L'avenir du développement de l'Intelligence Artificielle (IA) passe par la mise à disposition d'IA responsables, dignes d'humanité et de confiance

SOMMAIRE

- Eléments probants
- Limites de l'efficacité des outils disponibles

ii. Equité, non-discrimination, erreurs et biais

- Problématique générale d'équité et non-discrimination, d'erreurs et biais
- Une notion d'équité, de non-discrimination et de biais multiforme complexe à prendre en compte
- Erreurs de raisonnement et biais cognitifs : corrélations faussement trompeuses et causalités
- Erreurs de raisonnement et biais cognitifs : choix et calibration des variables
- Erreurs de raisonnement et biais cognitifs : données d'apprentissage non représentatives
- Erreurs et biais techniques : maturité d'apprentissage
- Erreurs et biais techniques : faux positifs et faux négatifs
- Erreurs et biais techniques : évènements exceptionnels ou cygnes noirs
- Erreurs et biais cognitifs : restitutions
- Erreurs de simplification et d'approximation
- Réponses aux biais
- Limite des outils
- Emettre une opinion sur ce type d'assertion : un des défis les plus critiques des IA

d. Les moyens nécessaires pour des IA responsables, dignes d'humanité et de confiance : référentiels, outils et bonnes pratiques pour créer de la valeur

i. Démarche et dispositif de gouvernance à mettre en place pour évaluer les options, arbitrer et fixer les orientations en matière d'IA responsables, dignes d'humanité et de confiance
ii. Dispositif de management et opérationnel performant
iii. Une démarche efficace qui se doit d'être intégrée, holistique, contextuelle et dynamique
iv. Une démarche simple, ouverte et flexible et alignée aux différents référentiels et standards basée sur des modèles adaptés
v. Des bonnes pratiques spécifiques pour répondre aux exigences et spécificités des IA

e. La nécessité d'être rassuré par un tiers indépendant

i. Définition de la proposition de valeur adaptée pour la certification des IA
ii. Convenir des éléments qui permettront de définir l'objectif d'un audit donné
iii. Convenir de la nature et du niveau des travaux d'audit
iv. Illustration des éléments à convenir pour un commissariat aux comptes
v. Démarche à mettre en œuvre concernant l'audit des IA

7. Les recommandations et contributions potentielles du groupe de travail

a. Définir un socle commun de compréhension et en particulier concernant les exigences attendues des IA

b. Elaborer une démarche efficace de développement et de mise à disposition d'IA responsables, dignes d'humanité et de confiance en portant une attention particulière sur le dispositif de gouvernance et d'affectation des responsabilités, sur celui de gestion de la prise de risque et sur les modèles spécifiques en soutien à cette démarche.

c. Définir les tests adaptés par catégorie d'IA pour la phase d'apprentissage et la phase opérationnelle et les outils permettant de réaliser ces tests. Définir aussi les tests d'algorithmes à effectuer dans le cade des audits.

d. **Identifier la manière de prendre en compte les exigences spécifiques aux IA et en particulier la transparence et, la non-discrimination et les biais. Concernant la transparence, requérir l'explicabilité « by design » et exiger la traçabilité et la piste d'audit (audit trail) des traitements. Concernant la non-discrimination et les biais, limiter les discriminations du fait de biais par application de bonnes pratiques de développement et d'analyse des données, rendre les « biais volontaires » transparents et mesurés et limiter les « biais techniques » en publiant le degré de maturité de l'apprentissage.**
e. **Obtenir une « certification » par un tiers de confiance pour les IA « sensibles », proposer des exemples d'assertions qui pourraient être certifiées et élaborer des guides d'audit (questionnaires, checklists, matrices, …).**

8. **De nombreuses contributions complémentaires par d'autres organisations sont en cours et pourront contribuer à traiter ces problématiques**

L'avenir du développement de l'Intelligence Artificielle (IA) passe par la mise à disposition d'IA responsables, dignes d'humanité et de confiance

1. Introduction

a. Le contexte des travaux

Un groupe de travail de l'Académie des Sciences Comptables et Financières a réuni de nombreux experts d'horizons très variés (scientifiques, avocats, commissaires aux comptes, auditeurs informatiques, universitaires, …) pour aborder toutes les facettes de la problématique des IA responsables, dignes d'humanité et de confiance. De telles IA sont, pour de nombreux Etats, organisations internationales et acteurs de l'écosystème des IA, un impératif pour le développement d'IA pleinement participatives tant à un développement économique mondial dynamique, innovant et durable qu'au progrès et au bien-être des individus.

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

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

b. Les objectifs du groupe de travail

L'objectif premier du groupe a donc été de comprendre le contexte qui incite cet impératif d'IA responsables, dignes d'humanité et de confiance en précisant la nature de ce besoin, les acteurs intéressés par cette démarche, son intérêt pour eux et les attentes associées.

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

Dans le cadre de ses travaux, le groupe a dégagé plusieurs thèmes spécifiques aux IA, essentiels au bon traitement de cette problématique. Le groupe a étudié ces thèmes nécessitant des approfondissements, a sélectionné les plus importants et ceux où se situaient les plus grandes difficultés selon lui et a tenté d'y apporter une contribution.

L'objectif in fine est de pouvoir disposer :

- d'assertions spécifiques quant aux qualités et caractéristiques attendues d'IA responsables, dignes d'humanité et de confiance
- de référentiels de bonnes pratiques qui si elles sont en œuvre doivent permettre de concevoir et mettre à disposition des IA responsables, dignes d'humanité et de confiance
- de référentiels de bonnes pratiques de certification d'assertions spécifiques à mettre en œuvre par les auditeurs (i.e. commissaires aux algorithmes)

c. Le contenu de ce rapport

Ce document a pour vocation de restituer notre compréhension du contexte et des conditions nécessaires à la mise à disposition d'IA responsables, dignes d'humanité et de confiance et de présenter les thèmes qui méritent des travaux complémentaires. Notre intention est de publier au fil du temps les résultats concernant ces différents autres travaux.

2. Une préoccupation majeure : de nombreuses organisations nationales et internationales se sont récemment mobilisées sur ce sujet

a. Les initiatives internationales

Au cours des derniers mois, plusieurs organisations, nationales et internationales, publiques et privées, représentants des fournisseurs de solutions ou d'utilisateurs des services d'IA, ont proposé des démarches, des principes et des recommandations visant à stimuler leur adoption tout en renforçant leur niveau de confiance. Il s'agit d'établir les fondements d'une approche responsable au service d'IA dignes d'humanité et de confiance.

Ces propositions ont pour vocation de compléter les principes et normes actuels dans les domaines du numérique, de la protection de la vie privée et de la gestion des risques en matière d'IA.

On peut citer notamment la publication d'un document de référence de l'**OCDE** en mai 2019, **« Recommandation du Conseil sur l'Intelligence Artificielle »** adopté par le G20 en juin 2019 et du livre blanc de la **Commission Européenne** en février 2020 **« Intelligence artificielle : une approche européenne axée sur l'excellence et la confiance ».**

b. Leurs constats

Etant donné que les IA pourraient occuper une place de plus en plus conséquente dans tous les aspects de la vie des citoyens, des consommateurs, des entreprises et de l'Etat, et activement contribuer au développement d'une activité économique dynamique et durable et à la résolution des grands défis planétaires qu'ils soient sociaux, sociétaux, démocratiques, de santé, climatiques ou environnementaux, il a été considéré que la concrétisation de ces nombreux apports potentiels ne pourra se faire que si les différentes parties prenantes pourront avoir confiance dans les résultats issus de cet écosystème.

De nombreux freins sont néanmoins présents et de nombreux nouveaux risques pourraient casser cette dynamique de confiance alors que cette confiance est un facteur clé à la diffusion et à l'adoption de ces IA. Sans confiance, pas de croissance.

c. Leurs premières orientations

Ces organisations ont donc énoncé les principes et exigences qui devraient permettre le développement d'IA et d'écosystèmes associés de confiance. Il s'agit par exemple de transparence, d'équité, de non-discrimination, de sécurité, de sureté, de protection de la vie privée, de respect des droits fondamentaux, de robustesse, de responsabilité, …

Mais elles ne précisent pas la nature exacte et le niveau attendu de ces exigences, le type de pratiques qui pourront favoriser l'atteinte de ces exigences, leurs faisabilités techniques et organisationnels, les difficultés de mise en œuvre, la manière dont on pourra se rassurer de manière raisonnable quant à la présence effective d'IA responsables, dignes d'humanité et de confiance, la manière dont cette confiance sera restituée, les démarches et référentiels associés, …

De nombreux travaux ont donc été entrepris pour permettre d'avancer sur tous ces points.

3. Les nombreuses opportunités de création de valeur vont en dépendre

Le développement des opportunités des IA découle de plusieurs facteurs concomitants :

- des progrès dans l'efficacité des **techniques et outils d'apprentissage** : apprentissage machine, apprentissage profond, apprentissage renforcé, …
- une forte augmentation de la **puissance de traitement** qui dépasse largement les capacités humaines. L'arrivée prochaine de l'informatique quantique ne fera que renforcer cette situation.
- une **quantité et une diversité phénoménales de données disponibles** pour l'apprentissage, du fait notamment de la multiplication des sources et des capteurs (IoT, drones, …) et de leur richesse, et de l'augmentation des capacités de stockage.
- une **multiplication de modèles, d'outils et d'offres de services** qui facilitent les différentes finalités d'usage : l'analyse, le diagnostic, la prédiction, la prescription, la décision et l'action en vue de la résolution d'un nombre croissant de problèmes variés et complexes.
- un **accès beaucoup plus aisé et abordable** à tous ces éléments notamment avec le développement du cloud et des services associés.

