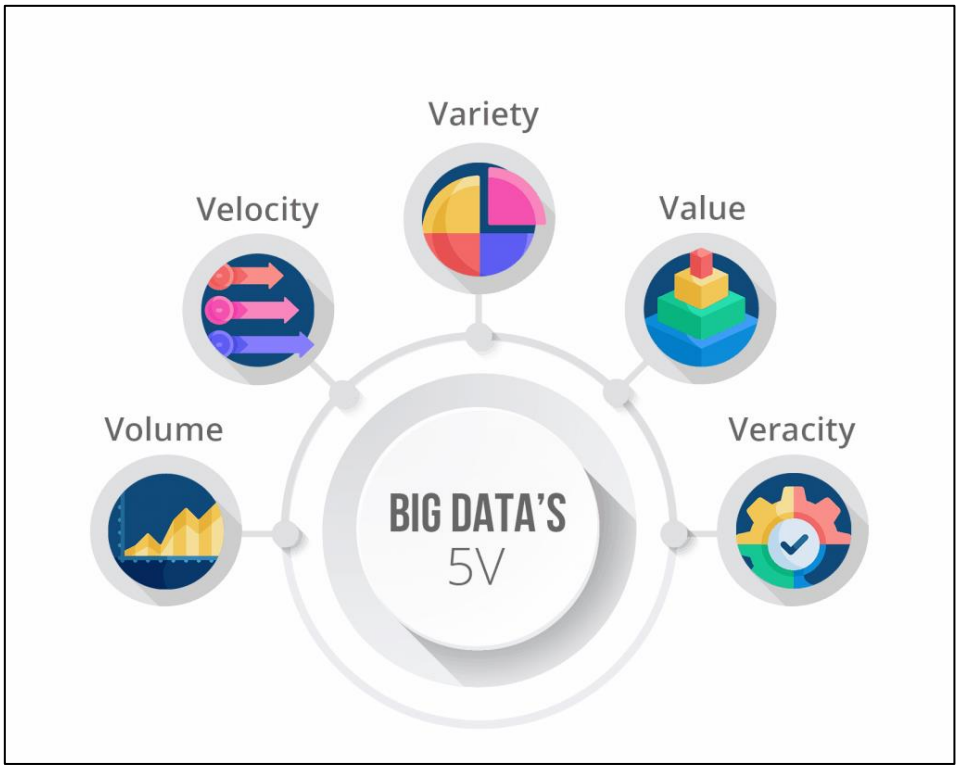


Figure 26 : Les 5 "V" du Big Data - Source : kanoppe.com

C'est en 2005 qu'on assista à une véritable prise de conscience collective de la quantité de données croissante, notamment autour de services en ligne tel que **Youtube** et **Facebook**. Il fallut alors trouver

une réponse technologique à cette croissance : Le **développement d'infrastructures « open source » telle que Hadoop** a été primordial pour la croissance du Big Data pour l'époque, car **permettant théoriquement de réduire les coûts de stockage ainsi que des traitements parallélisés**.

L'augmentation du volume des données est une aubaine pour l'Intelligence Artificielle car elle se nourrit de données : L'intelligence a donc besoin du Big Data et son essor récent converge avec celui de l'IA. **Le Big Data et l'Intelligence Artificielle sont deux technologies incontestablement liées.**

Néanmoins, cette technologie est peu à peu abandonnée au profit de solutions Cloud qui permettent notamment de décupler les possibilités de traitement et de stockage.

En effet, Les plateformes de « **cloud computing** » offrent des ressources informatiques abordables et évolutives, permettant aux chercheurs et aux développeurs d'accéder facilement à une puissance de calcul considérable pour l'entraînement et le déploiement de modèles d'IA. Cela a contribué à réduire les barrières à l'entrée et à démocratiser l'IA.

Si toutefois la rencontre (convergence) entre la puissance informatique et le Big Data avec l'IA ont contribué à son accélération, celui des capacités logicielles a été plus que nécessaire.

1.2.4.3 L'augmentation des capacités logicielles : de nouvelles avancées scientifiques et industrielles

Une autre raison qui a permis l'accélération de l'Intelligence Artificielle ces dernières années est l'avancée scientifique notamment autour des algorithmes ou modèles d'IA :

- **Amélioration des algorithmes d'apprentissage automatique** : Les chercheurs en IA ont développé de nouveaux algorithmes d'apprentissage automatique, tels que les **réseaux de neurones profonds**, qui ont la capacité d'assimiler des structures de données complexes et de surmonter des problèmes difficiles. Ces avancées ont permis des progrès significatifs dans des domaines tels que la vision par ordinateur, le traitement du langage naturel et la reconnaissance vocale.

- **Disponibilité d'outils en « open source » (source ouverte)** : Les outils essentiels au développement sont en accès libre, souvent gratuit sauf parfois dans leur version entreprise pour des questions de support. Cette ouverture de la communauté du libre permet des avancées significatives car ces outils "libres" sont plus faciles à modifier rapidement. **Ils permettent de rebondir plus vite et de rester compétitif. Exemples : certains outils d'OpenAI**, **TensorFlow**, **PyTorch**, **Scikit-Learn, etc**.

- **Disponibilité de grands ensembles de données annotées, notamment en « open data »** : L'apprentissage automatique nécessite souvent de grandes quantités de données annotées pour entraîner des modèles. Au fil du temps, des ensembles **de données étiquetés de haute qualité ont été créés dans divers domaines, ce qui facilite l'entraînement de modèles d'IA** plus précis et performants.

- **Avancées dans le traitement du langage naturel** : L'amélioration de la compréhension du langage naturel par les machines a été un moteur majeur de l'essor de l'IA. Les techniques telles que la modélisation des langages, les réseaux de neurones récurrents et les transformateurs ont permis des avancées significatives dans la traduction automatique, la génération de texte et l'analyse des sentiments, ouvrant ainsi de nouvelles opportunités pour les applications basées sur le langage.

Aujourd'hui, l'Intelligence Artificielle et les machines dotées d'IA peuvent accomplir les actes suivants (liste non exhaustive) :

- Reconnaissance vocale
- Apprentissage
- Planification
- Résolution de problèmes
- Raisonnement
- Perception
- Possibilité de manipuler et de déplacer des objets

Voici un aperçu rapide de ce qu'il est maintenant possible de faire grâce aux avancées sur les modèles (liste non-exhaustive) :

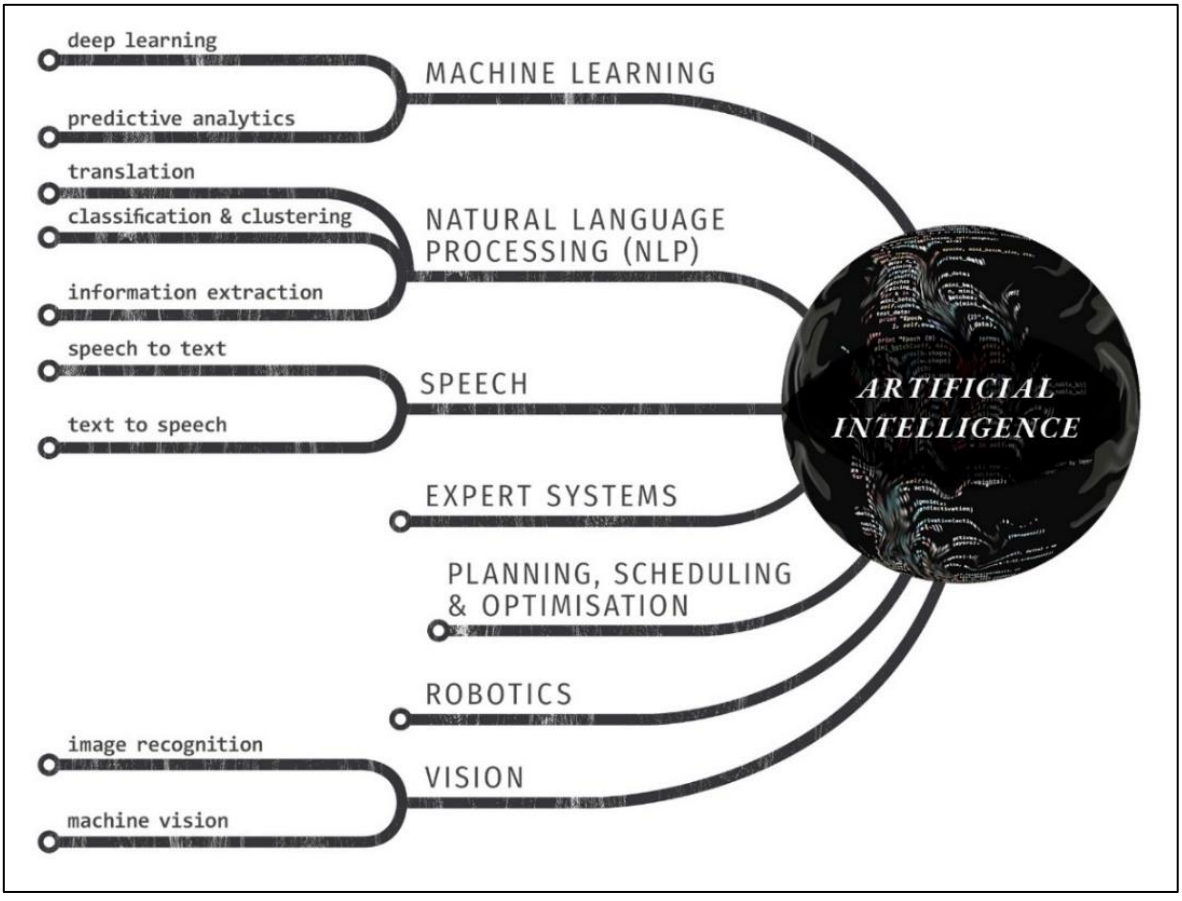


Figure 27 : Les possibilités offertes par les derniers modèles – Source : [syndicoop.fr]

1.2.4.4 L'augmentation des investissements et de l'intérêt pour l'IA

Le 21e siècle a été marqué par une augmentation significative des investissements dans le domaine de l'IA, tant du secteur privé que des organismes gouvernementaux et académiques. Cela a stimulé la recherche, favorisé l'innovation et créé un écosystème propice au développement de l'IA.

Au niveau de la France, **notre président Emmanuel Macron avait annoncé en mars 2018 un plan d'1,5 milliard d'euros pour développer l'Intelligence Artificielle** et attirer de nouveaux chercheurs, et cela afin que « La France ne rate pas le train de l'Intelligence Artificielle ».

De nos jours, l'Intelligence Artificielle déchaîne les passions des investisseurs, notamment sur le matériel comme les processeurs Nvidia, spécialisés pour l'IA comme vu ensemble précédemment, mais aussi sur le marché des logiciels d'IA dont voici un graphique du budget pressenti par an en US Dollars (USD) :

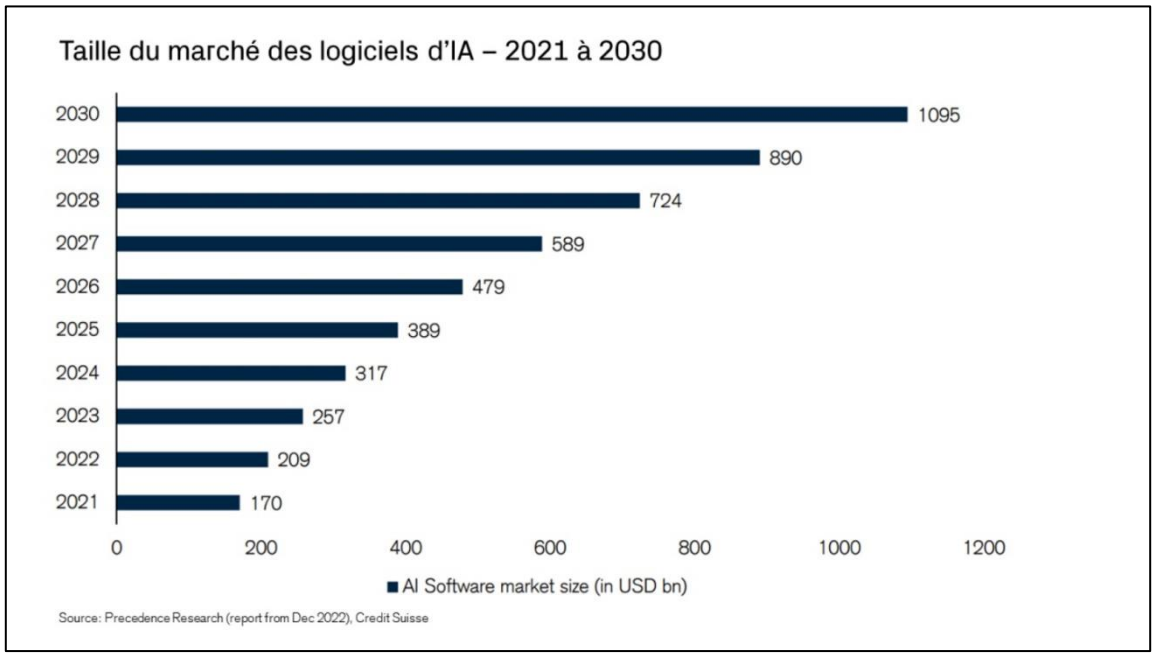


Figure 28 : évolution du marché des logiciels d'IA de 2021 à 2030 - Source : credit-suisse.com

Selon une étude de **Precedence Research (Credit Suisse)**, le marché des logiciels d'IA devrait connaitre une croissance à un taux annuel moyen de 23 % entre 2021 et 2030.

> *Avec l'explosion de l'accès aux données, de la puissance de calcul et de la rapidité que permet le cloud, les entreprises peuvent déployer l'IA à grande échelle plus rapidement, améliorer leurs résultats et accélérer le retour à une croissance durable.*

***Sanjeev VOHRA**, Global Business Lead of Accenture Applied Intelligence*

1.2.5 Les niveaux d'Intelligence Artificielle

Il existe principalement 3 types d'IA dont voici un rapide résumé :

- **L'Intelligence Artificielle étroite ou Artificial Narrow Intelligence (ANI)** : Son autre nom est « **l'Intelligence Artificielle faible** ». Elle met l'accent sur des tâches spécifiques et est conçue pour accomplir une seule tâche avec une grande efficacité, comme la reconnaissance vocale ou la prédiction des recommandations de films, les voitures sans conducteur, par exemple.
- **L'Intelligence Artificielle générale ou Artificial General Intelligence (AGI)** : Elle est également appelée « **l'Intelligence Artificielle forte** ». Elle vise à atteindre le niveau de l'intelligence naturelle humaine, à l'imiter ou à imiter les comportements de l'être humain ce qui la rendrait indiscernable vis-à-vis de l'humain.

- **La Super-Intelligence Artificielle ou Artificial Super Intelligence (ASI)** : L'ASI signifie qu'une machine est super-intelligente ou plus intelligente qu'un humain, bien que ce type d'IA ne soit pas actuellement développé. Elle fait référence à un niveau intellectuel supérieur à celui des humains.

Il est important de voir plus en détail chacun de ces types.

1.2.5.1 L'Intelligence Artificielle étroite (ou faible)

1.2.5.1.1 Quelques explications…

L'Intelligence Artificielle étroite est conçue pour résoudre un problème donné, spécifique, particulier. Sa qualification « étroite » ou « faible » fait référence à son champ d'application, dont la programmation est faite pour une tâche prédéfinie.

Elle est donc limitée dans ses fonctionnalités et n'est en aucun cas à la hauteur de l'intelligence humaine. Mais le fait d'être axé ou concentré sur un objectif précis font là sa valeur.

Elle reçoit généralement un entrainement, une formation sur un ensemble de données spécifique, ce qui lui permet de développer une compréhension du problème qu'elle doit résoudre. Une fois ce problème compris, elle peut alors utiliser les connaissances acquises pour prendre des décisions, prédire des résultats ou exécuter des actions.

Il peut s'agir d'une analyse prédictive, d'une tâche de classification ou de reconnaissance d'images et de voix, etc.

Toutes les applications d'Intelligence Artificielle existantes aujourd'hui appartiennent à cette catégorie d'IA, à savoir faible ou étroite.

1.2.5.1.2 Quelques exemples d'IA Etroite (ANI)…

Il existe un nombre très important de secteurs d'application ou de domaines pour ce type d'Intelligence Artificielle, rendant impossible de faire une liste complète d'exemples. Toutefois, il est possible de classer les cas d'usages de la manière suivante :

- **L'IA conversationnelle**

L'Intelligence Artificielle conversationnelle est l'une des technologies d'Intelligence Artificielle étroite les plus répandues à l'heure actuelle. Grâce à l'aide d'un programme informatique, il est possible de faire interagir les deux parties (l'Humain ⇔ IA) via le langage naturel. Cela a d'ailleurs fait émerger une technologie de Natural Language Processing (**NLP**) qui permet aujourd'hui d'améliorer la façon dont les humains et les machines communiquent entre eux.

L'IA conversationnelle peut ainsi répondre à des questions, imiter des conversations humaines, comprendre les sentiments, etc. Cela peut être à travers une conversation vocale ou écrite.

Plusieurs types d'IA conversationnelle existent comme par exemple :

- **Les chatbots** : utilisés principalement pour automatiser des services et le support client → mis en place par les entreprises,
- **Les assistants virtuels** : Alexa, Siri, Google Assistant, Cortana, etc. → utilisation par les particuliers.

Aujourd'hui, l'IA conversationnelle « **ChatGPT** » d'**OpenAI** serait la plus aboutie : présentée le 30 novembre 2022, il est multi-langues et se base sur le modèle **GPT-3.5**, avec un ensemble de données d'entraînement comprenant 300 milliards de mots, datant de 2021 en termes de fraicheur.

Voici un exemple de conversation avec ChatGPT :

Figure 29 : exemple de conversation avec ChatGPT - Source : writesonic.com

Les évolutions dans ce domaine sont phénoménales depuis déjà quelques années comme le montre cet exemple (cliquer pour lire la vidéo), où l'extension « Duplex » de Google Assistant appelait pour prendre un rendez-vous chez un coiffeur : la personne ne se rend alors pas compte qu'elle parlait avec un agent conversationnel.

- **Les systèmes de reconnaissance automatique**

Un autre domaine très exploité est la reconnaissance automatique. Ce domaine connait plusieurs applications tel que :

- La reconnaissance de sons (notamment de la parole) pour par exemple détecter le chant d'une espèce d'oiseau en milieu naturel,
- La reconnaissance de caractère (OCR) pour par exemple classer des documents scannés selon leur contenu,
- La reconnaissance d'image.

C'est ce dernier qui occupe la place la plus important en termes d'utilisation notamment dans le milieu de la santé, notamment pour le diagnostic médical dont voici quelques exemples :

- Aide au diagnostic du cancer par l'IA :

Figure 30 : Aide au diagnostic du cancer par l'IA – Source : radiofrance.fr

- Aide au diagnostic sur base d'images radiologiques :

Figure 31 : Aide au diagnostic sur base d'images radiologiques – Source : radiologie.fr

Enfin, un cas d'usage également important de ce domaine de la reconnaissance automatique est celui de l'authentification des personnes via des techniques de reconnaissance faciale, l'analyse d'empreintes digitales ou de la voir.

Notons que la reconnaissance automatique du texte et la parole jouent aussi un rôle dans l'IA conversationnelle que nous avons vu précédemment.

➢ L'analyse prédictive

L'analyse prédictive est une technique de « **Data Science** ». C'est une méthode d'analyse de données. Elle permet de prédire des résultats futurs en se basant sur des analyses de comportements passés issus d'ensemble de données historiques.

C'est donc une technique qui se base exclusivement sur la « data », et dont la volumétrie nécessaire au fonctionnement est très importante. La convergence avec le Big Data est donc ici essentielle et est poussée au maximum. Sans ces données, il est impossible de modéliser des prédictions utiles et efficaces. Il est cependant important de rappeler que ces prédictions issues de probabilités ne sont que des hypothèses.

L'analyse prédictive est notamment utile dans la maintenance prédictive, la prévision des futures ventes, la gestion des stocks, la gestion du trafic pour un site, etc.

Dans la maintenance prédictive, un cas d'usage connu est celui de **Rolls-Royce**, qui utilise les analyses prédictives à base de « Data Science » et en quasi-temps réel via des « sensors » (IOT) pour détecter d'éventuelles pannes en avance, proposer également des révisions ou passage en atelier de leurs célèbres moteurs (service **TotalCare**) mais aussi pour automatiser l'inspection de pièces et des moteurs ou prédire leurs performances.

Figure 32 : Maintenance prédictive chez Rolls Royce pour les moteurs d'avion - Source : tom.travel

- **Les systèmes de recommandation**

L'analyse prédictive est également utilisée dans les systèmes de recommandation notamment dans le domaine du **Marketing**.

En effet, en utilisant les informations constituant sa base de données « connaissance des comportements client », le système d'IA peut prédire les préférences d'un client. Cela lui permet alors de formuler des recommandations en conséquence.

Ces outils d'Intelligence Artificielle sont particulièrement courants sur les réseaux sociaux, comme par exemple **Facebook**, **Pinterest**, **Instagram** ou même **TikTok**. Mais ils le sont aussi tout particulièrement sur des plateformes d'achat en ligne tel que **Amazon**.

Les achats en magasin ne sont pas délaissés, loin de là et sont même les premiers à les avoir mis en œuvre. Prenons pour exemple la chaine américaine de supermarchés **Target**. Elle a établi un modèle d'analyse au début des années 2000's, lui permettant de détecter l'arrivée d'un enfant pour des parents (en particulier des femmes) et de proposer de la publicité ciblée en conséquence, tout 'simplement' en analysant les variations dans leurs comportement d'achat.

Le statisticien **Andrew POLE** prétendait même : « *We'll be sending you coupons for things you want before you even know you want them* » (« ***Nous allons vous envoyer des bons de réduction pour des objets que vous voulez avant même que vous sachiez que vous les voulez*** »).

Si l'éthique est ici discutable, les chiffres sont pourtant là : **entre 2002 et 2010, le chiffre d'affaires de Target est passé de 44 milliards à 64 milliards de USD soit +52%**. Pour comparer, au même moment, les grandes surfaces françaises sans système de recommandation à cet époque ont augmenté de 16 % environ.

Figure 33 : Moteur de recommandation de Target - Source : forbes.com

Néanmoins, l'affaire éclata au grand jour lorsque le père d'une fille de 17 ans contacta Target pour les informer qu'elle recevait des bons de réduction pas adaptés selon lui car selon lui, sa fille n'est pas enceinte. Alors que le service client le rappelait quelques jours plus tard pour s'excuser, ce père de famille s'excusa de son côté car il apprit finalement que sa fille était enceinte à la suite d'une discussion ayant eu lieu entre temps. Target fut donc au courant avant…

- **Les moteurs de recherche**

Certains moteurs de recherche utilisent également l'Intelligence Artificielle faible. L'IA a alors plusieurs utilités : comprendre le sens des requêtes réalisées par les utilisateurs finaux et classer les résultats en fonction de leur pertinence.

Elle interviendra également pour contrôler la qualité des résultats, empêcher les abus comme la suroptimisation SEO et la détection des spams. Mais surtout, **elle donne une dimension sémantique supérieure aux requêtes notamment vis-à-vis des moteurs sans Intelligence Artificielle**.

Ainsi, Microsoft Bing, Google Chrome et Baidu ont tous annoncé en début année 2023 leur « dopage » à l'IA.

Prenons un exemple : vous souhaitez apprendre à faire un plat et vous cherchez « recette de la blanquette de veau ». Là où un moteur de recherche sans IA ne vous donnera que des liens vers des

recettes, ces moteurs de recherche avec IA donnent la liste des ingrédients, la préparation, le temps de cuisson, rédigé et expliqué parfaitement.

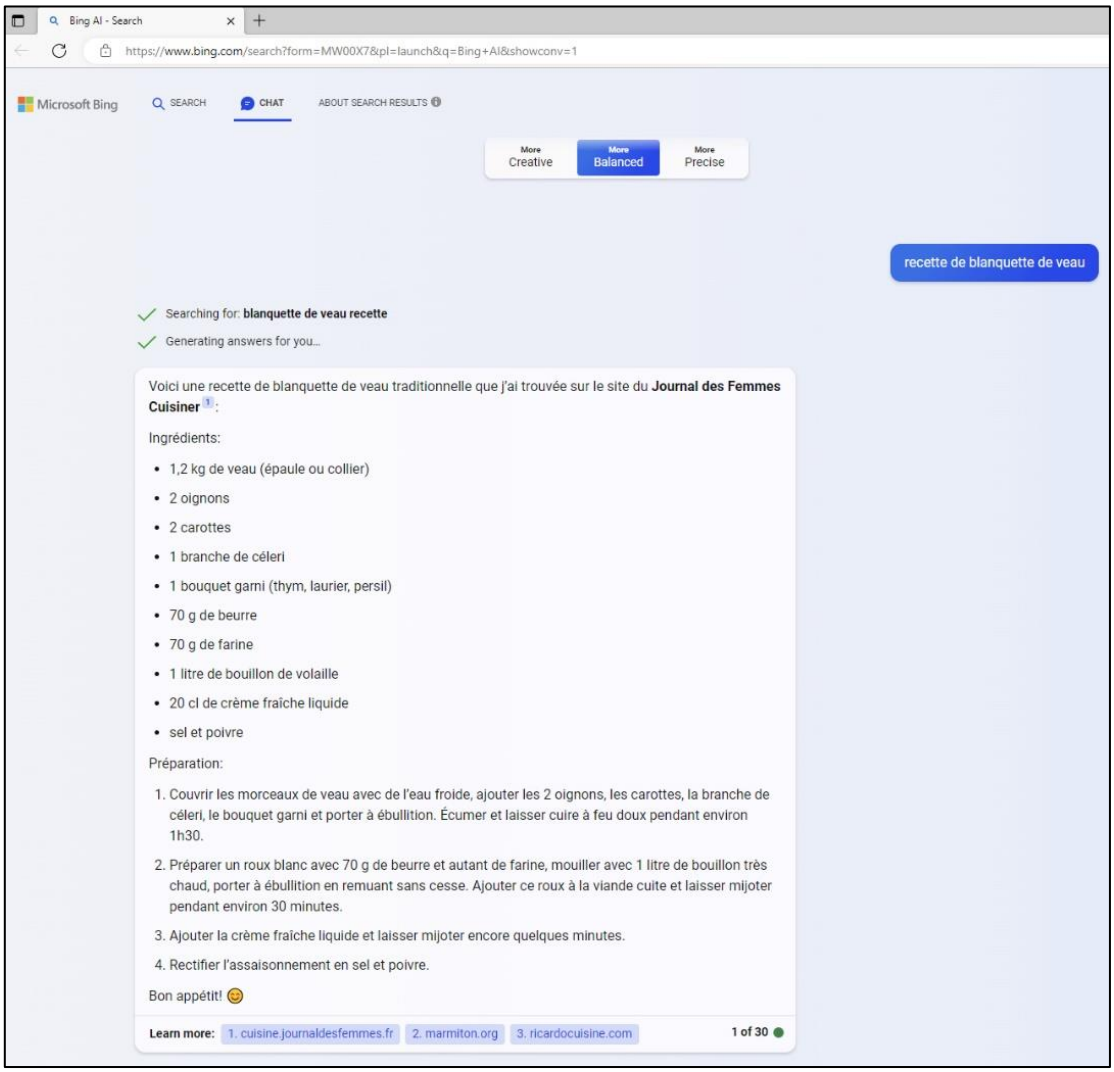


Figure 34 : Recherche d'une recette de blanquette de veau avec Bing & ChatGPT

Félicitations ! Vous avez maintenant une recette de blanquette de veau !

- **Les machines autonomes**

Il s'agit des dispositifs dont le fonctionnement se fait sans l'intervention humaine. Dans ce domaines, nous pouvons retenir :

- Les robots
- Les véhicules autonomes

Ce genre de dispositifs sont de plus en plus présents dans le monde technologique actuel et reposent très souvent sur l'Intelligence Artificielle étroite pour réaliser leurs actions sans supervision de l'être humain.

Pour être autonome, ces éléments nécessitent d'être dotés d'éléments leur permettant une perception de l'environnement qui les entoure, afin de réagir en conséquence. Il est surtout question de reproduire la vision humaine via la technologie « computer vision ».

Grâce à eux, les robots utilisent des capteurs et des caméras pour percevoir leur environnement et les objets qui s'y trouvent. De ce fait, ils peuvent éviter les obstacles, saisir des objets et interagir avec eux.

Par exemple, les aspirateurs-robots :

Figure 35 : Les aspirateurs-robots utilisent l'IA pour le fonctionnement autonome – Source : yankodesign.com

C'est également le cas pour les voitures autonomes :

Figure 36 : L'IA "Autopilot" est au cœur des voitures autonomes Tesla - Source : keley.com

1.2.5.1.3 Quelques avantages et limitations de l'IA étroite

Pour réaliser une tâche bien spécifique, **le système d'IA étroite est optimisé pour l'effectuer avec plus de rapidité et de précision qu'un être humain**. Nous pourrions donc dire qu'il peut être meilleur que les êtres humains mais cela jusqu'à un certain point.

Les capacités de type d'Intelligence Artificielle sont particulièrement utiles dans des tâches nécessitant des prises de décisions rapides et très poussées dans un domaine particulier.

Mais cette force est aussi la faible de ce type de système d'IA : En effet, ils ne peuvent pas résoudre d'autres problèmes. Par exemple, un modèle d'IA entraîné à reconnaître un visage ne peut pas faire de la reconnaissance de texte.

En théorie, si l'IA parvenait à se généraliser et donc à devenir une Intelligence Artificielle Générale (ou « AGI »), il serait probable de repousser cette limite. Néanmoins, pour l'instant, les « IA Engineer » doivent se contenter de trouver des moyens immédiats pour relever ces défis.

1.2.5.2 L'Intelligence Artificielle Générale (ou forte)

L'IA générale n'existe pas à l'heure actuelle : c'est une forme théorique qui serait capable de réaliser des tâches intellectuelles comparables à celles effectuées par les êtres humains. Ce serait un système par définition totalement cognitif, capable d'appliquer son intelligence à n'importe quel problème ou à plusieurs tâches pour les résoudre, contrairement à l'Intelligence Artificielle faible spécialisée dans une tâche mais un être humain le pourrait.

Il est clair qu'**une telle forme d'IA n'existe pas encore**. Aucun système actuellement qualifié d'intelligent n'a atteint ce niveau de cognition, surtout compte tenu du potentiel d'amélioration de l'IA faible.

Sa création, si possible un jour, aura des impacts majeurs sur la vie humaine, notamment le monde économique. Il ne serait plus question d'avoir une IA qui apporte une aide à l'humain dans la réalisation de ses actions ou de ses prises de décision. Les experts imaginent ici des IAs capables de remplacer des humains dans des emplois hautement qualifiés tel qu'un ingénieur, un médecin, etc.

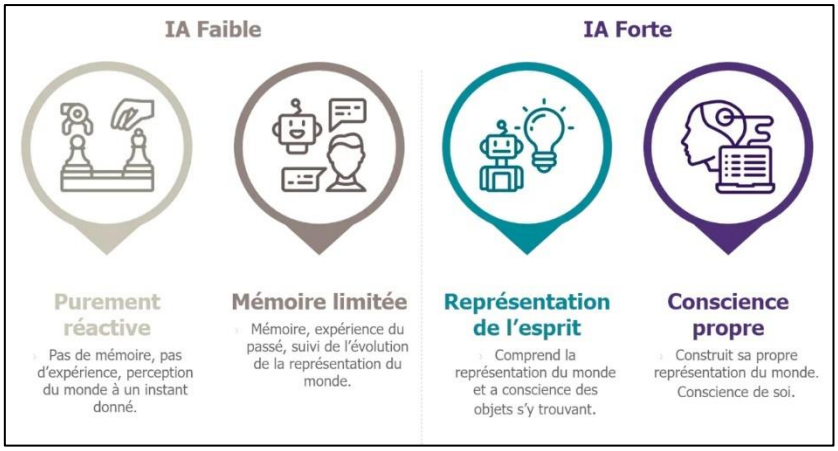


Figure 37 : Différences entre IA Faible et IA Forte – Source : digitalcorner-wavestone.com

Mais pour que cela puisse être possible, l'IA générale être en mesure de réussir plusieurs tests :

- **Le fameux test de Turing**, qui permet de vérifier la capacité d'une IA à agir comme un humain.
- **Le test du café**, qui consiste à laisser l'IA apprendre à faire le café seule dans un environnement domestique après l'avoir trouvé
- **Le test du Robot College**, qui veut que le robot s'inscrive à des cours et les suive avec succès
- Et potentiellement **passer un entretien d'embauche**, réussir des tests professionnels, des examens de conduite, etc.

L'Intelligence Artificielle forte n'est donc pas encore réelle.

1.2.5.2.1 Un esprit

L'Intelligence Artificielle générale ou forte se base sur la théorie de l'esprit, et donc qui aurait la capacité de discerner les émotions, les croyances, les processus de pensée et les besoins d'autres personnes intelligentes. L'Intelligence Artificielle basée sur la théorie de l'esprit ne consiste pas à simuler ou à reproduite mais devrait avoir un ensemble complet de capacité cognitives et devrait comprendre réellement les humains sur tous les aspects dont les émotions.

Il s'agit d'un immense défi notamment à la vue de la complexité du cerveau humain et le modèle de création d'une intelligence naturelle d'une manière générale. Or aujourd'hui, les chercheurs manquent déjà de connaissances approfondies sur le fonctionnement du cerveau humain, ce qui limite à l'heure actuelle les capacités à reproduire les fonctions de base de l'être humain, à savoir la vue, la parole et le mouvement.

Et combien même les connaissances seraient là, la limitation se trouve aussi au niveau des capacités des machines en termes de ressources. Il a fallu à « K », l'un des superordinateurs les plus rapides au monde et construit par Fujitsu, 40 minutes pour simuler une seule seconde d'activité neuronale de l'être humain. Dans ces conditions, **reproduire un « esprit humain » relève aujourd'hui du quasi-impossible et nul n'est capable de déterminer ou prédire si dans un avenir prévisible, une Intelligence Artificielle forte sera atteinte un jour**.

Figure 38 : IA forte & la théorie de l'esprit - Source : bbc.com

1.2.5.2.2 Un corps

Selon une autre théorie, l'esprit et le corps serait inséparable. Or, qui dit corps dit représentation des 5 sens de l'être humain qui lui permettent de percevoir son environnement et d'interagir avec lui. Pour une machine, la vue, le son, le toucher sont des valeurs numériques ce qui fait qu'il est très difficile de percevoir le monde de la même façon que l'être humain.

S'il existe aujourd'hui des avancées notamment médicales pour permettre de revoir à nouveau à un pourcentage très minime par rapport à la normale, ou même simuler l'audition, etc., il y a encore énormément de progrès à faire dans ces domaines pour « digitaliser » ces sens.

Grâce à la technologie de traitement du langage naturel (NLP), nous avons vu qu'il était possible de comprendre le langage humain. L'IA peut comprendre et interagir, voir même comprendre le sens des mots relatifs aux sensations ressenties par l'être humain grâce à son corps. Mais cela n'est que sémantique : pas de notion véritable d'émotion comme le chaud ou le froid. De plus, cela reste limité par rapport aux données d'apprentissage de l'IA. Il lui faudra des milliards de données sur tout ce qui est relatif au corps, **tant le corps humain est complexe**.

Ainsi, selon des chercheurs et expert en IA, elle devra disposer un équivalent en « assertion » (traduction en code interprétable par un ordinateur de certitudes ou d'éléments considéré comme vrai) de 2 siècles de travail environ. Impossible donc à l'heure actuelle…

1.2.5.2.3 Mais l'IA générale sera-t-elle possible un jour ?

Pour certains, elle ne sera jamais possible. Pour d'autres, il s'agit d'un objectif très difficile à atteindre mais pas irréalisable un jour. Nous sentons bien que cela dépend de chaque de point de vue.

En effet, si l'IA forte doit être une reproduction de l'intelligence humaine (naturelle) dans son intégralité, cela parait impossible à l'heure actuelle.

Néanmoins, s'il s'agit de construire un système d'IA capable de réaliser l'ensemble des tâches qu'un être humain peut accomplir, alors cela parait plus réaliste.

Si nous revenons à la définition de l'IA de **RUSSEL & NORVIG**, nous avons vu que ces deux chercheurs distinguent deux approches : **l'approche humaine et l'approche rationnelle (idéale). La première approche finalement correspond à notre premier point de vue sur l'IA forte, à savoir se comporter entièrement comme l'humain et cela parait donc impossible. La deuxième est celle correspondant à la simulation des comportements selon ce que l'IA connait. Nous avions vu que cette approche se base sur les mathématiques et l'ingénierie, soit une approche scientifique**.

Or, la science ne cesse de progresser et les avancées de la science d'aujourd'hui ne permettent pas de présager de celles du futur ! Rappelez-vous, nous sommes en tant qu'humain dans l'incertitude du futur.

Par conséquent, **il nous est donc impossible d'affirmer que les machines finiront ou non par avoir un raisonnement de même niveau que celui des êtres humains.**

1.2.5.2.4 Focus sur le test de Turing

Ce test a été imaginé par le mathématicien **Alan TURING** en 1950, dans un article dont le titre était « *Computing Machinery and Intelligence* ». **Il a pour but de mesurer la capacité de l'Intelligence Artificielle à être confondue avec l'intelligence humaine**, et cela dans le but de répondre à la fameuse question « **une machine peut-elle penser ?** ».

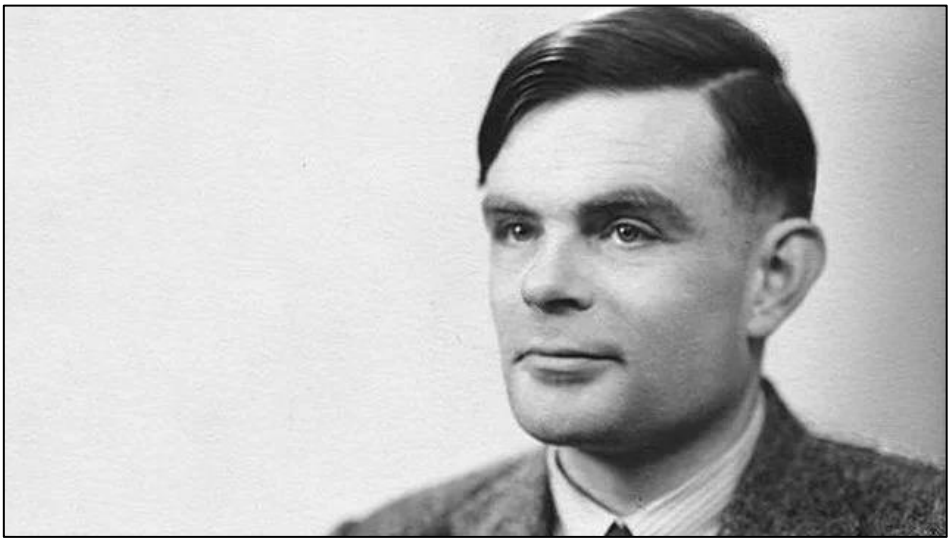


Figure 39 : Photo d'Alan TURING - Source : wallonica.org

Le mode opératoire est le suivant : **un humain ayant le rôle d'évaluateur va dialoguer avec deux interlocuteurs distincts, à l'aveugle sans savoir lequel est l'humain ou la machine**. Il sait cependant que l'un des deux est une machine. Après 5 minutes de conversation, l'évaluateur doit réussir à déterminer lequel est l'IA parmi ses deux interlocuteurs.

Pour la machine, il s'agit de tromper l'évaluateur en répondant le plus « humainement » possible sans se soucier de répondre correctement aux différentes questions posées. L'évaluateur doit donc distinguer les deux interlocuteurs : s'il se trompe, alors la machine a réussi le test.

Seule une poignée très réduite de programmes d'Intelligence Artificielle a été capable de réussir ce test. Voici les tests les plus connus :

- **En 1966 avec ELIZA** : considéré comme le premier à avoir réussi le test de Turing, a réussi à convaincre plusieurs interlocuteurs qu'il était une véritable personne en répondant de manière cohérente après avoir analyser et rechercher les mots clés nécessaires aux réponses.
- **En 1972 avec PARRY** : évalué par un groupe de psychiatre, le programme imitait un comportement schizophrène paranoïaque. Dans 52 % des cas, les psychiatres croyait avoir un dialogue avec un humain
- **En 2014 avec le chatbot Eugene Goostman (premier chatbot)** : simulant un jeune de 13 ans ne maitrisant pas totalement l'Anglais, il parvient à convaincre 33 % des juges (au nombre de 30) en 5 minutes, sur des sujets illimités.

A l'heure actuelle, il est admis que de nombreux systèmes d'IA seraient en mesure de réussir le test imaginé par Alan Turing. Cependant, il est important de rappeler qu'il n'existe aujourd'hui que des IA étroites, dites faibles, et donc spécialisées dans un domaine précis.

Devant l'évolution récente des système d'IA, le test de Turing dans son format original n'est plus appliqué : d'ailleurs, en 2019, le prix **Loebner** qui récompensait la réussite à ce test a été arrêté.

D'autres tests pourraient voir le jour, ou ont déjà vu le jour, comme celui imaginé par Google notamment pour tester l'agent conversationnelle **Meena** qu'ils ont développé en 2020.

Le test de Turing n'est plus utilisé sur les système d'IA mais persiste sous une autre forme, celles des **CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart)**, présent dans de nombreuses pages Web afin de vérifier que l'individu parcourant le web n'est pas un robot spammeur.

1.2.5.3 La Super-Intelligence Artificielle

Si l'Intelligence Artificielle forte ou générale est celle qui permettrait de simuler l'intelligence humaine ou son comportement, la **Super-Intelligence est l'IA imaginée dans l'hypothèse qu'elle dépasserait les aptitudes humaines en tout point**.

C'est l'Intelligence Artificielle de la science-fiction : dotée de ressources machines supérieures en termes de puissance et de vitesse de calcul, **elle surpasserait l'humain**. Cette IA serait en mesure d'avoir une conscience, un esprit, des émotions et un corps (forme robot) mais aussi de les discerner chez les autres. Elle aurait conscience d'elle-même et de son écosystème, aurait un esprit réel.

Elle serait meilleure dans tout ce que peut faire l'être humain : sciences (mathématiques, médecine, etc.), sports, art et même au niveau des relations émotionnelles. Ses capacités seraient bien meilleures notamment en termes de résolution de problèmes ou de prise de décision.

Cependant, tout comme dans la science-fiction, la Super-Intelligence Artificielle suscite de nombreuses interrogations, de débat et de craintes notamment vis-à-vis des conséquences sur la vie humaine.

Par exemple, si de tels être super-intelligents voyaient le jour, seraient-ils amenés à développer des idées comme celles de l'auto-préservation ? Il est clair que cela nous ramène aux fictions et aux spéculations, alimentant les craintes notamment vis-à-vis des impacts sur l'humanité, son mode de vie et sa survie.

Pour l'instant, **cette Super-Intelligence Artificielle existent donc que dans les romans et au cinéma** 🙂

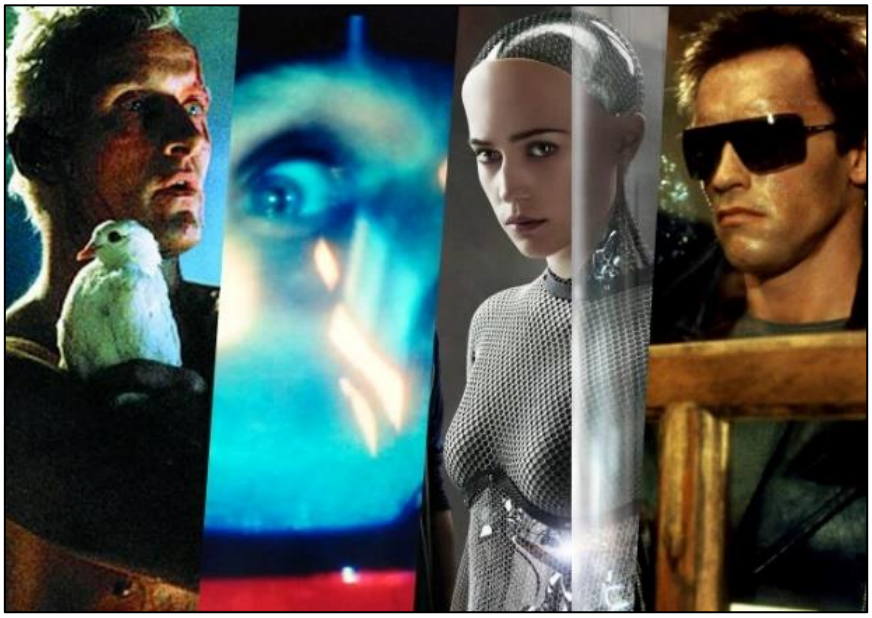


Figure 40 : Exemples de films abordant l'IA - Source : indiewire.com

1.2.6 Les approches de l'IA

Si l'Intelligence Artificielle repose sur des techniques de développement informatique capable de simuler l'intelligence humaine, notamment ses capacités de prise de décision, elle constitue une des 6 branches des **sciences cognitives**. Qu'est que les sciences cognitives ? Une définition :

> *Sciences cognitives : sciences qui ont pour objet de décrire, d'expliquer, voire de stimuler les processus de la connaissance.*

Dictionnaire Larousse - 2023

Ainsi, en termes de principe, pour les sciences cognitives, tout mécanisme de pensée peut être modélisé par le biais d'un système symbolique de traitement de l'information.

Au fur et à mesure du temps et de l'avancée de leurs travaux, les chercheurs ont donc défini différentes approches :

- **L'approche symbolique** : historique, basée sur le principe de système symbolique des sciences cognitives. Cette approche suppose que l'Intelligence Artificielle s'obtient grâce à la manipulation de symboles, plutôt qu'en traitant des données comme le fait l'IA moderne.

- **L'approche par apprentissage automatique** : cette approche **part de l'hypothèse** que la manipulation de données peut produire de l'Intelligence Artificielle. C'est l'approche récente.

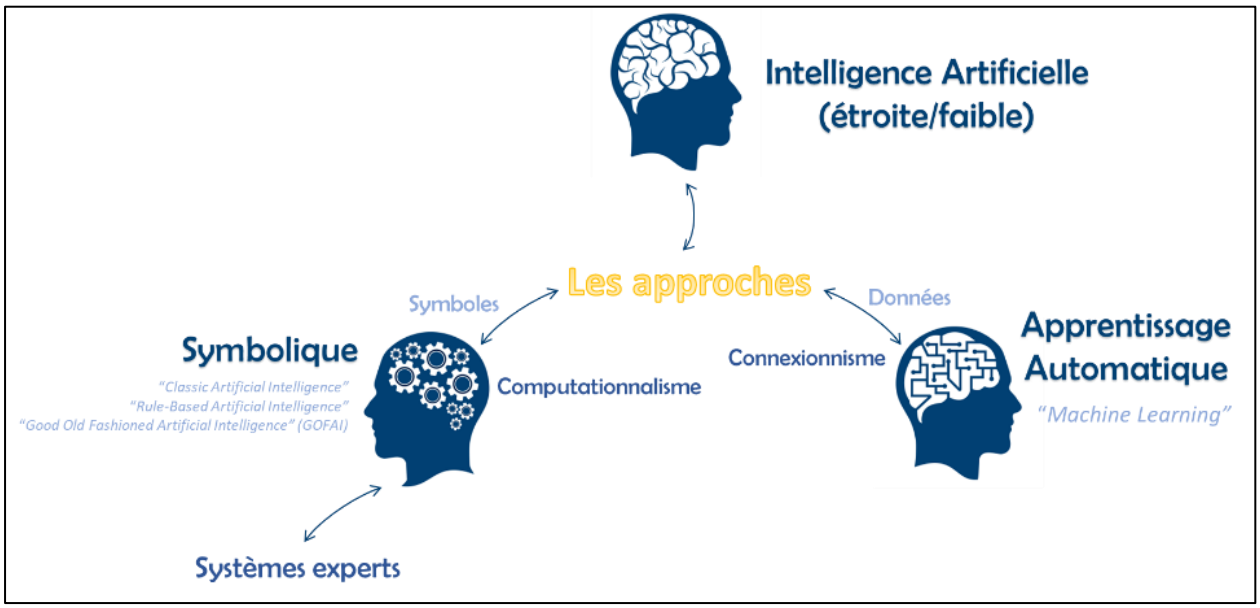


Figure 41 : Les approches de l'IA - Source : autrice

Deux mouvements de pensée des sciences cognitives peuvent être retrouvés derrière ces 2 approches de l'IA :

- **Le computationnalisme** ou **cognitivisme** : Grâce à des manipulations symboliques, des règles et des calculs peuvent être assimiler à des processus mentaux. Ce courant se trouve plutôt derrière l'approche symbolique.

- **Le connexionnisme** : les processus mentaux émergent de réseaux interconnectées d'unités simples (souvent des réseaux neuronaux). On retrouve ce courant dans l'approche par apprentissage automatique.

1.2.6.1 L'approche symbolique

Cette approche, également appelé « **approche cognitiviste »**, est celle historique. C'est celle évoquée par **Alan TURING** dans son fameux article « *Computing machinery and intelligence* » en 1950 qui étudie la possibilité de simuler une Intelligence via la manipulation de symboles à travers un ordinateur.

Un peu plus tard, en 1956 **John MCCARTHY**, **Marvin MINSKY** et d'autres scientifiques se réunissent sur le campus de Dartmouth. Au cours de leurs ateliers, ils abordent ensemble les premiers systèmes experts utilisant les méthodes symboliques, ainsi que des avantages des systèmes utilisant le raisonnement déductif ou inductif. Ces ateliers sont aujourd'hui considérés fondateurs de l'Intelligence Artificielle comme discipline académique.

Cette approche a plusieurs noms :

- « *Classic Artificial Intelligence* »
- « *Rule-Based Artificial Intelligence* »
- « *Good Old Fashioned Artificial Intelligence* » *(GOFAI)*

Les systèmes experts ont pour but de simuler les capacités d'un expert, dans un domaine donné. En utilisant un **moteur d'inférence** ainsi qu'**une base de faits et de règles**, le système choisit les règles selon les faits donnés en entrées et produit de nouveaux faits en sortie.

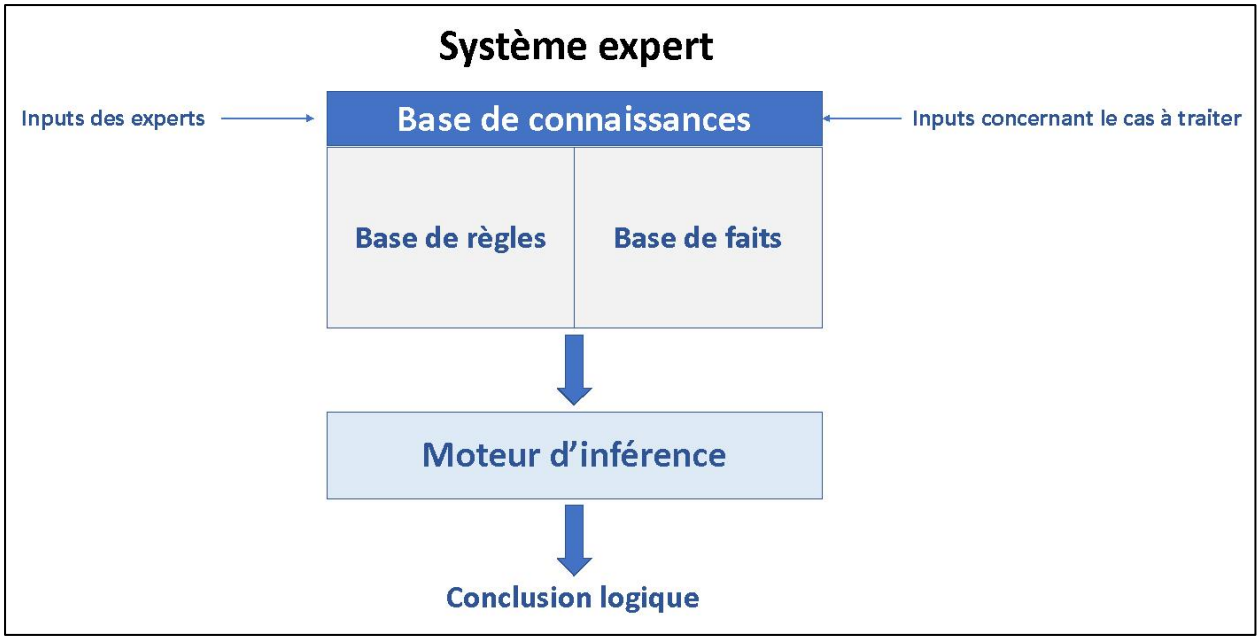


Figure 42 : Fonctionnement d'un système expert - Source : quantmetry.com

Voici un exemple basique de système expert :

Base de faits :

Nathalie est la sœur de Pierre

Pierre est le frère de Paul

Caroline est la mère de Nathalie

Base de règles :

si x est le frère ou la sœur de y alors la mère de x est aussi la mère de y

Déductions possibles du moteur d'inférence :

Caroline est la mère de Pierre

Caroline est la mère de Nathalie

Ces systèmes sont plus efficaces que l'utilisation de syntaxe « *if <règle> then <fait>* » lorsque dans des situations réelles, le système expert a pour bases un très grand nombres de règles et de faits. Lors de l'exécution, le moteur d'inférence prend toute son importance : son rôle est donc de détecter les règles vérifiées et d'établir la chaine de raisonnement optimisée permettant d'aboutir au fait en sortie, qui est la conclusion logique du raisonnement.

L'approche symbolique est particulièrement efficace pour les problèmes de **satisfaction de contrainte**, à savoir des solutions dont le contexte possède un nombre limité de critères ou de paramètres (règles/faits). Par exemple : cette IA serait efficace pour résoudre des grilles du jeu de Sudoku.

En revanche, l'approche symbolique est fortement plus limitée dans un contexte généralisé ou en dehors de cadre défini.

Les systèmes experts suivent la conception du **computationnalisme**, qui affirme que toute intelligence peut s'expliquer par des calculs descriptibles par des algorithmes. Cette thèse du computationnalisme a été vivement critiquée, notamment par le philosophe **Hubert DREYFUS** ce qui a permis l'émergence de nouvelles approches, notamment celle « **connexionniste** ».

1.2.6.2 L'approche par apprentissage automatique : le Machine Learning

Cette approche suit le courant de pensée du **connexionnisme**, vise à simuler l'intelligence en se basant sur les données. Les **probabilités** et la **Data Science**, le **Data Mining** et le **Big Data** sont les sources pour faire ici de l'apprentissage automatique, et plus particulièrement du **Machine Learning** (ou **ML**).

La Data Science et le Data Mining ont pour but d'établir des modèles issues d'observation d'un volume très important de données, et cela afin d'expliquer des phénomènes et de donner des probabilités d'occurrence d'évènements dans le futur. Ils permettent tous deux de comprendre les « data », ainsi que les relations entre elles (corrélation).

1.2.6.3 Les modes d'apprentissage

A l'heure actuelle, on compte 3 principaux modes d'apprentissages :

- **Apprentissage supervisé :**

Le système d'IA est **entrainé sur un ensemble de données étiquetées par un humain**. Ces données sont validées par l'être humain. Un modèle est obtenu par rapport à des vérités observés sur des valeurs de références dans le passé. Le modèle permet alors lors de la soumission de données en entrée de **réaliser des prédictions ou des classements**.

L'avantage principal de ce type d'apprentissage est que **l'être humain est en mesure de contrôler les prises de décisions du système d'IA**. Mais son principal inconvénient réside dans **la lenteur et les coûts d'apprentissage**, notamment par la nécessité que l'humain intervienne (pour l'étiquetage par exemple).

- **Apprentissage non-supervisé :**

Ici, les données ne sont pas validées ou étiquetées par un humain. **Le système d'IA analyse les données et utilise des fonctions de coûts pour calculer son éloignement par rapport au résultat**, le plus petit coût indiquant la probabilité que le système d'IA soit proche du résultat. **Il est libre dans sa prise de décision**.

Aucune « vérité » ou « valeur de référence » n'est connue et n'est recherchée. Ce mode est particulièrement utilisé pour faire de la segmentation.

L'avantage principal de ce type d'apprentissage se trouve dans **la rapidité d'apprentissage** du système d'IA qui ne nécessite pas d'intervention humaine. **Mais son principal inconvénient réside dans l'explicabilité de la prise de décision et le risque d'erreur non contrôlé**.

- **Apprentissage par renforcement :**

Là aussi, il n'y a pas de données étiquetées ou validées par un humain. **Le système d'IA s'améliore (est renforcé) en cas de réponse positive ou sanctionné en cas de réponse négative**. Cela permet au système d'IA d'apprendre de la même manière d'un être humain apprend au fil du temps de ses erreurs. Il apprend les actions à effectuer à partir de son expérience.

Par exemple, si la fonction de coût utilisée consiste à optimiser un nombre de points, toute mauvaise action pénalisera le modèle en retirant des points, et toute bonne action engendrera l'ajout de point. **L'algorithme apprend alors très vite à favoriser les bonnes actions et éviter les mauvaises**. Néanmoins, cela nécessite de définir des règles pour récompenser et pénaliser.

Ce mode est particulièrement utilisé dans les domaines des jeux et de la robotique.

L'avantage principal de ce type d'apprentissage se trouve dans **la rapidité d'apprentissage** du système d'IA qui ne nécessite pas d'intervention humaine. **Mais son principal inconvénient réside dans la nécessité de définir des règles d'apprentissage basée sur la récompense ou la sanction**.

Résumons les 3 modes de l'apprentissage automatique grâce à cette image, à travers laquelle nous pouvons voir les cas d'usages associés :

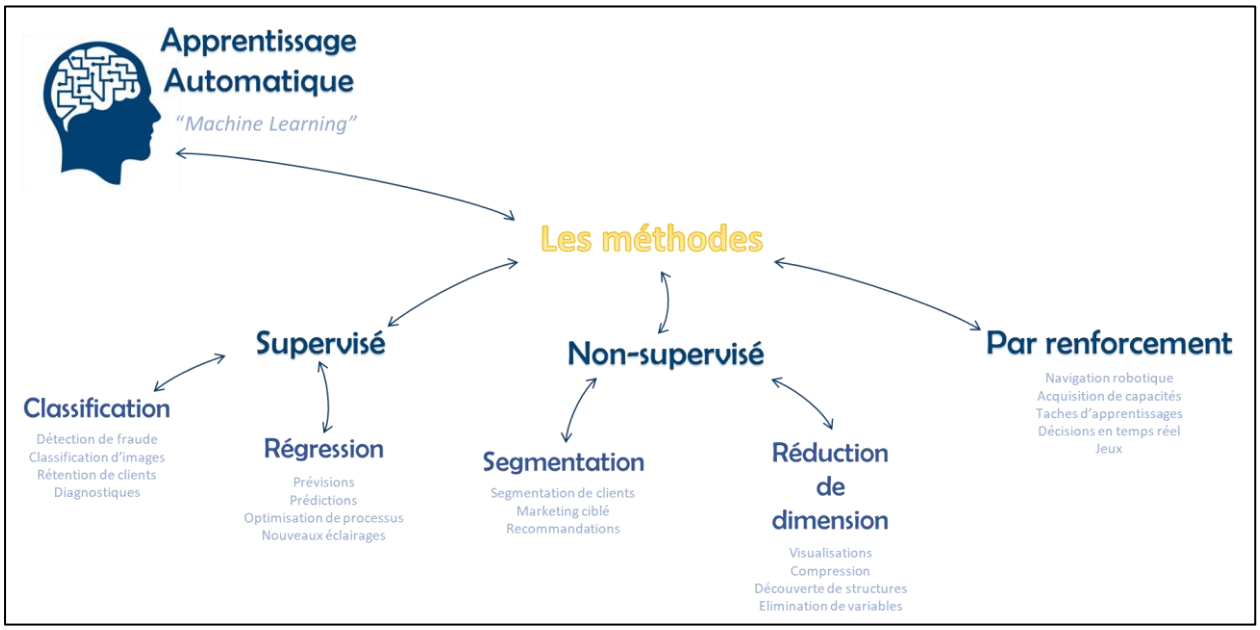


Figure 43 : Les 3 modes d'apprentissage automatique - Source : autrice

1.2.6.4 Les principaux modèles

1.2.6.4.1 Les principes généraux de construction d'un modèle

Dans la pratique, l'apprentissage d'un programme d'Intelligence Artificielle repose sur des principes mathématiques, plus exactement les probabilités et les statistiques pour établir des modèles, et sur des principes de l'ingénierie informatique pour réaliser l'implémentation d'algorithmes robustes et efficaces.

C'est une fois que l'implémentation de l'algorithme est réalisée que la machine peut apprendre, prédire des événements précis. Il peut également s'enrichir au fur et à mesure, lors de la réception de nouvelles « data ».

Les modèles en question reposent sur des équations mathématiques, donnant aux systèmes d'IA une compréhension de base des concepts appris par les êtres humains. Plus exactement, les concepts mathématiques utilisés dans ces modèles sont principalement les **corrélations** de variables et leur **régression**.

Une **corrélation** est l'observation de la dépendance de deux variables. Il y a co-relation quand deux variables évoluent ensemble. Si l'une augmente et que l'autre augmente aussi, nous parlons de **corrélation positive**. Si à l'inverse, l'une augmente et que l'autre diminue, alors nous parlons de **corrélation négative**.

La **régression** permet d'approcher une variable (dite expliquée) à partir d'autres qui lui sont corrélées (dite explicatives), dans le but de prédire la valeur de la variable « expliquée » à partir des autres variables.

Ces deux principes mis ensemble seule ne suffisent pas. Il faut donc les confronter à des données réelles et représentatives du cas d'usage. **Le modèle mathématique est alors établi en déterminant quelle équation mathématique permet d'obtenir une « distance » la plus petite possible par rapport aux données réelles**.

Pour l'expliquer, il est préférable de prendre un exemple, via un cas très simple dont beaucoup d'entre nous se souviennent sans doute, celui de la « droite d'ajustement » apprise à l'école. Supposons un nuage de points. La droite d'ajustement est justement obtenue grâce à cette équation qui permet de représenter au mieux ce que les données montrent. Nous pourrions illustrer la droite comme suit :

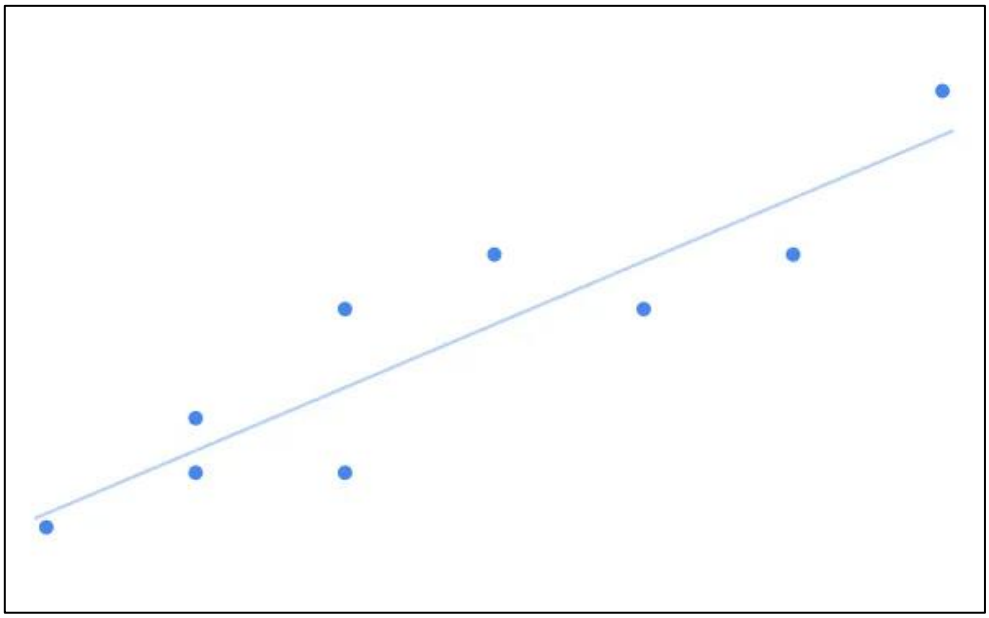


Figure 44 : Exemple de droite d'ajustement en régression linéaire – Source : Blog Aboutgoods Company

Imaginons que les données représentent la taille des personnes selon les pointures des chaussures et qu'il y a **corrélation positive** entre la pointure et la taille. En effet, généralement, plus une personne est grande, plus elle a de grand pieds :

- En bas à gauche, la personne la plus petite a de petite chaussure
- En haut à droite, la personne la plus grande a de grande chaussure.

Dans l'absolu, il existe une infinité de droite. Mais pour obtenir la droite d'ajustement la plus représentative de nos données, il s'agit de calculer la distance entre chaque point et la droite, mettre cette distance au carré pour n'avoir que des valeurs positives, puis en faire la somme.

La droite d'ajustement sera celle dont la somme des carrés des distances sera la plus petite par rapport aux différents points des données. C'est pour cela que cette droite est appelée la « **droite des moindres carrés** ». La minimisation de la somme des distances s'appelle une **fonction de coût** ou selon les cas une **fonction de perte**.

La droite d'ajustement est ici la représentation graphique de l'équation mathématique qui permet de prédire la taille du propriétaire grâce à sa pointure :

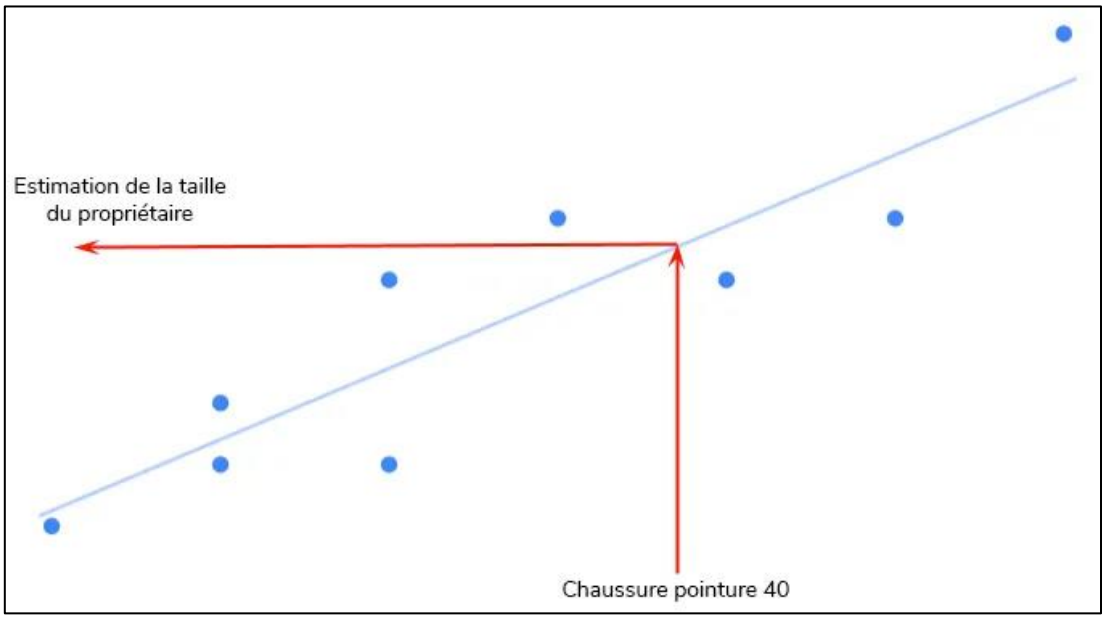


Figure 45 : Faire une estimation grâce à la droite d'ajustement – source : blog Aboutgoods Company

Ce principe est très important : le modèle est calculé en mesurant des distances. Cela ne change pas même pour les modèles plus sophistiqués. Tous les modèles mathématiques utilisés en IA sont établis sur les principes de fonction de coût et sur un calcul de distance. Ainsi, la régression, la corrélation et la fonction de coût permettent alors d'obtenir une équation mathématique de prédiction, simplification de la réalité.

1.2.6.4.2 Les performances d'un modèle

Le travail d'un modeleur consiste à choisir le ou les bons modèles selon le cas d'usage pour lequel il est souhaité de créer un système d'IA.

Néanmoins, les performances d'un modèle ou de ces modèles est influencé par plusieurs éléments qui contraignent « l'espace de travail » du modeleur :

- **Le degré d'apprentissage**
- **Le compromis précision-rappel**

Tout d'abord, intéressons-nous au **degré d'apprentissage**. Une des responsabilités du modeleur est d'éviter le **sous-apprentissage** et le **sur-apprentissage**.

En effet, si le ou les modèles utilisés n'ont pas été suffisamment entrainés, l'équation mathématique obtenue n'est pas représentative des données réelles et par conséquent, ne permettra en aucun cas de prédire quoi que ce soit ou bien fera beaucoup d'erreur.

Afin de mieux expliquer ces éléments, prenons un exemple avec des données réelles qui suivent une courbe sinusoïdale :

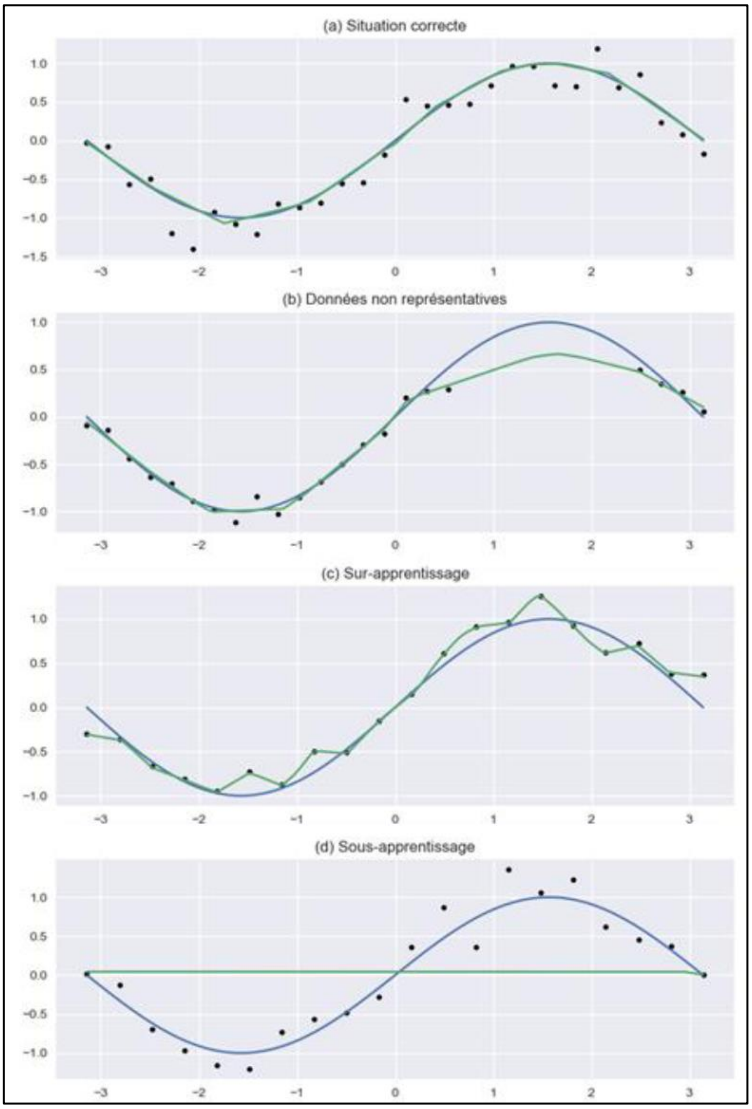


Figure 46 : Cas de figure selon degré d'apprentissage d'un modèle – source : [eduscol.education.fr]

Avec un **bon degré d'apprentissage**, la situation correcte est atteinte (cas « **a** » sur le schéma). La courbe sinusoïdale obtenue en fonction des données de test (en vert) épouse presque parfaitement la distribution des cas réels (en bleu). Il s'agira probablement du meilleur modèle.

Avec des **données non représentatives de tous les cas de la réalité** ou **peu diversifiées** (cas « **b** » sur le schéma), la courbe sinusoïdale obtenue en fonction des données de test (en vert) aura une « distance » importante avec les cas réels (en bleu).

Si en revanche, le modèle est entrainé plus qu'il ne faut (cas « **c** » sur le schéma), la courbe sinusoïdale obtenue en fonction des données de test (en vert) suit la distribution des cas du jeux de données et non pas une distribution mathématique des cas réels (en bleu). On dit que le modèle ne généralise pas. Lorsqu'un tel événement se produit, le modèle prédictif pourra donner de très bonnes prédictions sur les données d'apprentissage (les données déjà « vues »), mais il prédira mal s'il rencontre des données pas vues lors de sa phase d'apprentissage. Le mieux est donc ici l'ennemi du bien ! C'est le **sur-apprentissage** !

Enfin, sans entrainement ou avec peu d'entrainement (cas « **d** » sur le schéma), nous obtenons une droite en fonction des données de test (en vert) qui ne représente rien car toujours la même valeur. Elle n'est pas ou pas suffisamment représentative de la réalité car met une grande partie des nuances de la distribution des cas de tests et des cas réels. C'est le **sous-apprentissage** !

Autre contrainte pour le modeleur : dans le cas d'une modélisation d'une **classification**, il devra faire un **compromis** entre deux valeurs qui sont la **précision** et le **rappel**.

Idéalement, la prédiction du modèle sera conforme à la réalité. Si la prédiction est conforme à la réalité, tout va bien. Nous avons alors des vrais positifs ou vrai négatifs.

Malheureusement, nul modèle n'est parfait et le modèle va certainement commettre des erreurs selon deux cas de figure :

- **Cas 1** : Si la vérité est négative et la prédiction du modèle est positive, alors il s'agit d'un faux positif : nous l'appellerons "erreur de type 1".
- **Cas 2** : Si la vérité est positive et la prédiction négative, alors il s'agit d'un faux négatif : nous l'appellerons "erreur de type 2".

La précision est la proportion du nombre de vrai positif (ou « True Positive » - TP) sur le nombre de prédiction positif (« True Positive » - TP & « False Positive » - FP) :

$$\text{précision} = \frac{TP}{(\text{TP} + \text{FP})}$$

Le rappel est la proportion du nombre de vrai positif (ou « True Positive » - TP) sur le nombre de réel positif (« **True Positive** » - TP & « **False Negative** » - FN) :

$$\text{rappel} = \frac{TP}{(\text{TP} + \text{FN})}$$

Voici ici un tableau qui donne les différents cas de prédictions et leur catégorisation selon la réalité :

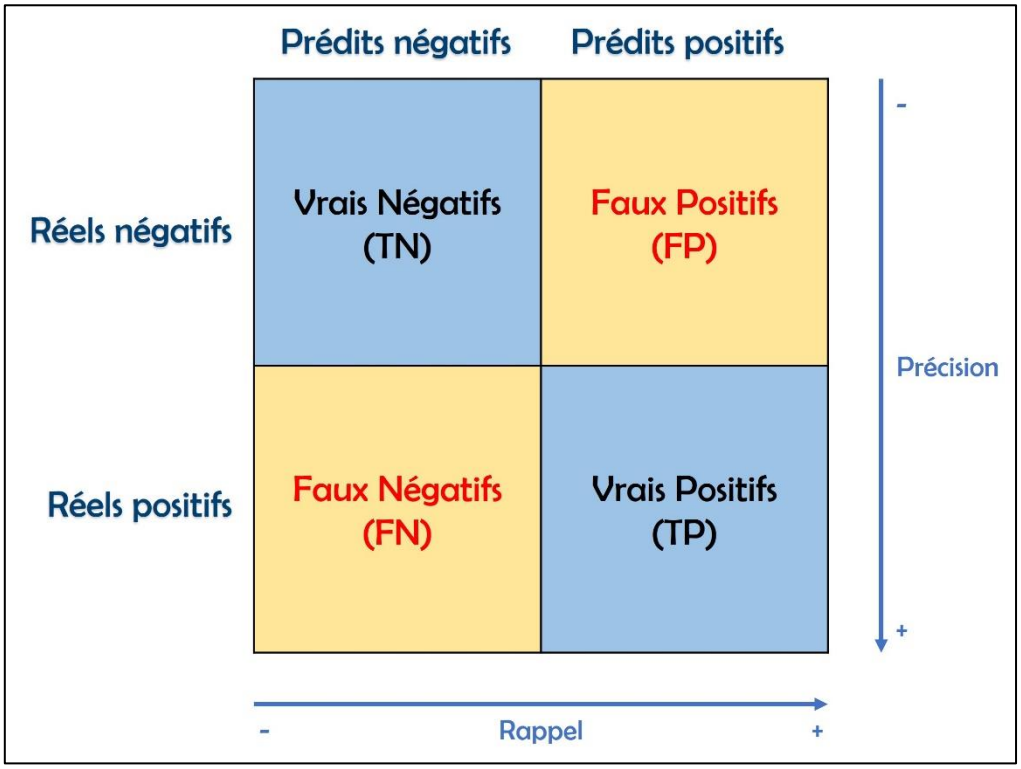


Figure 47 : Précision et rappel selon les prédictions et la réalité – Source : autrice

Malheureusement, ces deux valeurs sont antinomiques. En effet, soit le modèle a une très bonne précision et prédit toujours positif quand la vérité est positive : néanmoins, certains vrais positifs pourraient être en faux négatif.

Soit le modèle a un très bon rappel : tous les positifs ont été identifiés au prix d'un certain nombre de faux positifs donc d'une précision moindre.

Le modeleur devra donc faire un choix selon le cas d'usage selon le poids donné aux erreurs de type 1 (faux positif) vs aux erreurs de type 2 (faux négatif). En effet, ces erreurs n'ont pas toujours les mêmes conséquences.

Exemple dans le diagnostic médical :

- Si nous avons un faux positif, la personne recevant le diagnostic médical n'est en réalité pas malade. La personne subira probablement des tests complémentaires pour confirmer ou pas la maladie et aura certainement une bonne nouvelle au final.
- Si nous avons un faux négatif, le diagnostic de l'absence d'une maladie est donné à une personne qui est pourtant malade. Les conséquences peuvent donc être très grave : absence de traitement, développement de la maladie et dans le pire des cas, une conséquence possiblement fatale.

Dans cet exemple, le modeleur devra éviter à tout prix les erreurs de type 2 donc favoriser le rappel. Cela se fera malheureusement au prix d'une diminution de la précision et donc d'une augmentation des erreurs de type 1.

Ce choix de la proportion entre précision et rappel est donc une décision importante du modeleur en fonction du cas d'usage et peut dépendre de facteurs éthiques et culturels de celui-ci (voir le site « Moral Machine » qui aborde ce sujet, notamment les dilemmes moraux).

Le modeleur est donc soumis à deux décisions majeures qui l'amène à évoluer dans un espace délimité par le sous apprentissage et le surapprentissage, quel que soit le modèle et le cas d'usage. S'il doit traiter des cas de classifications, cet espace est d'autant plus délimité par la précision et le rappel. Ses choix sont majeurs car effectués très en amont et non remis en question par la suite. Leurs conséquences auront un impact tout au long de l'exécution du modèle.

1.2.6.4.3 Quelques modèles les plus connus

Il existe de nombreux modèles ou algorithmes qu'il est possible de dispatcher selon les modes d'apprentissage et cas d'usages associés que nous avons vus précédemment.

En voici une liste succincte :

- Apprentissage supervisé :
  - Régression :
    - Régression linéaire
    - Régression logistique
  - Classification :
    - KNN
    - Séparateur à Vaste Marge (SVM)
    - Arbre de décision
    - Réseaux neuronaux
- Apprentissage non-supervisé :
  - Segmentation
    - K-means
  - Réduction de dimension
    - Isolation forest
- Apprentissage par renforcement :
  - Temporal difference learning (TD-Learning)
  - Q-learning

Voici un dispatchage de ces quelques exemples de modèles connus selon les 3 modes de l'apprentissage évoqués précédemment :

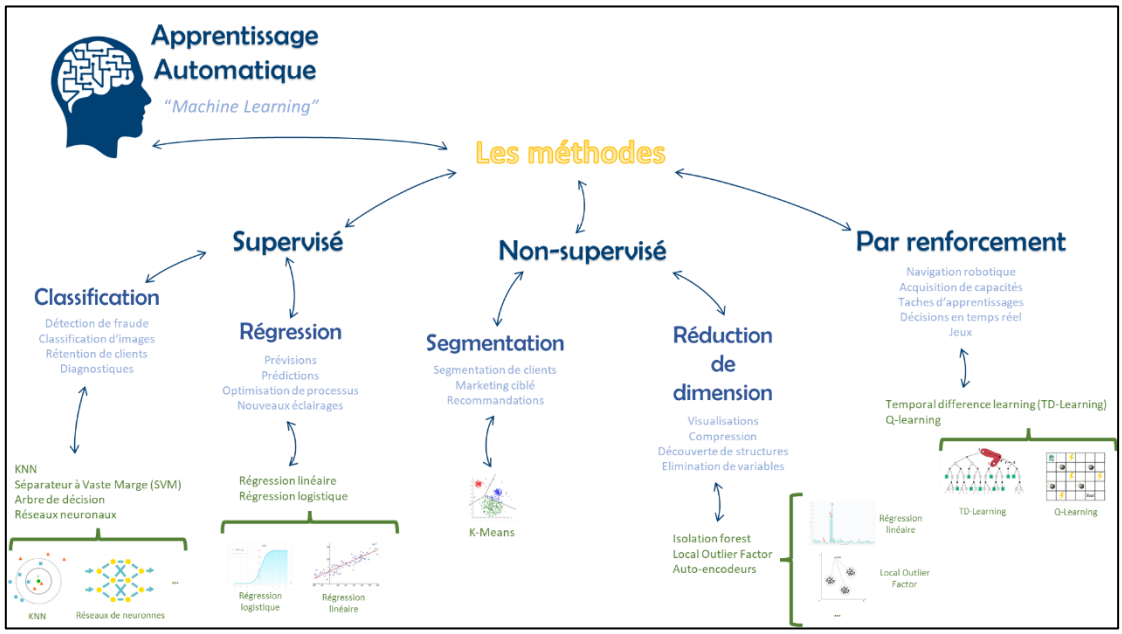


Figure 48 : Quelques modèles les plus connus - Source : autrice

Parmi tous ces modèles, les réseaux neuronaux sont sans doute les plus complexes. Le but ici est de reproduire le fonctionnement du cerveau humain et la transmission du signal de neurones en neurones.

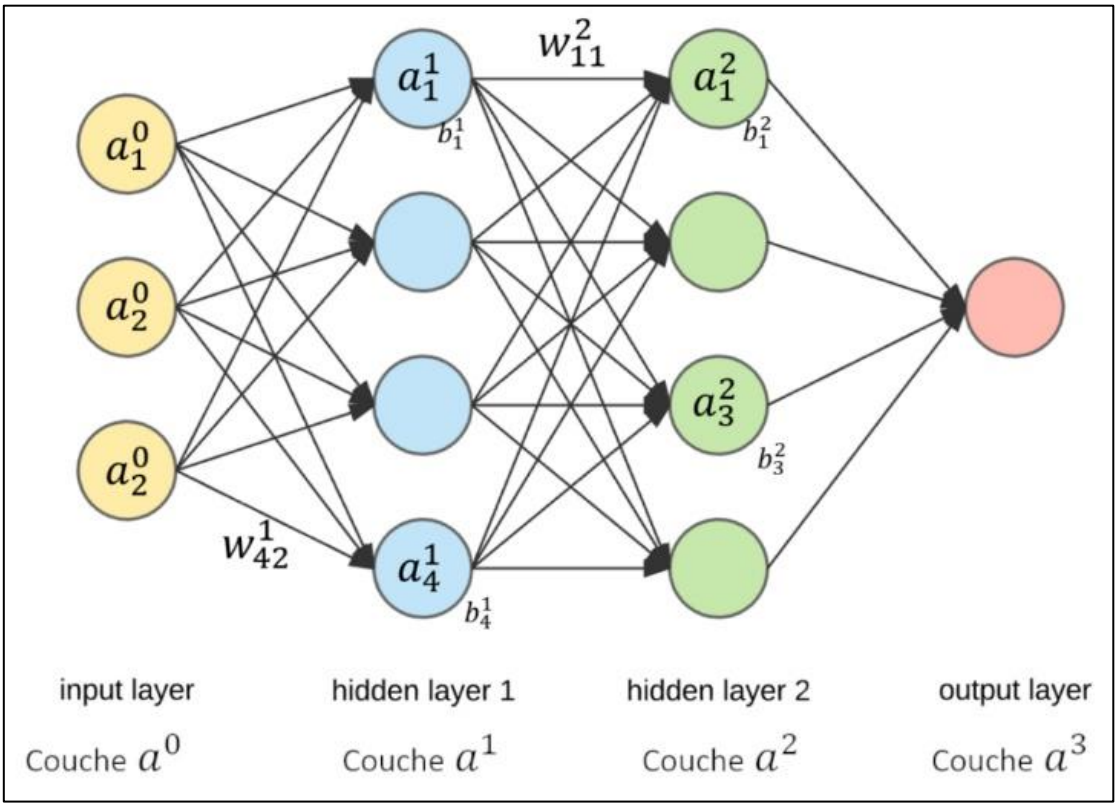


Figure 49 : Architecture d'un réseau de neurones de type "perceptron multi-couches" - Source : aspexit.com

Il est composé de neurones agencés en couches :

- Une couche d'entrée,
- Un ou plusieurs couches cachés,
- Une couche de sortie.

Afin de simplifier la compréhension, il est important de retenir que chaque neurone est une opération mathématique, qui prend des valeurs en entrées, les multiplie par les poids qu'on lui a défini, puis transmet la somme à travers la fonction d'activation aux autre neurones. Cela génère très souvent des probabilités comprises entre 0 et 1.

Chaque neurone réalise en réalité une opération mathématique pour laquelle il est programmé. Lorsqu'il reçoit en entrée une ou des données, le neurone calcule une somme pondérée en utilisant les valeurs des variables d'entrée grâce à une fonction de combination. La somme est dite pondérée car utilisant également un poids de connexion pour donner une importance supérieure ou moindre à cette opération dans le réseau neuronal. Ensuite, la données en sortie est définie via une fonction d'activation et un seuil de sortie.

Voici la structure d'un neurone artificiel qui explique les entrées, la pondération via les poids, la fonction d'activation et la sortie que le neurone donne :

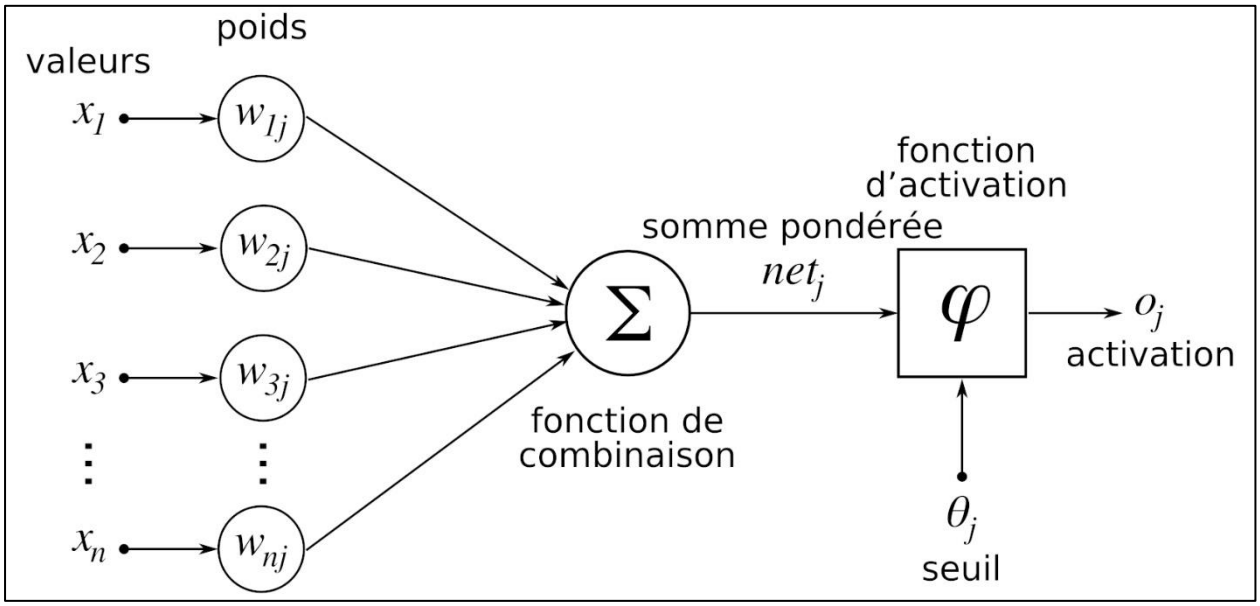


Figure 50 : Structure d'un neurone artificiel - Source : wikipedia.org

Selon la complexité du cas d'usage ou du problème à résoudre, le nombre de couche choisi par le modeleur peuvent ne pas suffire à traiter les problèmes complexes, nécessitant alors l'augmentation du nombre de couches. Seul le nombre de couches cachées peut être augmenté.

1.2.6.5 L'apprentissage profond : le Deep Learning

Le **Deep Learning** est une méthode d'apprentissage automatique qui est une spécification du Machine Learning. C'est l'Intelligence Artificielle dans sa version la plus aboutie à ce jour. Sa spécificité repose sur le fait que le Deep Learning utilise les réseaux de neurones dit « profond ».

Mais qu'est-ce qu'un réseau neuronal « profond » ? c'est l'augmentation et l'accumulation de plusieurs couches cachées.

Il est important de préciser que l'appréciation du nombre de couche cachées à partir duquel il s'agit ou pas d'un réseau profond (et donc de **Deep Learning**) est très empirique et varie selon la perception de chaque modeleur, selon mes recherches.

Nous retiendrons ici qu'il est admis que l'utilisation de réseaux neuronaux « profonds », soit le Deep Learning, se fait à partir de plus de cinq couches cachées de manière générale et pour certains Data Scientists, plus de trois couches cachées.

L'inconvénient principal de ces réseaux de neurones à grand nombre de couches cachées **est le fonctionnement « boîte noire » ce qui complexifie sa mise en œuvre et sa compréhension**.

Toutefois, l'augmentation du nombre de couches cachées et donc l'arrivée des réseaux profonds a révolutionné de multiples secteurs en permettant de résoudre des problèmes complexes qui ne pouvaient l'être auparavant.

- **Les grandes étapes de l'évolution des réseaux de neurones :**

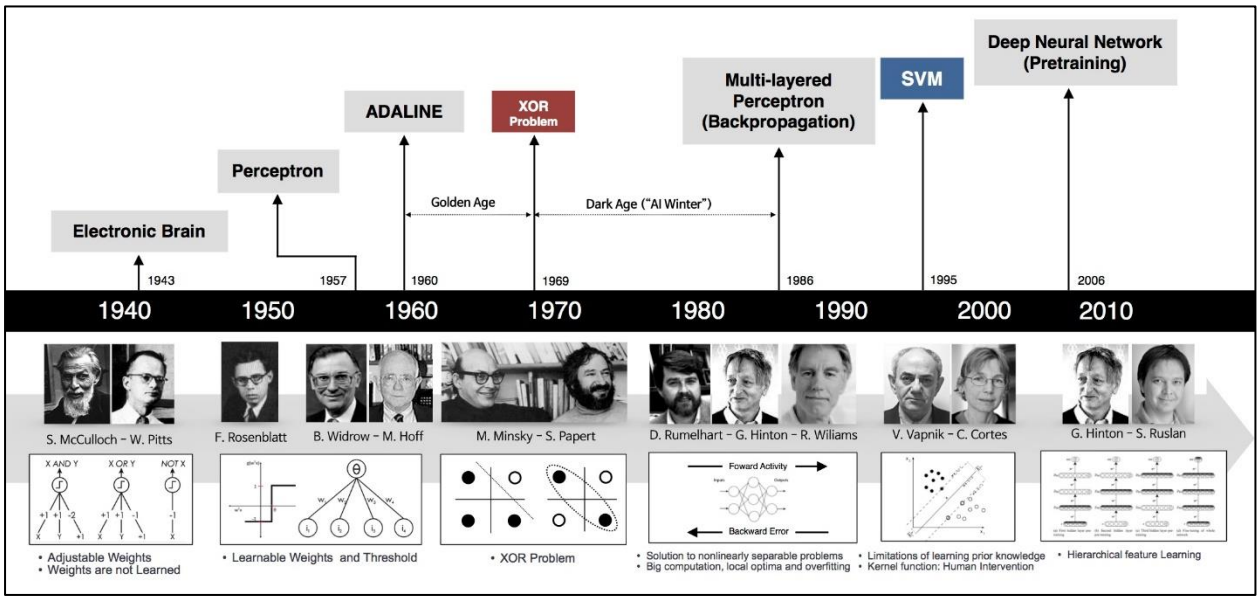


Figure 51 : Grandes étapes de l'évolution des réseaux de neurones - Source : stateofther.github.io

C'est **en 1943** que l'on entend parler pour la première fois d'un neurone artificiel, pseudo « cerveau électronique » qui n'était à ce moment qu'une simple fonction mathématique. En effet, **Warren MCCULLOCH et Walter PITTS** publient **le neurone formel**, premier modèle informatique et surtout mathématique du neurone biologique.

En 1957, apparait alors le « **Perceptron** » dans un papier expliquant le fonctionnement d'un neurone artificiel, premier réseau de neurones créé par **Frank ROSENBLATT**.

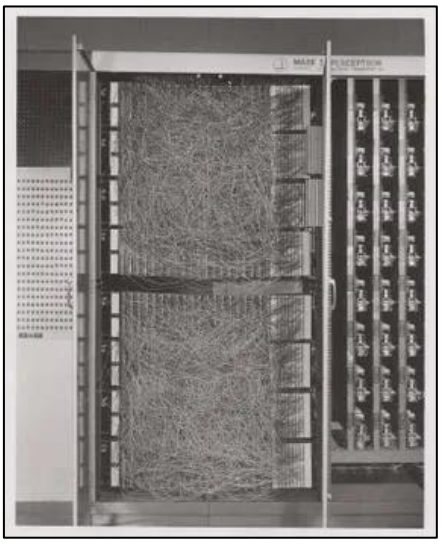


Figure 52 : « Mark I Perceptron machine » : première implémentation du Perceptron - Source : wikipedia.org

En 1960, **ADALINE**, développé par les professeurs **Bernard WIDROW** et **Ted HOFF**, est une amélioration du perceptron. **ADAptive LINear Element (ADALINE)** ouvre alors l'âge d'or des réseaux de neurones qui vont connaître une véritable frénésie pendant dix ans.

Malheureusement, en **1969**, **Marvin MINSKY** et **Seymour PAPERT** révèlent dans un article « *Perceptrons : an introduction to computational geometry* » que le Perceptron utilise un séparateur linéaire et qu'il ne peut donc pas implémenter le XOR (appelé "OU Exclusif"), provoquant alors par cette remarque « **la mort du Perceptron** ».

Ce problème est présenté dans les graphique ci-dessous :

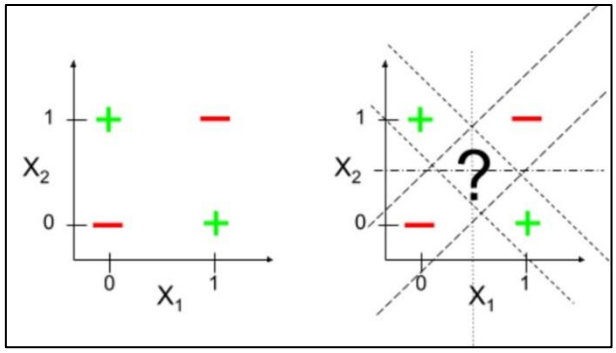


Figure 53 : Problème du XOR pour le Perceptron - Source : tech-quantum.com

Avec deux observations positives et deux négatives, il est impossible de les séparer par une seule de droite.

L'incapacité à l'époque des réseaux de neurones à résoudre ce problème va engendrer une défiance des investisseurs et le début de ce que l'on a appelé « **l'hiver de l'Intelligence Artificielle** » (déjà évoqué précédemment) qui durera plus de 20 ans.

En 1986, un évènement va initier un changement majeur : L'invention du **réseau de neurones à plusieurs couches avec rétropropagation** par **David RUMELHART**, **Geoffrey HINTON**, **Ronald WILLIAMS**. Celui-ci consiste, durant la phase d'apprentissage et une fois une observation passée à travers toutes les couches du réseau, à revenir en arrière et à modifier les paramètres afin que le résultat s'approche au plus près du résultat attendu. Cette méthode augmente considérablement les capacités d'apprentissage des réseaux de neurones.

Voici un schéma du fonctionnement d'un réseau de neurone avec couche caché avec rétropropagation :

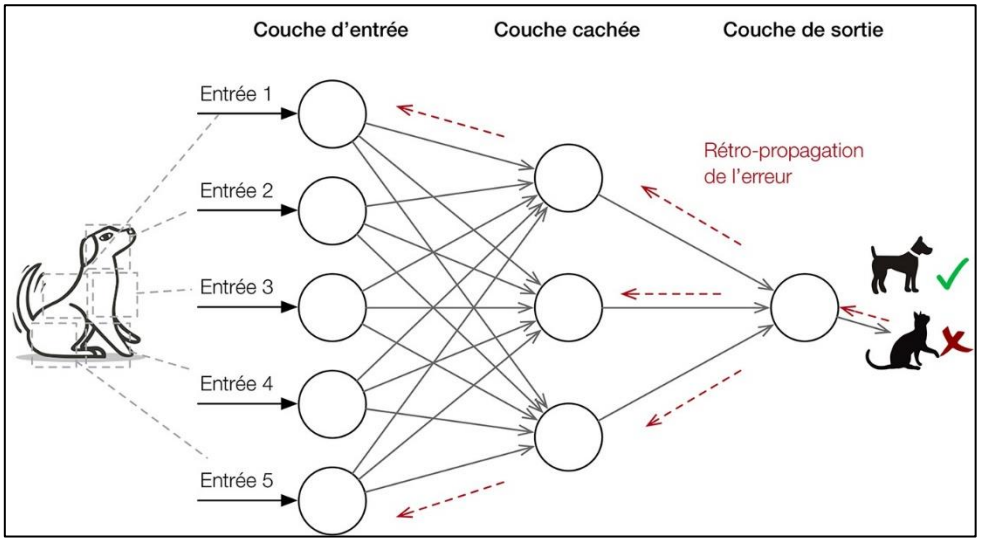


Figure 54 : Fonctionnement d'un réseau de neurone avec couche caché avec rétropropagation - Source : cairn.info

Un autre évènement majeur intervient **en 1995** lorsque **Vladimir VAPNIK**, un chercheur russe arrivé aux états unis à la suite de la chute de l'URSS, invente la technique **SVM**, **Support-Vector Machine** ou **Séparateurs à Vaste Marge** en français, avec **Corinna CORTES**.

Par une astuce mathématique digne d'un tour de « passe-passe » appelée « **Kernel Trick** », il ajoute, pour simplifier, une dimension supplémentaire à l'espace et lui donne une courbure qui permet alors de séparer les populations de façon linéaire permettant ainsi de résoudre le problème bloquant du **XOR** pour les réseaux de neurones.

C'est le retour en grâce des réseaux de neurones dont l'évolution ne s'arrêtera plus et se poursuit encore aujourd'hui.

En 2006, le terme de **Deep Learning** est forgé dans un papier "*A fast learning algorithm for deep belief nets*" écrit par **Geoffrey HINTON** et son équipe, dont **Ruslan SALAKKHUTDINOV**, notamment grâce à une série de stratégie et d'idées innovantes comme l'ajout de couches cachées et le pré-entrainement, mais l'utilisation des réseaux de neurones dans l'approche non-supervisé. Ainsi, HINTON et ses collègues démontrent que la profondeur d'un réseau de neurones est préférable à la largeur (nombre de neurones d'une même couche).

En 2010 démarre le challenge « **ImageNet** ». Il s'agit ici de comparer les performances des réseaux de neurones auxquels on soumet une bibliothèque d'images standardisés : 1,2 million d'images réparties en 1000 catégories en guise d'apprentissage. Puis 150 mille images sont à reconnaitre dont voici quelques exemples :

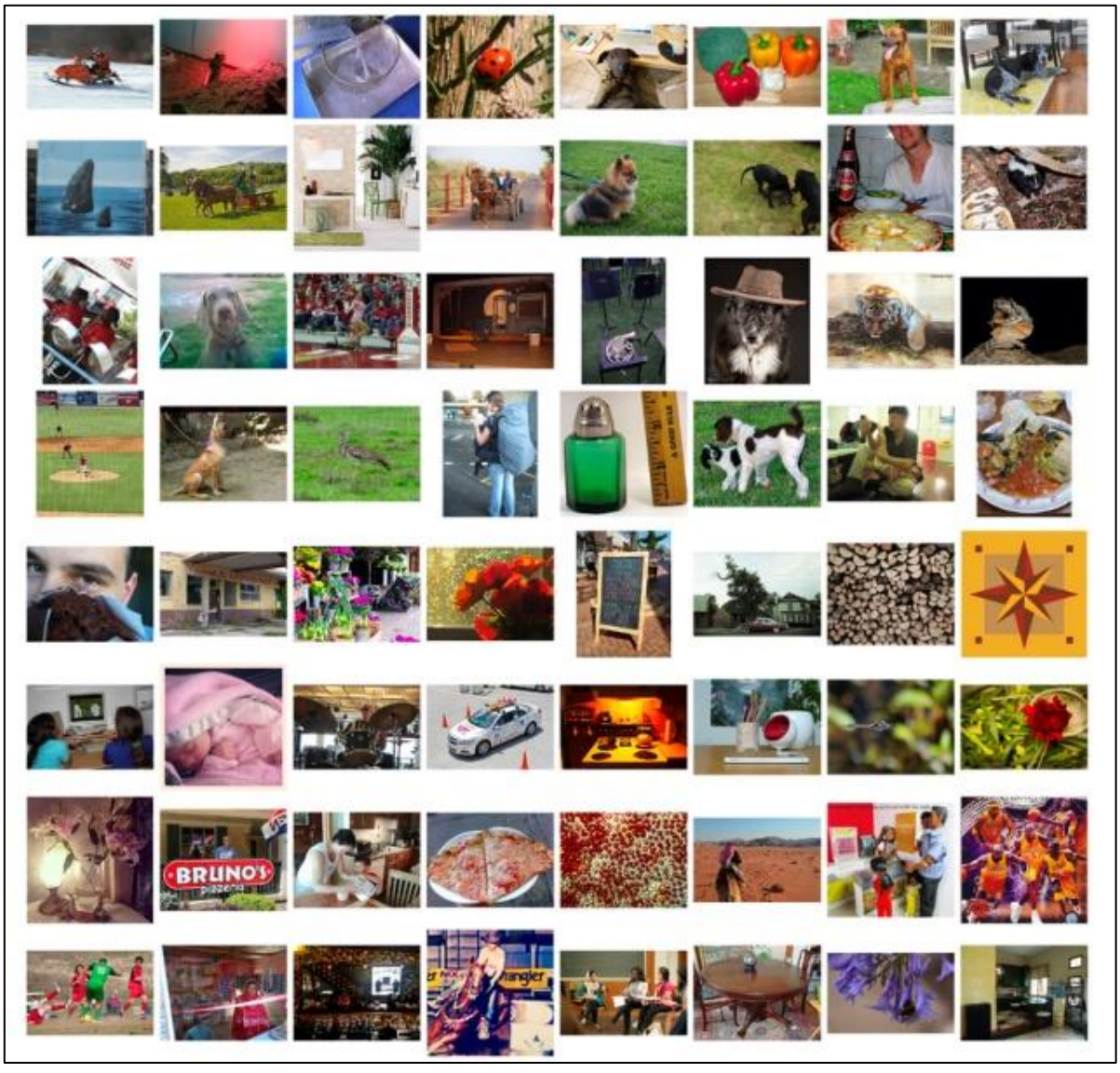


Figure 55 : Exemples d'images à reconnaitre dans le challenge "ImageNet" - Source : machinelearningmastery.com

A l'issue, un pourcentage d'erreur est calculé :

- **En 2010 et 2011**, avec des réseaux de neurones à moins de quatre couches, **les réseaux de neurones font 28 à 25 % d'erreurs**
- **En 2012 et 2013**, les premiers réseaux de neurones profond à huit couches sont soumis au challenge et **font 16 % puis moins de 12 % d'erreurs**
- **En 2014**, **VGG** avec 19 couches puis **GoogleNet** avec 22 couches, **le pourcentage d'erreur tombe respectivement à 7,3 % et 6,7 %**
- **En 2015**, **Resnet** opère une révolution avec **152 couches** et obtient un pourcentage d'erreurs de **3,57%**.