Tirer parti de ces opportunités nécessite d'importants efforts d'investissements qui ne pourront se matérialiser que si cet écosystème donne confiance.

4. C'est dans l'intérêt des différentes parties prenantes : sans confiance, pas de développement dynamique

De nombreux acteurs contribuent au développement de cet écosystème :

- des **fournisseurs de solutions IA** telles que des algorithmes, des modèles, des données d'entraînement, des données « opérationnelles », des technologies, des outils, des solutions, des techniques, …
- des **fournisseurs de services IA** qui vont intégrer ces différentes solutions pour proposer des services adaptés à leurs « clients » dans les différents secteurs d'activité.
- des **utilisateurs/consommateurs** de ces services : Acteurs économiques, Etat, Consommateurs/Citoyens
- l'**Etat** qui encourage l'innovation, la compétitivité et l'emploi et qui élabore la règlementation et la fiscalité de l'écosystème
- des **tiers de confiance** tels que des commissaires aux algorithmes qui devront donner confiance dans les résultats des IA

Ces différents acteurs vont faire appel à de nombreux types d'intervenants : sponsors, architectes, modélisateurs, développeurs, data scientistes, spécialistes sécurité, exploitants, auditeurs, …

Tous ces acteurs ont un intérêt commun, la mise en œuvre d'IA responsables, dignes d'humanité et de confiance. Chacun a une part de responsabilité. Pour y arriver, ces acteurs doivent donc définir le cadre qu'il convient de mettre en place. Ce cadre commun devra susciter l'adhésion de l'ensemble des parties.

5. De nombreux obstacles sont néanmoins présents qui risquent de freiner son développement

a. Un sujet mal maîtrisé menant à beaucoup de confusion

i. Pas de socle commun de compréhension

Il existe de nombreux concepts liés aux IA mais les différentes parties prenantes n'ont pas la même définition ou approche pour chacun d'entre eux. Les contours des IA sont incertains et imprécis.

Cela commence par la définition même de l'IA. L'IA est un objet virtuel, constitué d'un ensemble de données et d'algorithmes mais qui ne s'y réduit pas. Cela peut couvrir les systèmes d'IA dits faibles qui effectue des tâches précises telles que jouer aux échecs ou effectuer un diagnostic médical, aux systèmes dits forts basés sur des transferts de connaissance entre différents domaines et pouvant disposer d'une conscience que plusieurs considèrent relever plutôt de la science-fiction. Il peut s'agir de la voiture autonome fonctionnant sans conducteur et effectuant les choix éthiques à la place de l'humain.

Un autre exemple d'absence de socle de compréhension commun pourrait être la signification de la notion de transparence dans le contexte d'IA apprenantes à base d'apprentissage profond. Comment se positionnent les concepts d'explicabilité, justifiabilité, interprétabilité, traçabilité, auditabilité, jouabilité, loyauté, exactitude, responsabilité, etc… ?

ii. Faible maturité du sujet

Il n'y a pas aujourd'hui de consensus sur la règlementation, sur les référentiels ou sur les bonnes pratiques qu'il conviendrait d'adopter. Les outils et techniques qui pourraient faciliter la mise en œuvre d'IA responsables, dignes d'humanité et de confiance restent très limités. Leur efficacité n'est pas encore démontrée.

iii. Interrogation sur la faisabilité actuelle de pouvoir disposer d'IA responsables, dignes d'humanité et de confiance

De plus, que ce soit en termes de transparence, d'équité, de non-discriminations, est-on en mesure de fournir le niveau de confiance attendu ? A-t-on les outils qui permettront de donner une opinion circonstanciée sur le fait que les utilisateurs pourront comprendre et retracer la manière dont l'IA arrive aux résultats ou que les résultats sont équitables et ne présentent pas de discriminations ?

b. Des peurs et des fantasmes qui risque de mener à son rejet

La prise de contrôle ou la substitution des IA sur l'humain voire la disparition de l'humain, la perte du libre arbitre et de l'autonomie, la perte de savoir, de savoir-faire et la destruction de l'emploi, le déclassement d'un grand nombre de personnes jugées non utiles à la société, la mise en place d'un classement social en fonction de ses comportements, les risques potentiels accrus de cyberattaque ou d'atteinte à la vie privée, … sont autant de situations qui peuvent conduire au rejet pur et simple des IA ou tout au moins à fortement freiner leur adoption.

Ainsi, lorsque les IA sont plus fiables que l'humain, qu'elles vous connaissent mieux que vous-même y compris dans les aspects les plus intimes, accepterons-nous de leur déléguer nos décisions ou nos actions (voiture autonome, diagnostic médical, rencontres amoureuses, …) qui pourraient être bien plus performantes en la matière mais dont on aurait aucune explication du pourquoi de la décision finale et pour lesquelles nous perdrions notre libre arbitre et la responsabilité de nos actions.

Accepterions-nous de manière consciente des situations qui pourraient être sous-optimales pour nous permettre d'exercer cette liberté ?

En outre, les humains acceptent plus facilement les biais, les erreurs, les faux raisonnements, … s'ils résultent de processus humains que s'ils résultent d'algorithmes dont ils ne maîtrisent pas complètement le fonctionnement. Il est en effet admis qu'un humain puisse se tromper dans la mesure où il assumera les conséquences de ses actes. Mais dans le cas d'une voiture autonome, sur qui reposera la responsabilité d'une erreur qui aurait des conséquences lourdes ? Aussi, on hésitera à adopter la voiture autonome qui pourrait sauver des millions de vies en réduisant voire en éliminant les erreurs humaines mais qui pourrait provoquer par ailleurs des morts qui auraient été évités par l'homme, même si le solde de vies sauvées s'avérait largement positif.

De très nombreux acteurs considèrent que certains traits intrinsèques à l'humain tels que la biologie (système hormonal, protéines, …) ou la conscience (sentiments, sens, …) ne seront jamais couverts par les IA et que la place de l'humain sera toujours centrale. Mais de nombreux autres pensent que les solutions neuro-morphiques qui imitent les architectures neurobiologiques présentes dans le système nerveux, ou que les solutions qui simulent la conscience voire qui reproduisent des supports biologiques de la conscience sont en cours de développement et verront le jour dans les années à venir. Elles pourraient alors avoir la perception de leurs propres actions et des réponses apportées.

Dans ce contexte, il est difficile d'appréhender ce que les IA font aujourd'hui, ce qu'elles ne font pas aujourd'hui mais qu'elles pourront faire demain et ce qu'elles ne pourront jamais faire.

Tous ces points expliquent en grande partie pourquoi un nombre très important de personnes ne font pas confiance aux IA. Il convient donc de fournir des réponses acceptables à ces préoccupations pour permettre une mise à disposition acceptable d'IA responsables, dignes d'humanité et de confiance.

c. Des impacts sensibles et critiques qui peuvent déranger

Les IA touchent de nombreux secteurs critiques tels que la santé, les décisions de justice, le recrutement, etc. ou des sujets sensibles, tels que la reconnaissance visuelle, qui peuvent présenter des risques pouvant avoir des conséquences majeures sur les principes fondamentaux sur lesquels reposent nos démocraties tels qu'égalité d'accès, de traitement, … Si de tels garanties ne sont pas assurées, cela conduira au rejet des IA.

Les IA peuvent aussi prendre des chemins logiques pour répondre efficacement aux objectifs visés qui soient inacceptables pour les humains. Ainsi, une IA qui déciderait d'éliminer tous les humains pour traiter le problème du CO2 remplirait parfaitement sa mission. Mais dans ce contexte, quel serait le garde-fou pour l'éviter ?

Par ailleurs, de nombreux systèmes sont chaotiques par nature. De petites variations peuvent provoquer de grands changements. Ainsi, en matière de prévisions météorologiques, il suffit de quelques données mal mesurées pour remettre en cause les prévisions. Les IA sont mal armées aujourd'hui pour traiter les « cygnes noirs » qui n'ont pas vraiment d'antécédents. Il en découle une certaine fragilité.

Plusieurs types de réponses sont évoqués pour traiter ces situations. Il s'agit par exemple de convenir de principes et dispositifs spécifiques pour ces contextes sensibles tels que les principes :

- d'« **ALARA »** de minimisation des risques à un niveau aussi bas que raisonnablement possible

- de **précaution** pour des risques potentiels non avérés
- de **vigilance** conduisant à une obligation de moyens renforcée de questionnement régulier et méthodique des IA et de surveillance de leurs évolutions autonomes pour contrebalancer une confiance qui pourrait être excessive. Un esprit du doute permanent doit perdurer.
- de **proportionnalité** qui implique que les moyens et actions mis en œuvre soient pertinents et n'excèdent pas ce qui est nécessaire au regard des finalités et que les inconvénients causés ne soient pas démesurés par rapport aux buts visés
- de **non malfaisance** (primum non nocere) c'est-à-dire que face à un problème particulier, il peut être préférable de ne pas faire quelque chose ou même de ne rien faire du tout que de risquer de faire plus de mal que de bien.
- du **droit d'alerte** pour signaler les cas qui présenteraient une menace ou un préjudice grave pour l'intérêt général

Mais leurs contours restent incertains et ne rassurent pas toujours. Il convient de trouver un consensus équilibré entre les exigences des parties prenantes de l'écosystème des IA à inclure dans la loi sans entraver l'innovation et la compétitivité et celles qui doivent être internalisées ou faire l'objet d'accords entre les parties.

d. De nombreux risques spécifiques mal maîtrisés nécessitant des réponses innovantes : discrimination, sécurité, fragilité, vie privée, autonomie, responsabilités, …

L'IA introduit de nouveaux risques ou le renforcement de risques actuels tels que :

- le **renforcement des biais humains** du fait de la présence de biais dans les données d'entraînement ou dans le développement des algorithmes et de l'efficacité des IA pour les reproduire
- l'**enfermement algorithmique** dès lors que les contenus et services proposés sur les médias sont personnalisés et basés sur ses goûts et opinions et donc cantonnés aux mêmes vues. Cela peut conduire à une polarisation voire radicalisation de ses positions au détriment d'un débat plus éclairé au travers d'un accès à une diversité d'opinions et d'arguments
- la **non reproductibilité des résultats** du fait de l'opacité et de la complexité notamment dans le cas de l'apprentissage profond pour lequel les IA s'auto-développent et les décisions se prennent hors de la compréhension et du contrôle de l'humain
- la **non compréhension des causes et des effets** du fait de l'absence de transparence, d'explicabilité ou de traçabilité
- la **plus grande fragilité** pour traiter des imprévus ou des attaques dans des environnements IA incertains, instables ou volatiles, du fait notamment de la non maîtrise des IA par les humains et de la difficulté de mettre au point des IA en mesure d'appréhender les problèmes de manière contextuelle en utilisant le sens commun.
- la **non protection de la vie privée** du fait de la non connaissance de l'utilisation qui est faite de ses données personnelles que ce soit dans les phases d'apprentissage, de tests ou opérationnelles ou du possible non consentement libre et éclairé de l'utilisation qui en est faite. Par ailleurs, les IA permettent d'identifier des informations, des traits de caractères et des comportements intimes des individus qu'ils ne souhaitent pas révéler.

- la **non protection des données autres que les données personnelles** (économiques, industrielles, …) qui ne sont pas couvertes par la RGPD et qui concernent notamment le secret des affaires, le droit de propriété intellectuelle, le secret journalistique, les délits d'initié, …

- la **grande dépendance** à des monopoles de fait, industriels ou étatiques, actuels ou futurs, pour l'accès aux données (GAFAM) et aux nouvelles technologies (informatique quantique, réseaux 5G ou 6G, …) compte tenu du fort niveau de moyens à déployer et du fait d'un avantage concurrentiel très important pour le premier acteur ayant du succès (cf. notion du « winner takes all ») ou pour l'accès à l'énergie du fait d'une très forte consommation énergétique des IA. Cette dépendance peut conduire à la perte de souveraineté et donc au risque du non-accès à ces ressources.

- la **non responsabilisation des différents acteurs** du fait de l'aspect diffus des IA et donc d'une dilution des responsabilités, de la non traçabilité des actions, de la non traçabilité des résultats aux sources, de la difficulté d'apporter tous les éléments probants, de la difficulté d'affecter la responsabilité à un individu en cas d'action contraire aux règles et au droit, de l'insécurité juridique, … Dans ce contexte, il y a une interrogation sur la place de l'humain dans la prise de décision et donc dans l'attribution de responsabilités. En effet, le processus de décision lui-même dépend grandement de la responsabilité qu'oblige cette décision. Mais, même dans le cas où les IA prédictifs ne seraient que des aides à la décision et que l'humain conserverait le contrôle final sur toutes les décisions, qu'en est-il réellement de sa responsabilité dans le cas où ces IA prédictifs tiendraient compte par exemple des dernières techniques médicales et qu'il ne fonderait pas ses soins sur la base de leurs conclusions, conclusions dont il n'aurait pas une parfaite maîtrise du fait d'une certaine opacité dans son fonctionnement. Pourrait-il être tenu pour responsable d'une perte de chance pour le patient de guérir ? Certains proposent de donner un statut juridique à certains IA qui ont une interaction physique (ex. robots, voiture autonome, …) pour traiter ces problèmes de responsabilité. Mais ces entités auraient-elles le libre arbitre ? Serait-ce judicieux ?

Il convient donc de bien comprendre la nature de tous ces risques spécifiques aux IA et d'identifier les dispositifs qui seraient susceptibles d'apporter des réponses adaptées.

6. Les conditions nécessaires pour des IA responsables, dignes d'humanité et de confiance

a. Base de compréhension commune et identification de typologies pertinentes

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

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

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

b. Des exigences a minima à définir

Dans ce contexte, il convient en premier lieu d'identifier les types d'exigences auxquels devront répondre ces différentes IA. Il sera ensuite nécessaire d'identifier le niveau d'exigence requis a minima pour donner confiance et les parties prenantes concernées. Ce niveau pourra dépendre par exemple de l'importance et de la sensibilité des impacts et du niveau du risque acceptable en fonction de chaque situation. Le type et le niveau d'exigences a minima pourront être fixés par une règlementation spécifique ou par contrat entre les parties intéressées. Enfin, il sera nécessaire de fournir une opinion circonstanciée par un tiers de confiance indépendant quant à la mise en œuvre effective de ces types et niveaux d'exigences.

i. Exigences relatives aux finalités d'usage

Les exigences peuvent concerner les aspects de finalité d'usage. Il peut s'agir de finalité individuelle ou finalité collective. Il s'agit de s'assurer d'une part que les IA sont **efficaces** c'est-à-dire qu'elles font bien ce qu'elles disent faire et qu'elles ne fassent que ce qu'elles sont censées faire. On parle souvent dans ce cas de loyauté vis-à-vis des utilisateurs. Il s'agit d'autre part de s'assurer qu'elles font bien ce qu'elles disent faire de manière **efficiente** et utilisant les ressources de manière responsable. Il conviendra ainsi de choisir les méthodes les moins coûteuses en temps, en espace, en énergie, …

ii. Exigences relatives aux moyens et conditions de mise en œuvre

Les exigences peuvent aussi concerner les moyens et les conditions de mise en œuvre de ces IA. Il va s'agir de s'assurer de :

- sa **fiabilité**, sa **sécurité** et sa **robustesse** dans le temps,
- sa **transparence**,
- la **protection de la vie privée**,
- son **traitement équitable**, **non discriminant** et **respectueux des valeurs fondamentales de dignité**,
- sa **conformité** (i.e. sa régularité) aux lois et réglementations.

iii. Nature et niveau des moyens à mettre en œuvre pour répondre aux exigences

L'atteinte de ces exigences dépendra de la nature et du niveau de moyens mis en œuvre qu'ils soient techniques, organisationnels, comportementaux, humains, informationnels ou autres. Il conviendra donc de s'assurer de leur faisabilité technique et de leur soutenabilité financière.

iv. Niveau d'atteinte des exigences attendues

Le niveau d'atteinte de ces exigences dépendra de la nature et du niveau de risque acceptables. La nature même des IA, des espèces de boîte noire non maîtrisées, conduit à une demande plus forte d'être rassuré quant à la maîtrise du résultat. On voit bien les types de risques associés à la conduite autonome. Ainsi, une erreur de conception, une erreur d'interprétation visuelle, une prise de contrôle par des hackers, etc… pourraient impacter des millions de voitures et mener au décès de milliers ou centaines de milliers voire plus de personnes. Le niveau de risque tolérable est bien moindre que les risques liés à la conduite individuelle.

Par ailleurs, plusieurs exigences peuvent être incompatibles entre elles. Le respect strict de certaines exigences peut conduire à un résultat contraire à l'objectif visé. Ainsi, quel niveau d'exigence est attendu en matière de conformité dans le cas du franchissement d'une ligne blanche par une voiture autonome pour éviter un cycliste. Faut-il introduire une notion de tolérance au non-respect des règles ? Il conviendra donc de trouver un équilibre entre les différents niveaux d'atteinte des exigences attendues qui pourraient être incompatibles entre elles.

v. Responsabilités pour l'atteinte ou la non-atteinte des exigences attendues

Qui dit exigences attendues, dit responsabilités en cas de problème ou de contestation. De nombreuses parties prenantes sont impliquées dans la chaîne de valeur. Il convient donc de pouvoir imputer les responsabilités aux uns et aux autres. Ce n'est pas toujours facile d'imputer les responsabilités dans le contexte des IA. Ainsi, une discrimination peut résulter d'une erreur de conception du système, d'erreurs ou de biais dans les données d'entraînement, d'erreurs ou de biais dans le développement des algorithmes, d'erreurs ou de biais d'interprétation, d'erreurs d'exploitation, etc…

La manière même d'arriver aux résultats n'est pas toujours explicite ou compréhensible. La reproductibilité des résultats dans un contexte d'apprentissage mouvant et la traçabilité des actions ne sont pas évidentes. L'utilisation de techniques d'apprentissage profond mobilisant de nombreuses variables, peut obscurcir la chaîne de calcul en place. Alors, comment attribuer les responsabilités ? Peut-on s'en extraire par la contractualisation de contrats d'assurances couvrant les risques spécifiques ?

vi. Exigences dans un contexte sensible à confirmer par un tiers de confiance

La nécessité d'être rassuré par un tiers indépendant externe peut dépendre de l'importance de l'impact des IA concernées. Ainsi, les impacts liés plus particulièrement à certains secteurs d'activité (santé, transports, système judiciaire, …) et à certaines situations (produisant de effets juridiques, recrutement, …) pourraient conduire à l'obligation légale de fournir une opinion par un tiers externe sur la mise en œuvre effective d'IA responsables, dignes d'humanité et de confiance. Cette intervention pourrait alors être considérée comme une mission d'intérêt général pour le compte de l'ensemble des parties prenantes. Il conviendra néanmoins de définir les moyens à mettre en œuvre pour permettre d'établir cette opinion externe et de s'assurer de leur faisabilité technique et de leur soutenabilité financière.