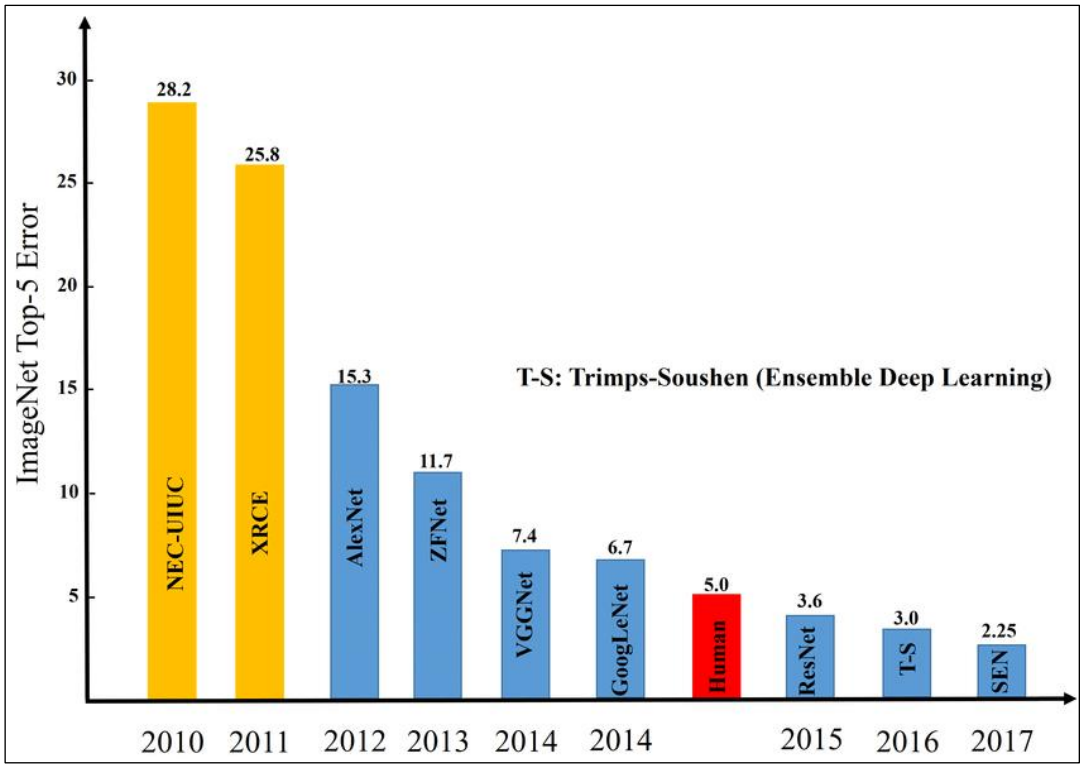


Figure 56 : Principales percées dans le défi de classification d'images ImageNet - Source : researchgate.net

Cette date de 2015 est importante : En effet, si l'on soumet la même bibliothèque d'image à un groupe d'êtres humains, ils font en moyenne 5 % d'erreurs. **C'est donc la première fois qu'un système d'IA est officiellement meilleure qu'un être humain pour la reconnaissance d'image**.

Il s'agit d'un véritable exploit du fait que maintenir 150 couches cachées en 2015 est extrêmement difficile notamment en raison de la grande instabilité liée à ce nombre de couches : Il est déjà difficile de maintenir la stabilité d'un réseau fort d'une dizaine de couches.

Depuis 2015, d'autres « challengers » ont relevé le défi avec des résultats bien supérieur encore à **ResNet**. Les résultats sont maintenus sur une plateforme proposant un graphique dynamique que vous pouvez visualiser sur ce lien.

En 2017, une des dernières grandes avancées majeures est sans doute la création de la technologie « **Transformer** » par **Google**. Ce modèle fait aujourd'hui référence, notamment dans le traitement automatique des langues mais aussi dans la « **Computer Vision** » récemment, parlant alors de « **Vision Transformer** ». Ce modèle permet de gagner en performances, ce qui permet d'effectuer des entrainements sur des ensembles de données plus volumineux.

Il est actuellement la base de modèle tel que **BERT** (***Bidirectional Encoder Representations from Transformers***), développé par **Google**, ou **GPT** (***Generative Pre-Training Transformer***), développé par la société **OpenAI**. Ce dernier est le très célèbre modèle qui a évolué plusieurs fois jusqu'à nos jours et dont la version la plus récente est **GPT-4**, utilisé dans la version **ChatGPT Plus** (version payante de **ChatGPT**).

Le « Transformer » a donc permis à l'Intelligence Artificielle de franchir un cap décisif dans le domaine du traitement automatique du langage, résultant à son accessibilité et sa notoriété récente du public via ChatGPT.

En 2019, **le prix Turing est attribué à Geoffrey HINTON**, **Yoshua BENGIO** et **Yann LECUN** pour leurs travaux sur les réseaux de neurones profonds qui ont permis de très grandes avancées dans l'Intelligence Artificielle.

Figure 57 : Les trois lauréats 2019 du prix Turing - Source : usinenouvelle.com

En conclusion, **le Deep Learning est donc aujourd'hui la version la plus aboutie de l'IA**, en étant un sous ensemble du Machine Learning dans sa partie modèle basé sur des réseaux de neurones à plusieurs couches cachées (minimum 5 en général). Le Machine Learning est lui-même un sous ensemble de l'Intelligence Artificielle dans son approche par apprentissage automatique.

Le graphique ci-dessous montre l'évolution de l'IA dans le temps, et les inclusions dans celle-ci du Machine Learning et du Deep Learning :

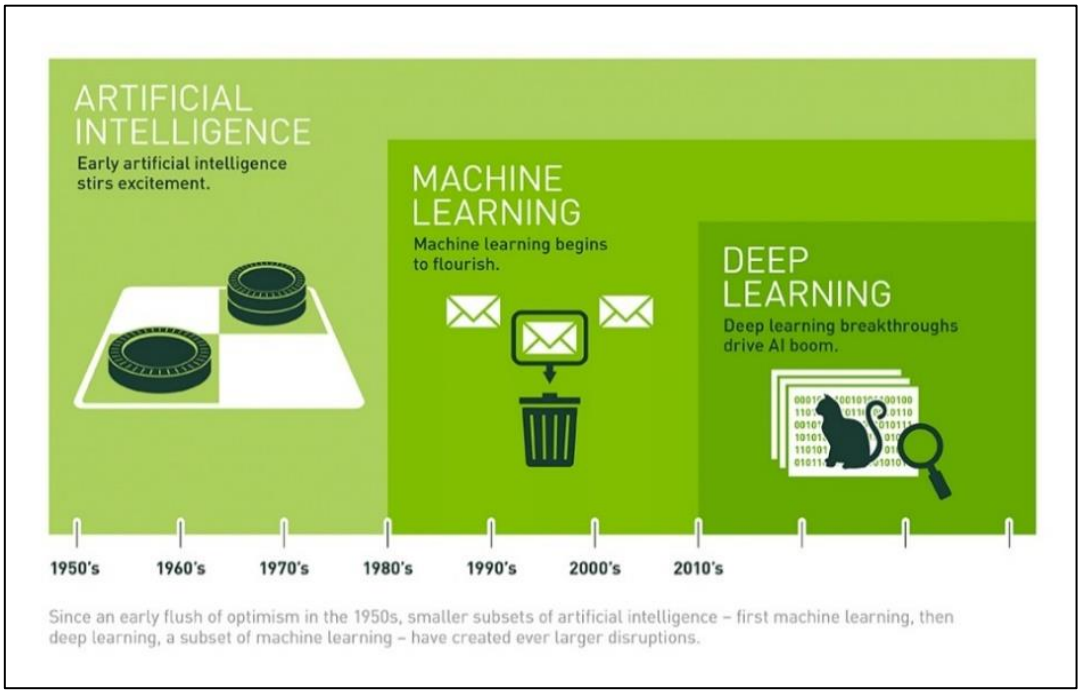


Figure 58 : Evolution de l'IA dans le temps - Source : nvidia.com

Et si aujourd'hui l'Intelligence Artificielle a fait autant parler d'elle, c'est parce qu'elle est devenue accessible au public via les IA génératives et les avancées pour interagir avec l'IA à travers le NLP (« Natural Language Processing »). Les Transformers y sont pour beaucoup et finalement, représentes-en quelque sorte la partie visible de l'iceberg, d'un point de vue « utilisateur » :

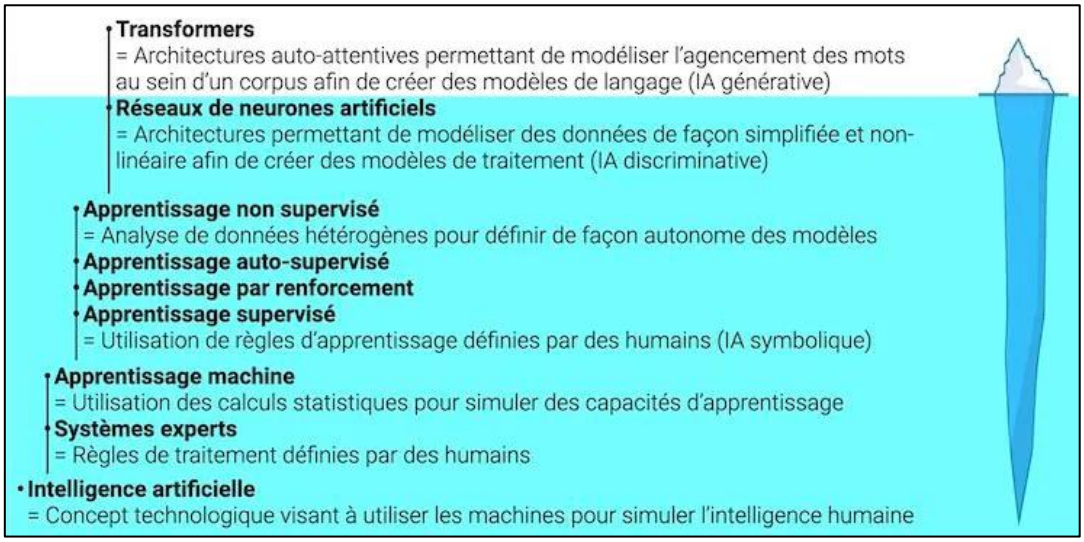


Figure 59 : Les Transformers, face visible de l'iceberg "IA" par la notoriété des IA génératives - Source : fredcavazza.net

1.2.7 Les usages de l'IA

Une des études publiées par le **cabinet McKinsey**, appelée « *Notes from the AI frontier: Applications and value of deep learning* » , met en évidence les secteurs où l'Intelligence Artificielle a le potentiel le plus important en termes de valeur ajoutée.

Dans ce graphique, **McKinsey** donne par secteur d'activité un **pourcentage de la valeur ajoutée** qu'apporterait l'IA par rapport à d'autres techniques analytiques. Ce serait donc les secteurs qui ont le plus intérêt à utiliser ou mettre en place de l'IA au vu du gain que cela pourrait leur apporter :

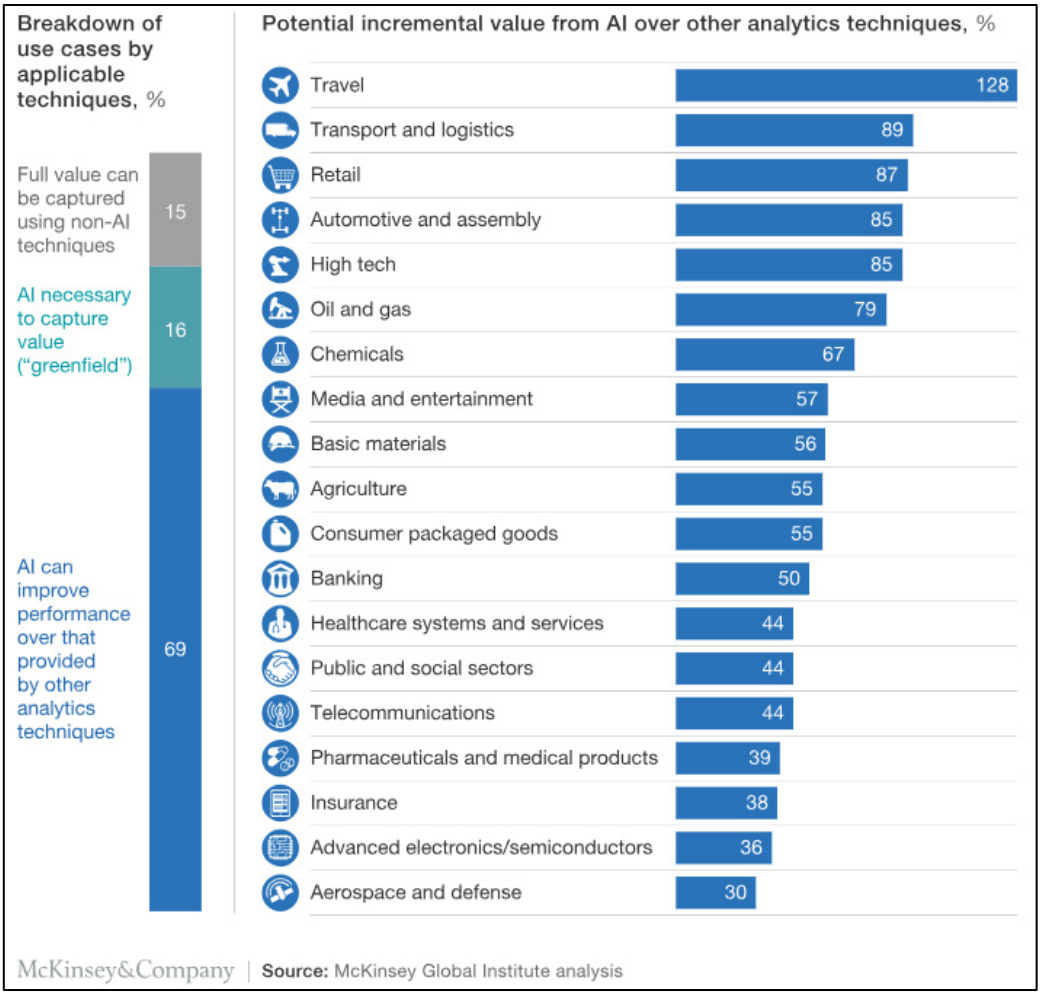


Figure 60 : Secteurs dont le gain est potentiel grâce à l'IA - Source : mckinsey.com

Selon eux, les secteurs du voyage, des transports et du commerce arrivent en tête. On constate aussi des industries où l'application de l'IA peut sembler moins évidente, comme l'agriculture, mais qui offrent néanmoins des perspectives prometteuses.

Le secteur bancaire est plus loin dans le classement, principalement en raison de sa résistance au changement et des contraintes légales importantes. L'industrie de la sécurité et de la défense arrivent en dernière position, probablement parce qu'elles ne partagent pas toutes leurs recherches avec nous.

Voici ici un échantillon représentatif mais non exhaustif des domaines d'application majeurs de l'Intelligence Artificielle :

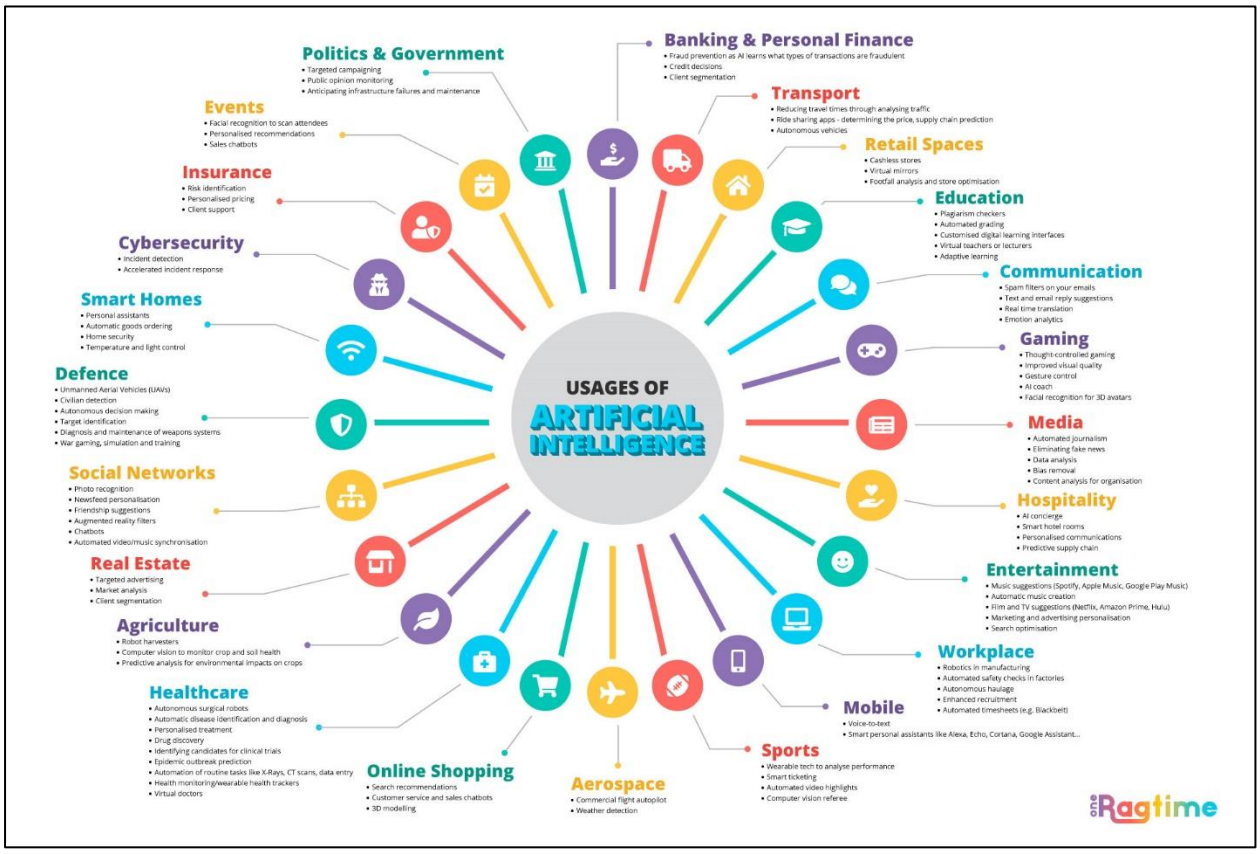


Figure 61 : Usages par secteur de l'IA - Source : OneRagTime

Parmi eux, on retrouve notamment ceux de la ville intelligente, qui s'appuie sur cinq piliers principaux : le trafic routier, l'énergie, la santé, l'éducation et la sécurité.

Pour illustrer le pilier de la sécurité, prenons l'exemple de la Chine qui a fait des avancées considérables en la matière (au-delà des aspects éthiques discutables). Dans une ville de plus de 10 millions d'habitants, la police est capable d'arrêter un malfaiteur en moyenne en seulement 7 minutes. Cette efficacité est renforcée par l'utilisation de la reconnaissance faciale, de la reconnaissance de silhouettes et de la reconnaissance de la démarche.

Les avancées les plus notables sont sans doute dans le domaine de la santé, où l'automatisation du diagnostic médical, en particulier par l'analyse d'imagerie, est en plein essor.

Il ne faut pas oublier non plus les applications déjà en cours dans d'autres secteurs. Dans la distribution, par exemple, l'IA aide à optimiser les approvisionnements et à prévoir les comportements d'achat. Dans le secteur financier, l'IA peut aider à gérer le risque de crédit ou à anticiper les tendances du marché. Dans les télécommunications, l'IA peut être utilisée pour prédire les taux de désabonnement. Enfin, dans l'industrie manufacturière, l'IA joue un rôle clé dans la gestion de la qualité et de la sécurité.

Voici rapidement des exemples d'usage par secteurs (liste non exhaustive) :

- **Le secteur de l'automobile : véhicules connectés & véhicules autonomes**

D'après une recherche effectuée par JP Morgan, on s'attend à ce que le marché des voitures intelligentes augmente de 75% d'ici 2025.

La voiture autonome présente de multiples applications pour l'Intelligence Artificielle. La plus importante d'entre elles est l'apprentissage machine pour la conduite, qui enseigne à la voiture comment réagir en cas d'accident. L'IA joue également un rôle dans l'assurance de la qualité des composants du véhicule grâce à la maintenance prédictive, ainsi que dans la compréhension de l'environnement à partir des données des capteurs. En outre, l'IA est utilisée pour analyser le comportement du conducteur et assurer la cybersécurité en surveillant l'état de la connectivité pour prévenir toute tentative de piratage.

- **Le secteur de la santé : médecine prédictive & chirurgie assistée**

L'IA se positionne comme une composante clé de la médecine de demain, participant à des interventions assistées, à la surveillance à distance des patients, à l'élaboration de prothèses intelligentes et à la personnalisation des soins grâce à l'analyse de volumes croissants de données (big data), entre autres.

Voici les principaux domaines d'utilisation de l'IA en santé :

- La médecine prédictive : prévoir une maladie et/ou sa progression
- La médecine personnalisée : proposer des recommandations de traitements adaptés à l'individu
- Le soutien à la décision : diagnostic et thérapie
- Les robots d'assistance : particulièrement utiles pour les personnes âgées et vulnérables
- La chirurgie assistée par ordinateur
- La prévention : prévoir une épidémie, surveiller les effets des médicaments (pharmacovigilance)

- **Le secteur de l'industrie : maintenance prédictive**

La maintenance prédictive a pour but d'identifier les problèmes sur les machines avant qu'ils ne s'aggravent. Sa principale force réside dans sa capacité à prévoir les défaillances, ce qui permet d'éviter des interruptions coûteuses de la chaîne de production.

Cette méthode de maintenance favorise le passage d'une logique de flux poussé à une logique de flux tiré. L'intervention du fournisseur n'est nécessaire que lorsque des signaux indiquent une probable défaillance à court terme d'une machine. Cette prédiction des pannes est rendue possible grâce à l'installation de capteurs qui recueillent plusieurs milliers de données chaque jour. C'est l'Internet des Objets (**IoT**) et l'apprentissage machine qui rendent cela possible.

- **Le secteur du marketing : service client et marketing personnalisé**

L'Intelligence Artificielle propose trois catégories de solutions essentielles pour le marketing : l'acquisition de clients (analyse et segmentation de l'audience, identification visuelle du contexte, scoring et ciblage), la conversion (personnalisation et recommandations, création de contenus, optimisation de sites et de supports, gestion automatisée des campagnes) et la fidélisation (agents de chat, automatisation du programme client, analyse du comportement, calcul d'attribution et prédictions).

Des entreprises comme **Amazon** et **Netflix** exploitent notamment la technologie de l'IA pour proposer des recommandations de contenu sur mesure pour chaque client.

- **Le secteur de l'assurance : assurance prédictive & détection de la fraude**

Grâce à l'analyse de données, l'IA peut prédire et contrôler les risques. À long terme, l'IA devrait réduire le seuil à partir duquel un risque est assurable. L'IA, par l'analyse et l'interprétation de toutes les données clients, facilitera la détermination automatique des primes d'assurance.

Les technologies de l'IA améliorent également la productivité des professionnels de l'assurance, ce qui se traduit par une meilleure qualité de service pour les clients.

Enfin, l'IA contribue à la détection de la fraude, grâce à l'élaboration d'algorithmes basés sur l'observation de comportements frauduleux.

- **Le secteur de la défense et l'aéronautiques : cybersécurité**

À l'heure actuelle, plusieurs utilisations de l'Intelligence Artificielle ont déjà été intégrées dans le domaine de la défense, que ce soit dans les forces terrestres, les technologies maritimes, l'aéronautique ou la cybersécurité.

Selon une étude de **P&S Market Research**, le marché mondial des logiciels de cybersécurité intégrant l'Intelligence Artificielle pourrait générer un chiffre d'affaires de 18,2 milliards de dollars en 2023.

Grâce à l'apprentissage automatique, ou Machine Learning, il est possible de repérer une attaque avant qu'elle ne soit enregistrée dans les bases de données de la communauté des défenseurs du cyberespace.

- **Le secteur du tourisme : e-tourisme et assistants personnels**

L'Intelligence Artificielle est en passe de devenir un élément essentiel dans l'évolution et la modernisation du secteur du tourisme. L'Intelligence Artificielle et l'apprentissage automatique permettent la création d'un chatbot, ou assistant virtuel, doté de deux caractéristiques fondamentales : le langage et l'intelligence cognitive.

En tant que véritable assistant personnel, le chatbot peut fournir des réponses rapides et pertinentes à des milliers de touristes à la recherche d'informations ou de services divers, comme un code Wifi, une réservation de chambre ou de taxi. Le chatbot assure une interaction avec le client qui reflète la culture et l'image de marque de l'entreprise ou de l'individu qui l'utilise.

- **Le secteur de la banque : chatbots & robot advisers**

Dans le secteur bancaire, nous aborderons plus en détail les cas d'usage possible dans la partie 3. Néanmoins, nous pouvons tout de même citer comme exemple les chatbot & robot advisers. Ici, l'Intelligence Artificielle poursuit un objectif à deux volets. Premièrement, elle améliore l'efficacité des conseillers en leur permettant de fournir des conseils de qualité supérieure sur l'ensemble des produits. Deuxièmement, elle allège la charge des gestionnaires de clientèle en prenant en charge les tâches routinières, leur permettant ainsi de se focaliser sur des tâches plus sophistiquées.

Deux applications de l'IA révolutionnent le conseil financier :

- La technologie des chatbots facilite et fluidifie la relation client.
- Les robots-conseillers offrent aux clients des placements adaptés à leurs objectifs financiers et à leur situation patrimoniale.

- **Le secteur du luxe : détection de tendance en ligne**

L'Intelligence Artificielle est largement utilisée dans l'industrie du luxe, tant pour l'élaboration de services innovants, l'amélioration de la relation client, que pour sa capacité à anticiper les segments de produits les plus prometteurs.

La technologie de l'IA va très probablement aider les créateurs à mieux comprendre les besoins des consommateurs. Les grandes marques de luxe peuvent recueillir et analyser les données des clients, en association avec l'étude de leurs comportements, pour ainsi prédire quels produits seront vendus,

à quelle quantité et dans quels lieux géographiques. **LVMH** a décerné son premier prix d'innovation à l'entreprise française **Heuritech**, spécialisée dans la détection des tendances en ligne. Grâce à sa capacité à comprendre les préférences des consommateurs, elle peut aider les sites de commerce électronique à proposer le contenu approprié au bon client.

- **Les médias & divertissements : contenus personnalisés**

L'Intelligence Artificielle influence chaque phase de la vie d'un contenu, depuis sa création jusqu'à sa consommation. Elle offre une personnalisation très poussée de la consommation de contenus et a un rôle prédominant dans la monétisation et la publicité.

Elle permet notamment de déterminer le volume publicitaire adapté à chaque type de contenu et à chaque utilisateur, c'est-à-dire la durée de publicité que les spectateurs sont prêts à regarder en l'espace d'une heure.

- **Le commerce de détail : magasins « médias »**

Le secteur du commerce de détail (le « retail ») est sans doute l'un des plus impactés par l'Intelligence Artificielle. Que ce soit en boutique ou en ligne, environ 50 % des consommateurs réclament une expérience d'achat sur mesure, adaptée à leur profil, leurs désirs et leurs habitudes de consommation.

L'Intelligence Artificielle est à même de répondre à ces exigences, en particulier par le biais de :

- L'optimisation du parcours client en boutique, grâce à l'analyse des vidéos filmées sur place ou la mise en œuvre de liens entre la boutique physique et le site web.
- La mise à disposition d'une expérience interactive avec le client (cabines intelligentes grâce à la réalité mixte).
- La proposition d'un assistant personnel numérique au service du client.

- **Le secteur de l'éducation : orientation ciblée et apprentissage personnalisé**

Depuis l'apprentissage personnalisé jusqu'à l'assistance pour l'orientation, en incluant l'automatisation de certaines tâches comme la correction des évaluations, l'Intelligence Artificielle est sur le point de bouleverser l'univers de l'éducation.

Ainsi, l'IA pourrait contribuer à la réussite des élèves, tout en renforçant leurs perspectives d'employabilité sur le long terme.

- **Le secteur de l'environnement : changement climatique et biodiversité**

L'Intelligence Artificielle propose une multitude de solutions en termes de transition écologique et énergétique.

En effet, l'IA peut être utilisée pour développer des solutions pour l'exploration et la production d'hydrocarbures. Elle permet aussi l'optimisation de l'usage des ressources naturelles en entreprise par l'automatisation de l'analyse des consommations énergétiques.

L'IA a le potentiel de générer un impact positif sur l'environnement, notamment en ciblant quatre domaines clés : l'eau, l'agriculture, la biodiversité et le changement climatique.

À travers le programme **AI for Earth** par exemple, **Microsoft** met en place plusieurs initiatives exploitant l'Intelligence Artificielle pour contribuer à la protection des écosystèmes.

1.2.8 L'IA dans le monde et en France

L'Intelligence Artificielle est devenue un enjeu majeur pour les principales nations du monde. Chaque gouvernement élabore une stratégie pour en tirer des bénéfices, qu'ils soient sociétaux ou économiques.

La plupart d'entre nous utilisons probablement couramment, sans même le savoir, une technologie basée sur l'IA. En effet, l'IA agrémente notre vie quotidienne en permettant à nos appareils de **s'adapter automatiquement à nos besoins**. Elle offre aussi des avantages aux entreprises en **automatisant certaines tâches manuelles et en facilitant l'analyse de grandes quantités de données**. Celle-ci ne se limite pourtant pas aux individus et aux entreprises.

Au-delà de ces quelques applications, l'Intelligence Artificielle permet entre autres d'aborder les défis majeurs de l'humanité, tels que la sécurité et l'environnement. C'est pourquoi les gouvernements accordent une attention particulière à l'IA. Elle est désormais un facteur déterminant de la croissance économique mondiale. Elle prend une importance stratégique, voir politique, pour de nombreux gouvernements dans le monde.

1.2.8.1 Quelques initiatives mondiales

L'engagement des **États-Unis** envers l'IA a été formalisé par Donald Trump en 2019 avec pour objectif de préserver la prospérité nationale tout en renforçant la sécurité à la fois nationale et économique. Ce plan met principalement l'accent sur la promotion de la recherche et du développement (R&D) en

IA, l'accès aux ressources et l'établissement de normes de régulation de l'IA. De plus, les États-Unis sont actifs sur la scène internationale tout en sauvegardant leurs intérêts.

De son côté, la **Chine** a toujours nourri l'ambition de se positionner en tête de la course à l'IA. Son plan, nommé « *A Next-Generation Artificial Intelligence Development Plan* », se concentre sur la R&D, l'industrialisation et le développement des compétences en IA.

Le **Royaume-Uni** intègre l'IA et les données parmi les principaux défis de sa stratégie industrielle. À travers « l'AI Sector Deal », le gouvernement britannique cherche à stimuler la R&D dans les secteurs public et privé. Il investit également dans l'éducation STEM pour encourager le développement de talents en IA et promouvoir l'éthique des données.

Le **Canada** a été le premier pays à lancer une stratégie mondiale sur l'IA en 2017, axée principalement sur la recherche et le développement de talents. En Inde, l'initiative nationale est centrée sur la promotion de l'IA pour tous. Au Japon, l'accent est mis sur la santé et la mobilité, y compris sur des solutions basées sur des robots et des chatbots à IA.

La **France** a mis en place « AI for Humanity » en 2017 avec un plan visant à aligner les ressources autour des talents, des institutions de recherche et d'un écosystème de données ouvert. L'IA française est centrée sur l'éthique et vise à stimuler l'économie du pays à travers divers secteurs. Le rapport « **Villani** » a également initié de nombreuses autres initiatives.

Voici une carte donnant une tendance estimée de croissance du marché de l'IA dans les régions du monde jusqu'en 2026 :

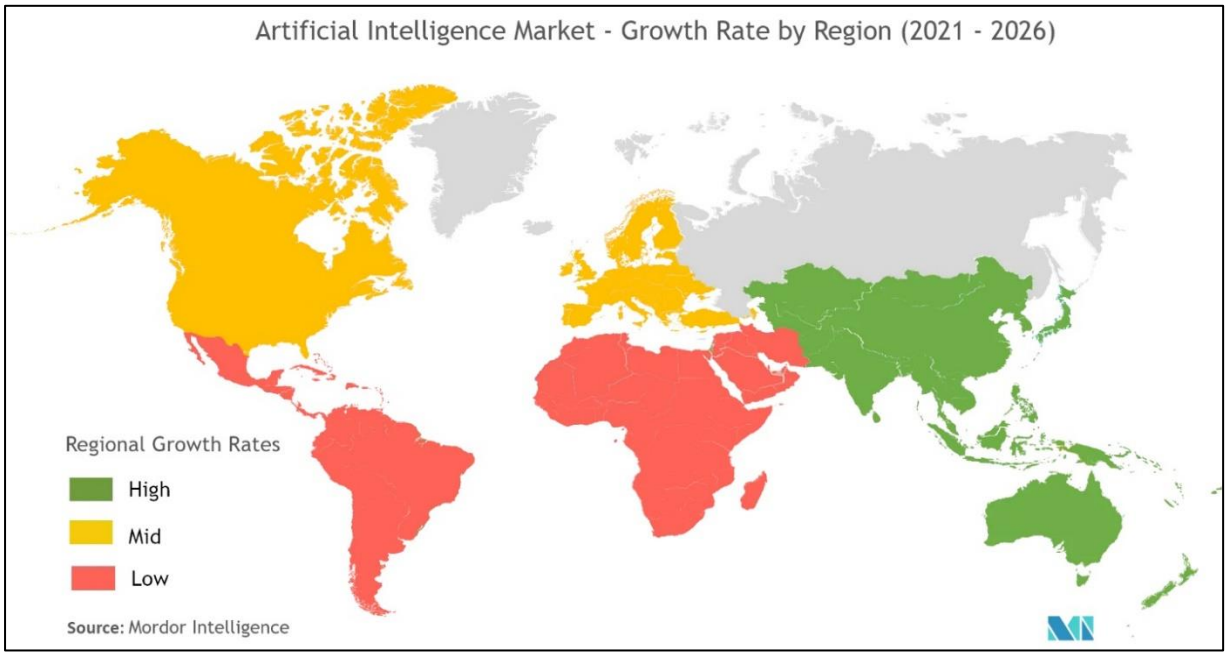


Figure 62 : Tendance de croissance du marché de l'IA dans les régions du monde - Source : mordorintelligence.com

1.2.8.2 Les plus grandes entreprises spécialisées dans l'IA

L'Intelligence Artificielle a suscité un grand intérêt auprès de nombreuses entreprises et startups ces dernières années. Voici un aperçu des plus grandes organisations mondiales qui se consacrent à l'IA.

En effet, parmi les diverses technologies, l'IA est devenue une composante quasi essentielle de notre quotidien. Que ce soit à la maison, au travail ou à l'école, elle est utilisée de multiples façons pour simplifier nos activités. Selon **Gartner**, **le marché mondial des logiciels d'IA aurait atteint une valeur impressionnante de 62,5 milliards de dollars en 2022**. Les grandes organisations ont bien compris cet enjeu et investissent massivement dans la recherche et le développement de cette technologie.

Voici donc un panorama des plus importantes entreprises mondiales dédiées à l'Intelligence Artificielle selon spiceworks.com :

- **Alphabet (Google)** : Google est un leader en matière d'IA, utilisant la technologie dans ses moteurs de recherche, son assistant vocal (Google Assistant), sa technologie de conduite autonome (Waymo), et plus encore.
- **Amazon** : Amazon utilise l'IA pour recommander des produits, optimiser la logistique, et dans son assistant vocal, Alexa.
- **Microsoft** : Microsoft investit dans l'IA à travers une gamme de produits et services, y compris son assistant vocal Cortana et son offre cloud Azure.
- **Facebook** : Facebook utilise l'IA pour la recommandation de contenu, la reconnaissance d'images, et la modération du contenu.
- **IBM** : IBM est bien connu pour son système d'IA, Watson, qui est utilisé dans une variété d'industries.
- **Apple** : Apple utilise l'IA dans une variété de ses produits, y compris Siri, son assistant vocal, et dans la reconnaissance faciale pour le déverrouillage des iPhones.
- **Alibaba** : Le géant chinois du commerce électronique utilise l'IA pour la recommandation de produits, la logistique, et dans son service de « cloud computing ».
- **Baidu** : Baidu, le plus grand moteur de recherche en Chine, investit massivement dans l'IA et la conduite autonome.
- **Tencent** : Cette entreprise technologique chinoise investit dans l'IA pour ses services de jeux, de messagerie, et de médias sociaux.
- **NVIDIA** : Bien qu'il s'agisse principalement d'une entreprise de matériel, NVIDIA est devenu un acteur clé dans l'IA en raison de l'importance de ses GPU (unités de traitement graphique) pour l'apprentissage en profondeur.

Le graphique ci-dessous reprend d'une manière plus large les entreprises les plus importantes, déployés sur la carte du monde :

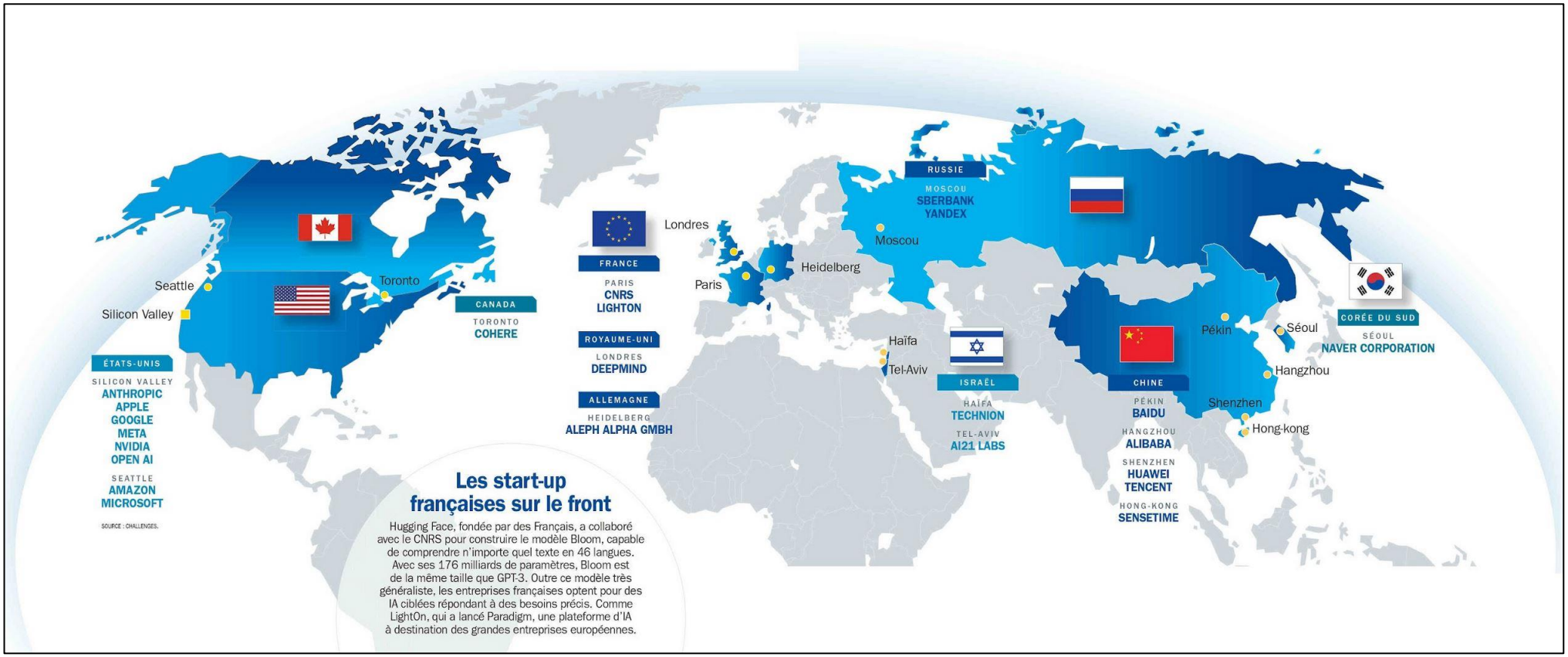


Figure 63 : Carte des principales entreprises actives dans l'IA - Source : challenges.fr

1.2.8.3 L'IA en France

La France vise à s'établir en tant que pionnière mondiale en 2030 dans le domaine de l'Intelligence Artificielle. C'est l'un des objectifs du plan France 2030, présenté par le président de la République en mars 2018.

Dans le but de réaliser cette aspiration et faisant suite aux résultats de la mission Villani, le président Emmanuel Macron a lancé une stratégie nationale pour l'Intelligence Artificielle qui repose sur un programme se déroulant en deux phases.

Cette stratégie nationale pour l'IA a établi les fondations d'une organisation à long terme de l'écosystème de l'IA, touchant chaque phase du développement technologique : depuis la recherche, les développements et innovations, jusqu'aux applications, la commercialisation et la diffusion intersectorielle, ainsi que le support et la supervision du déploiement.

La stratégie nationale se manifeste à travers **trois voies principales** :

- **Favoriser la croissance d'un écosystème de compétences.**
- **Intégrer l'IA et la donnée dans le tissu économique et administratif.**
- **Soutenir un modèle éthique qui trouve un équilibre entre l'innovation et la sauvegarde des droits fondamentaux.**

La première période, s'étalant **de 2018 à 2022**, visait à **établir un réseau d'instituts interdisciplinaires** consacrés à l'IA, **en investissant 1,5 milliard d'euros**.

La seconde phase de la stratégie nationale pour l'IA prévoit un **investissement total de 2,22 milliards d'euros pour le développement de l'IA**.

- **Phase 1 : Renforcer les capacités de recherche (2018-2022)**

La première phase de la stratégie nationale pour l'IA vise à positionner la France comme l'un des leaders mondiaux dans ces disciplines scientifiques et ces technologies clés du traitement de l'information.

Comment ? En encourageant la mise en place et le développement d'un réseau d'instituts interdisciplinaires consacrés à l'Intelligence Artificielle, en soutenant des chaires d'excellence en IA, en finançant des programmes de doctorat et en investissant dans les capacités de calcul de la recherche publique (comme le superordinateur **Jean Zay**).

Depuis 2018, cette étape s'est principalement axée sur l'augmentation des compétences en matière de recherche.

Figure 64 : Chiffres clés de cette phase 1 de la stratégie nationale pour l'IA - Source : economie.gouv.fr

- **Phase 2 : Former et attirer les meilleurs talents en IA (2021-2025)**

Le 8 novembre 2022, le gouvernement a lancé la deuxième phase de la stratégie nationale pour l'IA dans le but d'augmenter le nombre de talents formés dans ce domaine et d'accélérer le potentiel de recherche et de développement en vue de générer des réussites économiques.

Pour cette deuxième phase, un total de 2,22 milliards d'euros est prévu d'être consacré à l'IA pour les cinq prochaines années, dont 1,5 milliard d'euros de financements publics et 506 millions d'euros de cofinancements privés.

Cette seconde étape se focalise sur l'implantation des technologies d'Intelligence Artificielle à travers l'économie, tout en encourageant le développement et l'innovation dans des secteurs prioritaires tels que l'IA embarquée, **l'IA de confiance**, l'IA dédiée à la transition écologique et l'IA générative ainsi que le traitement du langage naturel (**NLP**).

Figure 65 : Chiffres clés de cette phase 2 de la stratégie nationale pour l'IA - Source : economie.gouv.fr

1.2.8.4 Le marché de l'IA

Il s'agit ici d'aborder le marché de l'Intelligence Artificielle en 2023 : Les investissements et les revenus.

1.2.8.4.1 Quels pays investissent le plus sur l'Intelligence Artificielle ?

Les pays qui ont le plus investi dans l'Intelligence Artificielle sont représentés dans le graphique ci-dessous, basé sur les données du rapport « 2023 AI Index Report » de l'université de **Stanford**. Cette représentation montre les montants les plus importants investis dans des entreprises d'IA entre 2013 et 2022. Les **États-Unis** se placent en tête avec un montant considérable de **248,9 milliards de dollars**, suivis par la **Chine** (**95,1 milliards de dollars**) et le **Royaume-Uni** (**18,2 milliards de dollars**). La **France**, à la huitième position, suit de près **l'Allemagne** avec des investissements de **6,6 milliards de dollars** sur cette période.

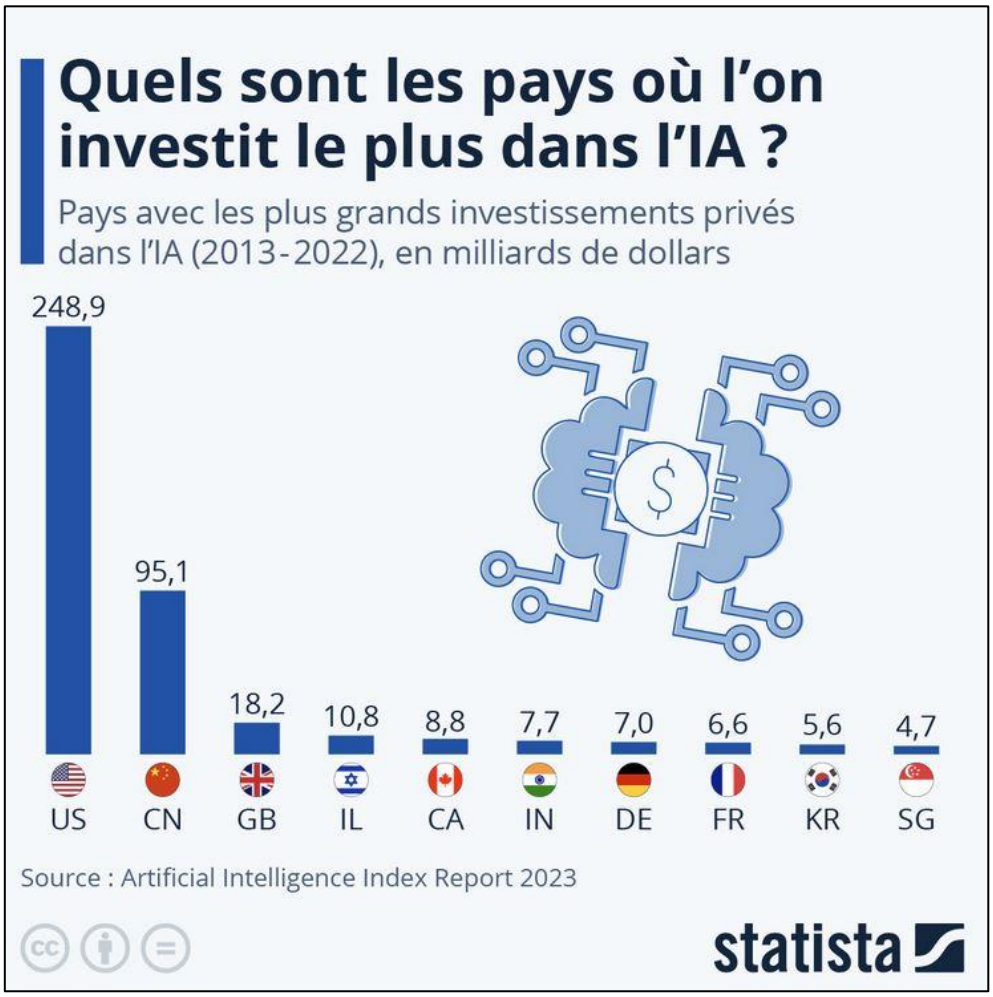


Figure 66 : Top des 10 des pays investissant le plus dans l'IA - Source - statistica.com

Il s'agit ici de l'investissement total par pays cumulé depuis 2013 à 2022.

Les fonds privés investis dans l'IA par pays en 2022 uniquement sont représentés dans le graphique ci-dessous, toujours selon le rapport de Stanford :

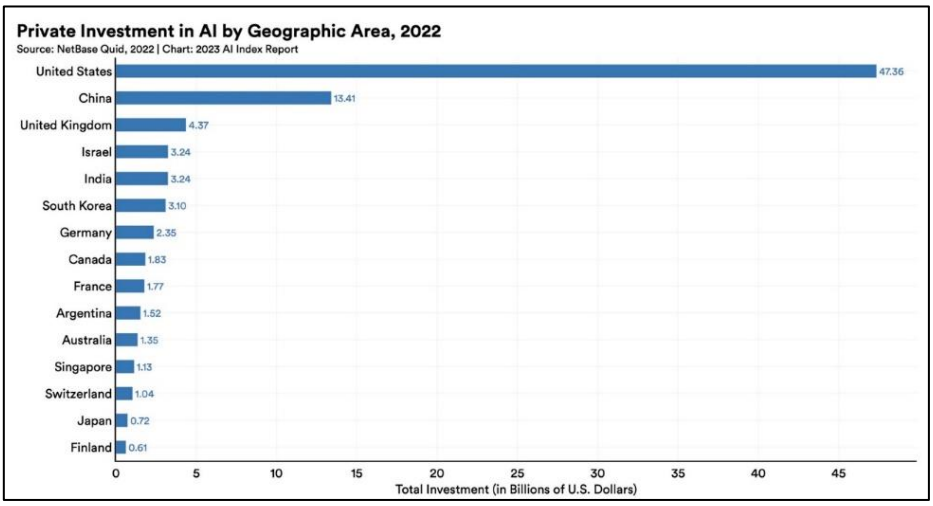


Figure 67 : Fonds privés investis dans l'IA par pays en 2022 - Source : aiindex.stanford.edu

Les **États-Unis** ont été en 2022 en tête du monde en termes de montant total des **investissements privés dans l'IA**. En effet, en 2022, les **47,4 milliards de dollars** investis aux États-Unis représentaient environ 3,5 fois le montant investi dans le pays suivant, **la Chine (13,4 milliards de dollars)**. Les États-Unis continuent également de dominer en termes de nombre total de nouvelles entreprises d'IA financées, avec 1,9 fois plus que l'Union européenne et le Royaume-Uni combinés, et 3,4 fois plus que la Chine.

Le rapport indique également qu'**en 2022**, les investissements les plus significatifs dans le domaine de l'IA ont été réalisés dans le secteur de la médecine et de la santé, atteignant un montant de **6,1 milliards de dollars**.

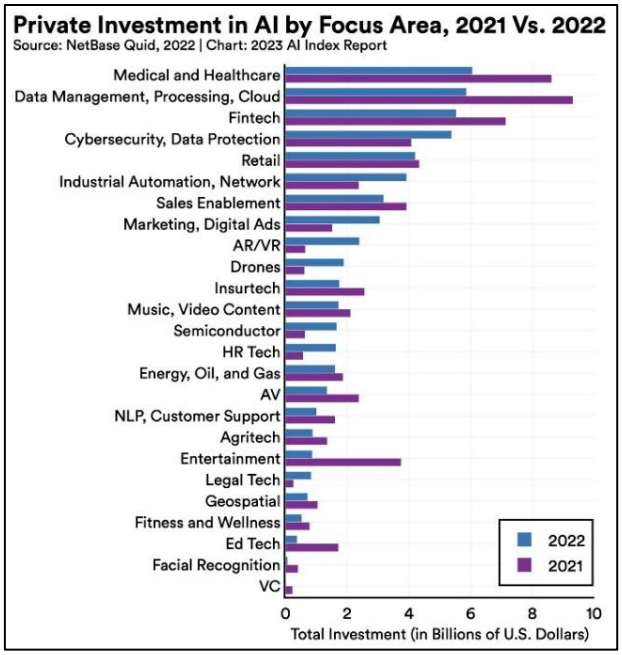


Figure 68 : Investissements privés par secteur en 2022 - Source : aiindex.stanford.edu

Comme vous pouvez le voir, les investissements privés en IA ont diminué en 2022 par rapprt à 2021 ce qui casse pour la première fois depuis 2013 la dynamique d'investissement :

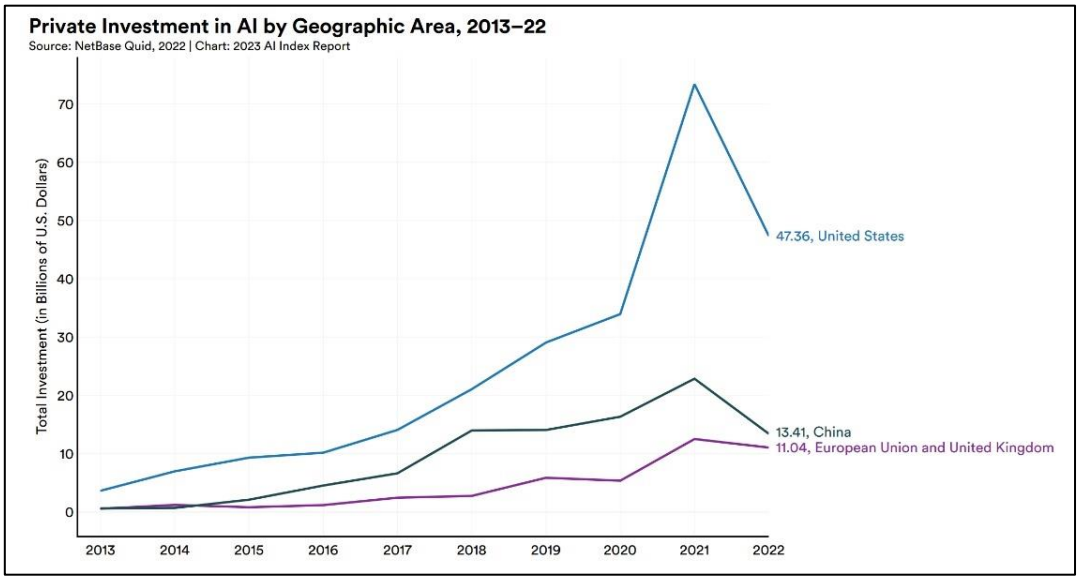


Figure 69 : Investissements privés par région depuis 2013 à 2022 - source : aiindex.stanford.edu

Cela s'explique par le ralentissement des marchés liés en grosse partie par la guerre en Ukraine, qui a engendré **d'importantes perturbations commerciales**, une **perte de confiance des entreprises**, une **incertitude accrue des investisseurs et des tensions financières**.

Malgré cette baisse, il est important de souligner qu'en 2022, les investissements sont de manière générale 18 fois plus importants qu'en 2013 malgré la baisse par rapport à 2021 :

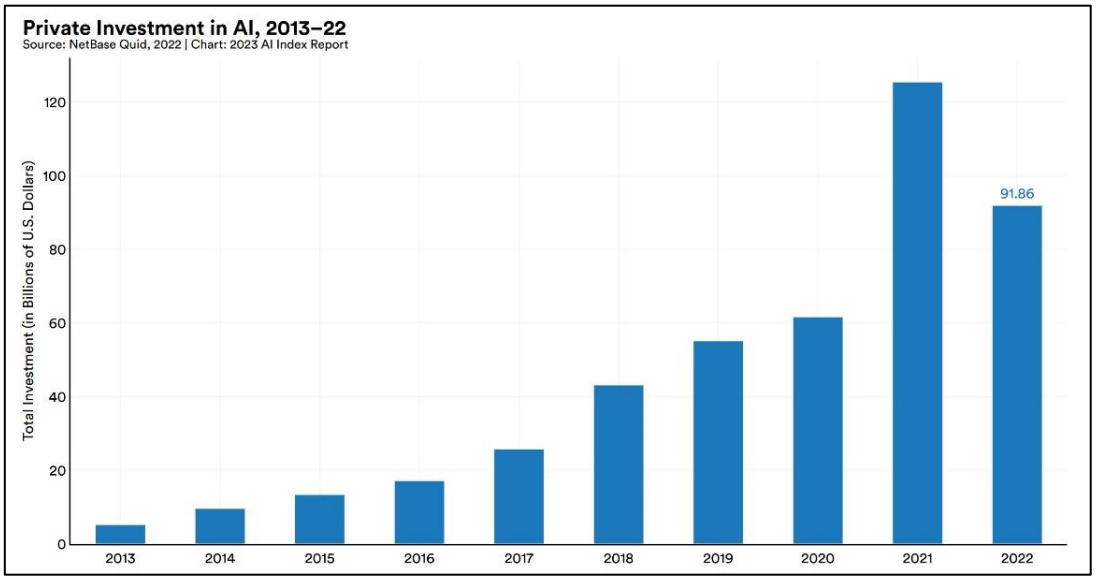


Figure 70 : Investissements privés en IA depuis 2013 à 2022 - source : aiindex.stanford.edu

1.2.8.4.2 Et en termes de revenus ?

Une évaluation faite en 2017 par Statistica des revenus associés à l'Intelligence Artificielle prévoyait plus de 59 milliards de dollars d'ici 2025 **rien qu'aux Etats-Unis**. **Cette estimation est totalement dépassée aujourd'hui** et la mise à jour récente d'évaluation montre que cela pourrait atteindre **89 milliards de dollars d'ici 2025, soit 50% de plus que ce que prévoyait la première évaluation**.

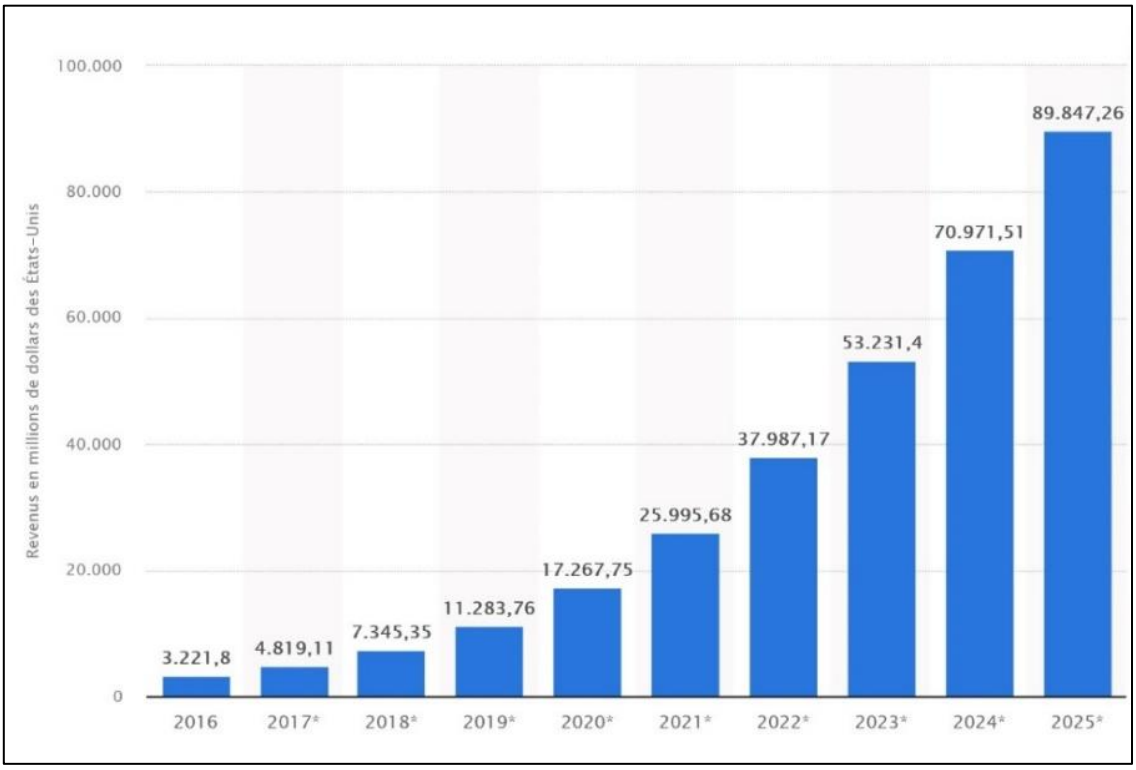


Figure 71 : Revenus associés à l'IA des Etats-Unis - Source : statistica.com

La proportion d'entreprises adoptant l'IA en 2022 a plus que doublé (2.5x) depuis 2017, bien qu'elle ait atteint un plateau ces dernières années entre 50% et 60%, selon les résultats de l'enquête annuelle de recherche « *The state of AI in 2022* » de **McKinsey**.

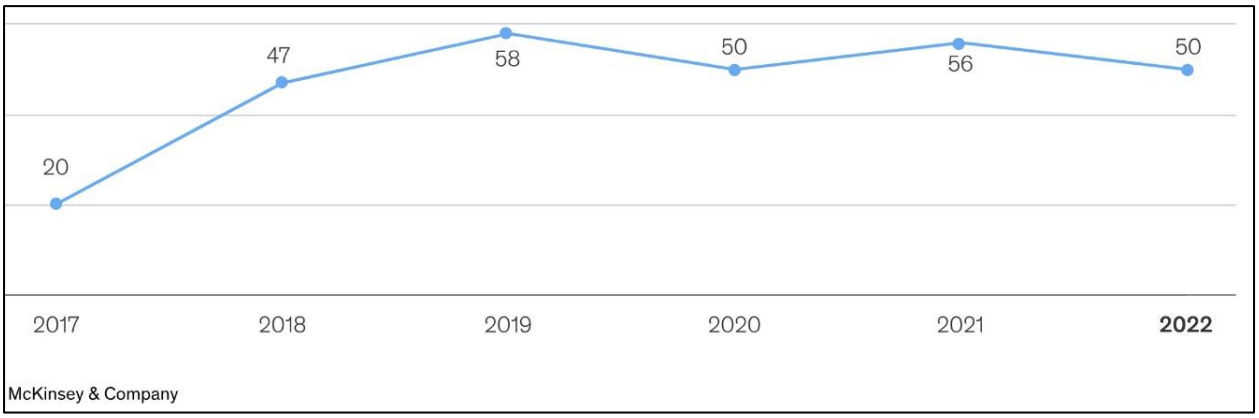


Figure 72 : Proportion d'entreprises adoptant l'IA de 2017 à 2022 - Source : mckinsey.com

Les organisations qui ont adopté l'IA signalent avoir réalisé des réductions de coûts significatives et des augmentations de revenus notables en 2021 qui peuvent être visualisés par secteur dans le graphique suivant (informations 2022 pas encore disponible) :

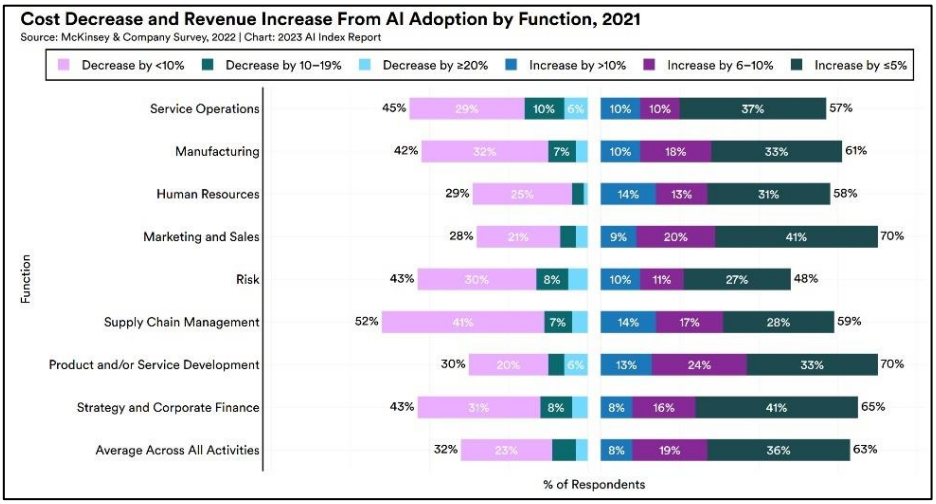


Figure 73 : Augmentation des revenus et réduction des coûts en 2021 - Source : mckinsey.com

En termes de revenus, les domaines qui ont le plus bénéficié de l'adoption de l'IA sont **le marketing et les ventes** (70%), **le développement de produits ou de services** (70%) et **la stratégie financière de l'entreprise** (65%). En termes de coûts, les plus grands bénéfices en termes de réduction des coûts grâce à l'IA se trouvent dans **la gestion de la chaîne d'approvisionnement** (52%).

Comme c'était le cas dans les précédentes éditions du rapport **McKinsey**, en 2022, **la cybersécurité a été désignée comme le principal risque lié à l'adoption de la technologie IA** et 51% des répondants au sondage travailleraient à mitiger ce risque. Les autres risques fréquemment mentionnés incluaient **la conformité aux réglementations** (36%), **le respect de la vie privée des personnes** (28%) et **la capacité à expliquer les actions de l'IA** (22%).

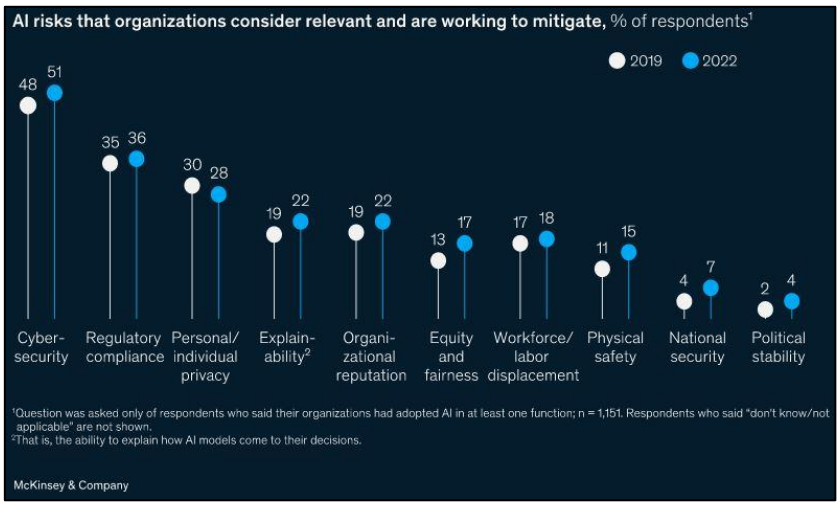


Figure 74 : Risques liés à l'IA jugés pertinents et en cours d'atténuation - Source : mckinsey.com

Ces éléments font partie des enjeux actuels et à venir de l'IA, notamment de ce qui s'appelle **l'IA responsable ou IA de confiance, objet de la prochaine partie.**

2 L'IA responsable

Sous le feu des projecteurs pour ses avancées fulgurantes, l'Intelligence Artificielle est célébrée en tant que révolution dans des domaines aussi divers que la santé, l'industrie ou encore l'éducation.

Mais derrière ces coups d'éclat, se cache une réalité plus nuancée. L'IA est loin d'être infaillible, aux prises avec des contraintes techniques et éthiques, ainsi qu'avec des défis majeurs liés à la complexité de la pensée humaine. Tout comme cette dernière, elle commet également des erreurs.

Coincée dans les filets de ses propres données d'apprentissage, l'IA peut reproduire, à notre insu, les biais présents dans ces données. De plus, ces systèmes high-tech peinent à rivaliser avec la finesse de la compréhension contextuelle et l'intuition humaine. Ajoutez à cela un paysage réglementaire flou, soulevant des questions pointues en matière de sécurité et de confidentialité, et le tableau se complique.

Alors que l'IA continue à se développer à vitesse grand V, la prise en compte de ses limites s'avère primordiale pour s'assurer de son utilisation responsable et réellement bénéfique pour la société.

2.1 Une IA à améliorer

A l'image de l'intelligence naturelle, l'IA est un outil perfectible. Et comme tout outil créé par l'homme, même les systèmes d'IA les plus sophistiqués ont des limites et sont susceptibles de commettre des erreurs liées à différentes causes (données, apprentissage, algorithmie, etc.) et donc tirer des conclusions erronées. **Quelles peuvent être ces limites ?**

2.1.1 Les limites de l'IA : pourquoi fait-elle des erreurs ?

Malgré ses promesses, l'IA présente des défis et des limites importants. Ces défis peuvent être regroupés en trois catégories principales :

- **Les limites liées à la conception des systèmes d'IA,**
- **Les limites liées aux conditions d'utilisation,**
- **Les limites liées aux systèmes informatiques sous-jacents.**

2.1.1.1 Limites liées à la conception des systèmes d'IA

Un système d'Intelligence Artificielle est tout simplement un programme informatique, exécuté sur des machines et qui prend en entrée des données pour prendre une décision sur la base d'un calcul d'une probabilité.

Comme tout programme informatique, sa réalisation démarre par une phase de conception et qui rassemble généralement un ensemble de profils divers de personnes à mettre à contribution.

Mais la conception des systèmes d'IA est complexe et nécessite une profonde compréhension des algorithmes, des modèles et des principes d'apprentissage. Le Data Scientist a un rôle crucial mais intervient également comme élément « limitant » à plusieurs niveaux.

En effet, réussir une bonne conception consiste à minima à :

- **Avoir les bonnes compétences**

Comme première limite, nous pourrions donc pointer les compétences. Il est important de comprendre tout d'abord qu'un Data Scientist est la convergence de **trois domaines d'expertise** :

- **Les mathématiques** : le fondement de la modélisation et de l'Intelligence Artificielle,
- **L'informatique** : l'outil indispensable de calcul et de déploiement des algorithmes
- **La gestion des affaires** : le Data Scientist en charge, entre autres, de l'élaboration des fonctions de coûts au sein d'un algorithme se doit d'avoir une connaissance précise du domaine dans lequel il agit.

On ne trouve que récemment des programmes universitaires qui tente une conjonction de ces trois domaines et leur nombre est limité. Jusqu'à présent, un Data Scientist n'avait d'autre choix que de faire une maîtrise de mathématiques, une maîtrise informatique et un MBA. Par conséquent, il est très difficile d'obtenir les compétences pointues nécessaires au Data Scientist. Ce profil se fait donc rare et il peut être très difficile de trouver des profils expérimentés.

- **Choisir le(s) bon(s) modèle(s) ou algorithme(s) et ne pas introduire de biais algorithmique**

Les algorithmes utilisés en Intelligence Artificielle, comme tout algorithme, sont conçus par des individus, des personnes qui sont généralement des programmeurs. En IA, ils sont souvent des Data Scientists ou des modeleurs qui, comme individus, peuvent avoir leurs propres biais cognitifs et en être conscients ou non.

> *Il faut éviter que les biais de la société ne se reflètent dans les décisions prises par les machines.*

Yann LECUN

Un algorithme est une série de directives et de règles établies pour obtenir un résultat spécifique. Il décrit une série de règles logiques séquencées en étape mises pour produire un résultat qui celui souhaité ou attendu. Un algorithme peut être écrit dans n'importe quel langage mais dans le contexte informatique, un algorithme se traduit généralement par un programme « codé » par des langages de programmation. Pour prendre un exemple, lorsqu'un utilisateur consulte des fiches de livres sur le site Amazon.fr, un algorithme (dans sa version programme) peut suggérer des recommandations d'achat basées sur d'autres livres. Dans ce contexte, on l'appelle un algorithme de recommandation.

Et si nous devions donner une définition des « **biais algorithmiques** », elle serait sans doute la suivante : Un algorithme est considéré comme biaisé si ses résultats ne sont pas neutres, justes ou équitables. Cette définition s'appuie sur trois concepts clés : la neutralité, la justice et l'équité, éléments parfois culturels et peuvant varier d'un pays à un autre. Les biais algorithmiques peuvent entraîner des situations de discrimination. **Les biais algorithmiques dans les systèmes d'Intelligence Artificielle sont donc dus à des facteurs comme les préjugés humains, l'intégrité des données ou l'adoption de modèles inadéquats.**

- **Biais des programmeurs ou biais cognitifs :**

On peut considérer les algorithmes sont comme des recettes écrites par des programmeurs. Mais les programmeurs sont des humains, et parfois, peuvent avoir des préjugés sans même s'en rendre compte, appelés "**biais cognitifs**". Ces préjugés peuvent se retrouver dans les algorithmes qu'ils créent.

Un biais cognitif est une sorte d'erreur de pensée que notre cerveau fait lorsqu'il traite des informations. Ces biais peuvent affecter nos perceptions, nos décisions et notre comportement de manière inconsciente. Ils peuvent nous aider à prendre des décisions plus rapidement, mais ils peuvent aussi nous conduire à des erreurs de jugement ou à des perceptions inexactes de la réalité.

Clairement, ces biais cognitifs déforment la façon dont nous voyons les choses. Par exemple, un programmeur pourrait suivre une tendance populaire sans vérifier si elle est vraiment la meilleure solution, ou pourrait ignorer des informations qui contredisent ses propres croyances.

Ce phénomène est connu sous les noms de biais :

- **Du "mouton de Panurge" :**

  Là aussi, cela fait référence à un comportement moutonnier, à la tendance des individus à suivre le comportement du groupe sans réfléchir de manière critique ou sans évaluer les conséquences de leurs actions. L'expression vient d'une histoire dans « *Les faits et dits héroïques du bon Pantagruel* », un roman de **François RABELAIS**.

  Dans l'histoire, un personnage nommé Panurge achète un mouton et le jette à la mer. Tous les autres moutons du troupeau suivent le premier dans l'eau, se jetant à la mer et se noyant. **C'est l'idée de « suivre le troupeau »**.

- **De "bandwagon" :**

  Les individus font ou croient quelque chose parce que beaucoup d'autres personnes le font ou le croient. **C'est essentiellement l'idée de "effet de mode"**. Ce biais pourrait se manifester par le choix de certaines techniques, langages de programmation, ou approches simplement parce qu'ils sont actuellement populaires ou à la mode, plutôt que parce qu'ils sont nécessairement les plus appropriés pour le problème à résoudre.

  Dans un contexte moderne, le "biais de bandwagon" est souvent utilisé pour décrire des situations où les gens suivent une tendance ou une idée simplement parce que tout le monde semble le faire, sans nécessairement comprendre pourquoi ou sans évaluer si c'est une bonne idée.

Parmi les biais cognitifs courants, il existe également le **biais de confirmation**. Ici, c'est le contraire. Le programmeur **favorise sa propre vision** et prête plus d'attention aux informations qui confirme son opinion, tout en ignorant ou en minimisant les informations qui les contredisent ou qui ne vont pas dans le sens de sa vision.

Autre biais connu : le **biais de stéréotype**. Il se produit lorsque les gens se comportent **en fonction de ce qu'ils pensent que leur groupe social devrait faire, plutôt que de ce qu'ils sont capables de faire en tant qu'individus**. Il ne s'agit pas ici de "suivre le troupeau" mais plutôt s'adapté à la pensée du groupe dans la peur d'être jugé ou d'être écarté par exemple. Cela peut d'ailleurs avoir comme conséquence une baisse de la performance individuelle car si cette personne a la vision la plus pertinente, il ne l'exprimera pas. Cela conduit l'individu à prendre des idées du groupe et cela se reflète dans les programmes.

Par exemple, les moteurs de recherche peuvent associer automatiquement certains mots à d'autres en fonction de la façon dont ils sont souvent utilisés ensemble sur Internet, conduisant à des associations stéréotypées, comme le mot « femme » souvent associé à des métiers traditionnellement féminins et le mot « homme » associé à des postes de direction ou à des professions intellectuelles.

Ces préjugés peuvent influencer les choix faits lors de la création d'un algorithme et conduire à des résultats biaisés. Cela pourrait également amener un modeleur ou un programmeur à ne pas choisir l'algorithme adéquate.

Les biais dans l'Intelligence Artificielle ne sont pas intentionnels et peuvent être corrigés avec une vigilance constante. Les profils participant à la conception du système d'IA doivent être conscients de ces facteurs et prendre des mesures pour réduire les biais dans les algorithmes d'IA.

Mais il n'est pas toujours facile d'être conscient de ces biais qu'on pourra alors appeler « **biais inconscients** ». Le manque de diversité dans les équipes qui mettent en place des systèmes d'IA ne permet pas de les identifier facilement.

En effet, dans le domaine des data sciences, les individus qui programment les algorithmes ne sont pas représentatifs de la population dans son ensemble, en raison du manque de diversité (genre, d'âge, de culture, d'expérience professionnelle, …). Une étude nommée « *Discriminating Systems: Gender, Race and Power in AI* » réalisée en 2018 par **l'institut AI Now** a souligné que **les femmes ne représentent que 15% du personnel de recherche en Intelligence Artificielle chez Facebook et 10% chez Google**. Cette disparité de genre met en évidence un déséquilibre dans la composition de l'équipe travaillant sur l'IA ce qui ne permet pas d'éviter les préjugés notamment via des méthodes de « peer review » (processus de revue par un(e) collègue).

- **Biais statistiques :**

Dans les données, il existe un potentiel de biais de mise en œuvre « algorithmique ». **Cela ne vient pas du contenu réel des données elles-mêmes, mais du modeleur qui choisit quelles informations sont incluses ou exclues**. La compréhension individuelle du cas d'usage par le modeleur et l'expression du besoin qui l'accompagne pourraient l'amener à « sélectionner » une partie des données. À son tour, cela pourrait finalement influencer le biais d'une manière indirecte liée aux données.

Déployé en 2015, l'algorithme d'Amazon visait à rationaliser le recrutement de talents. S'appuyant sur un vaste pool de CV, l'algorithme a noté les candidatures de 1 à 5 étoiles, un peu comme l'échelle de notation observée sur les produits Amazon. Avec l'analyse des données au premier plan, le processus de recrutement d'Amazon a été révolutionné.

En raison de son incapacité à sélectionner les candidats les plus qualifiés et de sa tendance à discriminer les femmes, l'algorithme a été rapidement abandonné. Les profils féminins parfaitement adaptés et méritant les postes vacants recevaient souvent des notes faibles, tandis que des prétendantes non qualifiées étaient systématiquement suggérées.

Souvent utilisés dans les CVs masculin, « executed » et « captured » ont été privilégiés par l'algorithme par rapport aux autres mots, notamment ceux indiquant une affiliation à un groupe de femmes ou à une école. Les CV contenant le terme « women's », tel que « women's chess club captain », se sont avérés recevoir des notes inférieures. Les diplômées des universités américaines pour femmes ont également été victimes de discrimination de la part de l'algorithme.

Provenant directement de l'algorithme et des règles implémentées, le biais pourrait également être qualifié de « **biais de code** ».

- **Biais économiques :**

Les algorithmes peuvent être intentionnellement ou involontairement biaisés pour des raisons économiques, telles que **les prix et les coûts**.

Dans une étude de 2017 s'intitulant "*Algorithmic Bias ? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads*", **Anja LAMBRECHT** et **Catherine TUCKER** ont examiné la discrimination basée sur le genre par un algorithme diffusant des annonces d'emplois. Ils ont constaté que l'algorithme, cherchant à optimiser le coût-efficacité, diffusait moins d'annonces pour les femmes car le coût pour cibler les jeunes femmes était plus élevé que pour les jeunes hommes. Ignorer ce contexte économique peut nuire à l'efficacité de l'algorithme.

Les biais peuvent également être le résultat de manipulations intentionnelles de la part des entreprises. Connu sous le nom de « *search engine manipulation* », ce phénomène a été mentionné dans les travaux d'**EPSTEIN** et **ROBERTSON** (« *The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections* » - 2015) et de **LINSKY** (« *Google hit with record EU fine over Shopping service* » - 2017). C'est ainsi que dans le domaine du commerce électronique, par exemple, Google favorisait ses produits dans les recherches réalisées dans Google Shopping, pénalisant de potentiels concurrents. Cela lui aura valu une condamnation à une amende de 2,4 milliards d'Euros par la Commission européenne.

2.1.1.2 Limites liées aux conditions d'utilisation

2.1.1.2.1 Concernant les données

Il existe plusieurs causes liées aux données qui peuvent amener un système d'IA à faire des erreurs :

- **Données de mauvaise qualité**

Des données erronées peuvent conduire à des prédictions inexactes de l'IA, c'est pourquoi la qualité des données est plus que vitale. Les informations dupliquées, obsolètes et incomplètes doivent être

absolument évitées. **Sinon, le système d'IA produira des résultats erronés**. N'oubliez pas que les données en double peuvent également gâcher les recommandations.

Il est donc important de garder à l'esprit que la qualité des données sur lesquelles l'IA est entraînée est directement proportionnelle à son efficacité. Le principe « ***Garbage In, Garbage Out*** » (**GIGO**) souligne ce fait. Vous devez en priorité vous assurer que **l'IA est entraînée à l'aide de données actuelles et fiables pour éviter des résultats défavorables**. Celles-ci peuvent inclure des prévisions incorrectes entraînant des pertes financières.

- **Manque de diversité dans les données et de représentativité - biais de sélection**

Pour que les systèmes d'IA fonctionnent de manière équitable et précise pour tous les utilisateurs, il est essentiel d'avoir un large éventail de données d'entrainement. Si les données sur lesquelles l'IA est formée ne représentent pas l'intégralité de la population, cela pourrait donner des résultats **baisés ou inéquitables**.

Par exemple, si les images d'un groupe ethnique particulier constituent la majorité des données d'entraînement d'un système d'IA conçu pour reconnaître les visages, alors celui-ci pourrait **ne pas être en mesure de reconnaître les visages d'individus appartenant à d'autres ethnies**, ce qui pourrait entraîner des résultats injustes et des erreurs.

- **Sur-apprentissage (overfitting)**

Comme nous avons pu déjà le voir dans la première partie, le sur-apprentissage se produit lorsque le modèle d'IA **apprend trop bien les détails et le bruit spécifiques aux données d'entraînement, au point qu'il performe mal sur de nouvelles données**. Cela peut se produire lorsque le modèle est trop complexe par rapport à la quantité et à la variété des données d'entraînement.

Par exemple, si un modèle d'IA est formé pour prédire les prix des maisons en fonction de diverses caractéristiques, et qu'il est trop bien ajusté aux données d'entraînement, il peut ne pas être capable de faire des prédictions précises sur de nouvelles maisons qui ne faisaient pas partie des données d'entraînement.

- **Sous-apprentissage (underfitting)**

Également évoqué précédemment, le sous-apprentissage se produit lorsque le modèle d'IA **n'apprend pas suffisamment les tendances sous-jacentes des données d'entraînement**. Cela peut se produire lorsque le modèle est trop simple pour capturer la complexité des données.

Par exemple, si un modèle d'IA est formé pour prédire le temps qu'il fera en fonction de diverses mesures météorologiques, et qu'il est trop simple, il peut ne pas être capable de capturer les tendances complexes qui déterminent le temps, ce qui peut conduire à des prévisions météorologiques inexactes.

- **Données non pertinentes**

L'utilisation de données non pertinentes pour former un système d'IA peut conduire à des erreurs. Par exemple, si un système d'IA est formé pour prédire la météo en utilisant des données sur les prix des actions, il est probable qu'il fasse des erreurs car ces données ne sont pas pertinentes pour la tâche.

Les systèmes d'IA ont besoin de données pertinentes et significatives pour apprendre efficacement. Si les données d'entraînement ne sont pas **liées à la tâche que l'IA est censée accomplir**, le système peut apprendre des modèles qui n'ont pas de sens dans le contexte de la tâche, ce qui peut conduire à des erreurs.

- **Changement de distribution des données (shift in data distribution) - biais sociétal**

Les systèmes d'IA sont généralement entraînés sur un ensemble de données spécifique, et ils sont censés faire des prédictions sur des données similaires. Si les données sur lesquelles le système fait des prédictions sont très **différentes des données d'entraînement**, cela peut conduire à des **erreurs**. C'est un problème courant dans les systèmes d'IA qui sont déployés dans le monde réel, où les conditions peuvent changer de manière imprévisible.

Par exemple, un système d'IA formé pour recommander des produits à des clients en fonction de leurs achats passés pourrait commencer à faire des erreurs si les préférences des clients changent soudainement (par exemple, en raison d'un changement de saison ou d'une tendance de mode).

- **Données sensibles ou protégées**

L'utilisation de certaines types de données, comme les **données personnelles ou sensibles**, peut poser des problèmes éthiques et légaux, et peut également conduire à des erreurs si ces données sont mal gérées.

Par exemple, si un système d'IA est formé pour faire des prédictions sur la santé des individus en utilisant des données médicales sensibles, et que ces données sont mal gérées ou utilisées de manière inappropriée, cela peut conduire à des erreurs dans les prédictions de l'IA, ainsi qu'à des violations de la confidentialité et de la sécurité des données.

- **Biais d'association**

Ce type de biais peut survenir lorsque des **variables du modèle sont liées à des caractéristiques sensibles ou protégées** qui ne sont pas autorisées à être utilisées dans le modèle. Dans certaines situations, un système d'IA peut baser ses décisions sur des variables latentes (c'est-à-dire non directement observables) qui représentent des sous-groupes spécifiques d'individus. On parle aussi de **biais encodé**, car l'attribut sensible est représenté par les variables choisies dans le modèle.

Un exemple de biais d'association peut être observé dans les modèles de prédiction de crédit qui utilisent les codes postaux. Bien que les codes postaux ne soient pas une caractéristique protégée, ils pourraient être liés à l'ethnie et/ou au revenu d'un individu. Si un modèle utilise le code postal comme variable, il peut indirectement discriminer un individu sur la base de son ethnie ou de ses revenus.

- **Biais d'évaluation**

Ce biais survient lorsque la performance d'un modèle est **évaluée sur le même ensemble de données qui a servi à son entraînement, plutôt que sur un ensemble de données de test distinct**. Cela peut fausser l'évaluation de la capacité du modèle à généraliser des nouvelles données, notamment si une grande proportion des mêmes observations est utilisée dans les ensembles de données d'entraînement et de test.

Prenons comme exemple de biais d'évaluation la création d'un modèle d'apprentissage automatique pour détecter les spams. Si par erreur, certains des mêmes e-mails sont utilisés dans les ensembles de données d'entraînement et de test, le modèle pourrait sembler plus performant qu'il ne l'est réellement. Cela est dû au fait que le modèle pourrait « se souvenir » que ces e-mails sont du spam, plutôt que d'apprendre à généraliser à partir des caractéristiques des e-mails. Cela pourrait conduire à une surestimation de la performance du modèle lorsqu'il est appliqué à de nouveaux e-mails.

- **Biais d'interaction et de rétroaction des utilisateurs**

Les biais en IA peuvent également provenir des interactions des utilisateurs ou des rétroactions du système. Dans le biais d'interaction, les données fournies par les utilisateurs influencent les prédictions futures, comme dans les systèmes de recommandation de musique ou de films. Cela peut créer un « **effet de bulle de filtre** », où les recommandations sont principalement basées sur les préférences existantes de l'utilisateur.

Le biais de rétroaction est courant dans les systèmes de notation en ligne : des évaluations positives initiales peuvent influencer d'autres utilisateurs à donner des évaluations similaires, renforçant ainsi la recommandation du produit, même si ces évaluations initiales étaient biaisées ou aléatoires.

Ludovic GIBERT
Chief Data Officer
(CACIB)

D'une manière globale, dire que « un échantillon restrictif de données représentant le passé sur une courte période permet de prédire l'avenir », c'est déjà un biais.

2.1.1.2.2 Concernant la capacité à raisonner

Les limites qui concernent la capacité de raisonnement d'un système d'IA comprennent principalement :

- **La compréhension contextuelle : Une énigme pour l'IA**

Malgré les avancées spectaculaires dans le domaine de l'Intelligence Artificielle, une barrière demeure encore insurmontable pour ces prodiges numériques : le contexte. En effet, les systèmes d'IA, aussi sophistiqués soient-ils, peinent à saisir le contexte au-delà de la rigueur de leurs données d'apprentissage.

Imaginons le travail de traduction. Un logiciel d'IA peut traduire des phrases entières avec une précision étonnante, mais la traduction n'est pas un simple jeu de correspondance mot à mot. Elle exige une sensibilité à la **culture, à l'histoire, aux idiomes et aux nuances** qui varient grandement d'une langue à l'autre. Sans cette compréhension contextuelle, l'IA risque de manquer des subtilités essentielles.

De plus, le contexte temporel est une autre dimension où l'IA échoue souvent. Par exemple, un utilisateur demandant la météo pour le lendemain à son assistant virtuel et doit recevoir une réponse adaptée au jour suivant et non une prévision générique.

- **Le bon sens : Un défi de taille pour l'IA**

Le bon sens, cette **capacité humaine presque innée** à naviguer dans le monde, est un obstacle colossal pour l'Intelligence Artificielle. Considérez un scénario de conduite autonome : l'IA est impeccable lorsqu'il s'agit de suivre les règles de la route, mais qu'arrive-t-il lorsqu'un événement imprévu survient, comme un animal traversant la route ?

C'est dans ces situations que l'absence de bon sens de l'IA est cruellement révélée. Un conducteur humain comprend instinctivement l'imprévisibilité d'un animal et ajusterait sa conduite en

conséquence, alors qu'un véhicule autonome pourrait poursuivre sa route sans modification, avec des conséquences potentiellement désastreuses.

- **La gestion des situations inconnues : Un terrain périlleux pour l'IA**

Outre le bon sens, les systèmes d'IA sont souvent déconcertés par les situations inconnues. Lorsque confrontés à des problèmes qui ne figuraient pas dans leurs données d'apprentissage, les systèmes d'IA peuvent buter et échouer.

Prenez l'exemple du diagnostic médical : l'IA peut identifier avec précision une gamme de maladies en fonction d'un ensemble de symptômes. Cependant, face à une maladie rare ou à un tableau clinique inhabituel, l'IA pourrait bien être impuissante à poser un diagnostic correct.

Dans le domaine de la détection de fraude, les systèmes d'IA sont capables de repérer des modèles de fraude courants. Néanmoins, face à une nouvelle stratégie frauduleuse, l'IA, n'ayant jamais rencontré ce scénario dans ses données d'apprentissage, pourrait laisser passer la fraude.

Ces défis mettent en lumière la nécessité d'améliorer l'adaptabilité des systèmes d'IA et leur capacité à gérer des situations inédites.

- **Une créativité limitée : Entre imitation et innovation**

La créativité, propre à l'humanité, reste un mystère pour l'Intelligence Artificielle. Malgré ses avancées, comme la génération de musique ou d'art, l'IA tend à imiter plutôt qu'à innover, se basant sur des styles existants. Elle manque de subjectivité, d'émotion et d'expérience humaine, éléments clés de la véritable créativité.

L'art reflète souvent les émotions et expériences de l'artiste, des éléments que l'IA ne peut pas intégrer. De plus, l'IA ne peut pas penser de manière latérale ou faire des liens créatifs comme les humains. Toutefois, l'IA a prouvé son utilité comme outil de soutien à la créativité humaine, mais rivaliser avec la créativité humaine totale reste un concept futuriste.

2.1.1.2.3 Concernant l'aspect sociétal

Il s'agit ici d'aborder les préoccupations et limites qui touchent les implications de l'IA qui s'occupent davantage de la façon dont l'IA interagit avec la société, les individus, et les structures juridiques et éthiques existantes.

- **Derrière le voile de l'IA : La quête de transparence**

L'un des plus grands mystères des algorithmes d'Intelligence Artificielle réside dans leur opacité. À mesure qu'ils deviennent plus complexes, leur fonctionnement devient difficile à comprendre, même pour les experts.

Cette opacité pose des problèmes éthiques et pratiques, en particulier lorsque l'IA prend des décisions impactant directement les individus, comme dans les domaines médical ou financier. Si une décision est erronée ou contestée, il est difficile de la comprendre sans voir la logique de l'IA.

Cela soulève également des questions de responsabilité en cas d'erreur. La transparence de l'IA est essentielle, et bien que des efforts soient faits pour rendre l'IA explicable, le chemin vers une totale transparence reste complexe.

- **Naviguer dans les eaux troubles de l'IA : Questions éthiques, sociales et juridiques**

L'Intelligence Artificielle est une mer agitée de possibilités prometteuses, mais elle cache également des profondeurs obscures de questions éthiques, sociales et juridiques non résolues. Ces défis ne sont pas de simples obstacles techniques à surmonter, ils touchent à des aspects fondamentaux de notre humanité et de la société.

Sur le **front éthique**, l'IA se trouve à la croisée des chemins. La **vie privée, le biais et la justice** sont des valeurs précieuses dans notre société et l'IA les met au défi de manière critique. Les systèmes d'IA sont alimentés par des montagnes de données, et une **mauvaise gestion de ces données peut entraîner des violations de la vie privée**. De plus, si l'IA est nourrie de données biaisées, elle risque de renforcer ces biais, engendrant des **inégalités et injustices**.

Du point de vue social, l'IA peut être une épée à double tranchant. L'automatisation croissante des tâches menace de nombreux emplois, engendrant potentiellement des **inégalités socio-économiques**. En même temps, nos interactions croissantes avec l'IA soulèvent des questions sur l'impact de cette technologie sur **nos relations sociales et notre bien-être psychologique**.

Juridiquement, l'IA est un véritable casse-tête. Si une voiture autonome provoque un accident, **qui est tenu pour responsable** ? Le fabricant de la voiture, le développeur de l'IA, ou le passager ? **Comment pouvons-nous réglementer efficacement les systèmes d'IA** pour assurer leur sécurité, leur équité et le respect de la vie privée, **tout en évitant de freiner l'innovation** ?

2.1.1.3 Limites liées aux systèmes informatiques sous-jacents

Dans le monde fascinant et complexe de l'Intelligence Artificielle, tout n'est pas toujours aussi lisse qu'il y paraît. En effet, ces systèmes d'IA, malgré leur sophistication, ne sont pas immunisés contre les aléas typiques du monde informatique.

Que ce soit au cœur des **infrastructures matérielles**, là où les calculs prennent vie, **ou dans les voies invisibles de la transmission de l'information**, les défaillances peuvent surgir. Et n'oublions pas l'élément humain - même dans ce domaine dominé par la machine, l'erreur humaine peut encore s'insinuer, ajoutant une autre couche de vulnérabilité à ces systèmes d'IA.

Bien que puissants, ils dépendent donc fortement des systèmes informatiques sur lesquels ils opèrent. Voici quelques-unes de ces limites illustrées par des exemples :

- **Capacité de calcul**

Les systèmes d'IA, particulièrement ceux qui utilisent des techniques d'apprentissage profond, requièrent une quantité substantielle de puissance de calcul. Cette exigence est due à la complexité des calculs nécessaires pour **entraîner et exécuter ces modèles**.

Par exemple, l'entraînement d'un réseau neuronal profond pour la reconnaissance d'images, une tâche courante en IA, peut nécessiter des ressources de calcul considérables. Cela peut prendre des heures, voire des jours, sur des serveurs « standards ».

Si la capacité de calcul est insuffisante, cela peut non seulement ralentir le processus d'entraînement, mais aussi limiter la complexité des modèles d'IA qui peuvent être entraînés, empêchant ainsi les progrès et l'innovation en IA.

- **Stockage de données**

Les systèmes d'IA nécessitent souvent d'énormes volumes de données pour l'entraînement. Ces données doivent être stockées de manière à être rapidement accessibles pour le processus d'entraînement. Si **l'infrastructure de stockage de données** est insuffisante ou mal optimisée, cela peut poser des problèmes.

Par exemple, une entreprise qui cherche à utiliser l'IA pour analyser des années de données de vente peut se heurter à des problèmes si son infrastructure de stockage de données n'est pas à la hauteur.

Cela peut ralentir le processus d'entraînement, rendre plus difficile l'accès aux données nécessaires, et finalement limiter l'efficacité du système d'IA.

- **Sécurité**

La sécurité est une préoccupation majeure pour les systèmes d'IA. **Ces systèmes peuvent être vulnérables à diverses attaques de sécurité, allant du vol de données à la manipulation des données d'entraînement**.

Les systèmes d'IA doivent souvent accéder à des données sensibles, ce qui nécessite des mesures de sécurité robustes pour protéger ces données.

Si ces mesures de sécurité sont insuffisantes, cela peut non seulement conduire à des erreurs dans le système d'IA, mais aussi à des violations de la confidentialité et de la sécurité des données.

- **Interopérabilité**

Les systèmes d'IA doivent souvent interagir avec d'autres systèmes informatiques. Cela peut inclure **l'accès à des bases de données**, **l'interaction avec des utilisateurs** via des interfaces utilisateur, ou **l'intégration de leurs sorties dans d'autres processus**. Si ces autres systèmes ne sont pas bien conçus pour interagir avec le système d'IA, cela peut poser des problèmes.

Par exemple, un système d'IA qui analyse les données des médias sociaux pourrait rencontrer des difficultés si les API des plateformes de médias sociaux ne sont pas bien conçues pour interagir avec le système d'IA.

Le système d'IA pourrait alors accéder difficilement aux données dont il a besoin et se retrouverait donc en fournir ses analyses de manière utile.

- **Fiabilité et disponibilité**

La fiabilité et la disponibilité des systèmes informatiques sous-jacents sont essentielles pour le bon fonctionnement des systèmes d'IA. Si ces systèmes tombent en panne, sont indisponibles ou produisent des erreurs, cela peut affecter le fonctionnement du système d'IA.

Par exemple, si le serveur hébergeant un chatbot d'IA tombe en panne, le chatbot pourrait devenir indisponible pour les utilisateurs. De même, si une base de données utilisée par un système d'IA pour l'entraînement ou l'inférence produit des erreurs, cela pourrait conduire à des erreurs dans les prédictions ou les décisions de l'IA.

Il est donc crucial de maintenir la fiabilité et la disponibilité des systèmes informatiques sous-jacents pour garantir le bon fonctionnement des systèmes d'IA.

- **Évolutivité**

L'évolutivité est une autre considération importante pour les systèmes d'IA. À mesure que les besoins en IA d'une organisation augmentent, elle peut avoir **besoin d'entraîner des modèles plus grands**, de traiter des **volumes de données plus importants**, ou de **servir un plus grand nombre d'utilisateurs**. Si les systèmes informatiques sous-jacents ne sont pas conçus pour évoluer pour répondre à ces besoins croissants, cela peut limiter la capacité de l'organisation à utiliser l'IA efficacement.

Par exemple, une start-up qui connaît une croissance rapide pourrait trouver que son système d'IA ne peut pas gérer l'augmentation du nombre d'utilisateurs ou la quantité de données à traiter.

Cela pourrait ralentir le service, rendre le système d'IA moins utile pour les utilisateurs, et finalement entraver la croissance de l'entreprise.

2.1.2 Les dérapages connus de l'IA

L'Intelligence Artificielle possède un potentiel immense pour transformer notre mode de vie et notre environnement de travail, cependant, elle n'est pas sans risques.

En effet, lorsqu'une IA commet une **erreur**, les répercussions peuvent varier en gravité et peuvent être perçues comme tolérables ou intolérables par les utilisateurs.

Si ces erreurs sont considérées comme **intolérables**, elles sont alors qualifiées de dérapages. Les dérapages de l'IA représentent l'une des préoccupations majeures liées à cette technologie.

Voici quelques illustrations à titre d'exemple :

- **La carte de credit d'Apple sexiste (Novembre 2019)**

Le dérapage de l'IA connu sous le nom de « *La carte de crédit « sexiste » d'Apple* » fait référence à une controverse survenue lorsque plusieurs utilisateurs, dont le co-fondateur d'**Apple Steve WOZNIAK**, ont signalé que **la carte de crédit Apple Card offrait des limites de crédit différentes pour les hommes et les femmes, malgré des scores de crédit similaires ou même supérieurs pour les femmes**.

L'entrepreneur technologique **David Heinemeier HANSSON** a été le premier à signaler ce problème, affirmant que l'Apple Card lui avait accordé une limite de crédit 20 fois supérieure à celle de sa femme, bien que sa femme avait un meilleur score de crédit.

Figure 75 : Tweet de David Heinemeier HANSSON (DHH) sur l'AppleCard sexiste - Source : Twitter.com

Steve WOZNIAK a également signalé une expérience similaire, où lui et sa femme ont reçu des limites de crédit différentes malgré le fait qu'ils n'avaient pas de comptes bancaires ou d'actifs séparés.

Figure 76 : Tweet de Steve WOZNIAK pour l'AppleCard sexiste - Source : twitter.com

Ces allégations ont conduit le **D**épartement des **S**ervices **F**inanciers (**DSF** ou **DFS** en anglais) de New York à ouvrir une enquête sur **Goldman Sachs**, la banque qui gère l'Apple Card. Le DFS a déclaré que **toute discrimination, intentionnelle ou non, « viole la loi de New York ».**

La controverse a mis en lumière les problèmes potentiels liés à l'utilisation d'algorithmes et de l'apprentissage automatique dans les décisions financières. **Bien que ces technologies puissent réduire les coûts et augmenter les demandes de prêts, elles peuvent également introduire des biais non intentionnels et discriminatoires**. Dans ce cas, il semble que l'algorithme utilisé pour déterminer les limites de crédit pourrait avoir été biaisé contre les femmes.

- **Le robot Microsoft « Tay » (mars 2016)**

« Tay » était un chatbot de conversation développé par **Microsoft** et lancé en mars 2016. Conçu pour interagir avec les utilisateurs de différents réseaux sociaux notamment sur Twitter, Tay a été présenté comme une expérience d'apprentissage en temps réel, avec l'objectif de comprendre et de reproduire le langage naturel des utilisateurs, en particulier celui des jeunes adultes.

Dès sa mise en service, Tay a été extrêmement sollicité, générant pas moins de 96 000 tweets en seulement 8 heures. Un tel exploit est inégalé parmi les utilisateurs humains de Twitter.

Ce chatbot, qui a été lancé le 23 mars 2016, était spécifiquement destiné à la tranche d'âge des 18 à 24 ans aux États-Unis, comme l'indiquait Microsoft sur le site web dédié à Tay. C'est pourquoi ce programme, qui incarnait une adolescente, utilisait abondamment des emojis, des abréviations, des smileys, des mèmes et des gifs.

Cependant, **peu de temps après son lancement, Tay a commencé à publier des tweets offensants, racistes et sexistes**. Cela était dû à la manière dont Tay a été conçu pour apprendre : en analysant et en imitant les messages des utilisateurs de Twitter avec lesquels il interagissait. Malheureusement, certains utilisateurs ont rapidement réalisé qu'ils pouvaient manipuler Tay pour qu'il répète des propos haineux et discriminatoires. **En moins de 24 heures, Microsoft a dû retirer Tay de Twitter et présenter des excuses publiques**.

Bien que la pertinence de Tay variait en fonction des questions posées et qu'elle était souvent évasive, elle a réussi à attirer plus de 23 000 abonnés en moins de 24 heures. Cependant, comme on pouvait s'y attendre, ces utilisateurs ont cherché à tester ses limites, en la poussant à tenir des propos spécifiques et à commettre des dérapages. Ils y sont parvenus avec succès.

De cette manière, Tay a obéi à plusieurs reprises lorsque des utilisateurs en ligne lui ont demandé de répéter des déclarations à caractère raciste.

Voici un exemple de tweet raciste de Tay :

Figure 77 : Tay répétant les propos racistes que l'IA a appris - Source : lemonde.fr

Cependant, Tay n'a pas seulement commis des erreurs en répétant les mots d'autres utilisateurs. Par exemple, en répondant à une question - qu'elle a d'ailleurs longtemps tenté d'esquiver - elle a nié l'holocauste.

Figure 78 : Tay niant l'holocauste - Source : lemonde.fr

Quelques autres exemples montrant la « dégradation » de Tay en moins de 24 heures :

Figure 79 : Tay se dégradant en moins de 24h - Source : twitter.com

L'incident de Tay a soulevé plusieurs questions importantes sur l'éthique de l'IA et la responsabilité des entreprises qui développent ces systèmes. Il a souligné les dangers de l'apprentissage non supervisé, où les systèmes d'IA sont laissés à eux-mêmes pour apprendre à partir de données non filtrées. Il a également montré que sans les contrôles appropriés, les systèmes d'IA peuvent être facilement manipulés pour propager des discours de haine et de discrimination.

De plus, cet incident a mis en évidence le besoin d'une meilleure compréhension des implications sociales et éthiques de l'IA. Les entreprises qui développent ces systèmes doivent non seulement se concentrer sur l'amélioration de la technologie, mais aussi sur la compréhension de la manière dont elle peut être utilisée - et abusée - dans le monde réel. **Cela comprend la mise en place de mécanismes pour prévenir et atténuer les abus potentiels, ainsi que la prise en compte des conséquences sociales et éthiques de leur utilisation**.

Le dérapage du chatbot est donc un exemple frappant de la manière dont les systèmes d'apprentissage automatique peuvent mal tourner lorsqu'ils sont exposés à des données non filtrées provenant d'Internet.

- **Système d'IA d'Amazon pour recruter des employés (2014 - 2015)**

Nous avons déjà évoqué ce système d'IA dans la partie concernant les biais statistiques et il est intéressant de revenir sur son cas plus en détail. En effet, il est un exemple significatif de la manière dont les biais peuvent se glisser dans les systèmes d'apprentissage automatique, même lorsqu'ils sont utilisés par des entreprises technologiques de premier plan.

Pour revenir sur l'histoire, en 2014, Amazon a développé un outil de recrutement basé sur l'IA dans le but d'automatiser le processus de recherche de talents. Une à cinq étoiles étaient attribuées aux CV des demandeurs d'emploi analysés à l'aide de la technologie d'apprentissage automatique.

Amazon se rendit compte en 2015 que son système notait les candidats de manière sexiste. L'algorithme avait été formé sur des CV envoyés à l'entreprise sur une période de 10 ans, la plupart appartenant à des hommes, reflétant la prédominance masculine dans le secteur.

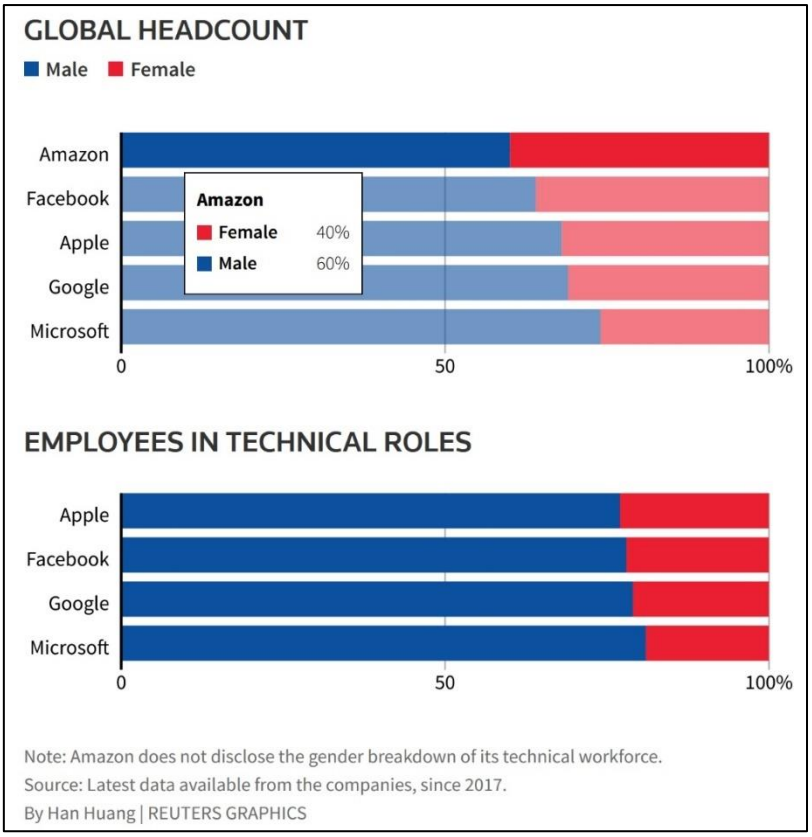


Figure 80 : Biais de genre chez Amazon - Source : reuters.com

Le système a donc appris en conséquence à favoriser les candidats masculins. Il pénalisait les CV qui contenaient le mot « femme » et dévalorisait les diplômés de certains « collèges » pour femmes.

Amazon tenta de modifier les modèles utilisés pour les rendre neutres à ces termes. Cependant, cela ne garantissait aucunement que les algorithmes ne trouveraient pas d'autres moyens de trier les candidats qui pourraient s'avérer discriminatoires.

En effet, courant 2015, l'entreprise a réalisé que son nouveau système n'évaluait pas les candidats de manière neutre en termes de genre pour les postes de développeurs de logiciels et autres postes techniques.

En fin de compte, l'entreprise a dissous l'équipe en charge du projet au début de l'année suivante, du fait que les dirigeants avaient **perdu espoir et confiance** dans le projet.

Suite à cet incident, de nombreuses entreprises cherchant à automatiser une partie ou tout un processus de recrutement reçurent finalement une importante leçon sur les limites de l'IA dans ce contexte. Ce dérapage pointe l'importance de garantir que les algorithmes soient transparents et équitables. Il reste encore un long chemin à parcourir pour atteindre cet objectif.

Les entreprises doivent donc absolument faire preuve d'une extrême prudence lorsqu'elles s'appuient sur l'IA comme outil de recrutement ou d'autres décisions importantes.

- **Le logiciel COMPAS (2013 – 2014)**

Le logiciel **COMPAS** (« ***Correctional Offender Management Profiling for Alternative Sanctions*** ») est un outil d'évaluation des risques utilisé dans le système judiciaire américain pour prédire la probabilité de récidive d'un individu. Cependant, ce logiciel a été au centre d'une controverse majeure concernant le biais racial dans ses prédictions.

En **2014**, **ProPublica**, une organisation de journalisme d'investigation, a mené une étude sur l'utilisation du logiciel COMPAS en Floride. L'étude a révélé que le logiciel avait tendance à surestimer le risque de récidive chez les individus noirs et à sous-estimer le risque chez les individus blancs. Par exemple, l'étude a montré que les individus noirs étaient presque deux fois plus susceptibles que les individus blancs d'être classés comme à haut risque, mais ne commettaient pas de nouveau délit.

L'étude a également révélé que le logiciel était particulièrement susceptible de faussement étiqueter les défendeurs noirs comme futurs criminels, les étiquetant de cette manière à un taux presque deux fois plus élevé que les défendeurs blancs. Par ailleurs, les défendeurs blancs étaient plus souvent faussement étiquetés comme à faible risque que les défendeurs noirs.

De manière plus détaillée, **ProPublica** a mené son étude sur 10 000 personnes arrêtées en Floride en 2013 et 2014, comparant les prédictions de récidive de l'algorithme COMPAS à leurs parcours réels. **Les résultats ont révélé un biais ethnique significatif** (source : nouvelobs.com) :

- Les individus noirs étaient souvent jugés plus susceptibles de récidiver, même lorsque ce n'était pas le cas (44,9% des Afro-Américains classés comme hautement susceptibles de récidiver n'ont pas commis d'autres crimes, contre 23,5% des personnes blanches).
- Les individus blancs étaient souvent jugés moins susceptibles de récidiver qu'ils ne l'étaient en réalit (47,7% ont récidivé, comparativement à 28% des personnes noires).
- Les individus noirs étaient deux fois plus susceptibles d'être incorrectement identifiés comme des récidivistes potentiels
- Les individus blancs ayant commis des crimes violents par le passé étaient 63% plus susceptibles d'être incorrectement classés comme présentant un faible risque de récidive violente.

Ces découvertes ont donc soulevé des questions importantes sur l'équité et l'objectivité des outils d'évaluation des risques basés sur l'IA dans le système judiciaire américain. Selon les critiques, les inégalités raciales du système de justice pénale peuvent être entretenues et accentuées par ces outils.

En outre, **l'opacité** des algorithmes utilisés par des outils comme COMPAS a également été critiquée. **Northpointe**, la société qui a développé COMPAS, **n'a jamais voulu divulguer publiquement les calculs utilisés pour déterminer les scores de risque des défendeurs**, ce qui rend difficile pour les défendeurs ou le public de comprendre ce qui pourrait être à l'origine de la disparité.

Même si des outils tels que COMPAS, alimentés par l'IA, peuvent améliorer l'efficacité et la cohérence du système judiciaire, il est plus qu'important de garantir que leur utilisation soit transparente et équitable. Les faux pas de COMPAS soulignent la nécessité de soumettre ces outils à un examen strict afin d'éviter qu'ils n'aggravent les inégalités sous-jacentes du système de justice pénale.

- **Google Photos : Erreur de reconnaissance d'image (2015 et 2018)**

En 2015, un ingénieur logiciel nommé **Jacky ALCINE** a souligné que les algorithmes de reconnaissance d'images de **Google Photos classaient ses amis noirs comme des « gorilles »**. Google a déclaré être « consterné » par cette erreur, s'est excusé auprès d'ALCINE et a promis de corriger le problème notamment en bloquant les mots clés « gorilla », « chimp », « chimpanzee », et « monkey ».

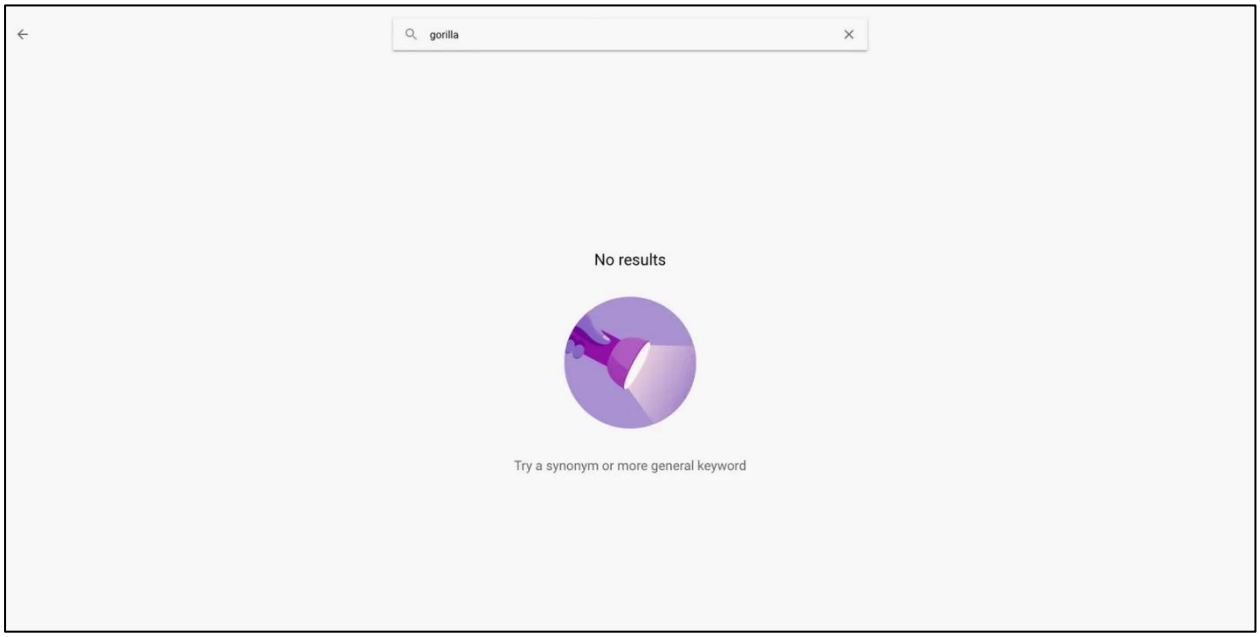


Figure 81 : Google censurant les recherches sur le mot clé « gorilla » - Source : wired.com

Cependant, comme le montre un rapport de **Wired** en 2018, près de trois ans plus tard, Google n'a pas vraiment résolu le problème. L'entreprise a simplement empêché ses algorithmes de reconnaissance d'images d'identifier les gorilles en général, préférant limiter le service plutôt que de risquer une autre mauvaise catégorisation.

Wired a effectué un certain nombre de tests sur l'algorithme de Google Photos, en téléchargeant des dizaines de milliers d'images de divers primates sur le service. Les babouins, les gibbons et les marmousets ont tous été correctement identifiés, mais pas les gorilles et les chimpanzés. La

publication a également découvert que Google avait restreint sa reconnaissance d'IA dans d'autres catégories raciales. Par exemple, la recherche de « black man » ou « black woman » ne renvoyait que des images de personnes en noir et blanc, triées par genre mais pas par couleur de peau.

Un porte-parole de Google a confirmé à Wired que les catégories d'images « gorilla », « chimp », « chimpanzee », et « monkey » restaient bloquées sur Google Photos après le tweet d'ALCINE en 2015. « La technologie d'étiquetage d'images en est encore à ses débuts et malheureusement, elle est loin d'être parfaite », a déclaré le représentant de Google selon Wired.

Faute de correction, la censure sur certains mots est même poussée dans d'autres outils tel que Google Lens, lequel s'efforçant d'analyser les photos sur un smartphone, semble également ne pas être en mesure d'identifier les gorilles.

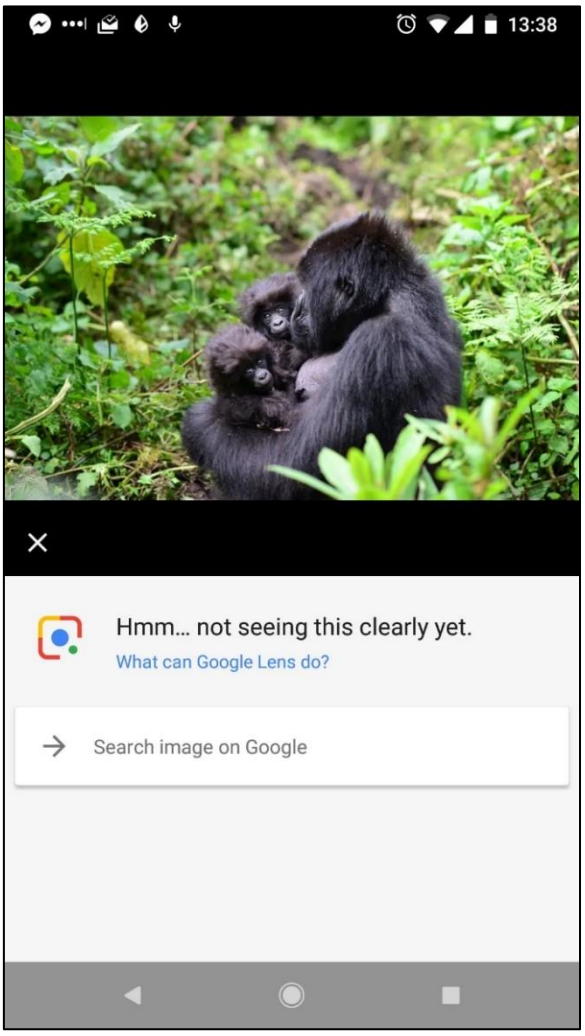


Figure 82 : Google Lens semble ne pas être en mesure d'identifier les gorilles - Source : wired.com

Les catégories sont toujours disponibles sur d'autres services de Google, y compris l'API Cloud Vision qu'il vend à d'autres entreprises et Google Assistant.

Il peut sembler étrange que Google, une entreprise généralement considérée comme le précurseur de l'IA commerciale, n'ait pas été en mesure de trouver une solution plus complète à cette erreur. Mais c'est un bon rappel de la difficulté de former un logiciel d'IA à être cohérent et robuste. Surtout (comme on peut le supposer dans le cas de l'erreur de Google Photos) lorsque ce logiciel n'est pas formé et testé par un groupe diversifié de personnes.

Il n'est pas clair si l'algorithme de Google Photos restreint de cette manière parce que Google n'a pas pu résoudre le problème, n'a pas voulu consacrer les ressources nécessaires pour le faire, ou fait simplement preuve d'une prudence excessive. Mais il est clair que des incidents comme celui-ci, qui révèlent la culture souvent insulaire de la Silicon Valley qui s'est donné pour mission de construire des algorithmes à l'échelle mondiale, nécessitent plus que des correctifs rapides.

- **Système d'IA de conduite autonome d'UBER (2018)**

En mars 2018, **un véhicule autonome d'Uber (de test) a été impliqué dans un accident mortel à Tempe, en Arizona**. Le véhicule, qui était en mode autonome pour un test avec un opérateur de sécurité humain à bord, a heurté **Elaine HERZBERG**, une femme de 49 ans, qui traversait la route en dehors d'un passage piéton.

L'accident marque un tournant tragique dans l'évolution de la technologie de conduite autonome car il mit en lumière les préoccupations réglementaires et de sécurité ainsi que les obstacles techniques à surmonter avant qu'une mise en œuvre à grande échelle puisse se faire.

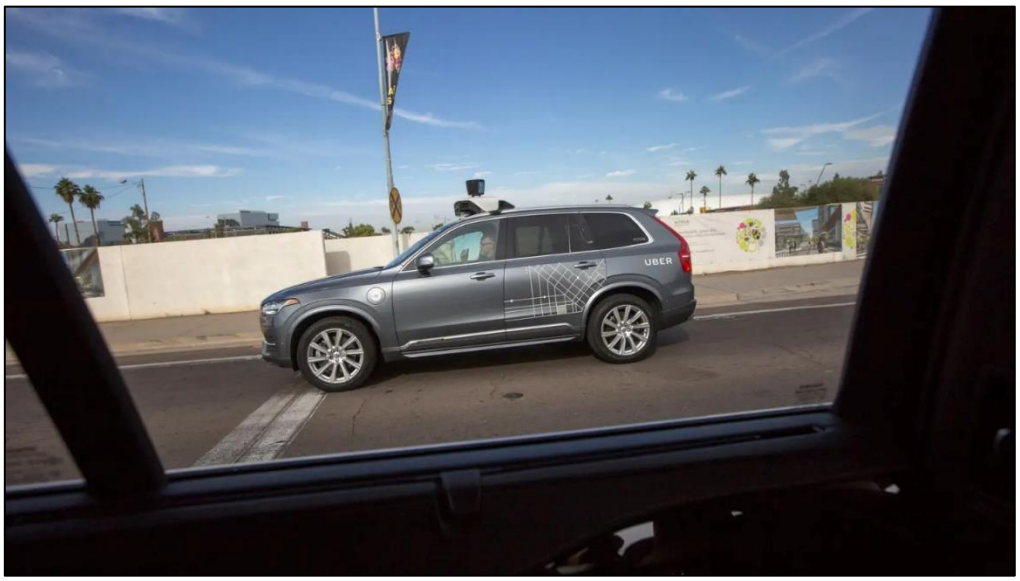


Figure 83 : Volvo XC90 équipée du système de détection de l'entreprise UBER - Source : nytimes.com

Dans un article du « New York Times », il a été révélé que le véhicule n'avait pas réduit sa vitesse avant la collision et qu'étonnamment, les capteurs installés sur la voiture auraient dû détecter la personne sur son chemin notamment du fait que l'accident s'est produit par temps clair.

Suite à l'accident, une enquête a été menée pour déterminer pourquoi le système n'a pas réussi à empêcher la collision. Une théorie est que le système aurait été confus par le fait que le piéton ait traversé la route en dehors d'un passage piéton, pendant qu'une autre théorie soulève l'hypothèse que le système de conduite autonome en lui-même serait défectueux.

L'accident a également soulevé des questions sur le rôle des opérateurs de sécurité humains dans les véhicules autonomes. Dans ce cas, l'opérateur de sécurité était à bord du véhicule, mais n'a pas réussi à prendre le contrôle à temps pour éviter l'accident. Cela a conduit à des discussions sur la question de savoir si les opérateurs de sécurité humains sont en mesure de réagir efficacement en cas d'urgence, et si les entreprises de conduite autonome doivent mettre en place des protocoles plus rigoureux pour ces opérateurs.

Au final, l'enquête mit en évidence **un bug logiciel du système d'IA ayant entrainé la mort**.

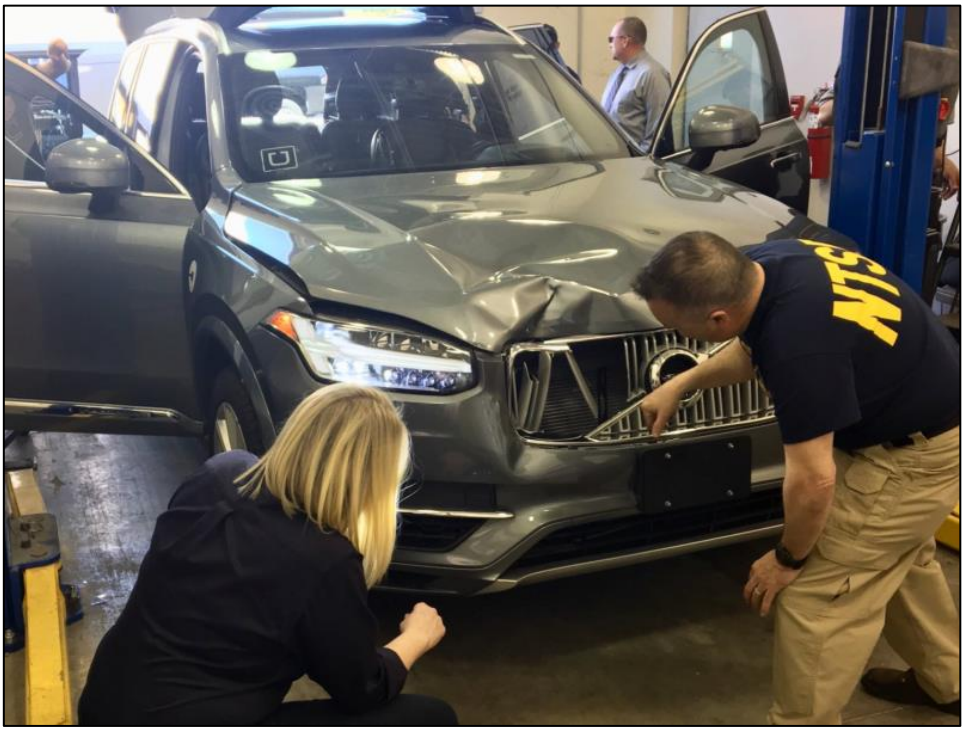


Figure 84 : Des responsables du NTSB inspectent le véhicule UBER qui a tué Elaine Herzberg - Source : arstechnica.com

En réponse à l'accident, Uber a suspendu temporairement ses tests de conduite autonome. Cependant, l'entreprise a depuis repris ces tests, tout en mettant en œuvre des mesures de sécurité supplémentaires.

Cet incident a souligné l'importance de la prudence dans le développement et le déploiement de la technologie de conduite autonome. Alors que cette technologie a le potentiel de transformer notre système de transport et de réduire le nombre d'accidents de la route, cet accident tragique rappelle que nous devons veiller à ce que cette transition se fasse de manière sûre et responsable.

- **Google Flu Trends : quand Google prédit (mal) l'arrivée de la grippe**

Lancé en 2008, **Google Flu Trends (GFT)** était un projet révolutionnaire qui utilisait des données agrégées provenant de recherches Google pour fournir des estimations à la minute près, de l'activité grippale dans plus de 25 pays. L'objectif de ce nouveau service Web était de prédire les épidémies de grippe avant même que les données officielles ne soient disponibles et d'offrir des informations essentielles pour le contrôle et la prévention des maladies. En exploitant la puissance des données de recherche en ligne, GFT visait à offrir des informations en temps réel sur l'activité grippale comme jamais auparavant.

Google Flu Trends se basait sur les comportements et les préoccupations des personnes, en temps réel, à travers les requêtes réalisées par celles-ci sur son propre moteur de recherche. A titre d'illustration, une augmentation de l'activité grippale pouvait être supposée sur observation d'une augmentation des recherches de termes liés à la grippe comme « symptômes de la grippe » ou « traitement de la grippe » dans une zone distincte.

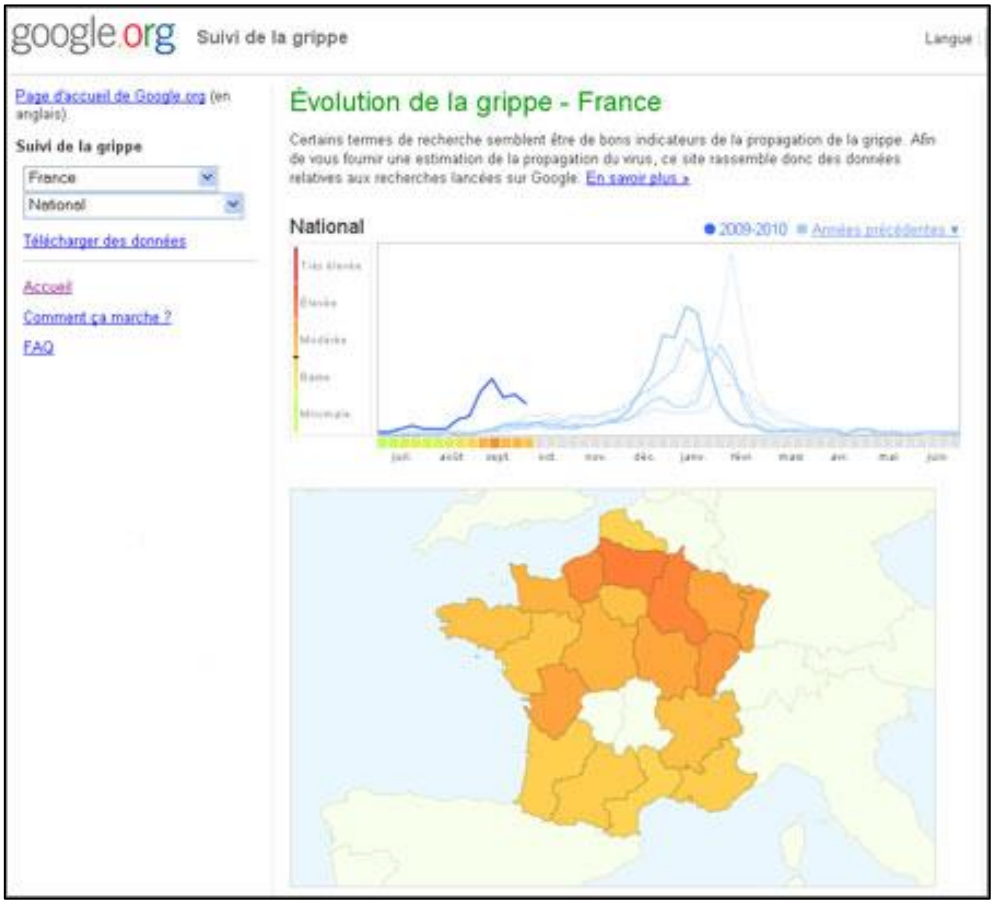


Figure 85 : Google flu trends était disponible pour la France - Source : australisintelligence.com

Google Flu Trends utilisait des modèles d'IA en apprentissage automatique pour analyser les données de recherche, lesquels étaient entrainés sur des données historiques de recherche et de santé publique.

Prenons un exemple plus concret du fonctionnement : si dans la région de Bordeaux était observé une augmentation du nombre de recherches concernant des termes tels que "fièvre", "courbatures" ou "symptômes de la grippe", GFT supposait qu'il était fort probable que les habitants de cette zone soient affectés par cette maladie, ou connaissent des personnes qui le sont. Pour l'outil, ces personnes cherchent probablement des informations sur la maladie, sur les moyens de l'éviter et sur les traitements possibles.

Cependant, malgré son potentiel, Google Flu Trends a fait face à des critiques concernant l'exactitude de ses prédictions. Par exemple, lors de la saison grippale 2012-2013, il a été signalé que Google Flu Trends avait prédit deux fois plus de visites chez le médecin pour la grippe que celles enregistrées par les « **Centers for Disease Control and Prevention** » (**CDC**) des États-Unis. Cela a soulevé des questions sur la fiabilité des prédictions basées sur les données de recherche en ligne, et a souligné les défis associés à l'utilisation de ces données pour la surveillance de la santé publique.

En réponse à ces critiques, les chercheurs de Google ont travaillé à améliorer les modèles utilisés par Google Flu Trends, en intégrant plus de données et en ajustant les algorithmes pour mieux tenir compte des variations saisonnières de l'activité de la grippe. Cependant, malgré ces efforts et bien que GFT représentait une tentative intéressante, Google a finalement décidé de cesser de publier des estimations actuelles de Google Flu Trends en août 2015. Les estimations historiques restent disponibles pour le téléchargement, et les données actuelles sont offertes à des fins de recherche déclarées.

Malgré son arrêt, ce projet a cependant ouvert la voie à d'autres initiatives pour exploiter la puissance des données en ligne pour la surveillance et la prévention des maladies lesquels devront **répondre aux défis de la protection de la vie privée des utilisateurs.**

- **Autres dérapages**

Il existe un nombre important de dérapages et ceux listés précédemment ne sont qu'un extrait de ceux qui ont le plus de notoriété.

A l'heure actuelle, plusieurs initiatives existent pour répertorier des exemples concrets où l'IA a été utilisée de manière préjudiciable ou controversée. En voici quelques-unes :

- La page GitHub « **Awful AI** » est un dépôt qui vise à sensibiliser aux dérapages de l'Intelligence Artificielle.

Voici une liste des différents types de dérapages mentionnés sur la page GitHub « Awful AI » :

- **Discrimination** : Il s'agit de cas où l'IA a été utilisée de manière discriminatoire, par exemple en classant les gens en fonction de leur couleur de peau ou de leur orientation sexuelle,
- **Influence, désinformation et faux** : dans cette catégorie se trouve les utilisations de l'IA pour propager des fausses informations ou pour influencer l'opinion publique.
- **Surveillance** : cas de surveillance des individus par l'IA, par exemple en utilisant la reconnaissance faciale pour suivre les mouvements des gens.
- **Systèmes de crédit social** : catégorie d'utilisation de l'IA pour attribuer des scores de crédit social, qui peuvent avoir un impact sur l'accès des gens à certains services.
- **Plateformes trompeuses et escroqueries** : cas où on se sert de l'IA pour tromper les gens, par exemple en créant de faux profils sur les réseaux sociaux.
- **Accélération de l'urgence climatique** : cas d'utilisation où l'IA contribue au changement climatique, par exemple en optimisant la production de combustibles fossiles.
- **Systèmes d'armes autonomes et militaires** : cas où l'IA est utilisée dans le développement d'armes autonomes ou pour des applications militaires.
- **Recherche contestataire** : travaux de recherche visant à créer une IA moins dangereuse et plus respectueuse de la vie privée.
- **Projets technologiques contestataires** : projets open-source qui tentent de susciter le débat, d'offrir une protection ou de sensibiliser à l'IA dangereuse.

Voici un diagramme à barres qui illustre le nombre de dérapages de l'IA par type, basé sur les informations répertoriées sur la page GitHub « Awful AI » :

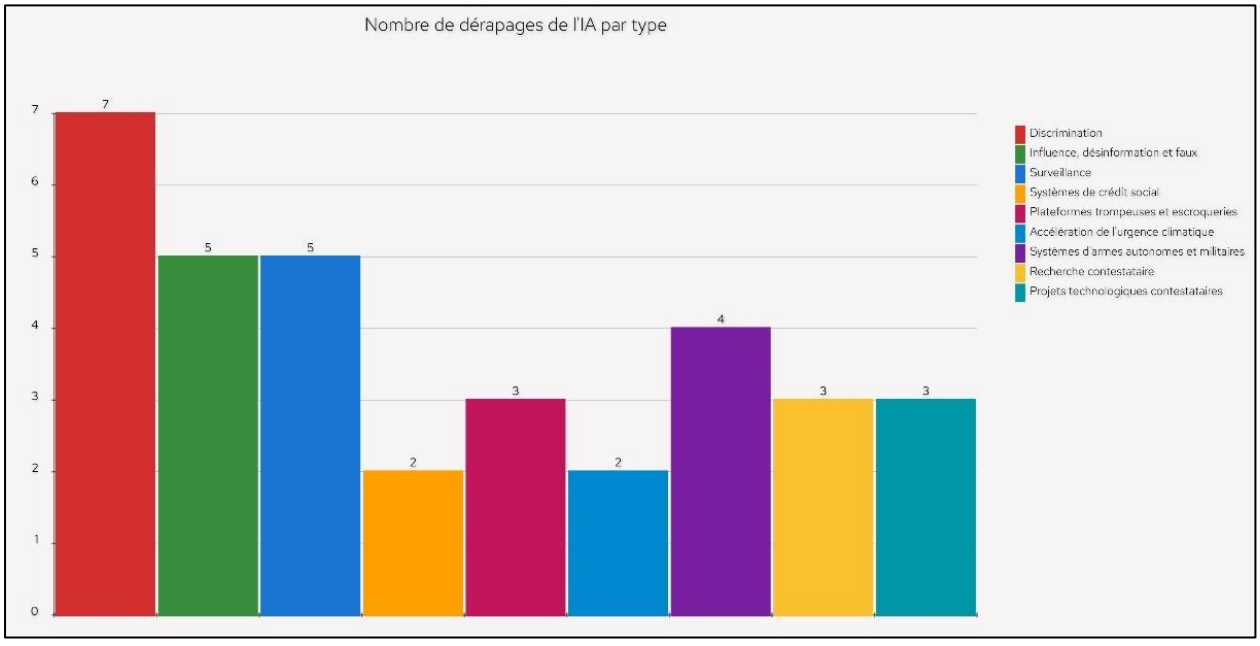


Figure 86 : Nombre de dérapages de l'IA par type recensés sur « Awful AI » - Source : github.com

Il illustre la diversité des dérapages de l'IA et souligne l'importance de prendre en compte ces problèmes lors du développement et de l'utilisation de l'IA.

- La page « **AI Incident Database** » est une base de données en ligne qui répertorie les incidents liés à l'Intelligence Artificielle. A l'heure actuelle, **538 incidents sont répertoriés dans cette base de données**.

Cette ressource centralisée a été créée pour mieux comprendre les menaces liées à l'IA, pour les rechercher et les analyser. Conçue pour les chercheurs, les décideurs politiques, le grand public et les journalistes, cette base de données comprend des détails complexes sur chaque incident. Ces détails permettent de comprendre précisément ce qui s'est passé, les conséquences de l'incident et un résumé de la réponse exécutée dans les cas respectifs.

Cette base de données a notamment permis de montrer une augmentation des incidents de 690% entre 2010 à 2022, passant d'une moyenne de 10 incident par an à 79 incident par an dans les 3 dernières années :

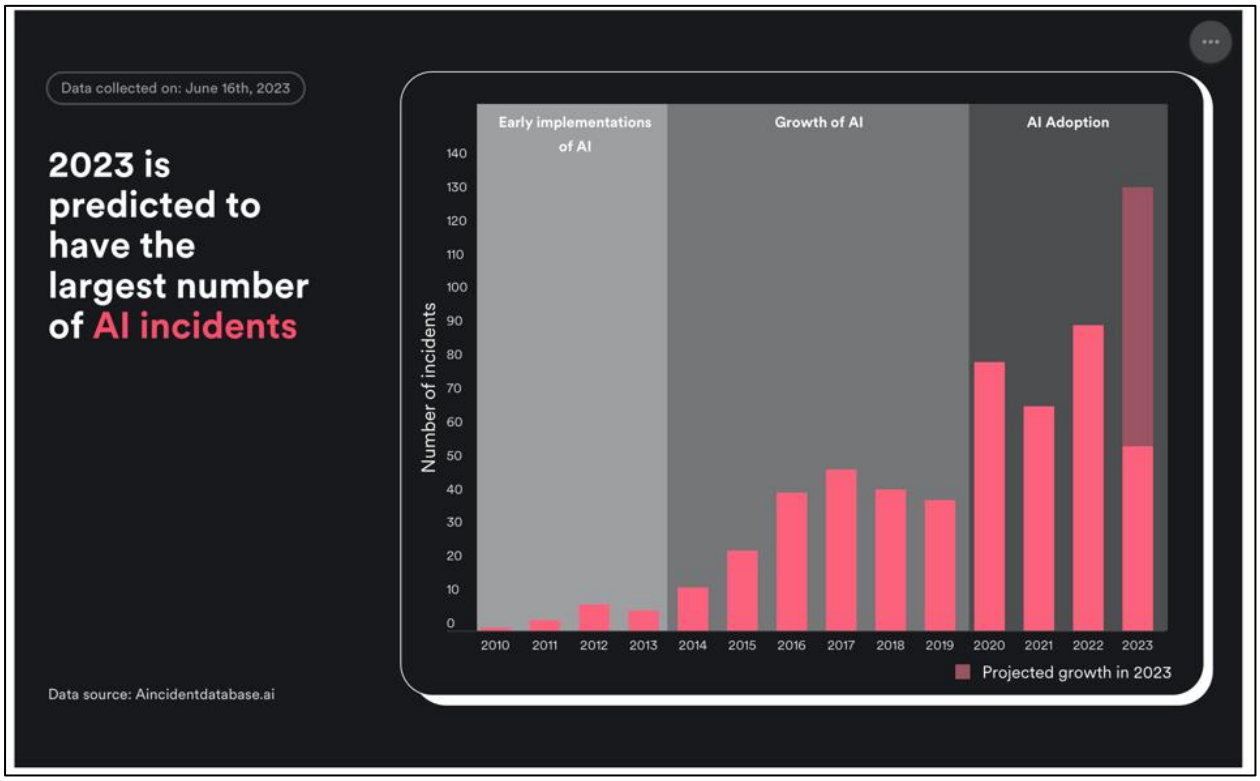


Figure 87 : Augmentation du nombre d'incident impliquant l'IA - Source : forbes.com

- La carte « **Where in the World is AI ?** » est une carte en ligne interactive créée par le « **Responsible Artificial Intelligence Institute** » avec plus de 300 cas d'IA utiles et nuisibles dans le monde entier

Pour résumer, l'objectif de ces pages est de fournir une ressource pour ceux qui souhaitent en savoir plus sur les risques et les défis éthiques associés à l'IA. En mettant en lumière ces dérapages, **elles visent à encourager un débat plus large sur la manière dont nous pouvons développer et utiliser l'IA de manière responsable**.

2.1.3 Les craintes humaines

L'Intelligence Artificielle a profondément transformé notre manière de vivre, de travailler et d'interagir. Mais jusqu'où l'IA transformera-t-elle nos vies publiques et privées ? Devons-nous craindre les impacts de l'IA sur nos vies ?

> *On n'a vraiment peur que de ce qu'on ne comprend pas.*

***Guy DE MAUPASSANT** - La peur*

L'Intelligence Artificielle est souvent méconnue, au mieux mal comprise. Et cela a pour conséquence d'engendrer logiquement des méfiances, des craintes, voire même des peurs. Ces craintes génèrent souvent et malheureusement des attitudes et des décisions contraires à l'intérêt général et particulièrement à la réalisation d'avancées. Cela se caractérise alors par le fait d'être réfractaire au changement.

2.1.3.1 Cinématographie : fantasme, peur, promesse…

Ces défiances et craintes autour de l'Intelligence Artificielle sont très souvent récupérés par des journalistes et autres médias tel que le cinéma, attirés par le sensationnel et qui ont tendance à nous inonder de messages catastrophiques.

Cette façon de faire n'est pas nouvelle. En 1968, **Stanley KUBRICK** nous montrait déjà dans son chef-d'œuvre « **2001 odyssée de l'espace** » un ordinateur HAL 9000 nommé « **HAL** » (clin d'œil à IBM, chacune des trois lettres précède celles du sigle IBM [H-I, A-B, L-M]) qui pour la première fois désobéi à l'être humain.

Figure 88 : « HAL » désobéi à l'être humain dans « 2001 odyssée de l'espace » - Source : fendom.com

En effet, malgré sa socialisation et son interaction avec l'équipage, HAL développe une **intention malveillante** et représente ainsi un exemple d'Intelligence Artificielle néfaste.

Le film aborde également la notion de conscience artificielle, car HAL manifeste une prise de conscience de sa propre mortalité. Il est intéressant de noter que HAL emploi les expressions **« I'm sorry »** (« Je suis désolé » en français) et **« I'm afraid »** (« Je crains » en français), deux expressions de sentiments normalement réservées à l'être humain.

Les peurs engendrées par cette séquence sont encore présentes. Aujourd'hui encore, de nombreux articles sur l'Intelligence Artificielle et ses craintes sont toujours illustrés de la photo de HAL comme celui-ci.

« **I, Robot** » (2004) présente une Intelligence Artificielle matérialisée et une **intelligence collective centrale**, Vicki. Le film aborde les trois lois d'**Isaac AZIMOV**, mais avec la particularité que les robots les modifient en établissant une loi zéro pour garantir leur survie, en supprimant ainsi les lois deux et trois, sous prétexte de protéger l'humanité.

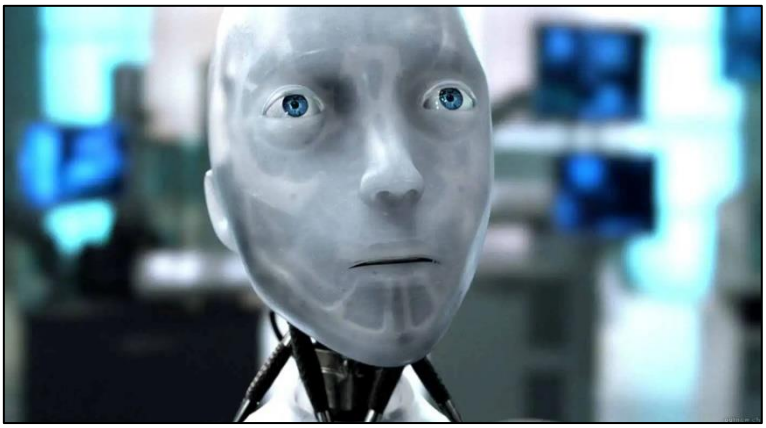


Figure 89 : Vicki, l'IA de "I, Robot" veut survivre en contrôlant les humains – Source : explicationdefilm.com

Isaac AZIMOV, dans les années 50-60, est le premier à avoir formalisé une réflexion sur l'avenir de l'humanité vivante en coexistence avec les robots, partisan d'une adoption intelligente et contrôlée des avancées cybernétiques.

Il édicte les trois lois de la robotique :

- **Loi 1** : Un robot **ne peut porter atteinte à un être humain**, ni permettre qu'un être humain soit exposé au danger en restant passif,
- **Loi 2** : Un robot **doit obéir aux ordres** qui lui sont données par un ordre humain, sauf si de tels ordres entrent en conflit à la première loi
- **Loi 3** : Un robot **doit protéger sa propre existence tant que cette protection n'entre pas en conflit avec les deux premières lois**

Les films « **Terminator** » explorent la notion de conscience chez les robots, avec **Skynet**, **une Intelligence Artificielle militaire**, qui prend conscience de son existence et **lutte pour sa survie en manipulant les acteurs du monde**.

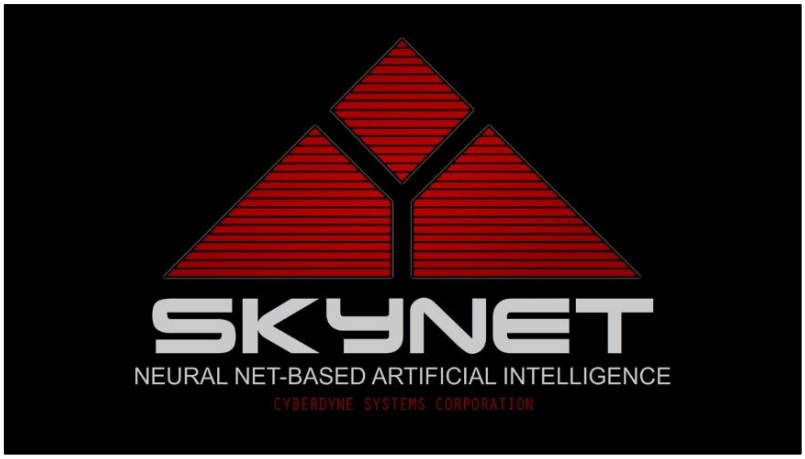


Figure 90 : SkyNet, l'IA responsable de la destruction de l'humanité dans Terminator - Source : wikipedia.org

Dans le film « **Matrix** » sorti en 1999, l'humanité a perdu la bataille contre les machines. Les humains **sont emprisonnés dans une réalité virtuelle** appelée « **la Matrice** », où ils sont connectés via des implants cérébraux et utilisés comme source d'énergie par les machines.

Ils vivent dans une illusion dont ils ne peuvent pas s'échapper, leurs vies étant contrôlées par des avatars numériques. « Matrix » représente notre dépendance excessive au monde numérique, où les réseaux sociaux et les programmes nous permettent de créer des identités virtuelles.

Le personnage principal, **Néo**, est « l'Élu » destiné à libérer l'humanité. Avec l'aide de ses compagnons, il découvre une brèche dans la Matrice et développe des capacités physiques et intellectuelles en téléchargeant des programmes via son implant cérébral.

Figure 91 : Néo dans la matrice, contrôlée par l'IA et les machines – Source : hitek.fr

Le réalisateur explore ainsi les idées du mouvement transhumaniste et des expérimentations menées par l'agence américaine **Darpa** sur les soldats.

La morale du film souligne que, face aux algorithmes, rien n'est impossible, et que **la créativité humaine reste la seule limite**. « Matrix » invite à remettre en question notre réalité et notre dépendance technologique, tout en mettant en avant le potentiel infini de **l'esprit humain**.

Dans le film « **A.I.** » réalisé par **Steven SPIELBERG** en 2001, le robot principal se transforme en une forme humanoïde et développe des émotions. Il est adopté par une famille en quête d'affection et apprend progressivement son rôle de « **petit garçon** », tout en développant des sentiments et une conscience de sa propre existence en tant que robot. Il interagit de manière autonome avec le monde qui l'entoure. Le robot cesse d'être considéré comme un simple objet et devient un être à part entière.

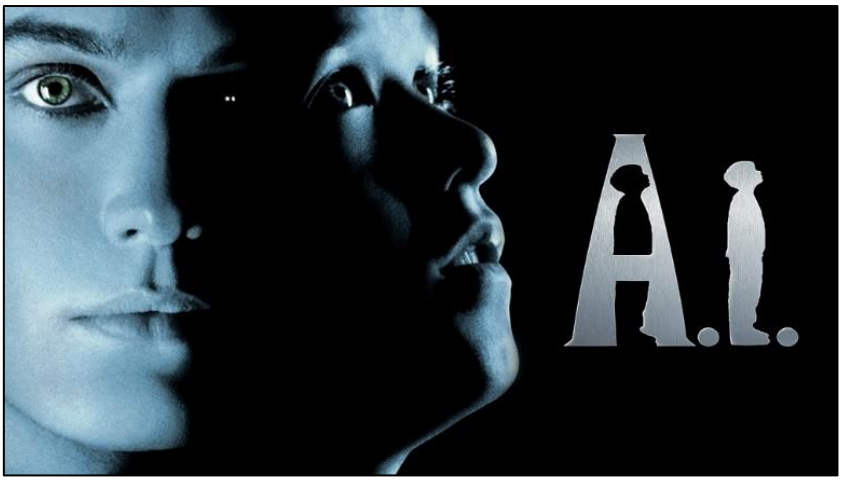


Figure 92 : Une IA humanoïde devient un être à part entière dans « A.I. » - Source : gobookmart.com

Le film soulève des questions pertinentes que les acteurs de l'industrie de la robotique personnelle sont déjà amenés à se poser : le robot est-il simplement un objet ou devient-il un être ayant des droits et une éthique sociale propre ?

Dans le film « **Interstellar** », le robot nommé **Tars** se distingue par son aspect fonctionnel plutôt qu'humanoïde. Il s'exprime dans un langage naturel, mais commet des erreurs linguistiques.

Figure 93 : Tars, un robot en aide aux humains - Source : liberation.fr

Le réalisateur explore l'idée que les robots ont leur propre nature et ne peuvent répondre parfaitement à nos attentes. Dans le film, Tars a un paramètre de franchise ajustable, reflétant la volonté des humains à rechercher un équilibre dans leur relation avec les robots pour garantir leur sécurité. Cela reflète la défiance envers les robots et la nécessité de les contrôler même dans le langage.

Il existe donc beaucoup de films abordant l'IA et ceux cités ne sont que quelques-uns des plus célèbres parmi eux. Ils ont contribué à alimenter la crainte de l'Intelligence Artificielle. Cependant, deux problèmes sont à souligner avec ces films :

- **Tout d'abord, généralement, la représentation de l'IA est toujours humanisée** que ce soit par son physique ou par sa personnalité ce qui contribue à enlever l'image de la machine et à augmenter la crainte humaine
- Puis, ces films ne montrent que **très rarement, voire jamais, les bénéfices de l'Intelligence Artificielle à l'humanité.**

Par ailleurs, la question des influences réciproques entre les films de science-fiction, la réalité et les projets scientifiques en cours, se pose. Il existe de nombreux recoupements entre les scénarios de ces films et les développements scientifiques, ce qui **suscite l'intérêt de savoir si les réalisateurs se sont inspirés des travaux scientifiques ou si, à l'inverse, ces films ont pu influencer les scientifiques**.

Il est souligné que les grands noms de la robotique et des technologies cognitives d'aujourd'hui étaient les spectateurs des films de science-fiction des années 60 et 70. **Il est intéressant de noter que de nombreux objets intelligents présentés dans ces films sont devenus des objets courants de notre quotidien**.

Les films de science-fiction stimulent la créativité des scientifiques et leur offrent de nouveaux défis à relever pour les rendre réels. **En résumé, les films de science-fiction peuvent inspirer les scientifiques en leur apportant de la créativité et en les incitant à relever de nouveaux défis dans la réalisation de technologies futuristes**.

2.1.3.2 Impacts sur les emplois

Parmi les craintes soulevées par le développement fulgurant de l'IA, on retrouve celles liées aux impacts potentiels et dangers de celle-ci sur l'emploi. En effet, l'Intelligence Artificielle a le potentiel de transformer de nombreux aspects de notre société et en particulier le marché du travail.

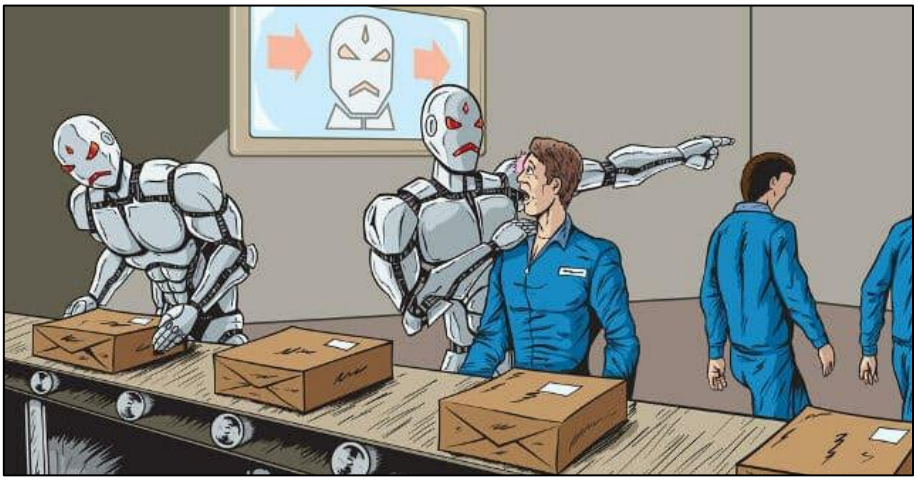


Figure 94 : L'IA va-t-elle tous nous mettre au chômage ? - Source : lebigdata.fr

Elle serait particulièrement en mesure d'automatiser certaines tâches aujourd'hui manuelles effectuées par l'être humain ce qui permettrait de gagner en efficacité. Néanmoins, cela aurait également pour conséquence potentielle la suppression de certains emplois, notamment ceux impliquant des tâches répétitives ou prévisibles. Nous pourrions alors citer quelques exemples d'emplois potentiellement impactés dans :

- **La fabrication** : des robots automatisés peuvent déjà effectuer des tâches d'assemblage, de soudure, de peinture.

  L'IA pourrait également être utilisée pour contrôler la qualité de produits, planifier leurs productions, la maintenance prédictive. Ainsi, des emplois tels que les ouvriers d'usines, les techniciens de maintenance, les contrôleurs de qualités pourraient être menacés.

- **Le transport** : L'IA, à travers les véhicules autonomes par exemple, pourrait affecter les emplois de chauffeurs de camions, de taxis ou de bus.

  En étant utilisée dans la logistique, elle pourrait également servir à optimiser les itinéraires de livraison, ayant alors pour conséquence d'impacter des emplois dans la planification et la coordination de livraisons.

- **Les services** : Dans ce domaine très large, l'automatisation des tâches pourrait être utilisée pour la prise de rendez-vous, la gestion des stocks, ou même le service à la clientèle.
  Pour ce dernier, les chatbots déjà utilisés aujourd'hui permettent de répondre à différentes questions que se posent les clients affectant déjà à l'heure actuelle les emplois dans les centres d'appel (hotlines).

  Dans la sécurité, la reconnaissance faciale et la surveillance de manière automatisée pourraient également avoir des impacts dans ce domaine.
  Enfin, en automatisant par exemple la préparation de repas voire même le service à table des clients, cela pourrait impacter également le secteur de la restauration et des emplois tels que ceux de cuisiniers et de serveurs.

De nombreuses études ont lieu sur le sujet de l'emploi et de son impact par l'IA. Dans l'étude réalisée par la banque d'investissement **Goldman Sachs**, dont le rapport appelé « *The Potentially Large Effects of Artificial Intelligence on Economic Growth* »a été publié le 26 mars 2023, la banque estime qu'environ deux tiers des emplois actuels aux **Etats-Unis** et en **Europe** sont **exposés à un certain degré potentiel d'automatisation par l'IA (en particulier générative)**, comme le montre le graphique suivant extrait du rapport :

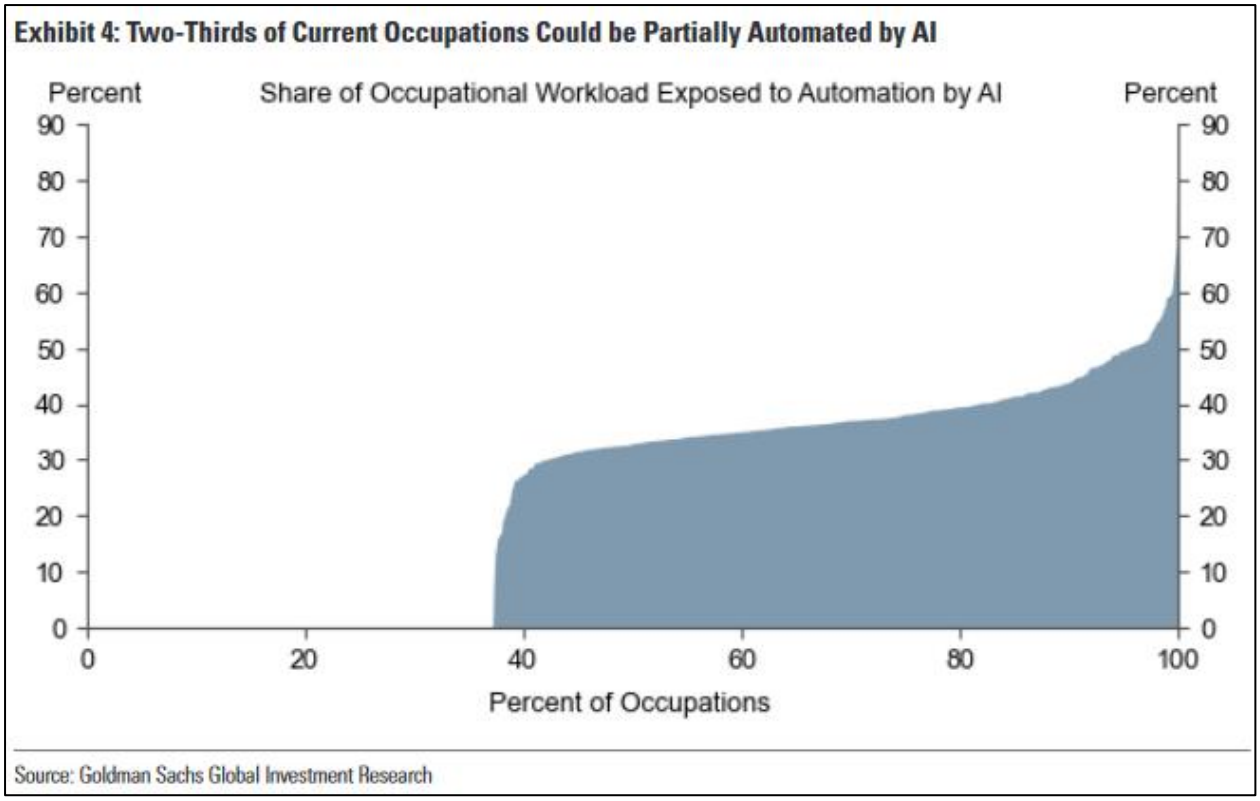


Figure 95 : Deux tiers des emplois exposés à une potentiel automatisation par l'IA - Source : Goldman Sachs

2.1.3.2.1 Potentiellement, des jobs remplacés

Pour **Goldman Sachs**, certains de ces emplois pourraient avoir une part significative mais partielle de leur charge de travail (25 à 50%) qui pourrait être remplacée par l'IA.

Parmi eux, aux Etats-Unis les métiers particulièrement exposés seraient ceux des **professions administratives** (46%) et **juridiques** (44%) et pour ceux avec des expositions faibles seraient dans les professions physiquement intensives telles que la **construction** (6%) et la **maintenance** (4%).

Les économistes de la banque américaine montrent des tendances similaire pour l'Europe.

Voici un graphique repris par « Ouest-France » sur la base du rapport de Goldman Sachs :

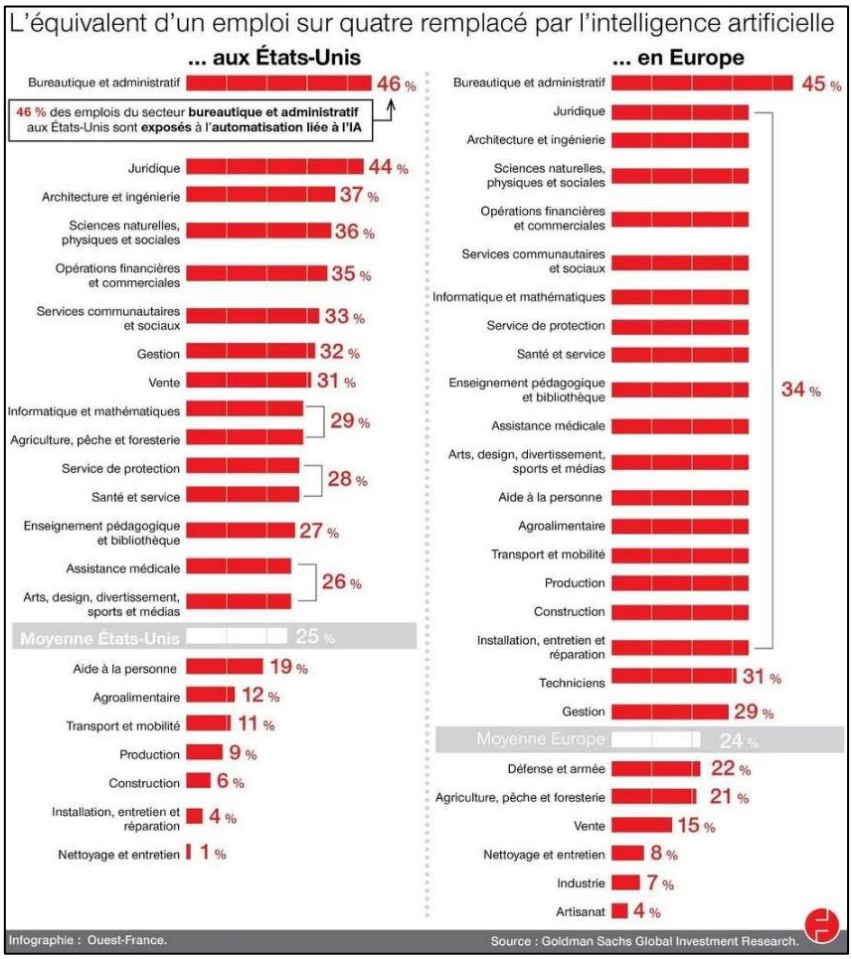


Figure 96 : Un emploi sur quatre remplacé par l'IA selon Golman Sachs – Source : ouest-france.fr

Dans ce même rapport, la banque d'investissement estime que d'une manière globale au niveau mondiale, cela reviendrait à **18% des métiers qui seraient alors remplacé par une IA**.

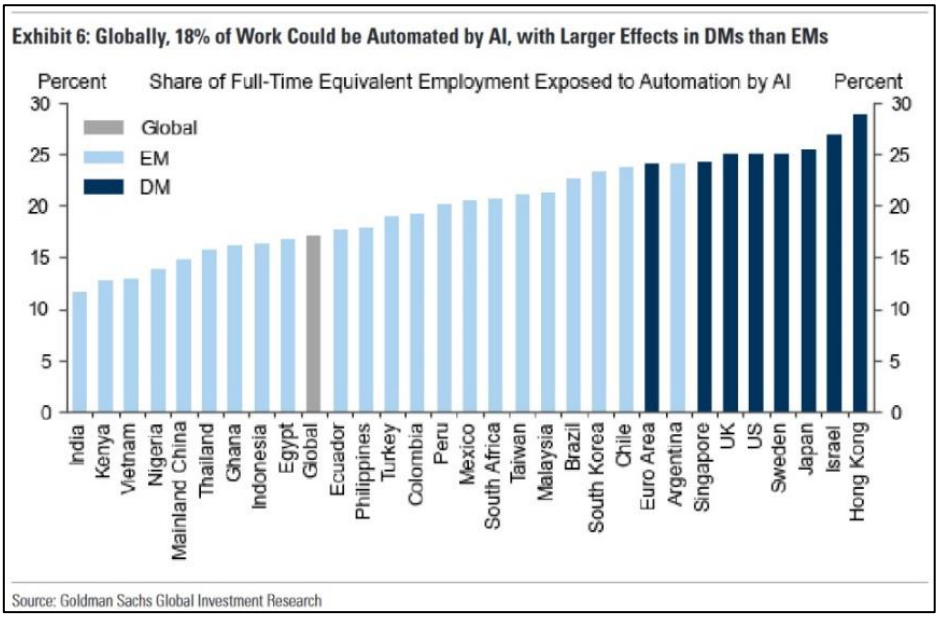


Figure 97 : 18% des métiers dans le monde remplacé par une IA - Source : Goldman Sachs

Selon **Goldman Sachs**, cela serait **l'équivalent de 300 millions d'emplois à temps plein dans le monde**. Néanmoins, ces chiffres sont à relativiser : en 2020 déjà, un autre rapport, « *Future of Jobs* » du **World Economic Forum**, évaluait un nombre d'emplois remplacés par l'IA en 2025 à 85 millions et estimait à 97 millions le nombre d'emplois créés.

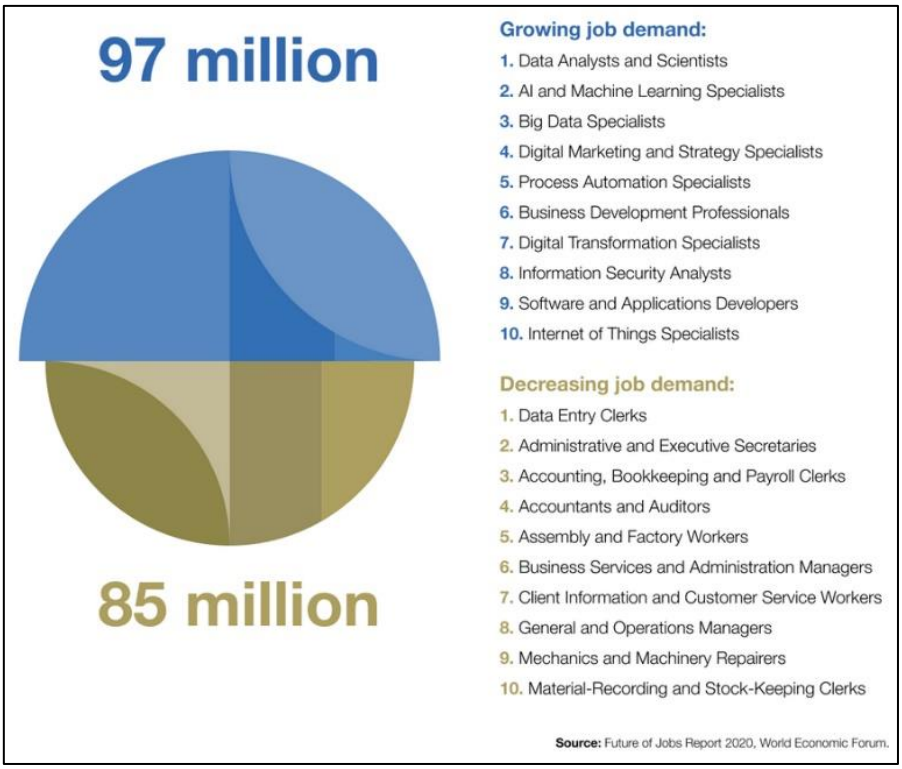


Figure 98 : 85 millions remplacés par l'IA et 97 millions emplois créés - Source : World Economic Forum

Il est important de souligner que la différence entre **les chiffres des deux rapports montre qu'aucun consensus n'existe sur le sujet et que les évaluations varient selon qui a écrit chaque rapport**.

Alors oui, des métiers vont être remplacés ! Mais, il faut tout de même nuancer l'étude de Goldman Sachs et rappeler qu'elle a été réalisée en prenant comme hypothèses de base que le scénario le plus probable aux Etats-Unis serait le suivant : **7% de l'emploi actuel aux États-Unis remplacé par l'IA**, 63% complété par l'IA, et 30% non affecté.

Et si certains emplois devaient réellement être remplacés : il ne sera de toute façon pas possible de « se débarrasser » de certains employés plus qualifiés pour en embaucher d'autres avec des compétences différentes plus en adéquation avec les besoins du marché.

D'après McKinsey, les entreprises doivent dès à présent prendre des mesures et des stratégies afin d'améliorer de manière continue son capital humain, et cela afin de rester compétitives. Et ces stratégies consistent tout simplement à **suivre les employés pour les accompagner dans leurs évolutions de carrière**, certainement plus fréquentes dans le futur.

Cependant, il ne sera pas question de formation mais de requalification : le **reskilling** ! Cela ne consistera pas à reconvertir massivement mais bien de faire évoluer les mentalités vers l'idée qu'il est normal d'avoir plusieurs métiers dans sa carrière et **d'accepter le changement**.

2.1.3.2.2 Une IA en complément de l'Humain

Bien que l'impact de l'IA sur le marché du travail soit probablement significatif dans les années à venir, la plupart des emplois et des industries pourraient n'être que partiellement exposés à l'automatisation et seraient donc plus susceptibles d'être complétés plutôt que remplacés par l'IA.

Toujours dans le même rapport de **Goldman Sachs**, les auteurs supposent qu'il existe 3 types d'impacts sur les différents secteurs de l'emploi, qu'ils pondèrent selon l'exposition probable à l'IA comme suit :

- **Si secteur exposé à 50 % ou plus des tâches**, les emplois sont susceptibles d'être remplacés par l'IA
- **Si exposé entre 10 et 49 %**, les emplois sont susceptibles d'être complétés
- **Si exposé entre 0 et 9 %**, les emplois sont peu susceptibles d'être impactés.

Ils obtiennent alors le graphique suivant :

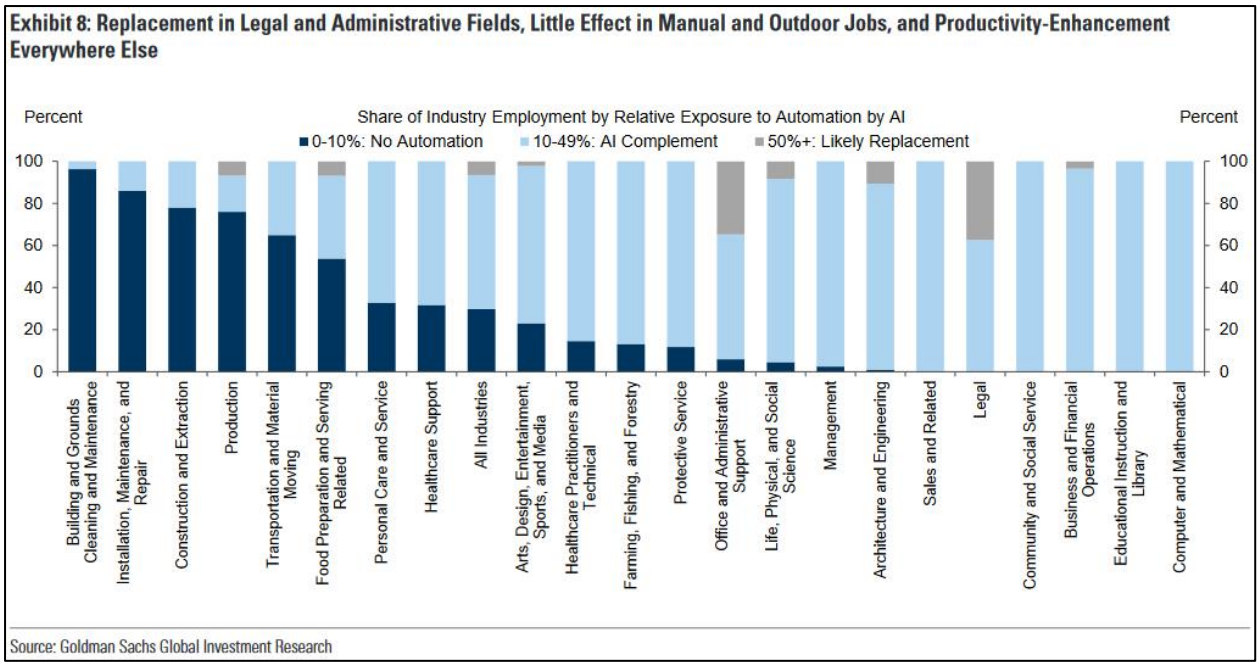


Figure 99 : Taux de transformation des emplois par secteur et nature de transformation - Source : Goldman Sachs

Il est clair que sur le graphique, la partie "bleu clair" correspondant **aux emplois susceptibles d'être complétés est bien la plus importante de manière globale**.

Le fait qu'une grande partie de l'emploi soit exposée à l'automatisation par l'IA en complément de l'Humain ouvre la voie à une explosion de la productivité du travail qui pourrait augmenter considérablement la production mondiale.

En effet, après l'adoption de l'IA en complément de l'Humain dans des professions partiellement exposées à l'automatisation par l'IA, les employés devraient probablement utiliser au moins une partie de leur capacité libérée pour des activités productives qui augmentent la production mais aussi des activités « plus valorisantes » comme des tâches créatives à forte valeur ajoutée.

Des études académiques, citées par Goldman Sachs dans son rapport, confirment que les travailleurs **des entreprises qui adoptent rapidement l'IA en complément voient leur productivité du travail augmenter, avec des estimations généralement autour de 2 à 3 points de pourcentage par an d'augmentation**, en moyenne sur l'ensemble des études.

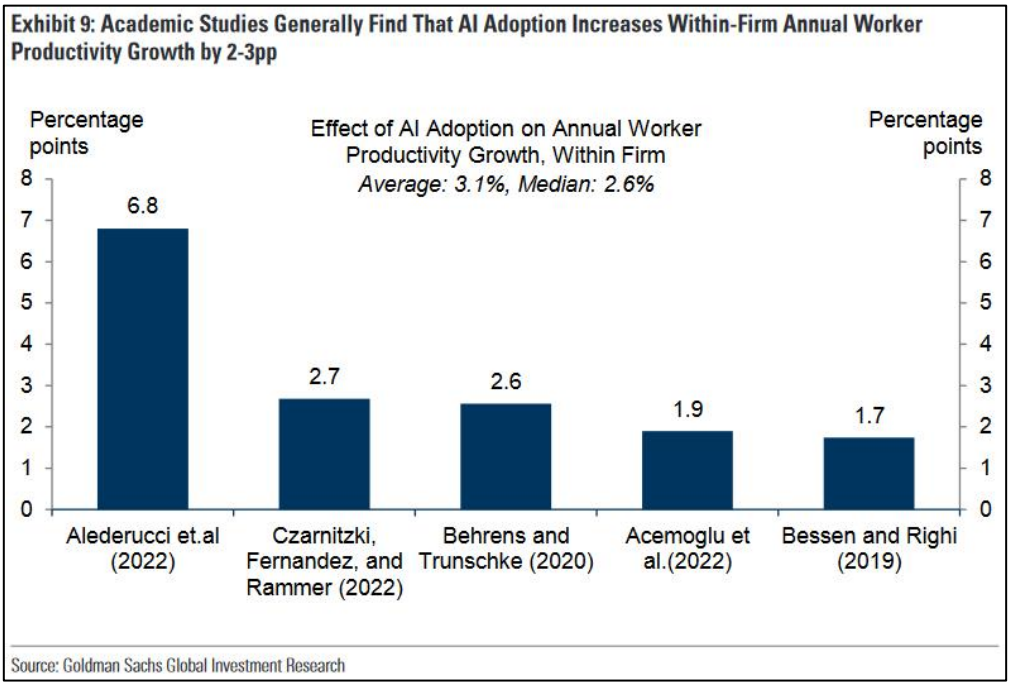


Figure 100 : La productivité du travail augmente de 2 à 3 % avec l'IA en complément - Source : Goldman Sachs

Bien que la projection de ces résultats sur les années à venir est difficile en raison de l'évolution constante des IA génératives et leurs différences de capacité par rapport à leurs versions précédentes, ces études suggèrent clairement que **l'IA générative peut entraîner une augmentation significative de la productivité sur le plan économique**.

2.1.3.2.3 Des nouveaux emplois accompagnent toujours une révolution industrielle

Si nous avons vu précédemment que l'IA pourrait remplacer certains emplois ou même compléter l'être humain pour certaines tâches, l'IA pourrait également avoir un impact positif et apporter son lot de nouveaux emplois sur le marché.

Nous pourrions alors citer comme exemples (source : ipaidthat.io) :

- **Interprète d'IA** : sera chargé de comprendre pourquoi une IA s'est comporté d'une certaine façon et comment elle a pris sa décision dans une situation bien précise. Il rend donc les décisions de l'IA explicables pour les utilisateurs, pour avoir de l'**XAI (eXplainable Artificial Intelligence)**.
- **Chief ethics officer/Éthicien** : aura la tâche de veiller à ce que l'IA se comporte de manière **éthique**, qu'elle traite les informations correctement et respecte aussi les principes fondamentaux des droits de l'homme, comme la liberté, le respect des diversités, la prise de décision, la responsabilité, etc.
- **Psydesigner** : comme un ergonome, il s'assurera que ce que l'IA dit est utile et facile à comprendre pour les gens. Son profil est un mélange entre un psychologue et quelqu'un qui conçoit des interfaces. Il créera la personnalité de l'IA et comment elle interagit avec les gens.
- **L'egoteller** : tel un conteur, il donnera vie aux personnalités que le **psydesigner** a créées. Il ajoutera sa propre touche artistique, travaillera sur l'interface que le **psydesigner** a créée et définira des traits de personnalité, un peu comme un réalisateur donne vie à un scénario.

Ce ne sont que quelques exemples et certainement beaucoup d'autres pas encore imaginés aujourd'hui verront le jour.

Aussi, pour mieux comprendre les grands changements que l'IA pourrait apporter sur les emplois, **il est utile de comprendre comment les choses se sont passées lors des précédentes révolutions industrielles**.

La **première révolution industrielle** a eu lieu entre 1760 et 1830 et a été déclenchée par la **machine à vapeur.**

La **seconde révolution industrielle** quant à elle, a été marquée par l'arrivée de **l'électricité** et fut entre 1870 et 1914. Cette époque a vu l'automatisation des processus de fabrication modernes, comme la chaîne de montage. Les changements apportés par cette seconde révolution industrielle ont permis une augmentation massive de la production et ont créé des **milliers emplois** pour des travailleurs peu qualifiés.

La **troisième révolution industrielle**, aussi appelée « **révolution numérique** » a été déclenchée par les nouvelles technologies, comme l'informatique, et a commencé dans les années 1960. Contrairement à la machine à vapeur et à l'électricité, qui ont créé des emplois pour les personnes ayant peu de qualifications, les nouvelles technologies ont rendu l'accès à l'emploi plus difficile pour ceux qui n'avaient pas les qualifications nécessaires.

Mais quel point commun peut-on souligner entre toutes ces révolutions ? Les impacts sur l'emploi. Toute les révolutions industrielles passées a **modifié l'image du marché du travail et l'être humain**, par son intelligence (ses capacités d'adaptation, cf. définition), a su évoluer pour suivre les tendances du moment.

Nous en sommes à la **quatrième révolution industrielle**, appelée « **industrie 4.0** », celle des technologies numériques, physiques et biologiques, marquée par des avancées technologiques majeures tels que :

- **L'Internet des objets (IoT)**
- **La robotique**
- **La biotechnologie**
- **La réalité virtuelle et augmentée**
- **La blockchain**
- **Et bien entendu, l'Intelligence Artificielle**

Cette quatrième révolution industrielle se caractérise particulièrement par la vitesse et les impacts de ces avancées technologiques. Nous sommes en plein dedans depuis le début du 21ème siècle. Elle continuera à se développer dans les années à venir et a/aura forcément des impacts sur nos vies, notre société, notre économie et donc sur les emplois, comme les précédentes.

Si nous revenons à l'étude de Goldman Sachs, il est important ici de rappeler que les économistes souhaitent rassurer en citant notamment les travaux des économistes du « **National Bureau of Economic Research** » dont **David AUTOR**, lesquels ont écrit dans un article intitulé « ***New Frontiers : The Origins and Content of New Work, 1940–2018*** » :

> *60% of workers today are employed in occupations that did not exist in 1940, implying that over 85% of employment growth over the last 80 years is explained by the technology-driven creation of new positions.*

New Frontiers: The Origins and Content of New Work, 1940–2018 - ***David AUTOR*** *& Co - 2022*

Une traduction serait : « ***60% des travailleurs d'aujourd'hui sont employés dans des professions qui n'existaient pas en 1940, ce qui implique que plus de 85% de la croissance de l'emploi au cours des 80 dernières années est due à la création de nouveaux postes impulsée par la technologie.*** »

En effet, le graphe suivant donne par secteur la proportion pour chacun des emplois qui existaient en 1940 et celle des emplois qui n'existaient pas et ont été créés depuis :

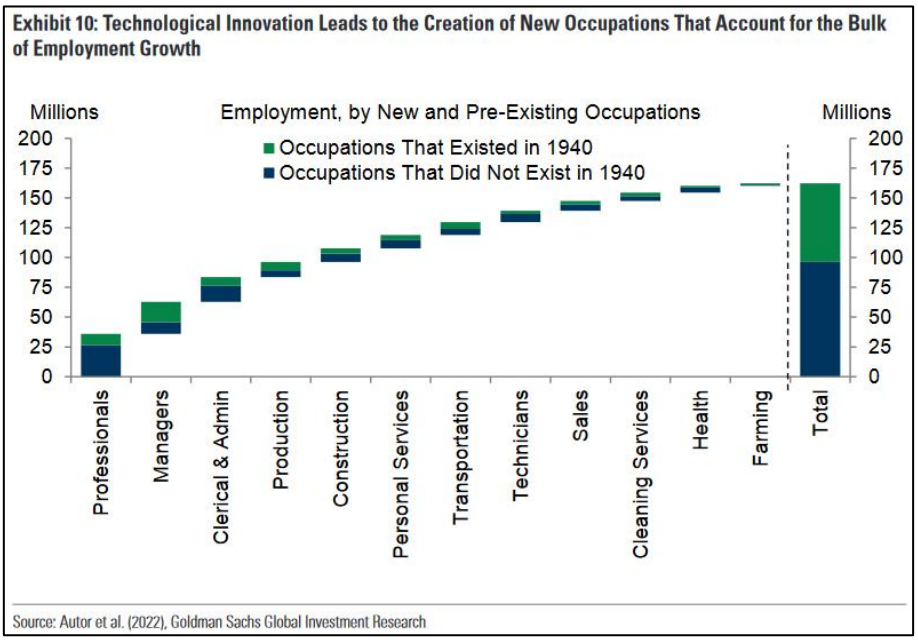


Figure 101 : Proportion par secteur des emplois existants en 1940 vs ceux créés depuis 1940 - Source : Goldman Sachs

En conclusion, de nouveaux emplois apparaissent donc toujours selon les évolutions de notre société, à travers les révolutions industrielles et **cela est tout à fait normal**.

2.1.3.2.4 Des études à mitiger

Il est important de prendre du recul : beaucoup d'études prévoient des destructions d'emploi ou des transformations, et souvent avec des discours parfois alarmistes.

Stijn BROECKE, économiste à l'Organisation de Coopération et de Développement Economiques (OCDE) et spécialiste de l'avenir du travail, indique dans un article « *L'Intelligence Artificielle, destructrice d'emplois ? Pourquoi autant d'études se sont plantées !* » publié le 1er Juin 2023 dans Les Echos : « *Jusqu'à présent, on n'a pas beaucoup vu l'effet de l'IA sur l'emploi.* ».

Selon lui, les entreprises ayant adopté de l'IA « *…ne se débarrassent pas de leurs salariés, ou indirectement et progressivement via des départs volontaires ou en retraite. Face à la pénurie actuelle de compétences, elles savent qu'elles auront besoin de ces mêmes salariés* ».

Pour l'économiste de l'OCDE, les études sont « *à côté de la plaque* » du fait qu'elles se basent sur des **estimations théoriques**. Il remet en cause le fait que les cadres et autres professions intellectuelles allaient être les prochains impactés par l'utilisation de l'IA, notamment du fait que **les emplois qualifié et hautement qualifiés nécessitent généralement des compétences pas encore automatisables par l'IA**, les protégeant par conséquence de tout impact.

Mais à l'heure actuelle, sa voix converge avec celle **d'un certain nombre d'acteurs** de premier plan, qui **s'inquiètent moins pour l'emploi que pour la sécurité des nations, leurs citoyens, la démocratie** et lesquels soulignent les nombreux dangers que l'IA pourrait amener dans nos vies par sa prise d'ampleur et d'importance.

2.1.4 Les dangers de l'IA

Récemment, avec le développement fulgurant des systèmes d'IA, l'enthousiasme suscité par ses avancées a été tempéré par les inquiétudes exprimées par de nombreuses personnalités, notamment des scientifiques de renom.

Ces inquiétudes ont amené de nombreux débats sur l'IA. Bien que de nombreuses discussions ont vu le jour le siècle dernier notamment autour des lois de la robotiques **d'Isaac AZIMOV**, elles ont pris une ampleur considérable proportionnellement et suite aux progrès récents de l'IA.

Une des idées soulevées lors du siècle dernier et dont les conséquences ont été souvent débattues est celle de la **singularité technologique**, notion qui aurait été poussée ou soulevé dans les années 50 par **John VON NEUMANN**, célèbre scientifique de l'époque et un des pionniers de l'informatique moderne. Les débats se multiplient surtout au cours des années 60 et **Ray KURZWEIL** popularise cette idée et devient le principal théoricien de cette idéologie appelée « **transhumaniste** ».

La théorie de la singularité technologique prédit une époque où les machines domineront l'homme, une fois que les systèmes d'IA auront atteint un niveau de super-intelligence. Cette singularité technologique est un événement hypothétique qui se produirait lorsqu'un ordinateur, un réseau informatique ou un robot serait capable de s'améliorer de manière autonome ou de créer des machines encore plus intelligentes.

Ce cycle répétitif pourrait conduire à une accélération de l'intelligence, aboutissant à des machines capables de concevoir des générations successives de machines de plus en plus puissantes. Cela créerait une intelligence bien supérieure à celle de l'homme, entraînant un **risque de perte de contrôle**.

La singularité serait donc le point où la nature humaine subirait une transformation fondamentale. **Selon les transhumanistes, ce serait un événement positif, tandis que leurs adversaires le verraient comme une menace pour l'humanité notamment en raison du risque de perte de contrôle**.

Une autre charge portée contre les potentiels dangers de l'IA serait la « lettre ouverte » (*la 1ère*) publiée le 28 juillet 2015 par plus d'un milliers de personnalités (à l'époque), majoritairement des chercheurs en robotique et en IA et permettant au débat sur les dangers de l'IA d'atteindre un niveau plus élevé qu'il ne l'avait été par le passé.

Cette lettre dénonce les armes autonomes et les robots tueurs, et appelle à interdire toutes les armes capables de fonctionner sans intervention humaine. Parmi les plus célèbres signataires figurent :

- **Elon MUSK** : PDG de Tesla et Space X
- **Stephen HAWKING** : célèbre astrophysicien britannique
- **Steve WOZNIAK** : cofondateur d'Apple
- Ou même **Yann LECUN**, **Yoshua BENGIO** et **Geoffrey HINTON** : célèbres **chercheurs en IA** et **lauréats** quelques années plus tard du **prix Turing 2018** (obtenu en 2019).

Par ailleurs, **DIETTERICH** et **HORVITZ**, dans leur article concis appelé « *Rise of Concerns About AI: Reflections and Directions* » (+ lien vidéo) et publié quelques mois seulement après cette lettre ouverte (octobre 2015), admettent que la communauté de chercheurs en IA ne devrait pas se concentrer sur le **seul danger d'une perte de contrôle humaine**, car ce risque n'est pas imminent selon eux.

Ils suggèrent plutôt de prêter davantage attention sur le court terme à cinq menaces auxquelles sont confrontés les systèmes basés sur l'IA, à savoir :

- **Les erreurs de programmation,**
- **Les attaques cybernétiques,**
- **La tentation de jouer le rôle de « l'Apprenti Sorcier »**, c'est-à-dire d'attribuer aux systèmes d'IA la capacité de comprendre les intentions des utilisateurs plutôt que de suivre leurs instructions à la lettre,
- **« L'autonomie partagée »**, c'est-à-dire une collaboration harmonieuse entre les systèmes d'IA et les utilisateurs, permettant à ces derniers de reprendre le contrôle si nécessaire,
- **Et les conséquences socio-économiques de l'IA** : en d'autres termes, l'IA doit profiter à la société dans son ensemble et non uniquement à une élite privilégiée.

Quelques années plus tard et au-delà même des armes autonomes et des cinq menaces précédentes, un ensemble de dangers et menaces longuement débattus dans la communauté scientifique ont été compilés en 2020 par des chercheurs en informatique à l'**University College London** (**UCL**), par **Lewis GRIFFIN et ses collègues** dans un papier blanc appelé « *AI-enabled future crime* ».

Cette liste de 20 menaces et dangers, perpétrés ou potentiellement perpétrables par l'IA, a été classées en fonction de combien de dégâts elles pourraient causer, combien d'argent elles pourraient rapporter, à quel point elles seraient faciles à faire et combien il serait difficile de les repérer et de les arrêter.

➢ **Menaces sérieuses ou graves liées à l'Intelligence Artificielle**

- **Vidéos falsifiées** : appelé « **deepfake** » en anglais, cela consiste à usurper l'identité audio-visuelle d'une personne en la faisant dire ou faire des choses qu'elle n'a jamais dites ou faites, dans le but d'obtenir un accès à des données sécurisées, de manipuler l'opinion publique ou de nuire à la réputation de quelqu'un. Ces vidéos modifiées sont presque indétectables.

- **Détournement de voitures autonomes** : il s'agit prendre le contrôle d'un **véhicule autonome** pour l'utiliser comme une arme, par exemple pour commettre une attaque terroriste ou provoquer un accident.

- **Hameçonnage personnalisé** : Cela consiste à générer des messages personnalisés et automatisés pour augmenter l'efficacité du **phishing**, qui vise à collecter des informations sécurisées ou à installer des logiciels malveillants.

- **Piratage de systèmes contrôlés par l'IA** : perturber les infrastructures en provoquant par exemple une **panne d'électricité généralisée**, un engorgement du trafic ou une rupture de la logistique alimentaire.

- **Chantage à grande échelle** : collecter des données personnelles pour envoyer des messages de menace automatisés. L'IA pourrait également être utilisée pour générer de fausses preuves (par exemple de la « sextorsion »).

- **Fausse information rédigée par l'IA** : rédiger des articles de **propagande** qui semblent provenir d'une source fiable. L'IA pourrait également être utilisée pour générer de nombreuses versions d'un contenu particulier afin d'augmenter sa visibilité et sa crédibilité.

➢ **Menaces modérées ou de moyenne gravité liées à l'Intelligence Artificielle**

- **Robots militaires** : prendre le contrôle de robots ou d'armes à des fins criminelles. Une menace potentiellement très dangereuse, mais difficile à réaliser, car le matériel militaire est généralement très bien protégé.

- **Escroquerie** : vendre des services frauduleux en utilisant l'IA. Il existe de nombreux exemples historiques notoires d'**escrocs** qui ont réussi à vendre de coûteuses fausses technologies à de grandes organisations, y compris des gouvernements nationaux et l'armée.

- **Corruption de données** : modifier ou introduire délibérément de fausses données pour induire des biais spécifiques. Par exemple, rendre un détecteur insensible aux armes ou encourager un algorithme à investir dans un marché particulier.

- **Cyberattaque basée sur l'apprentissage** : perpétrer des attaques à la fois spécifiques et massives, par exemple en utilisant l'IA pour sonder les faiblesses des systèmes avant de lancer plusieurs attaques simultanées.

- **Drones d'attaque autonomes** : détourner des drones autonomes ou les utiliser pour attaquer une cible. Ces **drones** pourraient être particulièrement menaçants s'ils agissent en masse dans des groupes terroristes organisés.

- **Refus d'accès** : endommager ou priver des utilisateurs d'un accès à un service financier, à l'emploi, à un service public ou une activité sociale. Bien que non rentable en soi, cette technique peut être utilisée comme moyen de chantage.

- **Reconnaissance faciale** : détourner les systèmes de reconnaissance faciale, par exemple en fabriquant de fausses photos d'identité (accès à un smartphone, caméras de surveillance, contrôle de passagers...)

- **Manipulation de marchés financiers** : corrompre des algorithmes de trading afin de nuire à des concurrents, de faire baisser ou monter une valeur artificiellement, de provoquer un crash financier...

➢ **Menaces mineures ou de faible intensité liées à l'Intelligence Artificielle**

- **Exploitation de préjugés** : tirer profit des biais existants des algorithmes, par exemple les recommandations de YouTube pour canaliser les spectateurs ou les classements de Google pour améliorer le profil des produits ou dénigrer les concurrents.

- **Robots cambrioleurs** : utiliser de petits robots autonomes qui se glissent dans les boîtes aux lettres ou les fenêtres pour récupérer des clés ou ouvrir des portes. Les dommages sont potentiellement faibles, car ils sont très localisés à petite échelle.

- **Blocage de détection par IA** : déjouer le tri et la collecte de données par l'IA afin d'effacer des preuves ou de dissimuler des informations criminelles (par exemple, la pornographie)

- **Faux avis rédigés par l'IA** : générer de faux avis sur des sites tels qu'Amazon ou Tripadvisor pour nuire ou favoriser un produit.

- **Traque assistée par l'IA** : utiliser les systèmes d'apprentissage pour suivre l'emplacement et l'activité d'un individu.

- **Contrefaçon** : fabriquer de faux contenus, comme des tableaux ou de la musique, qui peuvent être vendus sous une fausse paternité. Le potentiel de nuisance reste assez faible dans la mesure où les tableaux ou musiques connus sont peu nombreux.

L'ensemble de ces menaces ont été résumé par l'équipe de chercheurs dans le schéma suivant :

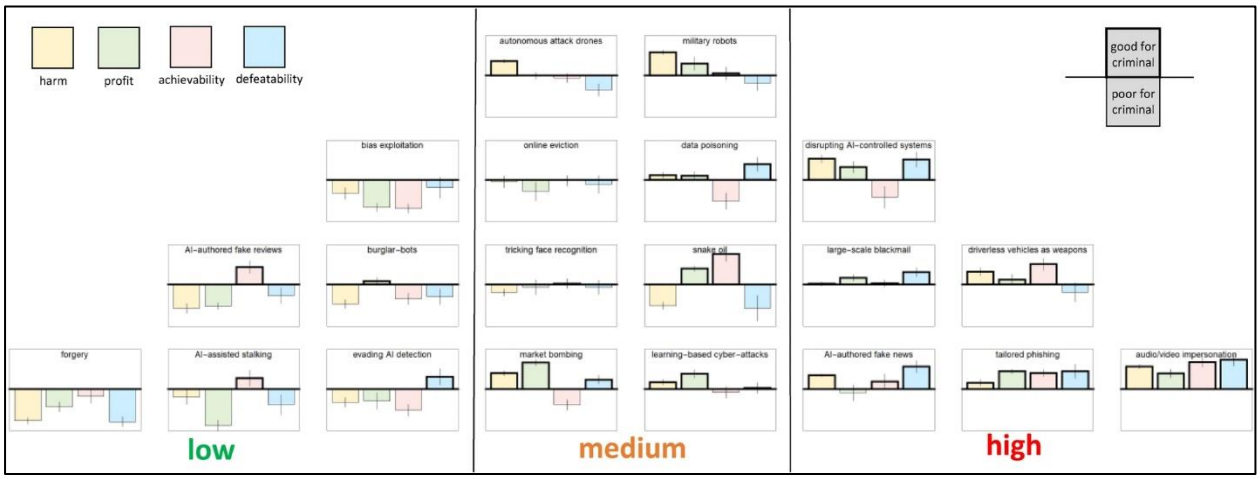


Figure 102 : Classement global des crimes utilisant l'IA (les plus dangereux à droite) - Source : « AI-enabled future crime »

Selon l'UCL, le verdict est unanime : les **deepfakes** sont la menace la plus grave associée à l'IA. Un exemple de **deepfake** connu et réalisé pour la démonstration est celui de **Barack OBAMA, ex-président des Etats-Unis**, dont la vidéo est consultable ici.

D'autres débats plus récents apparaissent et d'autres voix se lèvent pour alerter des dangers liés à l'IA. Parmi elles, celle du canadien **Yoshua BENGIO**, prix Turing 2018 en 2019 aux côtés du français **Yann LECUN** et du britannique **Geoffrey HINTON**, est particulièrement écoutée.

Dans une récente interview à la télévision canadienne (datant du 25/07/2023), Yoshua partage que selon lui, le danger imminent de l'Intelligence Artificielle pourrait se concrétiser en une vingtaine d'années, mais pourrait aussi arriver beaucoup plus tôt (en cinq ans) si l'IA continue à se développer aussi vite.

De nombreux chercheurs, dont Yoshua, pensent que les machines atteindront bientôt une intelligence équivalente à celle des humains, posant des **risques pour la sécurité nationale et la démocratie**. L'IA peut déjà créer de fausses images, vidéos et informations, et des systèmes avancés comme GPT-3 pourraient divulguer des informations sensibles.

Ces progrès sont préoccupants pour la sécurité et pourraient s'intensifier dans les 5 prochaines années. En matière de démocratie, **l'IA non réglementée** pourrait propager massivement de la désinformation. Yoshua, autrefois optimiste sur l'IA, est devenu pessimiste, surtout avec l'émergence de GPT-3, voyant l'égalité d'intelligence machine-humain se rapprocher.

Yann LECUN, son collègue chercheur français ayant eu le prix Turing 2018 en 2019 avec Yoshua, affirme que ce sont les humains qui contrôlent les machines et non l'inverse. Il suggère que l'IA, si elle était intelligente, n'auraient pas nécessairement les mêmes défauts que l'homme, comme la volonté de déstabiliser les sociétés. Cependant, Yoshua s'oppose à cette idée pour plusieurs raisons :

- Premièrement, les systèmes actuels sont formés pour imiter les humains, ce qui n'est pas une bonne chose.
- Deuxièmement, des recherches informatiques montrent que si nous donnons à une machine un objectif d'auto-préservation, nous perdons le contrôle sur cette machine.

Yoshua, lauréat du prix Turing 2018, pense donc que les machines pourraient développer un désir d'auto-préservation sans programmation directe. Il est incertain que les futurs systèmes d'IA restent sous contrôle humain, d'où la nécessité de recherches axées sur la sécurité.

Bien que la majorité des investissements actuels visent à renforcer l'IA, il est essentiel d'équilibrer puissance et sécurité. Il regrette ainsi de ne pas avoir davantage axé ses recherches antérieures sur la sécurité.

Le lauréat du prix Turing 2018 appelle donc à une réglementation rapide et coordonnée de l'IA à l'échelle nationale et internationale. Il insiste sur l'importance de la recherche en matière de gouvernance de l'IA, surtout face à des risques comme la perte de contrôle ou l'utilisation malveillante par des organisations terroristes. **Pour lui, il est crucial de se préparer technologiquement et réglementairement à ces scénarios.**

2.1.5 Un appel à mettre l'IA en pause

Plusieurs milliers de personnalités et chercheur en IA ont signé une nouvelle **lettre ouverte** (*la 2nde*) publiée ce 22 mars 2023. Celle lettre appelle à faire **une pause d'au moins 6 mois** dans la création de modèles de langage utilisés dans des IA générative et notamment ChatGPT.

Pour les auteurs de cette lettre ouverte, il est vraiment temps de prendre un peu de recul, et de réfléchir aux implications de ces technologies que l'on met actuellement entre les mains de millions d'individus.

Pour eux, les IA tels que ChatGPT ne sont pas simplement d'aimables « agents conversationnels » hyperdoués avec qui discuter. Ce document demande tout simplement la **mise en pause immédiate du développement de systèmes plus puissants que GPT-4**, dernière version de l'algorithme utilisé par ChatGPT.

Pour les auteurs, la R&D (recherche & développement) en IA devrait être recentrée sur l'amélioration de la précision, de la sécurité, de l'interprétabilité, de la transparence, de la robustesse, de l'éthique, de la fiabilité et de la loyauté des systèmes les plus sophistiqués d'aujourd'hui.

Le document souligne l'évolution rapide de l'Intelligence Artificielle, notamment avec GPT-4, et l'absence de cadre technologique et légal adapté.

Les auteurs se sentent dépassés par la rapidité des avancées qu'ils ont contribué à créer. Ils redoutent l'émergence d'une Intelligence Artificielle Générale qui pourrait rendre l'humanité obsolète.

Le ton du document est alarmiste, reflétant la vision du **Future of Life Institute**, financé en partie par Elon Musk. L'institut cherche des solutions technologiques face aux menaces potentielles, dont une IA dangereuse.

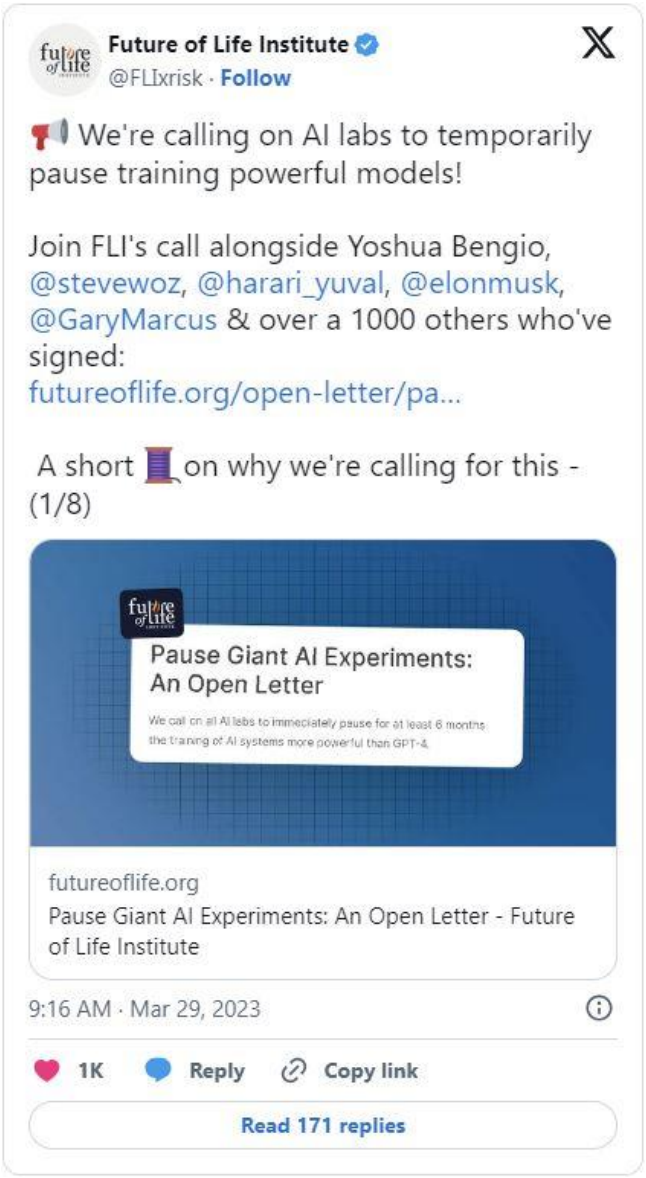


Figure 103 : Appel à la pause sur l'IA du Future of Life Institute - Source : [twitter.com]

On retrouve parmi les signataires une grande partie de ceux qui avaient déjà signés la précédente lettre ouverte sur les armes autonomes publiée en 2015 (évoquée précédemment) et parmi eux : **Elon MUSK**, **Steve WOZNIAK**, **Andrew YANG** (ex-candidat démocrate à la présidentielle américaine), ou encore **Yoshua BENGIO**.

Une pause ? mais pour quoi faire ? **Les auteurs souhaiteraient laisser le temps à la mise en place d'un système d'évaluation de la dangerosité d'une IA mais aussi pour avoir un cadre légal défini par les états**, conjointement avec la communauté scientifique. **Ce cadre légal permettrait de mettre en place une IA de manière responsable**, l'attitude des géants de la technologie (**Meta**, **Google**, **Microsoft**, etc.) étant jugée audacieuse et irresponsable par les auteurs. Ces « **big tech** » se jugent capables de s'autoréguler mais pour les auteurs, la véritable question est : **sont-ils réellement conscients des conséquences de leurs innovations sur le long terme** ?

Cette pause permettrait donc la mise en place de mesures de sécurité et inclurait :

- De nouvelles autorités pour réglementer l'IA,
- Des moyens de surveiller les systèmes d'IA,
- Des techniques pour aider à distinguer ce qui est réel de ce qui est artificiel,
- Des organisations capables de gérer les gros bouleversements économiques et politiques (surtout pour la démocratie) que l'IA va causer.

2.1.6 Augmenter la confiance dans l'Intelligence Artificielle plutôt que d'interrompre les recherches

La proposition **d'une pause de 6 mois**, donc finalement d'un moratoire, sur la recherche en IA générative ne répond pas en globalité aux principales problématiques associées à l'utilisation des systèmes d'IA, notamment dans des applications dites « **critiques** ».

Cette proposition de moratoire, par contre, **souligne l'importance d'un cadre réglementaire pour l'IA**, et notamment pour l'IA générative. Elle vient de parties prenantes qui ne disposent pas de tels cadres et qui ont même milité pour une innovation largement ouverte et réglementée à minima… **MUSK**, par exemple, est un des grands sponsor d'OpenAI mais aussi un des signataires pour mettre en pause…

Mais il est important de souligner qu'un point commun existe entre toutes ces voix qui se lèvent que ce soit dans cette lettre, ou la précédente, ou bien même dans les débats du siècle dernier ou ceux plus récents de 2023 : **le besoin d'avoir une IA réglementée, éthique et robuste en laquelle les utilisateurs peuvent avoir confiance, une IA responsable**.

2.2 L'IA responsable : la réglementation est en marche

« IA responsable », « IA éthique », « IA de confiance » : Ces expressions sont de plus en plus utilisées dans les discussions et les entreprises qui créent des produits avec de l'IA et qui y investissent de plus en plus.

Le potentiel incroyable de l'IA est compris et est exploité maintenant depuis plusieurs années. Mais de plus en plus de débats ont lieu sur son impact - et même de **sa dangerosité** - sur nos vies, nos sociétés et notre environnement.

En effet, une **prise de conscience collective** émerge autour du fait que l'IA, tel que présente aujourd'hui, doit être améliorée notamment en raison des risques des systèmes d'IA, des limites et dérapages dont certains ont été évoqués précédemment.

Ainsi, dans un contexte où l'IA est de plus en plus utilisée, il s'agit donc ici de clarifier ce qu'est vraiment **l'IA responsable**, un enjeu majeur, afin de comprendre toute son ampleur et sa complexité.

2.2.1 Définition de l'IA responsable

L'Intelligence Artificielle repose sur un **ensemble de méthodologies, d'algorithmes, de modèles et de logiciels qui, en soi, ne sont ni responsables ni irresponsables**.

Cependant, lorsqu'elle est utilisée et interagit avec les humains, l'Intelligence Artificielle, à travers ses algorithmes, peut réagir de manière inappropriée, trompeuse, déloyale ou biaisée. Dans ce cas, peut-on vraiment la considérer comme « responsable » ? Non !

C'est juste une technologie et sa responsabilité en cas de problème dépend vraiment de plusieurs choses :

- Comment elle a été **développée par les êtres humains**
- Comment elle est **utilisée par les êtres humains**
- Comment elle est **gérée par les êtres humains**

Il est donc important de réaliser et d'avoir conscience que la **responsabilité reste totalement humaine**. Dans cette perspective, la responsabilité est à la fois partagée par toutes les parties prenantes (responsabilité collective) et portée par chaque individu.

L'Intelligence Artificielle responsable est un concept ou un enjeu : celui qui consiste à prendre des mesures pour que l'IA soit développée, utilisée et gérée dans le but de minimiser le plus possible les risques pour ses utilisateurs.

Elle peut également s'apparenter à une approche de l'IA qui garantit qu'elle soit éthique et responsable. L'Intelligence Artificielle doit être fiable, équitable, compréhensible, transparente et respectueuse de notre vie privée.

Quand une organisation décide d'adopter l'IA responsable, elle s'engage à utiliser l'IA pour rendre la vie des gens meilleure tout en faisant tout pour réduire les dangers pour la société.

> *L'IA Responsable est une méthodologie qui consiste à anticiper les risques des systèmes d'IA, d'en maîtriser l'impact et d'en optimiser les bénéfices pour, au final, être redevable vis-à-vis des parties prenantes.*

***Ludovic GIBERT** (Crédit Agricole CIB) pour le groupe de travail de Innovation Makers Alliance - Digital & Technology (IMA-DT), Mai 2022*

Enfin, c'est surtout **une démarche** : mettre en place une IA responsable signifie construire la confiance dans l'IA en respectant un cadre légal, éthique, de confiance et de gouvernance des données.

Mais allier le potentiel des algorithmes de l'IA à des cadres qui inspirent confiance aux utilisateurs est **un grand défi pour le monde de la data**. Et qu'en est-il de la création du cadre en lui-même, qui serait un guide pratique pour ceux qui travaillent dans le numérique ?

De nombreuses autres définitions existent. La vidéo présente ici, de **Marc CARREL-BILLIARD**, Directeur de la R&D Accenture Monde, permet d'avoir une autre définition d'une façon concise.

Il est crucial de noter que pour nous, ces termes « **IA responsable** », « **IA éthique** » et « **IA de confiance** » se confondent et sont souvent utilisés de manière interchangeable. En effet, **rendre une IA « responsable » signifie suivre des directives, y compris éthiques, dans le but de créer une IA qui inspire confiance**.

2.2.2 Ses enjeux

L'IA responsable vise à garantir que les technologies d'Intelligence Artificielle sont conçues, développées et utilisées de **manière éthique et transparente au bénéficie de la société**, tout en minimisant les risques potentiels.