Dans d'autres cas sensibles pour les parties prenantes, on pourra contractualiser les types d'intervention et d'opinion permettant d'obtenir le niveau de confiance convenu entre elles.

c. Deux exigences avec de fortes spécificités IA posent problème : transparence et équité

i. Transparence

Problématique générale de transparence

Il existe un fort déséquilibre quant à l'accès aux informations concernant les IA par les différentes parties prenantes. Les IA apparaissent très souvent comme étant opaques notamment pour les personnes qui, in fine, sont concernées par les décisions qui sont prises à leur égard. Elles ne savent pas si elles sont directement concernées par ces décisions, si oui comment les décisions sont prises, si ces décisions sont pertinentes, fiables, équitables, non discriminantes et conformes à leurs droits, si elles peuvent s'y opposer et comment, qui est responsable en cas de problèmes, si leurs données personnelles sont protégées et ont été utilisées à mauvais escient sans leur consentement exprès, si des incidents ont eu lieu, etc.

Elles ont besoin de réponses claires à ces interrogations, sans lesquelles elles n'auront pas confiance dans ces IA et seront donc réticentes à les adopter. Mais leur fournir de manière transparente ces informations n'est pas aisé. Par exemple, quelles informations fournir, à qui et dans quelles circonstances ? Y-a-il des contraintes techniques ou légales pour ne pas les fournir ? Seront-elles compréhensibles et probantes ? Des outils existent-ils pour faciliter cette transparence ? Quelles sont leurs limites ? Qui doit rendre des comptes et à qui ? etc.

Nous avons tenté de préciser la nature de quelques-unes de ces interrogations ci-après.

Nature des informations susceptibles d'être communiquées et à qui

Les personnes directement impactées par la mise en œuvre d'une IA, ont-elles le droit de savoir si elles sont concernées, quelle est sa finalité, comment elles sont impactées, quels sont les types d'apprentissages, de modèles (modèles de boîtes blanches avec des règles connues, ou de boîtes noires avec des règles inconnues), d'algorithmes, de données d'entraînement et leurs critères de sélection et de représentativité, des variables utilisées et leur mode de captation, des paramètres influents, des contrôles et de tests effectués notamment pour contrer les biais, si des mesures permettant de mesurer des anomalies ou dérives éventuelles sont en place, à quel moment l'humain est impliqué et a son mot à dire dans la décision finale, si les exigences a minima sont respectées, si et comment elles peuvent contester les résultats, si des incidents ou évènements « néfastes » se sont produits, si des polices d'assurances ont été souscrites et pour quels types d'évènements et de dommages, … Ces informations doivent-elles être disponibles systématiquement ou pour certaines situations seulement ? Si oui, lesquelles (secteurs ou situations sensibles, …) ?

De même, les régulateurs peuvent-ils avoir accès par exemple aux différentes étapes qui ont abouti à un accident de voiture autonome ou aux standards qui ont été mis en œuvre en matière de sûreté (cf. concept de boîte noire) ?

Les concurrents peuvent-ils savoir si les bonnes pratiques annoncées par leurs concurrents ont été mises en place pour permettre une concurrence loyale, par la mise en place par exemple de missions d'intérêt général de certification du respect des exigences attendues et des pratiques mises en œuvre ?

Besoin de traçabilité et de piste d'audit

Par ailleurs, lorsque des décisions sont prises, existe-t-il une traçabilité suffisante des opérations effectuées et une piste d'audit qui permet de retracer toutes les actions qui ont été effectuées sur ces opérations y compris le traitement des incidents ? Pourra-t-on les imputer aux différents acteurs et ainsi affecter des responsabilités spécifiques ? Pourra-t-on auditer les conditions de mise en œuvre et les résultats et ainsi pouvoir fournir une opinion circonstanciée sur le respect des exigences attendues ?

Droit d'opposition et transparence

Lorsqu'une décision concernant une personne qui produit des effets juridiques est fondée exclusivement sur un traitement automatisé, elle a le droit de s'y opposer. A-t-elle cependant les moyens réels de pouvoir s'y opposer ? A-t-elle accès à la personne qui sera en mesure de modifier une décision la concernant ? Elle a aussi le droit le demander que des corrections soient apportées à ses données personnelles. Lorsqu'elle donne son consentement, ce consentement doit être libre et éclairé. Pour qu'il soit éclairé, elle doit disposer de certaines informations et il faut qu'elles soient suffisamment compréhensibles.

Limites des informations fournies

Le fait de rendre transparent le code des algorithmes ne répond pas nécessairement à toutes ces préoccupations. Il faudrait que les réponses aux questions ci-dessus soient compréhensibles, explicables, justifiables et loyales. Est-ce le cas ?

Ces droits ont été traités en partie dans la RGPD notamment pour les données personnelles. Mais dans le contexte d'algorithmes complexes « non linéaires » mouvants, ce type d'information n'est pas aisé à fournir notamment de manière compréhensible pour des non experts. Qui détermine si une explication est satisfaisante et sur la base de quels critères ?

De plus, il faut faire face à certaines contraintes telles que le secret des affaires qui peuvent limiter la capacité de fournir les informations utiles. Le fait de rendre transparents certains IA de contrôles (Fiscalité, Tracfin, …) ou certaines mesures de sécurité pourrait permettre de les contourner et donc poser problème. Pour préserver la confidentialité de ces informations, la communication de telles informations à des tiers de confiance est-elle envisageable ?

Une réponse : l'explicabilité « by design »

L'explicabilité « by design » semble être une nécessité pour permettre de répondre à la problématique d'une transparence responsable, digne d'humanité et de confiance. Il conviendra de communiquer les assertions des fournisseurs de services IA concernant la nature et le niveau des exigences qu'ils ont mis en œuvre et qu'ils sont prêts à rendre public : quel niveau de transparence, quels démarches, mécanismes, outils, référentiels, codes de bonne conduite, … en place permettant d'assurer la robustesse, la non-discrimination, la sécurité et la protection de la vie privée, etc…

Transparence par les résultats

Souvent, les humains ne sont pas intéressés à comprendre dans le détail le raisonnement appliqué mais plutôt à s'assurer que les résultats sont satisfaisants. Il s'agira alors d'identifier les mécanismes qui sont susceptibles de valider les résultats.

Une illustration pourrait être les chiens détecteurs. Ils apprennent de la même manière que les humains et les IA, par essais et erreurs et par renforcement au travers de récompenses. Les humains ne sont pas particulièrement intéressés par leur compréhension du raisonnement en œuvre dans le cerveau du chien mais dans le fait qu'il soit efficace dans sa recherche de drogues ou autres produits.

Plusieurs personnes considèrent que ce même raisonnement doit s'appliquer aux IA. Par exemple, pour traiter les discriminations, il est suggéré d'étudier la conformité et l'efficacité des résultats aux règlementations en vigueur en la matière et non d'expliquer le fonctionnement des IA qui permette d'y arriver.

Eléments probants

La confiance ne se déclare pas, elle s'obtient par des éléments probants. Les fournisseurs de services d'IA devront mentionner le niveau de confiance fourni pour chaque assertion et le démontrer au travers d'arguments et de preuves. Ils devront aussi indiquer si certaines assertions ont été certifiées par des tiers de confiance.

Limites de l'efficacité des outils disponibles

Des outils permettant d'identifier les paramètres influents, d'expliquer et d'interpréter les résultats, de rejouer les résultats, … sont-ils disponibles ? De nombreuses techniques sont proposées mais ne sont pas encore mures et ne traitent pas toutes les situations. Elles sont souvent basées sur des outils statistiques qui sont donc sujets à des incertitudes et erreurs. Il y a souvent un équilibre à trouver entre l'exactitude d'une explication et son interprétabilité. Dans le cas de l'apprentissage profond, il est plus difficile d'expliquer quelles variables et quel cheminement de traitement ont permis d'aboutir à une décision donnée que pour des algorithmes interprétables basés sur des régressions linéaires, des régressions logistiques, des arbres de décision, etc… ayant des relations linéaires plus faciles à décrire.

Dans quel contexte et avec quels outils, pourra-t-on donner une confiance suffisante dans le bon niveau de transparence ? Quels pourraient être les indicateurs simples à suivre ?

Dans le contexte des IA, souvent considérées comme des boîtes noires qui font peur, ce sujet de préoccupation qu'est la transparence est un élément essentiel dans l'obtention d'IA responsables, dignes d'humanité et de confiance et mérite d'être étudié plus en détail. Il est nécessaire de trouver le juste équilibre entre le droit et le besoin de savoir des uns et le droit et le besoin de ne pas divulguer ces informations des autres.

ii. Equité, non-discrimination, erreurs et biais

Problématique générale d'équité et non-discrimination, d'erreurs et biais

L'IA consiste, à partir de données, de faits observés, de situations, à appliquer des règles dans un contexte pour optimiser un résultat. Il peut s'agir d'aider à capter et percevoir les informations et situations, à améliorer les réflexions associées en les appréhendant et les interprétant mieux, en aidant à se projeter dans le futur avec de meilleures prévisions, prédictions et prescriptions, et en participant à leurs mises en œuvre.

Les deux principaux types de règles utilisées à cet effet sont :

- celles qui découlent de connaissances et de règles connues que l'on a modélisées. Les modèles partent du général pour aller au particulier. Il s'agit alors de raisonnements déductifs.
- celles qui découlent d'observations qui permettent de généraliser et de prévoir un résultat. L'apprentissage se fait par essais et erreurs ou par renforcement. Il s'agit par exemple de trouver des relations entre les faits, les données et les personnes. Les règles sont alors induites. L'apprentissage consiste à comprendre une situation, à trouver des similitudes malgré les

différences et des différences malgré les similitudes. Il s'agit le plus souvent de méthodes probabilistes.

Les humains alternent entre ces deux types de raisonnements.

Il existe d'autres types de règles : par analogie, statistiques, évolutionnistes, ….

Les résultats issus des IA peuvent néanmoins être erronés. Ces erreurs peuvent provenir de plusieurs sources : les données et leur interprétation, les règles et variables utilisées dans ces règles, l'interprétation des résultats.

Certaines erreurs sont assez classiques : données saisies de manière erronée, erreur de programmation, erreur dans le traitement, erreur de suivi, etc…

Certaines découlent de vulnérabilités de sécurité qui aussi sont assez classiques : introduction non valable d'informations fausses, cyber-attaques, …

Mais, d'autres erreurs sont plus spécifiques à l'environnement des IA.

Ainsi, certains de ces résultats erronés découlent d'erreurs de raisonnement. Ces erreurs de raisonnement sont multiples : erreurs d'interprétation ou d'appréciation, introduction de biais cognitifs, mauvais choix ou absence de choix de variables essentiels, … Ces erreurs peuvent avoir des impacts sur l'équité des résultats notamment en introduisant des discriminations qui peuvent être en opposition avec la loi et les réglementation ou inacceptables par la société.

D'autres erreurs résultent de l'application de règles probabilistes, qui est la base de l'apprentissage profond. Elles peuvent aussi découler de l'application de la notion d'aléa ou de hasard. On passe de modèles statistiques descriptifs à des modèles statistiques prescriptifs. Des erreurs peuvent aussi résulter de la manière dont ces règles probabilistes sont appliquées avec l'introduction de biais cognitifs. Les décisions qui découlent de cet apprentissage peuvent être biaisées de manière inconsciente.

Par ailleurs, plusieurs résultats considérés comme biaisés et discriminatoires découlent de biais volontaires qui ne sont donc pas des erreurs.

On pourrait s'attendre à ce que les IA permettent de proposer des solutions neutres et objectives non pourvues de biais cognitifs et donc sans discriminations. Mais, en réalité, de nombreux éléments peuvent contribuer à fausser les résultats : modèles biaisés, données incomplètes, non représentatives ou reflétant les biais présents dans la société, ... Aussi, les IA peuvent renforcer la prise en compte de biais cognitifs ou d'inéquités précédemment circonscrits en les généralisant souvent de manière opaque ou non intelligible.

Souvent, on confond différents types de biais et différents types d'erreurs.

Il convient donc d'étudier plus en détail la problématique concernant ces types d'erreurs, les biais et les discriminations associées plus spécifiques aux IA. Il conviendrait aussi a priori de mettre en place les dispositifs qui permettront d'éviter les erreurs et certains biais et ainsi fournir des résultats équitables ne favorisant pas les discriminations. Mais, cela n'est pas toujours très évident.

Une notion d'équité, de non-discrimination et de biais multiforme complexe à prendre en compte

La notion d'équité n'est en effet pas facile à décliner. Elle est multiforme.

S'agit-il d'une équité entre les individus ou entre les groupes, entre les opportunités ou les résultats ? Ces notions sont techniquement incompatibles. Il y a donc des choix à effectuer. Qui doit les effectuer et comment ? Ces choix sont le plus souvent éminemment culturels.

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

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

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

Erreurs de raisonnement et biais cognitifs : corrélations faussement trompeuses et causalités

Les IA vont permettre de capter, de filtrer, d'exclure, d'analyser, de structurer, de contextualiser et de classifier des informations ou des situations selon des règles prédéfinies ou auto-déterminées. On pourra alors identifier des corrélations notamment entre signaux faibles non perceptibles par les humains qui seraient susceptibles d'indiquer une causalité. Mais souvent, on relie des variables qui ne sont en fait pas liées entre elles. Il y a alors corrélation mais pas causalité. Ceci va mener à des schémas de pensée trompeurs faussement logiques et donc à des raisonnements erronés. Dans ces cas, qui sont très fréquents, on peut facilement faire des amalgames qui n'ont pas lieu d'être et qui vont se traduire par des discriminations.

Un exemple est la facturation des assurances automobiles : elle était beaucoup plus faible pour les femmes que pour les hommes jusqu'en 2013 lorsque cette pratique a été interdite étant considérée comme discriminatoire puisque basée sur le genre. Les analyses des données effectuées par les compagnies d'assurance montraient par ailleurs que cette corrélation liée au genre n'était en fait pas la causalité principale. Les causes identifiées étaient plutôt la distance parcourue, la nature des parcours (à proximité du domicile, sur grande route, autoroute, …), la période de conduite dans la journée (journée, soirées, nuits, weekend, …), etc…

Ainsi, pour illustrer ce point on peut le décliner sur une de ces variables. Les hommes et les femmes qui conduisaient moins avaient moins d'accidents que les hommes et femmes qui conduisaient beaucoup. Il y avait plus de femmes qui conduisaient moins d'où la corrélation favorable aux femmes. Mais la causalité n'était pas le fait d'être un homme ou une femme mais le fait de conduire beaucoup ou pas. La femme qui conduisait beaucoup était favorisé et l'homme qui conduisait moins était discriminé.

Il convient de noter que les pratiques évoluent dans le temps et que le poids des variables aussi. Ainsi, les corrélations du passé ne reflètent donc pas la situation d'aujourd'hui et leur utilisation pourrait donc fausser les résultats.

Dans ce contexte, comment trier les corrélations acceptables de celles qui ne le seraient pas ?

On doit aussi intégrer la notion d'alea ou de hasard. Il ne s'agit pas d'une absence de cause mais le fait que la cause ne détermine pas complètement le résultat. Dans ce cas, comment donner une réponse simple en intégrant et pondérant l'imprévu ou l'aléatoire dans la prise de décision.

Il conviendrait donc, une fois les corrélations identifiées, de procéder à une approche dite scientifique qui consiste à proposer des théories ou des modèles et à tester leur validité.

Erreurs de raisonnement et biais cognitifs : choix et calibration des variables

Le fait d'inclure dans un modèle certains critères pour un recrutement ou d'exclure certains candidats sur la base de prérequis peut conduire à exclure certaines catégories de personnes. La présence ou l'absence de certaines variables dans les modèles peut donc conduire à des situations que certains considéreront inéquitables. Le choix même des variables peut être conduire à des choix inéquitables. De plus, l'IA peut évoluer dans le temps, de nouvelles variables introduites, etc…

Souvent le choix de ne pas inclure certaines variables importantes découle de leur non-disponibilité ou de leur faible fiabilité et non de leur non-importance potentielle. Il convient de tenir compte de ces biais choisis et de faire part de leur impact potentiel sur la fiabilité des résultats qui en résulte.

Il sera nécessaire d'être vigilant et d'étudier et faire évoluer le cas échéant ces variables dans le temps pour éviter de nouvelles discriminations ou la perpétuation de celles qui existent.

Erreurs de raisonnement et biais cognitifs : données d'apprentissage non représentatives

Le fait de s'appuyer sur une représentativité statistique du passé, d'une région donnée ou d'une minorité active et militante avec des avis plus tranchés et la transformer en une condition systématique du présent, d'une autre région ou d'une population large plus silencieuse, introduit de ce fait un biais cognitif de statu quo historique, culturel ou de visibilité dans l'apprentissage.

Les données d'apprentissage ne sont pas erronées. Mais le choix de considérer que ces données comme représentatives alors qu'elles ne le sont pas a conduit à une erreur de raisonnement qui a été d'extrapoler les règles ainsi déduites à une autre population.

De même, si les règles sont en fait représentatives d'une situation existante et qu'elles sont extrapolées pour l'ensemble de la population mais que cette situation n'est pas acceptable du fait par exemple qu'elles traduisent des discriminations, alors cela conduira aussi à une erreur de raisonnement puisqu'elles conduiront au statu quo des discriminations ce qui ne serait pas l'objectif anticipé pour le futur.

Il conviendrait alors de définir les évolutions souhaitées et les moyens qui vont permettre de les intégrer dans les IA.

Il existe de nombreuses causes qui conduisent des populations de données à ne pas être représentatives. Il peut s'agir d'outils et techniques de captation, de mesure ou de perception des données, de données provenant de l'extérieur, de modèles, d'hypothèses, d'évènements rares qui peuvent fausser les analyses statistiques, … Il convient donc de les identifier, puis de les traiter en excluant, incluant ou corrigeant certaines données comme cela est fait pour les sondages.

De manière plus générale, l'introduction de nombreux autres types de biais cognitifs souvent de manière inconsciente vont conduire à des erreurs de raisonnement et possiblement à des discriminations.

Erreurs et biais techniques : maturité d'apprentissage

La mise en cause de l'objectivité des IA, sources potentielles de discriminations, trouve souvent son origine dans une méconnaissance du principe d'apprentissage et dans les ressorts de la théorie mathématique.