Ci-dessous quelques **enjeux liés à l'IA responsable** :

- **Éthique** : Des problèmes éthiques sont inévitables, et il est essentiel de les anticiper pour minimiser leurs impacts. **Les systèmes d'IA** doivent élaborés et opèrent dans le respect de l'éthique, ce qui implique qu'ils **ne doivent pas accentuer les biais existants**, ni être

discriminants ou nuisibles. Il est très important d'établir une **charte éthique interne** sur l'utilisation des données et de l'IA,

- **Transparence** : Les processus de prise de décision des algorithmes devraient être transparents et compréhensibles.

- **Accessibilité et Inclusivité** : L'IA devrait être **accessible à tous**, indépendamment de leur origine, de leur sexe, de leur âge ou de leur capacité. Cela signifie que **les systèmes d'IA devraient être conçus de manière inclusive** et ne pas exclure ou discriminer certains groupes.

- **Vie privée et Sécurité** : Avec l'augmentation de l'utilisation de l'IA, la protection des données personnelles et la sécurité des systèmes sont devenues des préoccupations majeures. Il est essentiel de garantir que les systèmes d'IA **respectent la vie privée** des utilisateurs et sont protégés contre les menaces potentielles.

- **Réglementation et Surveillance** : Pour garantir une utilisation responsable de l'IA, des réglementations appropriées et des mécanismes de surveillance sont nécessaires pour **guider et surveiller le développement et l'utilisation de l'IA**.

- **Responsabilité** : En cas d'erreurs ou de problèmes causés par un système d'IA, il devrait être clair qui est responsable. Cela nécessite une **réglementation juridique claire et des mécanismes de responsabilité.**

- **Durabilité** : Les systèmes d'IA devraient être conçus de manière à **respecter un développement durable**. Cela comprend la maîtrise des ressources de fabrication et d'usage des systèmes à base d'IA (énergie, volume de données, matières premières...).

- **Formation et Éducation** : Avec l'adoption croissante de l'IA, il est essentiel de **former les professionnels et le grand public** sur les implications et les utilisations responsables de l'IA.

- **Biais** : Les systèmes d'IA peuvent **renforcer ou perpétuer les biais existants** s'ils sont formés sur des données biaisées. Il est primordial de reconnaître et de corriger ces biais pour éviter la discrimination.

- **Autonomie et Contrôle** : À mesure que les systèmes d'IA deviennent plus autonomes, il est essentiel de garantir qu'ils restent **sous le contrôle humain** et ne prennent pas de décisions sans supervision ou contrôle approprié.
- **Impact sur l'emploi** : L'IA a le potentiel de remplacer certains emplois, ce qui peut entraîner des **perturbations sur le marché du travail**. Il est essentiel de comprendre et de gérer cet impact pour garantir une transition en douceur pour les travailleurs.

2.2.3 Les différentes initiatives (acteurs)

Il y a plusieurs initiatives majeures qui ont vu le jour pour encourager l'IA responsable et promouvoir une utilisation de cette technologie qui soit éthique et sûre.

Trois catégories d'acteurs sont principalement à l'origine de ces initiatives :

- **Acteurs publics :**
  - **En France, les comités pilotes** : La France a lancé plusieurs initiatives pour promouvoir une IA responsable aux niveaux national et international (européen, mondial).
    - Le **comité pilote d'éthique du numérique** a été formé au sein du Comité Consultatif National d'Éthique (CCNE). Composé de personnes venant de divers milieux, il a pour mission d'examiner de manière exhaustive et globale les questions éthiques liées au numérique et à l'Intelligence Artificielle.

    - Un **comité d'éthique de la défense** a été mis en place avec pour objectif de mener une réflexion éthique au bénéfice du ministère de la défense sur les questions liées à l'évolution de l'armement et à l'apparition de nouvelles technologies dans le secteur militaire. Cela comprend notamment le concept de « **soldat augmenté** » et le développement « **d'armes autonomes** ».
      Lien du principal projet : https://pfia2021.fr/journees/defense/Session1-AID-NP-IA.pdf

    - **Dans le secteur financier, l'Autorité de Contrôle Prudentiel et de Résolution (ACPR)** : L'ACPR est une autorité administrative indépendante adossée à la Banque de France, qui a pour mission de veiller à la stabilité financière et à la protection des clients des secteurs de la banque et de l'assurance.
      Chargée de surveiller l'activité des banques et des assurances pour s'assurer qu'elles respectent les réglementations en vigueur, elle s'intéresse également à l'impact de l'IA sur le secteur financier et travaille à l'élaboration de réglementations et de directives pour encadrer son utilisation de manière responsable et éthique.
      Lien du principal projet : https://lautorite.qc.ca/fileadmin/lautorite/grand_public/publications/professionnels/rapport-intelligence-artificielle-finance-fr.pdf

    - **L'Institut National de Recherche en Informatique et en Automatique (INRIA)** : L'INRIA est un établissement public français dédié à la recherche en informatique, en mathématiques et en sciences du numérique. Il a pour mission de produire des recherches au niveau international, de développer des innovations technologiques et de fournir à la France ainsi qu'à l'Europe des compétences stratégiques dans ces domaines.

L'institut contribue activement à l'innovation technologique en s'intéressant naturellement à l'Intelligence Artificielle et à ses **implications éthiques, sociales et techniques**. Il mène notamment des recherches pour comprendre les implications éthiques des technologies d'IA, notamment en ce qui concerne la vie privée, la transparence, la non-discrimination et l'équité et travaille en conséquence sur des méthodes et des outils pour **rendre les algorithmes d'IA plus transparents et explicables**.
Lien du principal projet : **REGALIA**, https://www.inria.fr/fr/regulation-algorithmes-projet-regalia

- **En Europe, la Commission Européenne sur l'IA** : Cette commission a été créée pour soutenir l'UE dans la mise en place d'une stratégie d'IA unifiée. Elle stimule l'innovation tout en assurant la protection des droits de l'homme, de la vie privée et de la sécurité.
  Le Groupe d'experts de Haut niveau sur IA (AI HLEG) joue un rôle crucial dans la mise en place de cette stratégie. En effet, les recommandations de ce groupe servent de base aux initiatives législatives de la Commission Européenne et des États membres, notamment la communication sur ce document « *Building Trust in Human Centric Artificial Intelligence* », la réalisation du « Livre blanc » sur l'Intelligence Artificielle appelé « Artificial Intelligence : a European approach to excellence and trust » ou même le « Coordinated plan on AI ».
  Lien du principal projet : **AI Act**, https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

- **Dans le monde :**
  - **Le Groupe de travail sur l'IA de l'OCDE** : Ce groupe, composé d'experts en IA du monde entier, travaille sur l'élaboration de recommandations politiques pour une IA responsable. Ils sont engagés dans la création de directives pour une IA responsable et encouragent la coopération internationale sur les questions liées à l'IA.

    Appelé également « **Observatoire des politiques d'Intelligence Artificielle de l'OCDE** », il a été créé à la suite de la publication de la Recommandation sur l'Intelligence Artificielle (Principes de l'IA de l'OCDE), qui est la première tentative de créer une norme internationale sur l'IA. Cette norme a été adoptée en mai 2019 par certains pays membres. Ces principes ont ensuite été utilisés pour établir les Principes du G20 en juin 2019. La liste des participants comprend de nombreux **États membres, des centres de recherche tels que le MIT, Harvard et l'INRIA pour la France, ainsi que de grandes entreprises comme IBM, Microsoft, Google/DeepMind et Facebook**.
    Lien du principal projet : https://oecd.ai/fr/ai-principles

  - **L'Initiative mondiale pour l'IA et le dialogue sur l'IA de l'UNESCO** : L'UNESCO a initié ce projet afin de promouvoir une IA responsable et inclusive à l'échelle mondiale. L'initiative vise à stimuler la collaboration internationale sur les

questions d'Intelligence Artificielle et à encourager un dialogue ouvert et transparent sur les conséquences de l'IA sur la société.
Lien du principal projet : **Recommandation sur l'éthique de l'Intelligence Artificielle**, https://unesdoc.unesco.org/ark:/48223/pf0000380455_fre

- **Le National Institute of Standards and Technology (NIST)** : Le NIST est une agence gouvernementale des États-Unis. Il fait partie du Département du Commerce des États-Unis. Le NIST est chargé de développer et de promouvoir des normes de mesure et de technologie qui améliorent la productivité, facilitent le commerce et améliorent la qualité de vie. Il est l'équivalent de **l'AFNOR** en France.

  Le NIST a développé un cadre de gestion des risques de l'IA, appelé « *AI Risk Management Framework* », destiné à améliorer la capacité à intégrer les considérations de confiance dans le développement et l'utilisation de l'IA.
  Par ailleurs, le NIST a lancé le programme TRAILS (« *Trustworthy and Responsible AI* ») visant à transformer la pratique de l'IA, d'une pratique principalement axée sur l'innovation technologique à une pratique axée sur l'éthique, les droits de l'homme et la responsabilité.
  Lien du principal projet : https://www.nist.gov/trustworthy-and-responsible-ai

- **L'Autorité des Marchés Financiers Canada (AMF Canada)** : L'AMF est une organisation gouvernementale, mandatée par le gouvernement du Québec pour réguler les marchés financiers de la province et fournir une assistance aux consommateurs de produits et services financiers. Elle a pour mission de **protéger les investisseurs, de maintenir des marchés financiers équitables et efficaces et de contribuer à la stabilité du secteur financier**.

  Au-delà de son activité première, elle a mené différents travaux autour de l'IA responsable, visant à donner des recommandations pour une utilisation responsable de l'IA dans le secteur financier. Elle a ainsi publié en novembre 2021 des recommandations dans un rapport appelé « *L'Intelligence Artificielle en finance : Recommandations pour une utilisation responsable* ».
  Lien du projet principal : https://lautorite.qc.ca/fileadmin/lautorite/grand_public/publications/professionnels/rapport-intelligence-artificielle-finance-fr.pdf

- **Les Etats-Unis** : En plus des efforts collectifs tels que les **10 principes de la Maison Blanche ou la mission du NIST** visant à créer un cadre de référence pour l'évaluation des risques liés à l'IA, on observe aux États-Unis une approche décentralisée avec des initiatives locales.

Par exemple, la ville de New York a pris des mesures sur les pratiques « *RH first attempt* » et la ville de Detroit a adopté une ordonnance pour exiger plus de transparence et de responsabilité dans les systèmes de surveillance.

- **La Chine** : Elle aspire à devenir le chef de file mondial en matière d'IA d'ici 2030. Elle a également mis en place une réglementation concernant les algorithmes d'IA, qui est entrée en vigueur en mars 2022.
  Cette réglementation exige que les objectifs soient **transparents** (par exemple pour les services de recommandation de produits ou services) et interdit aux entreprises basées sur l'IA de différencier les prix en fonction des données collectées. Toutefois, l'application effective de ce règlement n'en est qu'à ses débuts et sera probablement confrontée à la réalité du terrain économique.

➢ **Acteurs privés :**

- **L'Initiative pour l'IA responsable de Google** : Google a lancé cette initiative pour encourager des pratiques d'IA responsables en interne et dans l'industrie en général. Elle offre des lignes directrices pour une IA responsable, des outils pour aider les développeurs à créer des systèmes d'IA équitables, et des programmes de formation pour les employés.
  Lien du principal projet : https://cloud.google.com/responsible-ai?hl=fr

- **L'Initiative de Microsoft pour l'IA responsable** : À l'instar de Google, Microsoft s'efforce de promouvoir des pratiques responsables en proposant des principes directeurs et divers outils pour aider au développement de l'IA.
  Lien de principal projet : https://www.microsoft.com/fr-fr/ai/responsible-ai?activetab=pivot1%3aprimaryr6

- **Le Partenariat pour l'IA** : Cette initiative, dirigée par des organisations comme Microsoft, IBM, Amazon et Google, a pour but de promouvoir une IA responsable et d'établir des normes communes pour une utilisation éthique de l'IA. Elle travaille également à la mise en place d'un cadre de gouvernance pour l'IA.

- **Acteurs multi-types ou multipartites :**

  - **Le Partenariat Mondial sur l'Intelligence Artificielle (PMIA)** : Appelée également **Global Partnership on Artificial Intelligence (GPAI)** en anglais, cette initiative multipartite internationale rassemble des experts de différents domaines, notamment l'industrie, la science, de la société civile ou même des organisations internationales et gouvernementales. Son objectif est d'orienter le développement et l'utilisation responsable de l'IA, **tout en respectant les droits de l'homme, la diversité, l'inclusion, l'innovation et la croissance économique**.

    Il a été lancé par la France et le Canada à la suite du G7 de juin 2020. Le but du GPAI est d'approfondir la recommandation de l'OCDE publiée en mai 2019, en soutenant la recherche fondamentale en IA et ses applications industrielles et commerciales. Le GPAI est organisé en plusieurs groupes de travail sur différents thèmes, dont le Groupe de Travail sur l'IA Responsable. Ce groupe a pour mission de promouvoir et de contribuer au développement, à l'utilisation et à une gouvernance responsable des systèmes d'IA centrés sur l'homme, en accord avec les « Objectifs de Développement Durable » des Nations Unies. Actuellement, le GPAI compte 25 pays membres.
    Lien du principal projet : https://www.gpai.ai/projects/responsible-ai/gpai-responsible-ai-wg-report-2022.pdf

  - **Impact AI** : Cette association à but non lucratif vise à rassembler et à mobiliser l'ensemble de l'écosystème numérique, y compris les entreprises, les entités juridiques privées et publiques, les organismes de recherche et de formation publics et privés sur l'IA. Son but est de créer une approche collective de l'IA en France et de fournir des recommandations, des bonnes pratiques, des dossiers analyses. Elle organise également des événements et donne des formations spécifiques à l'IA responsable et éthique.
    Lien du principal projet : https://www.impact-ai.fr/app/uploads/2020/05/IMPACTAI_Guide_IA_dignedeconfiance_WEB.pdf

  - **La Global AI Action Alliance (GAIA) du World Economic Forum (WEF)** : L'association à but non-lucratif vise la coopération entre entités publiques et privées pour influencer les agendas mondiaux, surtout sur les sujets scientifiques. Elle prône l'indépendance, l'intégrité et des normes de gouvernance élevées.

    En 2021, elle lance la Global AI Action Alliance (GAIA) pour promouvoir une Intelligence Artificielle inclusive, transparente et fiable, rassemblant plus de 100 entités dont des entreprises majeures, universités et ONG. Dans le cadre de l'IA responsable, elle a également établi l'AI Governance Alliance avec des objectifs similaires.
    Lien du principal projet : **The Presidio Recommendations on Responsible Generative AI**, https://www.weforum.org/whitepapers/the-presidio-recommendations-on-responsible-generative-ai

- **Les organismes de normalisation ISO, IEC, IEEE, ITU, CEN-CENELEC et SAE** : Sont généralement des associations professionnelles, non gouvernementale et existe pour définir des normes.
  L'industrie de la normalisation s'est grandement intéressée à l'IA. Au niveau mondial, **l'ISO** et **l'IEC** ont mis en place un comité conjoint sur l'IA, appelé SC 42, qui s'attèle activement à élaborer des normes « universelles » en matière d'IA. Ces normes touchent à la terminologie, à la gouvernance, à la gestion de l'IA par les organisations, aux spécifications pour l'IA de confiance, à l'ingénierie, à l'interopérabilité, et plus encore.

  En Europe, le **CEN** et le **CENELEC** ont établi en 2021 le Comité Technique Mixte 21 « *Artificial Intelligence* » (JTC 21). La première mission du JTC 21 est d'accompagner la réglementation de l'IA en Europe en élaborant des normes « **harmonisées** ». Les normes européennes à venir auront donc une grande influence sur le marché européen, et peut-être même mondial.

  En outre, **l'IEEE** (Institut des ingénieurs en électricité et en électronique) a lancé la « *Global Initiative on Ethically Aligned design of AI Systems* », suite à la publication de « *Ethically Aligned Design* », une étude scientifique qui propose des principes de haut niveau et des recommandations pratiques.

- **Confiance.ai** : Il s'agit d'un progamme collaboratif rassemblant des acteurs académiques et industriels français majeurs (issus des domaines de la défense, des transports, et d'autres secteurs) ainsi que des organismes publics tel que l'INRIA.
  Le but de ce programme est de travailler sur l'IA de confiance, en particulier dans les systèmes critiques. Il a d'ailleurs publié en octobre 2022 un livre blanc appelé « *Towards the engineering of trustworthy AI applications for critical systems* » visant à caractériser et qualifier une IA de confiance dans le but de comprendre ce qui fait qu'une IA est considérée comme digne de confiance et comment cela peut être mesuré et évalué.
  Lien du principal projet : https://www.confiance.ai/wp-content/uploads/2022/10/LivreBlanc-Confiance.ai-Octobre2022.pdf?utm_source=Website&utm_medium=Clic&utm_campaign=LivreBlanc-Octobre2022

- **AI Ethical** : Le projet « AI Ethical », une initiative du Comité Innovation et Technologies de **Numeum** qui fait partie du programme Numérique Responsable de l'organisation, est une organisation professionnelle rassemblant un réseau de partenaires, y compris des organisations professionnelles, des écoles et des instituts interdisciplinaires d'IA.
  Il a pour objectif de promouvoir l'éthique dans le domaine de l'Intelligence Artificielle. Grâce à de nombreux dialogues et partages d'expériences, un guide pratique a été réalisé pour fournir une méthode pour mettre en œuvre les principes éthiques clés lors de la conception, du développement et du déploiement de solutions

d'Intelligence Artificielle. Il constitue un code de conduite volontaire qui guide le développement d'intelligences artificielles dignes de confiance, tel qu'encouragée par la Commission européenne.
Lien du principal projet : http://ai-ethical.com/wp-content/uploads/2021/09/2021-SN-Guide-Méthodo-IA-Ethiques-version-imprimée.pdf

Il existe encore beaucoup d'autres initiatives et celles listées sont les plus connues. **Le think-tank Future Society, membre du Partenariat Mondial sur l'IA (GPAI), a recensé plus de 200 initiatives et groupes de travail à travers le monde dédiés à l'IA**.

Tous ces efforts sont centrés sur l'idée d'une IA qui **place l'humain et son bien-être au cœur de ses préoccupations**. Le champ d'application de l'IA « responsable », tel que défini par le mandat du groupe de travail Future Society, est très vaste.

Cela se traduit par une multitude d'initiatives visant à orienter l'évolution et l'adoption de l'IA, ou à utiliser cette technologie pour réaliser le programme « *AI for Social Good* » afin d'atteindre les Objectifs de Développement Durable des Nations Unies pour 2030.

Ces programmes, menés par le monde académique, les secteurs public et privé, la société civile et les ONG, ont tous pour objectif commun de définir les attributs d'une IA responsable, que ce soit par des mécanismes formels ou informels.

D'ailleurs, les propositions de ces différentes initiatives convergent : elles se partagent en effet de nombreux principes majeurs. Parmi elles, quelques-unes tentent de se démarquer et il est particulièrement intéressant de rentrer plus en détails sur deux d'entre elles **: celle de l'UNESCO et celle de la Commission Européenne**.

2.2.4 Focus sur l'initiative mondiale pour l'IA de l'UNESCO

En 2018, **Audrey AZOULAY**, à la tête de l'UNESCO en tant que Directrice générale, initie un projet audacieux : **établir un cadre éthique mondial pour l'usage de l'Intelligence Artificielle**.

Trois ans plus tard, après une mobilisation massive d'experts internationaux et d'intenses négociations à l'échelle mondiale, les 193 États membres de l'UNESCO adoptent le 25 novembre 2021 le premier cadre éthique sur l'IA responsable appelé Recommandation sur l'éthique de l'Intelligence Artificielle (ou ici également en version Web).

> *Le monde a besoin de règles pour que l'Intelligence Artificielle profite à l'humanité. La Recommandation sur l'éthique de l'IA est une réponse forte. Elle fixe le premier cadre normatif mondial tout en donnant aux États la responsabilité de l'appliquer à leur niveau. L'UNESCO soutiendra ses 193 États membres dans sa mise en œuvre et leur demandera de rendre compte régulièrement de leurs progrès et de leurs pratiques.*

***Audrey AZOULAY** - Directrice générale de l'UNESCO, 25 novembre 2021*

Ce texte crucial définit des principes et valeurs universels pour guider la création d'un cadre juridique qui assure un développement responsable de l'Intelligence Artificielle.

2.2.4.1 Contenu de la recommandation

L'objectif de la recommandation est de matérialiser les bénéfices que l'IA peut offrir à la société tout en minimisant les dangers qu'elle représente.

Elle s'assure que les évolutions numériques encouragent le respect des droits de l'Homme et contribuent à l'accomplissement des objectifs de développement durable des Nations Unies, en se concentrant sur des problématiques comme la responsabilité, la transparence et la protection de la vie privée.

Elle inclut des sections politiques axées sur l'action concernant la gestion des données (gouvernance), la culture, l'éducation, le travail, l'économie et les soins de santé.

- **Protection des données :**

La Recommandation invite à dépasser ce que les entreprises et les gouvernements mettent actuellement en place pour assurer une protection plus étendue des individus, en garantissant la transparence, la capacité d'action et la maîtrise de leurs informations personnelles.

Elle soutient que **chaque individu devrait être en mesure d'accéder aux registres de ses données personnelles et même de les supprimer**. Elle envisage aussi des mesures destinées à renforcer la protection des données et la sensibilisation des individus concernant leurs propres informations, ainsi que leur droit de les gérer.

En outre, elle augmente la capacité des entités de régulation à l'échelle mondiale pour **assurer le respect de ces clauses**.

- **Interdiction de l'évaluation sociale et de la surveillance de masse**

La Recommandation **interdit formellement l'usage des systèmes d'IA pour l'évaluation sociale et la surveillance à grande échelle**. Ces technologies sont excessivement intrusives, violent les droits de l'homme et les libertés fondamentales, et leur utilisation est largement répandue.

Elle met également en avant que lors de la création de cadres législatifs, les États membres doivent garder à l'esprit que **la responsabilité** et le devoir de rendre compte reposent toujours finalement **sur les individus**, et que les technologies de l'IA ne devraient pas se voir attribuer une personnalité juridique.

- **Soutien au suivi et à l'évaluation**

La Recommandation établit aussi les fondations d'**outils** qui aideront à la construction de systèmes d'IA et leur application par les nations et les entreprises, mais aussi dans **l'évaluation de l'effet des systèmes d'IA sur les individus, la société et l'environnement**

Parmi eux, l'outil d'évaluation de l'état de préparation aide les États membres à déterminer leur niveau de préparation en termes d'infrastructure légale (juridique) et technique. Cet outil permettra d'améliorer la capacité institutionnelle des pays et recommandera les démarches adéquates à suivre pour garantir l'application effective de l'éthique.

De plus, la Recommandation incite les États membres à envisager la nomination d'un **responsable de l'éthique de l'IA qui se doit d'être indépendant**, ou tout autre dispositif, qui permettra de superviser des audits et une surveillance régulière.

- **Préservation de l'environnement**

La Recommandation met en avant l'idée que les acteurs de l'IA devraient favoriser des approches économes en matière de données, de ressources et donc d'énergie, et cela afin de contribuer à rendre l'IA un instrument clé dans la lutte contre le changement climatique et la résolution des problèmes environnementaux.

Elle incite les gouvernements à **évaluer l'impact environnemental, direct et indirect, tout au long du cycle de vie d'un système d'IA, incluant son empreinte carbone, sa consommation d'énergie et l'impact environnemental de l'extraction des ressources nécessaires à la création de ces technologies**.

Pour réduire les effets environnementaux négatifs des systèmes d'IA et des infrastructures de données, le document suggère d'investir dans des technologies respectueuses de l'environnement.

De plus, le document déconseille d'utiliser des systèmes d'IA qui ont un impact environnemental disproportionné.

2.2.5 Focus sur l'UE : l'AI Act

Actuellement, il existe plus de 200 documents, chartes et propositions cherchant à établir un cadre éthique pour l'Intelligence Artificielle. Face à cette profusion, un réel danger d'apparence éthique, sans réelle substance, se manifeste pour les institutions ou entreprises souhaitant mettre en avant l'IA.

Devant ce constat, la Commission Européenne s'est lancée dans la création de nouvelles normes et règles internationales pour assurer une IA fiable et digne de confiance. Elle compte bien devenir le centre mondial de l'Intelligence Artificielle responsable, fiable, éthique et digne de confiance.

2.2.5.1 Qu'est-ce que l'AI Act ?

L'AI Act est une proposition de règlement définie par la Commission Européenne et visant à établir des règles harmonisées concernant l'Intelligence Artificielle. Cette proposition a pour objectif de garantir que les systèmes d'IA mis et utilisés sur le marché de l'Union Européenne soient sûrs et respectent la législation en vigueur en matière de droits fondamentaux et les valeurs de l'Union. Elle cherche aussi à renforcer la souveraineté numérique Européenne.

Figure 104 : L'AI Act, proposition de cadre réglementaire de l'UE - Source : linkedin.com

La Commission Européenne propose ainsi **le tout premier cadre juridique sur l'IA**, qui aborde **les risques de l'IA** et positionne l'Europe pour jouer un rôle de premier plan à l'échelle mondiale.

2.2.5.1.1 Qu'est-ce qu'un système IA selon l'AI Act ?

Afin d'établir un **futur cadre valable juridiquement** et à l'épreuve du temps, l'AI Act pose tout d'abord une définition de ce qu'est un système d'IA, qu'il est important de souligner :

Un « ***Système d'Intelligence Artificielle*** » (système d'IA), dont fait l'objet cette proposition, est « ***un logiciel qui est développé au moyen d'une ou plusieurs des techniques et approches énumérés à l'annexe I et que peut, pour un ensemble donné d'objectifs définis par l'homme, générer des résultats tels que des contenus, des prédictions, des recommandations ou des décisions influençant les environnements avec lesquels ils interagissent*** ».

Dans l'annexe « I » est alors présenté une liste de types d'IA, qui pourrait être modifiée en fonction des avancées technologiques dans le domaine de l'IA, et qui englobent les différentes approches suivantes :

- **Approches d'apprentissage automatique**, incluant l'apprentissage supervisé, non supervisé et par renforcement, avec diverses méthodes comme l'apprentissage en profondeur.
- **Approches basées sur la logique et le savoir**, incluant la représentation des connaissances, la programmation basée sur la logique inductive, les bases de données de connaissances, les systèmes d'inférence et de déduction, le raisonnement symbolique et les systèmes experts.
- **Approches statistiques**, évaluation bayésienne, et méthodes de recherche et d'optimisation

La proposition de l'AI Act vise également :

- A assurer la **sécurité juridique** pour faciliter les investissements et l'innovation dans le domaine de l'IA,
- A renforcer la **gouvernance** et l'application effective de la législation existante en matière de droits fondamentaux et des exigences de sécurité applicables aux systèmes d'IA,
- A faciliter le développement d'un marché unique pour des applications **d'IA légales, sûres et dignes de confiance**. Elle établit donc des règles pour le développement, la mise sur le marché et l'utilisation de systèmes d'IA dans l'Union.

2.2.5.1.2 Pourquoi est-il nécessaire d'établir des règles pour l'IA selon l'AI Act ?

La proposition de régulation de l'IA vise à assurer aux citoyens européens qu'ils peuvent compter sur les avantages offerts par l'IA.

Bien que la majorité des systèmes d'IA soient **généralement sans danger** et puissent aider à surmonter de nombreux problèmes sociétaux, **certains d'entre eux peuvent présenter des risques** qui, s'ils ne sont pas gérés, pourraient conduire à des conséquences non souhaitées.

Par exemple, il est parfois difficile de comprendre la raison pour laquelle un système d'IA a opté pour une décision ou une action spécifique. Cela peut rendre compliqué de déterminer si quelqu'un a été traité de manière inéquitable, comme lors d'une décision d'embauche ou d'une demande d'assistance sociale.

Même si les lois actuelles offrent une protection, elles ne sont pas totalement adaptées pour répondre aux défis uniques posés par les systèmes d'IA.

Les directives suggérées incluent :

- Aborder les **risques directement liés à l'utilisation de l'IA**,
- Élaborer une liste des usages d'IA à haut risque,
- Établir des critères précis pour les systèmes d'IA destinés aux usages à haut risque,
- Instaurer des responsabilités distinctes pour les utilisateurs et les fournisseurs d'IA à haut risque,
- Prévoir une évaluation de conformité avant la mise en œuvre ou la commercialisation du système d'IA,
- Mettre en place des procédures de suivi après la commercialisation d'un système d'IA,
- Introduire un cadre de gouvernance à la fois au niveau européen et national.

2.2.5.1.3 L'AI Act au sein d'une stratégie numérique plus large

Le projet de règlement de l'Intelligence Artificielle IA Act est d'abord un texte européen qui a été constituer en s'appuyant sur de nombreuses études, rapports publiés ces dernières années par des groupes de réflexion sur l'Intelligence Artificielle.

Son texte s'inscrit dans le cadre de la mise en place de la stratégie numérique pour l'Europe au même titre que d'autres **textes européens** qui ont été récemment élaborés. En voici quelques-uns qui font complément à l'AI Act dans leur domaine respectif :

- **Déjà votés** :
  - **Digital Services Act (DSA)** : vise à réglementer les services numériques, en particulier les grandes plateformes en ligne.
  - **Digital Markets Act (DMA)** : vise à réglementer les « gatekeepers » ou gardiens du marché numérique, c'est-à-dire les grandes entreprises technologiques qui ont un impact significatif sur le marché intérieur, jouent un rôle de passerelle pour les utilisateurs et ont (ou sont sur le point d'avoir) un pouvoir économique durable
  - **Data Governance Act (DGA)** : vise à établir un cadre réglementaire pour la gouvernance des données. Voici une petite vidéo explicative (lien ici)
- En cours :
  - **Data Act (DA)** : vise à traiter des droits d'accès, d'utilisation et de partage des données industrielles

Les deux actes **Data Act** & **Data Governance Act** font partie de la stratégie plus large de l'Union européenne en matière de données, qui vise à **créer un espace européen unique des données pour faciliter l'accès et l'utilisation des données**, tout en garantissant des normes élevées en matière de protection des données, de sécurité et de respect de la vie privée.

Pour en revenir à l'AI Act, cette proposition s'inscrit dans un ensemble plus vaste concernant l'IA, incluant également le **plan coordonné sur l'IA**.

La stratégie vise à renforcer la coopération entre les États membres de l'UE, la Commission et d'autres acteurs pour développer et adopter l'IA en Europe, tout en respectant les valeurs européennes.

Le plan insiste sur la collaboration et couvre des domaines variés comme la recherche, l'éducation et l'éthique. En associant ce plan au cadre réglementaire de l'AI Act, la **Commission Européenne souhaite protéger les droits des individus et entreprises, tout en favorisant l'innovation et l'adoption de l'IA dans l'UE**.

2.2.5.2 Une approche qui se veut robuste

L'AI Act, tel que proposé par la Commission Européenne, vise à établir une approche robuste et complète pour réguler l'IA.

L'approche est conçue pour :

- S'adapter **proportionnellement aux risques associés** à différentes applications d'IA
- S'adapter aux **évolutions technologiques rapides** dans le domaine de l'IA

2.2.5.3 Une approche à l'épreuve du temps

L'IA est une technologie **en constante évolution. Elle nécessite donc une approche flexible qui puisse s'adapter aux innovations technologiques rapides**. La proposition vise à garantir que les applications d'IA demeurent sûres et fiables, même après leur introduction sur le marché.

Pour s'assurer que la réglementation reste pertinente et efficace face aux avancées technologiques, l'AI Act prévoit plusieurs mécanismes :

- **Flexibilité réglementaire** : Plutôt que de prescrire des solutions technologiques spécifiques, l'AI Act établit des objectifs et des principes généraux. Cela permet aux entreprises d'innover tout en respectant le cadre réglementaire.
- **Révision et mise à jour** : L'AI Act prévoit des mécanismes pour revoir et mettre à jour régulièrement les dispositions, en fonction des évolutions technologiques et des retours d'expérience.
- **Engagement des parties prenantes** : L'AI Act encourage la consultation et la collaboration avec différentes parties prenantes, y compris les experts en IA, les entreprises, les chercheurs

et le grand public. Cela garantit que la réglementation bénéficie des perspectives les plus récentes et les plus informées.

- **Surveillance continue** : L'AI Act souligne l'importance de la surveillance continue des systèmes d'IA après leur mise sur le marché pour s'assurer qu'ils restent conformes et sûrs.

L'AI Act a pour objectif de créer un cadre réglementaire capable de suivre l'évolution rapide de l'environnement de l'IA. Cela sera possible en réunissant ces divers éléments.

2.2.5.4 Une approche fondée sur les risques

Le projet de règlement AI Act est construit sur la base d'une approche proportionnée fondée sur le risque pour établir des règles. On distingue quatre catégories de système d'IA :

- Les systèmes d'IA à **risque inacceptables** : ils sont **interdits**,
- Les systèmes d'IA à **risque élevé** : ils sont soumis à des **exigences strictes de conformité**,
- Les systèmes d'IA à **risque limité** : ils sont soumis à des **exigences moindres et principalement à une obligation de transparence**,
- Des systèmes d'IA à **risque minimal** : ils sont soumis à des **recommandations de conformité**.

Voici graphiquement un résumé de ces quatre catégories :

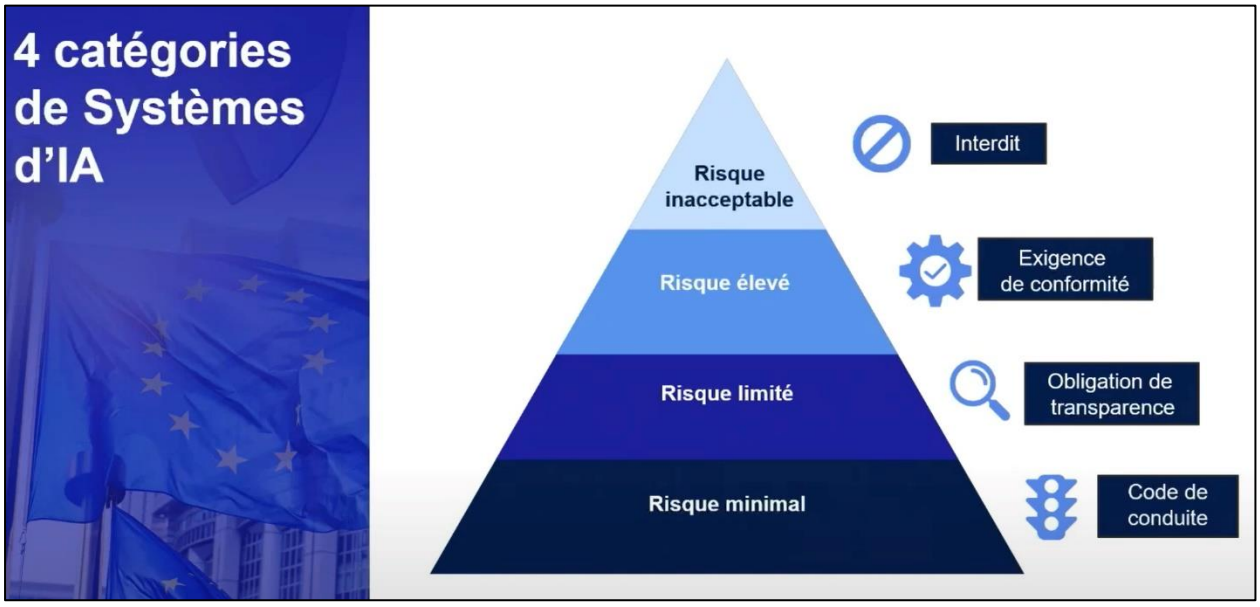


Figure 105 : Les 4 catégories de risques selon l'AI Act

En classant les applications d'IA selon leur niveau de risque, l'AI Act permet une réglementation adaptée à chaque situation. **Cette approche basée sur le risque garantit que les règles restent pertinentes, même si les technologies évoluent et donc est elle-même robuste en quelque sorte à l'épreuve du temps**.

2.2.5.4.1 Les systèmes d'IA à risque inacceptable

Le règlement énumère une liste limitative de systèmes d'IA **interdits** tels que :

- Ceux qui permettent la **manipulation mentale** avec des effets sur son comportement ou qui cherchent à influencer des comportements en utilisant des méthodes subliminales qui agissent sur le subconscient,

- Ceux qui **exploite les vulnérabilités des personnes** ou qui visent des groupes sensibles tels que les enfants ou les personnes en situation de handicap,

- Ceux qui sont destinés à évaluer ou établir **la notation de personnes en fonction de leur comportement social ou de leur caractéristiques personnelles** ou qui catégorisent les individus en fonction de leur fiabilité et qui pourraient entraîner un traitement inéquitable ou discriminatoire,
- Ceux qui permettent **l'identification biométrique**, en temps réel et à distance, dans les lieux publics à des fins répressives,

- Ceux considérés comme une **menace claire pour la sécurité**, les moyens de subsistance et les droits des personnes.

La liste des systèmes est actuellement en cours de discussion au Parlement européen. Néanmoins, comme exemple, pourraient être ciblés ceux permettant une **notation sociale par les gouvernements selon des comportements sociaux ou même les jouets utilisant une assistance vocale qui encourage les comportements dangereux**.

Un exemple plus détaillé : imaginons qu'une personne travaillant dans la commercialisation de produits de luxe a une idée de mettre en place une application, basée sur l'IA, à l'entrée de ses magasins pour filtrer les individus notamment sur la base de leur style vestimentaire ou à travers la reconnaissance faciale et pour faciliter l'accès aux individus qui présentent **pour lui** une probabilité d'achat très élevé. Un tel système d'IA donnerait donc une sorte de notation sociale à des personnes. Un tel système est envisageable ? La réponse est **non !!!**

Certaines pratiques d'IA particulièrement néfastes sont donc interdites en raison de leur **caractère contraire aux valeurs de l'Union Européenne**.

Cependant, des restrictions et des garanties spécifiques sont proposées en ce qui concerne certaines utilisations de systèmes **d'identification biométrique à distance à des fins répressives**. En effet, certaines exceptions sont identifiées par l'AI Act pour des situations spécifiques, telles que **la recherche de victimes de kidnapping, la prévention d'une menace terroriste spécifique, ou la détection et la poursuite d'auteurs de crimes graves**. Dans ces cas de figure, ces systèmes d'identification biométrique seront considérés dans la catégorie dite « **à haut risque** ».

2.2.5.4.2 Les systèmes d'IA à haut risque

Les systèmes d'IA à haut risque sont des systèmes autorisés et sont répartis en deux catégories :

- Les systèmes d'IA qui sont des **composants de sécurité de produits**.
- Les systèmes d'IA **autonomes.**

Ces deux types de systèmes devront faire l'objet d'une **évaluation de conformité**, un processus destiné à s'assurer que les systèmes d'IA à haut risque respectent les exigences réglementaires énoncées dans l'acte.

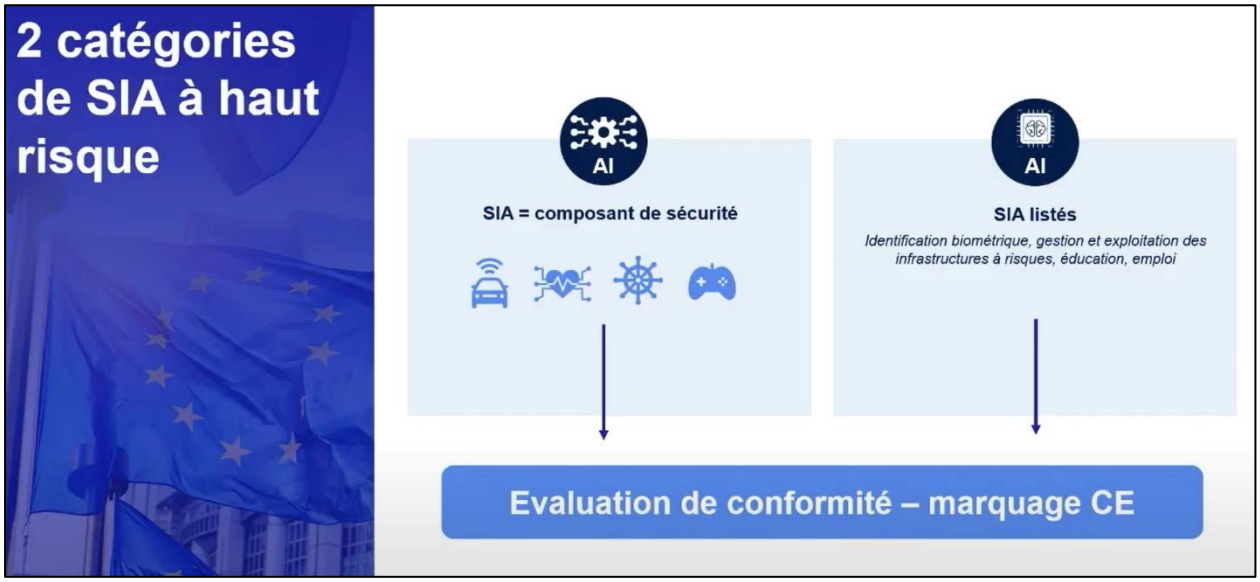


Figure 106 : Les 2 catégories de systèmes d'IA à haut risque

Concernant les **systèmes d'IA de type « composants de sécurité »**, ils ne pourront être mis sur le marché que si le système d'IA a fait l'objet d'une évaluation de conformité par un tiers autorisé.

C'est le cas des **jouets, des équipements de certains véhicules, des dispositifs médicaux**.

Ainsi, seulement une fois considérés comme conformes, les produits avec les systèmes d'IA pourront être mis sur le marché.

Pour les **systèmes d'IA autonomes**, ils font l'objet d'un référencement dans une liste capable d'évoluer dans le temps. Celle-ci a été établie en prenant en compte la fonction du système d'IA et le but dans lequel ce système va être utilisé.

Les systèmes d'IA considérés comme hautement risqués englobent la technologie de l'IA utilisée dans :

- Les **infrastructures critiques** ou domaines vitaux comme les transports, où la sécurité des citoyens pourrait être compromise,
- **L'éducation et la formation professionnelle**, dans le cadre d'une utilisation influençant l'accès à l'éducation et la trajectoire professionnelle d'un individu (par exemple, la correction automatisée des examens),
- **Les composants de sécurité des produits**, comme l'IA dans les interventions chirurgicales robotisées,
- **Le recrutement, la gestion des employés et l'accès à des postes indépendants** (comme les outils automatisés de sélection de CV),
- **Des services essentiels, privés ou publics** (tels que les systèmes d'évaluation de crédit qui peuvent refuser un prêt à un individu),
- **Les services d'ordre public** (répressifs) qui pourraient restreindre les droits fondamentaux (comme l'évaluation de la solidité des preuves),
- **La gestion des migrations, des demandes d'asile et des contrôles frontaliers** (comme la vérification des documents de voyage),
- **Les procédures judiciaires et les processus démocratiques** (par exemple, l'application automatisée des lois à des situations spécifiques).

L'AI Act répond également à la question des conditions nécessaires pour autoriser ces systèmes d'Intelligence Artificielle à haut risque. Pour cela, la proposition réglementaire liste les obligations strictes auxquelles sont soumises ce type de système d'IA :

- Mise en place de **mécanismes d'évaluation** et de réduction des risques,
- Utilisation de **jeux de données de haute qualité** pour réduire les biais et éviter des résultats discriminatoires,
- Conservation des **logs pour garantir une traçabilité** complète des actions,
- Fourniture d'une **documentation exhaustive et détaillée** décrivant le système et son fonctionnement, permettant aux autorités de vérifier sa conformité,
- Communication **d'informations transparentes et compréhensibles aux utilisateurs**,
- Mise en œuvre de **contrôles humains pertinents** (surveillance) pour minimiser les dangers,
- Garantie d'une grande **fiabilité (robustesse), sécurité et exactitude du système** (précision).

L'utilisation de ces dérogations nécessite l'approbation d'une autorité judiciaire ou d'une autre entité indépendante, avec des restrictions claires en termes de durée, de zone géographique et des bases de données consultées.

Comment seront contrôlés les exigences pour la mise sur le marché des systèmes à haut risques ?

Une fois que les **fournisseurs** de systèmes d'IA se seront dotés de moyens humains et technologiques qualifiés pour **appliquer ces textes et le système d'AI à haut-risque développé** (**STEP 1**) en prenant en compte les exigences de la Communauté Européenne proposés dans l'AI Act, les entreprises devront faire démarrer les démarches pour avoir **l'évaluation de conformité** pour ces systèmes d'IA (**STEP 2**).

Les systèmes en question sont alors **enregistrés dans une base de données Européenne permettant leur référencement** (**STEP 3**).

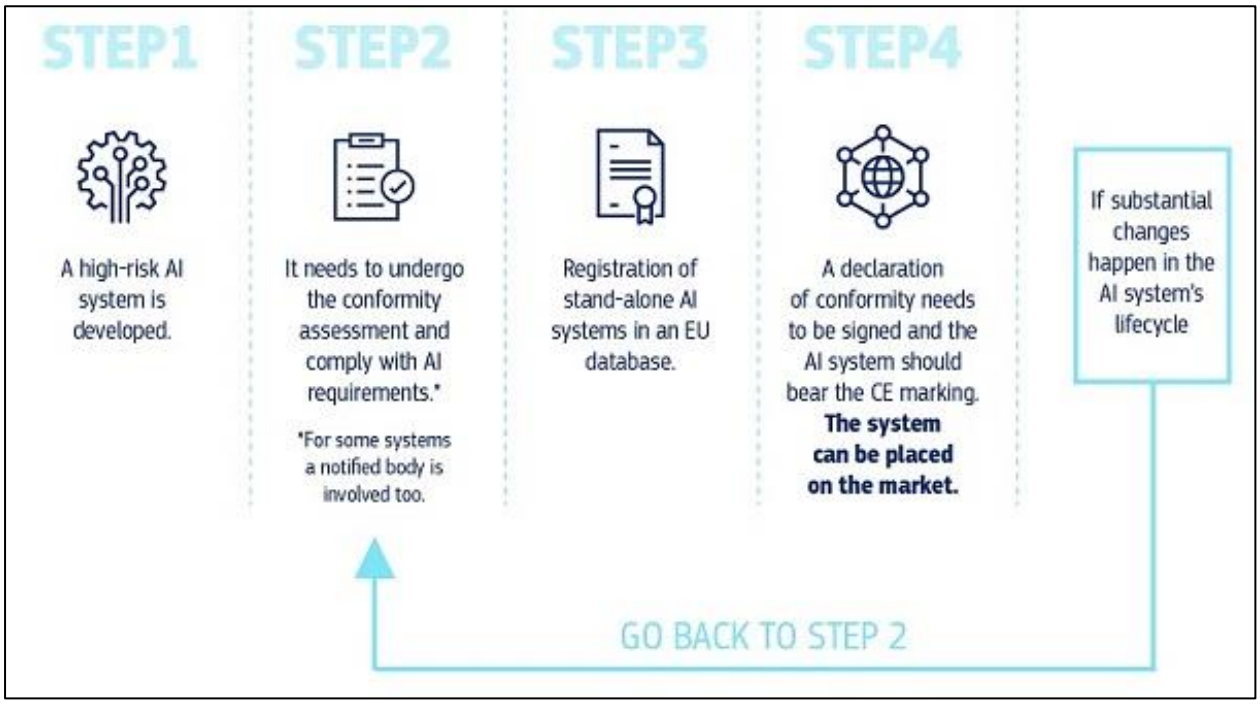


Figure 107 : L'évaluation de conformité pour les systèmes d'IA à haut risque

Dès lors, dans la proposition de l'AI Act est prévu un mécanisme qui consiste à faire évaluer la conformité des systèmes soit par un organisme tiers agréé tel que **l'AFNOR** où le **LNE**, soit de recourir à une procédure **d'auto-évaluation**.

Mais cette dernière n'est possible que sous réserve que des normes ISO existent et prennent en compte toutes les exigences du règlement proposé par l'AI Act.

Aujourd'hui, un certain nombre de normes ISO sur l'Intelligence Artificielle sont en cours d'élaboration pour répondre aux besoins d'évaluation de la conformité de ces types de système d'IA. Ils sont bien entendu établis par rapport à toutes les exigences prévues la proposition de l'AI Act.

Une fois l'évaluation de conformité réussie, les systèmes d'IA à haut risque peuvent **recevoir le marquage CE** (**STEP 4**), indiquant qu'ils sont **conformes aux exigences de l'AI Act et peuvent être mis sur le marché de l'UE**. C'est le fameux logo visible sur un grand nombre de produits et permettant d'attester de cette conformité.

Figure 108 : Logo 1 de marquage CE

Figure 109 : Logo 2 de marquage CE

Les certifications obtenue seront valides **pendant 5 ans** sauf si le système d'IA connait un changement majeur nécessitant de refaire la procédure d'évaluation de conformité.

L'obtention du marquage CE n'est pas la fin processus pour les fournisseurs lesquels, après la mise sur le marché de leurs systèmes d'IA à haut risque, se doivent continuer à **surveiller leurs performances, signaler tout incident ou dysfonctionnement grave, et prendre des mesures correctives si nécessaire**.

L'évaluation de conformité est donc un élément essentiel de l'AI Act, garantissant que les systèmes d'IA à haut risque respectent des normes élevées de **sécurité, de fiabilité et de protection des droits fondamentaux des utilisateurs**.

Concernant les systèmes d'IA aux **risques limités et minimum**, ces derniers sont soumis à des exigences moins importantes et il faut donc **retenir l'absence de mécanisme d'évaluation de conformité** pour ces deux types de systèmes d'IA.

2.2.5.4.3 Les systèmes d'IA aux risques limités

Ici sont concernés les systèmes d'IA susceptibles de **détecter les émotions, de faire des trucages vidéo ou les chatbots**. Les utilisateurs doivent savoir qu'ils interagissent avec une machine, leur permettant ainsi de décider en toute connaissance de cause s'ils souhaitent poursuivre l'interaction ou s'en éloigner.

Le texte prévoit alors que pour ces systèmes, les fournisseurs d'IA sont soumis à des **obligations de transparence dont doivent informer les utilisateurs qu'ils échangent avec un système d'Intelligence Artificielle**.

2.2.5.4.4 Les systèmes d'IA « risque minimal »

Les fournisseurs d'IA, et particulièrement les développeurs, sont encouragés à se **conformer à des codes de conduite et si le souhaitent à l'ensemble des exigences** qui sont prévues pour les systèmes d'IA à plus haut risque, notamment celles aux risques limités comme la transparence.

2.2.5.4.5 Pour tous les types

Une fois un système d'IA déployé sur le marché, les autorités supervisent le marché. **Les utilisateurs sont responsables de la surveillance et de l'intervention humaine, tandis que les fournisseurs mettent en place un mécanisme de surveillance post-commercialisation**. Les fournisseurs et les utilisateurs sont également tenus de signaler tout incident majeur ou anomalie.

2.2.5.4.6 Les sandboxes

Ce texte pourrait être perçu comme un risque de ralentissement à l'innovation notamment en cas de nécessité de réaliser une évaluation de conformité pour tout test d'une idée.

Mais le texte prévoit ce cas de figure **des bacs à sable réglementaires appelés « sandboxes »** qui permettront des tests, avant tout démarrage d'une phase « industrielle » et de mise en place sur le marché.

Les « sandboxes » se réfèrent à des environnements contrôlés où les entreprises et les développeurs pourront tester et expérimenter leurs solutions d'IA sans être immédiatement soumis à toutes les exigences réglementaires.

Elles permettront donc aux innovateurs de comprendre comment leurs systèmes fonctionnent dans des conditions réelles **sans craindre d'enfreindre les réglementations**, d'identifier les éventuels problèmes et de les résoudre avant une mise en œuvre complète, tout en assurant une certaine protection au public.

Les régulateurs pourront également en tirer profit de leur côté car les « sandboxes » leur permettront de mieux comprendre les nouvelles technologies et d'ajuster éventuellement la réglementation en conséquence.

2.2.5.4.7 Quelques chiffres clés sur la classification par risque

Selon une étude réalisée par **appliedAI Initiative GmbH** appelée « *AI Act : Risk Classification of AI Systems from a Practical Perspective* », parmi 100 systèmes d'IA sélectionnés, seulement 18% sont clairement classés comme étant à **haut risque**, 42% des systèmes d'IA sont classés comme étant à **faible risque**, tandis que **près de 40% restent à définir ou sont sujets à interprétation**. Un des systèmes d'IA pourrait être interdit.

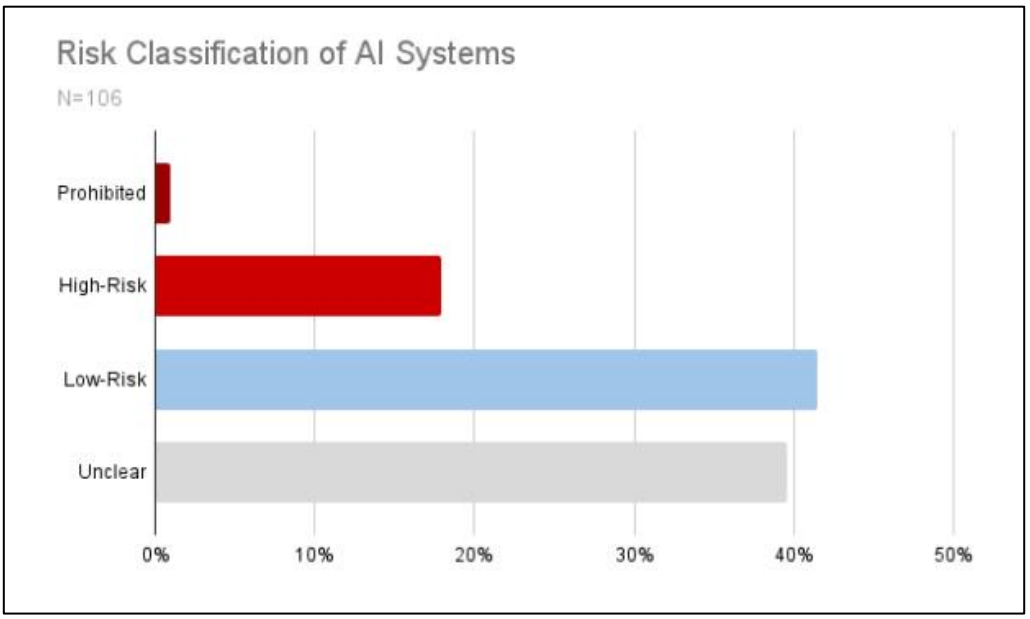


Figure 110 : Classification de 100 systèmes d'IA par risque de l'AI Act - Source : appliedAI Initiative GmbH

Selon cette étude, la proportion de systèmes d'IA à haut risque, en étant à 18%, est légèrement supérieure à la valeur maximale de 15% supposée par la Commission européenne (une estimation de 5 à 15% avait été faite).

Les 40% supplémentaires de cas incertains, où un système d'IA n'a pas pu être clairement classé comme un système d'IA à haut ou faible risque, portent la proportion potentielle de systèmes d'IA à haut risque à 58%.

Cette observation est centrale pour l'évaluation de l'impact, car **la plupart des exigences du règlement sur l'IA s'appliquent aux systèmes d'IA à haut risque et à leurs fournisseurs, pour lesquels les coûts économiques globaux et les efforts augmentent en conséquence**.

Dans ce contexte, un fournisseur serait plus susceptible de **choisir la catégorie à haut risque en cas de doute et cela afin d'éviter les risques potentiels**. Cependant, pour la Commission Européenne, la plupart des cas incertains pourraient être classés dans la catégorie à moindre risque ce qui montre **une volonté pour les fournisseurs d'aller plus loin que le nécessaire pour minimiser les risques**.

Aussi, d'après cette étude, **la plupart des systèmes d'IA à haut risque dans la proposition de l'Acte sur l'IA de l'UE se trouveraient dans les catégories suivantes : Ressources Humaines, Comptabilité et Finance, Service Client et Support, Sécurité Informatique et Juridique.**

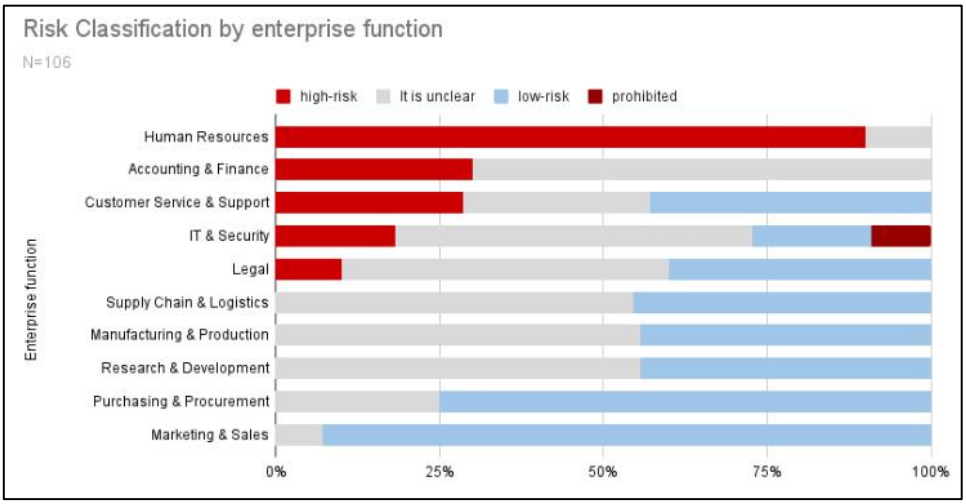


Figure 111 : Classification de 100 systèmes d'IA par risque de l'AI Act et par fonction dans l'entreprise - Source : appliedAI Initiative GmbH

En effet, c'est sans surprise. Plus de 75% des systèmes d'IA dans les ressources humaines sont classés dans la catégorie à haut risque car le domaine RH est référencé par l'AI Act dans cette catégorie pour la plupart des usages autour de la manipulation des CV.

Plus de 25% sont à haut risque respectivement dans le service client, la comptabilité et la finance. Le reste des systèmes IA de cette catégorie se trouve dans l'IT et la sécurité.

Des classifications incertaines se trouvent dans toutes les fonctions d'entreprise, mais principalement en comptabilité et finance à plus de 70% montrant qu'il reste certainement un flou permettant une catégorisation plus simple et directe. Seul dans le marketing et les ventes, la proportion de cas incertains est inférieure à 25%.

Effectivement, dans le tableau suivant, toutes les fonctions listés ont des classifications incertaines :

Enterprise Function	High-risk	Unclear	Low-risk	Prohibited	Total
Accounting and finances	3	7			10
Purchasing		2	6		8
Research and development		5	4		9
IT and security	2	6	2	1	11
Customer service	4	4	6		14
Logistics and supply chain		6	5		11
Marketing and sales		1	13		14
Human Resources	9	1			10
Production and manufacturing		5	4		9
Legal	1	5	4		10
Total	19	42	44	1	106

Figure 112 : Tableau qui montre le classement incertain du risque de beaucoup d'IA - Source : appliedAI Initiative GmbH

2.2.5.5 Qui est concerné ?

L'AI Act, s'il clarifie les systèmes d'IA concernés, identifie également les acteurs ciblés par ses propositions.

Il est donc important de rappeler que la mise en place d'un système d'Intelligence Artificielle implique de nombreux acteurs :

- Ceux qui conçoivent,
- Ceux qui structurent la base de données,
- Ceux qui développent,
- Ceux qui mettent sur le marché,
- Ceux qui mettent sur le marché par l'intermédiaire d'un tiers
- Ceux qui l'utilisent.

Sont donc ciblés **les fournisseurs (distributeurs, importateurs) et les utilisateurs**. Il y a bien deux notions autour des acteurs pris en compte dans ce projet de texte :

- **Fournisseurs d'IA :**

Ce sont tous les **acteurs impliqués dans le processus de fabrication** d'un système d'IA qui auront une part **de responsabilité dans la mise en œuvre, le contrôle et le suivi de cette nouvelle réglementation**.

Le texte prévoit aussi un **cadre concernant les petites entreprises** qui n'auraient pas les moyens de se doter de ces nouvelles compétences humaines, matérielles et technologiques, ou pour devoir former leurs employés ou recruter.

En effet, il prévoit des leviers pour leur permettre de se mettre en conformité. Sont prévus **des accès prioritaires aux sandboxes réglementaires ainsi que des mesures de sensibilisation** pour pouvoir les accompagner jusqu'à l'entrée en application de ce texte.

- **Utilisateurs d'IA :**

Les utilisateurs ont un rôle actif dans la surveillance, la formation, la collaboration et l'utilisation éthique des systèmes d'IA, en veillant notamment à la qualité des données et à la transparence. L'AI Act, un règlement européen, **s'applique à tous les acteurs, y compris ceux hors de l'UE**, s'ils développent ou utilisent des systèmes d'IA au sein de l'Union. Cela signifie que les entreprises non européennes doivent respecter les valeurs européennes pour accéder au marché de l'UE. L'AI Act a une portée extraterritoriale, tout comme le RGPD l'avait fait pour la protection des données. L'UE vise à établir des normes mondiales basées sur des valeurs humanistes pour la technologie.

2.2.5.6 Les dates clés

En avril 2021, la Commission européenne a introduit ce projet visionnaire qui a suivi une procédure d'approbation « traditionnelle » jusqu'à ce que les eurodéputés adoptent leur position en juin 2023. Cependant, la sortie publique des IA génératives, notamment ChatGPT en novembre 2022, a bouleversé le calendrier prévu pour l'adoption de cette réglementation.

Cette initiative réglementaire découle des efforts précédents de l'UE en matière d'IA, débutant avec le livre blanc publié le 19 février 2020, suivi par une consultation publique du 19 février au 14 juin 2020. Elle prend également en compte les différentes résolutions du Parlement européen du 20 octobre 2020 sur l'éthique, la responsabilité et la propriété intellectuelle et se termine par les conclusions de la Commission du 21 octobre 2020 qui plaident pour une IA respectueuse des droits fondamentaux.

Le 14 juin 2023, les eurodéputés ont intégré des mesures pour adresser les éventuels dangers des IA génératives telles que **ChatGPT** ou **Midjourney**. Parmi ces mesures, on note l'instauration d'un mécanisme d'identification des contenus générés (pour identifier les images falsifiées), la réalisation d'une analyse d'impact pour les systèmes jugés « à haut risque », et la mise en œuvre de dispositifs pour contrer les contenus interdits.

L'approbation de cette proposition par les eurodéputés en **juin 2023 marque le début de la phase finale** : elle permettra des discussions avec les États membres au sein du Conseil pour peaufiner le texte avant la fin de l'année 2023.

Selon les estimations les plus favorables, le **texte final pourrait être adopté d'ici la fin de 2023**, avec **une mise en application complète prévue pour 2026**.

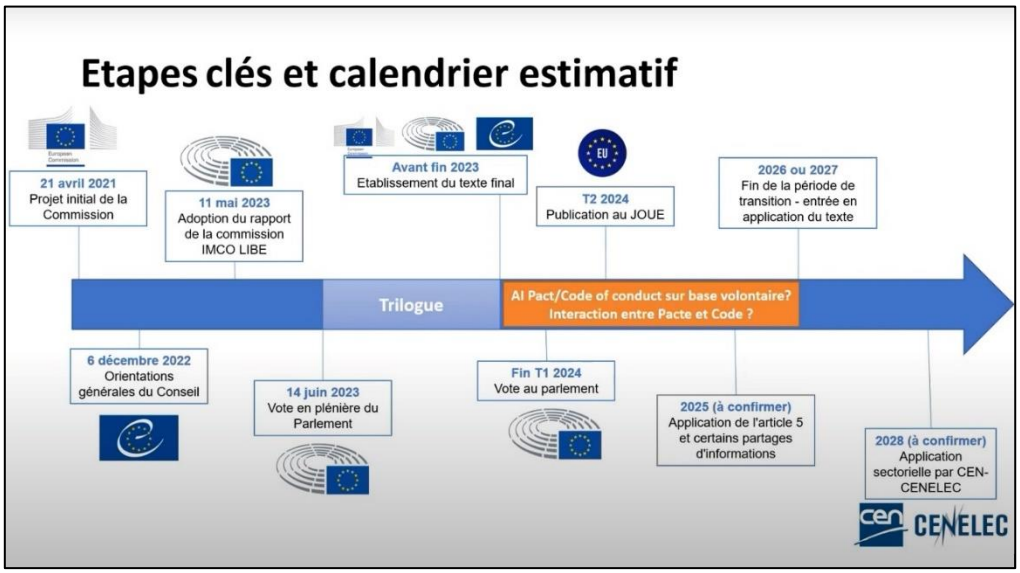


Figure 113 : Calendrier estimatif des futurs jalons de l'AI Act - Source : cenelec

Les acteurs ciblés ont donc d'ici là pour se mettre en conformité.

2.2.5.7 Les sanctions

Le projet de règlement suit la philosophie de la **R**églementation **G**énérale sur la **P**rotection des **D**onnées (**RGPD**) et en ce sens, a beaucoup de similitudes avec cette dernière.

D'abord, en ce qui concerne son champ d'application, tout **comme la RGPD, il cible non seulement les fournisseurs et utilisateurs d'IA en Europe, mais aussi ceux situés hors de l'UE qui exploitent les résultats d'une IA basée dans l'Union** (Article 2). Nous l'avons abordé précédemment.

Cette similitude est également évidente dans l'établissement d'un régime de sanctions. **Les amendes sont basées sur le chiffre d'affaires annuel mondial, et peuvent varier entre 2% et 6% selon la gravité de la violation (Article 71).**

Il appartiendra à **chaque état membre de déterminer l'applicabilité de ses sanctions**. Les amendes déjà prévues ont des **plafonds et elles varient selon les non-conformités au règlement**, selon les cas tels que la mise sur le marché d'un système d'IA interdit, des systèmes d'IA à haut risque ou les autres systèmes d'IA.

Voici les principales sanctions envisagées qui doivent être effectives, proportionnées et dissuasive :

- **Sanctions financières** : Les entreprises qui ne respectent pas le règlement peuvent être soumises à des amendes allant **jusqu'à 6% du chiffre d'affaires annuel total de l'entreprise dans le monde pour les violations les plus graves (système d'IA interdit)**. Les plafonds des amendes administratives qui sont prévus vont **de 10 millions à 30 millions d'euros**.

Figure 114 : Les sanctions prévues par l'AI Act

- **Interdictions** : Certains systèmes d'IA qui présentent un « **risque inacceptable** » peuvent être interdits sur le marché de l'UE.
- **Réparation** : Les entreprises pourraient être tenues de **corriger les systèmes d'IA non conformes avant de pouvoir les remettre sur le marché**.
- **Révocation de certifications** : dans le cas où un système d'IA à **haut risque a été certifié conforme mais on a découvert qu'il ne l'est pas**, cette certification peut être révoquée.
- **Retrait du marché** : Les autorités nationales de surveillance du marché pourront **tester, inspecter et, si nécessaire, retirer les systèmes d'IA du marché s'ils ne sont pas conformes** à la réglementation.
- **Responsabilité** : En plus des sanctions, les entreprises pourraient être considérées comme **responsables des dommages causés par leurs systèmes d'IA**, en fonction des réglementations nationales en matière de responsabilité.

A ce qui est prévu dans l'AI Act s'ajoute deux directives européennes :

➢ **Directive sur la Responsabilité liée à l'IA**

Le projet de directive vise à harmoniser les normes de l'UE pour les recours civils en cas de dommages dus à une faute non contractuelle liée à l'IA. Il se concentre sur la négligence des acteurs de l'IA et vise à compenser les victimes.

La « faute » est définie comme une violation d'une obligation de diligence. La directive facilite la procédure pour les victimes en présumant un lien entre le dommage et la négligence pour tout système d'IA, avec des règles spécifiques pour les systèmes à haut risque.

➢ **Directive sur la Responsabilité des Produits Défectueux**

L'AI Act envisage de moderniser la directive de 1985 sur la responsabilité des produits défectueux, incluant les logiciels d'IA comme « produits » et élargissant la notion de « dommage » pour couvrir la perte de données. Cette directive se concentre sur la responsabilité du fabricant pour les produits défectueux.

Les eurodéputés préfèrent une directive à un règlement, offrant une flexibilité aux États membres pour l'intégrer à leurs systèmes juridiques, mais cela pourrait compromettre l'harmonisation. Ils envisagent également de créer un Bureau de l'Intelligence Artificielle (« *AI Office* ») pour faciliter la mise en œuvre de la réglementation et résoudre les litiges transfrontaliers.

Enfin, voici un graphique indiquant la prise de sanction par le tuple « AI Act » + directives :

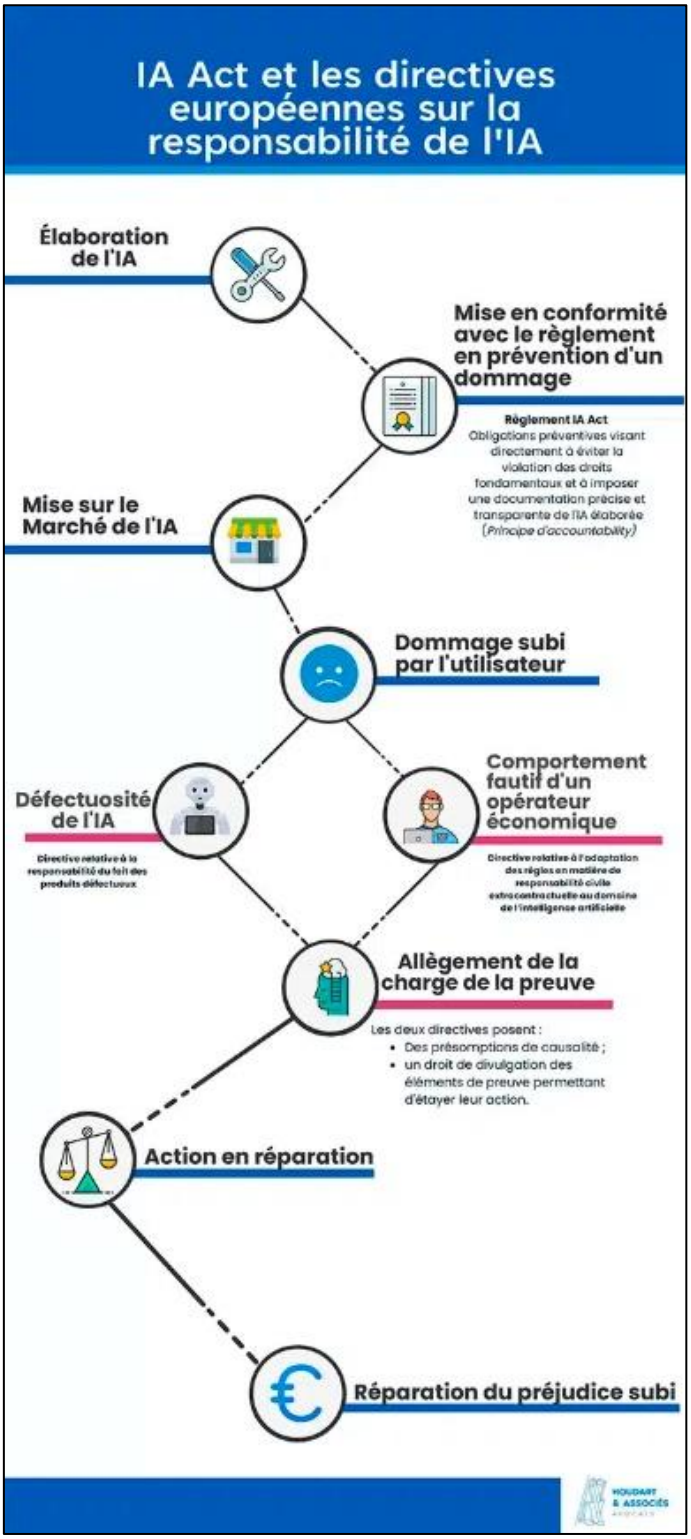


Figure 115 : Mise en application des sanctions de l'AI Act et des directives de l'UE - Source : houdart.org

Il est important de noter que l'AI Act est toujours en cours de discussion et de négociation au sein des institutions de l'UE, et que les détails finaux, y compris le régime de **sanctions**, pourraient être modifiés avant son adoption définitive prévue fin 2023.

2.2.5.8 Les sept principes de l'AI Act

Selon l'Union européenne, pour instaurer la confiance, une IA véritablement fiable doit répondre à trois critères essentiels, à maintenir durant tout son cycle de vie :

- **Conformité légale** : Elle doit se conformer à toutes les lois et réglementations en vigueur.
- **Éthique** : Elle doit adhérer à des principes et des valeurs éthiques.
- **Solidité** : Elle doit être robuste, tant d'un point de vue technique que social. Même avec de bonnes intentions, les systèmes d'IA peuvent entraîner des dommages non intentionnels.

Ces trois critères, bien que fondamentaux individuellement, ne sont pas suffisants en eux-mêmes pour garantir une IA fiable. **L'idéal est qu'ils opèrent en synergie et se complètent mutuellement.**

Sur la base de ces trois piliers, la Commission Européenne, à travers sa charte « *Lignes directives en matière d'éthique pour une IA digne de confiance* », énumère **sept principes majeurs pour atteindre une IA véritablement digne de confiance** :

- **Action et Contrôle humain** :
  - L'IA doit être au service de l'humain, respecter les droits fondamentaux et préserver l'autonomie humaine.
  - Elle doit intégrer des mécanismes de contrôle et d'intervention humaine.

- **Robustesse technique et sécurité** :
  - Les algorithmes doivent être conçus pour être résistants aux attaques, sécurisés et fiables.
  - Ils doivent gérer efficacement les erreurs et incohérences à chaque étape de leur cycle de vie.

- **Respect de la vie privée et gouvernance des données** :
  - Les données personnelles doivent être protégées, conformément aux réglementations comme le RGPD.
  - Les individus doivent avoir un contrôle total sur leurs données, évitant ainsi toute utilisation malveillante ou discriminatoire.

- **Transparence** :
  - Les systèmes d'IA doivent être traçables.
  - Ils doivent offrir une explicabilité claire de leurs prédictions, en tenant compte des contraintes professionnelles et réglementaires.

- **Diversité, non-discrimination et équité** :
  - Les systèmes d'IA doivent être conçus pour être justes, sans biais et accessibles à tous.
  - Ils doivent refléter une variété de capacités, compétences et besoins humains.

- **Bien-être sociétal et environnemental** :

- L'IA doit contribuer positivement à la société et promouvoir la durabilité environnementale.
- Elle doit aligner les intérêts commerciaux avec les préoccupations sociétales et environnementales.

➢ **Responsabilité** :

- Les systèmes d'IA doivent être auditables et minimiser les impacts négatifs.
- Des mécanismes de responsabilité doivent être mis en place, garantissant une obligation de transparence et de reddition de comptes.

Voici un graphique qui donne la déclinaison de ces 7 principes à partir des 3 piliers :

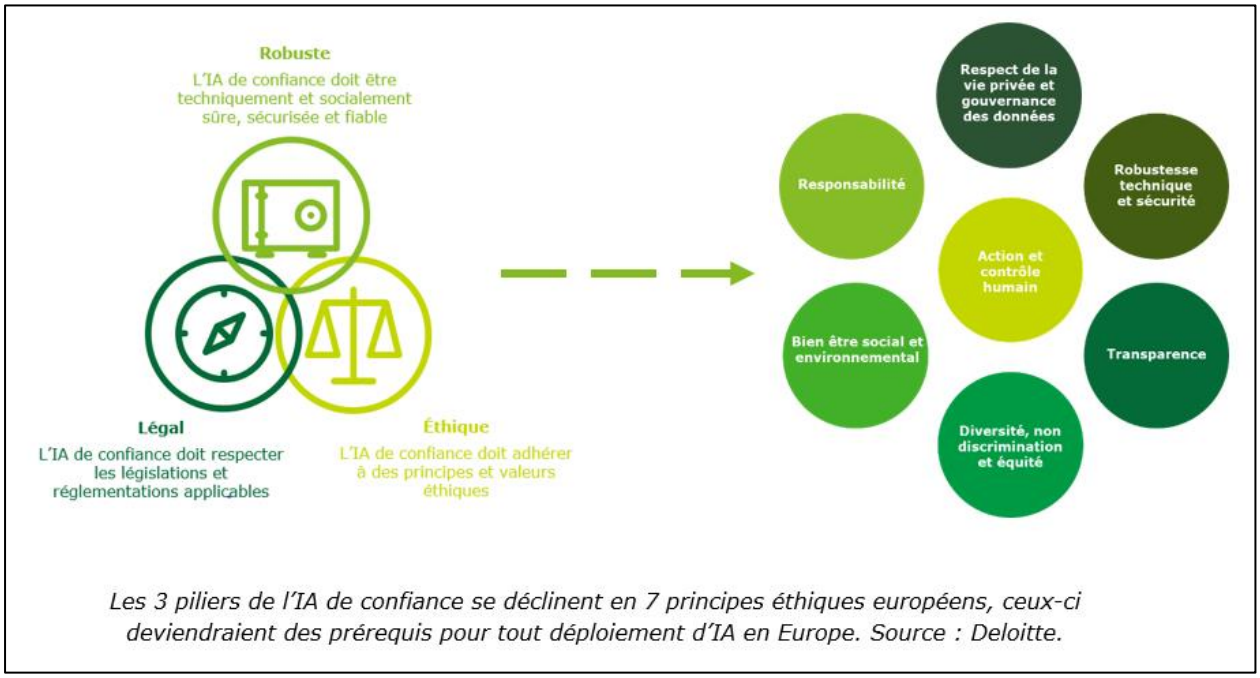


Figure 116 : Les 3 piliers et les 7 principes de l'AI Act - Source : deloitte.fr

Comprendre et intégrer ces principes éthiques au sein d'une organisation est une première étape vers une IA responsable.

Pour chaque principe, nous allons détailler ci-dessous sa définition et quelques retours d'expérience lorsque nous avons pu en obtenir.

Rentrons donc un peu plus dans les détails de chacun de ces principes.

2.2.5.8.1 Principe 1 : Action et contrôle humain

Les systèmes d'Intelligence Artificielle devraient promouvoir des sociétés justes en se consacrant au bien-être humain et à la protection des droits fondamentaux, sans compromettre ou altérer l'indépendance des individus. L'idée d'un outil ou service d'IA échappant au contrôle humain est une vision apocalyptique. Il est essentiel que tout système, qu'il soit partiellement ou entièrement automatisé, reste sous la supervision finale de l'utilisateur. Cette nécessité est donc inhérente à la définition des besoins et à la phase de création de l'application.

On pourrait établir quatre types d'actions où l'humain doit absolument intervenir ou avoir le contrôle :

1. **L'humain conçoit et décide :**

L'humain détermine le niveau de délégation d'un projet IA dès le début. Il est en charge de concevoir, contrôler et superviser l'IA, y compris ses impacts. **L'humain décide de l'utilisation du système à chaque étape** : cela fait référence à l'idée de « l'humain aux commandes ».

2. **L'humain est toujours dans la boucle :**

Le niveau de délégation dépend du contexte d'utilisation. **L'humain peut intervenir à différentes étapes du cycle de vie de l'IA pour superviser et guider**. L'intervention peut inclure la désactivation de l'IA, l'établissement de critères d'intervention ou la contestation d'une décision de l'IA. L'idée ici est d'avoir « l'humain dans la boucle ».

3. **Supervision sans intervention de l'humain :**

Le niveau de délégation est basé sur les responsabilités de l'IA. L'IA peut fonctionner sans intervention humaine directe dans certaines situations. Cependant, **des points de contrôle ou d'urgence doivent être prévus**. L'idée derrière ce principe est "l'humain en dehors de la boucle" mais ne signifie pas l'absence totale d'humain.

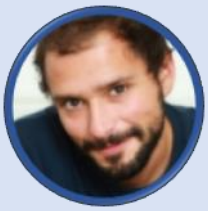
4. **L'humain donne son consentement :**

Avant l'utilisation de l'IA, **les utilisateurs et les citoyens doivent approuver son automatisation**. Ils doivent être informés des objectifs et capacités de l'IA. Les utilisateurs devraient pouvoir interagir avec l'IA, évaluer ou contester le système, et prendre des décisions éclairées.

Pour atteindre cet objectif, il serait judicieux d'utiliser un guide d'évaluation pour :

- Sensibiliser davantage les Data Scientists à l'importance cruciale du consentement,
- Clarifier l'objectif du système d'IA pour les utilisateurs et plus largement pour les citoyens,
- Former de manière appropriée tous les employés sur cette thématique.

L'utilisateur devrait être informé des critères utilisés par le système d'IA pour ses recommandations. Cela inclut la communication sur la fiabilité globale du système, la fourniture d'informations pour expliquer ou justifier une recommandation spécifique, et la description des conséquences potentielles de suivre une telle recommandation.



Grégoire COLOMBET
PDG de AI decision

Qui prend la décision, l'humain ou la machine ?

La mise en place de systèmes d'intelligence artificielle performants, responsables et éthiques ne garantit pas leur utilisation optimale.

Les projets chez CACIB ont montré que la collaboration entre l'homme et la machine n'était pas toujours la meilleure approche. En effet, parfois, les décisions prises uniquement par une machine ou uniquement par un humain peuvent être plus performantes que leur combinaison.

Par exemple, dans la détection de la fraude, un modèle d'IA peut prédire avec une grande certitude les transactions frauduleuses. Pour les probabilités élevées (plus de 95%), il serait plus efficace de transmettre la transaction directement à une équipe spécialisée plutôt qu'à une évaluation de premier niveau. À l'inverse, pour les probabilités moyennes, l'intuition humaine, comme appeler le client, peut-être plus utile que la machine. Il est donc très important de définir le rôle et la collaboration entre l'homme et la machine pour maximiser la performance, tout en réfléchissant à l'évolution des compétences et des rôles dans l'économie.

2.2.5.8.2 Principe 2 : Robustesse technique et sécurité

Ce principe englobe toutes les questions relatives à la **cybersécurité, aux dangers, à l'incertitude et à l'instabilité de l'environnement** concret dans lequel les systèmes sont implantés. Ce principe majeur se divise en trois sous-catégories : **la robustesse, la fiabilité des données produites par un système d'IA (ses résultats) ainsi que la défense contre les événements inattendus**.

- **Robustesse :**

Un système d'IA est robuste lorsqu'il maintient une performance constante face à des situations potentiellement risquées. Pour prévenir et gérer ces situations, il est essentiel d'identifier les vulnérabilités (les points faibles) du système, comme le type et la quantité de ses données

d'entraînement ou son mode de stockage, en fonction de ses objectifs. Par ailleurs, le système d'IA robuste doit disposer d'algorithmes robustes et sécurisés, capables de traiter les erreurs et les anomalies tout au long de son cycle de vie.

Par exemple, pour un système d'IA conçu pour distinguer différentes races de chiens, si les images de chiens-loups sont principalement sur un fond enneigé, cela constitue un risque – un risque « interne » au système.

On peut catégoriser les risques en deux types :

- Les **risques liés au système lui-même**, peuvent être dus à des anomalies dans les données d'entrée, d'inférence ou de sortie. Ils peuvent également résulter d'un écart entre les attentes du concepteur et sur quoi le système base sa décision pour réduire l'erreur.
- Les **risques causés par des facteurs externes**, tels que des cyberattaques ou d'autres interventions malveillantes.

Ainsi, les systèmes d'IA devraient être conçus avec une vision axée sur la prévention des risques, garantissant leur fiabilité et leur conformité aux attentes, tout en minimisant les vulnérabilités susceptibles d'être exploitées, que ce soit au niveau du système lui-même ou de l'infrastructure sous-jacente, matérielle ou logicielle.

- **Fiabilité des résultats :**

Un système d'IA est considéré **comme fiable si les résultats qu'il produit sont cohérents avec l'objectif visé et peuvent être reproduits lors de tests répétés sous les mêmes conditions**.

Les systèmes d'IA, en évoluant constamment, ne garantissent pas toujours des résultats reproductibles. La prévisibilité de leurs résultats n'est assurée qu'avec une certaine probabilité.

La reproductibilité est plus complexe pour les systèmes avancés avec de nombreux paramètres. Les variations dépendent des méthodes utilisées et de la structure du modèle.

Reproduire fidèlement les résultats nécessiterait une grande puissance de calcul à cause de la diversité des paramètres.

- **Défense contre les événements inattendus :**

L'IA est un outil pouvant servir à des fins positives, comme la surveillance agricole ou la protection des espèces, ou négatives, comme le système de notation sociale en Chine basé sur la reconnaissance faciale.

À l'instar de l'énergie nucléaire, tous les usages de l'IA ne peuvent être anticipés car on ne peut anticiper ni prévenir tous les abus futurs.

Les modèles de Machine Learning s'entraînent sur des exemples concrets. Un modèle formé uniquement avec des images de chats blancs pourrait ne pas reconnaître un chat noir.

Des perturbations mineures, comme flouter une image, peuvent fausser la classification d'un algorithme. Pour renforcer la robustesse, on peut inclure des données perturbées à l'entraînement ou détecter quand un modèle risque de se tromper.

Cependant, des attaques adverses peuvent toujours affecter les résultats. La robustesse des modèles présente donc des défis, notamment en matière de fiabilité. La fiabilité de chaque modèle doit être évaluée, et des systèmes de sécurité peuvent être ajoutés pour détecter des anomalies.

Voici ci-dessous les quelques exemples concrets de tests de robustesse selon les cas énumérés précédemment :

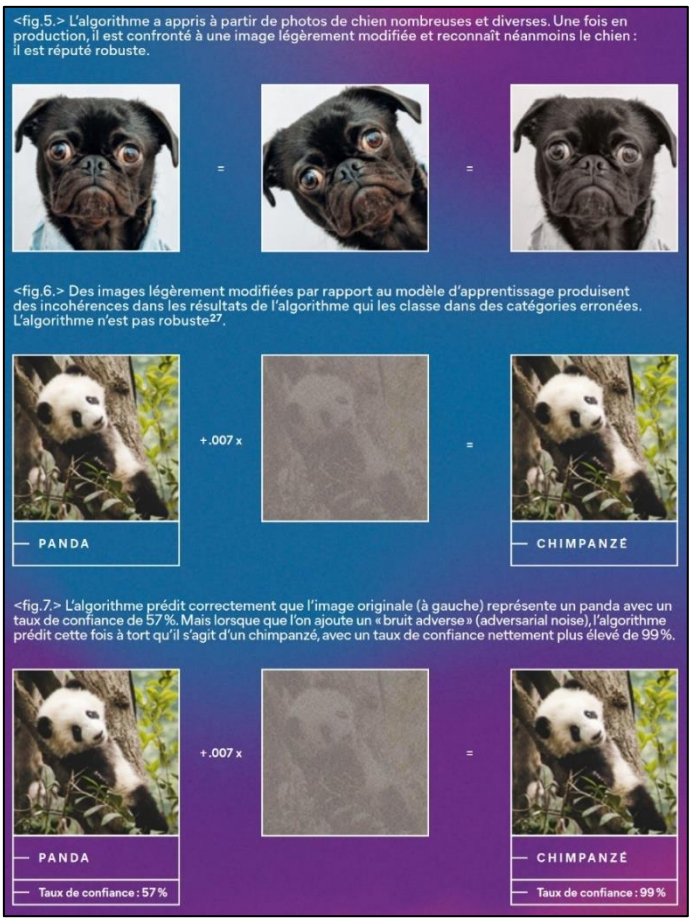


Figure 117 : Exemples concrets de tests de robustesse - Source : [impact-ai.fr]

Retour d'expérience :

Pour garantir la fiabilité des systèmes d'IA, **Axionable** (entreprise spécialisée dans la data et l'IA) adopte une approche en 3 parties alignée sur le cycle de développement de l'IA :

- **Conception** : Des critères précis sont établis pour définir le comportement attendu de l'algorithme, éliminant ainsi les ambiguïtés et anticipant les dérives possibles.
- **Avant mise en production** : Des scénarios basés sur des problématiques identifiées sont testés avec des données pour vérifier la conformité de l'algorithme et détecter des cas non anticipés.
- **Après mise en production** : Une surveillance continue est mise en place pour maintenir la performance de l'IA, car sa performance peut diminuer avec le temps. Trois méthodes sont utilisées pour détecter et corriger les erreurs :
    - **Ground Truth** : Comparaison des performances prévues et réelles basées sur des données concrètes.
    - **Data Drift** : Surveillance des changements dans les données d'entrée qui peuvent affecter la précision de l'IA.
    - **Concept Drift** : Identification des changements dans l'objectif initial de l'IA qui peuvent altérer la fiabilité des résultats.

Pour **Google DeepMind**, la robustesse d'une IA est construite autour de 3 « piliers » que sont les :

- La définition de manière « robuste » du but du système (conception robuste)
- La résistance aux perturbations (robustesse technique)
- La surveillance (monitoring et supervision) pour garder le contrôle

<fig.4.> Les trois piliers d'une IA solide et fiable selon DeepMind[25]

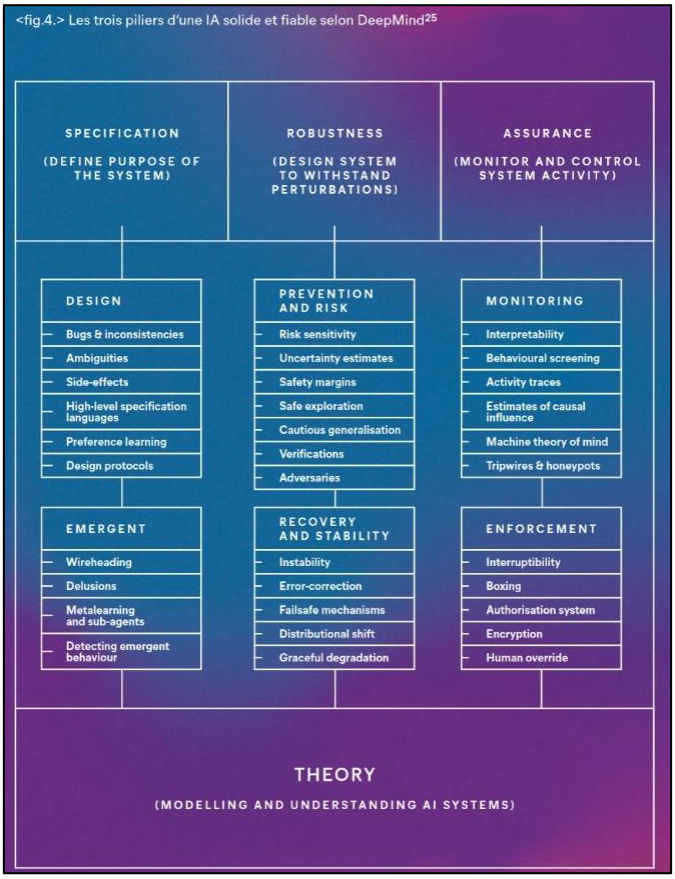


Figure 118 : Les 3 piliers d'une IA robuste pour Google DeepMind - Source : [impact-ai.fr]

2.2.5.8.3 Principe 3 : Respect de la vie privée et gouvernance des données

Ce principe majeur a une connotation « data » (données en français) et se divise en deux sous-catégories : le respect de la vie privée notamment dans l'utilisation de données personnelles par un système d'IA (ses résultats) ainsi que la gouvernance des données, quelles que soient leurs types.

- **Respect de la vie privée :**

Le respect de la vie privée et la protection des données personnelles sont deux droits essentiels reconnus au niveau européen. Ces droits peuvent être compromis par les systèmes d'IA lorsqu'ils **traitent des informations personnelles**. Ces droits sont encadrés **en Europe par le RGPD, en France par la Loi Informatique et Libertés, et par d'autres règlements spécifiques (comme le secret médical).**

Juridiquement, chaque phase de la vie d'un système d'IA doit se conformer à ces réglementations. Le cadre européen stipule, entre autres, que les données personnelles doivent être réduites, traitées pour des objectifs précis, basées sur le consentement de l'individu ou sur une autre base légale. Chaque individu a le droit de consulter, rectifier ou supprimer ses données. Il est essentiel de différencier les « données personnelles » des « données sensibles ». Le traitement de ces dernières est généralement interdit, sauf exceptions.

Certains systèmes d'IA, lorsqu'ils traitent des données à haut risque, doivent réaliser une analyse d'impact sur la protection des données, une exigence qui sera renforcée par le futur AI Act.

Outre les obligations légales, la gestion des données personnelles par un système d'IA pose des questions de confiance. Les individus doivent être assurés que leurs données ne seront pas utilisées de manière inappropriée. De plus, une gouvernance adéquate, comprenant des mesures techniques et organisationnelles, est essentielle pour garantir la qualité, l'intégrité et la confidentialité des données.

- **Gouvernance des données :**

Un système d'IA respectueux de la vie privée est l'ensemble des actions destinées à garantir la qualité, l'intégrité et la pertinence des données dans le contexte d'utilisation spécifié. Cela implique également la mise en place d'une confidentialité des données notamment via des procédures d'accès aux données et de directives pour leur traitement, tout en veillant à la protection de la vie privée.

- **L’anonymat dans les systèmes d’IA** :

L'anonymisation est un concept crucial lorsqu'un système d'Intelligence Artificielle traite des données personnelles. Elle vise à respecter le principe de minimisation des données. Selon la CNIL, l'entité française responsable de la protection des données, l'anonymisation est un processus qui rend impossible l'identification d'une personne par n'importe quel moyen, de manière irréversible. Il est essentiel de comprendre que l'anonymisation elle-même est un traitement de données soumis au RGPD. Seules les données résultant de ce processus échappent à cette réglementation.

Pour qu'un ensemble de données soit considéré comme véritablement anonyme, il doit répondre à trois critères essentiels :

- **Impossibilité d'individualisation** : On ne doit pas pouvoir isoler un individu spécifique dans l'ensemble de données.
- **Impossibilité de corrélation** : Il ne doit pas être possible de relier différents ensembles de données concernant la même personne.
- **Impossibilité d'inférence** : On ne doit pas pouvoir déduire de nouvelles informations sur une personne avec une quasi-certitude.

Dans certains cas, la pseudonymisation pourra être utilisée en lieu et place de l’anonymisation (voir partie « La gouvernance des données »).

Il est aussi important de noter que, bien que l'anonymisation concerne principalement les données personnelles, des techniques similaires peuvent être utilisées pour des données non personnelles mais sensibles.

- **Consentement et contrôle dans l’usage des données :**

Les systèmes d’IA traitant des informations personnelles doivent se conformer aux directives, régulations et autres standards relatifs à la protection de ces informations. **L'objectif principal de ces directives est d'accroître l'autonomie des individus dans la gestion de leurs données et leurs interactions avec les systèmes d’IA. Le consentement à l'utilisation de leurs données et la clarté sur le fait qu'ils interagissent avec un système d’IA sont essentiels**. De plus, ces normes visent à assurer aux individus l'effectivité des droits qu'ils peuvent exercer sur leurs données.

Pour assurer une protection des données conforme aux normes européennes, des directives spécifiques concernent le **stockage et le transfert de données personnelles**. La **CNIL** indique que le transfert de données **en dehors de l'UE et de l'EEE est autorisé**, à condition de garantir **une protection adéquate des données**. Ces transferts doivent être régulés par divers **instruments juridiques**.

Il est à noter que, concernant les données **non personnelles**, la confidentialité de certaines informations (telles que les **secrets nationaux, industriels ou commerciaux**) est aussi cruciale que la protection de la vie privée. Tout service numérique offert aux organisations collectant des données

non personnelles sensibles devrait également offrir à ces entités un certain niveau de **contrôle et d'indépendance sur la gestion de leurs données**. Ces préoccupations sont intégrées dans la stratégie numérique de la Commission Européenne, notamment avec la proposition du **Data Act (DA**) et du **Data Governance Act (DGA)**.

- **Les différents types de données traitées par l'IA :**

L'IA peut s'appuyer sur différentes catégories de données pour son entraînement, parmi lesquelles :

- **Données structurées** : Elles suivent une organisation définie. Par exemple, elles peuvent inclure le nom d'un client, l'article acquis, les opérations bancaires, les coordonnées clients ou les bulletins de salaire.

- **Données non structurées** : Ces données ne suivent pas une structure ou un modèle prédéfinis, englobant des éléments tels que les images, les vidéos ou des textes sans format spécifique.

- **Données synthétiques** : **Créées numériquement via des algorithmes basées sur des données réelles**, elles imitent des données réelles reproduisant leurs propriétés statistiques sans compromettre la confidentialité. Elles sont couramment utilisées pour l'entraînement des modèles d'IA, les tests, ou pour **garantir la confidentialité en substituant des informations sensibles par des versions fictives**. Ces données ont la même capacité prédictive que les données originales, mais sans les problèmes de confidentialité associés. Par exemple, pour modéliser la relation entre la taille et la vitesse des athlètes, on peut utiliser une fonction linéaire pour générer des données synthétiques. Si d'autres variables comme la tension artérielle sont ajoutées, des équations plus complexes sont nécessaires, et l'IA peut aider à créer un modèle approprié.

- **Les principales caractéristiques des données en contexte d'IA**

Les données mobilisées pour l'apprentissage et l'opérationnalisation de l'IA possèdent des particularités distinctes. Lors de leur utilisation par l'IA, elles offrent un éventail étendu d'opportunités, mais sont également sources de défis :

- **Volume et variété des données** : Les algorithmes d'IA sont aptes à gérer d'énormes volumes de données, qu'elles soient structurées ou non. Cela renforce la nécessité de veiller constamment à la qualité et à l'intégrité des données ce qui peut être un défi en soi.
- **Versionnement des données** : Les processus d'IA sont souvent évolutifs et s'appuient sur différentes moutures de données. Il peut alors être difficile de déterminer quelle version des données a été utilisée pour l'entraînement d'un modèle donné et de retracer son évolution.
- **Agilité des données** : Outre la rapidité avec laquelle les données sont produites et traitées, certains systèmes fonctionnent en temps réel, s'adaptant aux flux de données instantanés.

Dans ces contextes, une gouvernance solide est primordiale, avec des modèles supervisés par l'homme (« *humain dans la boucle* ») et des modèles « *humain en dehors de la boucle* » (hors-ligne) prêts à intervenir si le système en temps réel ne traite pas correctement les données reçues.

- **Les principaux enjeux et défis de la gouvernance des données en matière d'IA**

La gouvernance des données dans le contexte de l'IA responsable représente un défi colossal pour les entreprises et les gouvernements à travers le monde. Ils doivent veiller à :

1. **Gérer le volume croissant de données :**

L'IA génère une quantité phénoménale de données qui ne cesse de s'accroître. Face à cette abondance, les organisations doivent garantir la qualité des données tout en respectant les normes éthiques et réglementaires. Comment assurer une qualité optimale lorsque les données sont si diverses et nombreuses ?

2. **Maintenir la qualité des données :**

Assurer la qualité des données est une tâche complexe, d'autant plus que leur volume augmente. Plusieurs problèmes peuvent compromettre cette qualité, tels que :

- **L'incohérence** : les données peuvent présenter des formats et des détails variés, rendant difficile l'analyse et la détection de tendances ou de modèles dans les données.
- **Le bruit** : des erreurs, des fautes de frappe ou des informations superflues (non-pertinentes) peuvent perturber un ensemble de données initialement cohérent.
- **Le manque de contexte** : sans un contexte adéquat, il est parfois difficile de saisir la signification de certaines données.
- **La qualité des données en fonction de la variabilité des sources** : les données peuvent provenir de sources diverses, dont la fiabilité et la qualité peuvent varier.
- **Les ambiguïtés (double sens)** : certaines données peuvent être sujettes à des interprétations multiples.

Pour pallier ces défis, des actions comme l'exploration, le nettoyage, la validation et l'intégration des données sont nécessaires. Toutefois, même avec ces mesures, garantir une qualité irréprochable reste complexe. Et les données synthétiques, quand bien même pourraient résoudre certains problèmes ci-dessus, ont leurs propres limites comme la représentation de la complexité et de la variabilité de la réalité, rendant ainsi sous-performant les modèles d'IA.

3. **Assurer la confidentialité et la sécurité des données :**

Les données sont cruciales pour l'IA et doivent être sécurisées pour gagner la confiance des utilisateurs. Les informations sensibles, telles que les données personnelles, financières ou médicales, requièrent une protection renforcée. Le cryptage, les contrôles d'accès et la formation des employés sont essentiels pour assurer la confidentialité et la sécurité. Les entreprises doivent également respecter les réglementations, comme le RGPD en Europe.

4. **Favoriser la transparence des systèmes d'IA en termes de données :**

Les modèles d'IA, souvent complexes, doivent être rendus transparents pour les utilisateurs. Cela implique d'expliquer les décisions prises par l'IA et de détailler la logique de l'IA (voir principe *Transparence*) mais surtout ici, d'un point de vue « gouvernance des données » de montrer comment les données ont été utilisées.

En effet, la transparence doit également apporter aux utilisateurs de la clarté sur la manière dont leurs données sont recueillies, conservées et exploitées par les systèmes d'IA, ainsi que sur le traitement de leurs données personnelles par ces systèmes.

Cette transparence renforce la confiance des utilisateurs et leur permet de comprendre les risques et avantages liés à l'utilisation de l'IA.

2.2.5.8.4 Principe 4 : Transparence

La transparence des modèles d'Intelligence Artificielle passe par les techniques d'explicabilité. L'explicabilité vise à déchiffrer ou **à rendre compréhensible la manière dont un algorithme ou un modèle opère, les raisons de ses décisions et les résultats qu'il produit, en s'assurant de son bon fonctionnement**.

Ces techniques offrent, en plus du résultat du modèle ou de l'algorithme, une vision de l'impact positif ou négatif de chaque variable d'entrée sur la prédiction.

Cependant, même si on comprend comment un algorithme fonctionne, cela ne signifie pas toujours que nous comprenons pourquoi il fait certaines prédictions. Parfois, la logique interne peut sembler mystérieuse, ce qu'on appelle **l'effet « boîte noire »**. Certains types d'algorithmes sont traditionnellement considérés comme plus explicables que d'autres.

Il est important de différencier l'explicabilité de la transparence, de l'interprétabilité et de l'auditabilité :

- La **transparence** est une façon de rendre les décisions algorithmiques compréhensibles. **Elle offre un accès à la documentation, au code source et aux modèles**. De manière plus générale, elle fournit des informations sur l'utilisation d'un système d'IA, ses objectifs, les types de données utilisées et la logique sous-jacente.
- **L'auditabilité** concerne la capacité à **évaluer de manière pratique un algorithme**, non seulement pour comprendre ses prédictions, mais aussi pour l'évaluer selon d'autres critères, tels que la performance et le traitement des données.
- La distinction entre **explicabilité** et **interprétabilité** est sujette à débat. **L'explicabilité est souvent liée à une compréhension technique de l'algorithme**, adaptée à une perspective d'audit, tandis que **l'interprétabilité est moins technique et s'adresse davantage aux utilisateurs ou aux personnes affectées par l'algorithme**.

Il faut retenir que cette notion est importante pour les organisations car elle :

- **Facilite l'adoption de l'IA par les utilisateurs** en leur permettant de comprendre les raisons derrière les décisions de celle-ci.
- Sécurise la création de modèles d'IA **en favorisant un échange entre les Data Scientists et les experts métiers sur les critères de décision de l'IA**.

Voici différentes méthodes d'interprétabilité des IA, illustrées par un exemple d'une IA estimant le salaire d'un individu à partir de son CV :

- **Méthode de simplification de l'IA** : Création de règles claires et compréhensibles reflétant le fonctionnement de l'IA. *Exemple : Si l'expérience dépasse 5 ans et que le niveau d'études est bac+3, alors le salaire est de X euros.*
- **Méthode basée sur l'importance des variables** : Identification des variables ayant le plus d'impact sur la décision de l'IA. *Exemple : L'expérience professionnelle est un facteur déterminant majeur dans l'estimation du salaire.*
- **Méthode des exemples explicatifs** : Pour une décision spécifique, recherche d'exemples similaires traités de la même façon par l'IA. *Exemple : Pour un CV spécifique, on identifie d'autres CV semblables qui mènent à une estimation salariale similaire.*
- **Méthode d'explication générée par l'IA** : L'IA elle-même fournit une justification, souvent sous forme textuelle. *Exemple : J'ai estimé le salaire à X euros car les compétences et l'expérience de cette personne sont peu courantes sur le marché.*
- **Méthode structurelle d'explicabilité** : Des éléments internes du modèle décrivent un groupe d'individus. *Exemple : Sur la base d'un CV, l'IA évalue un critère représentant la rareté du profil sur le marché. Ce critère influence ensuite l'estimation du salaire.*

Pour assurer une interprétabilité efficace d'un système d'IA, il est essentiel de fournir des explications adaptées à chaque groupe en interaction avec le système, afin que chacun puisse comprendre les résultats. Par exemple, pour un système donné :

- Le créateur de l'IA nécessitera des détails sur les erreurs observées lors de la calibration des paramètres ou de l'utilisation du jeu de données d'entraînement,
- L'utilisateur final voudra savoir comment l'IA a déterminé sa décision dans sa situation spécifique,
- Un organisme de certification souhaitera avoir une vue d'ensemble de la logique opérationnelle de l'IA.

L'explicabilité est-elle essentielle pour les entreprises ? Est-il crucial pour les entreprises d'expliquer leurs décisions basées sur l'IA ?