L'apprentissage est un processus permanent d'essais et d'erreurs qui permet d'identifier la meilleure réponse à un problème de manière de plus en plus fiable. Au fil du temps de l'apprentissage, la probabilité de faire une erreur devrait se réduire. Lorsqu'une IA va apprendre à reconnaître le visage d'une personne, elle va trier, classer et éliminer certains critères qui permettront in fine d'aboutir à un résultat. Lorsqu'elle comparera un homme barbu à une femme, elle fera moins d'erreur que lorsqu'elle le comparera à un homme non barbu. Ces erreurs sont souvent appelées biais techniques. Mais au fur et à mesure de l'apprentissage, elle devrait faire de moins en moins d'erreurs, ce qui ne veut pas dire qu'elle ne fera plus d'erreur du tout. Si tel est le cas, à quel moment pourra-t-on considérer l'IA suffisamment fiable pour la rendre opérationnelle ? Un taux d'erreur de 1 pour 100, 1 pour 1 000, 1 pour 10 000, … ?

Par exemple, l'utilisation effective d'une IA prédictive qui vise à identifier des comportements suspects qui vont conduire à procéder à un contrôle et à une fouille pour les personnes identifiées comme étant à haut risque, et qui aurait un taux de succès de 10%, sera-t-elle acceptable ou sera-t-elle considérée comme trop intrusive ? Si un tel taux n'était pas acceptable, un taux de succès de 90% serait-t-elle alors plus acceptable aux vues de la finalité de sécurité et des inconvénients liés à cette intrusion et du risque d'interpellations répétées et abusives dans le futur ?

Il convient d'être vigilent sur le fait de voir les IA se nourrir de leurs propres prédictions ce qui pourraient conduire à ce qu'elles deviennent auto-réalisatrices.

Qu'en serait-il si l'action proposée en lieu et place du contrôle et de la fouille était une surveillance régulière et étroite de la personne identifiée à haut risque ? Quels seraient alors les taux d'erreurs acceptables (95%, 99%, …).

Erreurs et biais techniques : faux positifs et faux négatifs

Par ailleurs, s'agit-il de faux positifs ou de faux négatifs dont les impacts ne sont pas les mêmes ?

Dans le cas, cité ci-dessus, les erreurs sont des faux positifs. Leurs conséquences sont une intrusion d'une plus ou moins forte intensité.

Dans ce même cas, on pourrait aussi évoquer les faux négatifs. Ils correspondraient au fait de ne pas identifier les réels suspects qui auraient dû être identifiés, contrôlés ou surveillés étroitement. Si le taux de faux négatifs était de 99%, serait-il alors acceptable d'utiliser cette IA et si oui quel que soit le taux de faux positif ? En effet, si l'IA n'est susceptible d'identifier qu'un suspect à haut risque sur 100 et que pour l'identifier il faille procéder à 10 contrôles, comment s'appliquerait le principe de proportionnalité ?

Il convient donc de trouver le bon équilibre entre l'optimisation d'un bien commun qu'est la sécurité et les nuisances individuelles acceptables.

Qui doit prendre cette décision et en fonction de quels critères ? Qui apprécie ce qui est tolérable et dans quel contexte ?

Erreurs liées et biais techniques : évènements exceptionnels ou cygnes noirs

Le fonctionnement des IA permet d'apporter de fortes contributions à la résolution de problèmes. Mais, on s'interroge parfois sur certains résultats.

Les IA apprennent de manière efficace à partir de données d'évènements du passé. Mais, il arrive néanmoins que des situations inconnues et imprévisibles se présentent qui n'ont pas d'exemple dans le passé. On les appelle des cygnes noirs. Les IA ne seront alors pas en situation de traiter convenablement ces cas et pourront donc apporter des réponses inadéquates ou erronées.

Par ailleurs, parfois, une bonne décision ne dépend pas de l'analyse de la masse d'information mais d'un seul renseignement décisif. Dans ce cas aussi, les IA ne sont pas toujours les mieux à même d'y répondre de manière satisfaisante.

S'agissant d'incertitudes et non de risques avérés, les réponses probabilistes ne s'appliquent pas. Les humains quant à eux sont souvent plus à même de mieux capter et appréhender cette incertitude et son contexte, et d'en tirer des conclusions.

Comment ces situations peuvent-elles être identifiées et traitées ?

Erreurs et biais cognitifs : restitutions

Il existe aussi des biais dans la présentation des résultats. Une IA d'identification de traitement de cancer de la peau après un diagnostic de reconnaissance visuelle pourra proposer un taux de succès (i.e. guérir et rester en vie) de 70% ou un taux d'échec (i.e. mourir) de 30%. Le choix de la manière de présenter le résultat (i.e. guérir ou mourir) affectera fortement le choix du patient qui sera plus positivement influencé par une communication sur les chances de guérir que sur les risques de mourir, ce qui semble tout à fait irrationnel puisqu'il s'agit de la même situation. Les restitutions ne sont pas erronées mais il existe un biais de présentation. Qui doit faire ce choix qui sera en tout état de cause biaisé ? Il s'agit d'un problème de choix éthique.

De même, on peut présenter une pratique médicale qui multiplie par 3 le risque d'avoir un problème de santé ou expliquer de manière plus relative que le risque passe d'1 occurrence sur 100 000 à 3 occurrences sur 100 000. Les deux présentations sont exactes. Mais chacune de ces présentations du niveau de risque n'aura pas le même impact sur le niveau d'incitation de cesser la pratique. Ici aussi, il y a un choix qui a des conséquences éthiques.

Plus généralement, il faut tenir compte de tous les types de biais que peuvent avoir les destinataires des résultats.

Erreurs de simplification et d'approximation

Les IA utilisent des modèles. Pour être efficace, ces modèles représentent une vue simplifiée d'un contexte en vue d'un objectif. Il est alors considéré que le fait de ne pas intégrer toutes les données et toutes les règles peut ne pas fausser de manière significative les conclusions. Les résultats sont ainsi erronés mais jugés corrects pour l'objectif visé.

De même, certaines données ne peuvent pas être mesurées de manière précise du fait par exemple de capteurs imprécis. On pourrait néanmoins intégrer des modèles théoriques d'extrapolation et les valider par les résultats. Les données peuvent donc ne pas être exactes et les résultats jugés satisfaisants.

De même, il peut valoir mieux des IA qui prévoient vite que des IA qui prévoient mieux sur la base de tous les paramètres pertinents mais trop tard.

Les IA sont faites de modèles et d'hypothèses de grande précision mais aussi d'approximations et donc d'erreurs dont il faut trouver le bon dosage. Il n'existe pas de modèles parfaits. Les modèles ne sont que des approximations de la réalité.

Réponses aux biais

Compte tenu de tous ces éléments, doit-on corriger les algorithmes pour les rendre « plus » équitables ? Si oui, dans quels cas de figures et comment ? Tous les biais cognitifs et techniques sont-ils illicites ? Doivent-ils être tous corrigés ? Comment peut-on identifier qu'il s'agit de biais cognitifs et techniques ou pas ou qu'ils soient tolérables ou pas, ces notions évoluant avec le temps ? Il convient de noter que ces corrections de biais cognitifs et techniques pourraient se faire en pénalisant la performance des IA. Dans quel cas serait-il acceptable de privilégier la performance aux biais techniques et cognitifs ?

Limite des outils

La question qui se posera sera de savoir s'il est possible et ou souhaitable de s'assurer de l'absence de biais ? Des outils existent qui permettent par exemple d'exclure formellement toutes les décisions basées sur certaines variables telles que celles correspondant aux 25 critères concernant 7 situations protégés des discriminations par la loi française (outil d'anti-classification). Mais, il existe plusieurs autres informations non formellement exclues qui pourraient aboutir au même résultat tout aussi biaisé. Ces outils semblent aujourd'hui encore limités pour traiter correctement ce sujet. D'autres méthodes ou outils de collecte de données, d'étiquetage, de classification ou de calibration permettant de réduire le risque de biais existent mais ont aussi leurs limites qu'il convient d'identifier.

Emettre une opinion sur ce type d'assertion : un des défis les plus importants des IA

Par ailleurs, aujourd'hui, un tiers de confiance qui souhaiterait fournir une opinion sur la non-discrimination d'une IA devra-t-il s'appuyer sur ses propres choix probablement biaisés ou sur des règles précises communément acceptées et des outils performants associés ?

Le traitement satisfaisant de cette problématique est critique dans l'obtention d'IA responsables dignes d'humanité et de confiance mais il est aussi probablement le plus difficile à traiter. Sa faisabilité reste complexe et incertaine. Il conviendrait donc de l'étudier plus en détail avant de pouvoir répondre de manière satisfaisante à la mise à disposition d'IA responsables, dignes d'humanité et de confiance.

d. Les moyens nécessaires pour des IA responsables, dignes d'humanité et de confiance : référentiels, outils et bonnes pratiques pour créer de la valeur

i. Démarche et dispositif de gouvernance à mettre en place pour évaluer les options, arbitrer et fixer les orientations en matière d'IA responsables, dignes d'humanité et de confiance

Nous avons vu ci-dessus que plusieurs parties prenantes sont concernées par les IA. Chacun a des intérêts propres qui peuvent être parfois divergents, mouvants, conflictuels, contradictoires, à court ou long terme, … Chacun souhaite créer de la valeur qui lui soit propre c'est-à-dire qui lui confère des avantages ou bénéfices en maintenant un niveau de risque acceptable et mobilisant un niveau de ressources acceptable de manière responsable. Chacun aura donc ses propres attentes en matière d'exigences

Pour répondre à cette situation, il s'agira dans un premier temps

- d'évaluer les besoins, les conditions et les options de création de valeur pour les différentes parties prenantes, en vue de déterminer des objectifs et des exigences à atteindre convenus entre elles
- de fixer une orientation après une étape d'arbitrage, de priorisation et de prise de décision
- de piloter la performance et la conformité au regard des orientations, des objectifs et des exigences convenus préalablement.

Un dispositif de gouvernance performant devra ainsi être mis en place pour réaliser cette première étape.

ii. Dispositif de management et opérationnel performant

Il sera ensuite nécessaire de planifier, développer ou acquérir, exploiter et surveiller les activités et les moyens en cohérence avec l'orientation, les objectifs et les exigences convenus préalablement. Un dispositif de management et opérationnel performant devra ainsi être mis en place pour réaliser cette deuxième étape en liaison et en cohérence avec le dispositif de gouvernance.

iii. Une démarche efficace qui se doit d'être intégrée, holistique, contextuelle et dynamique

L'obtention d'IA responsables, dignes d'humanité et de confiance passe nécessairement par la mise en œuvre d'une démarche efficace de gouvernance et de management intégrée, holistique, contextuelle et dynamique :

- intégrée : la mise en œuvre de tous les dispositifs qui mis bout à bout sont nécessaires à l'obtention d'IA responsables, dignes d'humanité et de confiance quelles que soient leurs positions dans la chaîne de valeur. Il peut s'agir de fournisseurs de solutions technologiques, de données, de modèles, de prestataires de services d'IA, d'utilisateurs de ces services, d'instances règlementaires et de contrôle ou de tiers de confiance. Il ne s'agit pas uniquement de dispositifs au sein des fonctions numériques mais au sein de toutes les fonctions. La bonne intégration de l'ensemble de ces dispositifs doit permettre d'éviter des trous dans la raquette.
- holistique : la prise en compte des différents types de dispositifs inter agissants entre eux de manière systémique qui seront nécessaires à la réalisation des objectifs. Il s'agira par exemple :
  - de principes, de politiques, de directives, de référentiels et de standards
  - de structures organisationnelles
  - de processus
  - de compétences, d'aptitudes et de ressources humaines
  - de valeurs et de comportements, tant individuels qu'organisationnels
  - d'outils, d'infrastructures, de services
  - d'informations

  Il s'agit de couvrir l'ensemble du cycle de vie de ces types de dispositifs.

  Il s'agit aussi d'identifier des indicateurs de performance associés à ces types de dispositifs qui permettront de s'assurer d'une part que les objectifs et les exigences attendues des différentes parties prenantes ont été réalisés et que d'autre part les dispositifs ont été mobilisés de manière efficiente et efficace.
- contextuelle : les dispositifs devront s'adapter au contexte spécifique de chaque IA i.e. besoins et contraintes spécifiques, risques particuliers, priorités spécifiques, …

- dynamique : l'environnement des IA est amené à se modifier de manière très régulière. Il est donc nécessaire d'être vigilant et de réévaluer la situation et les dispositifs à mettre en œuvre chaque fois qu'un facteur de conception change.

iv. Une démarche simple, ouverte et flexible et alignée aux différents référentiels et standards basée sur des modèles adaptés

Cette démarche doit être

- basée sur un modèle conceptuel simple mais pas simpliste qui identifie les éléments-clés et les interactions entre eux pour permettre d'optimiser la cohérence de la démarche et son automatisation et des modèles associés pour couvrir les différentes thématiques telles que la création de valeur, la typologie des exigences, la typologie des risques, la gestion des risques, la typologie des dispositifs, la typologie des assertions, etc…
- ouverte et flexible pour permettre l'ajout de nouveaux éléments ou de nouvelles préoccupations si nécessaire tout en maintenant cohérence et permanence
- alignée aux principaux référentiels, standards et règlementations pertinents pour favoriser l'adhésion du plus grand nombre

v. Des bonnes pratiques spécifiques pour répondre aux exigences et spécificités des IA

Des bonnes pratiques génériques et contextuelles pour chacun des types de dispositifs devront être identifiées pour chacune des exigences ou préoccupations à traiter. De nombreux référentiels fournissent des exemples de bonnes pratiques usuellement utilisées pour traiter les préoccupations et exigences relatives aux systèmes d'information et au numérique. Ces bonnes pratiques pour la plupart d'entre elles peuvent s'appliquer aux IA. On peut citer COBIT 2019, ITIL V4 et ISO.

Néanmoins, il convient de les compléter de bonnes pratiques répondant aux exigences et préoccupations spécifiques aux IA telles que la transparence notamment l'explicabilité et la traçabilité ou l'équité et la non-discrimination. En effet, des techniques et des outils spécifiques ont été développés ou sont en cours de développement. Ils peuvent répondre à certaines exigences mais ont aujourd'hui leurs limites. Il est donc nécessaire d'identifier ces limites mais aussi les bonnes pratiques qui en l'état actuel permettraient néanmoins de fournir un bon niveau de confiance quant à certaines assertions.

e. La nécessité d'être rassuré par un tiers indépendant

i. Définition de la proposition de valeur adaptée pour la certification des IA

Les différents acteurs du système financier et économique qu'ils soient actionnaires, employés, banquiers, fournisseurs, clients, citoyens ou l'Etat ont besoin d'avoir confiance dans les éléments financiers qui sous-tendent leurs décisions d'investisseurs, de prêteurs, de partenaires, d'achats, etc. … Sans confiance, il ne peut y avoir de croissance.

Pour favoriser cette confiance, les commissaires aux comptes, tiers indépendants, sont mandatés pour certifier la sincérité et la régularité des états financiers des entreprises privées et publiques, des associations et de l'Etat, mission d'intérêt général pour le compte de l'ensemble de l'écosystème. Ainsi, les grandes entreprises du CAC40 sont prêtes à payer plusieurs dizaines de millions d'€ d'honoraires chaque année pour rassurer l'écosystème financier sur leurs comptes.

Compte tenu des enjeux de l'écosystème des IA et de l'impératif d'IA responsables, dignes d'humanité et de confiance, il apparaît nécessaire, tel que c'est le cas pour l'écosystème financier, de mandater un tiers de confiance indépendant, tel qu'un commissaire aux algorithmes, qui aurait une mission d'intérêt général de certification des assertions relatives aux exigences de qualité attendues des systèmes d'IA.

Cette certification doit pouvoir créer de la valeur. Il s'agit donc de trouver le bon équilibre entre

- la nature des assertions attendues (transparence, équité et non-discrimination, fiabilité, sécurité, robustesse, protection de la vie privée, régularité, …), la nature de l'opinion (certification, label, attestation, …) et le niveau de confiance souhaitée pour ces assertions,
- le niveau de risque acceptable de fournir une certification erronée par exemple certifier qu'une IA est équitable alors qu'elle ne le serait pas (i.e. un faux positif) ou certifier qu'elle ne serait pas équitable alors qu'elle le serait (i.e. faux négatif)
- le niveau d'efforts acceptable pour fournir ce type et ce niveau de confiance notamment un niveau d'honoraires soutenable

A cet effet, il est nécessaire de définir précisément la démarche et les éléments qui permettront d'identifier, d'évaluer et de fixer les options de niveau de confiance qu'il convient de fournir dans un contexte donné.

ii. Convenir des éléments qui permettront de définir l'objectif d'un audit donné

L'objectif d'un audit est de créer de la valeur en fournissant, à un certain nombre de **parties prenantes** (à déterminer), un certain **niveau de confiance** (à déterminer : confiance raisonnable ?) quant à une **opinion** (à déterminer) relative à certaines **assertions** (qualités/caractéristiques à déterminer) de l'**objet audité** (à déterminer) à une **date** ou pour une **période donnée** (à déterminer), opinion qui leurs serait utile compte tenu des **avantages/bénéfices** qu'elles en retireraient (à déterminer) et dans la mesure où le **niveau de risque** (à déterminer) que cette opinion soit erronée et le **niveau d'effort** nécessaire (à déterminer : financier ou autre) pour y arriver soient acceptables.

iii. Convenir de la nature et du niveau des travaux d'audit

Une fois le ou les objectifs et le périmètre d'un audit déterminés, c'est-à-dire une fois les différents éléments énoncés ci-dessus établis, il sera alors possible d'identifier la nature et le niveau des travaux d'audit nécessaires pour répondre aux attentes de qualité des parties prenantes.

En fonction de l'objectif fixé, les travaux d'audit pourront consister soit

- à valider la « qualité » d'une assertion ou d'un résultat précis (un standard de référence à déterminer),
- à valider la « qualité » des moyens mis en œuvre (bonnes pratiques de référence à déterminer) pour permettre d'aboutir au niveau de qualité de résultat attendu,
- à une combinaison des deux (qualité d'un résultat et qualité des moyens).

L'opinion fournie pourrait prendre plusieurs formes (à déterminer : certification, attestation, audit, opinion circonstanciée, etc.).

La nature et le niveau des travaux d'audit pourront alors être identifiés. Ils dépendront

- d'une part, du niveau de risque inhérent aux éléments audités (par exemple, le risque que les moyens/bonnes pratiques mis en œuvre dans l'organisation ne permettent pas d'obtenir le niveau de qualité attendu de l'objet audité),
- d'autre part, le niveau de risque inhérent aux travaux d'audit (le risque que les travaux d'audit conduisent à une opinion contraire à la réalité : opinion positive alors qu'elle devrait être négative ou opinion négative alors qu'elle devrait être positive).