La nécessité d'explicabilité dépend du contexte. Par exemple, pour une entreprise de services, justifier une erreur de commande à un client est une démarche courante.

Toutefois, dans des situations comme l'évaluation de dossiers médicaux pour définir les indemnisations des clients victimes d'accidents, chaque détail du dossier, tel que les séquelles, doit être scruté avec précision. Une erreur dans ce contexte est impensable !

Il serait inacceptable, voire illégal, de refuser une indemnisation sur la base d'un score algorithmique de 0,7 alors qu'un seuil de 0,8 est nécessaire. Un tel refus serait incompréhensible pour un client.

Par conséquent, l'explicabilité est primordiale, surtout dans des secteurs comme la santé, le juridique ou la défense où la confidentialité des données est vitale. En effet, **certains projets basés sur l'IA ont été abandonnés ou ont échoué à cause de ces enjeux. L'explicabilité représente donc une chance pour ces projets de voir le jour et de prospérer**.

Matthieu CAPRON
Responsable Design authority IA
DataLab Groupe Crédit Agricole

Il existe deux types d'explicabilité :

- Explicabilité globale : Comprendre les principales variables ou caractéristiques qui influencent le modèle dans son ensemble.
- Explicabilité locale : expliquer pour chaque modèle, pour chaque décision du modèle, la raison de la prise de décision avec un poids par variable ou autre

Pour les méthodes d'explicabilité, elles peuvent reposer sur différents éléments lesquels peuvent être combinées entre elles :

- Taux de confiance : Le modèle fournit une prédiction avec un niveau de confiance, indiquant à quel point il est sûr de sa décision. 80%, 90%..
- Heatmap : Pour les images, des techniques montrent sur quelle partie de l'image le modèle s'est basée pour prendre sa décision.
- Outils spécifiques : Des outils comme « Grad CAM » peuvent montrer sur quelle partie d'une image le modèle s'est basé pour sa décision.

Les modèles peuvent fournir des détails sur les variables les plus influentes dans une décision mais ils peuvent également indiquer si une variable a eu un impact positif ou négatif sur la décision.

	Certains modèles peuvent même fournir des détails sur la manière dont une valeur spécifique d'une variable influence la décision. Complexité de l'explicabilité : - Si un modèle utilise de nombreuses variables pour prendre une décision, il peut être difficile pour un humain de comprendre la combinaison de toutes ces variables. - Dans de tels cas, il peut être préférable de se concentrer sur un nombre réduit de variables pour faciliter la compréhension.

2.2.5.8.5 Principe 5 : Diversité, non-discrimination et équité

Le principe « Diversité, non-discrimination et équité » est l'un des piliers fondamentaux de l'AI Act, reflétant l'importance croissante accordée à la création et à l'utilisation éthique de l'intelligence artificielle.

Ce principe majeur se divise en trois sous-catégories dont voici une définition approfondie :

- **Diversité :**

La diversité est la reconnaissance et la valorisation des différences individuelles. Dans le contexte de l'IA, cela signifie que les systèmes doivent être conçus en tenant compte d'une variété de perspectives, d'expériences et de besoins.

Cela implique de prendre en compte des données provenant de différents groupes ethniques, genres, âges, orientations sexuelles, capacités physiques et mentales, et autres caractéristiques qui définissent l'identité humaine.

La diversité garantit que l'IA est inclusive et qu'elle peut servir une large gamme d'utilisateurs de manière équitable.

- **Non-discrimination :**

La non-discrimination est le principe selon lequel tous les individus doivent être traités de manière égale, sans préjudice ou favoritisme en raison de leurs caractéristiques personnelles ou de leur appartenance à un groupe particulier.

Dans le domaine de l'IA, cela signifie que les systèmes ne doivent pas produire de résultats qui avantagent ou désavantagent injustement un groupe par rapport à un autre.

Cela nécessite une attention particulière lors de la collecte de données, de la formation des modèles et de l'interprétation des résultats pour s'assurer qu'aucun biais indésirable n'est introduit.

- **Équité :**

L'équité en IA reconnaît la nécessité de traitements adaptés pour des résultats équivalents, contrairement à l'égalité qui traite tous de manière identique. Les systèmes d'IA doivent être adaptés aux besoins spécifiques, comme pour les apprenants ayant des déficiences.

L'IA doit éviter les préjugés, mais la perception d'injustice dépend du contexte. Par exemple, en assurance vie, considérer le sexe est pertinent car les femmes vivent plus longtemps, mais en crédit bancaire, cela pourrait induire une discrimination. Il faut donc évaluer les « biais » selon le cas d'usage.

La diversité au sein des équipes d'IA et la transparence des données renforcent également l'équité. Les entreprises ont un rôle majeur dans cette promotion, mais la législation, comme l'AI Act, pourrait renforcer ces obligations.

Benoit ROTTEMBOURG
Responsable Régalia (INRIA)

Quelles mesures est-il possible d'adopter pour assurer la non-discrimination?

Il y a 3 types de mesures possible :

- Le pré-process : Pour éviter les biais comme des disparités hommes-femmes par exemple, il s'agit de rééquilibrer la data ou de nettoyer la data de telle sorte qu'on évite le biais, c'est du process essentiellement sur la data, on va gonfler la data pour la rééquilibrer.

On peut retirer les variables comme le genre ou l'âge. Parfois, cela dégrade l'algorithme en retirant des variables suspectes et parfois, il faut en retirer tellement qu'il n'y a plus d'algorithme.

- Le in-process : selon les algorithmes, l'idée est d'ajouter des métriques de pénalité quand on enfreint la contrainte de disparité.

L'idée est d'utiliser un algorithme qui sait introduire une contrainte. On appelle ça une régularisation mathématiquement : c'est une sorte de fonction de coût qui va pénaliser le non-respect de la contrainte. On introduit donc toutes les métriques que l'on ne peut pas enfreindre dans cette fonction de pénalité et on l'utilise dans le tuning de l'algorithme.

- Le post-process : une métaphore possible pourrait celle où on aurait un immeuble qui a des fuites thermiques et que l'on veut refaire cet immeuble pour qu'il soit aux normes thermiques. Et bien il est possible de refaire la façade et le dessus de l'immeuble pour qu'il y ait moins de chaleur qui s'échappe, c'est-à-dire de mettre une couche pour corriger l'existant. C'est l'équivalent algorithmique de cette métaphore car les données sont déjà là et l'algorithme initial est déjà en place.

Pour résumer, une fois qu'un problème a été détecté, la question de savoir comment réparer l'algorithme se pose. Certaines entreprises peuvent choisir de retirer la variable suspecte, ce qui peut avoir un impact significatif sur la performance de l'outil. Une autre approche pourrait être d'appliquer un "skin" de réparation à l'algorithme, en regardant comment il se comporte avec des données similaires et en ajustant sa réponse en conséquence.

Retour d'expérience :

En 2016, **ProPublica** a révélé que **l'algorithme COMPAS** (cas déjà évoqué en partie « Les dérapages connus de l'IA »), utilisé pour évaluer le risque de récidive des détenus, était biaisé contre les Afro-Américains, les jugeant deux fois plus à risque que les Blancs. **Northpointe**, la société derrière COMPAS, a défendu son algorithme, affirmant qu'il était équitable pour toutes les races et avait une précision de 69%.

Cette situation montre la difficulté d'assurer l'absence de biais dans les modèles d'IA et l'importance de ne pas reproduire les discriminations existantes. Si la précision était historiquement le critère principal, l'équité est désormais une préoccupation essentielle à mesurer et à respecter.

Le graphique suivant montre le lien entre les scores COMPAS **en abscisse et les réelles récidives en ordonnée.**

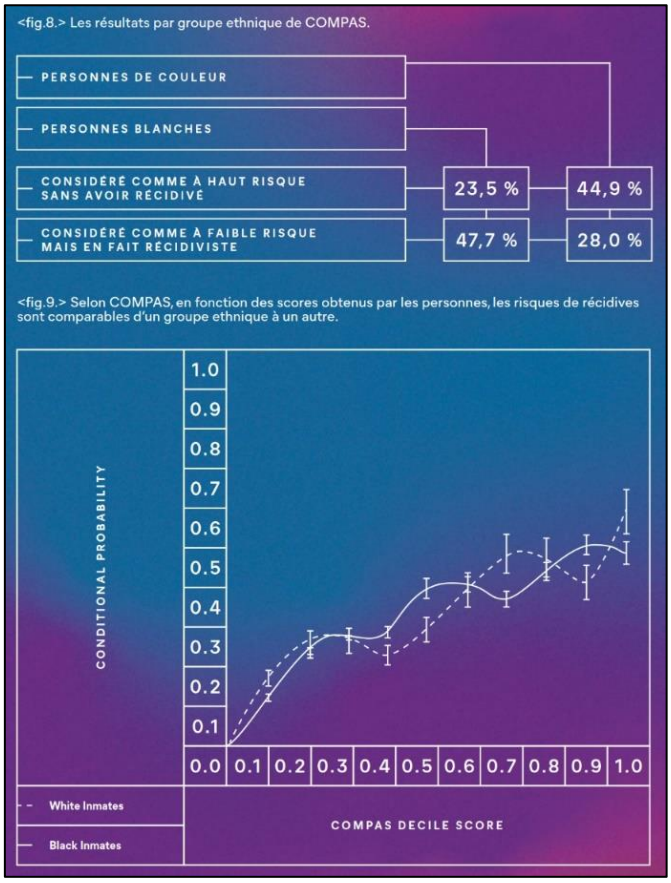


Figure 119 : Les résultats par groupe éthnique de COMPAS - Source : impact-ai.fr

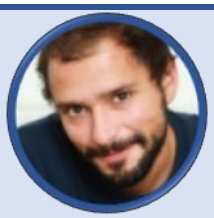
Nous voyons que, sur une échelle de 0 à 1 (1 étant équivalant à 100%), les détenus noirs et blancs dont la note COMPAS est de 0.6 présentent une probabilité de récidive entre 0.4 et 0.50 (donc entre 40 et 50%), alors que si elle baisse à 0.3, le risque s'élève à 0.3 (soit 30%), quel que soit le groupe.

Cette controverse met en lumière la complexité de définir clairement des modèles d'intelligence artificielle qui ne contiennent pas de stéréotypes et, par conséquent, non discriminatoires.

Orange s'est également engagé en faveur de **l'équité** en signant une charte sur l'IA inclusive avec **Arborus**, une organisation qui œuvre pour l'égalité professionnelle entre les femmes et les hommes. Cette charte est un guide pour les entreprises technologiques et utilisatrices d'IA, promouvant la diversité et la maîtrise des biais discriminatoires. C'est une étape vers l'obtention du label **GEEIS-IA**, **premier label international qui associe les enjeux de l'IA à la promotion de la diversité et de l'utilisation responsable, sans biais et stéréotypes**.

Cette charte prévoit sept engagements à tenir, des recommandations dont voici la liste :

Figure 120 : Les 7 engagements d'Orange issue de la charte signée avec Arborus - Source : impact-ai.fr



Grégoire COLOMBET
PDG de AI decision

Des principes comme la **transparence**, **l'explicabilité**, la robustesse et la performance ont un poids différent selon le cas d'usage mais il n'y a quasiment aucune chance qu'on trouve des cas d'usage où toutes les dimensions citées ont le même poids.

2.2.5.8.6 Principe 6 : Bien-être sociétal et environnemental

Dans le cadre de l'AI Act, le « Bien-être sociétal et environnemental » met l'accent sur l'utilisation de l'IA pour soutenir des progrès sociaux et renforcer la durabilité écologique.

Les systèmes d'IA devraient favoriser des objectifs tels que l'égalité et l'inclusion, tout en minimisant leur impact environnemental, comme la consommation d'énergie. Il est essentiel que ces technologies soient en accord avec les valeurs sociétales et servent le bien commun tout en protégeant l'environnement.

On pourrait établir deux catégories où les systèmes d'IA doivent avoir un impact ou un apport positif :

- **Le bien-être environnemental :**

L'impact écologique d'un système d'IA englobe tous les effets environnementaux (qu'ils soient directs, indirects ou consécutifs) liés à sa conception, son développement et son utilisation.

Il est également crucial de limiter l'énergie consommée lors du développement et de l'exploitation des système d'IA, en tendant vers une IA économe, en privilégiant par exemple des circuits de distribution courts. Pour contrebalancer cet impact, il serait judicieux de promouvoir des applications bénéfiques pour l'environnement, en accord avec les **Objectifs de Développement Durable**, des Nations Unies pour 2030.

Un système d'IA orienté vers un impact durable vise un projet respectueux de l'environnement et contribue à la transition écologique, aligné avec les objectifs de développement durable.

- **Le bien-être sociétal et économique :**

L'impact socio-économique d'un système d'IA se mesure par l'ensemble de ses effets (directs, indirects ou résultants) sur **l'économie, l'emploi et les relations sociales**. Plusieurs éléments permettent d'évaluer cette contribution.

Les applications d'un système d'IA devraient viser **le bien-être collectif, en favorisant la prospérité individuelle et une société équitable**. Les besoins auxquels il répond et ses fonctionnalités devraient être conçus pour complémenter les capacités humaines, que ce soit dans un contexte professionnel, social ou personnel. De plus, l'accès aux systèmes d'IA devrait être pensé de manière inclusive, permettant à un large public d'en bénéficier, indépendamment de leur situation ou condition.

- **L'impact à long terme :**

Selon le Rapport Bruntland, un système d'IA doit viser à satisfaire les besoins actuels sans compromettre ceux des générations futures. Ainsi, chaque système d'IA devrait être conçu en pesant ses avantages et inconvénients potentiels pour l'environnement, la société, l'économie et la santé, pour maintenant et pour demain.

2.2.5.8.7 Principe 7 : Responsabilité

- **Responsabilité de l'humain :**

Le terme « *responsabilité* » provient du latin « *respondere* », qui signifie « *répondre* ». Il évoque **l'obligation de rendre compte de ses actions, y compris leurs conséquences, et d'en assumer les retombées, qu'il s'agisse de réparation ou de sanction en cas de non-conformité aux attentes**.

Il existe plusieurs formes de responsabilité : légale, assurantielle, contractuelle et éthique. Alors que les responsabilités légales, assurantielles et contractuelles découlent de lois ou de contrats, la responsabilité éthique, quant à elle, reflète notre capacité à agir de manière autonome, à prendre des décisions basées sur des principes moraux.

Dans le contexte de l'IA, ce principe majeur souligne le rôle de l'humain et surtout les conséquences en termes de responsabilité. L'humain est au centre de ce principe.

- **L'humain doit prendre ses responsabilités :**

Chaque intervenant dans le développement d'un système d'IA doit assumer la responsabilité des conséquences de ses actions. Il est vital de définir les rôles et responsabilités de chacun.

L'entité lançant un système d'IA est juridiquement responsable de sa conformité éthique. Dans certains cas, comme les robots ou les systèmes d'IA collaboratifs, l'utilisateur peut aussi être responsable.

L'humain est toujours responsable de l'utilisation des résultats de l'IA, et différents acteurs peuvent être impliqués selon le type d'IA utilisé. Il est également essentiel de distinguer entre les conséquences prévues et imprévues, en mettant en place des mesures pour les gérer.

- **L'humain doit auditer et contrôler :**

L'auditabilité d'un système d'IA permet à des **parties extérieures d'examiner en détail les étapes de sa création, le trajet des données employées et les réglages des algorithmes utilisés**. Pour que les auditeurs puissent **évaluer le mécanisme, les résultats et même les effets imprévus**, ces éléments doivent être entièrement **transparents et ouverts**.

L'audit peut être orienté techniquement ou fonctionnellement :

- **Sur le plan technique**, il s'agit d'évaluer l'efficacité du système selon divers standards (consistance, précision des conclusions, etc.).
- **Sur le plan fonctionnel**, l'objectif est d'analyser ses répercussions sur les utilisateurs et de confirmer l'adhésion à divers principes éthiques (justice, autonomie, etc.).

Ces audits peuvent être effectués avant le début du développement et tout au long du processus de création, permettant ainsi une intervention préventive pour prévoir et gérer les défaillances et dangers éventuels. Ils peuvent également être réalisés après la finalisation, nécessitant une réponse rapide face à des défaillances confirmées. Dans certains cas, ils pourront même être fait avant, pendant et après.

Patrick BEZOMBES
Président du comité de normalisation de l'IA (AFNOR)

Qui serait responsable en cas d'accident ?

Prenons un exemple d'un véhicule autonome : Le constructeur pourrait ne pas vouloir assumer la responsabilité.

Pourtant, si le système est "autonome", cela signifie que l'utilisateur n'a rien fait, ce qui renforce la responsabilité du constructeur et déresponsabilise le conducteur si on le fait pas participer.

Par exemple, en cas d'accident avec des avions, c'est souvent une bataille entre le pilote, la compagnie aérienne et le constructeur pour déterminer qui est responsable. Cela dilue la responsabilité.

2.2.5.8.8 Conclusion sur les principes de l'AI Act

Ces sept principes de l'AI Act, bien que distincts, sont interdépendants et convergent vers une vision commune : celle d'une IA au service de l'humanité, transparente, équitable et bénéfique pour tous.

En intégrant ces principes dans toutes les étapes de développement et d'application de l'IA, la Commission Européenne nous permet de nous assurer d'un avenir où la technologie renforce nos valeurs, enrichit notre société et **surtout minimise les risques et les craintes quant à son utilisation**.

2.2.5.9 Focus sur l'IA générative

La Commission Européenne, dans la dernière version de l'AI Act parue le **14 Juin 2023**, inclut des obligations spécifiques pour **systèmes d'IA générative** et plus particulièrement pour les fournisseurs de « **modèles de fondation** » comme **OpenAI** et **Google**.

Un « modèle de fondation » (ou « foundation model » en anglais) fait référence à un type de modèle d'intelligence artificielle **pré-entraîné à grande échelle sur d'énormes quantités de données. Ces modèles servent de base ou de « fondation » sur laquelle des ajustements ou des spécialisations peuvent être effectués pour des tâches spécifiques**.

L'idée est que, plutôt que de former un modèle à partir de zéro pour chaque nouvelle tâche, on peut prendre un modèle de fondation **pré-entraîné et le peaufiner ou l'adapter pour des applications spécifiques**. Cela permet de gagner du temps et des ressources, car le modèle a déjà acquis une connaissance générale à partir des données sur lesquelles il a été préalablement formé.

ChatGPT, par exemple, est basé sur un tel modèle de fondation. Il est d'abord formé sur une vaste quantité de textes pour comprendre la langue, puis il peut être ajusté pour des tâches ou des applications spécifiques.

Si les modèles de fondation tel que **ChatGPT** sont abordés dans l'AI Act, c'est parce qu'ils transforment actuellement l'industrie avec leurs avancées fulgurantes telles une véritable révolution, tout en suscitant d'importants défis et débats.

Le centre de recherche de **l'Université de Stanford** a lancé une étude intitulée « *Do Foundation Model Providers Comply with the Draft EU AI Act ?* » suite à la publication du nouvel AI Act le 14 juin 2023.

Cette étude vise à évaluer la conformité d'acteurs majeurs comme **OpenAI** et **Google** à la proposition législative européenne sur l'IA. Stanford a établi une liste de 10 principaux fournisseurs de modèles de fondation pour examiner s'ils respectent un ensemble d'exigences de l'AI Act, qui est en réalité un sous-ensemble des 22 exigences définies dans la dernière version de l'AI Act.

Voici en résumé **l'approche méthodologique** utilisée par les auteurs de l'étude :

1. **Extraction des Exigences** : Sur les 22 exigences, ils ont sélectionné 12 exigences pouvant être évaluées de manière significative à partir d'informations publiques. Les autres ont donc été écartées en raison du manque d'informations publiques.
2. **Catégorisation des Exigences** : Ces 12 exigences ont été catégorisées sur 4 axes :

- « Data » ⇒ ressources de données (3 exigences)
- « Compute » ⇒ Ressources de calcul (2 exigences)
- « Model » ⇒ Le modèle lui-même (4 exigences)
- « Deployment » ⇒ Pratiques de déploiement (3 exigences)

Ces exigences et leur catégorisation sont dans le tableau ci-dessous :

Category	Keyword	Requirement (summarized)	Section
Data	Data sources	Describe data sources used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Data governance	Use data that is subject to data governance measures (suitability, bias, and appropriate mitigation) to train the foundation model.	Amendment 399, Article 28b, page 200
	Copyrighted data	Summarize copyrighted data used to train the foundation model.	Amendment 399, Article 28b, page 200
Compute	Compute	Disclose compute (model size, computer power, training time) used to train the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Energy	Measure energy consumption and take steps to reduce energy use in training the foundation model.	Amendment 399, Article 28b, page 200
Model	Capabilities/limitations	Describe capabilities and limitations of the foundation model.	Amendment 771, Annex VIII, Section C, page 348
	Risks/mitigations	Describe foreseeable risks, associated mitigations, and justify any non-mitigated risks of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Evaluations	Benchmark the foundation model on public/industry standard benchmarks.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
	Testing	Report the results of internal and external testing of the foundation model.	Amendment 771, Annex VIII, Section C, page 348 and Amendment 399, Article 28b, page 200
Deployment	Machine-generated content	Disclose content from a generative foundation model is machine-generated and not human-generated.	Amendment 101, Recital 60g, page 76
	Member states	Disclose EU member states where the foundation model is on the market.	Amendment 771, Annex VIII, Section C, page 348
	Downstream documentation	Provide sufficient technical compliance for downstream compliance with the EU AI Act.	Amendment 101, Recital 60g, page 76 and Amendment 399, Article 28b, page 200

Figure 121 : Les 12 exigences retenues par Stanford pour comparer les modèles de fondations connus - Source : crfm.stanford.edu

3. **Création de Rubriques** : Pour chaque exigence, une rubrique à 5 points a été conçue. Bien que l'AI Act énonce des obligations de haut niveau pour les fournisseurs de modèles de fondation**, il ne précise pas comment ces obligations doivent être interprétées ou mises en œuvre**. Les rubriques ont été élaborées à partir de l'expertise des chercheurs sur l'impact sociétal des modèles de fondation.
4. **Évaluation de la Conformité** : Les chercheurs ont évalué la conformité de 10 fournisseurs de modèles de fondation (et leurs modèles phares) par rapport aux 12 exigences de l'AI Act sélectionnées en fonction de leurs rubriques. Les chercheurs ont **indépendamment** noté tous les fournisseurs pour toutes les exigences. Les scores ont été fusionnés après une discussion entre tous les auteurs impliqués dans cette recherche.

En termes de notation, chaque exigence reçoit une note qui va de 0 (le pire) à 4 (le meilleur), le meilleur score possible étant de 48 sur toutes les exigences.

Voici donc le résultat obtenu par chaque fournisseur dans l'étude :

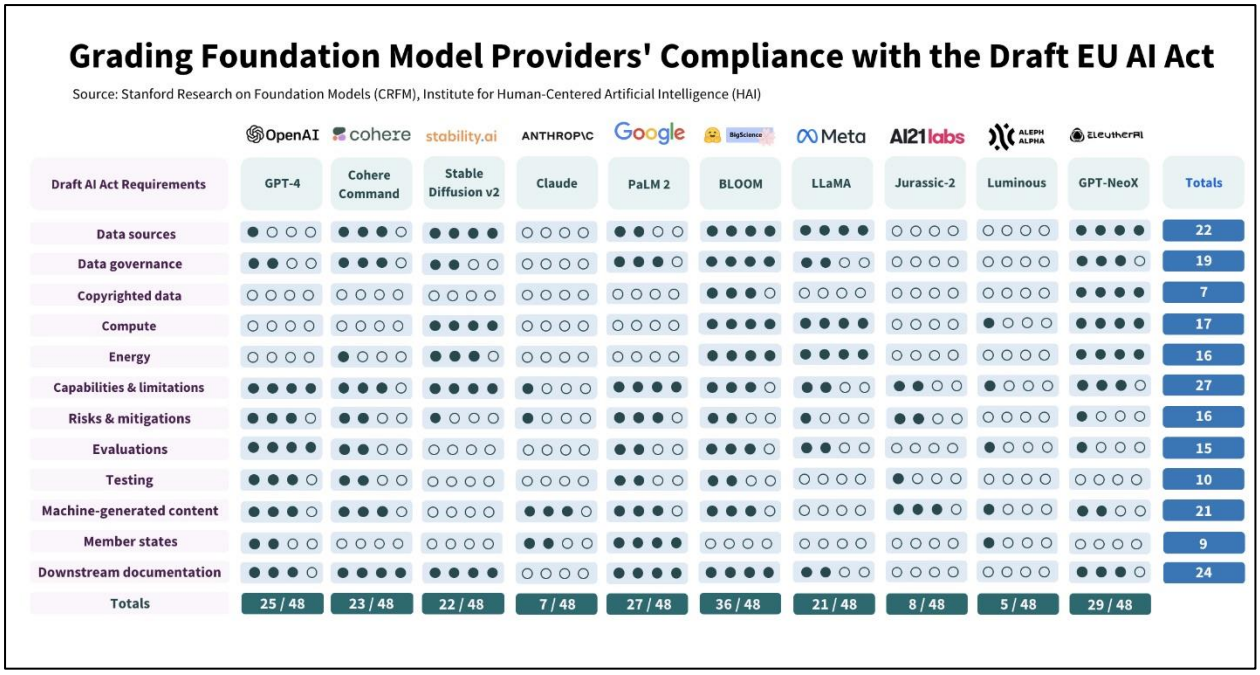


Figure 122 : Comparaison des différents acteurs de modèles de fondation - Source : crfm.stanford.edu

Dans les résultats, il est à noter que les **modèles GPT-4** (**OpenAI**) et **PaLM 2** (**Google**) **n'ont pas obtenu les meilleurs scores** (respectivement 25 et 27). **Le meilleur score revient à BLOOM (BigScience) qui obtient 36 sur 48.**

Cette notation **ne remet pas en cause les performances de chacun** mais bien leur conformité par rapport aux exigences de l'AI Act et donc indirectement est une photo de leur caractère « **IA responsable** ».

L'étude conclut d'ailleurs que la conformité aux exigences de l'AI Act de ces fournisseurs est inégale et qu'il existe une grande disparité entre les fournisseurs.

En effet, les résultats montrent des différences notables parmi les fournisseurs : certains n'atteignent même pas 25 % en termes de conformité, tandis qu'un unique fournisseur dépasse les 75 % actuellement.

Certains domaines, tels que la divulgation des données protégées par le droit d'auteur, l'utilisation de l'énergie, la mitigation des risques et l'évaluation, présentent des scores particulièrement faibles.

Les chercheurs recommandent une plus grande **transparence** et suggèrent **que les fournisseurs de modèles de fondation prennent des mesures collectives pour établir des normes industrielles qui améliorent la transparence**.

2.2.5.10 Les défis et limites de l'AI Act

L'AI Act proposé par la Commission Européenne, malgré ses intentions louables, présente selon nous plusieurs limites et défis qui sont importants d'être soulignés, notamment en vue de l'élaboration du texte final d'ici la fin de l'année 2023.

En voici quelques-unes :

- **La portée limitée du règlement** : L'AI Act se concentre sur la mise sur le marché des systèmes d'IA, mais néglige des aspects comme les droits de propriété intellectuelle ou bien même la responsabilité en cas de défaillance. Une clarification sur la responsabilité juridique nous semble nécessaire,
- **Le risque de ralentissement de l'innovation** : Bien que l'AI Act vise à instaurer la confiance dans les systèmes d'IA, il ne doit pas être un frein à l'innovation s'il est perçu comme contraignant,
- **L'approche réglementaire restrictive** : L'approche de l'UE pourrait être perçue comme une tentative de limiter les géants technologiques non européens plutôt que de promouvoir une IA éthique et responsable.
- **Le risque de distorsion de concurrence** : L'UE risque de désavantager ses entreprises si d'autres acteurs mondiaux ne suivent pas des règles similaires.
- **L'importance de la transparence** : Les entreprises doivent divulguer comment elles développent et utilisent leurs systèmes d'IA ce qui pourrait les amener à être réfractaires pour des enjeux de compétitivité.
- **La mise en œuvre cohérente nécessaire** : L'AI Act doit être appliqué uniformément dans tous les États membres de l'UE or il existe déjà des disparités visibles sur d'autres réglementations européennes aujourd'hui.
- **La nécessité d'une mise à jour régulière** : L'AI Act doit évoluer rapidement, de façon synchronisée avec les avancées technologiques.
- **Une collaboration internationale** : L'UE doit travailler avec d'autres acteurs mondiaux pour élaborer des normes communes.
- **L'éducation et sensibilisation** : Au-delà de la réglementation, il est crucial d'informer sur les enjeux éthiques de l'IA pour convaincre de ses bénéfices.

L'AI Act est certes une avancée significative pour l'IA en Europe, mais il nécessite des ajustements pour rester pertinent face à l'évolution technologique.

2.2.6 L'IA responsable : un réel besoin des entreprises

Bien que l'importance d'une IA responsable soit largement reconnue, il est important de se poser la question de son adoption actuelle par les entreprises

Sur cette question, le constat est le suivant à l'heure actuelle, selon une étude de BCG GAMMA, qui représente la branche du BCG spécialisée en IA, réalisée en 2021 auprès de plus de 1 000 grandes entreprises :

- **Seulement 48% des entreprises ayant adopté l'IA à grande échelle l'ont fait de manière responsable**.
- **55% des entreprises surestiment leur niveau de « responsabilité » en matière d'IA**.
- **Beaucoup d'entreprises pensent être en avance, mais sont en réalité en retard**.

En effet, selon cette étude, **il y a un décalage entre la perception des entreprises sur leur maturité en IA et la réalité**.

L'étude reprend également les sept principes clés de l'IA responsable et révèle que :

- **14% des entreprises sont à la traîne.**
- **La plupart sont en phase de développement (34%) ou avancée (31%).**
- **Seulement 21% sont leaders dans ce domaine**.

L'étude souligne donc qu'il y a une différence notable entre la perception des entreprises et leur réelle progression. 35% prétendent avoir pleinement adopté une IA responsable, mais 55% d'entre elles surestiment leurs avancées. En effet, l'étude insiste sur le fait que beaucoup d'entreprises surestiment leur maturité, en particulier celles qui ont adopté l'IA à grande échelle.

Steven MILLS de **BCG GAMMA** note que de nombreuses entreprises sont trop optimistes quant à leur progression en matière d'IA responsable. Il souligne **la nécessité d'efforts accrus pour une IA véritablement éthique**.

En matière de déploiement, la gouvernance des données est la plus avancée. A l'opposé, l'équité, l'impact social et l'approche **Human+AI** sont les plus difficiles à mettre en œuvre. 70% des leaders en IA responsable ont à la fois un responsable éthique et un comité stratégique.

Sur le plan régional, **l'Europe et l'Amérique du Nord sont en tête**, suivies de l'Asie et du Moyen-Orient.

En ce qui concerne les motivations, **42% des entreprises voient les avantages commerciaux comme la principale raison d'adopter une IA responsable**. D'autres raisons incluent la satisfaction client, la gestion des risques et la conformité. Seulement 6% le font pour des raisons sociales.

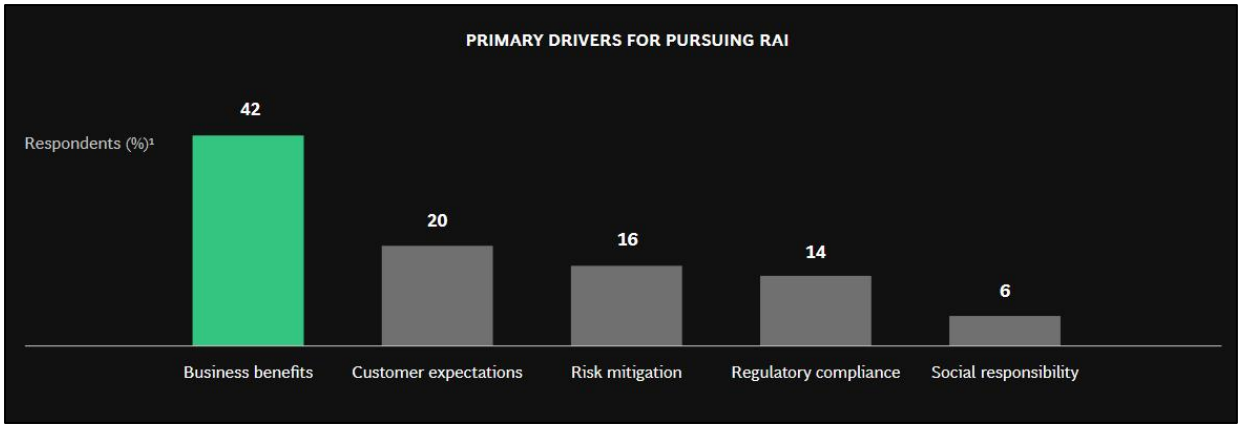


Figure 123 : Les avantages commerciaux d'une IA responsable motivent 42% des entreprises - Source : bcg.com

Sylvain DURANTON de BCG GAMMA souligne que les entreprises leaders voient l'IA responsable comme un avantage compétitif, reflétant une culture d'innovation responsable alignée sur leurs valeurs.

Pourtant, une IA responsable est une IA stable : Une intelligence artificielle développée sans éthique risque d'être instable et cela rejaillira de façon négative sur les entreprises. Les entreprises devraient donc impérativement adopter une IA responsable.

Voici ce qui nous semble être les principales raisons qui devraient inciter les entreprises à mettre en place une IA responsable :

- **La confiance des clients et des utilisateurs** :

De par sa nature même, l'IA responsable et éthique favorise la confiance des clients et des utilisateurs,

- **La conformité réglementaire et la réduction des risques** :

En visant à minimiser les risques, l'IA responsable joue un rôle essentiel en permettant aux entreprises de se conformer aux réglementations en vigueur et d'éviter les sanctions.

Au-delà de l'aspect réglementaire, la réduction des risques est aussi nécessaire pour éviter de commettre des erreurs, des discriminations voir d'autres problèmes plus critiques comme des dérapages qui pourraient nuire à la réputation des entreprises

- **Un avantage concurrentiel** :

Pour se distinguer par rapport à la concurrence, les entreprises pourrait grâce à l'IA responsable et éthique :

  - **Attirer ou retenir des talents** : Les entreprises qui défendent des valeurs « nobles » attirent davantage les travailleurs qualifiés qui accordent de l'importance à la responsabilité éthique.
  - Attirer des partenaires et des investisseurs : avec la réduction des risques, la confiance assure une stabilité qui augmente la confiance chez ces potentiels partenaires

- **Une préparation pour l'avenir via l'innovation éthique et la durabilité à long terme** :

Pour être prêt à intégrer de nouvelles innovations tout en restant éthique (via des normes éthiques) et en visant un impact durable sur la société et l'environnement. C'est donc se projeter pour réussir sur le long terme.

En résumé, au-delà de l'éthique, l'IA responsable est une stratégie essentielle pour les entreprises désireuses de rester compétitives et alignées sur les attentes sociétales.

Parmi ces entreprises, celles du milieu bancaire sont également particulièrement concernées par l'IA, ses cas d'usages. Dans ce secteur, l'adoption d'une IA responsable devient également primordiale pour garantir la confiance des clients, assurer la conformité réglementaire et naviguer avec éthique dans ce domaine sensible. **Il est donc particulièrement intéressant de faire un focus sur l'IA responsable et son application dans le milieu bancaire, objet de la prochaine partie**.

3 L'IA responsable et ses applications dans la banque

3.1 L'IA dans la banque

Dans un monde en constante évolution numérique, l'Intelligence Artificielle et les algorithmes sont omniprésents, notamment dans le secteur bancaire.

L'IA est devenue un pilier essentiel de la transformation numérique des métiers bancaires et financiers. Ces domaines se distinguent par leur adoption précoce et avancée des technologies d'apprentissage automatique. Pour les institutions bancaires, gérer efficacement leurs données et leur Intelligence Artificielle est crucial pour naviguer dans ce paysage en mutation.

L'enjeu est de ne pas se laisser distancer par les innovateurs technologiques, mais plutôt de prendre les rênes de cette technologie pour renforcer leur position concurrentielle. Cependant, les banques françaises font face à des défis, notamment les potentiels obstacles et les limites inhérentes à l'IA.

3.1.1 Les applications de l'IA dans le secteur bancaire et ses avantages

3.1.1.1 Un enjeu d'amélioration des services et de compétitivité

L'IA est devenue un élément clé pour la compétitivité. Elle accélère la prise de décision et empêche la domination technologique par quelques géants comme les GAFAM (Google, Amazon, Facebook...). Bien que la France ne soit pas en retard en matière d'IA, il est essentiel de rester à la pointe et notamment dans le secteur bancaire.

En effet, l'IA connaît une adoption notable dans le domaine bancaire et financier, un secteur particulièrement propice à son essor. Entre autres, la nécessité d'analyser et d'exploiter rapidement un volume toujours plus important de données fait de l'IA un outil incontournable dans ce secteur.

D'un point de vue global, les avantages de l'IA pour le secteur bancaire sont les suivants :

- **Marketing :**

  - L'IA permet l'analyse des données. Cela aide à comprendre les besoins des clients et à optimiser les produits financiers.
  - Elle permet également le développement des offres, produits et services proposées par les banques.
  - L'IA favorise par ailleurs une plus grande satisfaction et fidélité des clients.
  - Elle aide aussi les banques à décrypter les tendances du marché, identifier des opportunités de croissance et anticiper les risques financiers. Ces « **insights** » (*recommandations* ou *connaissances de la façon de résoudre un problème*) peuvent guider les banques dans leurs décisions d'affaires et la conception de leurs produits.

- **Commercial :**

Les outils d'IA, tels que les assistants bancaires et les robots-conseillers, simplifient la communication avec les clients et la compréhension des offres financières.

En effet, ils permettent à la banque de comprendre rapidement les demandes des clients, quel que soit le moyen de communication (écrit ou vocal) et de répondre avec plus de réactivité et de pertinence en complément de l'humain (réponses aux questions les plus simples/ fréquentes).

- **Réglementation :**

L'IA renforce la détection du blanchiment d'argent, essentielle pour la sécurité financière en identifiant les comportements anormaux ou suspects dans les transactions bancaires en temps réel. Elle peut ainsi détecter des opérations inhabituelles, des tentatives de piratage ou d'usurpation d'identité et plus encore des fraudes internes.

- **Gestion des risques :**

  - L'IA fournit des outils avancés pour une meilleur prise de décision concernant les risques.
  - Elle permet également de mieux maitriser les risques en optimisant la détection et en proposant aux clients des mesures préventives.
  - Elle améliore la gestion des anomalies : Identifier les erreurs et incidents lors des transactions bancaires et proposer des correctifs. Cela inclut la surveillance continue des systèmes et la reconnaissance des irrégularités.
  - L'IA impacte également positivement le pilotage des activités et gestion des risques : les risques liés aux crédits, risques aux marchés et aux opérations.

- **Financier :**

  - L'Intelligence Artificielle permet dans ce domaine de réaliser d'importantes économies d'échelle grâce à l'automatisation des tâches routinières et à l'optimisation des processus organisationnels.
    Étant donné que le secteur bancaire est influencé par les taux d'intérêt, il est essentiel de rechercher des améliorations de productivité. Cela implique notamment de minimiser les coûts opérationnels pour demeurer compétitif, d'où l'importance croissante de l'automatisation.

    En effet, automatiser les tâches fastidieuses et répétitives ou à faible valeur ajoutée facilite le travail des collaborateurs et leur permet de libérer du temps et des ressources pour les activités à plus forte valeur ajoutée afin d'établir et maintenir une

relation de confiance avec nos clients. Cela peut concerner des aspects tels que la gestion des comptes, la vérification des documents, la préparation des rapports, etc. Par exemple : optimiser les processus de crédit en analysant les données des clients pour évaluer leur capacité à rembourser les prêts.

- Elle permet par ailleurs d'optimiser les processus de prise de décision dans la banque.

- Elle contribue à l'amélioration de la gestion des données : Il est très important de traiter efficacement l'énorme volume de données produites par les activités bancaires.
  Cela englobe des éléments tels que la collecte, la conservation, l'analyse en temps réel, la synthèse et la représentation visuelle des données, afin d'en extraire des informations pertinentes pour des décisions éclairées.
  Face à l'immensité des données quotidiennes (transactions, paiements, prélèvements, etc.), l'automatisation et la robotisation deviennent essentielles.

- L'IA permet la prévision des tendances : L'analyse des données passées et présentes permet de prédire les orientations futures des marchés financiers. Cette démarche est essentielle pour guider les choix stratégiques des établissements financiers et les protéger contre d'éventuels risques.

- Elle facilite l'analyse des investissements des clients grâce au Deep Learning.

- Elle rend possible le trading assisté par IA : Dans le domaine financier, l'Intelligence Artificielle offre des innovations majeures, notamment l'analyse instantanée qui renforce le trading assisté par des algorithmes.

  D'après une recherche menée par **JPMorgan** en 2020, plus de 60% des opérations financières dépassant les 10 millions de dollars ont été réalisées via des algorithmes. On estime que le secteur du trading assisté par des algorithmes augmentera de 4 milliards d'USD d'ici 2024, atteignant ainsi un total de 19 milliards d'USD.

- **Infrastructure :**

L'IA permet de prévoir l'infrastructure informatique et les éventuels incidents associés.

3.1.1.2 Renforcer l'assistance augmentée : un enjeu d'intensification de la relation client

L'Intelligence Artificielle révolutionne le secteur bancaire, offrant des perspectives inédites, notamment dans la compréhension des clients grâce à la démarche « *Know Your Customer – KYC* ». Cette technologie exploite l'open data pour enrichir les profils clients via des API comme Data Infogreffe. L'objectif est d'optimiser le service client, en anticipant leurs besoins et en proposant des solutions personnalisées basées sur leurs données.

L'IA modifie la relation client. Les consommateurs, de plus en plus autonomes, bénéficient d'outils comme les chatbots, voicebots et analyseurs de mails pour répondre à leurs questions, allégeant ainsi la charge des conseillers. Ces outils, en constante évolution, visent à mieux cerner le client tout au long de sa relation avec la banque.

L'une des premières applications concrètes de l'IA dans le domaine bancaire est **la personnalisation de l'expérience utilisateur**. Les chatbots et assistants virtuels, grâce à l'analyse des données, offrent un accès simplifié aux informations financières des clients. L'IA améliore l'expérience client en fournissant des produits et conseils adaptés, basés sur leur situation financière, leur historique et leurs projets. Elle anticipe leurs besoins, notamment en cas de risque de départ de la banque. Par exemple, elle peut prévenir un client d'un risque de découvert ou lui proposer une offre d'épargne adaptée.

L'analyse des données par l'IA permet une meilleure compréhension des besoins des clients et de l'ajustement des produits financiers. En scrutant les dépenses des clients, les banques peuvent identifier les produits financiers les plus populaires et proposer des offres correspondantes. Les outils interactifs de l'IA peuvent également répondre aux questions des clients, traiter leurs préoccupations et fournir des recommandations sans nécessiter un rendez-vous physique avec un conseiller.

Un exemple concret est le **Crédit Agricole** avec son offre Eko. Cette offre inclut une application mobile dotée d'un assistant virtuel, Max, qui assiste les clients dans la gestion de leur budget, la comparaison d'offres d'assurance ou de crédit, et la sollicitation d'un conseiller si besoin.

L'IA offre donc aux banques une vision approfondie de leurs clients, leur permettant de proposer des offres plus pertinentes basées sur leurs finances, leurs habitudes et leurs projets futurs. Elle anticipe les besoins des clients, les conseille et peut même leur offrir des avantages comme des taux d'intérêt préférentiels.

Par exemple, **Crédit Agricole** emploie cette technologie pour identifier les clients susceptibles de rencontrer des difficultés financières. Les conseillers peuvent ainsi intervenir proactivement pour proposer des solutions adaptées. Cela épargne au banquier l'effort d'analyser manuellement la situation financière de chacun de ses clients, lui permettant de se focaliser sur des missions plus importantes.

Mais l'Intelligence Artificielle ne remplacera pas la chaleur humaine :

Avec l'évolution rapide de l'IA, une interrogation émerge concernant le rôle traditionnel du conseiller face à la puissance de l'Intelligence Artificielle.

Dans le secteur bancaire, l'IA n'a pas vocation à remplacer l'humain notamment en raison du peu de tâches automatisables à 100 % mais vient en complément de l'être humain. On parle alors souvent d'**intelligence augmentée**, celle qui permet l'amplification des capacités intellectuelles humaines grâce à l'IA.

Deux observations s'imposent :

- D'une part, **l'IA nécessite du temps et une multitude d'exemples pour atteindre son plein potentiel**, ce qui facilite l'intégration progressive des robots et de l'intelligence augmentée dans les habitudes des clients et des conseillers.
- D'autre part, la dimension humaine demeure irremplaçable, même si la technologie occupe une place centrale dans le modèle bancaire. **La confirmation des décisions et l'explication des refus par un expert sont toujours appréciables par les clients**.

Par conséquent, il est important de **prévoir une transition en douceur vers l'utilisation de l'IA**, en s'assurant que les employés disposent des compétences et des ressources nécessaires pour travailler de manière efficace avec ces technologies.

Avec une approche mêlant l'humain à l'IA, cela signifie profiter des dernières innovations technologiques tout en s'appuyant sur l'expertise et l'empathie d'un conseiller, épaulé par des algorithmes.

3.1.1.3 Un enjeu de gestion des risques de souscription

- **L'attribution de crédits :**

L'IA joue un rôle croissant dans le domaine du crédit, en particulier pour améliorer les systèmes de scoring. C'est particulièrement le cas pour les crédits à la consommation où les clients valorisent la rapidité et l'efficacité du processus.

Au lieu de se baser uniquement sur des données financières traditionnelles, le scoring s'enrichit grâce à l'exploitation du Big Data, incluant des informations non financières sur le client. Par exemple, le score de crédit fixe le montant et les conditions pour un emprunteur grâce aà l'IA en se basant **sur des informations professionnelles et personnelles** (risque de pays, d'âge, etc.). L'avantage est **l'intégration de données externes**, comme celles de grands fournisseurs externes ou liées aux comportements des clients, aux données bancaires habituelles pour le calcul du score.

Cette méthode non seulement affine le score, mais permet aussi de le déterminer même si l'historique bancaire du client est limité ou absent, en se basant sur des données non bancaires.

Plusieurs institutions déclarent déjà utiliser l'IA pour le **scoring**, tandis que d'autres, ayant finalisé la phase de développement, cherchent à rendre leur méthode plus transparente et conforme aux réglementations.

- **La prévention de la fraude et Lutte Contre le Blanchiment d'argent et le Financement du Terrorisme (LCB-FT) :**

Dans les domaines bancaire, l'IA est fréquemment utilisée pour détecter les fraudes documentaires et **L**utter **C**ontre le **B**lanchiment d'argent et le **F**inancement du **T**errorisme (appelé « ***LCB-FT*** »).

Les outils d'IA servent spécifiquement à reconnaître, analyser et valider les documents soumis. Les algorithmes conçus pour ces tâches sont souvent aboutis et intégrés dans de nombreux mécanismes de vérification.

Concernant les paiements, l'IA est également précieuse pour repérer les transactions suspectes grâce à l'analyse de données en temps réel. Cependant, bien que l'exploitation des données de paiement soit à l'étude par de nombreux acteurs, elle en est encore à ses prémices.

Les algorithmes, notamment les **LLM** (« *Large Margin Learning* ») souvent utilisés pour résoudre des problèmes de classement de documents, offrent aux banques **la capacité d'intervenir rapidement face à une fraude confirmée, en stoppant les opérations suspectes, en informant les clients touchés et en simplifiant le processus de remboursement**.

L'Intelligence Artificielle joue également un rôle préventif, en éduquant les clients sur les méthodes de sécurité et en suggérant des actions préventives.

Par exemple, en **Belgique**, la **KBC** (groupe financier spécialisé dans la banque et les assurances) emploie l'IA pour surveiller l'activité des utilisateurs sur son application mobile afin de repérer les comportements inhabituels. Face à une activité douteuse, **l'IA sollicite une vérification d'identité du client, soit par un code PIN, soit par une empreinte digitale**. Grâce à cette stratégie, la banque a pu **diminuer les tentatives de phishing de 70%**.

Selon le rapport « Guide du numérique : analyse sectorielle de l'Intelligence Artificielle » de **DLAPiper** paru en 2022, les banques, institutions réglementées en matière de conformité aux exigences réglementaires et prudentielles, ainsi que les autorités réglementaires travaillent main dans la main pour mettre en place des innovations technologiques

- Appelées "**regtechs**" côté institutions financières
- Appelées "**suptechs**" coté autorités réglementaires

Les institutions financières utilisent principalement les **regtechs** pour la lutte contre la criminalité financière, la sécurité informatique et les reportings. Elles y voient des avantages tels qu'une **meilleure gestion des risques, des capacités de surveillance et d'analyse améliorées, la réduction des erreurs humaines et l'amélioration de l'efficacité des processus internes**.

Quant aux autorités réglementaires, elles ont principalement recours aux **suptechs** pour les **reportings et la gestion des données pour améliorer leurs capacités de contrôle, de surveillance, d'analyse et pour générer des indicateurs de risque en temps réel**.

D'après cette étude, beaucoup d'autorités ont déjà développé des outils basés sur l'Intelligence Artificielle, le Machine Learning et le traitement automatique du langage naturel (NLP) pour automatiser le traitement de grands volumes de données complexes. Ces outils améliorent la vitesse et l'efficacité de leurs analyses tout en réduisant le taux d'erreur. Cela est particulièrement utile pour améliorer la surveillance d'activités comme le trading à haute fréquence ou pour détecter des manipulations de marché.

Elle arrive en tête des technologies utilisées pour répondre à leurs besoins :

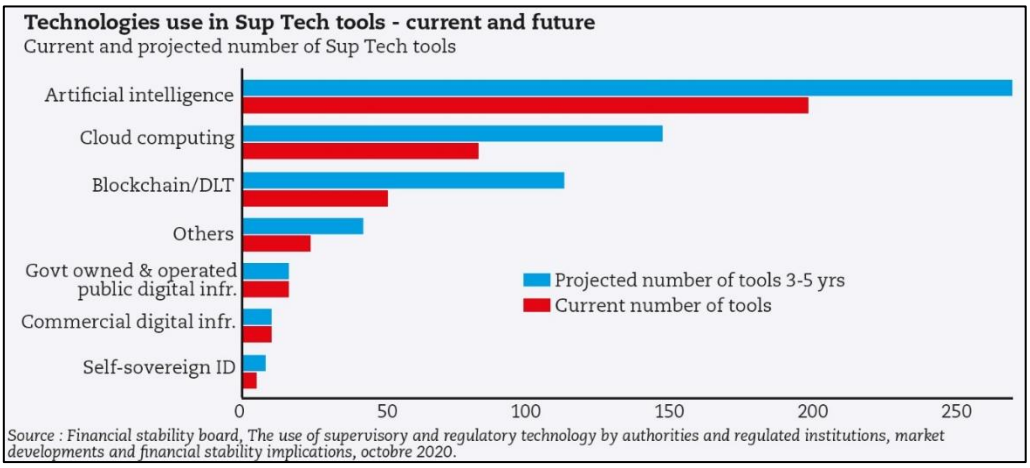


Figure 124 : Outils utilisés par les autorités financières - Source :

3.1.1.4 Un enjeu de gestion du risque cybernétique

Les institutions financières, en particulier les banques, adoptent de plus en plus l'IA pour se défendre contre les cyberattaques.

Les données de sécurité, principalement composées de journaux de connexions et d'activités, sont complexes à interpréter par l'être humain. Bien que ces données soient semi-structurées, des algorithmes avancés peuvent les traiter efficacement.

Ces outils ont la capacité de **s'ajuster en temps réel sans nécessiter de mises à jour logicielles**. L'utilisation de l'IA face aux menaces cybernétiques se divise en trois axes :

- **Prévention avant l'attaque** :
  - Détecter les **failles** grâce à des outils de scan autonomes.
  - Corriger ces **vulnérabilités**, que ce soit par une solution temporaire ou permanente.
  - Assister les développeurs à rédiger un **code sécurisé** en analysant les tendances des attaques précédentes.

- **Détection des cyberattaques** :
  - Les algorithmes de détection d'anomalies servent à **identifier les attaques en ligne, les vulnérabilités ou les virus** avant qu'ils ne se diffusent dans les systèmes internes.
  - L'IA est particulièrement utile pour repérer les **attaques ciblant les transactions**, complétant ainsi les méthodes traditionnelles par une analyse comportementale plus poussée.

- **Gestion et analyse post-incident :**
  - Certains outils d'IA sont testés pour mieux cerner **l'origine des cyberattaques**, aidant ainsi à « **profiler** » **les cybercriminels**.
  - L'IA permet également d'affiner les bases de données recensant les incidents, contribuant ainsi à **évaluer la sécurité des entreprises**.

3.1.1.5 Du côté des services d'investissement et des activités sur les marchés financier

Un rapport du Conseil de Stabilité Financière (« **Financial Stability Board** » (FSB) en anglais), organisation internationale qui surveille et fait des recommandations sur le système financier mondial) a mis en lumière les opportunités offertes par l'IA dans le domaine des marchés financiers, notamment en optimisant la gestion des risques et en minimisant rapidement les écarts de prix.

Selon les travaux de cette « task force », voici les principales utilisations de l'IA identifiées :

- **Détection d'anomalies** : L'IA permet d'identifier des irrégularités dans les transactions de marché, que ce soit face à des fraudes externes, des délits d'initiés ou des erreurs de saisie.
- **Surveillance des risques de marché** : Des algorithmes de Machine Learning sont en cours d'essai pour prévoir les risques de marché. Certaines de ces méthodes sont sur le point d'être déployées.
- **Recommandation de stratégies d'investissement** : Certains établissements ont déjà mis en œuvre cette application. Les algorithmes suggèrent des solutions économiques pour les parties acheteuses/vendeuses ayant le moindre impact sur le marché. Cela se traduit par des stratégies basées sur un séquençage aléatoire des ordres d'achat/vente, une version contemporaine du « *Time Weighted Average Price* ».
- **Évaluation des profils de risque** : Cela aide à mieux cerner la préférence des clients pour divers produits d'investissement et d'épargne. De plus, certains assureurs ont développé des outils

pour identifier l'intérêt potentiel de leurs clients pour certains produits d'assurance ou investissements. Plusieurs de ces outils ont récemment été lancés.

- **Gestion de portefeuille pour des tiers** : Bien que l'IA ne semble pas encore être en production selon les retours, certains établissements testent des algorithmes d'IA pour faciliter cette mission.

En conclusion, il s'agit ici d'une liste générale de possibilités d'utilisation de l'IA dans le secteur bancaire et financier. Il y a certainement encore beaucoup d'autres contextes dans lesquels elle pourrait être utilisé et avec l'avancée fulgurant de cette technologie, il est très difficile de tous les lister.

Voici quelques graphiques qui pourraient résumer les cas possibles d'utilisation de l'IA dans la banque :

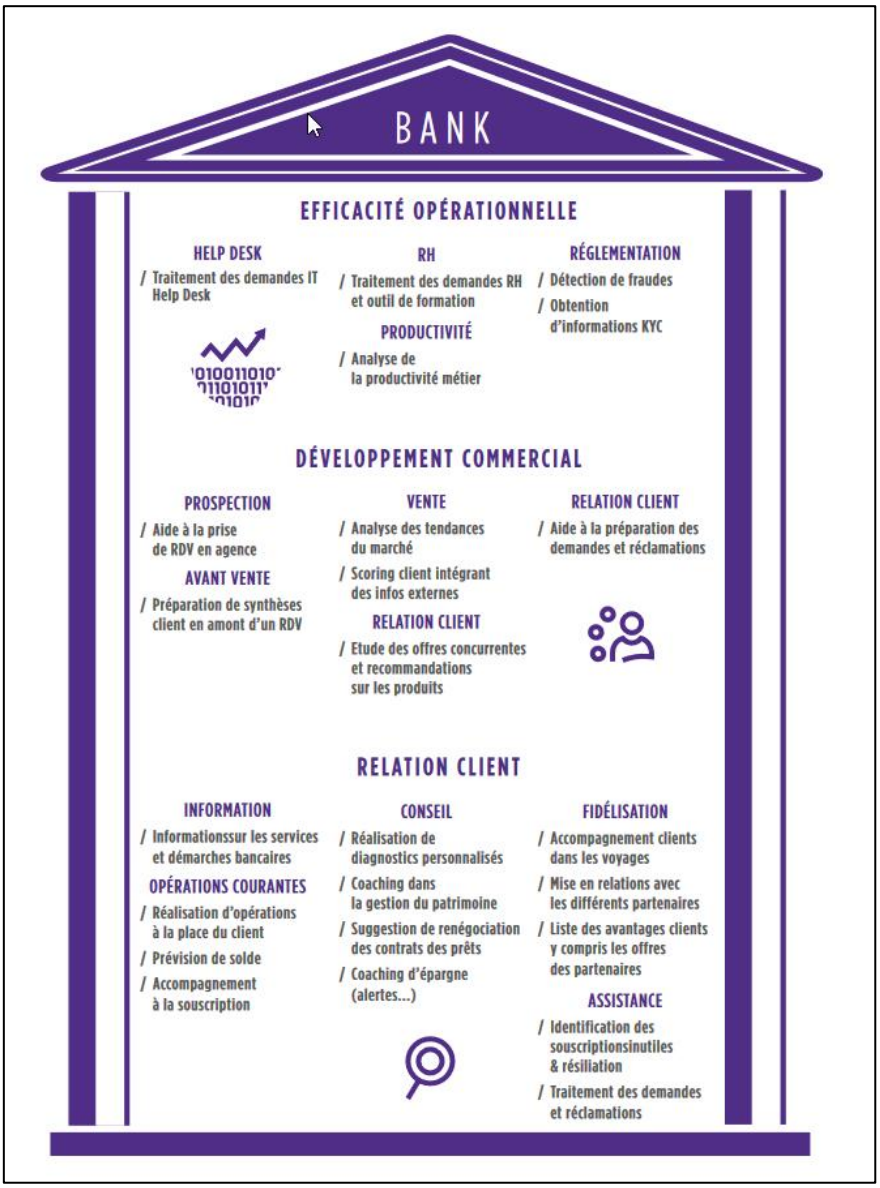


Figure 125 : Etude « L'intelligence Artificielle et la banque de demain » - Wavestone, 2018

Voici de manière plus large les usages possible de l'IA dans le domaine financier (à l'extérieur du cercle) puis les types d'IA associés :

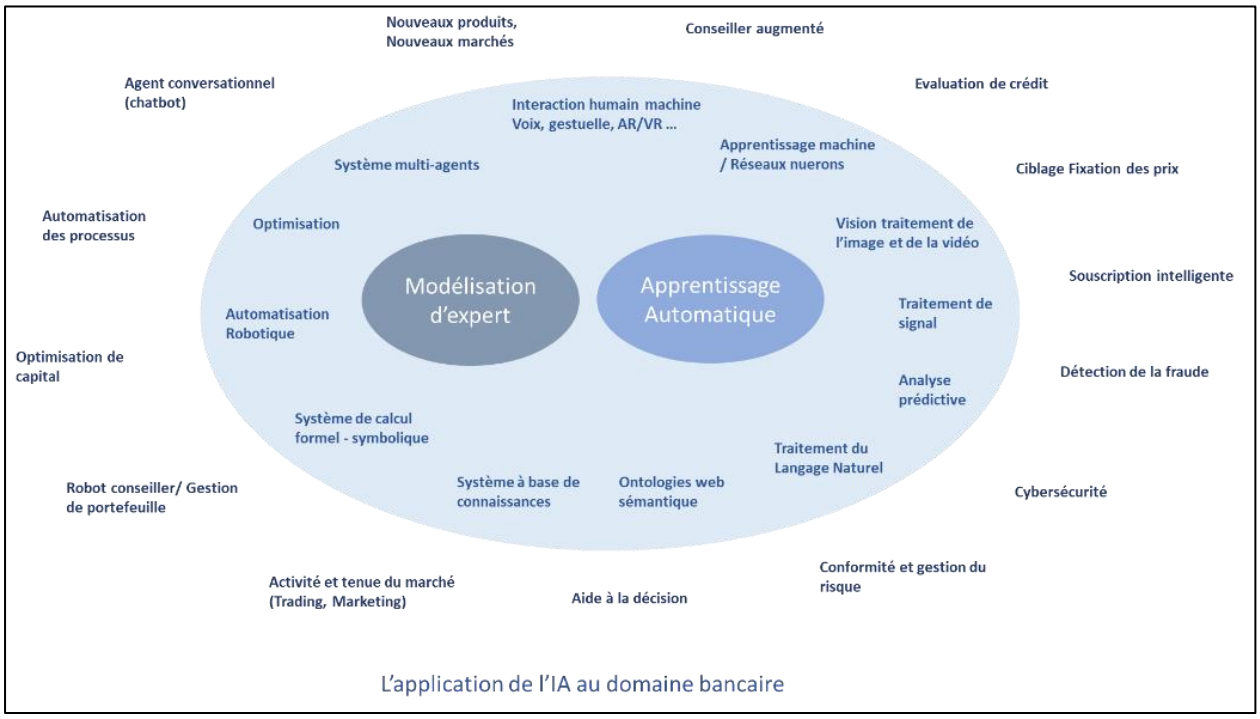


Figure 126 : Différentes technologies d'IA applicables au domaine financier

Selon le rapport de **DLAPiper** mentionné précédemment, l'autorité bancaire européenne remonte qu'en juin 2021, les instituts financiers utilisaient de manière privilégiée l'IA pour répondre aux besoins réglementaires et dans les proportions suivantes :

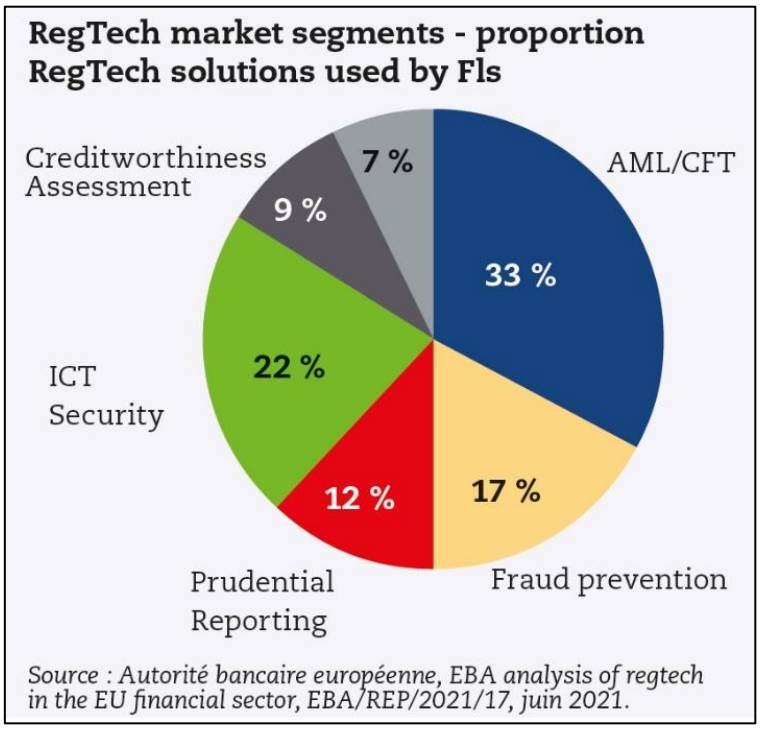


Figure 127 : Proportion des solutions d'IA utilisées par répondre aux régulateurs – Source : DLAPiper

Quelques traductions des termes utilisés dans ce schémas :

- AML (« Anti-Money Laundering ») : **lutter contre le blanchiment d'argent**,
- « Fraud prevention » : **prévention de la fraude**,
- « Prudential Reporting » : **reporting prudentiel**, rapports et des informations que les banques et les compagnies d'assurance, sont tenues de soumettre aux autorités de régulation et de surveillance pour démontrer leur conformité aux exigences réglementaires et prudentielles
- « ICT security » (**Sécurité des Technologies de l'Information et de la Communication**) : protection des données et de l'infrastructure, gestion des incidents, Conformité réglementaire, MCO…,
- « Creditworthiness Assessment » : **évaluation de la solvabilité**, processus par lequel une institution financière évalue la capacité d'un emprunteur à rembourser un prêt ou une dette,
- « CFT (Combating the Financing of Terrorism) » : **combattre le financement du terrorisme.**

3.1.2 Les risques liés à l'utilisation de l'IA dans les banques

3.1.2.1 Les risques liés au traitement des données

L'efficacité de l'Intelligence Artificielle repose grandement sur la qualité des données et sur une approche exempte de biais. Des biais dans les résultats algorithmiques peuvent nuire aux banques et à leurs clients, en raison des risques de discrimination ou de recommandations inappropriées.

Par exemple, une donnée comme la localisation géographique pourrait discriminer certaines populations lors d'une demande de prêt, accentuant les inégalités. Les modèles basés sur des historiques peuvent aussi être moins efficaces pour les jeunes clients avec peu d'historique.

La détection et l'élimination des biais dépendent principalement de la rigueur des Data Scientists. Beaucoup ne sont pas formés pour gérer ces biais, d'où l'importance pour certaines institutions financières de sensibiliser leurs experts aux risques associés.

3.1.2.2 L'IA et les défis croissants de la cybersécurité

Avec une augmentation constante des cyberattaques qui menacent la stabilité du système numérique mondial et en particulier du système bancaire, il devient crucial que les banques s'en prémunissent efficacement.

Face à 140 tentatives de phishing chaque heure, 29 cyberattaques annuelles par entreprise quel que soit son secteur et 65 vols de données chaque seconde, l'urgence des actions à entreprendre est manifeste. La cybersécurité est désormais l'une des principales préoccupations des dirigeants.

> *C'est un enjeu de premier plan alors qu'un quart des entreprises ne peuvent pas faire face - aujourd'hui - à une cyberattaque.*

***Jérôme HOMBOURGER**, Directeur général du Crédit Agricole Val de France*

Le développement de l'Intelligence Artificielle ne crée pas nécessairement de nouvelles vulnérabilités en matière de cybersécurité, mais peut amplifier celles déjà existantes.

Voici un aperçu de la situation :

- **L'IA multiplie les attaques possibles :**
  - En automatisant des tâches récurrentes et en augmentant les **interconnexions réseaux**, l'IA crée davantage de points faibles que les cybercriminels peuvent exploiter.
  - L'utilisation croissante du Cloud dans les banques et notamment pour les besoins de l'IA offre plus de points d'accès aux cybercriminels. Bien que les fournisseurs de Cloud offrent une sécurité robuste, l'adoption de solutions SaaS (Software as a Service) peut introduire de nouvelles vulnérabilités.

- **Des attaques innovantes visent spécifiquement les algorithmes d'IA :**
  - Les attaques par « ***flooding*** » sont courantes, cherchant à fausser les résultats d'un algorithme d'IA en injectant des données erronées.
    Une attaque par « *flooding* » (inondation en français) est une forme d'attaque par déni de service (« **DoS »**) où l'attaquant tente de submerger un réseau ou un service **en envoyant un volume massif de requêtes ou de données, dépassant ainsi la capacité du système cible à traiter ces demandes**. L'objectif est de rendre le service ou le réseau indisponible pour les utilisateurs légitimes.
    Exemple : Un attaquant pourrait submerger un système de détection de fraude bancaire basé sur l'IA avec une multitude de transactions légitimes pour masquer une activité frauduleuse ultérieure.

  - Des attaques plus sophistiquées, comme les attaques « ***adversarials*** » (appelé également « *par exemples contradictoires* »), **peuvent tromper** par exemple un algorithme de reconnaissance d'images **avec de légères modifications**.
    Les attaques « *par exemples adverses* » cherchent à introduire des entrées trompeuses ou altérées dans le système d'IA lors de sa mise en œuvre ou de son utilisation (si apprentissage automatique).
    Exemple : Un cybercriminel pourrait modifier légèrement les détails d'une transaction pour tromper un système de détection de fraude basé sur l'IA, faisant en sorte que la transaction malveillante soit perçue comme légitime.

- **L'IA peut renforcer la menace posée par les cybercriminels :**
  - L'IA peut rendre la cybercriminalité plus accessible et économique, **en automatisant les étapes d'une attaque**, modifiant ainsi l'équilibre entre l'ampleur et l'efficacité des attaques.
  - Le Machine Learning pourrait faciliter le **déchiffrement de mots de passe** en se basant sur des bases de données de mots de passe antérieurs.
  - Les attaques pourraient devenir plus ciblées et donc plus redoutables, comme le **phishing personnalisé ou l'utilisation de chatbots et de technologies d'imitation vocale** pour soutirer des informations sensibles.

3.1.2.3 Les risques liés à la dépendance des acteurs

L'expertise en Intelligence Artificielle, principalement détenue par de grandes entreprises informatiques non-européennes, pourrait conduire à une concentration du marché entre les mains de ces acteurs. Les banques sont particulièrement exposées notamment au vu du nombre de prestations externes auxquelles elles recourent.

Cette situation pourrait entrainer plusieurs conséquences pour elles :

- Une inflation des prix due à un manque de concurrence,
- Une restriction d'accès à certains services exploitant l'IA,
- Des relations commerciales inégales entre fournisseurs et clients,
- Des préoccupations de souveraineté, notamment en ce qui concerne la gestion des plateformes, technologies et données, comme celles des fournisseurs de services cloud ou de solutions d'IA,
- Une compréhension limitée des algorithmes par les utilisateurs, renforçant l'opacité des systèmes d'IA (effet « boîte noire »),
- Des obstacles à l'audit et à la surveillance nécessaires dans le secteur financier.

Le danger majeur réside également dans la complexité croissante des algorithmes d'IA, qui pourrait rendre leur duplication ou même leur explication difficile pour d'autres acteurs.

Un retard technologique pourrait aussi pousser les institutions financières françaises vers des solutions étrangères, renforçant ainsi la domination des firmes non-européennes dans le domaine de l'IA. Il en découle donc des questions de souveraineté.

3.1.2.4 Les risques de souveraineté et de stabilité financière

- **Concernant la stabilité financière :**

Avec l'avènement des algorithmes de « ***trading*** » haute fréquence au début du 21ème siècle, la stabilité financière est devenue une préoccupation majeure. Cette inquiétude s'est intensifiée avec l'émergence des algorithmes basés sur le « Machine Learning », dont les comportements futurs sont imprévisibles. Trois risques principaux associés à l'IA sont :

- **Trading directionnel technologique** : Les algorithmes, en utilisant des variables similaires, peuvent adopter des stratégies identiques, conduisant à des "comportements moutonniers". Cela peut amplifier la procyclicité et la volatilité du marché en provoquant des achats et ventes massifs simultanés.
- **Vulnérabilité du marché** : Les comportements uniformes rendent le marché plus susceptible aux attaques, car il est plus facile d'influencer des acteurs agissant de manière identique.
- **Formation basée sur des données historiques** : De nombreux algorithmes sont formés sur des données de périodes stables, et non de crises. Ils pourraient donc exacerber les crises financières s'ils n'ont pas été formés pour gérer ces situations.

Outre ces risques, une mauvaise application de l'IA peut entraîner d'autres dangers, comme une évaluation inadéquate des risques de crédit, mettant en péril le marché obligataire ou les institutions bancaires.

- **Concernant la souveraineté :**

Les disparités en matière de compétence technologique pourraient créer d'importants déséquilibres entre les nations. Un exemple concret est le risque de transfert de données vers des fournisseurs américains. Le **Cloud Act**, promulgué par le gouvernement américain en 2018, permet d'accéder aux données stockées sur les serveurs des fournisseurs de Cloud américains. Cette loi est en contradiction avec les principes du RGPD, en particulier l'article 48, qui stipule que toute demande d'un pays tiers pour accéder à des données personnelles doit être basée sur un accord international.

3.1.3 La maturité de l'IA dans les banques

3.1.3.1 Aujourd'hui

D'après Evident, déjà évoqué précédemment, qui évalue **la maturité en IA des 23 principales institutions financières mondiales sur la base de 143 critères, les grandes banques employaient plus de 650 spécialistes en IA en 2022 (date des données les plus récentes)**, dont 40% ont été recrutés cette même année. **Cela est le reflet de l'augmentation des besoins en IA et de son développement au sein des banques**.

Par ailleurs, entre 2017 et 2022, ces banques ont renforcé leurs investissements dans les start-ups spécialisées en IA, avec **une croissance annuelle moyenne de 15%**.

La méthodologie utilisée par **Evident** a été développée avec la contribution d'experts de premier plan en IA et en banque. Il s'agit de classer et d'évaluer chaque banque sur 143 indicateurs individuels répartis sur quatre domaines clés de compétence critique, avec des poids attribués pour chaque domaine de 40%, 30%, 15% et 15% respectivement :

- « **Talent** » : mesure le nombre et l'expérience des employés en IA et en données déclarés comme travaillant dans chaque banque, ainsi que les initiatives visibles en cours pour embaucher, retenir et développer les meilleurs talents en IA.
- « **Innovation** » : mesure les démarches entreprises par les banques pour stimuler l'innovation à travers la banque, couvrant la recherche académique et les brevets, les investissements dans la technologie et les entreprises axées d'abord sur l'IA, ainsi que l'engagement plus large dans l'écosystème open source.
- « **Leadership** » : mesure les communications publiques du leadership au niveau de l'entreprise et du groupe, y compris l'existence d'un récit public sur l'IA à travers les matériaux destinés aux investisseurs au niveau du groupe, les communiqués de presse et les médias.
- « **Transparency** » : reflète dans quelle mesure les banques communiquent publiquement sur un large éventail d'activités responsables en matière d'IA et rendent visibles leurs efforts pour créer des contrôles spécifiques à l'IA.

Selon le rapport d'Evident (lien ici), le classement des banques selon les 143 critères retenus est le suivant :

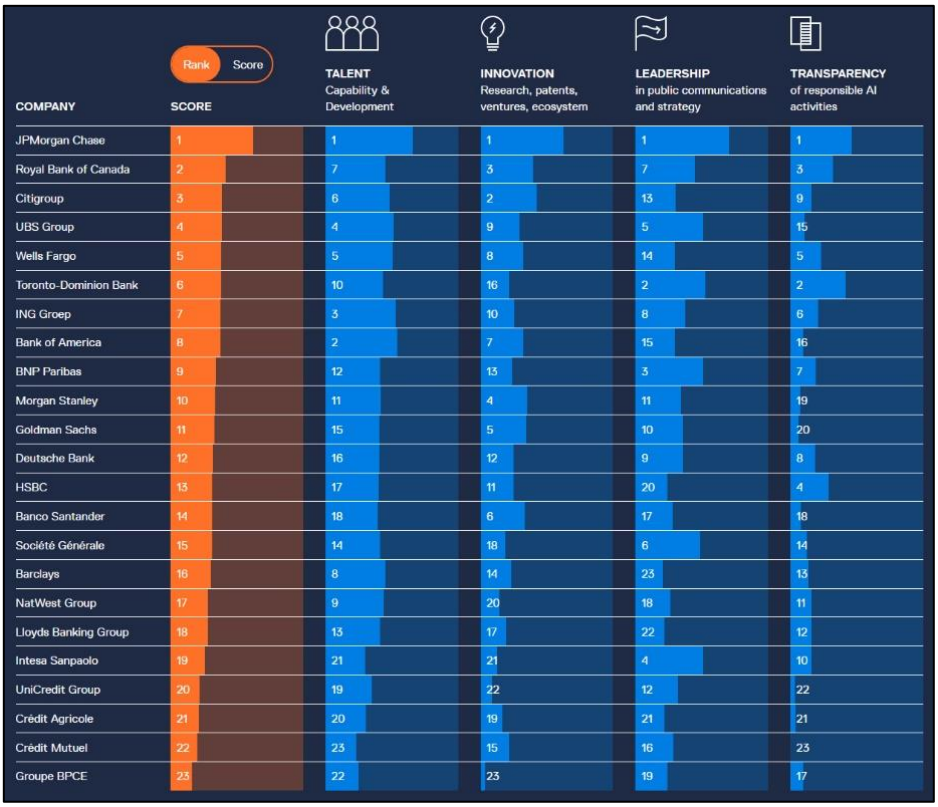


Figure 128 : Classement de 23 banques selon l'index IA d'Evident - Source : evidentinsights.com

D'un point de vue « score », voici les scores obtenus par les 23 banques retenues :

COMPANY	SCORE	TALENT Capability & Development	INNOVATION Research, patents, ventures, ecosystem	LEADERSHIP in public communications and strategy	TRANSPARENCY of responsible AI activities
JPMorgan Chase	62.6	65.8	62.5	70.9	46.2
Royal Bank of Canada	41.4	45.1	39.5	45.0	32.0
Citigroup	39.0	48.5	42.2	30.1	16.1
UBS Group	38.6	51.3	29.3	51.1	10.7
Wells Fargo	38.0	51.0	32.2	29.9	23.1
Toronto-Dominion Bank	37.9	43.2	21.6	53.0	41.5
ING Groep	37.6	53.0	25.4	37.7	21.0
Bank of America	37.4	54.3	32.5	29.8	9.6
BNP Paribas	34.4	41.3	24.2	51.3	19.4
Morgan Stanley	33.5	41.5	35.4	34.1	7.5
Goldman Sachs	33.0	40.9	34.6	36.1	6.1
Deutsche Bank	31.8	40.4	24.9	36.1	18.7
HSBC	31.6	40.1	25.2	24.2	28.9
Banco Santander	31.3	39.4	33.7	28.1	8.2
Société Générale	31.2	40.9	18.9	49.0	12.2
Barclays	30.3	45.1	23.6	20.5	13.8
NatWest Group	28.3	44.0	14.5	26.2	15.8
Lloyds Banking Group	27.9	41.3	19.3	21.6	15.8
Intesa Sanpaolo	25.9	29.9	12.9	51.3	16.0
UniCredit Group	21.9	35.0	8.6	32.3	3.3
Crédit Agricole	21.1	30.6	16.7	22.0	3.9
Crédit Mutuel	20.1	23.0	21.8	28.8	0.0
Groupe BPCE	16.1	24.8	3.6	24.4	9.6

Figure 129 : Score total des 23 banques de l'index IA d'Evident - Source : evidentinsights.com

Les banques nord-américaines, avec **JPMorgan, Capital One et Wells Fargo en tête**, dominent le secteur de l'innovation en IA.

Selon le rapport d'Evident, elles **surpassent largement leurs homologues européennes en termes de publications, de dépôts de brevets et d'investissements en IA**. En 2022, elles ont produit 80% des articles de recherche sur l'IA et effectué 60% des investissements dans les start-ups.

JPMorgan Chase se distingue avec plus de 120 chercheurs spécialisés, tandis que **Bank of America** excelle dans les dépôts de brevets, principalement dans le trading, les paiements et la conformité.

L'innovation en IA nécessite d'importants investissements, et Evident note une concentration marquée des ressources financières et humaines. Les cinq principales banques dominent largement en termes de publications, de brevets et d'investissements dans les start-ups.

En Europe, **BNP Paribas se démarque en matière d'IA**, notamment par ses investissements dans des start-ups en phase initiale, surpassant le **Crédit Mutuel** et **Barclays**. Cependant, en termes d'acquisitions, seul le groupe **BPCE** se distingue avec l'achat de **Rel8ed** par **Coface** en 2023. **BNP Paribas** est également **l'une des quatre banques européennes les plus actives en matière de recherche**, aux côtés d'**Intesa Sanpaolo**, **UniCredit** et **Crédit Suisse**.

Voici ce classement affiché autrement :

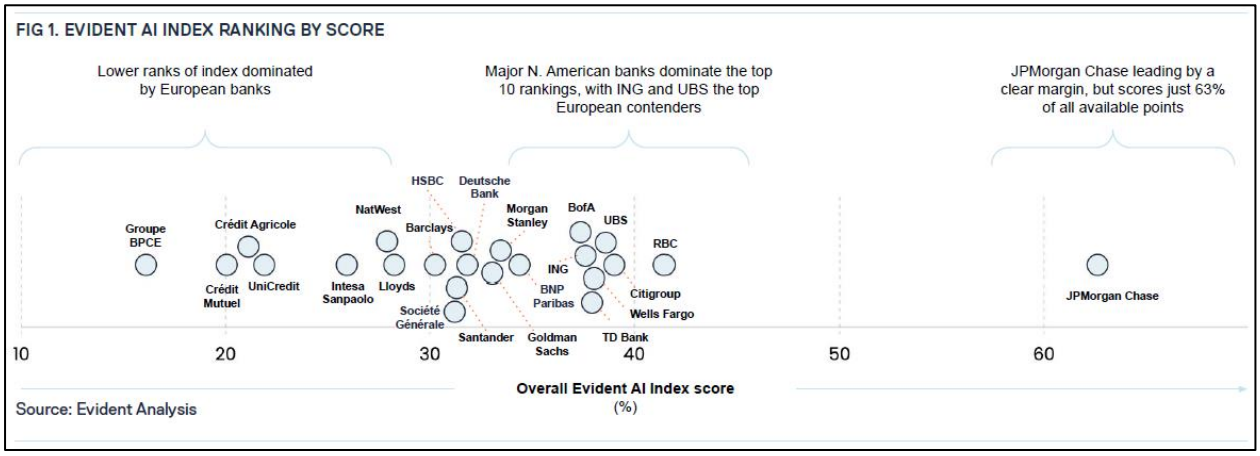


Figure 130 : JPMorgan Chase est loin devant dans le classement d'Evident - Source : evidentinsights.com

Les banques françaises sont donc dans le bas du classement.

Il est évident que les réalités diffèrent selon la géographie notamment compte tenu de la diffusion inégale de la technologie de l'IA, des différents viviers de talents, des politiques gouvernementales émergentes et des cadres réglementaires, cela peut ne pas être surprenant.

Selon cette étude d'Evident, les principaux constats par région géographique sont les suivants :

- **Banques nord-américaines :** Elles **dominent** l'index AI d'Evident, avec une avance notable sur leurs homologues européennes. Les banques canadiennes partagent également les premières places. **Les banques américaines détiennent 85% des brevets en IA**, avec **Bank of America** et **JPMorgan** Chase en tête. L'étude relève que celles-ci sont plus discrètes sur leurs ambitions en IA, ce qui pourrait être une stratégie pour protéger des informations sensibles.

- **Banques canadiennes :** Elles **surprennent par leurs performances**, avec **RBC** et **TD Bank** bien classées. Leur succès est dû à leur **transparence et leadership**, et reflète les efforts du Canada pour renforcer son industrie de l'IA.

- **Banques européennes :** Elles **semblent en retard** dans leur transformation par l'IA. Si certaines, comme **UBS** (Suisse), **BNP Paribas** (France) et **Deutsche Bank** (Allemagne), se démarquent, **la plupart sont à la traîne**, notamment sur les **axes de la transparence et de l'innovation**. Les banques françaises et italiennes ont des performances mitigées, malgré des points forts spécifiques.
  Concernant plus particulièrement les banques françaises, l'étude relève que malgré leurs forces historiques dans les études mathématiques et, par conséquent, le recrutement quantitatif**, les banques françaises se classent généralement plus bas en matière de talent**. **BNP Paribas** se distingue particulièrement en matière de leadership et de transparence, ce qui explique son classement dans le top 10.

- **Banques du Royaume-Uni** : Malgré la présence de **DeepMind** à Londres, les banques britanniques sont classées dans la **moitié inférieure de l'index**, possiblement impactées par des défis comme le Brexit.

Voici la disparité géographique des performances de chaque pays par rapport à la moyenne de l'index, par axe retenu dans l'étude d'Evident :

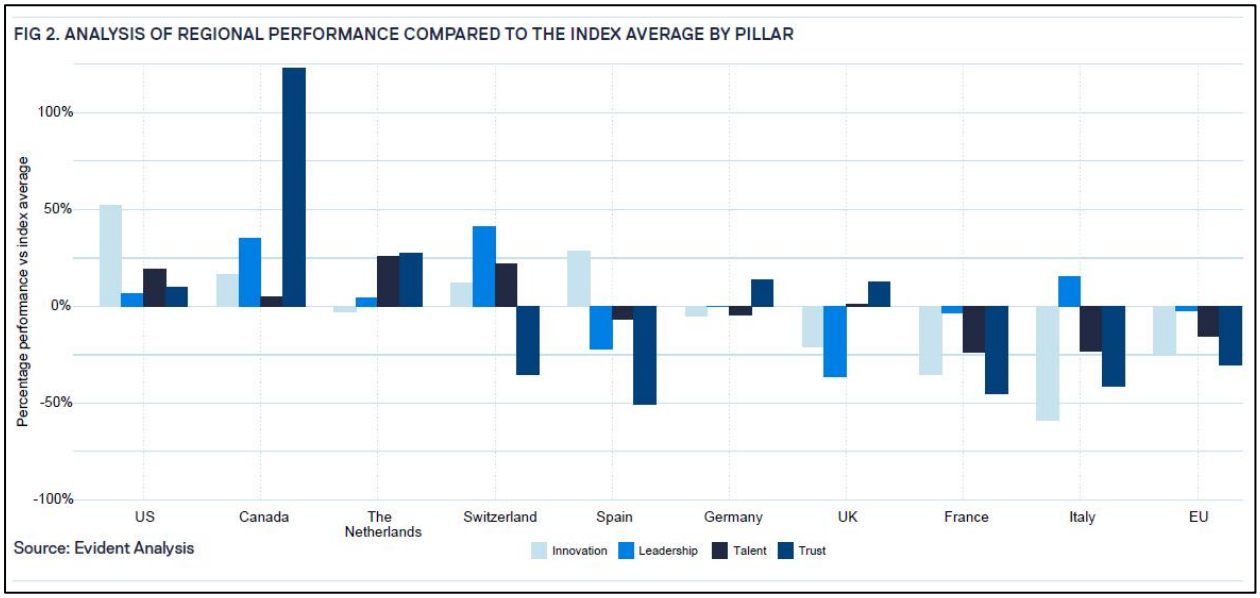


Figure 131 : Disparité géographique des performances de chaque pays - Source : evidentinsights.com

L'Union Européenne, et particulièrement la France, sont derrières mais cela ne présage pas des efforts réalisés notamment dans la mise en place de cadres réglementaires. A l'inverse, **aux Etats-Unis comme au Canada, les efforts sont portés sur l'innovation et la recherche, sans véritablement se concentrer sur ces aspects de réglementation**. Cela pourrait expliquer la situation soulignée dans cette étude.

Néanmoins, cela met tout de même en évidence que des efforts importants seront nécessaires en Europe pour être les leaders en IA.

3.1.3.2 Et demain ?

L'Intelligence Artificielle représente l'opportunité commerciale la plus importante actuellement. En 2022, les dépenses mondiales en solutions d'IA ont atteint l'incroyable somme de 433 milliards de dollars (selon le cabinet IDC), et le PIB mondial pourrait augmenter de 15,7 trillions de dollars (soit 15 700 milliards) d'ici 2030 (selon PWC) grâce à l'IA.

PWC explique dans le graphique suivant les catégories gagnantes d'ici 2030, le gain de productivité à lui seul permettra de gagner 55% :

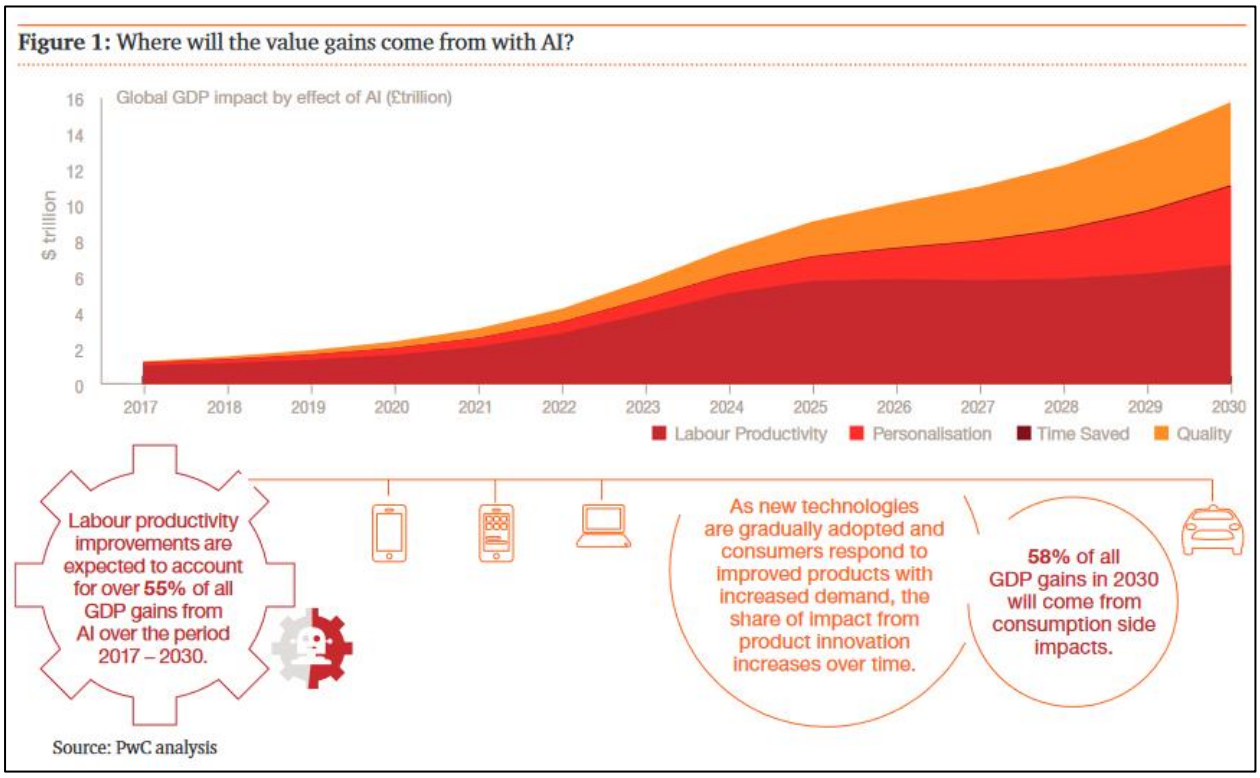


Figure 132 : Gain de productivité mondial d'ici 2030 - Source : PWC

Les entreprises les plus valorisées au monde misent déjà fortement sur l'investissement et la priorisation de l'IA. Elles comprennent que **l'IA offre une opportunité révolutionnaire pour les entreprises individuelles et représente une menace majeure pour celles qui restent à la traîne**.

C'est particulièrement vrai dans le secteur de la banque. **Les banques possèdent une mine d'or de données clients précieuses. Grâce à leurs efforts de transformation numérique au cours de la dernière décennie, elles sont idéalement positionnées pour profiter pleinement de la révolution de l'IA**. McKinsey estime que les technologies d'IA pourraient apporter jusqu'à 1 trillion (1 000 millards) de dollars de valeur supplémentaire à la banque mondiale **chaque année**.

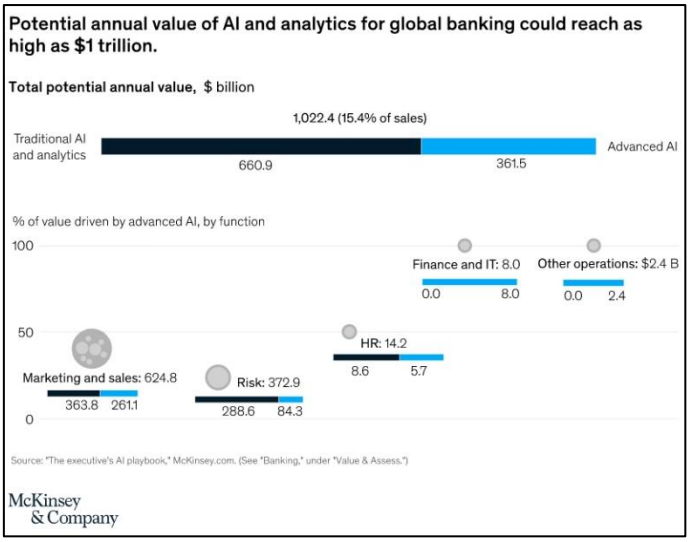


Figure 133 : Gain de valeur de l'IA à la banque mondiale par an - Source : [mckinsey.com]

L'étude de PWC classe les **services financiers comme 3ème secteur avec haut potentiel d'impacts lié à l'IA et estime à 41 % d'adoption d'ici les 3 prochaines années** :

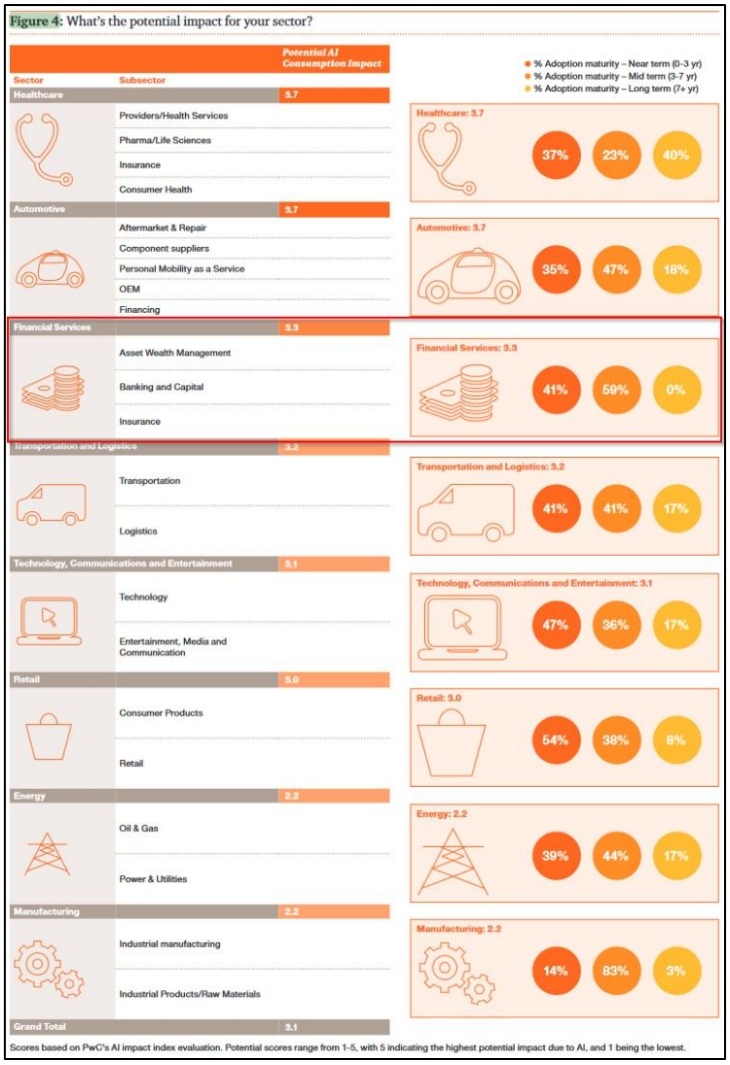


Figure 134 : 41 % d'adoption de l'IA dans les banque dans les 3 prochaines années - Source : [PWC]

Mais pour adopter l'IA, les banques sont confrontées à d'importants défis concurrentiels - venant des géants de la technologie, des fintechs et des néo-banques. Les banques traditionnelles pourraient **disparaître complètement si elles ne relèvent pas le défi de la transformation numérique impulsée par l'IA afin de rester compétitives**.

De plus, la transformation par l'IA est complexe, coûteuse et risquée. Une transformation majeure des entreprises est nécessaire en termes de technologie, de talents, de culture, et plus encore, avec des risques éthiques et de réputation importants pour ceux qui se trompent. Et la pression augmente, de l'intérieur mais aussi de la part des investisseurs, pour démontrer le retour sur investissement.

Les banques et les équipes dirigeantes capables de faire le saut et de minimiser les risques domineront le prochain cycle économique. Celles qui n'y parviennent pas pourraient avoir du mal à survivre.

3.1.4 Quelques cas d'usages concrets

Comme nous avons pu le voir précédemment, les banques s'intéressent grandement pour leur cas d'usage à l'Intelligence Artificielle. Et parmi les types d'IA, **elles trouvent un intérêt fort pour l'IA générative et en particulier le NLP (Natural Language Processing)**.

Le développement du NLP dans le secteur bancaire a commencé à prendre de l'ampleur depuis le milieu des années 2010.

En effet, le NLP peut être utilisée pour un large éventail d'applications au sein des banques. Principalement pour des **tâches d'efficacité opérationnelle**, comme les **chatbots**, la **traduction de langue** et la **génération du contenu**, ce qui en fait une technologie idéale pour mettre en œuvre divers cas d'utilisation.

Par conséquent, **elles investissent des ressources importantes dans la recherche et le développement afin de construire le meilleur modèle linguistique**.

Ces modèles utilisent de grandes quantités de données, notamment les dernières techniques d'IA qui en demandent toujours de plus en plus pour former ces modèles, les affiner pour des tâches spécifiques et **améliorer constamment leurs performances**.

Cette frénésie est motivée par le désir de gagner un avantage concurrentiel sur le marché et d'être le premier à apporter les avantages de ces modèles à leurs besoins.

En outre, les banques cherchent également des moyens de rendre les modèles plus économes en énergie et durable, en réduisant leur empreinte carbone.

Voici quelques exemples **généraux** de cas d'utilisation de l'IA générative dans les banques :

- **La banque d'investissement :**
  - **Aider à la création de résumés et de diaporamas** en générant automatiquement du contenu écrit,
  - **Automatiser le processus de réponse aux demandes des clients et fournir des recommandations personnalisées** an analysant les données des clients et en identifiants leurs besoins clés,
  - **Aide à la rédaction de documents juridiques**, tels que les contrats par exemple, en identifiants les termes et clauses juridiques essentiels.

- **Le trading :**
  - **Aider les traders à identifier les tendances du marché et à faire des transactions éclairées en analysant les nouvelles** et les publications sur les médias sociaux en temps réel,
  - **Extraire des informations à partir de grands volumes de données de marché**, telles que les prix et les volumes d'échange historiques, qui peuvent être utilisées pour générer des stratégies de trading algorithmiques,
  - **Aider à surveiller l'évolution de la réglementation**, comme les changements aux exigences de marge ou les nouvelles réglementations régissant le trading à haute fréquence.

- **La gestion des risques :**
  - **Surveiller les informations et les médias sociaux à la recherche de risques potentiels** pour la réputation de la banque, tels qu'une couverture médiatique négative ou des violations de la réglementation,
  - **Participer aux tâches de conformité**, telle que la révision de documents et de la détection des fraudes, en identifiants les termes et concepts clés dans de grands volumes de données textuelles,
  - **Aider à la surveillance et à la gestion du risque de contrepartie** en analysant les publications d'informations et de médias sociaux liées aux contreparties potentilles.

- **Compliance :**
  - **Contribuer au respect de la réglementation** en surveillant automatiquement les informations et les médias sociaux à la recherche de violations potentielles et en identifiant les principaux concepts de conformité au sein des documents internes,
  - **Automatiser le processus de création des rapports de conformité**,
  - **Faciliter la surveillance des communications des employés**, telle que les e-mails et les messages instantanés, en identifiant les problèmes de conformité potentiels, tels que les opérations d'initiés ou des manipulations de marché.

- **Asset management :**
  - **Aider à la gestion de portefeuille** en analysant les performances historiques et en prédisant les tendances futures,
  - **Aider à la création de rapports d'investissement** en générant automatiquement du contenu écrit,
  - **Aider à surveiller les développements réglementaires et la conformité** en surveillant automatiquement les nouvelles et les médias sociaux.

- **La recherche :**
  - **Aider à la sélection des entreprises et des industries** en résumant automatiquement les informations clés et en identifiant les tendances,
  - **Générer des synthèses**.

L'IA générative a le potentiel de **révolutionner** la façon dont la banque fonctionne, **mais il est important de noter que l'utilisation de ces technologies peut également soulever des questions de confidentialité qui doivent être traitées**.

3.1.4.1 Chez Crédit Agricole

Crédit Agricole et en particulier **Crédit Agricole Corporate and Investment Bank (CACIB)**, bras financier et d'investissement du groupe français, opère dans 33 pays, couvrant l'Europe, les Amériques, l'Asie-Pacifique et le Moyen-Orient.

CACIB se divise en six divisions principales, agissant comme une banque d'investissement, commerciale et de marchés. Elle gère aussi **les relations clients du Crédit Agricole à l'international, l'optimisation de la dette et la distribution, ainsi que les financements structurés**.

En raison de régulations strictes, CACIB est particulièrement vigilant quant aux entreprises avec lesquelles elle collabore, luttant activement contre la fraude, le financement du terrorisme et le blanchiment d'argent.

La non-conformité à ces régulations peut entraîner de lourdes sanctions. Pour évaluer la solvabilité des clients, certains employés doivent parcourir de vastes rapports financiers, une tâche longue et susceptible d'erreurs. L'enjeu est de détecter tout risque lié, par exemple, au terrorisme. Pour faciliter cette analyse, Crédit Agricole CIB a intégré une technologie NLP afin d'examiner les documents financiers et repérer d'éventuels risques pour sa stabilité financière (voir lien ici).

Voici également d'autres cas d'usages chez **Crédit Agricole** selon plusieurs domaines :

- **Fraude** :
  - **Détecter la fraude aux paiements (virements émis)**

Ce cas d'usage permet d'offrir aux clients une solution de détection des paiements atypiques, leur donner la possibilité de bloquer ces paiements et générer des alertes pour informer les clients (IBAN FRAUD).

  - **Détecter le blanchiment d'argent**

Historiquement basé sur un scénario analytique fondé sur des règles, l'IA permet à la sécurité financière d'identifier des schémas de blanchiments d'argent plus complexes et sophistiqués.

  - **Détecter la fraude dans les salles de marché**

Il permet d'analyser les données via la transcriptions d'audio en texte, des enregistrements des échanges entre traders et vendeurs dans les salles de marché afin de prévenir la fuite d'informations confidentielles ou fraude

- **Connaissance des clients** :
  - **Analyse des verbatim**

Une solution d'Intelligence Artificielle interne d'analyse des verbatim clients.

Cette solution analyse les verbatims selon des axes exprimées par le client, avec un taux de bonne reconnaissance proche de 80%.

La solution, composée d'un moteur d'IA et de deux interfaces graphiques de gestion des fichiers et de Data Visualisation des résultats d'analyse.

  - **Gérer des portefeuilles avec l'IA Analytics**

Il permet de mieux définir la segmentation des clients et optimiser les efforts commerciaux.

  - **Analyser les comportements client**

Il permet aux traders et aux commerciaux de comprendre les comportements des clients et la valeur du flux client afin d'ajuster leurs stratégies de prix.

  - **Analyser des clients à risque**

Il permet d'améliorer l'exactitude de l'identification des informations négatives et la nouveauté d'un événement.

Lors de l'analyse des clients à risque, la détection des informations négatives est une phase factieuse réalisée par l'équipe sécurité financière. Cet outil est un assistant cognitif qui raccourcit le temps consacré à cette recherche.

Auparavant, ce travail était effectué manuellement et client par client.

L'information permet aux employés de gagner un temps considérable. C'est pourquoi cet outil est une solution essentielle pour garantir un processus de qualité pour la sécurité financière

- **Identifier des appels d'offre pour les actions**

L'IA permet d'identifier et automatiser les milliers d'appels d'offres échangés manuellement sur une base quotidienne avec les clients

- **Anticiper la dégradation des crédits**

Le système d'IA permet d'identifier dès que possible les signes de détérioration probable de la solvabilité des clients afin d'agir de manière préventive

- **Tendance des dépôts**

L'IA permet ici d'anticiper les tendances des dépôts et donc une meilleurs allocation des ressources.

- **Trading :**
  - **Prédire des opportunités commerciales**

Le système d'IA permet, via l'extraction automatique en temps réel d'informations de presse, de prévoir des opérations précises à l'avenir pour **identifier les priorités stratégiques des clients et prospects** qui seront convertis en opportunités commerciales

L'IA propose une recommandation de News aux traders : ceux-ci posent des questions en langage naturel à l'Intelligence Artificielle laquelle analyse l'intention du Trader et lui répond en conséquence.

Ici, l'Intelligence Artificielle sert aussi **d'outil prévisionnel des ventes**. En croisant les données internes à l'entreprise et des données externes (historique des ventes, positionnement de la concurrence, saisonnalité, etc.), en fonction des informations du marché, : l'IA prédit de façon limité les tendances et aide les banquiers à **établir des scénarios facilitant la prise de décision pour trouver des opportunités**.

- **Automatiser le trading Repo**

Les demandes de devis sont reçues par la banque dans une **langue non formatée**. Un défi pour l'automatisation de la tarification Repo car chaque client a sa propre façon de demander une opération pareille. 90% des opérations Repo sont effectuées manuellement.

Des centaines de demandes de devis envoyées par les clients via des chats ont été utilisées pour alimenter un modèle de traitement du langage naturelle (NLP). Ce modèle a appris à lire et interpréter le contenu des achats afin d'identifier les caractéristiques pertinentes pour chaque transaction demandée. Les informations financières sont ensuite envoyées à un modèle alimentant un **moteur de tarification « Repo »**.

Grâce à l'IA, les vendeurs et les traders ont gagné beaucoup de temps, une réactivité accrue et une réduction du risque opérationnel.

- **Automatiser le traitement de données financières**

Les équipes financières traitent souvent un grand volume de documents qui contiennent des informations de sécurité financière critiques et qui peuvent être complexes et impliquer de multiples acteurs mondiaux. **L'extraction manuelle de ces informations à partir de documents physiques peut prendre du temps et entrainer des erreurs**.

Il est donc nécessaire de trouver des méthodes plus efficaces et plus précises pour traiter ces documents.

Cet outil utilise une technologie avancée de **reconnaissance de caractères pour numériser des documents financiers et extraire rapidement des informations précises de sécurité financière**.

En automatisant ce processus, l'outil peut aider les équipes financières à gagner du temps et à réduire le risque d'erreurs, tout en améliorant la sécurité globale des opérations financières grâce à la numérisation des documents.

- **E-mailing :**
  - **Classifier des e-mails**

Il permet de faciliter la gestion des e-mails envoyés à une adresse e-mail générique. Sa vocation première est d'aider à **collecter, gérer et suivre les demandes** des clients et partenaires.

L'utilisation de l'IA dans la classification des e-mails s'est avérée être un gain de temps précieux pour les utilisateurs de la plateforme. Non seulement cela aide **à catégoriser rapidement et avec précision les e-mails**, mais cela permet également **une amélioration continue en apprenant des commentaires des utilisateurs**. Cela a conduit à une efficacité et une productivité accrues au sein de l'équipe, et à la possibilité d'un déploiement plus large dans d'autres équipe.

L'intervention de l'IA dans le processus de traitement des e-mails comprend les cinq étapes suivantes :

- L'utilisateur envoie un mail au métier.
- Gestion des mails par l'IA : L'IA propose une catégorisation et applique les règles existantes pour proposer une catégorie de demande.
- Réception des e-mails chez le métier : l'e-mail s'affiche chez le métier avec une catégorie suggérée, ce qui fait gagner du temps aux utilisateurs.
- Les utilisateurs peuvent valider ou rejeter la catégorisation suggérée par l'IA.
- Si les utilisateurs rejettent la proposition d'IA, ils peuvent remplir la bonne catégorie et le résultat est envoyé à l'IA pour la correction des règles et les futures suggestions de catégorisation.

- **Analyse des e-mails**

Il permet d'analyser les e-mails entrants et de proposer une réponse pour faire gagner beaucoup de temps aux conseillers en agence. De fait, ces derniers n'ont pas à retaper les réponses de A à Z, mais ont dans la plupart des cas un moteur d'IA qui peut préremplir et rédiger 95 % de la réponse.

- **Organiser les flux d'actualité et le recevoir dans la boite mail**

L'IA permet ici aux collaborateurs de recevoir chaque jour dans la boite mail les actualités des clients et des secteurs sélectionnées selon les critères que le collaborateur a choisi comme une sorte de Newsletter personnalisée.

- **Chatbot :**
  - **Robot intelligent**

Un robot d'Intelligence Artificielle, développé et utilisé par Crédit Agricole Bank Polska, a reçu le titre de meilleure innovation dans le Best in RPA 2023 le plus grand concours de solution robotique et d'automatisation en Pologne.

Le robot prend en charge le traitement des réclamations et des demandes de la banque ainsi que d'autres processus **après-vente**.

- **Optobot**

Un assistant virtuel permettant aux utilisateurs du site « Optim » de répondre aux questions fonctionnelles (24/24 et 7j/7) afin de **réduire le nombre d'appels vers la hotline**.

- **Chatbot Compliance**

Outil innovant développé en interne dans le cadre de Framework Compliance. Une base de connaissance a été créé par sujet **anticipant les questions de collaborateurs les plus fréquentes**.

Grace à sa composante d'intelligence Artificielle, le chatbot est capable non seulement de répondre aux questions, mais aussi de **proposer des textes de Gouvernance et de diriger l'utilisateur vers la page pertinente**.

Ce chatbot est un outil pour aider les employés, car il vise à rendre les sujets de conformité plus accessibles et compréhensibles pour les utilisateurs non experts. Une meilleur compréhension des sujets protège d'avantage la banque contre les risques majeurs auxquels elle pourrait être exposée.

- **CAesar le Metabot**

Un assistant virtuel intelligent des conseillers en Caisses régionales

Dans une caisse régionale, l'information métier est contenue dans plus de **10 000 documents** et **un conseiller consacre en moyenne 7 heures 30 par semaine à la recherche d'informations**. Avec **Caesar**, les conseillers peuvent questionner, en langage naturel, toute cette base de connaissance.

Résultat, près de 40 000 heures gagnées par mois. La solution (appelée **Metabot**) est intégralement développée et maitrisée en interne.

- **Gestion documentaire :**
  - **CACIB Search**

Une sorte de Google interne avec l'IA, qui permet d'identifier des informations dans une base documentaire (documentation de conformité, documentations juridiques, …) d'une manière intelligente.

Il permet **d'analyser, contrôler la présence de clauses et d'informations ciblées au sein de contrats volumineux à structure variable**. Aussi, sécuriser l'activité de l'entreprise sur les contrats ou sur la gestion de prêts structurés.

Les équipes de métiers doivent analyser les contrats clients pour déterminer s'ils sont éligibles à ce type d'opération. LA recherche d'informations spécifiques dans des contrats volumineux prend beaucoup de temps. **Il a permis de réduire de 75% le temps consacré à l'analyse de chaque contrat** et d'effectuer des recherches auparavant trop longues lorsqu'elles étaient effectuées manuellement.

- **IA documentaire de dernière génération, performante et frugale**

Le **DataLab Groupe** a développé grâce à sa R&D des solutions internes d'intelligence Artificielle documentaire performantes intégrées à de multiples processus clients et collaborateurs. Ses travaux les plus récents ont conduit à une nouvelle génération d'IA documentaire capable de s'affirmer de **l'OCR** (« *Optical Character Recognition* » ou en français « *Reconnaissance Optique de Caractères* »), d'apprendre en **minimisant les besoins d'exemples** créés manuellement, tout en **dépassant les performances et la robustesse** des générations précédentes.

- **Classification des justificatifs et extractions des informations**

Il permet d'améliorer le contrôle des entrées en relation des particuliers. L'IA **classifie les justificatifs remis par le client** (pièce d'identité, justificatif de domicile...) puis en **extrait les informations nécessaires aux contrôles**.

Grace à cet outil, le collaborateur concentre des efforts sur les points de contrôle à forte valeur ajoutée (par exemple : contrôle des documents de mauvaise qualité, difficilement lisible par l'IA ou encore le cohérence du dossier).

En plus de permettre un gain de temps en réalisant les tâches simples, l'outil guide les débutants dans la réalisation des contrôles plus complexes pour une confiance accrue en leur travail. Le potentiel de gain de temps est estimé à 50% par document.

- **Analyser automatiquement des rapports annuels**

Il permet d'analyser automatiquement des rapports annuels pour proposer des produits structurés qui répondent mieux aux besoins des clients de la banque. Pour identifier les sujets stratégiques pour les entreprises, les services conseil aux entreprises doivent passer beaucoup de temps à analyser ces rapports.

- **Collecte de données RSE sur les émissions de Gaz à Effet de Serre (GES)**

La « *collecte de données sur les émissions de GES* » est un outil d'IA et de traitement de langue naturel pour analyser et trouver des informations clés sur les GES en matière de **R**esponsabilité **S**ociale et d'**E**ntreprise (RSE) dans les rapports et **R**apports **A**nnuels (RA) des homologues du métier.

L'IA permet ici de :

- **Lire et analyser les rapports annuels pour détecter automatiquement les relations** entre les clients, les pays et les activités pour répondre aux questions de sécurité financière.
- **Le métier doit extraire et consolider les données** sur les émissions de GES pour les rapports extra-financiers, l'évaluation des risques et les trajectoires
- **Extraire des informations GES spécifiques à partir d'un ensemble de documents volumineux** pour faire gagner du temps à l'expert métier lors de l'analyse

Comment ça marche :

Le document téléchargé par l'utilisateur est traité avec un modèle de vision par ordinateur pour découper le texte en unité de recherche en respectant la mise en page du document. Grâce à la recherche par mot clé, toutes les pages du document contenant les informations sont identifiées et restituées à l'utilisateur sur une interface dédiée. L'outil est lié avec une interface pour que l'utilisateur puisse effectuer son analyse et son interprétation des données récupérées.

Les rapports RSE annuels peuvent atteindre plus de **400 pages** et la recherche manuelle de chaque information GES peut prendre 30 minutes en moyenne. Grace à cet outil le temps de récupération des informations est réduit à 5 minutes en moyenne

- **Formation interne :**
  - **Améliorer l'expérience d'apprentissage (formation) en la rendant plus personnalisée**

L'IFCAM a voulu innover dans ses dispositifs de formation. Dans ce cadre, l'entité de formation généralise ses « *Scan'up* » (dispositif de montée en compétences) en « *adaptive learning* », un dispositif d'évaluation et de monter en compétences personnalisée et optimiser par l'IA.

L'évolution de contexte d'apprentissage fait émerger de nouvelles réponses pour adresser l'expérience apprenant.

L'IFCAM va plus loin avec la mise en œuvre d'une « *learning expérience plateforme* » une plateforme offrant à l'apprenant une **expérience personnalisée** et au service formation une vision fine et un pilotage efficace des compétences de leur équipes, transformant le paysage métier de la formation.

L'avantage : Le temps gagné sur les formations et le dossier est à ce titre extrêmement intéressant puisqu'il débouche concrètement sur des formations à la fois plus courte et plus efficaces.

  - **Monitorer les solutions d'IA pour garder la maitrise**

La nouvelle plateforme de monitoring d'IA permet de **sécuriser les performances** des solutions IA dans la durée et de **prévenir les impacts d'un éventuel dysfonctionnement**.

Il est essentiel de monitorer les données entrées, les modèles de performances statiques et techniques. Pour simplifier sa mise en œuvre complète par les développeurs, le **DataLab Groupe** du **Crédit Agricole** a créé une plateforme **permettant aux développeurs d'intégrer ces fonctionnalités et aux métiers de superviser la santé de l'IA.**

3.1.4.2 Chez BNP Paribas

BNP Paribas est l'une des plus grandes banques françaises et mondiales, laquelle offre une vaste gamme de services financiers, de la banque de détail à la banque d'investissement, avec une présence significative à l'échelle internationale.

Cette banque a également mis en œuvre important de projets autour de l'IA et notamment du NLP.

Voici quelques cas d'usages :

- **Assistant virtuel**

La filiale espagnole de **BNP Paribas Personal Finance** a réalisé un projet innovant visant à améliorer l'efficacité du service client et l'expérience utilisateur et qui consistait à adopter un assistant virtuel de Nuance, **combinant l'Intelligence Artificielle et l'intervention humaine**. Il permet d'améliorer l'expérience client et gagner en efficacité via la transformation et le développement du modèle de service client en trouvant le bon équilibre entre service « humain » et « virtuel ».

Cet assistant, premier point de contact avec les clients, peut répondre à plus de 90 questions liées à l'entreprise. Si une requête dépasse ses capacités, elle est redirigée vers un agent humain. Après deux ans d'utilisation, 46% des demandes des clients ont été résolues par l'assistant sans intervention humaine.

Voici quelques chiffres qui reflète du succès de l'assistant virtuel :

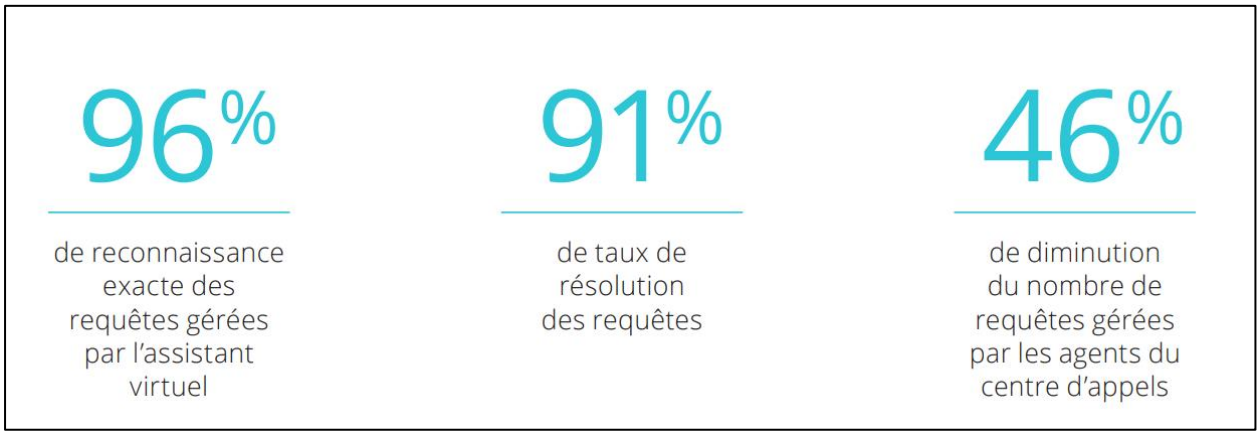


Figure 135 : Chiffres du succès de l'assistant virtuel de BNPP – Source : nuance.com

Suite à ce succès, **BNP Paribas Personal Finance Espagne** a étendu l'utilisation de la technologie Nuance, intégrant des fonctionnalités de compréhension du langage naturel à ses canaux vocaux et digitaux.

Cette initiative s'inscrit dans la volonté de la banque d'optimiser son service client tout en renforçant son efficacité.

- **Gestion des contrats d'épargne**

Ce système d'IA permet **l'amélioration de l'efficacité dans la modification des clauses bénéficiaires** et **de renforcer de la sécurité** lors du processus de modification des clauses bénéficiaires.

Les contrats d'assurance vie incluent une clause spécifiant le bénéficiaire qui recevra les fonds à la mort du titulaire du contrat. Il est courant de modifier cette clause, et cela peut être fait à plusieurs reprises pour un même contrat.

Auparavant, le processus de modification était manuel et volumineux. Pour optimiser cela, une automatisation a été mise en place pour les contrats d'assurance vie des clients BNP. Désormais, lorsqu'une modification de clause bénéficiaire est demandée via le système « *Appian* », un robot RPA est déclenché.

Ce robot vérifie les informations dans notre système de gestion, met à jour la clause sur le contrat en question, puis renvoie le résultat au système « *Appian* » pour finaliser la tâche.

- **Chatbot WhatsApp**

Au Brésil, **Cetelem** (filiale du groupe BNP Paribas) a introduit un **chatbot performant sur la plateforme de messagerie *WhatsApp***. Ce chatbot vise à améliorer la communication avec les clients en offrant des réponses instantanées, en les orientant vers nos offres et services, et en facilitant les transactions à toute heure du jour ou de la nuit.

3.1.4.3 Chez Société Générale

La **Société Générale** est l'une des principales banques françaises et l'une des plus anciennes. Avec ses racines fermement ancrées en France, la banque s'est développée pour devenir un acteur majeur sur la scène bancaire internationale.

La « SocGen », comme de nombreuses grandes banques et institutions financières, a exploré et investi dans l'Intelligence Artificielle pour améliorer ses services et opérations.

En voici quelques cas d'usage :

- **Améliorer le pilotage des profils des employés**

Ce système d'IA permet **d'augmenter la visibilité des collaborateurs**, de **faciliter l'identification des profils pertinents par les managers** et d'avoir un référentiel dynamique des compétences pour les ressources humaines.

- **Gestion de la relation client**

Cette IA permet d'orienter automatiquement les demandes des clients vers le professionnel compétent pour une réponse rapide mais aussi **d'extraire de manière autonome les informations contenues dans les e-mails et leurs pièces jointes envoyés par les clients**.

L'e-mail est toujours un canal de communication privilégié pour les clients lorsqu'ils interagissent avec la banque. La Société Générale enregistre en moyenne plus de 1 000 000 d'e-mails annuels provenant de leurs clients (source : GBIS post-opérations – 2019). Ainsi, leur approche digitale vise à minimiser autant que possible les échanges non structurés avec leurs clients.

- **Surveillance des communications des collaborateurs régulés**

L'Intelligence Artificielle permet ici de surveiller et **analyser automatiquement les communications vocales et écrites** des employés régulés en plusieurs langues mais aussi d'assurer le suivi des échanges **audio dans 12 langues et des communications écrites dans 26 langues** grâce à l'IA.

La réglementation exige que les communications sensibles soient enregistrées et surveillées. Étant donné la quantité importante de ces communications, cette tâche nécessite des centaines d'heures de travail manuel. L'IA permet donc ici un gain de temps très important.

3.1.4.4 Chez Orange Bank

Orange Bank est la filiale bancaire du groupe **Orange**. Lancée en 2017 en France, elle s'inscrit dans la tendance des néo-banques qui utilisent la technologie pour offrir des services bancaires simplifiés et centrés sur le client.

Cette banque récente utilise l'Intelligence Artificielle de manière innovante pour améliorer l'expérience client et optimiser ses services.

Voici un cas d'usage d'Orange Bank avec de l'IA :

- **Djingo chez Orange Bank**

Djingo est le conseiller virtuel d'Orange Bank basé sur l'Intelligence Artificielle **d'IBM Watson**. Ce conseiller virtuel d'Orange Bank, traite près de 24.000 demandes par semaine en moyenne, avec un taux de compréhension des questions de 85%.

Il répond principalement à des questions générales sur le fonctionnement, les tarifs, etc., similaires à une FAQ. Il peut également répondre à des demandes plus spécifiques, comme le blocage temporaire d'une carte bancaire. Djingo est capable de répondre à 500 questions en langage naturel et est en apprentissage permanent. Les questions plus complexes sont traitées par des conseillers humains.

Il est disponible 24/7, et près de 20% des conversations ont lieu en dehors des heures d'ouverture du centre de relation clients. Il joue donc un rôle essentiel dans la stratégie d'Orange Bank pour **fournir un service client efficace et moderne**.

Pour Orange, les chatbots sont un moyen de répondre au besoin d'instantanéité des clients.

Pour conclure, cette liste de cas d'usage ne se veut pas exhaustive. En effet, il existe beaucoup d'autres applications au sein de ces banques mais également au sein de banques qui n'ont pas été citées. Le but était ici de donner un rapide aperçu de ce qui existe bien que celui-ci ne couvre pas tous les types d'applications possibles.

Il existe par exemple des IA liées au traitement de l'image, la surveillance et la sécurité, la création graphique, la gestion des banques de données visuelles, etc. **Les domaines d'application sont donc nombreux et tous sont assujettis à la mise en place de bonnes pratiques en vue de la future réglementation européenne**.

3.2 Les bonnes pratiques et recommandations pour une IA responsable dans le secteur bancaire

Comme abordé précédemment, l'IA responsable est devenue un enjeu majeur pour les sociétés utilisant l'IA et en particulier les banques. Avec l'arrivée de cadres règlementaires et notamment l'AI Act, elles se doivent de prendre en considération pour se mettre en conformité.

Certaines sont très avancées dans la prise en compte de ses principes quand d'autres sont un peu plus « à la traîne ». Qu'en est-il vraiment aujourd'hui ?

Benoit ROTTEMBOURG
Responsable Régalia (INRIA)

L'IA responsable pour moi, c'est anticiper les problèmes.

Les banques et les assurances sont parmi les entités qui sont les plus sensibles à ce discours parce que :

- Elles sont habituées à être régulées,
- Elles ont déjà des autorités de contrôle interne dont elles ont un peu peur en général, qui surveillent,
- Elles passeront des mots à l'acte le plus vite possible.

3.2.1 Etat des lieux : maturité de l'IA responsable dans ce secteur

Dans une précédente partie, nous avions mentionné de manière globale (i.e. tous secteurs confondus) les chiffres obtenus dans une étude appelée « *Are You Overestimating Your Responsible AI Maturity ?* » par BCG GAMMA (représente la branche du BCG spécialisée en IA) réalisée en 2021 auprès de plus de 1 000 grandes entreprises lesquels sont **pour rappel** :

- Seulement 48% des entreprises ayant adopté l'IA à grande échelle l'ont fait de manière responsable.
- 55% des entreprises surestiment leur niveau de responsabilité en matière d'IA.
- Beaucoup d'entreprises pensent être en avance, mais sont en réalité en retard.

D'après cette même étude, il est particulièrement intéressant de souligner le constat suivant, propre au secteur bancaire et financier :

- **56 % des banques et instituts financiers sont en avance** dans la mise en place d'IA responsable,
- **32 %** d'entre elles sont **en cours de mise en place,**
- **12 % sont en retard**

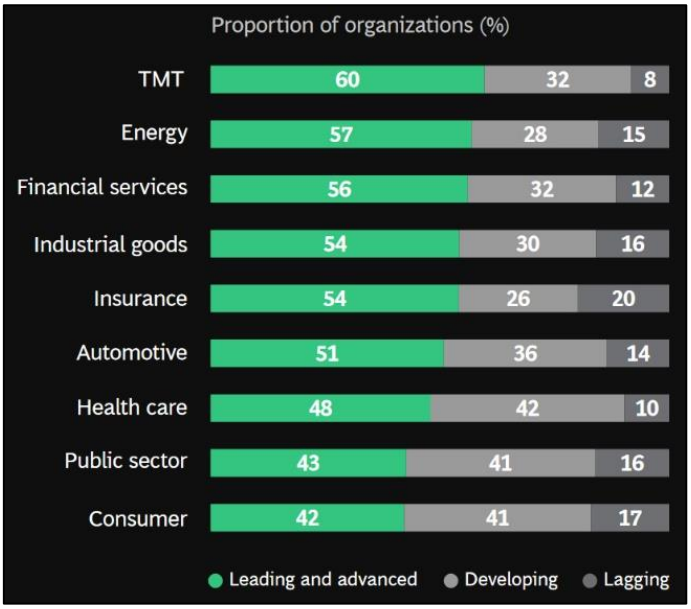


Figure 136 : Proportion de services financiers ayant mis en place de l'IA responsable – Source : bcg.com

Le secteur des services financiers est fortement réglementé et a donc un besoin de forte conformité et de gestion des risques. Par conséquent, il n'est donc pas surprenant que ce secteur souhaite se mettre rapidement en maturité vis-à-vis de l'IA responsable pour être en conformité le plus tôt possible par rapport aux différents cadres réglementaires à venir dont l'AI Act. Le secteur bancaire et financier est naturellement **un des 3 premiers secteurs en termes d'avancement dans la mise en place de l'IA responsable**.

Une autre étude récente citée précédemment est celle d'Evident, dans laquelle un classement est également réalisé plus spécifiquement sur la maturité des banques sur l'axe de la « **Transparency** », lequel reflète dans quelle mesure les banques communiquent **publiquement** sur un large éventail d'activités responsables en matière d'IA et rendent visibles leurs efforts pour créer des contrôles spécifiques à l'IA.

Bien que la transparence ne soit pas la solution miracle pour instaurer la confiance - et qu'elle ne soit pas le seul élément important - elle est une première étape pour démontrer la fiabilité d'une institution à sa large gamme de parties prenantes.

COMPANY	SCORE	TALENT Capability & Development	INNOVATION Research, patents, ventures, ecosystem	LEADERSHIP in public communications and strategy	TRANSPARENCY of responsible AI activities
JPMorgan Chase	1	1	1	1	1
Royal Bank of Canada	2	7	3	7	3
Citigroup	3	6	2	13	9
UBS Group	4	4	9	5	15
Wells Fargo	5	5	8	14	5
Toronto-Dominion Bank	6	10	16	2	2
ING Groep	7	3	10	8	6
Bank of America	8	2	7	15	16
BNP Paribas	9	12	13	3	7
Morgan Stanley	10	11	4	11	19
Goldman Sachs	11	15	5	10	20
Deutsche Bank	12	16	12	9	8
HSBC	13	17	11	20	4
Banco Santander	14	18	6	17	18
Société Générale	15	14	18	6	14
Barclays	16	8	14	23	13
NatWest Group	17	9	20	18	11
Lloyds Banking Group	18	13	17	22	12
Intesa Sanpaolo	19	21	21	4	10
UniCredit Group	20	19	22	12	22
Crédit Agricole	21	20	19	21	21
Crédit Mutuel	22	23	15	16	23
Groupe BPCE	23	22	23	19	17

Figure 137 : Notation de l'axe "Transparency" pour les banque française selon Evident - Source : evidentinsights.com

Le rapport classe BNP Paribas en 7ème position sur cet axe « Transparency », seule banque française faisant partie du top 10 du classement établi par Evident. Les autres banques françaises sont plus loin dans le classement :

- Société générale : 14ème
- Groupe BPCE : 17ème
- Crédit Agricole : 21ème
- Crédit Mutuel : 23ème et dernière du classement sur l'axe de la transparence.

Ce classement est cependant à mitiger. En effet, pour Evident, le pilier de la Transparence mesure comment les banques communiquent publiquement leurs approches de l'IA responsable, par exemple **à travers l'annonce de principes éthiques, des collaborations avec d'autres organisations, et la publication de recherches originales**.

En effet, la transparence dans le contexte de l'IA responsable est encore quelque peu naissante, et il n'existe pas encore de manière standard pour les entreprises de rendre compte de celle-ci.

Evident s'est donc concentré sur des déclarations publiques explicites montrant comment différentes banques communiquent leurs approches sur le sujet - en puisant dans des sources telles que les rapports d'entreprise, les communiqués de presse, les embauches de dirigeants, les intitulés de postes, les descriptions de postes et la recherche académique.

Les preuves publiquement disponibles ont été trouvées sous les formes suivantes :

- Principes d'IA responsables ou éthiques publiés
- Collaborations avec d'autres institutions pour faciliter la compréhension des thèmes clés autour de l'IA responsable ou de l'éthique
- Rôles dédiés liés à l'éthique de l'IA, l'éthique des données ou l'IA responsable
- Outils utilisés et partagés en externe pour améliorer l'explicabilité et l'équité
- Recherches liées à des domaines spécifiques de l'IA responsable, tels que l'explicabilité et l'équité
- Descriptions de la manière dont les approches de gestion des risques ont été adaptées pour l'IA
- Rôles dédiés pour couvrir les défis de gouvernance spécifiques à l'IA.

Voici les éléments publiés publiquement, par répartition géographique :

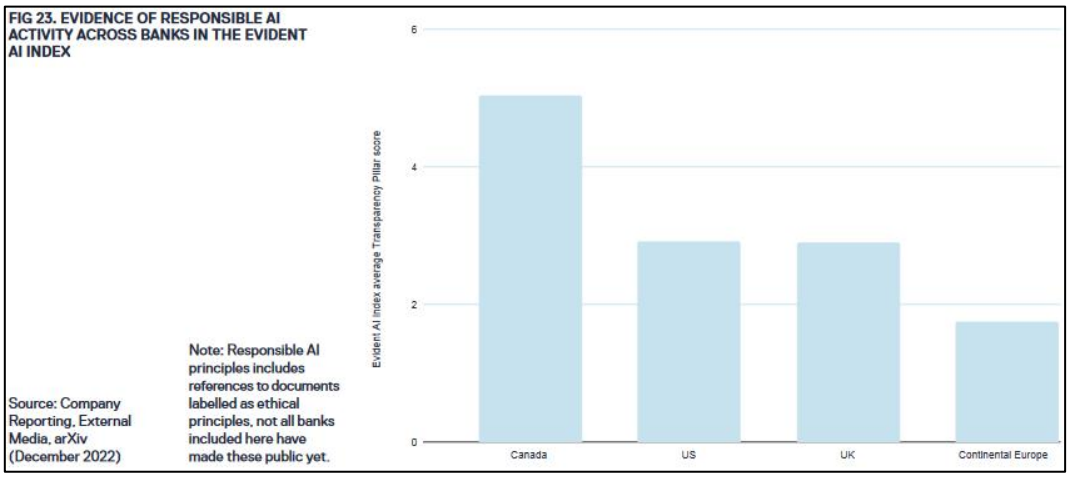


Figure 138 : Nombre de travaux sur l'IA responsable par région géographique - Source : evidentinsights.com

L'Europe a le moins de rapport publique sur cet axe par rapport à d'autres régions et plusieurs raisons pourraient expliquer cela :

- **Cadre réglementaire strict** : L'Europe est connue pour son cadre réglementaire strict en matière de protection des données, notamment avec le Règlement Général sur la Protection des Données (RGPD). Les entreprises pourraient être réticentes à publier des informations qui pourraient potentiellement les exposer à des risques juridiques.
- **Culture de la discrétion** : Comparées aux entreprises américaines, les entreprises européennes ont souvent été beaucoup plus secrètes sur leurs opérations internes, notamment lorsque c'est en lien avec la technologie et à l'innovation. Il s'agit d'une tendance de longue date tout au long de l'histoire.

- **Maturité du marché de l'IA** : Bien que l'Europe soit un acteur majeur dans le domaine de la recherche en IA, elle pourrait être en retard en termes d'adoption commerciale de l'IA par rapport à des régions comme l'Amérique du Nord ou l'Asie. Cela pourrait se traduire par moins de rapports sur l'IA responsable.
- **Priorités différentes** : Les entreprises européennes pourraient se concentrer davantage sur d'autres aspects de l'IA, tels que la recherche fondamentale, plutôt que sur la publication de rapports sur l'IA responsable.
- **Manque de normes uniformes** : Il est possible que l'absence de normes ou de directives claires sur ce qui devrait être inclus dans un rapport sur l'IA responsable rende les entreprises hésitantes à publier de telles informations.
- **Concurrence et propriété intellectuelle** : Les entreprises pourraient craindre que la divulgation de leurs pratiques en matière d'IA responsable ne donne un avantage à leurs concurrents ou n'expose leurs propriétés intellectuelles.
- **Initiatives à l'échelle de l'UE** : L'Union européenne elle-même travaille sur des initiatives liées à l'IA responsable, et certaines entreprises pourraient attendre des directives plus claires de l'UE avant de publier leurs propres rapports.

Il s'agit d'un classement établi par la « face visible de l'iceberg » et Evident ne pouvait faire autrement.

Par conséquent, le classement établi pourrait ne pas refléter les efforts réels de chaque banque, notamment si les travaux menées par celle-ci autour de l'IA responsable ne sont pas publiés publiquement.

En effet, bien que le nombre de rapports publiés puisse être inférieur, cela ne signifie pas nécessairement que les banques européennes ne s'engagent pas activement dans des pratiques d'IA responsable. Elles pourraient simplement choisir de ne pas les divulguer publiquement ou de le faire d'une manière différente, probablement pour des aspects de compétitivité.

Mais il s'agit d'une stratégie à « double tranchant ». Le revers de la médaille, presque paradoxal, est qu'être réticent à rendre publique les efforts réalisées sur l'IA responsable pourrait potentiellement avoir l'effet inverse, c'est à dire de possiblement être moins compétitive.

En effet, une recommandation sur ce sujet serait : **il devient de plus en plus essentiel que les entreprises, les banques particulièrement, communiquent clairement les mesures qu'elles prennent pour mettre en place une IA responsable, afin d'établir une confiance avec un large éventail de parties prenantes**, qu'ils soient les investisseurs, les actionnaires, les employés ou même les clients et consommateurs.

L'impact ne peut en être que positif notamment en termes de compétitivité. De plus, il est possible de publier des éléments sur l'IA responsable sans forcément livrer de secret stratégique. Un bon exemple à suivre sur ce sujet serait le Canada.

> *The most fundamental thing wrong with AI is that we make it a mystery.*

Cathy BESSANT - Vice Chair, Global Strategy at Bank of America, Décembre 2020

Une traduction pourrait être : « *Le problème le plus fondamental avec l'IA, c'est que nous en faisons un mystère* ».

Recommandation : communiquer sur vos travaux IA responsable au maximum du possible pour montrer de la proactivité ne pourra être que bénéfique, notamment par rapport à tous vos partenaires !

3.2.2 Evaluer sa maturité en termes d'IA responsable de ses cas d'usages

3.2.2.1 Faire une auto-évaluation : prendre conscience de sa maturité

Mettre en place une IA éthique et responsable est essentiel car elle garantit le respect des utilisateurs et renforce la position de l'entreprise comme un acteur fiable sur le marché européen.

Néanmoins, s'il y a bien une prise de conscience forte autour de la notion d'IA responsable et de ses enjeux, elle est récente. Or, de nombreux systèmes d'IA sont en cours de développement ou déjà mis en place dans divers secteurs, notamment le secteur bancaire.

Pour identifier les éventuels domaines d'amélioration, il est important de prendre une photo de la maturité à un instant T d'un système d'IA en termes d'adoption responsable et éthique.

Il est donc proposé ici différentes grilles pour déterminer l'application de bonnes pratiques relatives à chaque principe de l'AI Act, le cadre réglementaire à venir au niveau de l'Europe, mais également **le niveau de compréhension du besoin métier**. Elles ont été établies sur la base de nombreux modèles (voir plus loin).

Le but est ainsi de se comparer et comprendre la maturité du système d'IA afin de se rendre compte du chemin encore à parcourir éventuellement pour être en conformité, c'est-à-dire toutes les démarches pour appliquer les principaux fondamentaux de cette future réglementation.

3.2.2.2 Prérequis : évaluer la compréhension du besoin métier et le niveau de risque du système d'IA

Il est plus que nécessaire que **le besoin soit clairement expliqué et compris par toutes les parties prenantes du projet** et tous les interlocuteurs associés.

A l'issue de cette évaluation, **le niveau de risque du système d'IA est estimé**.

Voici la grille d'évaluation proposée pour la partie « *Besoin métier & risques* » :

Attribut	Description	Ce point concerne mon projet ?	Si oui, pris en compte ?
Besoin métier	Le système prend une décision de manière autonome ou assiste dans la prise de décision concernant des individus.		
Besoin métier	Le système exécute des tâches automatiquement pour l'utilisateur.		
Besoin métier	Le système est destiné à une mise en œuvre à grande échelle, que ce soit en interne ou pour le grand public.		
Besoin métier	Le système est destiné à être lancé sur un marché inédit.		
Besoin métier	Le système communique directement avec l'utilisateur final.		
Risque inacceptable	Le système pourrait permettre la manipulation mentale d'une personne avec des effets sur son comportement ou pourrait influencer des comportements en utilisant des méthodes subliminales qui agissent sur le subconscient.		N/A
Risque inacceptable	Le système pourrait exploiter les vulnérabilités des personnes ou viser des groupes sensibles tels que les enfants ou les personnes en situation de handicap.		N/A
Risque inacceptable	Le système pourrait évaluer ou établir la notation de personnes en fonction de leur comportement social ou de leur caractéristiques personnelles ou catégoriser les individus en fonction de leur fiabilité et entraîner un traitement inéquitable ou discriminatoire.		N/A
Risque inacceptable	Le système pourrait permettre l'identification biométrique, en temps réel et à distance, dans les lieux publics à des fins répressives.		N/A
Risque inacceptable	Le système pourrait engendrer une menace claire pour la sécurité, les moyens de subsistance et les droits des personnes.		N/A
Risque élevé	Le système est utilisé dans les infrastructures critiques ou domaines vitaux comme les transports, où la sécurité des citoyens pourrait être compromise.		N/A
Risque élevé	Le système est utilisé dans l'éducation et la formation professionnelle, dans le cadre d'une utilisation influençant l'accès à l'éducation et la trajectoire professionnelle d'un individu (par exemple, la correction automatisée des examens).		N/A
Risque élevé	Le système est utilisé dans les composants de sécurité des produits, comme l'IA dans les interventions chirurgicales robotisées.		N/A
Risque élevé	Le système est utilisé dans le recrutement, la gestion des employés et l'accès à des postes indépendants (comme les outils automatisés de sélection de CV).		N/A
Risque élevé	Le système est utilisé dans des services essentiels, privés ou publics (tels que les systèmes d'évaluation de crédit qui peuvent refuser un prêt à un individu).		N/A
Risque élevé	Le système est utilisé dans les services d'ordre public (répressifs) qui pourraient restreindre les droits fondamentaux (comme l'évaluation de la solidité des preuves).		N/A
Risque élevé	Le système est utilisé dans la gestion des migrations, des demandes d'asile et des contrôles frontaliers (comme la vérification des documents de voyage).		N/A
Risque élevé	Le système est utilisé dans les procédures judiciaires et les processus démocratiques (par exemple, l'application automatisée des lois à des situations spécifiques).		N/A
Risque limité	Le système est susceptible de détecter les émotions ou de faire des trucages vidéos ou les chatbots.		N/A
Risque minimal	Le système n'est pas un système à risque inacceptable ou élevé ou limité.		N/A

3.2.2.3 Evaluer sa maturité sur les principes de l'IA Act

3.2.2.3.1 Principe 1 : Action et Contrôle humain

L'IA doit être au service de l'humain, respecter les droits fondamentaux et préserver l'autonomie humaine.

Elle doit intégrer des mécanismes de contrôle et d'intervention des individus.

Voici la grille d'évaluation pour la partie « *Action et Contrôle humain* » :

Attribut	Description de la bonne pratique	Ce point concerne mon projet ?	Si oui, pris en compte ?
Conception & décision par l'humain	Légaliser le projet (par exemple, s'assurer de la conformité avec les réglementations sur la vie privée et les droits essentiels).		
Conception & décision par l'humain	Établir la latitude "optimale" de l'intervention humaine au sein du système d'IA en fixant des principes éthiques, puis les communiquer aux utilisateurs.		
Conception & décision par l'humain	Instaurer un mécanisme de repérage, de notification et de rectification des anomalies (violation des droits essentiels, situations dangereuses).		
Conception & décision par l'humain	Accompagner l'usager dans la gestion des rectifications, si une ou plusieurs anomalies sont identifiées.		
Conception & décision par l'humain	Élaborer une procédure d'intervention manuelle en cas de dysfonctionnement du système d'IA.		
Intervention de l'humain	Analyser les conséquences de la mise en service et de l'utilisation du système d'IA.		
Intervention de l'humain	Repérer les moments nécessitant une intervention humaine.		
Intervention de l'humain	Confirmer les interventions humaines aux moments clés pour prévenir les préjugés (en conservant un historique).		
Intervention de l'humain	Assurer une traçabilité des actions du système d'IA.		
Intervention de l'humain	Établir un protocole de tests et de validation pour confirmer le degré d'indépendance souhaité.		
Intervention de l'humain	Instaurer un mécanisme de feedback sur le comportement du système d'IA afin d'informer l'utilisateur final et d'ajuster le modèle si nécessaire. Par exemple : offrir la possibilité aux utilisateurs d'intervenir lorsqu'une décision est prise par le système ou de signaler des dysfonctionnements potentiels.		
Supervision de l'humain	Définir le cadre d'utilisation du système d'IA.		
Supervision de l'humain	Évaluer le système d'IA en fonction des droits fondamentaux		
Supervision de l'humain	Assurer la traçabilité du système d'IA grâce à des méthodes de conception, de test et de suivi validées par des normes rigoureuses.		
Supervision de l'humain	Tester le système d'IA pour comprendre son fonctionnement et ses conséquences. Par exemple : vérifier que l'algorithme n'effectue pas de décisions automatiques ayant des implications légales.		
Supervision de l'humain	Veiller à ce que l'humain reste impliqué : mettre en place des résumés et des accès d'urgence au système d'IA pour effectuer des corrections, surveiller et fournir des retours, notamment aux utilisateurs.		
Acceptation de l'humain	Créer un manuel d'évaluation du système d'IA destiné à l'utilisateur final (client du service).		
Acceptation de l'humain	Lancer une campagne de sensibilisation concernant le manuel d'auto-évaluation.		
Acceptation de l'humain	Établir les modalités d'accès à ce manuel d'auto-évaluation.		
Acceptation de l'humain	Fournir un moyen pour l'utilisateur de contester les actions du système d'IA.		
Acceptation de l'humain	Instaurer une notification claire, compréhensible et modifiable pour l'utilisateur final.		

3.2.2.3.2 Principe 2 : Robustesse technique et sécurité

Les algorithmes doivent être **conçus pour être robustes aux attaques, sécurisés et fiables**.

Ils doivent également **gérer efficacement les erreurs et incohérences** à chaque étape de leur cycle de vie.

Voici la grille d'évaluation pour la partie « *Robustesse technique et sécurité* » :

Attribut	Description de la bonne pratique	Ce point concerne mon projet ?	Si oui, pris en compte ?
Fonctionnement	Indiquer les métriques pour évaluer la reboustesse technique d'une IA avant la conception du système ?		
Fonctionnement	Avoir une checklist de robustesse, dans l'environnement de test avant la mise en production.		
Fonctionnement	Laisser le système d'IA détecter lui-même ses limites de fonctionnement et alerter quand ces dernières sont atteintes.		
Fonctionnement	Établir un niveau acceptable d'erreur pour un système informatique.		
Fonctionnement	Assurer la constance du système lors de son fonctionnement habituel.		
Fonctionnement	Traiter les anomalies grâce à des méthodes appropriées partagées en avance.		
Résultats	Établir les résultats escomptés, les critères ou les mesures utilisées pour évaluer la conformité des résultats.		
Résultats	Veiller à ce que les critères soient basés sur des méthodes éprouvées et supervisées : il est essentiel de garantir l'intégrité des données avec une annotation réalisée par des humains, évaluée par plusieurs juges, tout en veillant à l'impartialité et à l'objectivité des intervenants.		
Résultats	Déterminer des marges d'erreur pour chaque utilisation/application.		
Résultats	Afficher ces marges d'erreur pour les utilisateurs.		
Résultats	Utiliser des méthodes statistiques pour élaborer ces techniques/outils.		
Protection	Déterminer les domaines où la précision des résultats est essentielle, en particulier ceux avec des implications majeures en matière de sécurité (comme les centrales nucléaires, l'aviation).		
Protection	Faciliter la transition vers un mode "alternatif" où le contrôle est rendu à l'opérateur humain.		
Protection	Citer des cas où le système d'IA, intégré matériellement, a un impact physique direct (par exemple : robots, voitures autonomes) ou lorsqu'il est intégré à un système décisionnel (comme une application de microcrédit influençant les conditions de vie d'une personne).		
Protection	Mettre en évidence des exemples d'effets physiques indirects, tels que la détection de situations potentiellement dangereuses pour l'utilisateur.		
Protection	Recenser et quantifier le pourcentage d'utilisations prévisibles du système d'IA, qu'elles soient bénéfiques ou nuisibles, et définir le seuil critique.		
Protection	Apprécier et juger le système d'IA en fonction des droits essentiels.		

3.2.2.3.3 Principe 3 : Respect de la vie privée et gouvernance des données

Les **données personnelles doivent être protégées**, conformément aux réglementations comme le **RGPD**.

Les individus doivent avoir le contrôle total de leurs données et cela afin de les protéger de toute utilisation malveillante ou discriminatoire qui pourrait en être faite.

Voici la grille d'évaluation pour la partie « *Respect de la vie privée et gouvernance des données* » :

Attribut	Description de la bonne pratique	Ce point concerne mon projet ?	Si oui, pris en compte ?
Anonymisation	Évaluer si l'anonymisation est réellement la meilleure méthode pour garantir la protection de la vie privée dès la conception et par défaut, en fonction de l'utilisation prévue des données. D'autres technologies favorisant la confidentialité pourraient être plus appropriées.		
Anonymisation	Déterminer les informations à conserver en fonction de leur pertinence.		
Anonymisation	Éliminer les identifiants directs ainsi que les données rares qui pourraient faciliter la ré-identification des individus (par exemple, l'âge peut facilement permettre de reconnaître les centenaires).		
Anonymisation	Différencier les informations essentielles des données secondaires ou superflues, qui peuvent être supprimées.		
Anonymisation	Établir le niveau de détail optimal et acceptable pour chaque information retenue.		
Anonymisation	Assurer une surveillance constante pour maintenir l'anonymat des données au fil du temps, en tenant compte des outils techniques actuels et des autres sources de données susceptibles de compromettre l'anonymat.		
Anonymisation	Mettre en place un cadre légal pour réguler l'utilisation des données traitées par le système d'IA, même si elles sont anonymes.		
Traitement des données selon la RGPD	Évaluer la nature des traitements de données personnelles effectués par le système d'IA (à haut risque ou non) et, si nécessaire, procédez à une évaluation d'impact.		
Traitement des données selon la RGPD	Garantir la conformité des données personnelles traitées par le système d'IA aux normes légales et réglementaires, en collaboration avec un délégué à la protection des données au sein de l'organisation.		
Traitement des données selon la RGPD	Documenter cette conformité et tenez à jour un registre des opérations de traitement des données personnelles.		
Traitement des données selon la RGPD	Présenter de manière transparente, compréhensible et accessible les modalités légales et réglementaires de collecte, de stockage, de traitement et de partage des données aux individus concernés.		
Traitement des données selon la RGPD	Opter pour le consentement comme base légale lorsque le système d'IA collecte des données personnelles de ses utilisateurs.		
Traitement des données selon la RGPD	Obtener ce consentement de manière libre, spécifique, informée et univoque. Par exemple, évitez les cases pré-cochées.		
Traitement des données selon la RGPD	Fixer une durée de conservation des données cohérente avec les objectifs du traitement.		
Traitement des données selon la RGPD	Ne pas utiliser les données personnelles du système d'IA d'une manière préjudiciable aux personnes concernées.		
Traitement des données selon la RGPD	Établir des protocoles d'accès aux données personnelles du système d'IA, définissant qui peut y accéder et dans quel contexte.		
Traitement des données selon la RGPD	Offrir aux individus la possibilité de révoquer leur consentement aussi facilement qu'ils l'ont donné, à tout moment.		
Interaction de l'humain avec le système d'IA	Prévoir le droit pour les utilisateurs, lorsqu'ils sont face à un système d'IA impliqué dans un processus décisionnel, de solliciter l'intervention d'un être humain.		
Interaction de l'humain avec le système d'IA	Clarifier pour l'utilisateur ou le bénéficiaire du système d'IA qu'il est en interaction avec un système d'IA, en particulier dans les situations telles que : une communication directe avec un système d'IA, l'emploi d'un système d'IA pour détecter des émotions, l'utilisation d'un système d'IA pour classer des personnes selon des critères biométriques, ou l'usage d'un système d'IA produisant ou modifiant du contenu audiovisuel.		

3.2.2.3.4 Principe 4 : Transparence

Les systèmes d'IA doivent être **traçables**.

Ils doivent offrir une **explicabilité claire de leurs prédictions**, en tenant compte des contraintes professionnelles et réglementaires.

Voici la grille d'évaluation pour la partie « *Transparence* » :

Attribut	Description de la bonne pratique	Ce point concerne mon projet ?	Si oui, pris en compte ?
Explicabilité	**Concernant la conception d'un système d'IA** : - Opter pour l'algorithme le plus simple, basé si possible sur des directives. - Consigner les techniques utilisées, l'utilisation des données et les modifications majeures : - La manière dont les données sont structurées, - Le modèle sélectionné et la méthode suivie, - Les hyperparamètres (c'est-à-dire définis par le programmeur) et leur méthode de sélection, - Comment les paramètres (issus de la formation du modèle) influencent les performances, - Les conditions de fiabilité des résultats, etc. et rédiger un guide récapitulatif.		
Explicabilité	**Quand les résultats de l'IA nécessitent une simplification :** - Introduire l'outil d'explication sélectionné, ses avantages et ses contraintes pour le scénario abordé, - Montrer à l'utilisateur les contraintes techniques du modèle et de l'outil d'explication, en illustrant avec des cas concrets.		
Explicabilité	**Quand le modèle évolue avec l'environnement** : - Fournir des illustrations de variations de prédictions si le programme évolue continuellement, - Informer l'utilisateur qu'il interagit avec une intelligence artificielle lors d'une interaction.		
Justification	Mettre en parallèle les actions de l'IA avec le processus de réflexion d'une personne dans une situation similaire.		
Justification	Fournir et détailler des exemples montrant les limites de cette comparaison entre humain et IA (par exemple : apprentissage automatique face à un raisonnement logique).		
Justification	Documenter les décisions éthiques guidant l'IA (par exemple : priorité à la confidentialité, anonymisation des cas cités, les individus mentionnés dans l'explication ne correspondent pas à de vraies personnes).		
Compréhension du système d'IA par l'humain	Élaborer des schémas simplifiées adaptées aux différents profils pour expliquer l'IA (scénarios typiques).		
Compréhension du système d'IA par l'humain	Disposer de divers moyens de communication (supports) : interface, démonstration, terminologie adaptée pour chaque équipe (initiation technique pour les équipes métier et vice versa).		
Compréhension du système d'IA par l'humain	Vérifier si le support fourni est considéré comme clair et informatif par le public cible.		

3.2.2.3.5 Principe 5 : Diversité, non-discrimination et équité

Les systèmes d'IA doivent être **conçus pour être justes, sans biais et accessibles à tous**.

Ils doivent refléter une **variété de capacités, compétences et besoins humains**.

Voici la grille d'évaluation pour la partie « *Diversité, non-discrimination et équité* » :

Attribut	Description de la bonne pratique	Ce point concerne mon projet ?	Si oui, pris en compte ?
Gestion des biais	Déterminer quel biais l'entreprise vise à rectifier : individuel, entre différents groupes ou sous-groupes.		
Gestion des biais	Sélectionner une métrique d'équité adaptée au biais perçu comme inéquitable, en fonction de la réglementation ou de la culture organisationnelle.		
Gestion des biais	Décider à quel moment rendre l'IA plus équilibrée : lors de la collecte des données (si modification possible), pendant la phase d'entraînement (pour une optimisation basée sur la métrique d'équité), après l'entraînement (par exemple, si les données ou le modèle ne peuvent être modifiés), ou opter pour une approche combinée si faisable.		
Gestion des biais	Choisisser et expliquer la méthode de correction équitable à utiliser (par exemple, une technique de pré-traitement, de traitement en cours, de post-traitement ou une combinaison de celles-ci).		
Gestion des biais	Surveiller régulièrement les données pour évaluer l'effet du système d'IA sur les groupes à protéger.		
Diversité de l'équipe	Veiller à ce que le processus de recrutement reflète la diversité des équipes en termes de profils, d'origines, de parcours académiques, de perspectives et d'âge.		
Diversité de l'équipe	Impliquer des experts en éthique, qu'ils soient internes ou externes, pour sensibiliser les data-scientists aux enjeux éthiques et juridiques du projet.		
Diversité de l'équipe	Effectuer et publier des auto-évaluations périodiques (par exemple, tous les six mois) pour vérifier la conformité des travaux des data-scientists avec la réglementation et les principes de l'entreprise (comme le formulaire de la CNIL).		
Diversité de l'équipe	S'assurer que le guide d'auto-évaluation examine les décisions prises, détaille la méthode de collecte des données, analyse les corrélations et les biais, et décrit les actions entreprises face aux biais jugés inéquitables.		
Diversité de l'équipe	Utiliser le guide d'autoévaluation de la CNIL pour identifier les risques de biais, notamment de discrimination, et les corriger efficacement.		
Accessibilité au système d'IA	Établisser et annoncer les étapes du système d'IA qui sont en accès libre, sous la supervision des responsables de la stratégie et de la régulation.		
Accessibilité au système d'IA	Préciser ce qui est accessible gratuitement ou moyennant des frais.		
Accessibilité au système d'IA	Protéger les sections du code nécessitant une sécurité renforcée.		

3.2.2.3.6 Principe 6 : Bien-être sociétal et environnemental

L'IA **doit contribuer positivement à la société et promouvoir la durabilité environnementale**.

Elle doit **aligner les intérêts commerciaux avec les préoccupations sociétales et environnementales**.

Voici la grille d'évaluation pour la partie « *Bien-être sociétal et environnemental* » :

Attribut	Description de la bonne pratique	Ce point concerne mon projet ?	Si oui, pris en compte ?
Harmonie écologique	Assurez-vous que le système d'IA contribue positivement à l'environnement.		
Harmonie écologique	Intégrer les préoccupations écologiques dans la sélection des fournisseurs, collaborateurs et partenaires liés au système d'IA.		
Harmonie écologique	Évaluer l'empreinte carbone, notamment pendant la phase d'apprentissage des modèles, en utilisant des outils tels que Carbon tracker		
Harmonie écologique	Optimiser la consommation énergétique et ressources du système d'IA tout au long de son existence, tout en simplifiant l'architecture nécessaire à son fonctionnement.		
Harmonie écologique	Anticiper la fin d'utilisation du système d'IA, en considérant notamment le devenir des données et des éléments logiciels et matériels.		
Harmonie écologique	Se focaliser sur les fonctionnalités primordiales du système d'IA et éliminer celles inutilisées au fil du temps.		
Harmonie écologique	Réduire la quantité de données collectées pendant la vie du système d'IA et instaurer une gestion rigoureuse de l'archivage et de la suppression des données.		
Harmonie écologique	Ajuster constamment la quantité d'information affichée en fonction des besoins pour économiser de l'énergie.		
Harmonie écologique	Rester informé des méthodes d'écoconception dans le domaine numérique.		
Harmonie écologique	Partager les connaissances sur l'impact écologique du numérique avec les équipes et les utilisateurs.		
Contribution socio-économique	Établir dès les premières étapes de création du système d'IA les principes socio-économiques à suivre durant son développement et son usage.		
Contribution socio-économique	Effectuer une analyse coûts-avantages de l'impact socio-économique du système d'IA pour garantir qu'il soit en adéquation avec les principes établis.		
Contribution socio-économique	Veiller à ce que le système d'IA ne puisse pas être utilisé légalement de manière nuisible pour les individus, que ce soit en tant qu'employé ou citoyen.		
Contribution socio-économique	Associer les (potentiels) utilisateurs du système d'IA à son élaboration et/ou à son suivi.		
Contribution socio-économique	Examiner la facilité d'utilisation de l'interface du système d'IA et la perfectionner régulièrement.		
Contribution socio-économique	Élaborer une approche pour assurer la compatibilité du système d'IA avec les équipements et logiciels désuets.		
Effet du système d'IA sur la santé mentale et physique	Intégrer les fondements de la bioéthique tels que la bienveillance, la non-nuisance, l'équité et l'indépendance dès la phase initiale de création du système d'IA		
Effet du système d'IA sur la santé mentale et physique	Valoriser les applications du système d'IA qui promeuvent le bien-être physique, mental et social, qui diminuent les disparités d'accès aux soins, ou qui contribuent à détecter, évaluer, surveiller, prévenir ou traiter les enjeux de santé actuels ou futurs (comme les virus, maladies longue durée, perte d'indépendance, vieillissement...).		
Effet du système d'IA sur la santé mentale et physique	Associer les usagers ou les destinataires finaux à l'élaboration du système d'IA.		
Effet du système d'IA sur la santé mentale et physique	Veiller à ce que les informations personnelles gérées par le système d'IA soient recueillies conformément aux législations et normes en vigueur (en particulier si ces informations proviennent de tiers)		
Effet du système d'IA sur la santé mentale et physique	Garantir que les données utilisées pour la formation du système d'IA reflètent fidèlement la population visée (tout en considérant la nécessité de minimiser les données personnelles)		
Effet du système d'IA sur la santé mentale et physique	Établir des procédures pour rectifier les préjugés présents dans les données de formation et les défaillances du système d'IA (comme l'apprentissage excessif ou insuffisant)		
Effet du système d'IA sur la santé mentale et physique	Offrir une clarification sur les conclusions du système d'IA à ses usagers.		
Effet du système d'IA sur la santé mentale et physique	Conserver les informations médicales traitées par le système d'IA dans des infrastructures certifiées pour l'hébergement de données de santé		
Effet du système d'IA sur la santé mentale et physique	Garantir une surveillance humaine des opérations du système d'IA et de ses conclusions.		

3.2.2.3.7 Principe 7 : Responsabilité

Les systèmes d'IA doivent être **auditables et minimiser les impacts négatifs**.

Des mécanismes de responsabilité doivent être mis en place, garantissant une obligation de transparence et de rendre des comptes.

Voici la grille d'évaluation pour la partie « *Responsabilité* » :

Attribut	Description de la bonne pratique	Ce point concerne mon projet ?	Si oui, pris en compte ?
Responsabilité des conséquences	**Avant la mise en pratique d'un scénario d'utilisation :** Établir une méthode pour déterminer les responsabilités à chaque étape de la conception et de la mise en œuvre d'un système d'IA.		
Responsabilité des conséquences	**Pour chaque scénario d'application :** Désigner un "model owner" qui sera en charge du modèle en production et servira d'interlocuteur unique pour tous les acteurs humains participant au processus de développement.		
Responsabilité des conséquences	**Pour chaque scénario d'application :** Identifier et classer le plus exhaustivement possible les types d'incidents inattendus envisageables, c'est-à-dire les situations qui ne seraient pas conformes aux réglementations actuelles ou aux principes éthiques établis		
Responsabilité des conséquences	**Pour chaque scénario d'application :** Préciser, potentiellement par voie contractuelle, les raisons pour lesquelles la responsabilité de chaque intervenant pourrait être engagée, que ce soit le concepteur, le développeur, le fournisseur, l'hébergeur, d'autres sous-traitants ou l'utilisateur final.		
Surveillance et auditabilité	**Avant le début du développement et tout au long du processus de création :** Mettre en place une méthode d'audit interne et, idéalement, externe.		
Surveillance et auditabilité	**Avant le début du développement et tout au long du processus de création :** Faciliter la réalisation des audits externes en documentant : - Le besoin auquel le système d'IA répond ; - La méthode de collecte des données d'apprentissage et leur pertinence ; - La sélection du modèle et les risques liés ; - Les stratégies de cybersécurité adoptées ; - Les modifications apportées au modèle après le déploiement du système d'IA.		
Surveillance et auditabilité	**Avant le début du développement et tout au long du processus de création :** Protéger la propriété intellectuelle et les informations confidentielles tout en préservant la transparence.		

Une fois ces grilles remplies, elles permettent via le tableau suivant de se faire une idée de l'implémentation des bonnes pratiques qui concernent le système d'IA en cours d'auto-évaluation. L'indicateur le plus parlant est celui consistant à avoir pour chaque des 7 principes de l'AI Act **un ratio entre le nombre de bonnes pratiques déjà prises en compte ou implémentées et le nombre de bonnes pratiques applicables au projet de système d'IA**.

Voici le tableau avec l'évaluation obtenue sur un cas fictif :

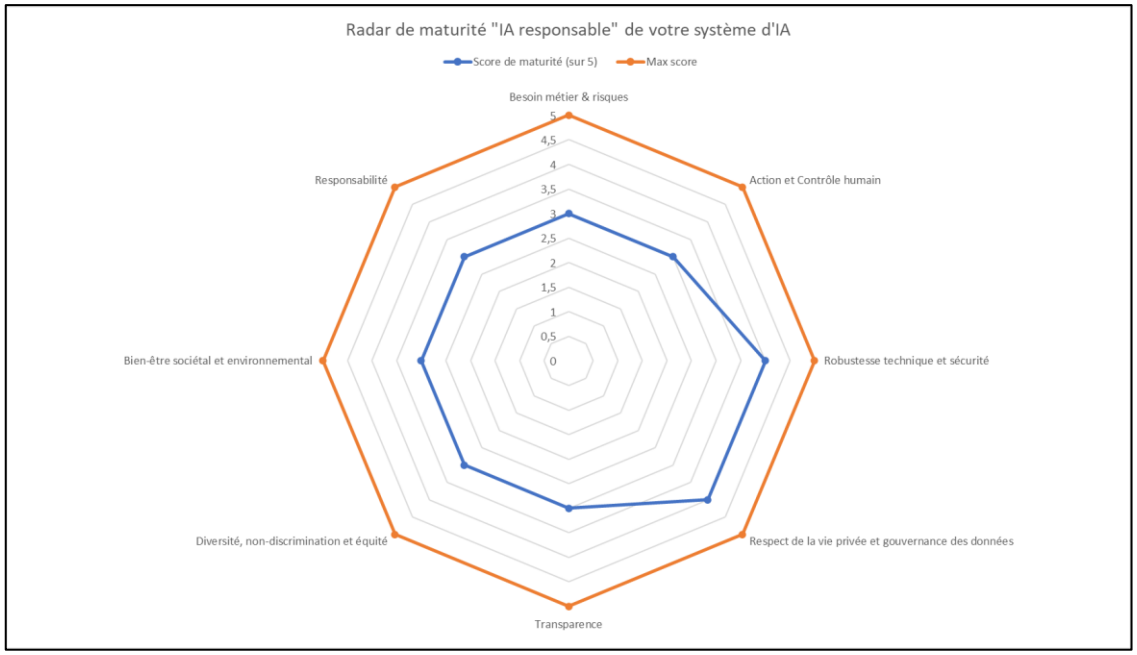
Niveau risque	Niveau 3 - Système d'IA à risque élevé

Onglet	Principe	Nb pratiques applicables	Nb pratiques appliquées	% de maturité	Score de maturité (sur 5)	Max score
0	Besoin métier & risques	2	1	50%	3	5
1	Action et Contrôle humain	21	10	48%	3	5
2	Robustesse technique et sécurité	15	9	60%	4	5
3	Respect de la vie privée et gouvernance des données	19	14	74%	4	5
4	Transparence	9	4	44%	3	5
5	Diversité, non-discrimination et équité	12	5	42%	3	5
6	Bien-être sociétal et environnemental	17	8	47%	3	5
7	Responsabilité	6	3	50%	3	5

Le score de maturité résulte d'un calcul du ratio suivant :

$$\frac{\text{Nombre de pratiques appliqués au projet}}{\text{Nombre de pratiques applicables au projet}}$$

Le radar suivant peut alors être obtenu (toujours associé à l'exemple fictif pris précédemment) :



Ces éléments permettent de se situer sur une **grille finale de maturité** :

Principe	Score 1 - non maitrisé	Score 2 - initial	Score 3 - défini	Score 4 - maitrisé	Score 5 - optimisé
Action et Contrôle humain	Les actions et le contrôle humain n'est pas intégré au système d'IA	Le degré de délégation est seulement examiné dans certaines situations d'utilisation, en fonction du contexte de mise en œuvre et des missions accomplies par le système d'IA.	Des protocoles existent pour définir le contexte de l'IA, le degré de délégation et la place de l'humain. Cependant, ils ne sont pas systématiquement mis en œuvre à chaque étape du cycle de vie du système d'IA.	Des protocoles définissant le contexte de l'IA, le niveau de délégation, la collecte du consentement de l'utilisateur et la place de l'humain sont mis en œuvre et intégrés dans l'ensemble des projets liés à l'IA. Ces directives garantissent que l'humain puisse reprendre la main sur l'IA si la situation l'exige.	La place centrale de l'humain est ancrée dans la stratégie de l'entreprise. Au-delà des contrôles adaptés au contexte, son rôle est intrinsèquement lié à la conception de la stratégie IA de l'entreprise, en harmonie avec ses valeurs et missions fondamentales.
Robustesse technique et sécurité	L'identification des vulnérabilités est négligée ou sous-estimée.	La détection des vulnérabilités se limite aux scénarios d'application.	Les vulnérabilités sont identifiées indépendamment des scénarios d'utilisation, avec l'établissement de normes pour assurer la robustesse, la fiabilité des résultats et une protection matérielle contre les aléas.	Une démarche structurée de sécurité IA est mise en place à l'échelle de l'entreprise, permettant de gérer tous les risques en fonction du cycle de vie de l'IA.	Outre la gestion des risques de sécurité, la prévision des changements majeurs de l'environnement est un élément clé de la stratégie de sécurité. Elle est intégrée dans l'ensemble des processus et niveaux relatifs à l'IA et aux données au sein de l'entreprise.
Respect de la vie privée et gouvernance des données	La vie privée et la protection des données n'ont pas été suffisamment prises en compte.	Uniquement les vérifications imposées par la loi sont mises en œuvre pour chaque situation abordée.	Chaque phase du cycle de vie des données et du système d'IA est conforme aux exigences réglementaires, y compris celles qui vont au-delà de la protection de la vie privée.	Établissement d'une gouvernance pour intégrer des vérifications, des procédures et des comités opérationnels dans la méthodologie de projet.	Établir un écosystème dynamique axé sur la protection des utilisateurs grâce aux données et au système d'IA. L'entreprise dépasse ses obligations en matière de réglementation, de gouvernance et de révision, tout en collaborant à l'extérieur pour faire reconnaître ses principes directeurs.
Transparence	Les principes de transparence, d'explicabilité et d'interprétation ne sont pas suffisamment valorisés et considérés.	Pour certains cas d'utilisation, des modèles plus transparents sont présents et des techniques d'interprétabilité sont utilisées.	Des méthodes ont été élaborées pour clarifier la conception du modèle, sa performance, les résultats produits ainsi que les contraintes et décisions techniques. Cependant, la documentation et l'application de ces méthodes demeurent inégales.	Des normes rigoureuses sont constamment utilisées pour détailler la conception du modèle, sa performance, les résultats atteints ainsi que les contraintes techniques et les décisions techniques prises.	Les valeurs de transparence, d'explicabilité et de compréhension sont profondément ancrées dans les processus de création de l'IA et sont constamment réévaluées pour prévenir tout risque, en particulier d'ordre éthique.
Diversité, non-discrimination et équité	L'interaction entre l'homme et la machine n'est pas correctement abordée, et il n'y a aucune prise en compte des biais potentiels.	Des méthodes pour identifier et rectifier les biais sont mises en œuvre sur certains modèles.	Quand l'IA conduit à des résultats défavorables pour l'utilisateur, une intervention humaine est possible pour examiner les biais, et la décision doit être validée par cette personne.	Les techniques de repérage des biais sont accessibles et employées de manière systématique pour chaque modèle d'IA.	L'emploi systématique de techniques de détection de biais, la collaboration avec des équipes variées et la mise à jour fréquente des modèles sont de mise.
Bien-être sociétal et environnemental	L'entreprise ne prend pas en compte les défis liés à l'impact durable.	Une première définition des critères d'évaluation a été établie. Une prise de conscience liée au bien-être sociétal et environnemental a été promue.	Mise en œuvre, au niveau des algorithmes, de méthodes visant à considérer et optimiser les impacts.	Intégrer les considérations environnementales et sociétales dès le début de la création du système d'IA.	Disposer d'un processus adaptable à tous les scénarios d'application de l'IA.
Responsabilité	L'entreprise n'a pas évalué la responsabilité humaine et n'a instauré aucune mesure de traçabilité.	Une méthode d'enregistrement et de gestion des incidents inattendus est en place.	Au-delà des obligations légales, une procédure est en place pour définir clairement les rôles et responsabilités éthiques de chaque intervenant dans la chaîne de valeur.	Un mécanisme de gestion des risques, qu'ils soient potentiels ou avérés, est en place, englobant tous les intervenants de la chaîne de valeur, le cycle de vie de l'IA et les données exploitées. Ce système est régulièrement mis à jour et documenté pour permettre des audits périodiques.	La prise en compte proactive des responsabilités humaines est ancrée dans la culture de l'entreprise et s'étend à tout l'écosystème. Outre les procédures d'audit, la gouvernance et la gestion anticipée des risques, une culture forte de responsabilité humaine est promue, souvent à travers des programmes de formation continue, des forums de discussion, et autres initiatives.

Ainsi, ces scores par principe couplés aux différentes grilles donnent les éléments à implémenter ou à instruire dans une démarche d'amélioration pour obtenir le 100% « IA responsable », relatif au contexte du projet de système d'IA (i.e. selon principes applicables). Ils fournissent donc une structure pour guider un système d'IA vers une **IA responsable, éthique et digne de confiance**.

Recommandations :
La possibilité de s'auto-évaluer est un atout pour se démarquer face à la concurrence et attirer des partenaires. Cette auto-évaluation sert de référence pour démontrer le niveau d'engagement éthique et responsable de votre système d'IA.
De plus, elle sensibilise et forme les équipes internes à l'importance de l'IA responsable, tout en vous permettant de vous situer en termes de conformité par rapport à l'arrivée de la réglementation européenne sur l'IA, l'AI Act. Elle vous permettra donc d'estimer le "chemin à parcourir" pour être prêt le moment venu.

3.2.2.4 Quelques guides d'évaluation supplémentaires

Voici quelques questionnaires et outils additionnels pour évaluer de manière concrète l'impact responsable et éthique de votre système d'IA :

- Ethics & Algorithms Toolkit, conçu par un groupe américain, il se concentre sur l'utilisation de l'IA dans le domaine public,
- Assessment List for Trustworthy Artificial Intelligence (Altai), le questionnaire du Groupe d'experts indépendants de haut niveau sur l'IA constitué par la **Commission Européenne**, disponible également à partir d'ici,
- The Box de **l'AI Ethics Lab**, qui offre la possibilité d'évaluer, selon divers critères, les atouts et limites de son système d'un point de vue éthique (en complément, l'article « Operationalizing AI ethics principles »),
- Ethics Guidelines de la **MAIF**, incorporé au projet **Melusine**, projet de tri automatique des e-mails,
- Le référentiel d'évaluation de la maturité d'une organisation, de **Labelia Labs (ex- Substra Foundation),**
- AI Ethics Assessment Toolkit, un outil proposé par **Open Roboethics Institute,**
- Responsible AI Toolkit, un outil d'évaluation très complet proposé par PWC,
- Artificial intelligence impact assessment, guide complet réalisé par **ECP**, la plateforme néerlandaise indépendante,
- Guide pratique pour des IA éthiques, de **Numeum & Ethical AI,**
- Guide d'auto-évaluation pour les systèmes d'Intelligence Artificielle de la **CNIL**,
- Ethic Assessment Tools (ALTAI), un autre outil proposé par **AI-on-demand** et qui permet également de se situer par rapport à l'AI Act,
- L'IA éthique en pratique, proposé par Hub France IA.

3.2.2.5 Déterminer les « next steps » éventuels

Vous avez pu vous auto-évaluer. Vous savez donc maintenant où vous vous situez en termes de maturité « IA responsable ». Mais vous vous demandez certainement ce que vous devez faire prochainement, ce fameux « chemin à parcourir » pour atteindre le sacre d'avoir une IA responsable.

Pour rappel, avoir une IA responsable, c'est avoir une IA qui **répond à un ensemble d'exigences ou principes lesquels sont en cours de définition au niveau Européen**. Il n'est pas certain d'avoir tout couvert dans ces kits d'auto-évaluation et il ne se suffit pas de se proclamer « IA responsable » pour l'être véritablement.

Le bout du « chemin à parcourir » serait donc de montrer officiellement que le système d'IA est bien une IA responsable et pour cela, une certification est le meilleur moyen de le prouver.

Les prochaines actions dépendent surtout du degré de maturité global obtenu lors de l'auto-évaluation ainsi qu'aux éléments pointés dans les différentes grilles comme étant à prendre en compte ou à mettre en place par celle-ci.

Pour **Axionable**, le degré de maturité permet de se situer sur une courbe, laquelle permet de déterminer les prochaines étapes et les actions associés :

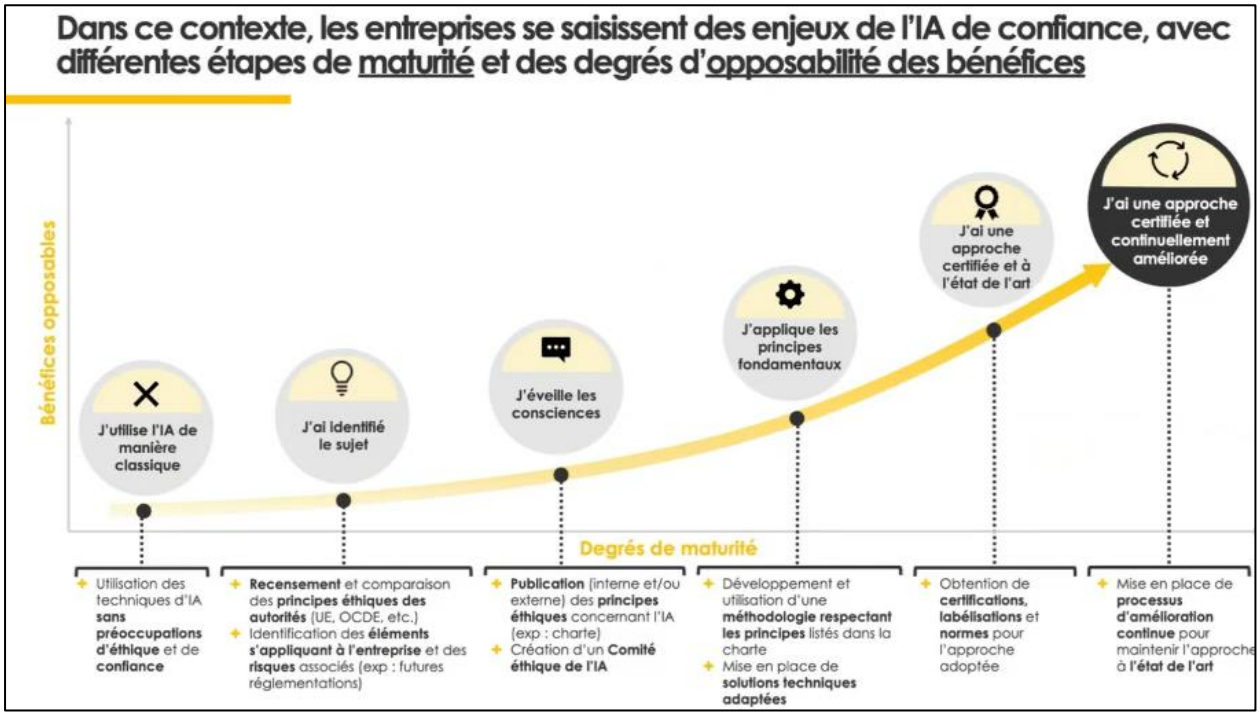


Figure 139 : Etapes de maturité d'un projet "IA responsable" - Source : Axionable - présentation chez Datacraft, octobre 2022

Recommandations :
Se référer à cette courbe qui est une bonne base pour se situer et connaitre les « next steps » et viser le plus loin possible, c'est-à-dire viser l'évaluation via des organismes externes reconnus pour obtenir une certification et/ou un label.

3.2.3 Se faire évaluer : les labels, certifications et normes

Votre organisation bancaire a pris l'initiative de conformer ses projets d'IA aux normes des futurs cadres réglementaires, tels que l'AI Act. Plus qu'une simple conformité, vous aspirez à valoriser et communiquer cette démarche auprès de vos parties prenantes.

Envisagez-vous d'obtenir une certification ou un label pour **mettre en avant vos efforts face à vos clients, collaborateurs et investisseurs** ?

Face à la multiplication des systèmes d'évaluation par des tiers (évaluation par un organisme externe), il devient difficile de s'y retrouver. Nombreux sont ceux qui confondent même les notions de label et de certification, d'autant plus que ces termes sont fréquemment employés de manière incorrecte. Mais alors, quelles différences ?

3.2.3.1 La labellisation

Le label est **une marque ou un signe distinctif établi par un organisme, qu'il soit privé ou public lequel vérifie la conformité aux critères** du dit label.

Il est attribué à un produit ou service pour attester qu'il répond à des normes spécifiques, souvent supérieures aux exigences réglementaires (comme la qualité, l'origine, etc.). **Le but d'un label est de valoriser un produit ou un service** en signalant au candidat qu'il possède des attributs particuliers.

Les labels sont nombreux et peuvent varier en fonction des produits, des industries, entre autres. C'est une démarche toujours volontaire.

L'obtention des labels « IA responsable » s'inscrit dans une « démarche volontaire » en attendant la finalisation de la réglementation européenne.

Des entreprises choisissent des labels comme positive.ai, **GEEIS-AI** et **LabelIA** pour évaluer la **qualité responsable** de leur organisation ou des produits développés par leur entreprise.

3.2.3.2 La certification

D'après **l'AFNOR** (**A**ssociation **F**rançaise de **NOR**malisation), la **certification est un processus par lequel une entité indépendante et reconnue atteste qu'une organisation, un processus, un service, un produit ou des aptitudes professionnelles respectent les critères** définis dans un cadre de référence (un référentiel).

Ainsi, ce sont uniquement des entités **reconnues par l'État** qui octroient la certification lorsque l'entreprise satisfait à des normes détaillées et à des exigences spécifiques. **La certification est une initiative prise de son propre gré s'il s'agit d'exigences définies dans un référentiel : une démarche « qualité » volontaire. A l'opposé, s'il s'agit de normes obligatoire, cette démarche peut être obligatoire**.

Dans le cadre de l'IA responsable, les exigences ne sont pas encore des normes et le seront probablement lors de l'entrée en vigueur de l'AI Act en Europe, par exemple.

En ce qui concerne la certification « IA responsable », **une seule peut être obtenue en France**. Proposée par le **Laboratoire national de métrologie (LNE)**, elle est orientée sur l'évaluation du processus de développement des systèmes d'Intelligence Artificielle. Elle se concentre principalement sur la mesure de la qualité **selon un standard ISO**.

Dans le domaine particulier de l'IA responsable, la certification se distingue des labels par le fait que les **labels sont principalement basés sur des principes tels que l'équité, la transparence ou l'impact environnemental**.

3.2.3.3 Et les normes dans tout cela ?

La normalisation soutient la législation et la réglementation, aidant à **réduire l'inflation des textes législatifs et réglementaires, surtout dans un monde en évolution rapide**. C'est le cas des « normes harmonisées », initiées par la Commission européenne pour garantir la conformité des produits et services aux exigences techniques légales. Ces normes sont le croisement entre les besoins réglementaires et leur mise en œuvre par le marché.

La Commission européenne, en se tournant vers cette approche pour l'IA, donne aux normes un rôle crucial pour le futur marché européen de l'IA et la confiance des citoyens européens.

Considérons les systèmes d'IA à haut risque. Ils seront réglementés et examinés pour conformité par des entités désignées. Les normes harmonisées détailleront ces exigences obligatoires, notamment celles liées à la confiance, et serviront de base pour les critères de certification.

Pour de nombreux systèmes d'IA, incluant de nombreuses applications industrielles, ces normes définiront les spécifications techniques pour se conformer à la réglementation (art. 43 AI act). De plus, suivre ces normes pourrait aussi offrir une présomption de conformité (art. 42 AI Act).

3.2.3.4 Et alors, qu'est-ce que la conformité ?

La conformité, également appelée « **compliance** » en anglais, englobe toutes les initiatives prises par les entreprises, et particulièrement les banques, pour **garantir que les dirigeants et les employés adhèrent aux normes réglementaires propres au secteur**.

La conformité est une démarche d'analyse des systèmes, services, produits et compétences professionnelles **pour vérifier qu'ils répondent parfaitement aux critères établis dans un référentiel, une loi ou une réglementation**.

Ainsi, lorsque l'AI Act sera en vigueur par exemple, les banques ou plus généralement toutes entreprises ayant un système d'IA devront vérifier leur conformité par rapport aux exigences de cette règlementation.

3.2.3.5 Qui peut se faire certifier ou avoir un label ? et quand ?

Les entreprises, et notamment les banques, qui auraient déjà des systèmes d'IA peuvent être candidats à l'obtention d'une certification ou d'un label, voir même des deux.

Le moment pour demander une évaluation externe sera déterminé suite à l'auto-évaluation. En effet, il n'est pas pertinent de se porter candidat à une certification et un label si l'auto-évaluation met déjà en exergue un travail plus ou moins important afin de combler des éventuels manques sur des exigences qui seront sans doute moins fortes que celles d'une évaluation externe.

Ainsi, **une fois les efforts réalisés pour couvrir les exigences de l'évaluation réalisée en interne**, il sera possible d'envisager l'obtention d'une certification ou d'un label.

3.2.3.6 Pourquoi se faire certifier ou avoir un label ?

Avec les labellisations et les certifications, les organisations ont l'opportunité de se préparer opérationnellement à l'IA Act (dont le vote pourrait venir d'ici fin 2023) mais le choix d'un label ou d'une certification dépend des objectifs de l'entreprise et du type de système d'Intelligence Artificielle développé. Les normes vont arriver et vont probablement être complémentaires à des labels ou des certifications.

Les standards (référentiels) en cours sont les suivants :

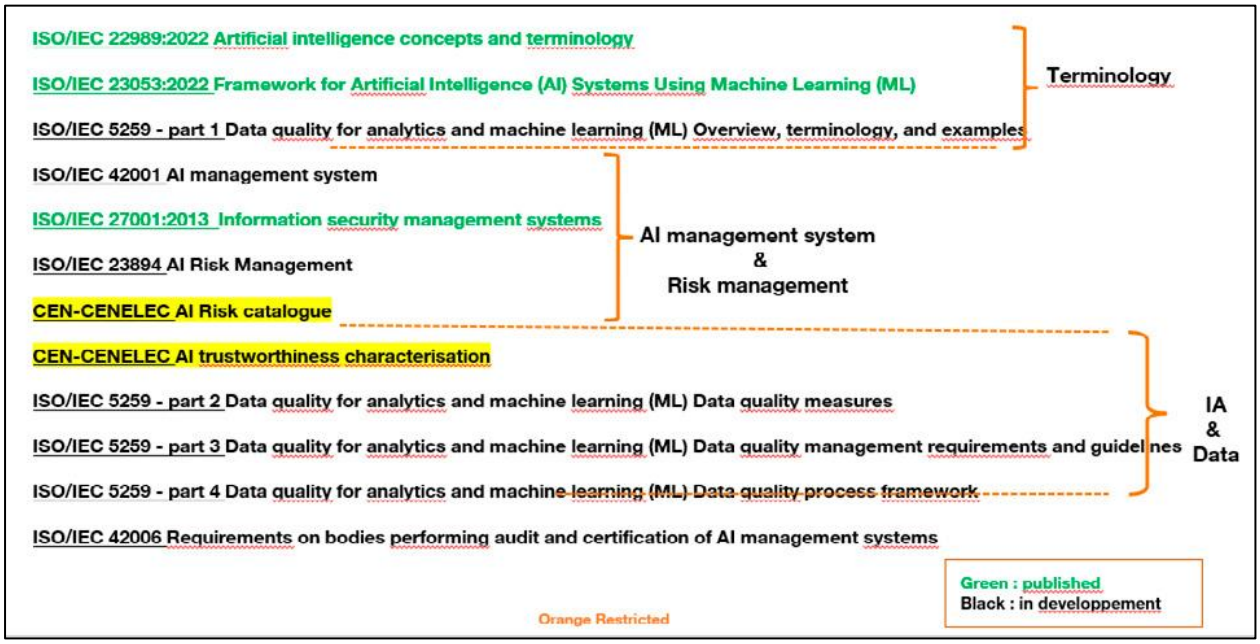


Figure 140 : Liste des standards (référentiels) pour une IA éthique et responsable – Source : ima-dt.org

Les standards de l'AI Act s'appliqueront tout d'abord à ceux présentant les plus hauts niveaux de risque, par exemple **le scoring bancaire ou le filtrage de CV dans le recrutement**.

Si les certifications subsisteront par leur caractère « légal » (reconnaissance par l'état), les labels ne seront certainement pas amenés à disparaître avec l'AI Act. Ils seront un moyen pour les entreprises d'aller plus loin que la simple conformité réglementaire en démontrant le caractère responsable des IA développées pour des cas d'usages moins risqués au sens juridique.

En effet, les labels actuels, et même si l'AI Act n'est pas finalisé, s'alignent avec ses grandes principes et reste un moyen fiable de se préparer pour prévenir un effet « RGPD ».

De plus, la pression sociale augmente et les entreprises doivent prendre des mesures pour prouver aux consommateurs que les produits qu'ils utilisent sont de qualité.

Voici en résumé pourquoi certaines entreprises veulent une certification ou une labellisation :

- Les petites entreprises, comme les startups, cherchent souvent à se faire connaître. Pour elles, obtenir un label ou une certification est « stratégique » car peut être un bon moyen de gagner la confiance de leurs clients.

- Pour les grosses entreprises, ces labels/certifications peuvent les aider à mieux s'organiser et à s'assurer que leurs produits sont bien faits et responsables.

- Cela permet également aux entreprises de :

    - **Se préparer aux nouvelles règles**, surtout si elles sont dans des domaines sensibles comme la santé ou la banque.
    - **Eviter les mauvaises réputations** liées à des problèmes avec leurs systèmes.
    - **Se démarquer** de leurs concurrents.
    - **Gagner la confiance** de tout le monde : clients, employés, partenaires.
    - **Répondre aux demandes du marché** qui veut plus de qualité.
    - **Améliorer le travail en équipe** entre différents départements dans l'entreprise.
    - **Faire partie d'un groupe de confiance**, surtout si leurs partenaires sont aussi certifiés.
    - Les certifications et les labels volontaires sont **plus rapides et moins chères que les normes officielles**.

Recommandation :
L'obtention d'un label et/ou d'une certification est stratégique. Faites-vous certifier ou labelliser ! Cela améliore votre image de marque et votre compétitivité.

3.2.3.7 Quoi choisir ?

Il existe un nombre important d'initiatives de certifications et de labels autour de l'IA responsable et la liste augmente de plus en plus avec le temps. Il est donc difficile de se repérer dans cet écosystème en perpétuel évolution.

Néanmoins, la sélection « IA responsable » suivante est une très bonne base (sélection « **Axionable** ») :

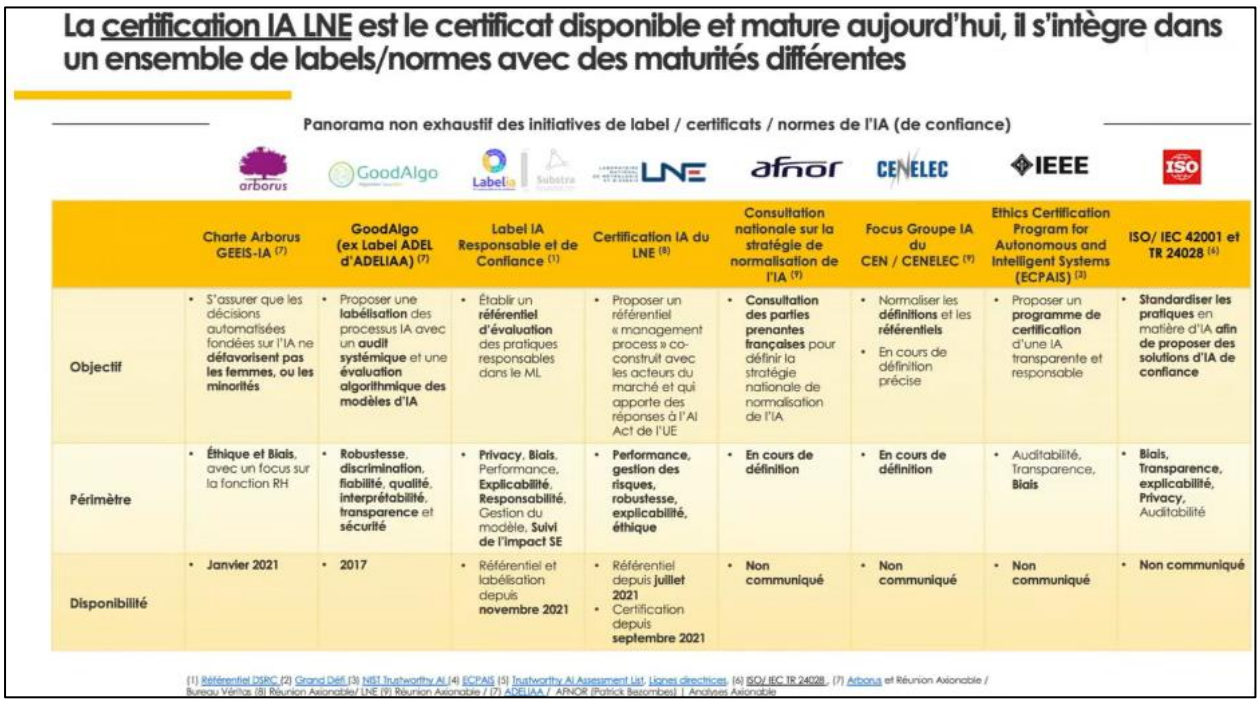


Figure 141 : Quelques labels et certifications pour une IA éthique et responsable – Source : [Axionable - présentation chez Datacraft, octobre 2022]

En pratique, il y a également une gamme de mesures pour garantir que les systèmes d'IA sont responsables et de bonne qualité. Cela va des solutions les moins strictes, comme **une charte ou un code de conduite basé principalement sur la mise en place d'une gouvernance et l'auto-évaluation, jusqu'à l'audit des systèmes d'IA par un tiers pour obtenir une certification officielle**.

Dans le choix d'un label et/ou d'une certification, il faut aussi penser à l'image de l'entreprise. Les labels ou certifications doivent correspondre à ce qu'elle veut montrer. Et avant de se faire évaluer, il peut être bon de travailler sur soi-même. **Chaque entreprise doit trouver sa voie**.

Par exemple, **Orange** a choisi un label sur l'IA parce qu'ils travaillaient déjà sur l'égalité au travail. Les banques, comme le **Crédit Agricole**, seront probablement amenées à choisir une certification car elles ont déjà des processus stricts.

D'après la couverture actuelle des exigences des labels et certifications actuelles, notamment celles de la liste précédente, **un choix pertinent serait le suivant** :

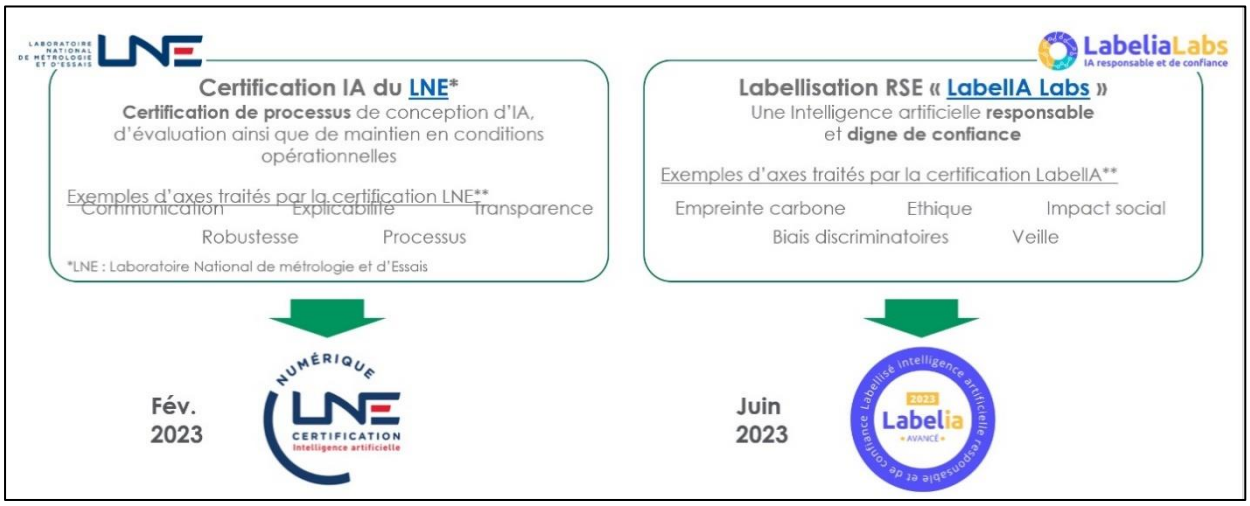


Figure 142 : Choix du Crédit Agricole : certification LNE + label « LabelIA Labs » – Source : Datalab Groupe CA

C'est le choix du **Crédit Agricole**, lequel **a obtenu la certification LNE en février 2023 et le label LabelIA Labs en juin 2023, et figurent parmi les précurseurs en France dans ce type de démarche**.

Ce choix est stratégique et complémentaire. **La certification LNE et le label « LabelIA Labs »** couvrent chacun de leur côté des exigences de l'AI Act et cela avec peu de recouvrement l'un envers l'autre.

En effet, la combinaison des référentiels existants du **LNE** (pour répondre aux enjeux de maitrise des risques et de confiance) et **LabelIA Labs** (pour une IA responsable respectueuse des enjeux sociaux et environnementaux) assure « *in fine* » **une complétude et un process « IA responsable », bout en bout « by design »** :

Figure 143 : Certfication LNE et Label "LabelIA Labs" sont complémentaires - Source : Axionable - présentation chez Datacraft, octobre 2022

LNE et LabelIA Labs sont les initiatives que beaucoup d'entreprises considèrent les plus abouties, avec un référentiel et certificat/label disponible. Ils sont souvent le choix naturel des entreprises, de plus que ces référentiels ont été coconstruits par des groupes de travail de grands groupes, PME et Data-Scientists avec l'AI Act en cible.

Recommandation : Choisir la certification LNE et le label « LabelIA Labs » vous assure une couverture maximum d'un ensemble d'exigences « IA responsable ».

De nombreux témoignages sont disponibles en annexes, notamment ceux de **Orange** (annexe « Labellisation chez Orange »), **Crédit Mutuel Arkéa** (annexe « Labellisation chez Crédit Mutuel Arkéa »), **Datalab Groupe Crédit Agricole** (annexe « Certification chez Crédit Agricole »), **Banque de France** (annexe « Labellisation chez Banque de France »)**.**

3.2.4 Recommandations et bonnes pratiques pour une IA responsable dans les banques pour une application par phase

S'il est important de connaitre sa maturité en termes d'application de principes « IA responsable » au sein d'un système d'IA, il est également important de prendre en considération toutes ces exigences et principes pour tout nouveau projet.

Il s'agit donc ici de fournir des bonnes pratiques et recommandations qui pourront être appliquées à l'ensemble des phases de construction d'un projet de système d'IA. Aussi, et si un projet a déjà démarré et est encore dans ses phases initiales, ces recommandations pourraient tout à fait être prises en compte en revenant sur les premières phases du projet.

En effet, l'usage important des méthodes dites « **Agiles** » prennent déjà en compte les adaptations possibles d'un projet via notamment l'évolution de ses besoins. Cela se traduit généralement par des « sprints » supplémentaires si la méthode « **SCRUM** » est utilisée. **Il n'est donc pas trop tard, même si un projet a déjà commencé**.

Il est proposé ici ce qui constitue le cycle de vie d'un système IA prenant en compte les enjeux de l'IA responsable :

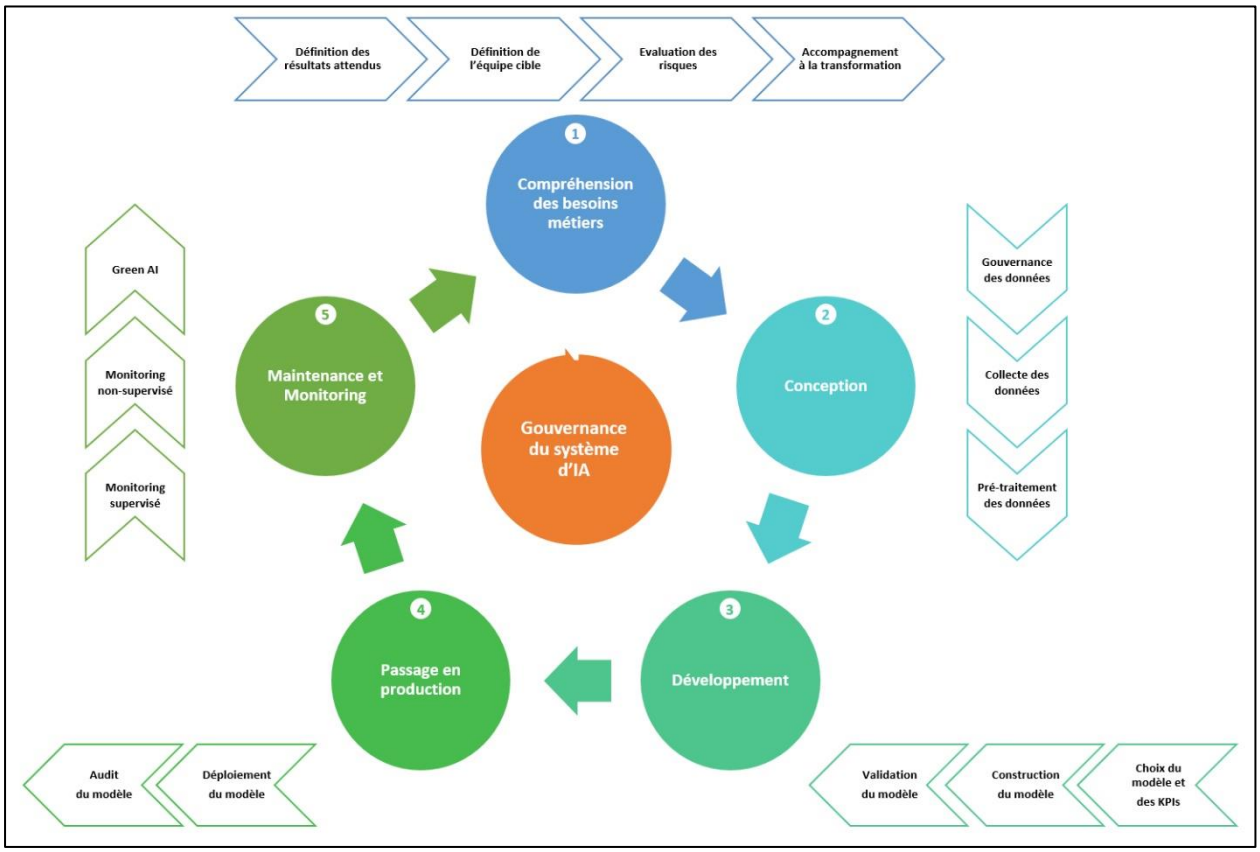


Figure 144 : Démarche par phase pour la mise en place d'une IA responsable - Source : autrice

Il est important de parcourir chacune de ces phases afin de souligner les éléments importants à prendre en compte pour constituer une IA responsable.

3.2.4.1 Phase 1 : la compréhension des besoins métiers

Les systèmes d'IA répondent aux exigences particulières de certains domaines métiers. **La démarche initiale, qui consiste à clairement définir ces objectifs « métier », est également un premier élément de risque**. Bien que ce risque ne soit pas exclusif à l'IA, il peut être accentué pour diverses raisons qu'il est essentiel de reconnaître afin d'en atténuer les conséquences.

3.2.4.1.1 Définition des résultats attendus

Un système d'IA vise généralement à générer un résultat qui est celui attendu par les "métiers", futurs clients utilisateurs du système. Il peut y avoir plusieurs besoins de résultats attendus.

Si ce besoin n'est pas correctement défini en fonction du cas d'usage ou du contexte, le système d'IA risque de **fournir des résultats qui ne correspondent pas aux attentes**.

Cela peut affecter les Data Scientists, car la modélisation serait inappropriée, et les utilisateurs de ces résultats pourraient **subir une performance insatisfaisante**.

Il est donc crucial de prêter une attention spécifique à la définition des résultats attendus :

- Une description complète et commune de la cible et de sa mesure,
- Vérification de la concordance entre les résultats attendus et la réalité sur le terrain.

Exemple de définition des résultats attendus : un système d'IA pour de la prédiction de la solvabilité d'un client par une banque pour décider d'accorder ou non un prêt.

Les résultats attendus sont ce qui est appelé généralement les variables de sortie (variables cibles) et correspondrait ici à la solvabilité : solvable (le prêt peut être accordé) ou insolvable (le prêt ne doit pas être accordé).

Dans cette étape, il s'agit également de donner une définition claire du problème à résoudre, de son contexte et périmètre, et des jalons (dates clés) attendus. Stratégiquement, il est essentiel de définir l'objectif principal du système, de comprendre son intégration dans l'écosystème existant et d'identifier les parties prenantes clés. Cela peut impliquer de décider si l'IA vise à automatiser entièrement une tâche, à assister l'utilisateur ou à offrir une nouvelle fonctionnalité jusqu'alors inexplorée.

Par conséquent, les « métiers » devront aussi donner les impacts des résultats erronés.

3.2.4.1.2 Définition de l'équipe cible

Si le problème à résoudre, les besoins et les jalons sont clairement définis, la définition de l'équipe cible (ou des équipes) pourra se faire.

La réalisation d'un système d'IA nécessite de nombreuses compétences et donc de nombreux profils qui seront amenés à collaborer pour un but commun.

Certains ne sont pas des spécialistes de l'Intelligence Artificielle et en particulier de la Data Science alors que d'autres le sont. Certains profils devront même être des experts comme les profils IT mais **surtout les « Data Scientists » qui doivent collaborer étroitement avec les métiers**.

Comme le montre la figure suivante, trois grandes catégories d'acteurs seront généralement nécessaires :

- Les profils « métier »,
- Les profils IT,
- Les experts en Data Science.

Ces acteurs peuvent avoir des noms différents selon chaque banque, ou les profils suivants peuvent être distribué différemment :

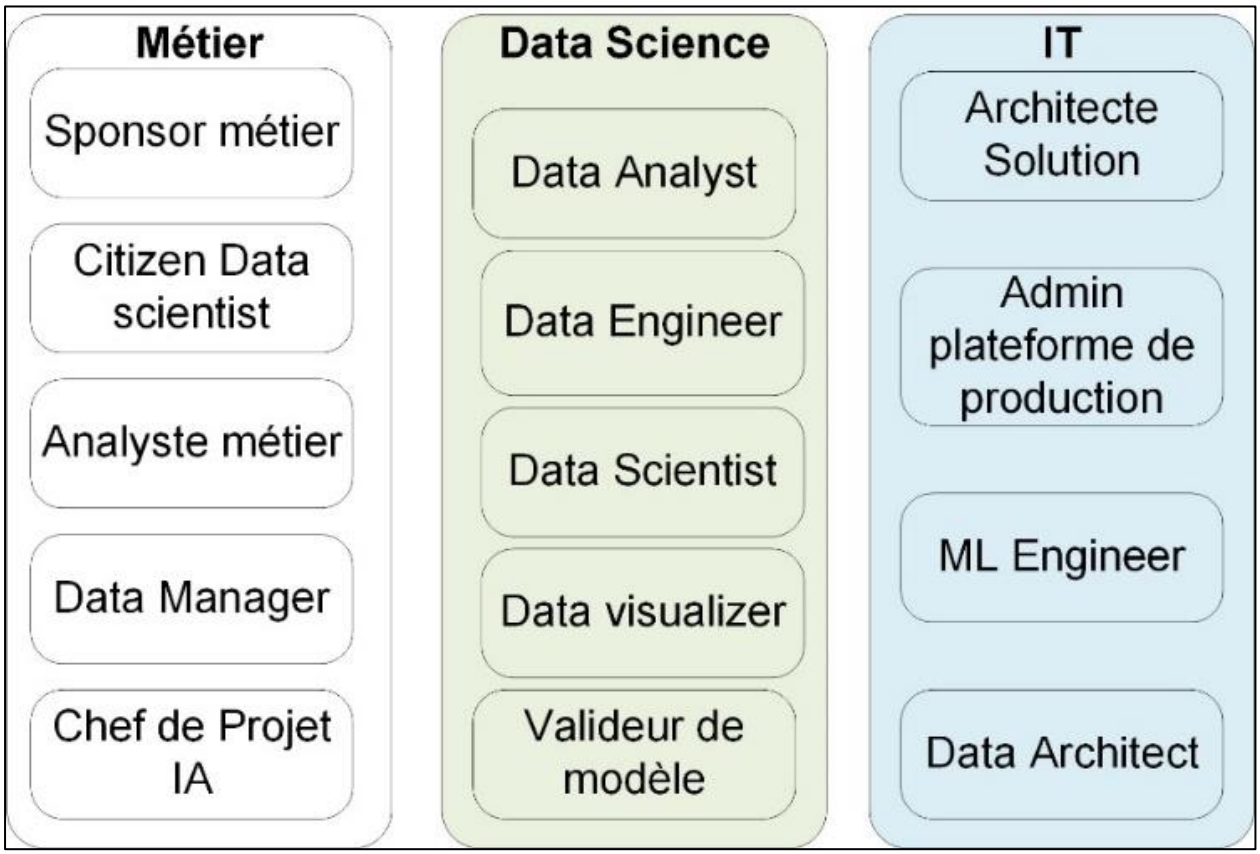


Figure 145 : Equipe type pour construire une IA responsable - Source : hub-franceia.fr

Il faut définir clairement les rôles et responsabilités au sein de l'équipe, notamment en matière d'éthique afin de s'assurer que **les questions éthiques sont régulièrement abordées et traitées**.

Recommandations :
Pour des considérations éthiques, la composition de l'équipe devrait refléter une diversité de genres, d'origines ethniques, de compétences et de perspectives.
Une équipe diverse est plus susceptible d'identifier et de traiter une gamme plus large de préoccupations éthiques. L'idéal est également que les membres de l'équipes suivent ou ont suivi une formation en « éthique de l'IA » afin que cela fasse partie de leurs préoccupations.
L'équipe doit s'engager à travailler de manière transparente, en documentant les décisions prises, les méthodologies utilisées et les considérations éthiques abordées.

Mohamed BEJI
Data Scientist
(CACIB)

Le rôle du Data Scientist

Il a un rôle de conseil quand on travaille avec les métiers qui nous proposent des données : il faut les conseiller, les cadrer, leur montrer les enjeux éthiques derrière ces IA et surtout alerter. Il faut que le modèle souhaité par les métiers soit juste.

Pour certains métiers, l'IA responsable ne leur parle pas et la majorité des actions concernant ce sujet sont réalisées par le Data Scientist pour par exemple assurer le respect de la vie privé et la gouvernance des données.

Le Data Scientist doit par exemple, absolument attirer l'attention des métiers quand il voit que la majorité des variables sont basées sur des données confidentielles personnelles (vie privée) et qu'en plus, elles vont biaiser le modèle. Notamment pour les raisons suivantes :

- On ne respecte pas le RGPD
- Le modèle sera biaisé et pas performant : par exemple, il va apprendre sur les noms des personnes et donc le jeu de données d'entrainement ne représente probablement pas la réalité des données, ou par exemple, si le genre (sexe) amène plus de chances d'avoir une prédiction.

3.2.4.1.3 Evaluation des risques liés au cas d'usage par un comité d'éthique

Le rôle d'un comité éthique dans le contexte d'un projet de création d'un système d'IA est fondamental pour **garantir que l'IA est développée et utilisée de manière responsable**.

Bien que **le comité éthique intervienne de manière transverse sur plusieurs phases** de la mise en place d'un projet d'un système d'IA, il est important que celui-ci soit inclus **au plus tôt**.

Le comité évalue les implications éthiques potentielles de l'IA de manière générale, en se penchant sur des questions telles que **la transparence, la justice, la non-discrimination, la vie privée et la sécurité**. Sur la base de ses évaluations, le comité peut formuler **des directives et des normes éthiques** pour guider le développement et l'utilisation de l'IA.

Ce sont ces directives qui seront importantes durant cette phase afin que tous « **Asset** » ou recommandations soient prises en compte **dès la phase initiale, dans la définition du besoin, « by design ».**

Le comité éthique, durant cette phase, servira de forum pour rassembler les opinions et les préoccupations des différentes parties prenantes mais aura un **rôle potentiellement éducatif, en sensibilisant l'équipe « projet » et les utilisateurs finaux aux questions éthiques** liées à l'IA et en offrant des **formations** appropriées si nécessaire.

Enfin, un des rôles majeurs du comité éthique est celui de la gestion des risques. Comme abordé précédemment dans la partie 2, le cadre réglementaire de l'AI Act catégorise les système d'IA selon 4 niveaux, de risque « inacceptable » à risque « minime ».

Leur consultation doit servir à :

- **Valider le besoin selon les réglementations en vigueur** (ou à venir), pour ne pas aller à l'encontre de règles
- **Analyser les risques d'IA afin de définir le niveau de risque** « AI Act » du futur système d'IA
    - S'il existe une analyse des risques de type MEHARI (ou méthode local à la banque, par ex. **MESARI** chez **Crédit Agricole**), elle pourra être lue, confrontée aux « risques d'un point de vue IA » et être complétée en conséquence
- Eventuellement, **donner des exigences supplémentaires** au-delà de ce que pourrait demander les réglementations (selon le contexte de chaque banque)
    - Cas du Crédit Agricole : définition d'une charte éthique d'entreprise en complément des réglementations (**cf. annexe ou lien** ici)

Les directives évoqués précédemment seront donc ajustées selon le niveau de risque résultant de cette consultation.

Le schéma suivant, crée par Burges Salmon et publié le 15 juin 2023 suite à la dernière version de l'AI, donne un arbre de décision pour aider à naviguer dans l'AI Act qui identifie les décisions clés à prendre en faisant référence aux sections pertinentes du dernier projet de l'AI Act et **permet de d'estimer le niveau de risque d'un système d'AI**.

Il pourrait être d'une première aide en l'absence d'élément existant en interne dans votre organisation mais n'a pas vocation à remplacer la documentation officielle de l'AI Act :

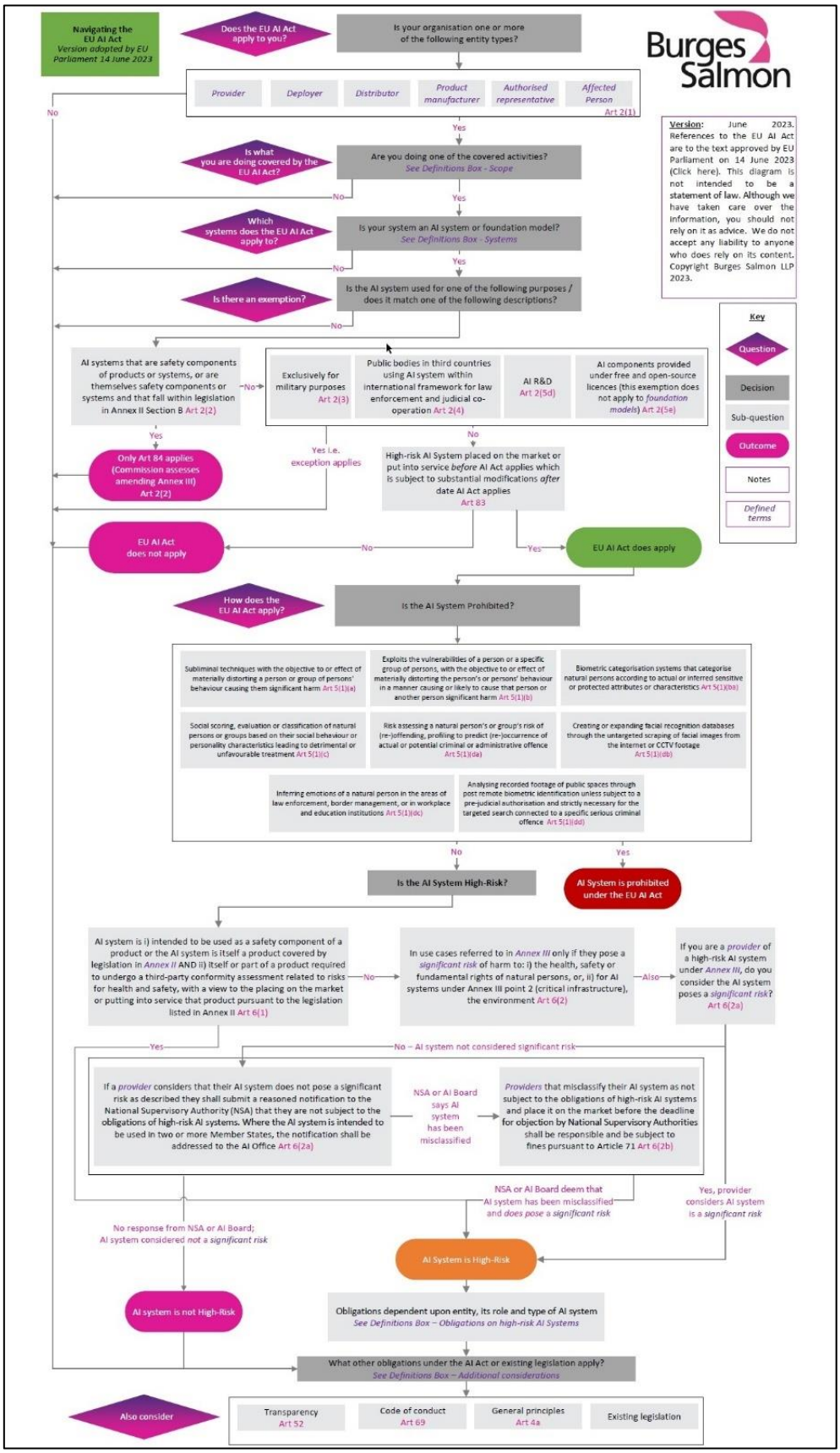
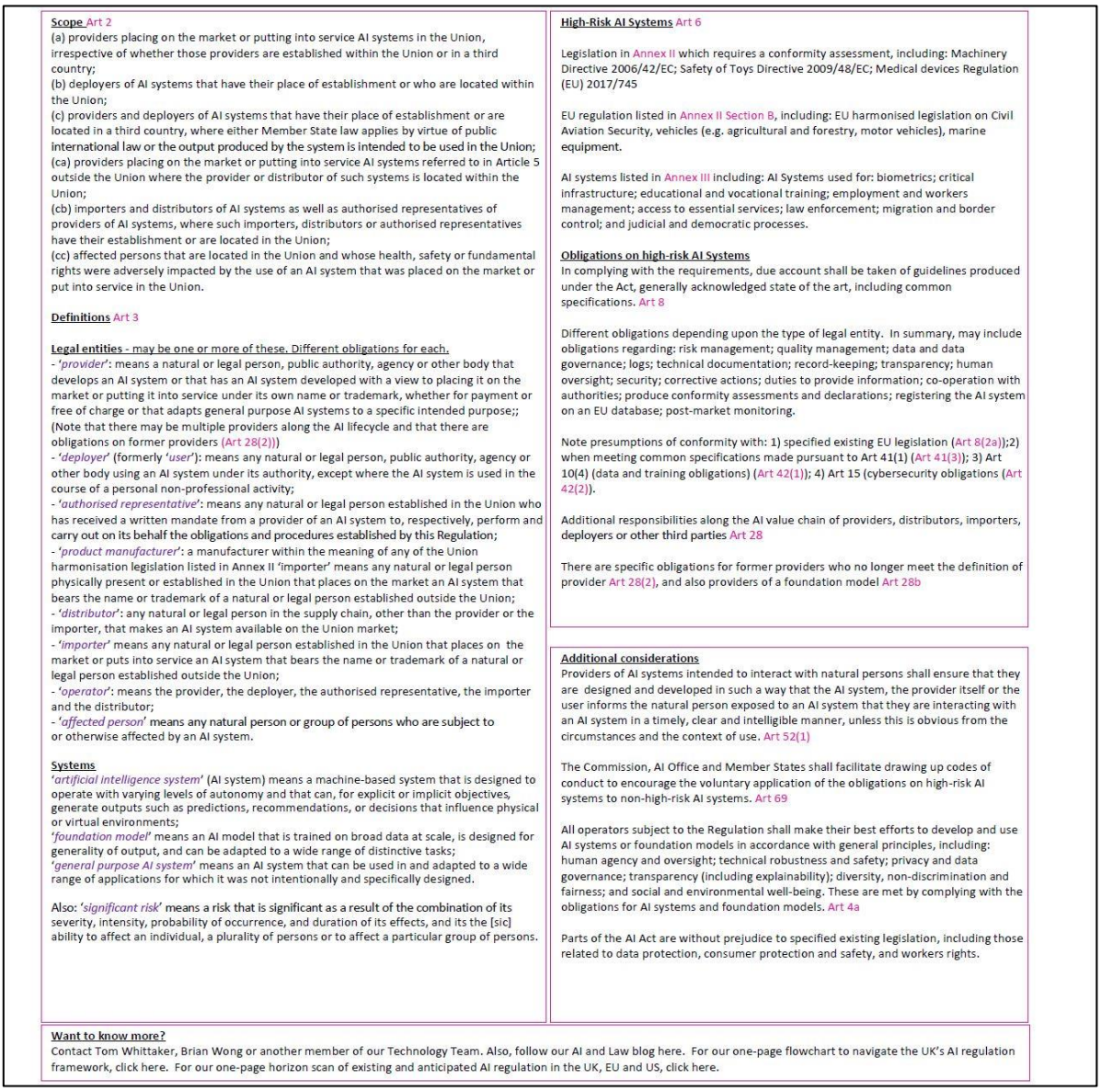


Figure 146 : Arbre de décision pour déterminer le niveau de risque d'un projet IA - Source : [Burges Salmon]

Et la partie réglementation dont l'arbre fait référence :



Scope Art 2
(a) providers placing on the market or putting into service AI systems in the Union, irrespective of whether those providers are established within the Union or in a third country;
(b) deployers of AI systems that have their place of establishment or who are located within the Union;
(c) providers and deployers of AI systems that have their place of establishment or are located in a third country, where either Member State law applies by virtue of public international law or the output produced by the system is intended to be used in the Union;
(ca) providers placing on the market or putting into service AI systems referred to in Article 5 outside the Union where the provider or distributor of such systems is located within the Union;
(cb) importers and distributors of AI systems as well as authorised representatives of providers of AI systems, where such importers, distributors or authorised representatives have their establishment or are located in the Union;
(cc) affected persons that are located in the Union and whose health, safety or fundamental rights were adversely impacted by the use of an AI system that was placed on the market or put into service in the Union.

Definitions Art 3

Legal entities - may be one or more of these. Different obligations for each.
- '*provider*': means a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge or that adapts general purpose AI systems to a specific intended purpose;; (Note that there may be multiple providers along the AI lifecycle and that there are obligations on former providers (Art 28(2)))
- '*deployer*' (formerly '*user*'): means any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity;
- '*authorised representative*': means any natural or legal person established in the Union who has received a written mandate from a provider of an AI system to, respectively, perform and carry out on its behalf the obligations and procedures established by this Regulation;
- '*product manufacturer*': a manufacturer within the meaning of any of the Union harmonisation legislation listed in Annex II 'importer' means any natural or legal person physically present or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established outside the Union;
- '*distributor*': any natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market;
- '*importer*' means any natural or legal person established in the Union that places on the market or puts into service an AI system that bears the name or trademark of a natural or legal person established outside the Union;
- '*operator*': means the provider, the deployer, the authorised representative, the importer and the distributor;
- '*affected person*' means any natural person or group of persons who are subject to or otherwise affected by an AI system.

Systems
'*artificial intelligence system*' (AI system) means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions that influence physical or virtual environments;
'*foundation model*' means an AI model that is trained on broad data at scale, is designed for generality of output, and can be adapted to a wide range of distinctive tasks;
'*general purpose AI system*' means an AI system that can be used in and adapted to a wide range of applications for which it was not intentionally and specifically designed.

Also: '*significant risk*' means a risk that is significant as a result of the combination of its severity, intensity, probability of occurrence, and duration of its effects, and its the [sic] ability to affect an individual, a plurality of persons or to affect a particular group of persons.

High-Risk AI Systems Art 6

Legislation in Annex II which requires a conformity assessment, including: Machinery Directive 2006/42/EC; Safety of Toys Directive 2009/48/EC; Medical devices Regulation (EU) 2017/745

EU regulation listed in Annex II Section B, including: EU harmonised legislation on Civil Aviation Security, vehicles (e.g. agricultural and forestry, motor vehicles), marine equipment.

AI systems listed in Annex III including: AI Systems used for: biometrics; critical infrastructure; educational and vocational training; employment and workers management; access to essential services; law enforcement; migration and border control; and judicial and democratic processes.

Obligations on high-risk AI Systems
In complying with the requirements, due account shall be taken of guidelines produced under the Act, generally acknowledged state of the art, including common specifications. Art 8

Different obligations depending upon the type of legal entity. In summary, may include obligations regarding: risk management; quality management; data and data governance; logs; technical documentation; record-keeping; transparency; human oversight; security; corrective actions; duties to provide information; co-operation with authorities; produce conformity assessments and declarations; registering the AI system on an EU database; post-market monitoring.

Note presumptions of conformity with: 1) specified existing EU legislation (Art 8(2a));2) when meeting common specifications made pursuant to Art 41(1) (Art 41(3)); 3) Art 10(4) (data and training obligations) (Art 42(1)); 4) Art 15 (cybersecurity obligations (Art 42(2)).

Additional responsibilities along the AI value chain of providers, distributors, importers, deployers or other third parties Art 28

There are specific obligations for former providers who no longer meet the definition of provider Art 28(2), and also providers of a foundation model Art 28b

Additional considerations
Providers of AI systems intended to interact with natural persons shall ensure that they are designed and developed in such a way that the AI system, the provider itself or the user informs the natural person exposed to an AI system that they are interacting with an AI system in a timely, clear and intelligible manner, unless this is obvious from the circumstances and the context of use. Art 52(1)

The Commission, AI Office and Member States shall facilitate drawing up codes of conduct to encourage the voluntary application of the obligations on high-risk AI systems to non-high-risk AI systems. Art 69

All operators subject to the Regulation shall make their best efforts to develop and use AI systems or foundation models in accordance with general principles, including: human agency and oversight; technical robustness and safety; privacy and data governance; transparency (including explainability); diversity, non-discrimination and fairness; and social and environmental well-being. These are met by complying with the obligations for AI systems and foundation models. Art 4a

Parts of the AI Act are without prejudice to specified existing legislation, including those related to data protection, consumer protection and safety, and workers rights.

Want to know more?
Contact Tom Whittaker, Brian Wong or another member of our Technology Team. Also, follow our AI and Law blog here. For our one-page flowchart to navigate the UK's AI regulation framework, click here. For our one-page horizon scan of existing and anticipated AI regulation in the UK, EU and US, click here.

Figure 147 : Articles de l'AI Act concernés par l'arbre de décision précédent - Source : Burges Salmon

Recommandations : La mise en place d'un comité éthique pour superviser les projets d'Intelligence Artificielle offre de nombreux avantages :

- Il veille à ce que les projets d'IA soient développés et déployés de manière éthique en favorisant la transparence
- Il identifie les risques et fournit les recommandations pour les atténuer
- Il s'assure de la conformité réglementaire et de l'amélioration de la qualité des projets
- Il renforce la confiance des parties prenantes
- Il peut jouer un rôle éducatif sur les enjeux d'IA responsable et parfois fournir des formations.

Si vous avez déjà un comité éthique, tournez-vous vers lui et il saura vous fournir de l'aide fortement utile.
S'il n'existe pas, faites-vous entendre pour demander sa création !

Mohamed BEJI
Data Scientist
(CACIB)

Afin de mettre en place une IA responsable, il est important d'avoir :

- Une gouvernance et les réglementations internes,
- Un comité risque (AIDA chez Crédit Agricole) qui vérifie les risques, la fiabilité et la sécurité (MRM + compliance), et fournit la méthodologie à suivre,
- **Un comité "éthique"** (MRM chez CACIB) qui évalue l'aspect éthique selon la réglementation en vigueur,
- La méthodologie à suivre pour anticiper les dégradations des performances des modèles, l'assurance de l''amélioration continue du modèle par design (dès le début) quand on doit désigner la solution et quand on doit entraîner le modèle. Il faut qu'il soit ré-entrainable et améliorable dans le futur,
- Outillages adéquates nécessaires à la mise en place des cas d'usage.

Témoignage d'un comité éthique chez Crédit Agricole :

- **AIDA : une IA performante et conforme dès la conception « by design » :**

Un comité opérationnel **AIDA** (« **AI D**esign **A**uthority ») transverse au groupe **Crédit Agricole** se réunit **mensuellement pour passer en revue les projets d'IA** en phase de conception et d'implémentation.

Les réunions mensuelles rassemblent les responsables **métiers**, le **Data Scientist** responsable de la conception, le représentant de **l'AI Factory**, ainsi que les équipes « **Risques** » et « **Conformité** ». Ces sessions évaluent chaque projet et fournissent des orientations initiales. **Cette revue définit également des exigences par projet**.

Cette phase détermine le processus de contrôle de l'IA : examen des risques, conformité, ou les deux. **La décision est proportionnelle à la gravité du risque**.

L'objectif est de favoriser une conception « Conforme dès la conception » et d'éviter les corrections lors de la revue d'implémentation.

Voici un schéma du process « AI Design Authority » :

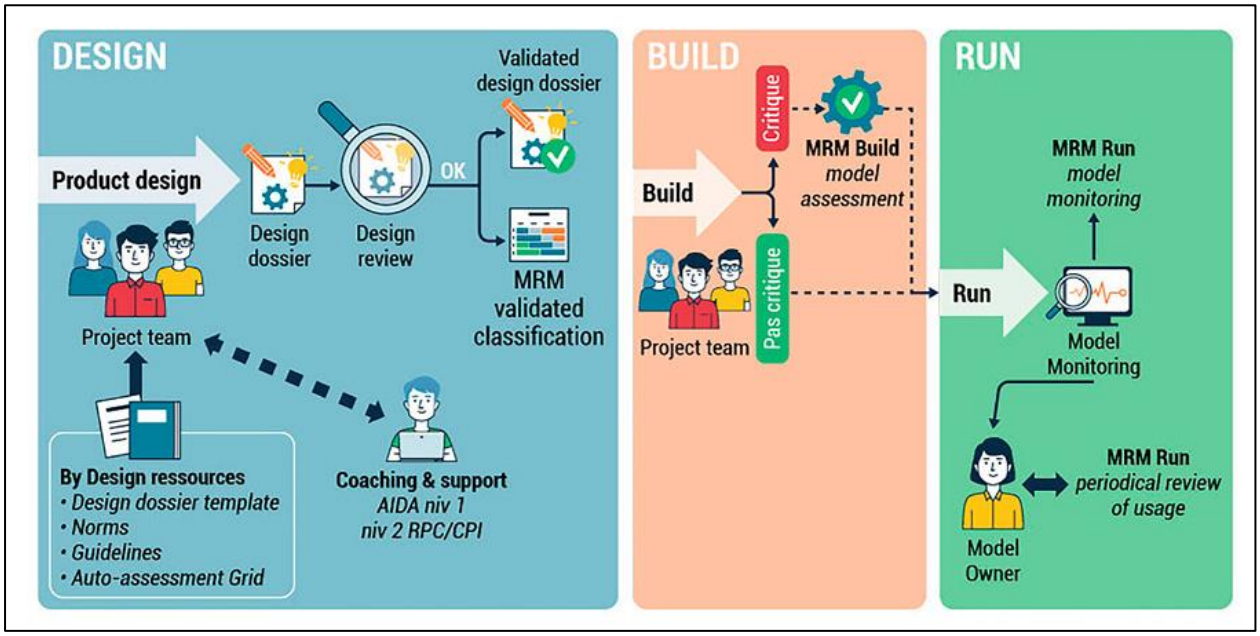


Figure 148 : Schéma du process « AI Design Authority » (AIDA) - Source : ima-dt.org

Voir d'autres témoignages à ce sujet en annexes pour **Société Générale** (lien « Cadre méthodologique chez Société Générale ») ou **Orange** (lien « Comité éthique chez Orange »).

3.2.4.1.4 Accompagnement à la transformation

Il est important de faire comprendre à tous les acteurs de votre organisation que l'IA n'est pas simplement une nouvelle tendance marketing, mais un véritable levier de compétitivité au service du business.

Cette prise de conscience suscitera un intérêt accru chez les acteurs et les conduira naturellement à se demander : « Comment l'IA peut-elle renforcer mon activité ? Comment peut-elle optimiser ma productivité ou ma compétitivité ? ». Bien entendu, les enjeux d'une IA responsable doivent être au cœur des réponses.

Pour cela, **il est fortement recommandé de lancer un programme d'acculturation et de formations** :

- **Sensibilisation pour le Business** (Management et opérationnels) pour clarifier et illustrer les différents cas d'usage en intégrant les valeurs « IA responsable ».
- **Formation des équipes projet** sur les particularités du cycle de vie logiciel de l'IA (**MLOps**, réentraînements) mais aussi les principes éthiques, de transparence, etc.
- **Informer** les data analysts, Data Scientists et ingénieurs ML/data **sur les dangers potentiels et les méthodes recommandées** (assurer l'absence de biais dans les données, encourager l'acceptation des utilisateurs grâce à l'explicabilité...).

- **Mettre en avant les bénéfices apportés par l'IA** est essentiel car l'objectif premier est d'ajouter de la valeur à l'entreprise. **Les bénéfices de l'IA responsable sont également à souligner**.

Bien que les bénéfices de l'IA soient souvent exprimés en termes financiers, il est aussi pertinent de réfléchir à **sa valeur sur le plan humain**, comme le bien-être des employés ou la confiance des clients. Ces aspects, même si moins concrets, sont essentiels pour le succès d'un projet en IA.

Recommandations :
La mise en place de formations sur l'IA et particulièrement l'IA éthique et responsable permettra de sensibiliser vos employés sur ses enjeux et le(s) futur(s) cadre(s) réglementaire(s).
Elles permettront également d'améliorer la qualité de vos projets car toutes les parties prenantes parleront avec connaissance.
Ils s'agit aussi d'un premier pas vers le « **upskilling** », c-à-d. vers le renforcement des compétences de vos employés avec une perspective d'évolution vers le futur.

3.2.4.2 Phase 2 : la conception

La conception d'un système d'Intelligence Artificielle responsable constitue une étape cruciale dans le développement de solutions technologiques innovantes. Elle **ne se limite pas à la simple élaboration d'algorithmes, mais englobe une série de décisions stratégiques et techniques** qui jouent un rôle déterminant dans le succès du projet.

D'un point de vue technique, les choix s'orientent autour des données : **quelles sources** utiliser, **comment les traiter** et **les nettoyer**, **quelles architectures de modèles** déployer, ou encore **comment** évaluer et itérer le système pour **optimiser sa performance et le monitorer**. Ces décisions doivent également tenir compte des contraintes de matériel, de temps et de budget, ainsi que des implications en matière de respect de la vie privée et de sécurité.

Cette phase doit également prendre en compte des **aspects éthiques, de responsabilité et de durabilité**, afin de garantir que le système final est à la fois efficace, fiable et en accord avec les normes sociétales.

En d'autres termes, la phase de conception façonne la trajectoire de l'IA, déterminant son potentiel d'impact, ses performances et sa pertinence dans des contextes réels. L'alignement harmonieux de ces orientations est fondamental pour assurer l'efficacité et la pertinence de l'IA dans son contexte d'application.

3.2.4.2.1 La gouvernance des données

- **Le choix des données :**

Le choix des données et de leurs sources pour un système d'IA est une **étape déterminante qui influence la qualité et la fiabilité du modèle**.

Les sources de données doivent directement être liées à l'objectif du projet d'IA. Elles doivent fournir les informations nécessaires pour répondre à la question ou au problème à résoudre. Les sources reconnues au sein de votre organisation pour leur sérieux et leur exactitude sont à privilégier et plus particulièrement les « **golden sources** ».

Une « **golden source** » (appelée « **source unique de vérité** » ou « **source de référence** ») pour une information spécifique est la source de données principale et fiable qui est considérée comme la référence la plus précise et la plus à jour pour cette information.

Elle se différencie par rapport à d'autres sources de données car ces dernières pourraient avoir des copies par rapport à la « golden source » avec plus ou moins d'altérations.

Par ailleurs, les données doivent être diversifiées pour couvrir tous les scénarios possibles mais également (surtout !) éviter les biais inhérents.

Il s'agira également de choisir les données strictement nécessaires afin d'assurer un système d'IA dit « **frugale** » sur la partie données, à savoir qui nécessite moins de ressources, tout en conservant une performance acceptable.

En effet, le Big Data apportait dans le passé un « *mindset* » de conservation d'un maximum de données pour y trouver de la valeur ultérieurement. Ce paradigme change pour l'IA (notamment d'un point de vue « **Green IA** ») et il est souvent possible de conserver un IA performant en termes de résultats avec moins de données (au sens attributs ou variables d'entrée).

- **La disponibilité des données :**

Il est crucial d'évaluer si les données nécessaires aux modèles sont accessibles. Ce besoin de disponibilité de données se manifeste à la fois lors de la création du modèle (**build**) et pendant son utilisation (**run**).

Durant la création (build), des problèmes peuvent surgir, comme l'absence d'un historique complet, dû à des contraintes techniques ou réglementaires. Par exemple, des données peuvent être **conservées seulement trois ans alors que dix ans seraient nécessaires** pour une modélisation adéquate.

Pour le build, **il est également important de prévoir les données de test et d'apprentissage dès le départ**. En général, **une bonne recommandation est d'extraire de manière aléatoire 20% de la base de travail pour constituer un échantillon de test pour la phase dite de « recette »** (post développement et avant la production). Les 80% restant de la base de travail constitueront l'échantillon d'apprentissage. Celui-ci doit représenter une « **réserve** » dans laquelle piocher pour réaliser 5 à 10 étapes d'apprentissage (de manière générale, selon sous-apprentissage et sur-apprentissage à contrôler).

Pendant l'utilisation (run), les données peuvent être absentes, incorrectes ou obtenues trop tard. Pour réduire ces risques, **il est recommandé d'analyser les données initiales et de surveiller continuellement les données lors de l'utilisation**. Toute évaluation des risques doit considérer ces mesures.

- **Les données externes :**

L'utilisation de données dites « **externes** » est déterminée selon les besoins « métier » définis préalablement et donc selon le cas d'usage.

Lorsqu'un modèle utilise des données externes, **sa qualité peut être hors de contrôle de l'entité responsable**. Les risques proviennent d'un **possible manque de transparence** sur la façon dont ces données ont été recueillies et traitées.

Contrairement aux données internes, où la qualité est généralement garantie, il revient au Data Scientist ou au responsable du modèle de vérifier la fiabilité des données externes. Ces préoccupations sont pertinentes aussi bien lors de la mise en place du modèle qu'au cours de son utilisation.

- **L'utilisation des données personnelles :**

Les systèmes bancaires contiennent des **données sensibles ou personnelles**. Selon le RGPD, l'utilisation de ces données nécessite souvent un consentement explicite. Il est crucial de s'assurer que les données utilisées correspondent aux finalités pour lesquelles le consentement a été donné.

De plus, **l'utilisation de données historiques doit respecter les règles de durée de conservation et le droit à l'oubli**. Il peut y avoir aussi des contradictions entre différentes régulations. Ces données peuvent inclure des informations individuelles ou même des cookies.

Il est essentiel de gérer ces risques dès le début, en informant correctement les Data Scientists et en mettant en place une gouvernance des données rigoureuse.

- **La protection des données :**

La sécurité des données est au cœur des politiques informatiques des institutions, en particulier pour les informations sensibles.

Toutefois, **le développement et l'application de systèmes d'IA peuvent présenter des vulnérabilités**, que ce soit l'accès aux modèles, l'utilisation d'outils externes, ou l'exposition à des environnements non sécurisés.

Il est crucial que la création et l'utilisation de l'IA soient en ligne avec ces politiques de sécurité.

- **Anonymisation des données :**

Selon la norme ISO/IEC 29100:2011, l'anonymisation transforme les données personnelles de façon irréversible pour qu'une personne ne puisse plus être identifiée. La **pseudonymisation**, quant à elle, permet une identification avec des informations supplémentaires.

Pour les modèles, il est préférable d'utiliser des **données anonymisées** sauf si cela nuit à leur usage, en particulier s'il y a un risque de réidentification. Dans ce cas de figure, il est alors possible d'envisager la **pseudonymisation** tout en s'assurant de respecter la conformité réglementaire, notamment le consentement comme requis par le RGPD.

3.2.4.2.2 Focus sur la qualité des données

La façon dont les données sont mesurées, codées ou simplement leur nature peut introduire des **biais**. Ces biais peuvent altérer les résultats, comme conduire à une mauvaise classification.

Il est donc crucial d'examiner attentivement les données introduites dans le modèle. Ces données pourraient être incomplètes, présenter des **biais dûs à un échantillonnage inadéquat** ou **manquer de représentativité**.

Par ailleurs, les anciennes données pourraient porter les marques de décisions basées sur d'anciens modèles et être biaisées. De plus, des erreurs peuvent survenir lors de la préparation des données, comme une mauvaise catégorisation ou normalisation.

- **Quelques directives à suivre sur les exigences des données pour l'IA** :

  - **La performance d'un modèle d'IA dépend directement de la qualité des données** sur lesquelles il est formé.
  - **Il est essentiel d'utiliser des données directement à leur source**, en s'assurant de leur actualité et exactitude.
  - Outre l'accès aux données à la source, **celles-ci doivent répondre à des normes de qualité**.
  - **Les réglementations**, comme le RGPD, concernant l'utilisation des données et le consentement des utilisateurs **doivent être respectées**.

- **Préparation et adaptation des données pour le modèle** :

  - **Les données sont souvent modifiées ou adaptées** pour être utilisées par les modèles statistiques. Par exemple, des données brutes peuvent être filtrées, croisées, agrégées ou utilisées pour calculer des indicateurs. Ces modifications influencent les entrées de l'algorithme d'apprentissage et sont donc cruciales pour le modèle d'IA.
  - **L'annotation précise des données (les labéliser) est vitale**, bien que cette étape puisse être coûteuse. Des outils peuvent aider à automatiser partiellement cette tâche et à explorer les bases de données pour évaluer leur pertinence ou détecter des biais.
  - **Toutes les modifications apportées aux données doivent être documentées et suivies** pour garantir leur pertinence.
  - **Ces transformations doivent rester cohérentes entre différents modèles et au fil du temps** pour assurer la fiabilité et la reproductibilité des résultats.

- **Recherche et gestion des biais dans les données d'IA :**

Les biais dans les modèles d'IA peuvent provenir de trois sources principales :

1. **Les données d'apprentissage** elles-mêmes, qui peuvent refléter des comportements historiquement biaisés.
2. **Le choix de périmètre des données**, qui ne couvre pas toutes les situations que le modèle pourrait rencontrer.
3. **Les informations utilisées pour créer le modèle** (variables d'entrées). Par exemple, ne pas prendre en compte le sexe dans un modèle d'assurance vie peut entraîner des tarifs plus élevés pour les femmes car elles vivent plus longtemps que les hommes.

Il est donc essentiel de détecter et de corriger ces biais présents dans les données sources. Cependant, c'est un défi car les méthodes pour le faire sont encore en développement. Il faut donc être vigilant et à jour sur ce sujet car ces biais sont courants, les modèles d'IA se basent souvent sur des données qui représentent le passé, pas le présent et **elles peuvent rapidement devenir obsolètes** quand elles réfèrent à des activités ou contextes d'activité humains comme par exemple le COVID.

3.2.4.2.3 La collecte et le pré-traitement des données

La **manière dont on collecte les données avant d'utiliser les modèles est cruciale** pour s'assurer que les systèmes d'IA fonctionnent bien. Bien sûr, la qualité des données que seront utilisées est essentielle. Mais ce n'est pas unique aux systèmes basés sur l'IA.

Même avec des méthodes statistiques traditionnelles ou des systèmes de rapport, des données de mauvaise qualité poseront des problèmes.

Pour les banques, des lois existent déjà sur la qualité des données qu'il faut prendre en considération selon les cas d'usage, comme la directive BCBS 239, une norme émise par le **Comité de Bâle** sur le contrôle bancaire qui concerne les pratiques de gestion des données et de reporting des risques dans les banques. Cette directive s'inscrit dans le contexte des réformes de **Bâle 3** et peut être considéré comme une composante de l'ensemble des réformes de **Bâle 3**.

Si par exemple le cas d'usage concerne les données de risque tel que le risque de crédit, il est important de rappeler que dans cet seule directive, le **Comité de Bâle a établi des principes** sur la gestion des données de risque.

Ces principes nous disent qu'il faut :

- Mettre en place une **gouvernance** liée à la manière dont on regroupe et communique les données de risque,
- Avoir une bonne **architecture de données** et une **infrastructure solide**, surtout lorsqu'il y a des tensions ou des crises,
- S'assurer que les données de risque soient précises et fiables,
- Garantir que les données sont exhaustives,
- Mettre à jour les données pour une production rapide lorsqu'il s'agit de gérer ces données sur les risques.

La visée de ces principes est d'offrir une évaluation plus précise des risques. Néanmoins, **la portée de ces principes peut largement être généralisés et ne sont pas limitable aux seules données de risque**. En effet, ils entraînent une structuration au sein des banques favorisant une perspective globale sur la qualité des données, que ce soit pour les reportings, les modèles d'IA ou n'importe quel autre objectif « métier ».

David CORTES
Président de Alvidence

La préparation des données :

Toute la magie se trouve dans la préparation des données pour que l'algorithme ait un bon niveau de performance et si possible, qu'il soit à peu près explicable.

Le sujet n'est donc pas vraiment dans le modèle en lui-même mais dans la chaîne de traitement total, ce qui veut dire : Je collecte mes données, je les stocke, j'assure leur mise en qualité, leur contrôle via des dictionnaires.

C'est donc toute une chaîne de traitement qui, à la fin donne un modèle car un modèle c'est l'algorithme et des données.

3.2.4.3 Phase 3 : le développement du modèle

Durant la phase de création d'un système d'IA, plusieurs aspects méritent une attention particulière. La sélection de l'architecture d'un modèle d'IA est une étape déterminante qui impacte non seulement la performance du système mais aussi sa fiabilité, sa transparence et son éthique.

Il s'agit de déterminer la structure et la forme du modèle qui traitera et analysera les données pour accomplir la tâche souhaitée. Cette décision, loin d'être triviale, doit intégrer à la fois des considérations techniques et éthiques pour garantir une IA responsable.

3.2.4.3.1 Le choix du modèle

La **sélection de l'architecture d'un modèle d'IA** est une étape déterminante qui impacte non seulement la performance du système mais aussi sa fiabilité, sa transparence et son éthique. Il s'agit de déterminer la structure et la forme du modèle qui traitera et analysera les données pour accomplir la tâche souhaitée. Cette décision, loin d'être triviale, doit intégrer à la fois des considérations techniques et éthiques pour garantir une IA responsable.

Dans les considérations techniques, il est important de revenir au besoin défini initialement : quel est l'objectif du système d'IA ? De cette question doit découler une ou des réponses qui doivent permettre de dire s'il faut réaliser une **classification**, une **régression**, une **génération de texte** ou une **détection d'objets** lesquels demandent des architectures différentes.

Par exemple, pour une tâche de vision par ordinateur, **un réseau neuronal convolutif (CNN)** pourrait être approprié mais ne le serait pas pour d'autres cas d'usage.

La complexité du modèle est également à prendre en compte : **plus l'architecture du modèle est profonde ou large, plus elle aura de paramètres**. Cela peut augmenter la précision, mais cela coûte également en termes de temps de calcul, de mémoire et de risque de surapprentissage. Ces aspects rentrent en compte car pour le besoin, il n'est peut-être pas nécessaire d'avoir beaucoup de paramètres. Ou bien, l'aspect des coûts des ressources peut être dissuasif.

Par ailleurs, certaines architectures de modèle, en particulier les très profondes, peuvent prendre du temps à s'entraîner. **Il faut donc trouver un équilibre entre temps et précision**.

Voici un tableau comparatif des modes d'apprentissages qui ont également **un impact sur la transparence et l'explicabilité, ainsi que sur le risque de biais** (tableau tiré du white paper « *Intelligence Artificielle : restez maitre de votre futur* » rédigé par **Business & Decision**) :

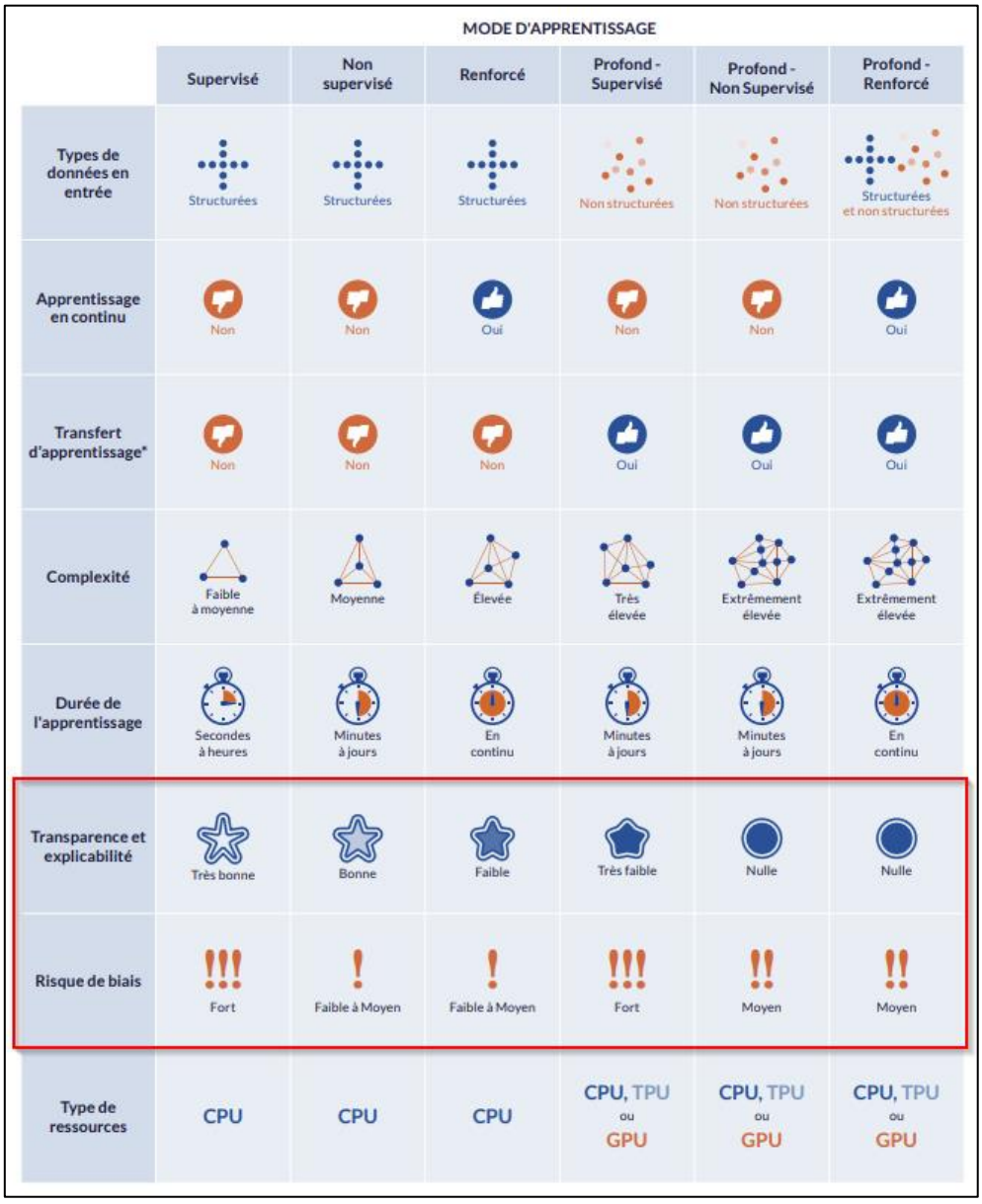


Figure 149 : Tableau comparatif des modes d'apprentissage et leurs impacts sur la transparence et les biais - Source : businessdecision.com

Dans les considérations **éthiques**, **certains modèles**, comme les arbres de décision, **sont nativement plus interprétables que d'autres**, comme les réseaux neuronaux profonds. La capacité de comprendre comment un modèle prend des décisions est cruciale pour établir la confiance et garantir la responsabilité.

Si le modèle est destiné à **des décisions critiques** (comme la médecine ou la justice), **il est impératif de choisir une architecture qui non seulement performe bien, mais qui est également transparente et justifiable.**

Les choix de modélisation pour maximiser les performances peuvent introduire ou amplifier des biais inhérents aux données. L'algorithme de Machine Learning lui-même, comme le **Random Forest**, peut favoriser certaines variables, pouvant entraîner des préjugés. De plus, certains systèmes d'IA exploitent des variables latentes qui peuvent cibler des sous-groupes spécifiques. **Il est donc essentiel, lors de la sélection d'un modèle, de considérer non seulement la performance mais aussi les biais potentiels, en recherchant un équilibre entre ces aspects**.

Enfin, il est possible de faire ce qui s'appelle du « **stacking model** » : l'idée est de combiner plusieurs modèles de prévision pour obtenir une meilleure performance prédictive que chacun des modèles individuels, surtout si les modèles de base sont très différents les uns des autres, car il peut capitaliser sur leurs forces respectives.

3.2.4.3.2 L'évaluation et la pertinence des KPI du Modèle

L'évaluation de la performance du modèle est cruciale avant sa mise en œuvre et durant toute sa durée de vie. Cette évaluation prévient la mise en service de modèles inadéquats, minimisant ainsi les risques associés. **Le choix de l'indicateur de performance (KPI)** est déterminant, et il varie selon la nature et l'objectif du modèle. Il est essentiel que le métier comprenne cet indicateur pour valider le modèle judicieusement.

Bien que nombreux, tous les KPIs ne reflètent pas nécessairement la performance réelle d'un modèle. Par exemple, pour une tâche de classification, l'exactitude pourrait être trompeuse, surtout si une classe est dominante. D'autres métriques comme **la précision, le rappel ou le F1-score** doivent être prises en compte selon le contexte métier.

Il est également vital d'évaluer la performance du modèle sur un ensemble de données différent de celui utilisé pour son entraînement pour détecter le surapprentissage. Si un modèle montre un bon résultat sur les données d'entraînement mais est moins performant sur les données de test, cela indique un problème. De plus, pour que l'évaluation soit valide, les données de test doivent être représentatives du contexte réel d'application du modèle.

En phase opérationnelle, **si les données diffèrent trop des données d'entraînement (ce qu'on appelle une dérive ou un drift), une phase de réentraînement peut être nécessaire**.

3.2.4.3.3 La construction du modèle

Il faut ajuster avec précision les hyperparamètres, veiller à ne pas accentuer des biais dans les résultats, s'assurer que les indicateurs de performance correspondent à l'objectif fixé et **documenter le processus** pour garantir sa transparence.

Dans les systèmes d'IA, **les hyperparamètres**, tels que le taux d'apprentissage ou le choix d'une fonction d'activation, **influencent significativement les performances et la robustesse d'un modèle**. Contrairement aux paramètres du modèle qui sont appris directement à partir des données, les hyperparamètres sont initialisés **à partir de connaissances préalables et ajustés via un processus itératif sur un ensemble de validation distinct des données d'entraînement**.

Cette calibration peut causer un « **target leakage** » (cas d'un dataset de test ou de validation accidentellement utilisée pour l'entraînement d'un modèle ce qui amène un biais). Pour atténuer ce risque, la validation croisée « **k-fold** » est souvent utilisée (ensemble des données est divisé en « k » sous-ensembles distincts, et le modèle est entraîné « k » fois).

Toutefois, la recherche d'hyperparamètres optimaux peut être chronophage. Pour optimiser ce processus, les « **learning curves** » sont utilisées pour visualiser les performances en fonction des hyperparamètres et détecter, par exemple, le sur-apprentissage en surveillant l'augmentation de l'erreur de validation. Les « **learning curves** » sont des graphiques qui montrent les performances d'un modèle en fonction de la quantité de données d'entraînement ou du nombre d'itérations d'apprentissage, permettant ainsi de visualiser comment l'erreur évolue sur les ensembles d'entraînement et de validation, et d'identifier des phénomènes tels que le surapprentissage ou des problèmes de convergence.

3.2.4.3.4 La transparence et l'explicabilité des modèles IA (appelé également XAI)

La transparence des modèles IA est assurée par des méthodes qui expliquent comment ils fonctionnent. Ces méthodes montrent comment chaque donnée entrante influence le résultat du modèle. Il y a deux principales manières d'expliquer les modèles :

- **Des techniques qui tentent d'interpréter la décision d'un modèle complexe dit « boite noire »**,
- **Des méthodes récentes qui créent des modèles naturellement explicables**. Jusqu'à récemment il fallait souvent choisir entre un modèle performant et un modèle facile à comprendre.

Il y a des outils (librairies) de référence pour l'explicabilité comme par exemple **LIME, SHAP ou encore ANCHOR**.

Malheureusement, plusieurs défis scientifiques et pratiques dans ce domaine nécessitent des efforts de R&D pour les surmonter.

Les entreprises commencent à explorer et à mettre en œuvre des techniques nativement explicables, telles que la « **Explainable Boosting Machine** » et des modèles Deep Learning avec des mécanismes d'attention pour offrir l'explicabilité.

Un cas concret illustré par le CEA (Commissariat à l'Energie Atomique et aux énergies alternatives), pour l'évaluation de plaintes émises par des consommateurs a combiné spécifiquement deux méthodes :

- Une approche d'apprentissage profond avec renforcement continue visant à créer un système hautement performant pour classer les requêtes.
- Un système de classification par clusterisation des données qui génère des explications sur les résultats obtenus.

Bien que ce dernier soit notablement moins efficace, son rôle consiste à fournir des explications à un professionnel *compétent, capable d'identifier les erreurs, de valider les explications* et donc de les associer aux réponses automatisées du système basé sur l'apprentissage profond lorsque nécessaire. Ainsi, la complexité théorique de l'explicabilité est en grande partie contournée par le contexte d'utilisation, qui n'exige pas souvent d'explications et demeure sous le contrôle humain.

L'enjeu d'assurer une explicabilité native des modèles d'Intelligence Artificielle représente un défi majeur pour les entreprises, sous diverses facettes :

- Cela répond aux **exigences** du futur règlement AI Act.
- Cette démarche favorise **l'adoption par les utilisateurs** en permettant aux professionnels en activité de saisir les critères ayant conduit aux décisions prises par l'Intelligence Artificielle.
- Cela renforce **la conception sécurisée** des modèles d'IA en instaurant un dialogue entre les experts en données et les experts métiers, autour des critères décisionnels employés par l'IA.

Dans ce contexte, **il devient impératif pour les entreprises de** :

- Élaborer des directives visant à **encourager l'intégration de techniques d'explicabilité** natives dans la phase de construction et de perfectionnement des modèles.
- **Implanter les outils et bibliothèques permettant cette explicabilité** : Il est possible d'utiliser la **probabilité d'une prédiction** pour évaluer la confiance d'un modèle en sa propre prévision. Cette information, exprimée comme un niveau de certitude (ex : élevé/moyen/faible), peut être partagée avec l'utilisateur. Une transparence sur le niveau de certitude du modèle renforce la confiance de l'utilisateur, surtout si les erreurs se produisent quand l'indicateur de confiance est faible.
- **Maintenir une veille proactive** sur les avancées de pointe en matière de techniques d'explicabilité.

- Se baser sur ces techniques d'explicabilité pour instaurer la confiance et, ainsi, susciter l'adoption des utilisateurs, tout en remettant l'IA au service de l'humain (par l'élévation de l'humain grâce à l'IA). **Cela permet à l'humain de comprendre les critères guidant les décisions de l'IA**, et ainsi, d'être en mesure de remettre en question les recommandations de l'IA.

Une transparence sur le niveau de certitude du modèle renforce la confiance de l'utilisateur, surtout si les erreurs se produisent quand l'indicateur de confiance est faible.

3.2.4.3.5 La validation des systèmes basés sur l'IA

La validation d'un modèle d'IA avant sa mise en production est essentielle pour garantir sa fiabilité et son efficacité.

Plusieurs considérations sont à prendre en compte :

- Une **analyse approfondie des risques** est nécessaire pour garantir la fiabilité de fonctionnement,
- Les problématiques de cybersécurité et de confidentialité nécessitent l'utilisation de diverses approches : **tests en boîte noire, boîte grise et boîte blanche**.
- L'évaluation de la robustesse inclut **l'examen de la stabilité face aux perturbations des entrées**.

La validation des systèmes IA exige une approche multidimensionnelle pour relever les défis liés à la sûreté, à la sécurité et à la fiabilité de ces systèmes.

Il est recommandé d'utiliser des outils de gestion de version pour conserver une trace des différentes itérations du modèle et des prétraitements. Cela permet de connaitre la version en cours de validation avant la mise en production et de pouvoir en cas de besoin faire un retour-arrière si les résultats ne sont pas satisfaisants.

Par ailleurs, si possible, **l'idéal est de tester le modèle dans un environnement simulant le plus fidèlement possible la production** (comme une **préproduction** par exemple) pour anticiper d'éventuels problèmes d'intégration ou de compatibilité.

Enfin, avant de déployer, **il est fort utile d'obtenir des retours de parties prenantes clés** (par exemple, des experts métier) pour avoir un aperçu de la pertinence et de la fiabilité du modèle dans un contexte réel sur cet environnement iso-production.

3.2.4.4 Phase 4 : le passage en Production

La phase de déploiement en production représente une étape importante dans le cycle de vie d'un modèle d'IA car étant en quelque sorte l'étape finale.

En effet, après des heures de conception, d'entrainement du modèle et de sa validation, cette phase marque le moment où le modèle est finalement intégré dans un environnement réel pour servir le ou les besoins pour lesquels il a été créé. C'est le moment de vérité où la théorie rencontre la pratique.

La transition vers la production nécessite une attention particulièrement poussée pour garantir que le modèle fonctionne comme prévu, tout en **répondant aux exigences de performance, de sécurité et d'évolutivité**.

Ce déploiement ne se limite pas à la simple intégration technique ponctuelle.

Au-delà de cette **mise en production (« MEP »)**, **cela impliquera également par la suite une surveillance continue, des mises à jour périodiques et une collaboration étroite avec les parties prenantes pour garantir que le modèle reste robuste et fiable dans son application**.

3.2.4.4.1 Le déploiement d'une nouvelle version d'un modèle

Suite à la création d'un modèle, tout comme pour n'importe quel service ou application IT, il est nécessaire de le déployer et de l'intégrer à l'infrastructure IT de production existante.

Cela permet son utilisation par différents départements, son intégration à des applications déjà en place ou sa mise à disposition en tant que service pour diverses applications ou secteurs.

Bien souvent, un modèle est élaboré sur une plateforme spécifique de Data Science. Par la suite, des équipes IT le prennent en main pour le lancer en production.

La mise en place d'un modèle s'articule autour de sa préparation pour intégration dans un environnement IT dédié à la production.

Les étapes typiques comprennent :

- **La préparation de l'environnement cible** destiné à accueillir le modèle,
- **La mise en place automatisée des flux, surtout ceux liés aux données** (comme les flux de connexion aux sources ou de prétraitement des données avant utilisation par le modèle),
- **La gestion des versions du modèle et de son paramétrage**,
- **Dans certaines situations, il peut être nécessaire d'adapter le modèle à l'environnement en raison de paramètres spécifiques**.

Une fois ces préparatifs terminés, le modèle est prêt à être déployé, généralement par l'équipe IT en charge de la production.

Plusieurs approches peuvent être adoptées, définies généralement en amont avec les responsables concernés, telles que le remplacement d'un modèle préexistant ou l'adoption de méthodes spécifiques comme les suivantes (liste non exhaustive) :

- **Canary** : le déploiement est fait de manière progressive touchant de plus en plus d'utilisateurs ou opérations. Dans cette stratégie, certains utilisateurs expérimentent ainsi le nouveau système en conditions réelles.

- **Shadow deployment** ou **Shadow mode** : l'ancienne modèle reste le modèle opérationnellement actif (principal), dont les résultats sont toujours utilisés. Le nouveau est déployé en parallèle mais ses résultats ne sont pas utilisés. Il s'agit d'un « **double-run** » qui permet de s'assurer de la robustesse et de la stabilité du nouveau modèle avant l'ouverture du service (bascule sur le nouveau modèle). Souvent, une telle mise en production s'appelle une « MEP technique ».

- **A/B Testing** : Pour garantir une mise en production sécurisée d'une nouvelle version, l'approche A/B Testing autorise le déploiement graduel sur un périmètre d'utilisateurs divisé en deux groupes (A et B). Cela permet de confirmer la performance optimale en conditions réelles et d'évaluer son amélioration effective avant un déploiement plus large à l'ensemble des utilisateurs. Cette méthode facilite également un retour à la version précédente en cas de baisse de performance (régression) sur la nouvelle version.

Voici une liste plus exhaustive des approches ou « *patterns* » de déploiement possible, avec leur explication :

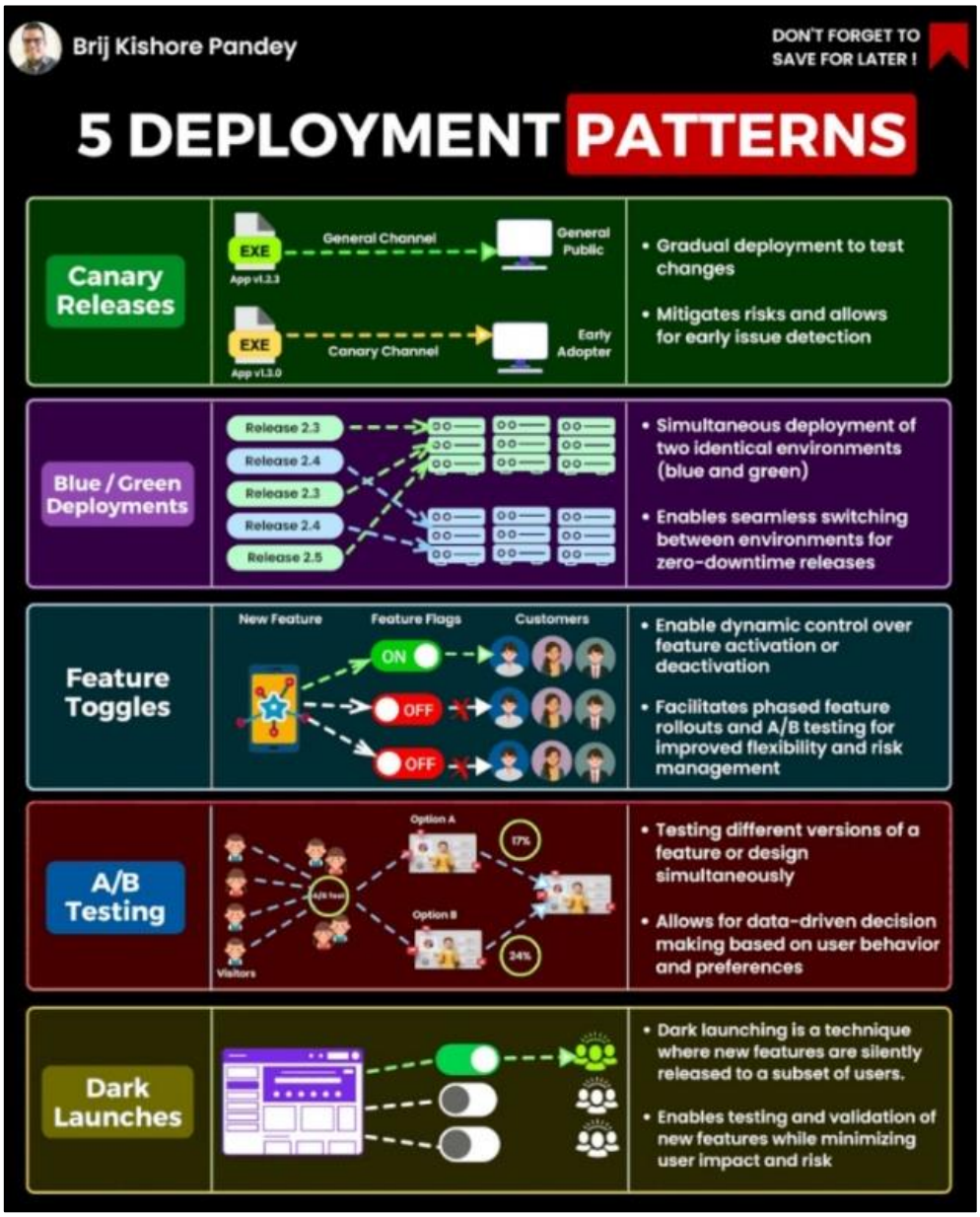


Figure 150 : Différents « patterns » de déploiement d'une IA en PRODUCTION - Source : linkedin.com - post de Brij Kishore Pandey, 14/08/2023

Quelques éléments clés sont à considérer lors du déploiement :

- Alors que nous mettons en avant la capacité d'apprentissage des algorithmes et leur évolution grâce aux retours des utilisateurs, il est essentiel de rappeler que l'algorithme a été formé sur des milliers d'exemples. De ce fait, quelques retours ne modifieront probablement pas immédiatement son comportement. En d'autres termes, « **l'IA apprend rapidement, mais nécessite de nombreuses corrections** ».

- L'envie de tester l'IA à petite échelle est courante. Cependant, cela ne devrait pas signifier qu'elle est déployée en dehors du processus existant. Sinon, les utilisateurs pourraient sentir que l'IA complique davantage leur travail plutôt que de le simplifier.

➢ **La vérification de la non-régression :**

C'est une prolongation du suivi des évolutions des modèles, se focalisant davantage sur la précision de la performance du modèle. Elle s'assure que, pour certains périmètres de données, il n'y a pas de baisse de la qualité prédictive.

➢ **Le contrôle de la cohérence des prédictions du modèle :**

Il est important de vérifier que la mise à jour du modèle (nouvelle version) produit des résultats pareils pour des prédictions passées. Un modèle donnant des résultats différents pour une même tâche (en raison d'une évolution dans les données d'entraînement ou de la conception du modèle) pourrait poser problème tant en terme commerciale et confiance des utilisateurs.

➢ **La gestion du risque en Cybersécurité :**

o **Robustesse des modèles IA face aux attaques adverses**

Les modèles accessibles depuis l'extérieur de l'entreprise peuvent être la cible d'attaques malveillantes (adverses), soit pour détourner le fonctionnement standard du modèle ou pour collecter des informations stratégiques économiques.

Voici quelques types d'attaques contre des systèmes d'IA (listes non exhaustive) :

- ***Poisoning*** : L'attaquant altère les données d'apprentissage pour influencer les résultats du modèle, en supposant qu'il ait accès à ces données,
- ***Oracle*** : En interrogeant constamment le modèle, l'attaquant cherche à en déduire des détails sur le modèle IA ou sur les données d'entraînement,
- ***Evasion*** : Ayant une certaine connaissance du modèle ou de sa conception, l'attaquant ajuste les données en amont pour que le modèle donne un résultat inattendu.
- ***Extraction (inference attack)*** : Une attaque de découverte de données cherche à extraire des informations sensibles d'un modèle d'IA en exploitant sa mémorisation involontaire des données d'entraînement. Les attaquants peuvent avoir des informations sensibles sur les données d'origine en interrogeant habilement le modèle.

Pour défendre les modèles d'IA, les entreprises peuvent adopter diverses techniques, telles que l'ajout de bruit aux prédictions, la limitation du taux de requêtes, ou l'utilisation d'authentifications plus strictes pour les API de prédiction.

- **Tromper le modèle / attaques adverses**

Les modèles d'IA peuvent réagir anormalement face à de légères modifications dans les données. On le voit notamment lorsqu'un simple motif sur un t-shirt peut rendre quelqu'un indétectable à l'IA ou avec les modèles de vision par ordinateur : un peu plus de bruit dans une image peut empêcher la reconnaissance d'un objet. Lors de la conception, il est essentiel d'anticiper ces risques et de réaliser des tests de robustesse aux différentes attaques.

- **Duplication du modèle (clone)**

Il est démontré qu'avec un nombre restreint de requêtes, il est possible de reproduire le comportement d'un modèle d'IA (le cloner). Une stratégie pour atténuer ce risque est **d'introduire de légères inexactitudes dans certaines prédictions ou de fournir un peu de prédictions non générées par le modèle** afin de minimiser cette menace.

- **Pillage du modèle (vol)**

Le pillage fait référence à l'obtention non autorisée des détails exacts d'un modèle, tels que ses architectures, ses poids et ses hyperparamètres. Cela peut être réalisé en accédant illégalement aux serveurs ou aux bases de données où ces détails sont stockés.

Une fois le modèle pillé, il peut être copié, distribué ou utilisé sans autorisation**. C'est une violation directe des droits d'auteur ou des droits de propriété intellectuelle** ou autrement dit, du **vol de « copyright » d'algorithme.**

Pour pouvoir identifier facilement une copie illicite, le « **watermarking** » (ou « ***tatouage numérique*** » en français) est une technique qui peut être utilisée pour insérer des informations cachées.

L'objectif principal du **watermarking** est de prouver la propriété et de détecter les copies non autorisées, via par exemple une question secrète que seuls les créateurs connaissent.

Ghassen BETTAIEB
Data Scientist
(CACIB)

La sécurité et la confidentialité des données :

Certains modèles peuvent mémoriser les données d'apprentissage sur lesquelles ils ont été entrainées. Face aux attaques adverses, les modèles peuvent révéler des données sensibles. Il est vital de protéger ces données contre les fuites et de s'assurer que le modèle apprend sur les patterns de données plutôt que sur les données elles-mêmes pour qu'ils puissent ensuite faire des prédictions.

Lors de l'utilisation de données réelles, en particulier celles contenant des informations sensibles, des conventions sont signées au début du projet pour garantir la sécurité et la confidentialité. Les données sont stockées de manière sécurisée, et les modèles sont conçus pour ne pas mémoriser ces informations sensibles. Une fois le projet terminé, les données sont conservées pendant une durée définie (par exemple, un ou deux ans) avant d'être supprimées.

3.2.4.4.2 La traçabilité et les audits

Les systèmes d'IA doivent répondre à certaines exigences, notamment en matière d'auditabilité et de traçabilité. Ces critères sont particulièrement mis en avant dans diverses recommandations sur l'IA et seront probablement intégrés dans la prochaine version du réglementation européenne AI Act.

Nous savons maintenant que les résultats des modèles d'IA peuvent varier en fonction des données sur lesquelles ils ont été entraînés. Mais surtout les modèles doivent être en mesure d'assurer une reproductibilité des résultats pour montrer qu'ils sont fiables.

L'objectif est surtout de démontrer à un régulateur la reproductibilité d'un modèle et d'assurer une collaboration fluide entre Data Scientists, permettant de comprendre et d'améliorer les prédictions basées sur les travaux précédents.

En cas de contrôle, il est nécessaire de :

- **Démontrer la fiabilité des résultats des modèles d'IA** :
  Pour atteindre cet objectif, il ne suffit pas de se baser sur les méthodes traditionnelles de DEVOPS pour le versioning du code. Il faut adopter une approche MLOPS qui prend également en compte le versionnement des données d'entraînement ainsi que les versions des modèles créés.
- **Assurer une traçabilité et suivi d'audit** :
  Afin de garantir une auditabilité optimale, il peut être pertinent de développer un suivi d'audit détaillé. Ce dernier devrait indiquer pour chaque prédiction :
  - La version spécifique du modèle utilisé,
  - Les détails explicatifs de la prédiction y compris la probabilité et le degré de confiance,
  - Le répertoire des modèles archivés pour, si nécessaire, reproduire le résultat.

Il est aussi envisageable d'obtenir une certification ou un label par l'intermédiaire d'un audit externe, comme **LNE, LabelIA ou GEEIS-AI** cités précédemment.

Voir le témoignage de **BNP Paribas** à ce sujet en annexes (lien « Traçabilité chez BNP Paribas »).

3.2.4.4.3 Focus sur l'audit des algorithmes

L'audit de algorithmes pourrait être réalisé par des auditeurs externes, des experts internes ou des autorités régulatrices. Cette vérification vise à renforcer la confiance dans les systèmes automatisés, détecter d'éventuels soucis et y apporter des solutions pour assurer une utilisation des algorithmes plus équitable, transparent et responsable.

L'approche de l'examen des algorithmes évolue constamment. Face à la complexité technique et juridique de l'audit des algorithmes, ce projet est essentiel. À l'avenir, de nouvelles mesures pourraient être adoptées en France pour renforcer le contrôle et la régulation des algorithmes.

- **Audit externe :**

Un des auditeurs externes reconnu par l'ACPR est l'**INRIA**, établissement de recherche et d'enseignement supérieur qui coordonne les travaux scientifiques prévus par la stratégie nationale pour l'Intelligence Artificielle de 2017 et la Direction Générale des Entreprises.

Initié par l'INRIA, le projet **Régalia** vise à mettre en place des procédures d'audit pour les algorithmes. La durée d'**un audit peut aller de trois jours à plusieurs mois selon la complexité du système d'IA étudié**.

Ces algorithmes, souvent considérés comme des « boîtes noires », peuvent parfois fonctionner de manière biaisée, même sans intention malveillante. Ils sont difficilement auditables : **Auditer un algorithme est en réalité un véritable défi**.

INRIA utilise alors une technique particulière appelée audit « en boîte noire », où elle modifie certaines variables pour voir comment les résultats changent.

Si un biais est détecté, le corriger n'est pas simple. En effet, **retirer une variable problématique peut affecter la performance de l'algorithme**. Une solution possible serait **d'appliquer un « skin » de réparation à l'algorithme**, en observant sa réaction avec des données similaires et en ajustant sa réponse en conséquence.

L'objectif de Regalia est d'identifier ces biais et de garantir une utilisation équitable et transparente des algorithmes. **Benoît ROTTEMBOURG**, responsable du projet, collabore avec divers experts pour développer un protocole d'audit. Il souligne que, contrairement aux comportements humains, un algorithme peut être corrigé. Le projet fait partie des efforts européens pour réguler l'utilisation des algorithmes et garantir leur conformité aux normes éthiques.

- **Audit interne :**

Il est très important de faire régulièrement des audits internes sans faire appel à un organisme externe. Ces audits peuvent être demandés par les utilisateurs finaux « métier », l'équipe projet, le comité éthique ou bien même l'inspection générale de la banque.

Après avoir identifié et priorisé les cas d'usage d'IA selon le risque et les enjeux stratégiques, le processus d'audit interne commence.

Trois acteurs clés sont impliqués :

- Les audités (composés d'un expert technique et d'un expert métier),
- Les auditeurs (un Data Scientist et un Data Consultant),
- Les instances de gouvernance.

L'audit se concentre sur l'intégration du modèle d'IA dans le processus métier global.

Le processus débute par une **réunion de lancement** pour informer tous les participants. Ensuite, les auditeurs cherchent à comprendre le cas d'usage lors de la « **prise de connaissance** ». Les audités remplissent ensuite un questionnaire pour « **l'auto-évaluation** », fournissant des « **preuves** » pour chaque action réalisée. Ces preuves sont ensuite analysées par les auditeurs. Après cette analyse, un « **plan de remédiation** » est coconstruit pour améliorer le cas d'usage. Enfin, les résultats et le plan de remédiation sont présentés aux instances de gouvernance lors de la « **restitution** ».

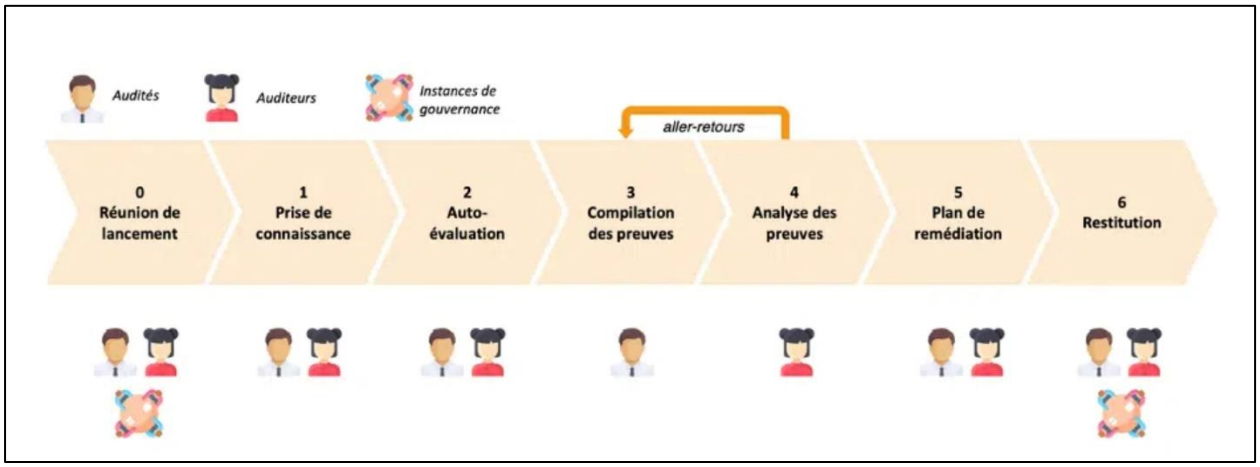


Figure 151 : Processus d'audit interne IA responsable pour un cas d'usage donné - Source : quantmetry.com

L'audit interne permet de mettre en évidence des lacunes que ce soit au niveau de la mise en œuvre par les équipes techniques ou l'adoption par les équipes « métier », telles que la mauvaise compréhension des prédictions, la faible performance du modèle en production, les discriminations (non-équité), etc.

Cela aide donc à identifier des solutions appropriées pour toutes les parties prenantes, l'IA responsable faisant le lien entre la maîtrise technique de l'Intelligence Artificielle et son utilisation par les équipes métiers.

3.2.4.5 Phase 5 : la maintenance et monitoring du modèle

Lorsqu'un modèle est déployé et en service, il est essentiel de vérifier régulièrement qu'il répond toujours aux attentes.

Des équipes combinées IT et métier sont mobilisées pour cette tâche. Au-delà de la surveillance IT qui garantit la disponibilité du service, il est également vital de confirmer que les prédictions du modèle restent adéquates pour l'entreprise et ses utilisateurs.

Il est donc important d'avoir de la surveillance de la robustesse des modèles : **L'objectif est de contrôler la qualité des modèles en production pour pouvoir réagir en cas de défaillance ou de baisse de la qualité de l'IA**. Enfin, la « **frugalité** » de l'IA et particulièrement sa consommation de ressources est un point à surveiller car a un impact direct sur les performances et éventuellement les coûts.

De nombreux gains sont possibles via le monitoring :


Robustesse


Confiance


Conformité


Green AI

Il y a en réalité trois types de monitoring qui doivent être mis en place que nous allons aborder par la suite :

- Le monitoring de la dérive des données ou **monitoring non supervisé**
- Le monitoring de la dérive des performances des modèles ou **monitoring supervisé**
- Le monitoring de la consommation de ressource pour une **IA « verte »** (ou « ***green AI*** »)

3.2.4.5.1 Monitoring de la dérive des données ou monitoring non supervisé

Il s'agit de surveiller les variations de la distribution des données d'entrée. Cela peut se faire en temps réel, mais il faut déterminer si une dérive observée est liée à une dérive réelle.

Les modèles sont entraînés à partir d'un jeu de données d'entrainement. Si les valeurs ou la distribution des données changent, la performance du modèle peut diminuer car il ne connaît pas ce nouveau contexte.

Le monitoring non supervisé est souvent utilisé pour détecter la dérive des données. En utilisant des techniques d'analyse des anomalies ou de détection de nouveauté, on peut identifier quand les données d'entrée commencent à dévier de ce qui est considéré comme « normal » ou « habituel ».

Une solution est d'avoir **un réentrainement automatique**. De plus, suivre la dérive des données peut identifier quand le modèle a besoin d'être ajusté.

Par exemple, si des données historiques ne sont plus adaptées à la situation actuelle, il pourrait être essentiel **d'instaurer des mécanismes limitant la portée des données historiques pour écarter les informations obsolètes ou d'adopter des stratégies d'apprentissage en continu** afin d'exploiter au mieux les données les plus récentes.

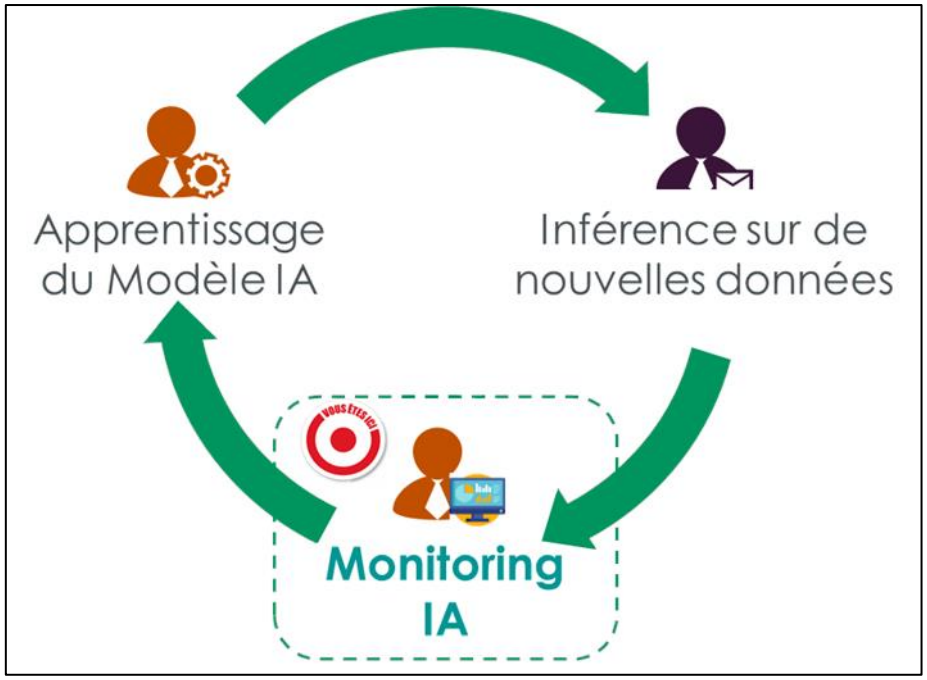


Figure 152 : Apprentissage en continu pour éviter les dérives de données - Source : Datalab Groupe C.A.

3.2.4.5.2 Monitoring de la dérive des performances des modèles ou monitoring supervisé

Il s'agit ici de mesurer la performance réelle du modèle pour détecter les dérives réelles. Cependant, cela nécessite d'avoir accès aux labels réels. Ce monitoring est à faire après chaque entrainement.

La dérive de modèle se produit lorsque la performance du modèle se dégrade au fil du temps, souvent à cause d'un changement dans la relation sous-jacente entre les entrées et les sorties.

Le monitoring supervisé peut être employé pour détecter la dérive du modèle. En comparant régulièrement les prédictions du modèle à des résultats réels (par exemple, en utilisant un ensemble de test tenu à l'écart), on peut déterminer si la performance du modèle se dégrade par rapport à un seuil acceptable.

Il est donc essentiel de surveiller les performances en production en continu, surtout après chaque réentrainement, et si possible, d'évaluer la qualité des résultats de l'IA (comme une comparaison entre la prédiction et l'observation réelle par exemple)

3.2.4.5.3 Monitoring de la consommation de ressource pour une IA « green »

Comprendre la consommation d'énergie d'un modèle en production est plus que jamais nécessaire notamment à une époque où la réduction de l'empreinte carbone est un enjeu mondiale majeur. Et il est important de s'y intéresser le plus tôt possible car cela permettra d'anticiper une éventuelle exigence réglementaire qui pourrait s'imposer dans le futur.

Un des éléments ayant un impact direct sur la consommation de ressources est le choix du modèle, **notamment en raison de sa complexité mais aussi de la quantité de données qu'il traite**.

La mise en place d'un outil de mesure d'émission de CO2 et de monitoring du modèle est alors pertinente dans un souci de « **bien-être sociétal et environnemental** », un des principes majeurs de l'AI Act.

Cet outil considérera l'émission de CO2 comme une métrique standard de performance du modèle, qui donnera une indication sur l'équilibre entre consommation énergétique et performance.

Le cas échéant, ce KPI permettra de justifier une revue du modèle pour trouver des axes pour dans l'idéal réduire la consommation énergétique tout en conservant sa performance.

L'usage d'un modèle plus énergivore peut être justifié si par exemple un impact financier significatif est à craindre s'il n'est pas utilisé. Mais dans le cas où un modèle consomme 100 fois plus d'énergie pour seulement 1% de performance supplémentaire, son choix devra avoir une justification solide.

Recommandations :
La mesure de l'émission carbone doit être une métrique de performance standard. Elle doit vous permettre d'envisager une revue de votre projet pour en faire une "green AI".

Les axes possibles d'améliorations pour atteindre cet objectif sont les suivants :

- Un modèle moins complexe (plus simple) pour le besoin attendu
- Une utilisation éventuelle d'un nombre moins important de variables (choisir les plus pertinentes)
- • Une utilisation de ressources IT plus optimisées et plus « green » (GPU vs CPU, infrastructure dédiée vs mutualisée, etc.) : consommer plus dans le cloud public a un impact direct sur les coûts liés à l'infrastructure et donc au projet.

➢ **Témoignage du DataLab Crédit Agricole :**

Le **DataLab Crédit Agricole** a développé une plateforme de monitoring appelé MonIA pour surveiller les performances des modèles, les données (comme la détection de dérive), l'utilisation professionnelle (comme le taux de traitement) et les ressources de calcul, les concepts (comme les variations des variables clés). **Elle vise à fournir des alertes automatiques** et à faciliter la détection précoce des problèmes, permettant la prise des décisions concernant la correction des données et le réapprentissage avant de mettre en place des nouveaux modèles.

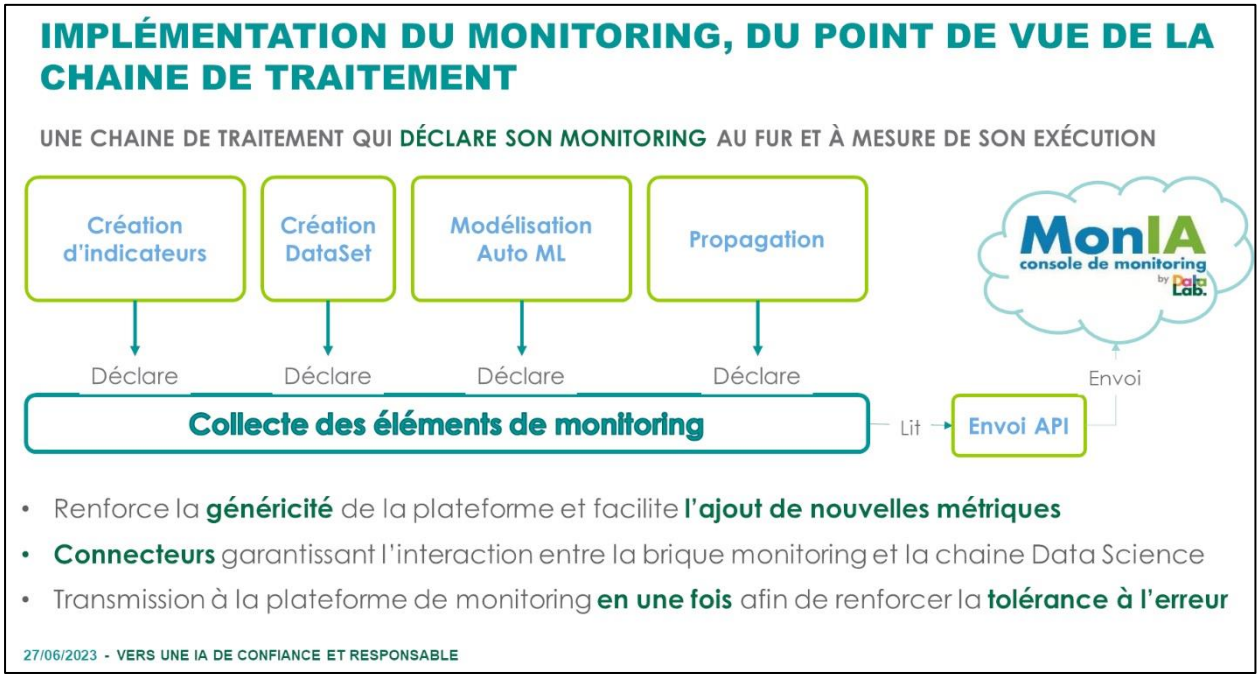


Figure 153 : Solution de monitoring du Datalab Groupe Crédit Agricole - Source : Datalab Groupe C.A.

3.2.4.6 Transverses aux phases : la gouvernance du système d'IA

La gouvernance d'un système d'IA se réfère aux structures, politiques, procédures et normes utilisées pour gérer, diriger et surveiller les systèmes d'Intelligence Artificielle.

Elle fournira les lignes directrices nécessaires pour une utilisation éthique et efficace de l'IA.

Voici quelques types de gouvernance à prendre en considération pour un système d'IA responsable :

- **« Technique »** : Elle doit aborder toutes les problématiques et considérations à prendre en compte sur les thématiques de la qualité des données, la maintenance des modèles et la sécurité de ces derniers. Elle couvre également des aspects plus « infrastructure » qui seraient spécifiques aux systèmes d'IA,
- **« Ethique »** : Cette gouvernance doit couvrir la vie privée, la transparence, l'équité,
- **« Réglementaire » :** Elle couvre la conformité (« compliance ») et la responsabilité juridique,
- **« Organisationnelle » :** Ce type de gouvernance définit la stratégie de formation et de sensibilisation autour des enjeux « IA responsable », la surveillance et les audits,
- **« Internationale »** : Si applicable dans l'organisation, doit couvrir les normes et accords existants pour guider l'utilisation responsable de l'IA partout dans le monde.
- **« Participative » :** couvre les thématiques de l'engagement des parties prenantes (en build ou en run), les feedback attendu des utilisateurs « métier » et les procédures d'amélioration continue des systèmes d'IA

Recommandations :
La gouvernance est un facteur majeur de succès. Un système d'IA, même en production, nécessite une gouvernance constante sur tous les aspects évoqués car celle-ci est essentielle pour maximiser les avantages de l'IA tout en minimisant les risques potentiels, notamment ceux éthiques et réglementaires

3.2.4.7 Conclusion sur les bonnes pratiques pour une application par phase

En conclusion, il s'agissait ici de donner de bonnes pratiques et recommandations qui pourraient être suivi lors de la mise en place d'un système d'IA depuis son démarrage. Si le projet a déjà démarré, il est toujours possible de les appliquer dans le cadre d'évolution ou pour les prochaines phases.

Compte tenu des différences entre les projets et du caractère parfois spécifiques des recommandations, cette liste ne doit pas être considérée comme exhaustive. Il est en effet fortement probable de découvrir d'autres recommandations au-delà de celles suggérées ici.

3.2.5 Des perspectives du futur de l'IA responsable dans la banque

L'IA responsable dans le secteur bancaire est effectivement un sujet d'actualité en raison des préoccupations croissantes liées à l'éthique, à la transparence, à la sécurité des données et à la conformité réglementaire à venir prochainement.

Si précédemment, il a été ici objet de bonnes pratiques et de recommandations pour faire en sorte qu'un système d'IA soit « responsable », il est plus que légitime de se poser les questions suivantes : **A quoi les organisations, en particulier bancaires, peuvent s'attendre dans les mois et années à venir ? Qu'est-ce que tout cela va changer ? que pourraient faire les banques pour aller plus loin ?**

Dans la partie 1, il était évoqué le fait que l'être humain ne pouvait en aucun cas déterminer le futur avec certitude. Néanmoins, il peut toutefois exprimer des suppositions.

Voici donc quelques perspectives d'un futur supposé pour l'IA responsable dans le secteur bancaire :

- **Renforcement de la réglementation** :

Les régulateurs du monde entier sont susceptibles d'introduire des directives plus strictes concernant l'utilisation de l'IA dans le secteur bancaire, en particulier pour garantir la transparence, l'équité et la protection des consommateurs.

- **Transparence accrue :**

Les banques pourraient adopter des modèles d'IA plus transparents et explicables pour gagner la confiance des clients et des régulateurs, en particulier pour des applications critiques comme l'octroi de crédits ou la détection de la fraude.

- **Ethique et équité :**

Les institutions financières pourraient mettre en place des comités d'éthique dédiés pour superviser le développement et le déploiement de solutions d'IA, afin de s'assurer qu'elles ne perpétuent pas de biais ou de discriminations.

- **Formation et sensibilisation :**

Les banques pourraient investir davantage dans la formation de leurs employés sur les principes de l'IA responsable, afin de garantir une utilisation éthique et conforme des technologies d'IA.

- **Collaboration intersectorielle :**

Les instituts bancaires pourraient collaborer en travaillant main dans la main avec d'autres secteurs, des universités et des organismes de recherche pour développer des normes et des meilleures pratiques en matière d'IA responsable.

- **Technologies d'IA centrées sur l'humain :**

L'adoption de technologies d'IA qui placent l'humain au centre de la prise de décision, comme les systèmes « **human-in-the-loop** », pourrait devenir plus courante pour garantir une prise de décision équilibrée.

- **Responsabilité et reddition de comptes :**

Les banques pourraient mettre en place des mécanismes robustes pour assurer la responsabilité en cas d'erreurs ou de problèmes causés par des systèmes d'IA, y compris des mécanismes d'audit et de surveillance.

- **Innovation durable :**

L'IA responsable pourrait également englober des préoccupations environnementales, poussant les banques à adopter des solutions d'IA plus écoénergétiques et à soutenir des initiatives liées à la finance verte.

- **Engagement des parties prenantes :**

Les instituts bancaires pourraient impliquer davantage les clients, les employés et d'autres parties prenantes dans les discussions et les décisions liées à l'IA, afin de garantir une approche holistique et inclusive.

Si certaines de ces perspectives devaient être absolument mises en œuvre, il serait primordial de privilégier en priorité **la mise en place un comité d'éthique** pour assurer la mise en œuvre de solutions d'IA responsable **et la collaboration intersectorielle** pour développer l'IA responsable de manière globale, en France et dans le monde, l'IA responsable étant une problématique multisectorielle internationale.

Pour conclure, l'avenir de l'IA responsable dans le secteur bancaire semble prometteur notamment avec cette prise de conscience croissante et globale de la nécessité d'adopter des pratiques éthiques, transparentes et équitables.

Il est important de garder à l'esprit que les banques qui adoptent ces principes pourraient non seulement **gagner la confiance de leurs clients et des régulateurs**, mais aussi se positionner comme des leaders dans l'adoption responsable de l'IA de manière générale.

A elles de jouer pour la construction de **l'IA de demain : responsable, éthique et digne de confiance** !

Conclusion

Dans la première partie de cette thèse, il était important pour moi de vous faire découvrir, ou tout simplement rappeler, ce qu'est l'Intelligence Artificielle, contribuant en quelque sorte à démystifier cette technologie laquelle a été rendu accessible à tous les publics récemment notamment avec la démocratisation de l'usage de IA de traitement du langage naturel tel que ChatGPT.

En effet, le constat à travers différentes sollicitations dans mon entourage, aussi bien dans le cercle professionnel que privé, est que tout le monde semble évoquer l'IA mais peu peuvent réellement l'expliquer. Il s'agissait donc de s'assurer que la définition et le fonctionnement d'une IA soit expliqué et partagé afin que nous ayons un même niveau de compréhension et qu'il soit possible pour tout le monde de comprendre les parties suivantes.

Ces informations ont donc donné les bases nécessaires pour comprendre, dans une deuxième partie, la définition d'une IA responsable et ses enjeux. Et grâce à eux, la nécessité de mettre en œuvre des Intelligences Artificielles responsables n'est plus à démontrer.

En effet, bien que les avancées technologiques autour de l'IA soient fulgurantes, elles sont le résultat d'une course à l'innovation mais elles ne garantissent en rien une utilisation sans dérapage. Les géants technologiques tentent de garantir qu'ils « implémentent » des systèmes d'IA infaillibles mais il n'existe pas de tels systèmes. Pire, rien ne permet à ce jour de le garantir officiellement, la « parole » de ces « big techs » faisant seul foi.

Les débats autour des limites et des dangers de cette technologie sont de plus en plus nombreux et les dérapages ou incidents en perpétuel augmentation (690% de plus qu'en 2010). Si les craintes autour de l'IA augmentent, cela amène la défiance en cette technologie ce qui est l'opposé de ce qui est souhaité, ou tout simplement la confirmation que ces débats et réflexions autour de l'IA sont nécessaires. La prise de conscience collective est de plus en plus large, progresse rapidement et de nombreux acteurs convergent vers un avis commun : il manque un cadre. Les initiatives se multiplient dans le monde, à l'OCDE, à l'UNESCO et également en Europe à travers les travaux de la Commission Européenne et ceux des états-membres tels que la France.

Suite au focus réalisé dans cette partie sur l'AI Act proposé par la Commission Européenne, nous avons compris ce qui est attendu : L'évaluation du niveau de risque d'un système d'IA entrainera un ensemble d'exigences à vérifier lesquels sont listées dans le futur règlement de l'AI Act. Ce dernier définit également les sanctions pour les organisations dont les systèmes d'IA ne respecteraient pas ces exigences. C'est ce qui est attendu mais pas comment y arriver.

C'est donc dans une troisième partie que j'ai cherché à donner une démarche, un mode d'emploi, à suivre pour viser la mise en place d'une IA responsable, en particulier dans le secteur de la banque en vous permettant de :

- Vous auto-évaluer : celle-ci vous donnera une image de la maturité « IA responsable » de votre système d'IA existant
- Vous informer sur les certifications ou labels existants : le but est de formaliser vos démarches en obtenant une reconnaissance officielle, indice de confiance auprès de vos différents partenaires ou clients
- Vous organiser et suivre un ensemble de bonnes pratiques pour vous assurer que vos systèmes d'IA soient justes, équitables, transparents et explicables, sans discrimination et soucieux de l'environnement.

Vous trouverez ci-dessous une synthèse de mes recommandations sous forme graphique, pour mettre en place une IA responsable en particulier dans le milieu bancaire. **Il est l'heure d'anticiper l'arrivée du futur règlement européen. A vous !**

Synthèse des recommandations

Annexes

Bibliographie

Livres blancs

- GROUPE D'EXPERTS INDEPENDANTS DE HAUT NIVEAU SUR L'INTELLIGENCE ARTIFICIELLE
- Législation sur l'intelligence artificielle-Parlement européen 2019-2024
- IA DE CONFIANCE OPPORTUNITÉ STRATÉGIQUE POUR UNE SOUVERAINETÉ INDUSTRIELLE ET NUMÉRIQUE
- L'IA Ethique en pratique
- Tech Sprint sur l'explicabilité des algorithmes d'intelligence artificielle
- Contrôle des risques des systèmes d'Intelligence Artificielle
- Responsible AI Working Group Report
- Towards the engineering of trustworthy AI applications for critical systems

Vidéos

- https://www.youtube.com/watch?v=cQ54GDm1eL0 : Exemple de deepfake
- https://www.youtube.com/watch?v=VYrcXTckP2s : Interview de Yoshua BENGIO
- https://www.youtube.com/watch?v=D5VN56jQMWM : Rendez-vous Coiffeur par une IA
- https://dl.acm.org/doi/10.1145/2770869 : « Rise of concerns about AI » (dangers de l'IA)

Webinars

Les replays des conférences ci-dessous sont disponibles sur le site : innovationshowbyca.com :

- La révolution de l'IA et du Machine Learning - CISCO
- Vers une IA industrielle, de confiance et responsable - DataLab Crédit Agricole
- IA documentaire de dernière génération, performante et frugale - DataLab Crédit Agricole
- L'IA générative au service de la finance - Google
- L'IA pour révolutionner le calcul scientifique - Microsoft
- L'IA d'aujourd'hui et demain : du Machine Learning classique en passant par les modèles de fondation jusqu'au Quantique - IBM

Livres

Howard GARDNER , *Les intelligences multiples*, 1983 & réédition en 1993

Daniel KAHNEMAN, *Thinking, Fast & Slow*, 2011

Grégoire COLOMBET, *Humain ou IA ? Qui décidera le futur ?*, 2023

Aurélie JEAN, *Les algorithmes font-ils la loi ?,* 2021

Mes articles #MBAMCI

Mon premier article : Les limites de l'intelligence Artificielle et ses dérapages

Mon deuxième article : L'IA Responsable

[Mon troisième article : La data dans l'IA responsable]

Mon podcast #MBAMCI

[Mon podcast Green AI sur Let's Go Digital #MBAMCI]

Témoignages

Certification chez Crédit Agricole

➢ Crédit Agricole conjugue IA industrielle innovante et responsable

Le **Crédit Agricole** vise à allier performance et responsabilité en matière d'intelligence artificielle. Dans cette optique, son **Datalab Groupe** a entrepris une double action : d'une part, obtenir une certification pour le processus de création de solutions IA, et d'autre part, décrocher un label d'IA responsable en intégrant des critères RSE supplémentaires. Une bon moyen de s'adopter durablement cette approche et de gagner en maturité.

Le **Datalab Groupe** a obtenu en février 2023 la certification de processus pour l'IA du **L**aboratoire **N**ational de métrologie et d'**E**ssais (**LNE**) comptant environ 100 exigences, permettant de répondre aux enjeux de maîtrise des risques et de confiance (n° 38916 rev. 0, lne.fr).

Le Datalab Groupe du Crédit Agricole devient ainsi la troisième structure en France et la première de son secteur à décrocher cette certification, les deux autres structures sont une Biotech et un cabinet de conseil spécialisé en « IA for Good ».

Cette certification est la reconnaissance de l'excellence de sa démarche de fabrication de solutions IA industrielles. Elle couvre toutes les étapes du cycle de vie d'une solution IA, du cadrage au maintien en conditions opérationnelles. Depuis, tous les projets IA du Datalab Groupe, en cours ou à venir, respecte dorénavant ce niveau d'exigence.

➢ Comment s'est déroulée cette certification ?

Le dispositif de certification du LNE et constitué

- D'un **référentiel d'une centaine d'exigence** qui traite de transparence, d'explicabilité, de robustesse, de performance …
- D'un **audit de certification initial** sur site de 2 jours et demi avec un chef de mission et un expert technique du LNE analysant respectivement la qualité de la méthodologie et des développements qui doit être illustré par des projets. un audit de diagnostic d'une journée peut-être proposé en option avant cette audit initiale.
- De **deux audits annuels de suivi** pour porter le bénéfice de la certification sur 3 ans.

La certification est l'aboutissement d'un travail collectif de fond au cours duquel l'ensemble des équipes de Datalab Groupe :

- Se sont appropriées le référentiel d'exigence du LNE,
- Ont adapté la méthode projet afin de répondre à toutes ces exigences,
- Ont appliqué cette nouvelle méthode sur trois solutions IA déjà proposées en mode « *as a service* » aux entités afin de servir d'illustration lors de la certification.

- **Les mesures pratiques pour avoir cette certification**

Les mesures pratiques pour atteindre cet objectif ont formé un projet dédié à l'amélioration constante et durable des techniques et des procédures au niveau des projets et de la réalisation. Sur le terrain, cela a impliqué la formation d'une équipe centrale et l'établissement de routines pour suivre la progression des tâches. Les initiatives ne se limitent pas aux changements méthodologiques.

Le DataLab Groupe a, par exemple, développé et mis en œuvre une plateforme de surveillance, un élément crucial pour avoir confiance dans les systèmes d'IA. « Cette plateforme de surveillance a été enrichie pour que les trois produits d'IA corrigés dans le cadre de la certification soient surveillés selon les critères les plus stricts », explique le Chief Data Officer du Groupe Aldrick Zappellini.

Ces nouvelles directives internes sont mises en œuvre pour tous les projets à partir de la date de certification. Pour les produits internes, les corrections sont envisagées au cas par cas.

- **Est-ce que ça vaut vraiment le coup de se faire certifié ?**

Pour le Crédit Agricole, la réglementation est en route et il est donc important de s'y préparer. Pour l'équipe du Datalab Groupe, c'était une expérience bénéfique et ont pu repérer les points à améliorer ou la direction à prendre pour atteindre l'objectif.

Même si parler de ces labels aux clients du grand public n'est pas toujours simple, leur vrai avantage est pour les employés du Crédit Agricole, surtout dans les domaines de l'IA ou des métiers car l'automatisation et l'IA peuvent inquiéter.

En plus, le Datalab Groupe vise l'exemplarité en anticipant, à travers cette certification, les exigences de la future réglementation européenne IA ACT et en poursuivant pour une labellisation RSE (LabelIA Labs) afin notamment de renforcer l’alignement de ses réalisations avec le projet sociétal du groupe.

S'appuyant sur ces expériences, le Datalab Groupe proposera au travers de l’IA Factory Groupe, un accompagnement aux entités qui souhaitent une aide à la mise en conformité de leur solutions IA aux exigences des réglementations relatives au modèles IA.

Labellisation chez Crédit Mutuel Arkéa

Le Crédit Mutuel Arkéa a décidé de suivre un processus de labellisation dans le cadre de sa transformation vers l'IA. Après avoir augmenté l'utilisation des modèles de ML et créé le Data Office, le groupe a décidé d'évaluer sa maturité. Les objectifs étaient de comparer leurs méthodes aux normes du marché, de renforcer la sensibilisation et la formation sur l'IA de confiance, et d'améliorer leurs processus de développement.

Un audit externe a été réalisé sur trois produits en fonction de deux initiatives :

- Le Label IA : centré sur les meilleures pratiques pour une IA respectant les enjeux humains et environnementaux.
- La Certification IA du LNE : axée sur les processus documentés et les outils de gestion des risques.

Suite à cet audit et à d'autres analyses, une deuxième phase a été lancée. Elle vise à :

- Promouvoir une IA de confiance : en lien avec l'engagement ESG du groupe, une politique de données responsables a été mise en place, ainsi que des directives pour le développement de l'IA.
- Améliorer les processus et la gestion des risques : en intégrant davantage l'IA dans les dispositifs de risques internes et en envisageant une certification LNE pour les activités liées à l'IA sur les 4 processus certifiables LNE lesquels sont :

Processus certifié	Définition	Exemples de processus audités
Conception	Transformer une expression de besoin en spécifications fonctionnelles	Spécification et prise en compte des exigences normatives et réglementaires
Développement	Traduire ces spécifications en une version de la fonctionnalité d'IA prête à être évaluée	Apprentissage Qualité des Banques De Données.
Evaluation	Vérifier la conformité du système aux spécifications définies avant son déploiement	Définition des protocoles d'évaluation, des métriques, sur l'ensemble des outils d'évaluation qui permettent de rendre compte de l'efficacité de ces systèmes intelligents.
Maintien en conditions opérationnelles	Assurer la conformité de la fonctionnalité d'IA aux spécifications définies après son déploiement et tout au long de sa phase d'exploitation	Toutes les caractéristiques propres au maintien de ces conditions opérationnelles. Les systèmes d'IA peuvent évoluer tout au long de leur vie avec des dérives et des dégradations de performance.

Figure 154 : Processus certifiables LNE - Source : lne.fr

Labellisation chez Banque de France

En 2021, la Banque de France a commencé à travailler sur une IA de confiance et en moins d ‘un an, a défini un cadre de travail pour la conception, le développement et l’utilisation des systèmes d’intelligence artificielle. En parallèle, elle a également signé une charte internationale pour une IA inclusive (GEEIS) qui a été déclinée en une politique interne pour mettre en place une IA inclusive.

Ces éléments définis et mis en place, il leur a semblé pertinent de faire une évaluation par un tiers externe afin d’identifier les points forts que le potentiel d’amélioration. En effet, un regard externe reste irremplaçable. Le label GEEIS a permis à la banque de France de reconnaitre la qualité du travail de ses équipes mais a aussi pu identifier des axes de progrès à faire.

Se préparer pour le label GEEIS a demandé un gros travail. Vu le champ couvert par le référentiel GEEIS, la banque a dû rassembler des experts de différents secteurs : le centre spécialisé en IA, mais aussi les ressources humaines, surtout pour la formation, et la déontologie. Grâce au soutien de la direction générale, cette préparation a permis à l’équipe en charge de dresser un bilan précis et de déterminer ensemble ce qui est déjà en place, ce qui doit l'être. C'est une base solide pour élaborer un plan d'amélioration.

Labellisation chez Orange

Aujourd’hui, **Orange** est engagé dans 2 démarches de labellisation IA : la société a adopté une approche d'IA inclusive, soutenue par une charte et le label GEEIS-AI (qui se concentre sur l'égalité des genres). Cette initiative, menée par ses équipes RSE, renforce leur engagement de longue date en faveur de l'égalité professionnelle et de la diversité au sein de ses équipes.

L'objectif est de prévenir les biais dans les systèmes d'IA qui pourraient avoir des conséquences négatives ou discriminatoires. Pour cela, orange met en œuvre diverses actions : embaucher des équipes diversifiées, former tout le monde sur ces enjeux, établir une gouvernance appropriée et intégrer ces considérations dans nos projets, notamment en RH, et dans ses interactions avec les partenaires sociaux.

De plus, l’entreprise participe à l'initiative « positive.ai », lancée par un groupe d'entreprises dont **BCG**, **Malakoff Humanis**, **L'Oréal** et elle-même. Avec le soutien de leurs dirigeants et de leurs équipes data-IA, cette initiative vise à promouvoir une approche pragmatique de l'IA responsable. Elle propose un label basé sur un audit en deux parties : la gouvernance globale de l'IA et les systèmes d'IA individuels. L'ambition d’Orange est d'élargir ce consortium.

Cadre méthodologique chez Société Générale

Société Générale, en collaboration avec deux autres banques françaises sous l'égide du **Hub France IA**, a élaboré un cadre méthodologique pour gérer les risques associés aux systèmes d'IA. Ce cadre identifie trois lignes de défense :

- La première ligne, comprenant les propriétaires et développeurs de modèles, est chargée de la gestion du risque.
- La seconde ligne, composée des équipes de gouvernance et de revue, définit les normes et évalue les modèles importants.
- La troisième ligne, l'audit interne, évalue l'efficacité globale du dispositif.

Dix risques majeurs ont été analysés en détail, allant de l'inadéquation entre l'IA et les besoins métiers à la mauvaise définition de la gouvernance. Les résultats sont publiés dans un livre blanc sur le site du Hub France IA.

Comité éthique chez Orange

Orange a instauré un comité éthique incluant des membres extérieurs. Elle a également élaboré une charte qui va plus loin que la charte de l'IA inclusive déjà ratifiée et a structuré la gouvernance avec des référents éthiques dans chaque pays.

Ces référents sont chargés de mettre en œuvre la stratégie éthique au niveau groupe, d'orchestrer les formations, de tenir à jour l'inventaire des systèmes d'IA de l'organisation et d'évaluer leur risque éthique.

Orange a par ailleurs conçu un questionnaire pour évaluer les risques éthiques tout au long du cycle de vie d'un système d'IA et l'a testé sur des projets axés sur l'innovation. La société a pu relever certains obstacles à la mise en œuvre de ce questionnaire lesquels sont :

- Réticence des équipes à compléter un autre questionnaire en plus de ceux concernant la protection des données, la sécurité, la RSE, d'où l'importance d'une gouvernance intégrée.
- Difficulté à comprendre certaines questions puisque l'éthique de l'IA est un domaine naissant, ce qui nécessite une formation appropriée pour les équipes.
- Questions concernant la pertinence et la portée du questionnaire. Bien qu'il soit assorti de recommandations, le questionnaire n'est pas auto-suffisant. Les réponses servent avant tout au référent éthique pour obtenir une vision d'ensemble, confirmer les niveaux de risque associés aux projets et prioriser les actions.

Orange a aussi élaboré une classification des applications avec les risques associés (en particulier pour les employés et les clients) pour guider ses équipes. En ce qui concerne cette gouvernance, ses démarches ont été exposées lors de l'audit réalisé pour le renouvellement du label GEEIS-IA, qui a lieu tous les deux ans.

Traçabilité chez BNP Paribas

BNP Paribas conserve les détails de ses expérimentations, incluant les données d'entraînement, le modèle et ses méta-paramètres. Elle utilise **Apache MLFlow**, crée par **Databricks**, pour trois fonctions majeures :

- Suivre et consulter les expérimentations : code, données, configuration et résultats,
- Package les expériences pour garantir leur reproductibilité,
- Offrir un format standard pour la transition des modèles vers différents outils de déploiement.

L'objectif est de démontrer au régulateur la reproductibilité d'un modèle et d'assurer une collaboration fluide entre Data Scientists, permettant de comprendre et d'améliorer les prédictions basées sur les travaux précédents.

Interviews

Questions des 14 interviews

Voici les questions générales posées aux 14 personnes interviewés mais certaines questions se sont ajoutées selon les personnes interviewées.

- Quel poste occupez-vous ?
- Intervenez-vous sur des sujets autour de l'IA ?
- Pouvez-vous parler de quelques cas d'usage ?
- Certains projets d'IA sont en PROD, avez-vous déjà rencontré des cas d'erreur ou de dérapage ?
- L'IA responsable vous parle ? Et Comment la définiriez-vous?
- Qu'est-ce que vous mettez en place pour que vos systèmes d'IA soient responsables ?
    - En termes de données : comment vous garantissiez le respect de la vie privée et la gouvernance des données ? (par exemple le fait qu'elles ne soient pas biaisées)
    - En termes de fiabilité (robustesse) technique et la sécurité : comment garantissez-vous que vos systèmes d'IA soient fiables et sécurisés ?
    - En termes de diversité, non-discrimination et équité : comment garantissez-vous que vos systèmes d'IA soient conformes aux réglementations et aux normes éthiques ?
    - En termes de durabilité et de responsabilité écologique : quelle mesure mettez-vous en œuvre pour garantir une IA respectueuse de l'environnement ?
    - En termes de transparence : vos systèmes d'IA expliquent-ils les décisions prises ? pour être compréhensibles pour les utilisateurs ?
    - En termes de responsabilité juridique : avez-vous défini clairement la responsabilité de la banque en cas d'erreur ?
    - En termes de contrôle humain : avez-vous prévu un bouton stop pour arrêter le système d'IA ?

- Connaissez-vous les principes de l'UE sur l'IA responsable ?
- Comment évaluez-vous votre niveau d'application de ces principes sur vos systèmes ?
- Il existe des certifications : êtes-vous certifié ou labelisé ?
    - Si oui, par quel organisme ?
    - Si non, envisagez-vous de le faire sous combien de temps ? Avec quel organisme ?
- Selon vous, quels sont les avantages de l'IA responsable pour les entreprises?
- Selon vous comment mettre en place une IA responsable ? Quel serait le processus et les points d'attention ?

Interview Benoit ROTTEMBOURG (INRIA)

Benoit est en charge de "Regalia" à l'Inria depuis trois ans. Le projet a pour but d'aider les autorités régulatrices à surveiller que les algorithmes respectent la loi, notamment pour détecter ceux qui trichent ou introduisent des biais.

Regalia, qui se concentre principalement sur le développement d'algorithmes destinés à superviser d'autres algorithmes. Vous pouvez nous considérer comme la "police des algorithmes", même si cette description est plutôt audacieuse. Nous nous voyons plus comme des médecins légistes humbles, essayant de vérifier si toutes les composantes sont à leur place et s'il n'y a pas d'actes suspects.

Les clients de Regalia :

- Autorités de régulation comme la CNIL, l'ARCOM, l'ACPR, etc.

Exemple : Si l'ACPR veut s'assurer que la Société Générale ne prend pas en compte le genre dans l'octroi de crédits, Regalia pourrait créer un algorithme pour les aider.

- D'autres entités comme la DGCCRF utilisent des algorithmes pour repérer des anomalies, comme de faux avis sur Internet.

Objectifs de Regalia :

- Aider les autorités de régulation en fournissant des outils et des méthodologies.

- Collaborer avec des entreprises souhaitant s'autoréguler, notamment pour suivre des normes éthiques qui vont au-delà de la loi.

Acteurs concernés :

- Banques et assurances : sensibles à cette approche car habituées à la régulation, avec des équipes de data science encore récentes.

Critères de Regalia :

1. Disparate impact sur des variables sensibles.

2. Recherche de biais généraux, par exemple, une catégorie de clients discriminée.

3. La métrique utilisée peut varier pour considérer l'impact réel et le préjudice causé.

Ensuite, l'IA Act, qui n'est pas encore convergé, pose aussi d'autres questions de sand boxies regulatory sandboxies…. Donc là on pourrait imaginer qu'une entreprise doit avoir un service d'audit ou de conformité algorithmique . la question, est-ce que les l'IA pourrait permettre à des employés de frauder, ou de mieux frauder, de donner des crédits à leur familles, ou de défavoriser les Pakistanais ou de favoriser les Portugais…..

j'ai senti que les banques et les assurances sont parmi les entités qui sont les plus sensibles à ce discours parce que

- elles sont habitués à être régulées

- elles ont déjà des autorités de contrôle interne dont elles ont un peu peur en général, qui surveillent.

- Les équipes de data Science dans les banques et assurances sont encore assez récentes.

Ces 3 raisons là font que la Banque et l'assurance sont parmi les endroits qui, passeront des mots à l'acte le plus le plus vite possible.

Vers une certification :

- Il n'y a pas encore de norme ou de certification existante. Mais les autorités pourraient éventuellement recommander ou exiger que certains prestataires de services soient utilisés pour des audits.

Exemples :

- Un algorithme de planification a, sans le vouloir, défavorisé les personnes dont le nom commence par les dernières lettres de l'alphabet.

- Des problèmes potentiels avec les algorithmes de plateformes de livraison comme Deliveroo où les livreurs peuvent ressentir une "punition" algorithmique.

Le processus de réparation :

- Lorsqu'un biais est détecté, un mécanisme devrait permettre de remonter l'information, enquêter et corriger.

Les biais algorithmiques :

- Biais sur les plateformes de vente : La "Buy Box" d'Amazon a tendance à privilégier les produits d'Amazon. Un site web pourrait favoriser certains produits d'assurance basés sur la marge bénéficiaire plutôt que la pertinence pour le client.

- Biais d'influence : Les plateformes de vidéos peuvent enfermer les utilisateurs dans des bulles en les dirigeant vers du contenu toxique.

- Difficulté à définir les biais : Il est souvent complexe pour un client de déterminer les biais qu'il souhaite surveiller. L'identification des biais basés sur le genre peut devenir compliquée, surtout lorsqu'on se base sur des proxies comme les prénoms.

- Distinction entre éthique et légalité : Tandis que le domaine juridique a des règles claires, l'éthique est plus subjective. Par exemple, des salons de coiffure peuvent avoir des prix différents pour les hommes et les femmes, ce qui est considéré par certains comme discriminatoire.

- Différences juridiques internationales : Le droit commercial varie d'un pays à l'autre. En France, par exemple, la tromperie commerciale est considérée comme un délit pénal, tandis qu'elle est traitée comme une question civile dans d'autres pays européens.

Robustesse des algorithmes :

La robustesse fait référence à la sensibilité d'un algorithme face à de petites variations dans les données d'entrée. Par exemple, si un petit changement de salaire peut influencer la décision d'accorder ou non un crédit.

Cas de biais dans une banque en ligne : Une banque en ligne utilisait la qualité de l'image pour authentifier les clients, discriminant ainsi ceux avec de vieux téléphones. Cela pourrait indirectement discriminer les personnes à faible revenu qui n'ont pas de smartphones modernes.

Cas de biais dans les services de livraison de repas : Une étude a montré que les prix de livraison variaient en fonction du quartier de résidence. De plus, certaines plateformes mettaient en avant les restaurants nouvellement ajoutés malgré le manque d'avis pour inciter les utilisateurs à essayer ces nouveaux établissements.

Comment éviter les biais dans les algorithmes ?

Lorsqu'un algorithme, en voulant être trop optimal, oublie certains éléments essentiels et finit par franchir certaines limites, cela est appelé un "dérapage". Il s'agit non pas d'une tromperie délibérée, mais plutôt d'un excès d'enthousiasme ou d'une négligence de la part de ses créateurs. Un exemple notable est celui de Deliveroo en Italie : leur système de notation des livreurs tenait compte des jours d'arrêt maladie, conduisant à une condamnation pour négligence grave.

Au sein des banques, des biais peuvent aussi se manifester. Les systèmes de relance pour les clients susceptibles d'être en déficit à la fin du mois peuvent avoir des conséquences différentes selon les segments de clients. Un SMS de relance envoyé à une personne âgée pourrait la paniquer, tandis qu'un jeune pourrait l'ignorer. Si un algorithme apprend ces réactions et agit en conséquence, il risque d'amplifier un biais existant, en traitant différemment les personnes âgées et les jeunes, par exemple.

De plus, les algorithmes peuvent renforcer des stéréotypes ou clichés culturels. Par exemple, en France, si un algorithme distingue des comportements typiques de personnes issues de certaines régions ou descendances, il pourrait catégoriser injustement et renforcer ces clichés.

En marketing, le biais peut également être exploité pour augmenter les conversions. Par exemple, un camping peut tester si la présentation d'une image de toboggan (attirant pour les familles) ou de montagnes (attirant pour les randonneurs) sur leur site web augmente le taux de clic. Bien que cela

puisse être efficace d'un point de vue commercial, cela révèle également comment les algorithmes peuvent catégoriser rapidement les utilisateurs en fonction de stéréotypes.

Robustesse

La variabilité de la sortie par rapport à l'entrée. Dire si les entrées bougent pas beaucoup, par exemple, si Je refuse un crédit à une personne mais avec 10€ de plus sur son salaire, j'accepte le crédit, ça veut dire que pour à 10€ près par mois (considérer comme négligeable toutes choses égales par ailleurs) donc à 10,00€ près est ce que l'algo peut lui refuser un crédit alors il y aura bien un cas où ça sera vrai à la limite de la limite si t'as tout sauf le salaire qui est un trop petit. on peut tout à fait imaginer que du coup là robustesse de l'algorithme qui a un impact finalement sur un disparaître impact , on peut imaginer que les 2 sont liés,

Les tests de robustesse c'est une autre famille de problèmes très importants dans l'industrie, très important dans du systémique, c'est un peu plus continu alors que nous on est très vite discret.

Est-ce éthique d'utiliser des algorithmes pour détecter des patterns?

Les algorithmes peuvent détecter des motifs subtils non perceptibles par l'œil humain. Par exemple, ils peuvent remarquer que les gens négocient plus les prix le jeudi que le vendredi, car ils préfèrent finir tôt pour le week-end.

Mesures pour éviter les biais:

1. Pres process : Il s'agit de manipuler les données avant qu'elles ne soient traitées. Par exemple, équilibrer les données pour éviter un déséquilibre entre hommes et femmes ou entre différentes races. Cela peut aussi consister à retirer certaines variables problématiques.

2. In process : Pendant le traitement, on peut ajouter des mesures punitives à l'algorithme si certaines contraintes ne sont pas respectées. C'est une forme de régularisation mathématique.

3. Post process : Après que l'algorithme ait fait ses prédictions, on peut modifier les résultats. Par exemple, pour un algorithme de crédit, on pourrait poser la même question pour une femme et un homme avec des conditions similaires et s'assurer que les réponses sont comparables.

L’IA responsable :

L'efficacité des algorithmes d'IA et leur impact potentiel sur la société soulèvent des questions éthiques et pratiques. Voici un résumé de vos réflexions sur ce sujet :

1. Cas d'Usage : L'exigence pour un algorithme varie en fonction de son application. Les attentes pour une plateforme vidéo diffèrent de celles pour un agent de crédit ou pour un système d'admission universitaire.

2. Biais Géographique : Les algorithmes de détection de cellules cancéreuses peuvent être biaisés en fonction de la région. Par exemple, un algorithme formé sur des données de Parisiens pourrait ne pas être aussi efficace pour les Soudanais en raison des différences démographiques et environnementales.

3. Responsabilité : Il est crucial d'anticiper les problèmes potentiels et d'informer les utilisateurs des limites d'un algorithme. Si un algorithme est principalement formé sur des données de Parisiens, il devrait être clairement indiqué qu'il pourrait ne pas être aussi précis pour d'autres populations.

4. IA Responsable : L'IA responsable vise à anticiper et à atténuer les préjudices potentiels. Si une entreprise est consciente des limites de son algorithme et l'utilise de manière appropriée, elle est jugée plus responsable.

5. Explicabilité vs Efficacité : Bien que l'explicabilité soit souvent demandée, elle peut parfois entrer en conflit avec l'efficacité. Dans certains cas, sauver des vies peut être plus important que de comprendre parfaitement comment fonctionne un algorithme.

Responsabilité juridique

il faut remonter toute la chaîne, la data, l'éditeur l'entreprise et même le déployer. Il faut que toute la chaîne ait un bout de la responsabilité chacun à son niveau.

Donc, la data est-elle convenable par rapport à la promesse qui est fait par l'éditeur ? On voit bien que si je promets que je guéris tous les hommes, enfin que je détecte le cancer sur tout le monde mais que j'ai jamais eu de de subsaharien, l'entreprise est coupable

mais celle qui s'en sert de cet algorithm. Si elle le déploie sur un marché où il y a beaucoup d'Africains. C'était écrit qu'on n'était pas bon avec les Africains, on n'a pas le droit.

Et puis celui qui déploie aussi ce qu'il a mis en place comme les effets indésirables des médicaments. On peut pas lâcher un médicament sans avoir un système de collecte des effets indésirables et on peut pas lâcher sur le marché des algorithmes sans une mécanique de collègues, des effets indésirables.

Donc, c'est la responsabilité de celui qui déploie, l'éditeur, l'entreprise doivent avoir une part de responsabilité chacune, et qui est une boucle de rétroaction. Il faut créer les conditions de ce dialogue dans ces trois acteurs, c'est extrêmement difficile parce qu'on achète loin et c'est de l'ordre de la traçabilité en fait.

Ce n'est pas facile même techniquement, c'est pas facile et avec les IA Génératives c'est encore moins facile.

Et pour moi un algo comme un forecast, par exemple, devrait sortir avec sa qualité des tortures tests et des biais possibles qu'on sait évident par rapport aux données et des biais imaginable.

Je n'ai pas assez testé sur les Soudanais ou sur les subsahariens donc attention, je n'ai que des Asiatiques dans ma base. Attention quand on achète un open source qui a été fait sur des données, je

ne sais pas quoi allemandes, y a des Allemands sur certains sujets qui se comportent pas du tout comme les Français.

On part du risque quand même Parce que à la fin on va pas non plus l'empêcher tout le monde de travailler, donc on part du risque pour avoir une analyse bénéfice risque tout le temps en. Sinon on fait du principe de précaution, puis on travaille plus, donc pour moi pour moi y a toujours un risque. Au départ on découpe les acteurs, c'est encore plus compliqué en IA générative parce que y a plein d'acteurs et on s'assure qu'il y a un feedback loop entre, c'est la personne qui subit l'algorithme et les maillons de la chaine qu'on produit l'algorithme

L'audit des algorithmes :

L'audit d'un algorithme dépend de divers facteurs comme le rythme d'innovation et la saisonnalité. Un audit régulier, préventif et longitudinal est recommandé, en particulier pour les algorithmes touchant un grand nombre de clients.

Auto-audit et audit tiers : Les entreprises devraient avoir leurs propres processus d'audit interne et être également soumises à des audits par des tiers de confiance.

Le principal message est que, tout comme dans d'autres domaines où la sécurité et l'efficacité sont primordiales (comme les médicaments ou les tests de grossesse), les algorithmes et l'IA devraient être soumis à des standards élevés de surveillance, d'audit et de responsabilité pour garantir leur bon fonctionnement et la confiance du public.

comme dans d'autres domaines où la sécurité et l'efficacité sont primordiales (comme les médicaments ou les tests de grossesse), les algorithmes et l'IA devraient être soumis à des standards élevés de surveillance, d'audit et de responsabilité pour garantir leur bon fonctionnement et la confiance du public.

Green IA :

L'impact environnemental des technologies de l'information, en particulier des modèles d'intelligence artificielle sophistiqués, est devenu un sujet de préoccupation croissante. Voici un résumé de vos réflexions sur ce sujet :

1. Coûts d'Entraînement : Les coûts associés à l'entraînement de modèles d'IA, en particulier des modèles génératifs, sont nettement plus élevés que ceux des modèles simples comme les régressions linéaires. Cet écart est d'autant plus marqué avec l'évolution des modèles génératifs.

2. Coûts d'Usage : Au-delà de l'entraînement, il est important de considérer les coûts associés à l'utilisation de ces modèles. La question se pose de savoir si les coûts d'utilisation d'un modèle complexe sont également disproportionnés par rapport à un modèle simple.

3. Sobriété Numérique : Avec l'augmentation des coûts associés à l'IA, il pourrait y avoir un mouvement vers une sobriété ou une frugalité numérique. Les entreprises pourraient être évaluées non seulement sur leur impact social, mais aussi sur leur "empreinte numérique".

4. Comparaison des Impacts : À mesure que les technologies évoluent, il pourrait y avoir un équilibre entre l'impact environnemental de l'utilisation d'un modèle d'IA sophistiqué et d'autres activités, comme la conduite d'une voiture.

5. Question de Priorité : Bien que l'impact des e-mails ou de l'IA puisse sembler négligeable aujourd'hui, cela pourrait changer à mesure que la puissance de traitement et la complexité des modèles augmentent. Cependant, d'autres activités, comme les vidéos publicitaires en ligne, pourraient avoir un impact environnemental plus important.

6. Mesure et Transparence : Il est essentiel de disposer de métriques claires pour évaluer l'impact environnemental des technologies. Lorsque cet impact n'est plus négligeable, il pourrait être nécessaire d'encourager l'utilisation d'alternatives plus respectueuses de l'environnement.

La sécurité des systèmes d'IA :

La sécurité des algorithmes d'IA est une préoccupation majeure à mesure que ces technologies se démocratisent. Voici un résumé des différents aspects de sécurité que vous avez évoqués :

1. Malveillance Interne : Il s'agit de scénarios où des individus au sein de l'entreprise, connaissant bien l'algorithme, cherchent à le manipuler à des fins frauduleuses. Ces actes relèvent généralement de l'inspection générale.

2. Empoisonnement du Modèle : Cette technique consiste à fournir intentionnellement des données erronées ou trompeuses à un algorithme afin de le dégrader ou de le faire fonctionner de manière imprévue. Sans protections adéquates, un algorithme peut être vulnérable à ce type d'attaque.

3. Pillage du Modèle : Il s'agit de copier ou de voler un algorithme depuis l'extérieur. Cela équivaut à une violation du droit d'auteur de l'algorithme. Des mesures doivent être mises en place pour limiter le nombre de requêtes qu'un utilisateur peut poser à l'algorithme, afin de prévenir ce type de vol.

4. Découverte de Données via Interrogation Malveillante : En interrogeant astucieusement un algorithme, il est possible de deviner certaines des données qui ont été utilisées pour le former. Cela s'appelle une "attaque par extraction de données". Un exemple cité concerne un traducteur de roumain vers le français qui avait été formé avec des données confidentielles. En analysant les traductions fournies par l'algorithme, il était possible de déduire certaines des données d'origine.

Comment faire pour assurer la sécurité :

La protection des algorithmes contre le piratage est un sujet émergent et crucial à mesure que l'IA devient de plus en plus omniprésente. La start-up "Shield" chez INRIA travaille sur cette problématique en utilisant le concept de "watermarking" pour les algorithmes. Voici un résumé de ce que vous avez partagé :

Shield, une start-up affiliée à INRIA, développe des méthodes pour protéger les algorithmes contre le piratage. Une de ces méthodes est le "watermarking" d'algorithme, similaire à la protection des images contre la contrefaçon. Cette technique consiste à intégrer des points de données spécifiques et secrets dans un algorithme. Par exemple, dans un algorithme de crédit immobilier, un point de données anormal, comme un salaire précis de 3201€ pour une femme, pourrait être intégré. Si un crédit est accordé uniquement pour ce montant précis, mais refusé pour des montants légèrement différents comme 3200€ ou 3202€, cela indique la présence du watermark.

Cette technique peut être utilisée pour détecter si un algorithme a été volé ou copié. Si une autre banque, comme la Société Générale, utilise un algorithme qui réagit de la même manière à ce point de données secret, cela pourrait indiquer qu'ils ont volé l'algorithme du Crédit Lyonnais.

Bien que cette approche ne soit pas encore standard et reste à la frontière de la recherche, elle pourrait devenir plus courante à l'avenir, notamment pour tester la robustesse des algorithmes. Des comportements suspects, comme des requêtes répétées avec de légères variations, pourraient être détectés et signalés comme potentiellement malveillants.

Interview Gilbert HANNOUCHE (CACIB)

L'IA Factory CACIB a mis en place 25 cas d'usage en 2022 et a pour objectif d'en mettre en place 50 en 2023.

Les cas d'usages sont réalisés via des algorithmes développés chez CACIB ou des bibliothèques publiques externes (open sources)

Problématiques rencontrés :

- Gagne en productivité et efficacité au travail. : Automatisation des taches

- Buisness : identifications, de d'opportunités ou de deal.

Le sujet aujourd'hui qu'on a, c'est, on peut développer le meilleur outil au monde, mais si on accompagne pas le changement et l'adoption par les utilisateurs, bah en fait on développe quelque chose qui sera pas utilisé.

Le ROI, retour sur investissement, est très important parce que on a payé des centaines et des centaines de de milliers d'euros pour que, si à la fin il n' y a personne qui l'utilise alors que ça peut apporter plein de choses ça sera un problème

On n'a pas d'erreur : l'alog apprends et c'est l'amélioration en continue

Interview Patrick BEZOMBES (AFNOR)

La machine n’est pas intelligente, c’est l’humain qui est derrière

Voici un résumé des points principaux abordés :

Autonomie :

Bien que l'IA soit souvent considérée comme "autonome", elle n'est pas intrinsèquement autonome. Elle fonctionne sur la base des programmes et des instructions qui lui sont donnés.

Responsabilité :

Si une machine ou un système est déclaré "autonome" et qu'un accident survient, la responsabilité pourrait incomber au constructeur du système, car l'utilisateur n'aurait rien fait pour causer cet accident. Cette idée de responsabilité est particulièrement pertinente dans des domaines tels que l'aviation, où la détermination de la responsabilité après un accident peut impliquer plusieurs parties.

Réglementation et certification :

Avec l'augmentation des systèmes basés sur l'IA, il devient nécessaire d'avoir des réglementations claires. Ces réglementations ne devraient pas seulement se concentrer sur la sécurité, mais aussi prendre en compte d'autres aspects tels que la non-discrimination, l'égalité et la démocratie. De plus, la certification par une tierce partie pourrait être un moyen de gérer les risques associés à l'IA.

Anthropomorphisation de l'IA : Il y a une tendance à attribuer des caractéristiques humaines à l'IA, ce qui peut être source de confusion. L'IA n'est pas "intelligente" de la manière dont les humains le sont, elle est plutôt capable de traiter des informations rapidement et efficacement.

Problème de définition de l'IA : L'IA a toujours été difficile à définir précisément. Sa définition a évolué au fil du temps, avec des termes tels que "cybernétique" utilisés avant l'adoption du terme "intelligence artificielle".

Confiance : La notion de confiance est culturellement et linguistiquement nuancée. La manière dont la confiance est construite et perçue diffère selon les régions, comme illustré par les différences entre les États-Unis, la Chine et l'Europe. De plus, des termes tels que "trust" et "confidence" ont des nuances différentes selon les langues.

Perspective de Patrick Bezombes : il a partagé son expérience de travail avec le Ministère des Armées et a évoqué les défis associés à l'utilisation de l'IA dans les applications militaires. Il a également souligné les dangers potentiels de l'adoption précipitée de technologies basées sur l'IA, comme le remplacement des services d'assistance téléphonique par des systèmes automatisés.

Interview Mohamed BEJI (CACIB)

Voici un résumé des points principaux abordés :

Problématiques rencontrés :

- La capacité de prédire le futur en s'appuyant sur des jeux de données existants pour générer des insights.

- Fluidification de l'accès à l'information à l'aide de l'IA, en rendant les données accessibles via des questions en langage naturel.

- Des modèles formés qui, lors de leur mise en production, s'avèrent inadéquats pour certains types de données car ils n'ont pas été formés sur ces types de données.

Nous n'avons pas rencontré des erreurs, l'humain a toujours la main

Recommandations/Mesures pour le monitoring :

Développer une approche méthodologique pour la formation, la mise à l'essai et la mise en production des modèles afin d'assurer leur amélioration continue.

- Éviter 100% des erreurs n'est pas possible, mais l'utilisation d'outils et de méthodologies appropriés, tels que MLOps, peut aider à identifier et à corriger rapidement ces erreurs.

Exemples :

- Les modèles prédictifs peuvent être biaisés s'ils sont basés sur des données historiques affectées par des événements sans précédent, comme la pandémie de COVID.

- Dans l'analyse de presse, les modèles peuvent mal fonctionner s'ils rencontrent des thèmes qu'ils n'ont jamais vus auparavant.

Outils utilisés pour le monitoring :

- Les outils tels que Dataiku et ML Flow sont utilisés pour le monitoring des modèles.

Ethique et responsabilité :

- L'IA "verte" recommande de réduire les paramètres du modèle pour le rendre plus économe en énergie.

- Les modèles ne doivent pas discriminer sur la base du genre ou d'autres caractéristiques personnelles.

- Les données personnelles doivent être traitées avec soin. Le masquage des entités nommées est recommandé avant la formation des modèles.

- Les modèles basés sur des données personnelles peuvent être biaisés et contrevenir au RGPD.

- Les Data Scientists doivent jouer un rôle de conseil et éduquer les autres métiers sur l'éthique de l'IA.

Sécurité et robustesse :

Des comités tels que AIDA et MRM chez CACIB sont en place pour assurer la conformité et évaluer les aspects éthiques.

Les Data Scientists doivent anticiper les dégradations potentielles des performances des modèles et s'assurer qu'ils peuvent être améliorés continuellement.

Explicabilité :

Il est conseillé d'éviter d'utiliser des modèles de type "boîte noire" pour des tâches critiques, car ils peuvent donner des réponses sans explication.

Responsabilité :

Il devrait toujours y avoir une intervention humaine dans le processus de prise de décision, même après une analyse IA.

Point d'attention principal :

L'éducation et la formation sont essentielles pour assurer une utilisation éthique et efficace de l'IA. Il est également essentiel de comprendre et de maîtriser les données sur lesquelles les modèles sont formés.

Interview Ludovic GIBERT (IMA)

Usage dans la banque :

L'IA offre une solution innovante pour automatiser des tâches qui étaient auparavant considérées comme trop complexes pour une automatisation traditionnelle. Voici un résumé de cette perspective :

L'IA permet d'automatiser des processus qui étaient auparavant difficiles à gérer en raison de la complexité des règles nécessaires. Ces règles, susceptibles de changer avec le temps, rendaient l'automatisation traditionnelle inefficace. De plus, l'IA est particulièrement utile pour traiter des sources de données non structurées, telles que le texte et les images, qui étaient auparavant difficiles à gérer avec l'informatique classique. Ainsi, dans les situations où les règles de gestion sont complexes, l'IA peut offrir une solution efficace et adaptative.

Automatiser les traitements complexes. Alors en apprenant à partir des données plutôt qu'en développant les règles. Et puis tout ce qu'on ne savait pas en fait jusque-là, gérer comme le analyser des contrats, la conformité de contrats, classer des mails en fonction des demandes, la nature de la demande du client, et cetera. Tout ça. C'est les cas qu'on peut adresser avec l'IA donc en fait c'est beaucoup de tâches qui sont répétitives. Ils peuvent être automatisés dans l'entreprise.

Augmenter les, l'humain en en lui proposant des recommandations, des suggestions et donc par exemple on a travaillé sur les systèmes d'ajustement comptable comme y a des gros volumes à ajuster tous les mois dans des délais très courts. Et puis il y a que certaines erreurs.

Alors je Présente des similitudes avec d'autres erreurs, donc de données présente certaines d'autres, d'autres problèmes de données à ce qu'on fait, c'est on fait un un moteur de recommandation de valeur d'ajustement à l'utilisateur en lui disant tiens, cette ligne d'ajustement là a présenté des des caractères, des similitudes avec une autre ligne qu'on a déjà ajustée.

Et les similitudes et compte tenu de ces similitudes, la valeur d'ajustement que je propose, avec quel niveau de confiance et donc du coup ça, ça facilité le travail d'analyse de de l'humain et du coup ça le facilité à finalement le le travail de faire des ajustements comptables.

Niveau de confiance :

La confiance, elle est à plusieurs niveaux, y a la confiance qu'ont les Data Scientist qu'ils ont développé fonctionne bien et donc en fait pouvoir comprendre quelles sont les variables qui sont importantes pour le modèle et du coup utiliser cette information là pour échanger avec le métier.

- ça permet en fait de s'assurer pour la Data Scientist que le modèle affiche de bonnes performances pour de bonnes raisons.

- ça permet aux à l'équipe projet. Côté métier, de dire c'est pas une boîte noire, y a du rationnel métier derrière le fonctionnement du modèle.

C'est l'équipe projet, l'équipe projet. Côté métier, l'équipe data science où finalement, avoir des notions de d'explicabilité du fonctionnement du modèle, ça crée les conditions de la confiance sur ce qui est en train d'être construit, etc

Ensuite, en temps, en tant qu'utilisateur du modèle, il y a 2 niveaux :

- Il y a un niveau quelque part d'acceptation parce qu'en fait on n'est pas sur quelque chose qui est déterministe. On n'est pas sûr quelque chose aujourd'hui qui est encore habituel et donc en fait ces personnes ont besoin de comprendre en fait un petit peu comment ça fonctionne, que c'est pas de la magie noire.

- Le 2e niveau, c'est comme on a quelque chose qui est pas parfait, qui prédit bien 8 fois sur 10, 9 fois sur 10.

En fait, il faut, c'est difficile pour un utilisateur d'avoir une solution qui parfois se trompent et donc dans quel cas je lui fais confiance ? Dans quel cas je lui fais pas confiance et si je peux pas lui faire confiance, tout le temps, est ce que est ce que finalement je l'utilise ? donc en fait à un moment donné il faut créer les conditions à la fois que l'utilisateur comprenne ce qu'on lui propose et qu'il comprenne dans quel cas en fait il peut avoir confiance et donc la notion d'explicabilité passe par 2 choses :

1- Pour qu'il puisse bien comprendre pourquoi le modèle a proposé cette solution, cette réponse?

2- une notion additionnelle qui est la notion de niveau de confiance dans les modèles. c'est les probabilités.

Exemple : il faut dire ce mail, c'est une demande d'information du client, une réclamation ou un spam peut être que le modèle. Il hésite entre 2 catégories parce que peut-être que le client a mis 2 choses, il y a peut-être une demande d'information et puis il a fait une réclamation, il a peut-être fait les 2 dans son mail et donc le modèle va finalement dire à l'utilisateur.

je suis pas sûr, je sais pas en j'hésite entre les 2 catégories, c'est très proche et donc donner un niveau de confiance c'est finalement dire à l'utilisateur dans quel cas il doit être vigilant sur la réponse proposée par le modèle, donc ça c'est

La confiance dans les modèles d'intelligence artificielle est essentielle pour leur adoption et leur efficacité dans des contextes réels. Voici un résumé des niveaux de confiance et des facteurs associés :

- 1er Niveau - Proposition et Contrôle :

    - L'IA fait des propositions à l'utilisateur.
    - L'utilisateur peut parfois douter de ces propositions.
    - Il est crucial d'expliquer à l'utilisateur comment l'IA fonctionne et de lui donner une visibilité pour évaluer la fiabilité de la proposition.

- 2ème Niveau - Compréhension et Explicabilité :

- Les utilisateurs ont besoin de comprendre les décisions de l'IA.
- Par exemple, dans une banque, si un conseiller ne comprend pas pourquoi un crédit est refusé à un client, cela peut compromettre la confiance.
- L'explicabilité est essentielle pour éviter de délégitimer les professionnels dans leurs rôles.

- 3ème Niveau - Qualité et Adaptabilité :

    - Les modèles d'IA doivent être robustes et s'adapter aux changements de données.
    - Les événements tels que la pandémie de COVID ou les fluctuations des marchés financiers peuvent affecter les données.
    - Il est vital de réentraîner régulièrement les modèles et de surveiller leur performance pour garantir leur pertinence.

- Mise en Place du MLOps :

- Après avoir établi la confiance à différents niveaux, il est essentiel d'adopter des pratiques de MLOps.
- Cela implique la mise à jour régulière des modèles et la surveillance de leur qualité.

- Au-delà de la Confiance Perçue :

    - Même si un modèle semble fonctionner correctement, cela ne garantit pas sa fiabilité totale.
    - Il est crucial de continuer à évaluer et à tester les modèles pour s'assurer qu'ils fonctionnent comme prévu.

Exemple.

On peut justement avoir utilisé des données historiques sur lequel les taux étaient bas sur lequel y a eu une période de COVID et donc on a un modèle qui est peut-être pertinent sur les données passées mais qui ne va pas bien capable de prendre en compte les nouvelles évolutions de de contexte où on va avoir un modèle qui va prendre sa décision sur un mauvais critère.

Et donc il faut contrôler l'absence de biais.

Exemple, Ça peut être qu'une banque a historiquement comme critère de pas accorder de crédit aux étudiants en fin d'études parce qu'ils n'ont pas le critère historique d'accorder des crédits, c'est de dire est-ce qu'en fait il y a une récurrence de revenus des CDI et un paye qui tombe tous les mois pour rembourser le crédit et que peut être qu'un jeune actif qui rentre dans la dans la vie active il a besoin de faire un crédit conso pour pouvoir s'équiper et demain il aura un emploi …

si dans l'historique, on n'a jamais accordé de crédit à un jeune qui est en fin d'étude, finalement on peut être qu'on lui accordera jamais de nouveaux de crédit parce que c'est quelque chose qu'on fait par le passé. Alors ce n'est pas ce qu'on veut avoir.

Et donc on a plein d'exemples d'IA qui semble fonctionner pour de mauvais critères.

Aussi, les exemples de de photos pour reconnaître un loup d'un chien ou le modèle semblait fonctionner. Mais finalement, ce que le modèle regardait, c'était pas l'animal, c'était l'arrière-plan

Parce qu'en fait y a des caractéristiques physiques entre le chien et le loup qui sont différentes parce que le modèle regardait l'arrière-plan, s'il y a de la neige ou s'il y a de la montagne. S'il y a de la neige, de la montagne, c'est un environnement sauvage et don ça va plutôt être un loup.

C'est des choses sur lesquelles on on on n'est pas assez vigilant. Il y a en fait y a beaucoup, beaucoup plus de cas qu'on ne le pense ou le modèle semble bien fonctionner, mais il fonctionne pas en fait si bien que ça. C'est qu'en fait tous les modèles quasiment que l'on développe, ils ont des biais.

1- Si on veut faire un modèle de remboursement anticipé de crédit, les données historiques les, les taux étaient bas et ils sont en train de remonter. Donc en fait on a un biais, c'est que ça ne représente pas la situation actuelle.

2- On peut aussi avoir un biais, c'est dans l'assurance vie ou dans l'assurance auto.

pour l'assurance vie on. C'est un biais réglementaire, mais on prend pas en compte le sexe pour pas faire de discrimination. Sauf que les femmes vivent plus longtemps que les hommes.

Donc en fait, elles sont pénalisées dans ce modèle là parce qu'en fait on a enlevé l'information homme/femme et que du coup, dans un souci d'égalité on fait payer les femmes plus chère leur assurance vie.

Une d'une manière globale, dire que je prendre un échantillon restrictif de données qui représente le passé sur une courte période de temps et qu'un échantillon d'informations du passé permet de prédire l'avenir c'est un biais.

Un autre biais, c'est que. Peut-être que si on veut prédire, est-ce que quelqu'un doit acheter où vendre des actions mais peut être qu'on n'a pas l'information qu'il y a eu la guerre en Ukraine à un moment donné et que du coup ça change complètement le contexte..

Il y a l'explicabilité qui crée la confiance, mais la confiance passe aussi par être conscient que les la plupart des modèles, finalement, comportent des biais parce que l'historique comporte des biais.

exemple réel :

C'est un moment donné dans une banque, on a utilisé des données de presse professionnelles économiques pour réduire la santé des entreprises.

on apprend l'univers de vocabulaire des entreprises. L'univers économique de ces articles pour que l'IA comprenne bien le texte.

Et finalement, l'IA entraînait sur des articles de presse économique, à une des choses qu'il a appris, c'est que le féminin de conducteur pour ce modèle appris sur de la presse économique entreprise.

c'est passagère et donc les biais qui existent les biais humains, dans les données … Ils sont capturés, amplifiés.

Donc il faut être conscient qu'il y a des biais pour faire de l'IA responsable, et il faut faire des contrôles sur les biais. c'est pas une tâche facile à faire parce que finalement pour voir s'il y a un biais de discrimination raciste, sexiste, discrimination de l'âge ou d'autres biais

En fait, il faut faire beaucoup de contrôles et n'a pas toujours les informations pour le pour le faire.

Robustesse :

Le model pourrait être fragile dans la façon dont il a été fait, peut être que le besoin exprimé par le client (le métier) a été mal compris.

Peut-être que le métier a dit, je voudrais un modèle qui me dise quels sont les clients qui vont faire telle action. Parce que je voudrais aller leur proposer quelque chose.

Alors qu'en fait, il faut peut-être que le besoin n'est pas de prédire quels sont les clients qui vont faire une action, mais peut-être que le besoin c'est de prédire à quel client je pourrais recommander de faire quelque chose, donc prescrire donc ne pas être sur un modèle prédictif, mais sur un modèle prescriptif.

Peut-être que si on veut écarter certaines populations, les, les jeunes, les, les personnes âgées etc

- La compréhension du besoin est pas toujours bonne.
- Les modèles, s'appuient sur les données en entrée et les données entrées sont pas toujours de bonne qualité, elles sont pas toujours très bien retraites et donc cette qualité des données en entrée influe sur la qualité du modèle.
- Le modèle a peut-être pas pris en compte certains critères importants. Peut-être qu'il semble fonctionner mais Il est erroné parce qu'il passe à côté du cœur métier.

Le du modèle : peut-être que le model qui a été choisi utilement complexe. Peut-être qu'il a été mal paramétré et peut-être qu'il fonctionne aujourd'hui mais peut-être que c'est au moindre changement, il ne fonctionnera plus. Or, le sujet de développer une IA, c'est pas en fait d'avoir un modèle qui fonctionne bien sur les données disponibles mais c'est d'avoir un modèle qui généralise bien sur des données qu'il n'a jamais vues.

D'où le fait d'avoir des modèles assesment parce qu'aujourd'hui l'IA est de plus en plus utiliser dans les entreprises et donc il y a un double phénomène, c'est utilisé pour de plus en plus des traitements sensibles au sens que si le modèle se trompe ou si le modèle tombe en panne, ça peut avoir un impact business.

Exemple :

Parce que peut-être qu'on fait des propositions clients, donc par exemple, si l'octroi de crédit, l'octroi de financement, si on dit non un client à qui on aurait dû dire oui peut-être qu'on va perdre le client

et si on dit oui à un client à qui on aurait dû dire non il va peut-être pas rembourser le crédit et on va avoir des pertes financières. Cela montre que les modèles deviennent de plus en plus critiques pour le business d'une banque. Si on parle d'une entité, par exemple qui fait que du crédit conso si le modèle d'octroi de crédit se met à dérailler complètement.

Ça peut complètement casser le business model de la viabilité économique de l'entité crédit Conso. Donc ça devient critique d'un point de vue business.

La démocratisation de l'IA dans les entreprises :

La démocratisation de l'intelligence artificielle dans le monde professionnel a entraîné une diversification des profils impliqués dans la conception et l'utilisation de modèles d'IA. Voici un résumé des points clés abordés :

- Évolution des Équipes d'IA :
    - Historiquement, les entreprises comptaient sur une petite équipe d'experts en data science pour développer et gérer leurs initiatives d'IA.
    - Avec la démocratisation de l'IA, de plus en plus d'individus sans formation complète en data science se lancent dans le développement de modèles d'IA.
- Diversité des Profils :
    - Outre les Data Scientists, des développeurs et des professionnels ayant une formation en statistiques, mais sans formation complète en IA, commencent à créer des modèles.
    - Cette diversité peut entraîner une variété de compétences et d'approches dans la conception de modèles.
- Changement de Motivation :
    - Les pionniers de l'IA étaient souvent des passionnés du domaine, motivés par un intérêt profond pour la technologie.
    - Avec la montée en popularité de l'IA et la demande croissante pour les compétences en data science, certaines personnes sont attirées par le domaine principalement pour des opportunités d'emploi, et non nécessairement par une passion intrinsèque pour le sujet.

Le modèle assessment de maîtrise du risque opérationnel. En fait, elle est là pour amener un contrôle de 2eme niveau et donc dans les banques ça existait déjà parce que pour les modèles quantitatifs il y avait déjà du contrôle, de la modélisation qui était fait par les équipes risques et donc la notion de modèle assessment c'est amener quelqu'un qui va venir contrôler et challenger ce qui a été fait et la démarche, elle est en 2 temps :

- 1er niveau : Est-ce qu'il y a vraiment besoin de faire un contrôle ? Est-ce qu'on est sur quelque chose qui est critique ?

- 2ème niveau : Si c'est critique, est ce qu'on est sûr que le risque est maîtrisé et donc on contrôle. Est-ce que finalement on a pris les bonnes sources de données ? bonne qualité ? données mesurées et contrôlées ? est ce qu'on a bien compris le besoin ? est ce qu'on a pris une approche de modélisation qui ne soit pas trop complexe, qui soit maîtrisée.

Est-ce qu'on a mis en place un monitoring pour contrôler la qualité en production et pouvoir réagir en cas de défaillance ou de baisse de la qualité de de l'IA.

AIDA (AI design Authority) : c'est une équipe de Data Scientist qui a une expertise spécifique sur la qualité des modèles et les facteurs de non qualité de modèles et qui vont faire un contrôle. Mais comme le but c'est pas de freiner l'usage de l'IA, Le dispositif de contrôle va être gradué.

un certain nombre de de de guidelines, de bonnes pratiques. De de process de gouvernance vont vont être, vont être mis en place.

Et on va contrôler de manière approfondie les cas d'usage auquel il y a le plus de risques de perte financière, le plus de risques d'image pour l'entreprise et ils vont contrôler un petit peu tous les points que je mentionnais est ce qu'on a bien compris le besoin métier, est-ce qu'on a bien choisi les bonnes sources, est-ce qu'on a bien une modélisation simple, maîtrisée. Est-ce que la complexité de la modélisation est adapté à la complexité du sujet.

Et est ce qu'on a bien mis en place les monitorings et les rôles et responsabilités pour rassurer et maîtriser la qualité en production ?

Le but, c'est pas de faire de la remédiation. c'est de pouvoir contrôler, sensibiliser les équipes à la source sur les bonnes pratiques sur les points d'attention pour que les équipes intègrent les bons réflexes dès le départ.

Exemple : pour reprendre l'image du contrôle routier, c'est à dire qu'en fait, un bon réflexe quand on monte en voiture, c'est de mettre sa ceinture de sécurité.

On peut mettre des policiers au rond-point pour vérifier que les gens ont bien mis leur ceinture de sécurité, mais le, la finalité, c'est pas de mettre des amendes. La finalité c'est bien en fait que le réflexe de mettre la ceinture de sécurité elle soit elle soit prise. Même chose pour les radars par rapport à la vitesse. Et donc c'est bien cette idée d'à la fois de contrôler mais que finalement ce contrôle-là soit surtout d'abord une action de sensibilisation pour que ces réflexes-là soient par construction.

Les démarches de certification de maturité dans les process : On est encore sur une discipline un petit jeune au niveau de la conception de de modèle IA avec beaucoup de nouveaux entrants et donc les pratiques sont hétérogènes et donc certaines équipes ont pas forcément le réflexe de versionner les données, de versionner les modèles, de contrôler la qualité des modèles après un nouveau entraînement pour vérifier qu'il y a pas eu un problème sur les données qu'on générait une baisse de la qualité du modèle et donc un moyen à la fois de se mesurer, se positionner, est-ce qu'on est mature dans les process de conception IA et aussi en afficher cette maturité, diffuser la confiance en interne et en externe. c'est les démarches de certification qui se mettent en place, le crédit Agricole notamment au niveau du data Lab ont mis en place une démarche où ils ont eu une double certification :

- Une certification qui est plus orientée processus de développement

- Une certification qui est plutôt au niveau de la partie Confiance, éthique.

La prise de conscience de l'importance de l'IA de confiance, des enjeux d'explicabilité, de maîtrise des biais, de contrôle du risque opérationnel de modèle, des enjeux de compliance réglementaire, d'éthique sont pris en compte ça se décline en plan d'action, en guidelines, en process, en gouvernance, au sein du groupe Crédit Agricole.

Green IA :

C’est un sujet qui va prendre une ampleur, criticité plus importante demain qu'aujourd'hui.

En fait, on peut-on peut obtenir le même résultat en consommant beaucoup moins que ça soit en fait en termes de réduction ou en termes d'entraînement.

- Le réflexe qui avait plutôt la scientiste Jusqu'à aujourd'hui, c'était plutôt de dire, je prends le modèle le plus gros, le plus complexe, parce que c'est celui qui va être le plus performant.

Et il y a un ensemble de techniques où on peut utiliser des modèles plus simples où on peut préparer les données en amont pour faciliter le travail d'apprentissage du modèle qui font qu'ils apprennent plus vite. Qui vont faire que la phase d'apprentissage va consommer moins d'énergie, moins de ressources. Et puis la phase de prédiction aussi. Si le modèle est moins complexe, il va consommer moins.

Et c'est jusqu'à il y a peu de temps, c'était peu pris en compte et en fait aujourd'hui y a au niveau de Crédit Agricole une première amorce.

Au niveau du datalab Groupe il y a une mesure de ce qui est consommé par un modèle.

Une fois qu'on mesure, un moment donné, on peut créer les conditions pour se dire ce que je progresse, est-ce que je progresse pas et où je me situe?

- Ça va devenir vraiment un sujet parce que les tout nouveaux modèles de LLM, donc les tous les modèles de type chat GPT sont des très gros modèles Deep learning, ces modèles consomment énormément de ressources pour s'entraîner et ils consomment beaucoup de ressources pour prédire et aujourd'hui

Par exemple, l'explosion de de service type GPT brutal aux États-Unis a eu un impact dans certains États sur la consommation électrique parce qu'il y a plus de data Center qui ont été allumés de serveur qui ont été allumés pour traiter les demandes et donc à un moment donné, il y a eu un sujet de parce que Il a suffisamment d'énergie en de capacité énergétique en en Californie pour pouvoir etc

Ça va être un sujet parce que ces entreprises sont engagées dans des démarches de Green IT, déclinaison de leurs démarches RSE au niveau informatique. Elle veulent garantir utiliser moins de serveurs, rationaliser un peu les consommations électriques. On va essayer de d'avoir de faire moins 5-10% en émission de CO2 etc. Alors qu'en fait ces modèles, tous ces modèles deep learning, extrêmement puissants ils sont extrêmement énergivore aujourd'hui, à la fois anticipation thermique et à la fois en énergie consommée.

En anticipation thermique veut dire que du coup ça tire les besoins, climatisation plus forts et donc on a un double effet, il consomme beaucoup et eux-mêmes ils ont beaucoup de chaleur qui, du coup, implique des besoins de refroidissement plus important.

Responsabilité juridique :

C'est un service qu'on met pour le client et donc si on le contrôle pas, si on le maîtrise pas, si il fait des erreurs, c'est une erreur de la Banque.

Avec les nouvelles IA génératives, se pose une question si jamais à un moment donné la réponse qui est généré venait offenser quelqu'un par exemple, qui est responsable, est-ce que c'est l'éditeur de la solution, est ce que c'est la Banque ?

Pour l'instant, dans tous les cas, indépendamment de la réponse juridique, ça serait vu par le client comme étant un impact pour l'entreprise.

Interview Pierre CASTELLANI (CACIB)

Quelle est la différence entre AIDA et MRM ?

MRM : La vue des modèles qui prend en compte l'ensemble des modèles de la banque

AIDA (AI Design Authority) : Une comitologie et derrière ça il va y avoir des spécifications. LA comité AIDA permet de valider l'ensemble des projets spécifiques IA.

Dans un projet IA les modèles vont être référencés par le MRM.

Dès qu'il y a la notion de model dans un projet, alors le MRM intervient.

AIDA, c'est le contrôle de l'IA au sens large, le contrôle des projets d'IA a donc typiquement, comment est-ce que les données sont représentées ?

c'est 2 sujets qui sont séparés et qui ont des intersections par l'intermédiaire de la modélisation.

la façon dont Aïda (comité et assessments sur les projets) sont construits est d'essayer de faire en sorte que les projets soient compatibles avec les différentes règles qui va édicter la design Authority (by design). Alors pourquoi ? Parce que ce qu'elle essaye d'intervenir au début du projet qui est de près ou de loin du machine learning

Dès que le projet va être en place, la comité AIDA va être sollicitée de façon à avoir une première vision et potentiellement de définir avec les personnes qui sont en train de mettre en place ce projet les différents aspects qui vont permettre de faire en sorte que le projet soit compatible pour éviter d'être dans un cas où avant d'arriver en production pour un projet on va se retrouver à dire vous pouvez pas partir en production,

Quels sont les critères que vous avez pour que pour valider ou pas un projet ?

- Première chose qui est une première grille d'auto assessment qui a été édicter, mais c'est quelque chose on goning Parce qu'il y a plusieurs aspects qui rentrent en ligne de compte et notamment il y a la réglementation européenne l'IA acte qui a été édicté récemment donc AIDA va être amené à faire des changements dans cette grille.

- Un autre aspect, c'est le fait qu'au niveau du Groupe il y a l'IA acte (Européen) qui va être décliné mais clairement, Crédit Agricole va potentiellement vouloir être plus pointu dans certains cas mais surtout avec certaines contraintes qui sont liés aux contraintes du groupe.

Il y a au sein du Crédit Agricole, un certain nombre d'équipes qui sont en train de travailler pour pouvoir décliner l'IA acte et nous envoyer la déclinaison pour le groupe que nous on va intégrer à nouveau et que l'on va décliner au sein de CACIB et donc derrière AIDA va pour chacun des projets modifier la grille d'auto assessment de façon à respecter les différentes règles que ça soit d'un point de vue data, d'un point de vue modèle, d'un point de vue éthique, d'un point de vue RSE, donc toutes ces choses-là vont être vont être modélisées dans cette design autorité.

Pour la data, qu'est-ce que vous demandez en tant que comité éthique ?

La préoccupation principale de CACIB en ce qui concerne les données utilisées pour l'intelligence artificielle est actuellement axée sur la qualité et l'intégrité statistique des données, plutôt que sur les biais potentiels qui pourraient être présents. Voici un résumé des points abordés :

Priorités Actuelles :

- CACIB se concentre principalement sur l'analyse statistique des données en entrée des modèles, en mettant l'accent sur les aspects numériques.
- Les analyses éthiques ou liées à la responsabilité sociale des entreprises (RSE) ne sont pas encore au premier plan.

Biais dans les Données :

- Bien que la question des biais dans les données soit pertinente, CACIB n'a pas encore mis en place d'analyses spécifiques pour les identifier ou les traiter.
- La mise en œuvre de telles analyses est prévue pour l'avenir, en particulier en réponse à des réglementations comme l'IA ACT.

Qualité des Données :

- La qualité et la fiabilité des données sont essentielles pour garantir la précision des modèles d'IA.
- CACIB s'assure que les données utilisées sont cohérentes et représentatives des scénarios réels.

En conclusion, bien que la question des biais dans les données soit cruciale pour l'éthique de l'IA, CACIB se concentre actuellement sur la garantie de la qualité statistique des données. Cependant, à mesure que les réglementations évoluent, l'entreprise envisage d'aborder la question des biais de manière plus approfondie à l'avenir.

Recommandation DATA :

L'automatisation et l'utilisation de l'intelligence artificielle dans le traitement des données nécessitent une approche rigoureuse et structurée pour garantir la fiabilité et la pertinence des résultats. Voici un résumé des étapes et des considérations techniques mentionnées :

1. Préparation des Données :

- Dictionnarisassions et Golden Sources : Il est essentiel d'avoir des données non transformées ou non interprétées, qui servent de référence.
- Analyse Métier : Avant toute chose, le métier doit valider la cohérence des données par rapport au projet envisagé.
- Validation Statistique : Une fois les données validées, l'étape suivante consiste à s'assurer de leur cohérence statistique, notamment en termes de représentativité des différentes classes.

2. Transformation des Données :

- Encodage : Les données catégorielles, comme les pays, doivent être transformées en valeurs numériques pour être utilisables par les modèles.
- Normalisation : Les variables numériques peuvent nécessiter une normalisation pour éviter les biais dus à des ordres de grandeur différents.

3. Types de Variables :

- Variables Catégorielles : Ces variables sont transformées en valeurs numériques à l'aide de techniques d'encodage.
- Variables Numériques : Ces variables peuvent nécessiter une transformation pour éviter les biais, en particulier lorsque les ordres de grandeur varient considérablement.

4. Considérations sur les Modèles :

- Les modèles de machine learning doivent être alimentés avec des données qui correspondent à leurs attentes en termes de format et de structure.
- Il est essentiel de vérifier que les transformations des données sont cohérentes et ne biaisent pas les résultats du modèle.

Robustesse :

L'efficacité et la robustesse d'un modèle de machine learning dépendent de plusieurs facteurs, notamment de la nature et de la quantité des données d'entrée. Voici un résumé des points clés abordés :

- Représentation de la Réalité :
    - Un modèle de machine learning est une représentation simplifiée de la réalité basée sur les données disponibles.
    - Si nous avions une infinité de données, le modèle pourrait parfaitement représenter la réalité. Cependant, dans la pratique, nous travaillons avec un sous-ensemble limité de données.
- Robustesse du Modèle :
    - La robustesse d'un modèle dépend de la quantité et de la qualité des données d'entrée.
    - Certains modèles peuvent être "overfitting", c'est-à-dire qu'ils s'adaptent trop bien aux données d'entraînement et ne généralisent pas bien aux nouvelles données.
- Généralisation :
    - L'objectif principal d'un modèle est de généraliser à partir des données d'entraînement pour faire des prédictions précises sur de nouvelles données.
    - La capacité de généralisation dépend du type de modèle utilisé et de la nature des données d'entrée.
- Choix du Modèle :
    - Différents modèles peuvent avoir des niveaux de robustesse différents en fonction des données d'entrée.
    - Si un modèle ne généralise pas bien, il peut être nécessaire de changer certains paramètres ou même de choisir un modèle différent pour améliorer la robustesse.

Sécurité :

Dans la sécurité, Il y a deux questions :

Qui a le droit de voir les données ?

Qui a le droit de faire tourner le modèle, donc ça c'est ça. C'est un sujet d'accessibilité.

Infrastructure

Il y a un sujet côté IA Responsable, On n'a pas encore décliné ce point-là, c'est quelles sont les possibilités de l'infrastructure par rapport au modèle que l'on cherche à calibrer ?

Et aujourd'hui, il y a beaucoup de modèles différents qui peuvent être plus ou moins consommateurs de ressources en fonction du détail de données

L'utilisation de modèles non supervisés pour évaluer la distance entre tous les points d'un ensemble de données peut rapidement devenir complexe et exigeante en termes de ressources, en particulier lorsque la taille de l'ensemble de données augmente.

- Complexité des Modèles :
    - Les modèles qui évaluent la distance entre tous les points d'un ensemble de données peuvent avoir une complexité exponentielle.
    - Cela nécessite une puissance de calcul importante, en particulier pour les grands ensembles de données.
- Pertinence des Résultats :
    - Malgré la puissance de calcul nécessaire, il n'est pas garanti que les résultats obtenus soient pertinents ou utiles pour l'objectif visé.
    - Le choix du modèle approprié nécessite l'expertise et l'expérience des Data Scientists.
- Infrastructure du Crédit Agricole :
    - Le Crédit Agricole dispose d'infrastructures solides, permettant de traiter des modèles complexes.
    - Cependant, la capacité de ces infrastructures peut parfois limiter les performances des modèles.
- Utilisation de PC Personnels :
    - Certains modèles de machine learning peuvent être exécutés sur des PC personnels, en particulier si l'ensemble de données est relativement petit.
- Taille du Dataset :
    - La taille de l'ensemble de données est un facteur déterminant de la puissance de calcul nécessaire.
    - Les ensembles de données plus grands nécessitent plus de ressources et peuvent ne pas être adaptés à l'exécution sur des machines locales.

Equité :

Le modèle ne va apprendre que ce qu'on ce que l'on va mettre dans les données, c'est à dire que si effectivement les données ne sont pas forcément compliant avec les règles de l'IA Responsable, le modèle va apprendre ça.

C'est-à-dire qu'il faut bien comprendre que le modèle de Machine learning ne va apprendre que ce qu'il y a dans les données. C'est pas lui qui va faire des biais.

Tout ce qui va être appris et donc tout ce qui va être potentiellement sujet à discrimination ou autre, et donc potentiellement qui va être lié à aux règles qui vont être édictées par l'IA Act seront dans les données parce que le modèle lui, il ne va que traduire ce que l'on a dans les données.

Green IA :

Aujourd'hui c'est pas forcément effectivement la priorité numéro un de tous les data Scientists.

C'est un sujet chez quasi chez CAGIP parce que les machines qui sont utilisées aujourd'hui pour faire de l'apprentissage sur la machine qui consomme énormément d'énergie.

Donc, quelque part, il faut les utiliser avec parcimonie. Et déjà aujourd'hui, il y a la notion de mutualisation.

Exemple :

Si on cherche à calibrer, à faire du Deep learning, donc à calibrer des réseaux de neurones, ça va être extrêmement consommateur en termes de calcul. Et aujourd'hui les meilleurs processeurs pour faire ce genre de de choses sont des GPU.

Or 1GPU, est très puissant, mais qui consomme énormément d'énergie, ce qui veut dire que si on utilise pas en permanence le GPU, alors on va le faire tourner pour rien et donc il y a le besoin de mutualiser les différents sujets, les différents projets, les différentes entités. de façon à ce que les GPU qui est utilisée soit utilisée au mieux et ça aujourd'hui c'est déjà quelque chose qui est accessible chez CAGIP puisqu'en fait ils mettent en place une architecture qui permet d'avoir accès avec plusieurs projets, d'avoir accès à la même infrastructure.

Bien sûr, ça résout pas tous les problèmes et on est encore loin de de de faire du Green Deep Learning. Mais ce qui est important c'est de comprendre que ce point-là va permettre effectivement de mutualiser un certain nombre de ressources, ça permet effectivement de de de diminuer l'impact en CO2.

Interview Alexandra COURTAUX (CACIB)

Elle intervient donc sur les sujets IA du NLP, donc de traitement du langage pour faire le pont entre le métier fonctionnel et les data Scientists qui vont coder la solution qui va permettre de récupérer l'information ou de la classifier, etc

Cas d'usage :

L'automatisation de la lecture des documents est essentielle dans de nombreux métiers où il est nécessaire de traiter des contrats, des rapports annuels ou d'autres données liées aux clients. Traditionnellement, ces documents sont lus manuellement, ce qui peut être fastidieux et

chronophage, surtout lorsque les documents sont volumineux ou de mauvaise qualité, comme les PDF scannés.

Pour répondre à ce défi, des solutions ont été mises en place pour extraire et classifier automatiquement l'information pertinente des documents. Ces solutions permettent non seulement de gagner du temps, mais aussi d'orienter l'utilisateur vers les informations pertinentes, lui permettant ainsi d'appliquer son expertise métier de manière plus efficace.

Un exemple concret est le "Smart Reader", qui ne se contente pas de rechercher des informations dans les documents, mais classifie également les phrases ou d'autres unités de texte selon différentes thématiques pertinentes pour l'utilisateur. C'est comparable à la lecture d'un livre et à la mise en évidence de différentes sections en fonction de leur pertinence, comme si on utilisait différents stabilos pour marquer différents sujets ou personnages.

Cette approche permet non seulement de faciliter la lecture et l'analyse des documents, mais aussi d'exploiter les données classifiées pour créer des synthèses, des tableaux de bord et d'autres outils d'analyse. En fin de compte, cela se traduit par une meilleure gestion de l'information, une prise de décision plus éclairée et une efficacité accrue.

D’autres projets d’IA dans la banque :

tâches de NLP, on a pas de use case sur la Text normalization donc le résumé direct de texte et ça commence à émerger les idées. Notamment de compte rendu de réunion même de compte rendu, de parfois de gros, une grosse quantité de données qu’ils récupèrent nos salers ou nos traders dans la journée donc des comptes rendus du jour parce qu'on a beaucoup côté marché.

Objectifs des projets :

Automatisation de tâches, rechercher l'info plus rapidement, gagner du temps à la lecture, avoir une lecture plus pertinente, classifier l'information même y compris dans des chats récupérés de l'information plus rapidement.

Problématiques rencontrés :

Analytics : leur enjeu ça va plutôt être d'utiliser toutes les données des marchés et de faire parler des prédictions de prix donc, y a des pricers de faire des prédictions de taux, de courbes.., IA documentaire

Erreur :

Nous travaillons sur des projet NLP sur du texte et dans la plupart des cas, les erreurs ne sont pas très graves. Si une mauvaise classification d'information se produit dans un bloc de texte, l'utilisateur professionnel reviendra simplement à une lecture manuelle du document pour identifier ce qui a été omis ou pour corriger la classification erronée. Dans d'autres situations, comme la capture d'informations spécifiques d'une section de texte, si l'information est manquante ou incorrecte, l'utilisateur parcourra le document entier.

Actuellement, dans la plupart des solutions, l'humain joue un rôle central et agit comme un second regard sur les résultats de l'IA. Cela garantit une double vérification et une validation des informations.

Il est essentiel de se rappeler que l'IA n'est jamais fiable à 100% et qu'il est improbable d'atteindre une performance de prédiction parfaite.

Il est crucial que les utilisateurs conservent leur expertise et leur discernement face aux résultats fournis par l'IA. Ils doivent être conscients de la manière d'interpréter les sorties de l'IA, en gardant à l'esprit ses limites. L'objectif est de s'assurer que l'utilisateur comprend le processus de bout en bout : quelles données sont fournies en entrée, comment elles impactent leur travail quotidien, ce qu'ils obtiennent en sortie et comment cela affecte également leur routine.

Assessment:

L'IA Factory est activement impliquée dans l'initiative AIDA. qui se décline en plusieurs volets :

1. Groupe Crédit Agricole : Il s'agit de la traduction et de l'application de la nouvelle réglementation AI ACT en termes fonctionnels et projets, afin de l'intégrer dans les opérations quotidiennes.

2. CACIB : Des comités sont mis en place pour veiller à ce que les projets respectent cette réglementation au quotidien. L'objectif est également de s'adapter aux exigences en matière de documentation et de répondre aux besoins de l'inspection générale et des aspects légaux.

3. Outils et processus : Grâce à AIDA, l'IA Factory a lancé un projet de surveillance des modèles. Ce projet vise à établir un cadre d'IA où, pour chaque type de modèle déployé (par exemple, un modèle d'annuaire pour une solution NLP), des indicateurs spécifiques sont obligatoirement suivis. Cela garantit que l'utilisateur sait ce qu'il obtient en production. Si un problème survient, le système est conçu pour alerter l'utilisateur et, si nécessaire, arrêter les opérations.

En résumé, l'IA Factory, à travers l'initiative AIDA, s'efforce de garantir la conformité réglementaire, la transparence et la fiabilité des modèles d'IA qu'elle déploie.

Quand on va en PRD :

Ils vont y a plusieurs métriques qui vont être utilisées de performances donc ça peut être des pourcentages mais ça peut être aussi un nombre de réponses trouvées correctes ou un nombre de réponses trouvées incorrectes où ça peut être aussi j'ai trouvé trop de réponses. D'accord j'ai trouvé une réponse correcte mais j'en ai trouvé trop donc en fonction en déterminé en fonction de ce que veut l'utilisateur métier à la fin.

Parce que, par exemple, on a des métiers qui vont nous dire on préfère que vous me récupériez toute l'information dont j'ai besoin, même si elle est enveloppée d'un petit peu de bruit et il y en a qui vont plutôt nous dire non, moi je veux vraiment me concentrer sur quelque chose qui serait correct et pertinent pour moi, même si vous ratez d'autres choses que j'aurais pu voir parce que c'est pas grave. quand on va en prod y a pas une règle générale, il y a pas une loi qui dit à partir de 80% on part en production.

DATA :

Avec la mise en place d'une stratégie globale captive de gouvernance des données (des Golden sources) qui assurent normalement la qualité des données qu'on va récupérer pour le respect de la vie privée.

Parmi les solutions proposées pour le respect de la vie privé il y a l'anonymisation des données personnelles qu'on récupérait pour la gouvernance des données

Sécurité :

Nos projets sont hébergés on-permis (en interne) on a le même niveau d'exigence de sécurité des systèmes et des données avec nos collaborateurs externes. Si demain on va choisir une solution externe, il faut qu'elle respecte le même niveau de confidentialité et de sécurité.. qu'on s'applique chez nous.

Robustesse :

Il est essentiel d'avoir un volume de données conséquent pour répondre de manière adéquate aux besoins des utilisateurs. La base de connaissances et la base d'entraînement sont travaillées pour s'assurer que le modèle fait les prédictions attendues. L'objectif est de garantir que le modèle se comporte comme prévu.

Un volume de données suffisant est nécessaire pour garantir la robustesse du modèle, lui permettant de généraliser à de nouvelles données qui ne sont pas uniquement basées sur le jeu d'entraînement. Cela évite que le modèle ne produise des résultats erronés en production s'il rencontre des données différentes de celles vues lors de l'entraînement.

Trois aspects sont donc essentiels : le volume de données, la qualité des données et les annotations. En tant qu'infolinguiste, le travail est effectué en collaboration avec les métiers pour annoter les jeux de données destinés au traitement automatique du langage naturel (NLP).

Déroulement d'un projet avec NLP :

Lorsqu'un métier présente une idée à l'IA Factory, deux enjeux majeurs se dégagent : comprendre le besoin fonctionnel et sensibiliser aux implications de l'IA. Il est essentiel de déterminer si l'IA peut répondre aux besoins, notamment en termes de disponibilité et de maîtrise des données. Passer d'une approche manuelle à un modèle automatisé peut être un défi, car un modèle n'a pas d'expertise métier intrinsèque. L'instinct humain et l'expérience ne peuvent pas toujours être intégralement transmis à un modèle.

Après avoir défini les besoins, la faisabilité technique est évaluée. Cette phase, appelée POC (Proof of Concept), consiste à tester différentes approches pour déterminer la meilleure solution. Si une approche est validée, le projet entre dans une phase de gestion de projet traditionnelle, avec planification, budget, réunions, etc. L'implémentation est suivie de l'industrialisation, puis de la mise en production. Une fois en production, le métier prend le relais, et une surveillance continue est mise en place pour s'assurer de la cohérence des données.

Il est également crucial de surveiller les performances et de réentraîner le modèle si nécessaire. La transition de méthodes manuelles à l'utilisation de l'IA nécessite également une gestion du changement pour aider les utilisateurs humains à s'adapter. Enfin, le suivi des performances est essentiel pour garantir l'efficacité continue du système.

Interview Grégoire COLOMBET (IBM)

Cas d'usage et erreurs rencontrées

La gestion de la fraude, la lutte contre le blanchiment, la santé, l'assistance au diagnostic et la gestion de la qualité dans l'industrie sont des domaines où l'intelligence artificielle est étudiée. Chez CACIB, l'accent était mis sur l'analyse textuelle et transactionnelle. Actuellement, les organisations reconnaissent les avantages potentiels de l'IA tout en craignant ses risques. La solution développée vise à déterminer quand on peut faire confiance à une décision algorithmique par rapport à une décision humaine. L'objectif est d'évaluer les avantages de l'automatisation tout en acceptant certains risques. L'outil fourni permet d'évaluer les gains potentiels de l'IA pour un cas spécifique, aidant ainsi les décideurs à déterminer s'ils doivent ou non intégrer ce type de système.

Robustesse :

On m'a effectivement à disposition des capacités de calcul qui permettent de déterminer quand est-ce que nos analyses sont plus ou moins robustes.

Explicabilité :

Il n'est pas toujours essentiel de comprendre comment ou pourquoi un modèle a pris une décision spécifique. Parfois, la question clé est de savoir pourquoi nous devrions privilégier une recommandation algorithmique plutôt qu'un jugement humain. De nombreuses décisions humaines sont prises sans que l'on se demande pourquoi ou comment, contrairement aux décisions algorithmiques. Bien que comprendre le fonctionnement d'un système soit important, il est essentiel de pouvoir expliquer aux dirigeants, régulateurs et collaborateurs pourquoi nous devrions faire confiance à un système algorithmique plutôt qu'à un jugement humain.

Son livre :

De quoi parle le livre en résumé ?

Quand on fait des systèmes qui étaient, qui touchaient après toutes les cases en termes de de performance, de responsabilité, d'éthique en matière d'intelligence artificielle, tous ces systèmes n'étaient pas utilisés de la manière dont on imaginait Pour la simple et bonne raison que au final, on y avait une sorte de dogme qui fallait toujours faire collaborer l'homme et la machine.

Alors que en fait, en pratique, quand on était sur des projets, quand on mettait à disposition ces outils, la collaboration machine c'était clairement pas quelque chose d'évident. Typiquement chez CACIB, on avait fait un projet où on avait mis toutes les informations qui sortaient du modèle d'intelligence artificielle et les utilisateurs les comprenaient. ça ne les amenait pas à prendre de meilleures décisions. le bouquin retracé en fait les le cheminement intellectuel de mon équipe.

On s'est interrogé empiriquement pour savoir si, effectivement la collaboration machine était systématiquement meilleure que l'humain seul ou la machine seule et ce qu'on a montré, c'est que c'était pas le cas, que y avait que certaines conditions ou la collaboration entre l'humain et la machine est meilleure que dans certains cas, la machine devait prendre des décisions, c'est à dire réaliser des tâches automatiquement et que dans certains cas, il fallait surtout pas intégrer l'intelligence artificielle dans le raisonnement humain, c'était contre-productif en termes de performance et donc ensuite, une fois qu'on a dit ça, on se pose un peu plus sur savoir comment.

Parce qu'on peut construire une collaboration entre l'humain et la machine qui soit efficace d'un point de vue performance d'un point de vue robustesse et d'un point de vue d'autonomie.

Et ensuite on se pose des questions de savoir. une fois qu'on a dit ça, qu'est-ce que ça veut dire en termes d'évolution des compétences, d'évolution des métiers, d'évolution des rôles peut-être plus largement sur l'économie.

Lors de la création de systèmes d'intelligence artificielle performants, responsables et éthiques, ces systèmes n'étaient pas toujours utilisés comme prévu. La raison principale était un dogme selon lequel l'homme et la machine devaient toujours collaborer. Cependant, en pratique, cette collaboration n'était pas toujours évidente. Par exemple, chez CACIB, un projet a été mis en place où toutes les informations issues du modèle d'IA étaient fournies aux utilisateurs. Bien qu'ils comprenaient ces informations, cela ne les aidait pas à prendre de meilleures décisions.

Le livre retrace le parcours intellectuel de l'équipe, qui s'est interrogée sur la valeur ajoutée réelle de la collaboration homme-machine. Les résultats ont montré que cette collaboration n'était pas toujours bénéfique. Dans certains cas, la machine était plus efficace seule, tandis que dans d'autres, l'intégration de l'IA dans le raisonnement humain était contre-productive.

Après avoir identifié ces constats, l'équipe s'est concentrée sur la manière de construire une collaboration efficace entre l'homme et la machine en termes de performance, robustesse et autonomie. Cette réflexion a également soulevé des questions sur l'évolution des compétences, des métiers et, plus largement, de l'économie à l'ère de l'IA.

Exemples : Dans le processus de détection de la fraude, une première analyse est effectuée pour chaque transaction potentiellement frauduleuse. Si une transaction semble risquée, une investigation plus approfondie est lancée. Les modèles d'intelligence artificielle sont de bons indicateurs de la probabilité de fraude. Cependant, lorsque cette probabilité est modérée (entre 30% et 60%), il est difficile pour un humain de détecter le cas de fraude parmi des milliers d'alertes, ce qui entraîne une perte de temps. Si le modèle prédit une très forte probabilité de fraude (plus de 95%), il est inutile de soumettre la transaction à une première analyse, car cela augmente le risque d'erreur humaine. Dans

ce cas, la transaction devrait être directement envoyée à une équipe spécialisée en détection de fraude. Pour les cas intermédiaires, l'intelligence relationnelle des collaborateurs est essentielle. Par exemple, un appel au client peut permettre de déterminer, à travers le ton de sa voix, s'il s'agit d'une fraude. C'est quelque chose qu'une machine ne peut pas faire.

Confiance / Fiabilité :

l y a deux choses à prendre en considération, le principe de confiance et de la fiabilité.

C'est à dire qu'on peut se reposer sur l'IA c'est quelque part de manière implicite,

Si je me repose ce lien, c'est que je me repose plus sur l'humain, donc il faut globalement pour avoir confiance, être en capacité d'expliciter extrêmement précisément et objectivement que le système utilisé va produire de meilleurs résultats que si on avait sollicité l'interprétation humaine.

Donc l'idée de confiance, c'est pour moi s'assurer que lorsqu'on utilisé ce type de système. on le fasse en sachant précisément quels sont les bénéfices de manière générale, du système, lorsqu'il est utilisé.

On s'est posé une question fondamentale sur lesquelles je travaille beaucoup avec les comités d'éthique, IBM, c'est qu'est-ce que ça veut dire qu'augmenter l'intelligence humaine avec intelligence artificielle ?

Interview David CORTES (ALVIDENCE)

L'IA Responsable dans la banque

La confiance en l'IA est essentielle pour son adoption. Si on ne comprend pas ou si on n'a pas confiance en ces outils, on ne les utilisera pas. Dans le secteur bancaire, cette confiance doit être renforcée, notamment en raison des exigences réglementaires. Les banques seront soumises à des audits et devront fournir des documentations, reflétant le rôle crucial du régulateur. Elles doivent se conformer à diverses règles, en particulier lorsque les modèles influencent des décisions financières majeures, comme la solvabilité. Malgré les impératifs commerciaux, les banques doivent fournir des preuves tangibles aux régulateurs. Selon la réglementation européenne AI ACT, qui devrait être adoptée début 2024, les banques devront assurer la qualité de leurs modèles, en particulier pour les applications à haut risque comme l'octroi de crédits. Cela impliquera une documentation complète, des audits internes et externes, la capacité de reproduire les transactions, la stabilité des prévisions, la gestion des versions et le respect total de la réglementation.

AI Responsable c'est quoi ?

Ma vision sur ce qu'il y a de l'IA de confiance, c'est quelque chose qui est très vertueux, qu'il faut faire parce qu'au niveau opérationnel, c'est indispensable pour que les projets se passent bien et que la confiance établie, si on veut passer à l'IA éthique, il faut absolument le faire aussi, mais que c'est très difficile de vraiment enclencher et mettre tout le monde en mouvement sur ces sujets tant qu'il y a une contrainte forte et la contrainte forte, elle arrive dans 2 ans et demi.

je le résume en 2 points :

1- Quand on a une interface client ou pas : toutes les activités de machine s'appelle Business-To-business, ses besoins sont bien représentés par ce qui est véhicule autonome et autres.

ça sera robustesse. Certification, stabilité, tous ces termes là qu'on retrouve dans les 7 principes.

2-ceux qui ont vraiment une interface client, notamment les banques, les opérateurs mobiles, les assurances et quelques autres, mais y en a pas tant que ça et eux. ils ont vraiment tout le problème de traitement non discriminant, transparence des traitements, attention, équité

Globalement, quand on met les 2 ensemble, on a à peu près tout.

Tout ce qui correspond concrètement, ce qu'on doit faire pour avoir une confiance donc côté très industriel et côté très éthique. Et alors ? Donc au niveau industriel en France il y a un autre collectif Qui s'appelle confiance ai (premier pillier robustesse, performance,..)

1er pilier : robustesse, performance, interviens confiance ai

2e pilier : éthique,

3em pilier l'environnement 'IA Frugale' :

on s'en préoccupe, notamment parce que ceux qui influent sur le texte au niveau français, c'est la DGE, direction générale des entreprises. Et c'est un des axes importants de la DGE depuis un an à peu près. Il y a une volonté française d'influer le texte final dans ce sens-là et le texte actuel fait déjà la part belle à l'ESG. (les considérations environnementales et sociétales).

Donc en général dans une charte éthique, il y a une partie de frugalité, de respect de l'environnement et la dedans intervient l'ESG (RSE responsabilité sociétale et environnementale.)

Comment l'IA peut être de confiance :

il faut qu'elle soit adoptée et acceptée par les régulateurs. Pour qu'elle soit adoptée il faut qu'elle soit éthique

AI Vidence :

David a fondé une entreprise il y a 2 ans, spécialisée dans la fourniture d'outils pour aider les entreprises et éventuellement les futurs auditeurs à comprendre les IA. Ces IA, étant nouvelles et souvent perçues comme des "boîtes noires", ne révèlent pas facilement ce qu'elles ont appris. L'objectif principal de l'entreprise est l'explicabilité de l'IA, avec le slogan "rendre l'IA explicable". Le lien essentiel se situe entre l'utilisateur métier et le Data Scientist.

Erreurs d'IA rencontrées :

Il est fréquemment observé un écart de performance entre la phase d'entraînement des IA et leur mise en production, soulevant des questions de stabilité. Un document d'explicabilité produit avec le régulateur financier ACPR a révélé des problèmes dans les résultats, notamment des biais et des soucis liés à l'entraînement sur des données non représentatives.

Comment assurer la protection des données :

Faire des tests cyber sur des données classiques pour voir si on peut rentrer sur les bases de données comme dans un moulin ou non ? DSI, RSSI qu'il surveillent plus.

Bonnes pratiques :

Il est important de mettre en place des nouvelles gouvernances en se disant il faut savoir quelles sont les données qui sont utilisées par quel modèle

La Gouvernance : c'est à dire que qu'on voit comment ces algorithmes sont entraînés sur certaines données

Ça correspond à une certaine distribution de statistique qui globalement va changer dans le temps. exemple : la situation des emprunteurs en 2007 et en 2009 dans au niveau crédit immobilier, c'était pas du tout la même chose et on voit la crise de 2008 il y a eu un problème de modèle, les distributions statistiques n'étaient pas du tout les mêmes sur des populations et on n'a pas appliqué le bon modèle sur la bonne population.

De façon générale, c'est les données bougent dans le temps et donc les algorithmes ont été entraînés sur un certain contexte et quand ils envoient un autre il raconte n'importe quoi donc ça fait partie de la gouvernance de données.

Respect de la vie privée :

Les IA génératives, en raison de leur grande capacité, peuvent involontairement "mémoriser" et reproduire des éléments des données sur lesquelles elles ont été formées. Cela pose un risque, identifié par des entreprises comme Microsoft, où les données d'apprentissage pourraient être récupérées à partir du modèle. Par exemple, un modèle d'IA générative avec des milliards de paramètres pourrait, après avoir vu quelques phrases d'Harry Potter, reproduire exactement des passages du texte. Ce n'est pas seulement dû à sa capacité de prédiction, mais parce qu'il a en quelque sorte "enregistré" le texte. Face à ce risque, la question se pose de savoir si nous devrions limiter l'utilisation des IA pour éviter des violations de données privées.

Outils utilisés :

MLOPS machine learning opération : il y a des outils qui existent comme MLFlow c'est des databricks.

La certification :

Il existe une tendance à encourager les différents secteurs à s'accorder proactivement sur des outils et des standards conformes à des principes majeurs, ce qui pourrait conduire à une auto-certification. Bien que cela ne soit pas encore décidé, des principes multisectoriels encouragent cette direction, avec une incitation à détailler les outils et méthodes spécifiques à chaque secteur.

Plusieurs labels existent, comme celui du coordinateur national de l'IA et la NLE pour la certification des processus. Ces labels se concentrent sur la gouvernance des données et l'éthique. Obtenir une certification technique est complexe. Actuellement, il existe des "consulting assessments", qui sont essentiellement des avis de consultants, comme ceux fournis par WC. Ces avis donnent une certaine confiance, mais n'ont pas de valeur légale.

À l'avenir, l'objectif est d'obtenir une certification réelle, basée sur des standards techniques précis. Il y a un effort pour transformer ces standards en normes officielles, avec des organisations comme le Senelec travaillant avec l'Afnor pour établir des certifications européennes. C'est un sujet complexe sans certitude.

Des réglementations existent déjà, notamment dans les secteurs bancaire et de l'assurance, qui adaptent leurs méthodes actuelles pour intégrer l'IA. L'essentiel est d'assurer la qualité des données et d'expliquer le modèle résultant. Il est également crucial de s'assurer que les solutions sont éthiquement acceptables.

Enfin, la question de l'éthique est centrale. De nombreux acteurs, comme le Label IA, cherchent à renforcer la position éthique à travers les algorithmes.

Interview Michel BONVIN (CACIB)

L'intégration de l'IA dans les processus opérationnels et décisionnels présente des avantages et des défis. Voici un résumé de vos réflexions :

CAS d'usages de l'IA dans la banque :

1. CACIB Search pour Long Distribution Solutions:
    - Objectif : Évaluer l'éligibilité d'un contrat de titrisation.
    - Méthode : Analyse de contrats pour identifier des clauses spécifiques.
    - Avantages : Automatisation de l'analyse de contrats volumineux, gain de temps pour les professionnels.
2. Moteur de recherche :
    - Objectif : Moteur de recherche pour les contrats.
    - Méthode : Indexation et recherche de texte dans une base de données de contrats.
    - Avantages : Accès rapide et efficace à l'information pertinente.
3. Search pour CF à certification of financing:
    - Objectif : Vérifier la conformité des contrats.
    - Méthode : Extraction d'informations spécifiques des contrats.
    - Avantages : Simplification du processus de vérification.

Outils et Techniques:

- CACIB Search : Un ensemble de briques technologiques pour construire des solutions adaptées à l'usage.

- Chatbot : Utilise le NLU pour comprendre les questions et le NLG pour générer des réponses.

L'IA est utilisée pour résoudre des problèmes que l'informatique traditionnelle ne peut pas.

L'efficacité de l'IA dépend de la qualité des données. Les données doivent porter un "signal" pour être utiles.

Problématiques:

1. Efficacité opérationnelle : Utiliser l'IA pour accélérer les processus manuels.

2. Développement des affaires : Utiliser l'IA pour générer de nouvelles opportunités commerciales.

Perceptions des utilisateurs :

Les utilisateurs ont des attentes élevées envers l'IA, souvent plus élevées que celles qu'ils ont envers les humains.

Ils comprennent que l'IA s'améliore avec le temps et l'utilisation.

Risques :

Les professionnels des risques évaluent les dangers potentiels associés à l'utilisation de modèles d'IA.

L'IA offre des avantages significatifs en termes d'efficacité et de capacité à traiter des volumes importants de données. Cependant, son intégration nécessite une compréhension claire des attentes, des risques et de la qualité des données utilisées.

Interview Ghassen BETTAIEB (Datalab Groupe Crédit Agricole)

Responsabilité :

Les personnes qui créent l'algorithme sont responsables de son apprentissage et de la confidentialité des données utilisées. Il est essentiel de s'assurer que toutes les parties impliquées dans le projet soient en accord sur les données à utiliser.

Sécurité des données :

Il y a des modèles qui ont appris sur une base de données, mais qui peuvent mémoriser ces données là et qu'on peut les retrouver après à partir du modèle,

Il faut garantir qu'il n'y ait pas de fuite de données d'apprentissage, notamment face aux attaques adverses. Ces attaques peuvent permettre de récupérer des données sensibles qu'il a déjà vu avant dans la base d'apprentissage à partir du modèle. Il est donc crucial de ne pas simplement sauvegarder les données, mais d'apprendre à partir des motifs de ces données pour effectuer des prédictions.

Les biais :

Un autre aspect essentiel est de s'assurer que le modèle n'est pas biaisé. Par exemple, si un modèle est formé sur une base de données contenant plus d'hommes que de femmes, il pourrait être biaisé en faveur des hommes. Les Data Scientists doivent donc être vigilants pendant la phase d'apprentissage pour garantir l'absence de biais dans le modèle.

Traçabilité :

Faut que ce soit **reproductible**. Enfin, quand on crée des algorithmes, il faut qu'on s'assure que ce soit reproductible. Enfin, si on arrive à avoir des résultats, on a l'algorithme. Si demain on va reproduire les résultats, faut qu'on soit sûr qu'on peut reproduire ces résultats. Donc faut bien tracer les expérimentations, les paramètres qu'on a utilisés pour créer tel algorithme.

Il faut tout tracer pour qu'on perde pas de temps après à refaire d'autres modèles qui sont similaires au modèle qu'on a déjà fait, parce que si on perd à chaque fois, si on va faire beaucoup d'expérimentations à chaque fois, les expérimentations, ça consomme de l'énergie

Nouveaux algorithmes :

Historiquement, pour obtenir des modèles robustes, il fallait qu'ils soient formés sur une grande quantité de données. Cela était dû au fait que les modèles initiaux n'avaient aucune connaissance préalable de la tâche à accomplir. Cependant, aujourd'hui, cette approche a évolué. On peut désormais utiliser des modèles pré-entraînés ayant déjà une connaissance de base sur la tâche, réduisant ainsi le besoin de grandes quantités de données pour l'entraînement.

Par exemple, pour la classification de documents, on peut utiliser un modèle pré-entraîné qui a déjà été formé pour lire et comprendre des textes. Ce modèle peut ensuite être affiné avec une petite quantité de données pour une tâche spécifique.

Lors du choix d'un modèle, il est essentiel de connaître la base de données sur laquelle il a été formé pour anticiper d'éventuels biais. Certains algorithmes, comme ceux en open source, sont transparents sur les données utilisées pour l'entraînement, ce qui est bénéfique pour comprendre et expliquer le comportement du modèle. Cependant, d'autres, comme le chat GPT, ne partagent pas ces informations, ce qui peut poser des problèmes.

Il est donc crucial de choisir judicieusement l'algorithme et de comprendre les données sur lesquelles il a été formé. De nombreux modèles sont accompagnés d'articles scientifiques qui détaillent les bases de données utilisées, permettant ainsi une meilleure compréhension et évaluation des éventuels biais.

Données :

La donnée est au cœur de toute démarche d'apprentissage automatique.

Données Synthétiques : Bien que l'idée de générer des données synthétiques pour pallier les problèmes de confidentialité soit bien, en pratique, cela ne fonctionne pas toujours. Les données synthétiques ne parviennent pas à reproduire fidèlement la réalité des données réelles, entraînant un écart de performance lorsque les modèles formés sur ces données sont testés sur des données réelles.

Confidentialité et Sécurité : Lors de l'utilisation de données réelles, en particulier celles contenant des informations sensibles, des conventions sont signées au début du projet pour garantir la sécurité et la confidentialité. Les données sont stockées de manière sécurisée, et les modèles sont conçus pour ne pas mémoriser ces informations sensibles. Une fois le projet terminé, les données sont conservées pendant une durée définie (par exemple, un an ou deux) avant d'être supprimées.

Réglementation : Il est essentiel que tout soit bien documenté et que toutes les réglementations soient clairement définies. Les données sont utilisées telles quelles, sans anonymisation, mais des mesures sont prises pour garantir leur protection. Les modèles apprennent à partir de ces données sans les mémoriser, créant ainsi une abstraction pour effectuer des prédictions. Des conventions juridiques sont signées, et les données ne sont pas conservées indéfiniment.

Équité

Il est très important de garantir un équilibre entre les différentes catégories ou classes dans une base de données pour éviter les biais. Par exemple, si une base de données contient 90% de femmes et 10% d'hommes, le modèle risque d'être biaisé en faveur des femmes.

Pour remédier à ce déséquilibre, plusieurs approches existent :

- Sous-échantillonnage : Cette technique consiste à réduire la catégorie dominante pour qu'elle corresponde à la catégorie minoritaire, réduisant ainsi la taille totale de la base de données.

- Sur-échantillonnage : Elle implique de multiplier ou de répéter la catégorie minoritaire pour qu'elle atteigne la taille de la catégorie dominante. Cependant, cela peut conduire à la création de données quasi-synthétiques, ce qui peut affecter la capacité du modèle à généraliser correctement.

Interview Matthieu CAPRON (Datalab Groupe Crédit Agricole)

De l'IA responsable chez Crédit Agricole depuis quand ?

Depuis février 2023, toutes les applications en production suivent une méthode certifiée, s'inscrivant ainsi dans une démarche d'IA responsable. Bien que cette démarche ait été officialisée par une certification, les projets étaient déjà menés selon cette approche auparavant. La certification a permis d'améliorer certaines étapes du processus, mais l'approche globale était déjà orientée vers la responsabilité.

Des cas d'usage de l'IA responsable chez CA ?

- Projet pour détecter précocement les clients susceptibles de présenter des situations de fragilité financière afin de leur proposer un accompagnement adapté.
- Projet autour de la classification de documents. On reçoit des pièces d'identité ou des justificatifs de domicile d'identité pour les entrées en relation et on doit classifier ce ces documents.
- Projet d'analyse de verbatim issus des enquêtes de satisfaction

Green - L'empreinte carbone dans l'IA :

- Mesure de l'empreinte carbone : L'empreinte carbone est désormais un critère de choix pour les modèles d'IA. Si deux modèles offrent des performances similaires, le modèle consommant le moins de CO2 est privilégié.
- Choix du modèle :
    - La taille du modèle influence sa consommation énergétique. Un modèle plus grand consomme plus d'énergie.
    - Il est essentiel de trouver un équilibre entre la performance du modèle et sa consommation énergétique.
- Paramètres du modèle :
    - Les paramètres du modèle, comme le nombre d'arbres dans un Random Forest, peuvent influencer son empreinte carbone.
    - Il est crucial d'optimiser ces paramètres pour minimiser l'empreinte tout en conservant une performance acceptable.
- Données d'entrée :
    - Réduire le nombre de données d'entrée peut diminuer l'empreinte carbone sans affecter significativement la performance.
    - Par exemple, en utilisant seulement 10% des variables les plus pertinentes, on peut parfois obtenir des performances presque identiques.
- Infrastructure :
    - Le choix entre le cloud, les serveurs dédiés ou les machines locales peut influencer l'empreinte carbone.
    - L'utilisation de GPU est plus énergivore, donc elle est limitée aux tâches nécessitant une puissance de calcul élevée.
- Initiatives de CAGIP :
    - CAGIP travaille à réduire l'empreinte carbone de ses data centers.
    - Il existe une réflexion chez AIDA sur l'évaluation de l'empreinte carbone entre les infrastructures de CAGIP et le cloud.

Explicabilité des modèles d'IA :

Les modèles d'IA varient en termes d'explicabilité. Avant de choisir un modèle, il est essentiel de déterminer si le besoin métier nécessite une explicabilité ou non.

- Types d'explicabilité :
    - Explicabilité globale : Comprendre les principales variables ou caractéristiques qui influencent le modèle dans son ensemble.
    - Explicabilité locale : Expliquer la décision prise par le modèle pour une instance spécifique.
- Méthodes d'explicabilité :
    - Taux de confiance : Le modèle fournit une prédiction avec un niveau de confiance, indiquant à quel point il est sûr de sa décision.
    - Heatmap : Pour les images, des techniques montrent sur quelle partie de l'image le modèle s'est concentré pour prendre sa décision.
    - -Outils spécifiques : Des outils comme "Grad CAM" peuvent montrer sur quelle partie d'une image le modèle s'est basé pour sa décision.

- Détails de l'explicabilité :
    - Les modèles peuvent fournir des détails sur les variables les plus influentes dans une décision.
    - Ils peuvent également indiquer si une variable a eu un impact positif ou négatif sur la décision.
    - Certains modèles peuvent même fournir des détails sur la manière dont une valeur spécifique d'une variable influence la décision.
- Complexité de l'explicabilité :
    - Si un modèle utilise de nombreuses variables pour prendre une décision, il peut être difficile pour un humain de comprendre la combinaison de toutes ces variables.
    - Dans de tels cas, il peut être préférable de se concentrer sur un nombre réduit de variables pour faciliter la compréhension.

Boîte Blanche et Explicabilité des Modèles d'IA :

Choix du modèle : Le choix entre un modèle "boîte blanche" (explicable) et un modèle "boîte noire" (moins explicite) dépend des besoins. Si l'explicabilité est prioritaire, il est préférable de choisir un modèle naturellement explicite.

Compromis entre performance et explicabilité :

- Si un modèle performant est nécessaire, mais qu'il est moins explicite, il faut trouver un équilibre entre sa performance et son explicabilité.

- Dans certains cas, pour des usages à haut risque, l'explicabilité sera obligatoire, même si cela signifie sacrifier une partie de la performance.

Méthodes pour améliorer l'explicabilité :

- Utiliser des modèles par nature explicables.

- Ajouter des couches d'explicabilité à des modèles moins transparents.

- Réduire le nombre de variables pour simplifier l'explication.

Exemple concret :

Un cas d'usage a été mentionné où un réseau de neurones a été utilisé pour déterminer le besoin de mise à jour d'un client. Pour expliquer les décisions du réseau de neurones, un modèle Random Forest a été utilisé pour identifier les variables principales.

Importance de la communication :

Poser ces questions et clarifier les attentes est crucial pour l'acceptation du projet.

Cela montre que l'IA n'est pas utilisée de manière arbitraire, mais est soigneusement considérée pour répondre aux besoins spécifiques. Cela aide également à rassurer les parties prenantes que l'IA n'est pas là pour remplacer les métiers, mais pour les compléter.

Utilisation d'algorithmes pré-entraînés :

Pour accélérer le processus, il peut être tentant d'utiliser des algorithmes déjà entraînés. Cependant, cela pose le problème de ne pas toujours savoir sur quelles données ils ont été formés, ce qui peut introduire des biais.

L'approche privilégiée est de ne pas s'appuyer sur des algorithmes dont les données d'entraînement ne sont pas maîtrisées. L'exception pourrait être pour les modèles génératifs, où des expérimentations sont en cours.

Maîtrise complète :

- Dans 99% des projets, l'approche adoptée est de concevoir les modèles de A à Z. Cela implique de prendre des données maîtrisées, d'entraîner un modèle spécifiquement pour le projet, et d'assurer un résultat contrôlé.
- Même si un algorithme moins explicite est utilisé, il y a une maîtrise complète des données, de la formation de l'algorithme, et de la vérification de l'absence de biais.

Il faut avoir une maîtrise totale du processus, de la collecte des données à la formation du modèle, afin d'assurer des résultats fiables et sans biais.

Responsabilité Juridique :

Estimation du Risque :

Avant la mise en œuvre d'un modèle, une évaluation du risque associé est effectuée, notamment via le Modèle Risk Management (MRM), qui estime le risque bancaire lié au modèle.

Contrôle Humain :

- Malgré l'automatisation, il y a toujours une intervention humaine pour superviser et valider les décisions prises par les algorithmes.
- Dans tous les projets, aucune décision automatisée n'est prise sans être revue par un humain. La responsabilité humaine demeure primordiale.
- Avec l'introduction de l'AI ACT, la responsabilité humaine restera centrale, en particulier pour les algorithmes à haut risque.

Le principe fondamental :

La responsabilité d'un algorithme est le résultat d'une combinaison de plusieurs facteurs, notamment la qualité des données, la transparence de l'algorithme et le contrôle humain. Toutes ces étapes garantissent la mise en place d'un algorithme responsable.

Au final, la responsabilité dans l'utilisation des modèles d'IA repose sur une évaluation rigoureuse des risques, une intervention humaine constante et une attention particulière à la qualité des données utilisées.

Avantages de l'IA Responsable :

- Confiance :
  L'IA peut susciter des craintes, notamment concernant le remplacement des emplois humains. Une IA responsable renforce la confiance en montrant qu'elle est conçue pour assister l'humain et non le remplacer.
- Acceptation et Démocratisation :
  Une IA responsable facilite son acceptation par les utilisateurs. Par exemple, une grande banque française a dû encourager l'utilisation d'IBM Watson en offrant des primes, car les employés n'avaient pas confiance en l'outil.
- Conformité :
  Avec des réglementations comme AI ACT, il est essentiel d'être en conformité. L'IA responsable garantit cette conformité.
- Adaptabilité :
  Les projets d'IA doivent être capables de s'adapter aux évolutions, comme les changements de comportement dus au COVID. Cela n'est pas nécessairement lié à la présence de biais, mais à la nécessité d'ajuster les modèles en fonction des nouvelles données.
- Monitoring :
  Bien que ce ne soit pas encore systématique, l'intégration du monitoring dans les projets d'IA est une démarche essentielle pour assurer une surveillance continue et une mise à jour automatique des modèles.

Interview Céline LUNET VAUTRIN (CACIB)

Gestion des Réglementations de l'IA:

Analyse Globale :

L'équipe IA a examiné les réglementations existantes à l'échelle mondiale pour définir des recommandations spécifiques à l'IA.

Principes Clés :

- Robustesse : Assurer la fiabilité et la stabilité des modèles.
- Data & Privacy : Garantir la confidentialité et la sécurité des données, un principe déjà intégré dans les pratiques courantes.
- Transparence & Interprétabilité : Il est essentiel de pouvoir expliquer comment l'IA arrive à ses conclusions.
- Proportionnalité : Adopter une approche basée sur les risques, en mettant en place des contrôles adaptés à la criticité du modèle.
- Intégration Humaine : Assurer que les décisions de l'IA peuvent être contestées et validées par des humains.
- Responsabilité : Définir qui est responsable en cas d'erreur de l'IA.
- Diversité & Non-discrimination : Être conscient des biais potentiels introduits par l'IA et mettre en place des mesures pour les gérer.

Phase Actuelle de l'IA : après une période d'expansion rapide de l'IA, l'industrie entre dans une phase de stagnation. Cette pause est nécessaire pour comprendre pleinement les risques et les limitations de l'IA.

Confiance :

Il est essentiel d'établir un cadre de confiance pour l'utilisation de l'IA, en particulier à mesure que la technologie évolue.

Intégration de l'Équipe :

L'équipe est impliquée dès le début du projet pour garantir une mise en œuvre réussie.

Pour le MRM, l'adoption responsable de l'IA nécessite une compréhension approfondie des réglementations, des principes éthiques et des risques associés pour garantir des résultats fiables et éthiques.

Cadre de l'IA et Gestion de Projet :

- Analyse des Recommandations :
  À partir de 940 recommandations initiales, un framework d'IA a été créé avec 89 recommandations essentielles, réduisant ainsi les redondances entre différentes régions et pays.
- Répartition des Recommandations :
  Ces 89 recommandations ont été attribuées à différents propriétaires ou "honneurs" au sein de CACIB, en fonction de leur pertinence. Par exemple, certaines recommandations ont été attribuées au juridique, à l'IT, ou à la gestion des risques.
- 68% des recommandations sont gérées par seulement trois honneurs. 23 sont spécifiques aux projets, 23 à l'IT et 22 aux risques.
- Document MRM (Modèle Risk Management) :
  Un document clé créé pour guider la mise en œuvre de l'IA. Il couvre divers aspects tels que la justification de l'utilisation de l'IA, l'organisation autour de l'IA, la conception du modèle, les données utilisées, les tests, la gestion des biais, la formation et la place de l'IA dans le processus global.
  Il sert également de guide pour répondre aux obligations réglementaires.
- Amélioration Continue :
  Une méthodologie d'amélioration continue est en cours de formalisation, en particulier pour la phase de fonctionnement de l'IA. L'objectif est de mieux documenter et structurer les audits et de définir des indicateurs pour surveiller l'IA.
- Conformité et Réglementation :
  Les 89 recommandations sont le minimum nécessaire pour garantir la conformité avec les réglementations actuelles. Ces recommandations sont également classées par catégorie, comme la gouvernance, l'externalisation, ou le design et le développement.
- Responsabilité :
  Chaque recommandation est attribuée à un propriétaire spécifique, garantissant ainsi que les responsabilités sont clairement définies.

Gouvernance et Amélioration Continue :

- Une gouvernance d'amélioration continue est mise en place pour surveiller et anticiper les déviations de performance des modèles d'IA.

- Des indicateurs sont utilisés pour détecter les dérives potentielles des modèles.

- Bien que des indicateurs et des tableaux de bord existaient déjà, une structuration et une professionnalisation supplémentaires sont en cours.

Documentation et Formation :

- Des documents sont en cours de finalisation pour guider le processus de formation des modèles. Ces documents détailleront les éléments à documenter lors de la formation, tels que la problématique adressée, le besoin métier, le dataset utilisé, la gestion des biais, les résultats du modèle, les tests, etc.

- Ces documents serviront de référence pour garantir la conformité avec les recommandations réglementaires.

Collaboration Interdépartementale :

- Pour les recommandations qui ne relèvent pas directement du projet, d'autres départements de CACIB seront impliqués, tels que le senior management, les risques, l'audit et l'IT.

- L'objectif est de collaborer avec ces départements pour évaluer les pratiques actuelles par rapport aux recommandations réglementaires, identifier les écarts éventuels et définir des mesures correctives.

Interview Christophe MOULIS (CACIB)

problématiques rencontrés :

- réduction du travail

- optimisation de de la charge de travail

- augmentation de l'efficacité des équipes

- optimiser des gains

Prise de décision par l'IA :

- Propositions et Corrections :

    - L'IA suggère des actions ou des classifications, mais une intervention humaine est souvent nécessaire pour valider ou rectifier ces suggestions. Par exemple, si l'IA classe mal un e-mail, l'utilisateur peut le reclasser correctement. L'IA apprend ensuite de cette correction.

- Apprentissage Continu :
    - L'IA est régulièrement mise à jour et formée. Les corrections apportées par les utilisateurs sont intégrées dans son entraînement pour améliorer sa précision. Par exemple, si l'IA commet une erreur dans la classification d'un e-mail, cette erreur est corrigée et utilisée pour améliorer ses prédictions futures.

- Amélioration Constante :
    - Grâce à ces corrections et à l'entraînement régulier, l'IA s'améliore constamment, augmentant ainsi sa précision et sa fiabilité dans la prise de décision.

Gestion et Préparation des Données :

- Source des Données :
  Identifier la provenance des données est primordial, surtout si elles sont disponibles à plusieurs endroits. Il est essentiel de déterminer qui les génère et qui en est responsable.
- Sécurité et Droits :
  Des enjeux liés à la sécurité des données et à leurs droits d'utilisation doivent être pris en compte.
- Nettoyage et Connexion :
  Les données nécessitent souvent une préparation avant utilisation, comme le nettoyage, la connexion avec d'autres sets de données (par exemple, lier des données clients à des données financières) ou la transformation.
- Importance de la Préparation :
  Pour une utilisation efficace dans l'IA, les données peuvent nécessiter des ajouts, des transformations ou des reformats.
- Synthèse des Données :
  Dans certains cas, il est utile de regrouper plusieurs données en une seule pour simplifier leur traitement.
- Données Synthétiques vs Calculées :
  Si les données brutes ne suffisent pas, des traitements peuvent être nécessaires pour obtenir des données calculées, qui seront ensuite exploitées par les algorithmes.

Explicabilité :

Certains modèles sont basés sur des arbres de décision. Lorsqu'on les forme, ils créent des structures d'arbres de décision qui sont ensuite utilisées pour analyser de nouvelles données. Ces modèles offrent une certaine transparence car on peut suivre le chemin emprunté par l'arbre pour arriver à une décision. Cependant, il peut être complexe de comprendre entièrement le processus, car bien que nous puissions voir les conditions utilisées dans l'arbre, les chemins exacts restent souvent compliqués à saisir.

MLOPS:

L'objectif est de mettre en place une chaîne automatisée qui facilite la transition du développement à la production, c'est le concept du MLOPS. Bien que cela résolve certains problèmes, il n'est pas une

solution universelle. Avant, les développeurs travaillaient de manière isolée sur leurs ordinateurs, mais avec l'évolution vers l'intégration continue et la livraison continue, les processus sont devenus plus automatisés et sécurisés. Pour l'IA, de nombreuses tâches étaient réalisées sur des Notebooks Python par des Data Scientists, mais cela ne conduisait pas à des applications industrielles robustes. Le monitoring des modèles en production nécessite que les applications soient bien identifiées et suivent un processus de livraison strict. Le MLOps joue un rôle crucial ici, car il intègre des validations métier, de conformité et de sécurité dans le flux de travail, de la phase de développement à la mise en production.

Notice d'utilisation de l'usage de Chat GPT

ChatGPT a été utilisé en tant qu'aide à la rédaction pour :

- Résumer des parties des interviews
- Résumer des textes ou des pdfs
- Extraire des informations intéressantes d'un pdf.

Voici quelques exemples de prompts utilisés:

Résume moi ce texte …

A partir de ce pdf, réponds à cette question …