Il faut donc déterminer quel niveau de risque est acceptable par les parties prenantes sachant que pour réduire le niveau de risque (à un niveau acceptable par exemple), il faudrait allouer des moyens supplémentaires (en général de manière exponentielle par rapport au gain du niveau de risque) qui soient soutenables.

iv. Illustration des éléments à convenir pour un commissariat aux comptes

Pour illustrer ce point, on peut se référer à la mission d'audit dans le cadre d'un commissariat aux comptes. L'ensemble des éléments mentionnés ci-dessus ont été déterminés :

- parties prenantes concernées (actionnaires, Etat, fournisseurs, créanciers, etc…),
- niveau de confiance raisonnable (95% de confiance),
- opinion (certification sans réserve ou avec réserve, certification négative, refus de certifier),
- assertions : qualités/caractéristiques (sincérité, régularité),
- objet (états financiers, annexes),
- période (comptes annuels),
- avantages / bénéfices (confiance dans le marché financier),
- risques acceptables (95% de confiance et obligation de moyens),
- efforts (honoraires annuels),
- bonnes pratiques concernant l'objet audité (Principes comptables),
- bonnes pratiques concernant les moyens mis en œuvre par les organisations conduisant à un objet de qualité (contrôle interne : COSO, COBIT 5, etc…),
- bonnes pratiques concernant les travaux d'audit (standards d'audit)

Ainsi, les bonnes pratiques d'audit qui conduisent à une opinion de qualité satisfaisant a priori les attentes des parties prenantes à une niveau de risque acceptable peuvent conduire à des honoraires annuels de plus de 50 M€ pour les grands français ou européens. Il convient de noter que le niveau de risques acceptable pour un audit des états financiers n'est pas très élevé en comparaison à ce que l'on pourrait s'attendre d'un audit de certaines IA très sensibles.

Le même type de démarche doit être conduit pour l'audit des IA.

v. Démarche à mettre en œuvre concernant l'audit des IA

De même que pour le commissariat aux comptes, il est nécessaire de déterminer l'objectif d'un audit d'IA, la nature et le périmètre de l'opinion, pour en déduire la nature et le niveau des travaux d'audit nécessaires.

Dans ce contexte, il est nécessaire dans un premier temps

- De déterminer l'objet ou les objets à auditer : un algorithme, un groupe d'algorithmes, les données d'apprentissage, les données de tests, les données en entrée, une combinaison de ces éléments, le résultat d'un ou plusieurs de ces éléments compte tenu de l'environnement dans lequel ils sont exploités (technique, humain, organisationnel, …)

- D'identifier les parties prenantes qui seraient susceptibles d'être intéressées par une opinion sur cet objet : aussi bien le(s) prescripteur(s) (i.e. celui ou ceux qui commandite(nt) un audit et éventuellement paie(nt) pour l'audit) que ceux qui l'utilisent (le fournisseur, les utilisateurs, l'Etat, etc…). Cette opinion pourrait découler d'une obligation légale ou d'un engagement contractuel.

- D'identifier la nature de l'opinion ou des opinions qui pourraient les intéresser : attestation, certification, quels types de réserves, etc…

- D'identifier les assertions (qualités/caractéristiques) qui seront auditées. Il peut s'agir de qualités/caractéristiques générales (fiabilité, sincérité, régularité, etc…), de qualités/caractéristiques spécifiques inhérentes (exactitude, conformité, objectivité, …), de

qualités/caractéristiques spécifiques contextuelles (pertinence, actualité, transparence, compréhension, facilité d'utilisation, équité, non-discrimination, conformité à des règles/droits, …), de qualités/caractéristiques spécifiques d'accès/sécurité (disponibilité opportune, restriction d'accès, …).

- D'identifier les contributions de valeur que pourraient obtenir les différentes parties prenantes d'une opinion/label concernant les objets et qualités/caractéristiques audités (sélection des éléments mentionnés ci-dessus). Plusieurs contributions sont possibles, ces contributions de valeur résultant des avantages/bénéfices découlant de l'obtention d'une opinion/label dans la mesure où le niveau de risque que les avantages/bénéfices attendus ne soient pas réalisés soit acceptable (ex. risque d'une opinion d'audit erronée) et que le niveau d'efforts (financier et autres) pour l'obtenir soit acceptable (ex. coût de l'audit pour le niveau de risque souhaité soutenable). Il conviendra alors d'identifier les contributions acceptables.

En fonction des contributions acceptables identifiées, il conviendra ensuite

- D'identifier le référentiel de bonnes pratiques de l'objet audité et les assertions visées (ex. les exigences)
- D'identifier le référentiel de bonnes pratiques concernant les moyens mis en œuvre qui permettraient de répondre aux attentes de qualité auditées. En fonction de l'objet et des assertions visées (qualités / caractéristiques auditées), ces bonnes pratiques pourraient concerner
  - Les principes, directives, référentiels et standards utilisés
  - Les structures organisationnelles mises en œuvre
  - La culture, l'éthique et les comportements (incitations et aspects dissuasifs) mis en place
  - Les informations utilisées
  - Les outils et services à disposition (infrastructures, applications, …)
  - Les aptitudes, compétences, savoir-faire mobilisés
  - Les processus de gouvernance, de management et opérationnels mis en œuvre
- Ces bonnes pratiques doivent couvrir l'ensemble du cycle de vie (planification, conception, acquisition/développement, exploitation, suivi/évaluation, mise à jour/destruction)
- Il s'agira aussi d'identifier les indicateurs de performance de résultats et de moyens permettant de s'assurer que les attentes et objectifs des parties prenantes ont été atteints et que les moyens mis en œuvre sont efficaces et efficients
- D'identifier la démarche et les référentiels de bonnes pratiques concernant les travaux d'audit qui permettraient de fournir une opinion/label de qualité de l'objet et des qualités/caractéristiques auditées avec un niveau de risque acceptable quant à la fiabilité de cette opinion/label et avec un niveau d'efforts acceptable (coût soutenable).
- En fonction de la nature de l'opinion à fournir, ces bonnes pratiques d'audit pourraient concernées
  - Les principes, directives, référentiels d'audit utilisés
  - Les structures organisationnelles d'audit mises en œuvre
  - La culture, l'éthique et les comportements (incitations et aspects dissuasifs) d'audit mis en place
  - Les informations utilisées dans l'audit
  - Les outils et services à disposition de l'audit (infrastructures, applications, …)
  - Les aptitudes, compétences, savoir-faire d'audit mobilisés (certifications nécessaires : commissaire aux algorithmes ?)
  - Les processus de gouvernance, de management et opérationnels d'audit mis en œuvre

Ces bonnes pratiques doivent couvrir l'ensemble du cycle de vie de l'audit (planification, conception, acquisition/développement, exploitation, suivi/évaluation, mise à jour/destruction)

Il s'agira aussi d'identifier les indicateurs de performance de résultats et de moyens permettant de s'assurer que les attentes et objectifs d'audit des parties prenantes ont été atteintes et que les moyens d'audit mis en œuvre sont efficaces et efficients

Ces deux points (bonnes pratiques concernant l'objet audité et bonnes pratiques concernant l'audit) vont largement dépendre des choix concernant le périmètre de l'audit (pour quelles parties prenantes, quel objet audité, quelles assertions auditées (qualités/caractéristiques)) déterminés à l'étape précédente. Ils pourront néanmoins largement s'inspirer des bonnes pratiques incluses dans COBIT 5 concernant aussi bien les objets audités que l'audit.

7. Les recommandations et contributions potentielles du groupe de travail

Le groupe de travail de l'Académie a identifié plusieurs sujets d'intérêt mentionnés dans ce rapport qui mériteraient un approfondissement en vue de permettre la mise à disposition d'IA responsables, dignes d'humanité et de confiance et pour lesquels elle pense pouvoir contribuer efficacement. En effet, ce groupe pluridisciplinaire apporterait ses expertises et ses expériences dans le domaine de l'information, du système d'information et des technologies associées notamment dans les IA qui couvre les problématiques telles que gouvernance, référentiels, bonnes pratiques, exigences business, exigences règlementaires, fourniture d'assurance, certifications, …

L'objectif serait de compléter ce document par des ajouts au rapport qui découleraient de travaux complémentaires du groupe de travail au fil du temps.

Les recommandations du groupe concernant les thèmes qu'il considère comme critiques au traitement de la problématique d'IA responsables, dignes d'humanité et de confiance et sur lesquels il estime pouvoir faire pour certains d'entre eux des contributions et des recommandations utiles, sont :

a. Définir un socle de compréhension commun et en particulier concernant les exigences attendues des IA

Dans les environnements d'IA, il existe de nombreux concepts et définitions associées. Mais, les différentes parties prenantes n'ont pas la même appréciation de leur contenu, de leur périmètre et de leurs limites. Ils sont donc sources de confusion. Ceci ne favorise pas un traitement efficace des IA et peut ne pas donner confiance dans leurs résultats. Il nous paraît donc utile de les passer en revue et de définir un socle de compréhension commun à ces différentes parties prenantes.

L'un des points le plus important pour lequel une clarification est nécessaire est la notion d'exigences. Il serait donc très utile de définir la liste et le contenu des **exigences attendues** des IA.

En effet, des dizaines de termes et de concepts ont cours. Ils recouvrent souvent des notions similaires avec néanmoins beaucoup de nuances et de spécificités. Il convient donc de les définir précisément, de les différencier, de les regrouper et de les hiérarchiser.

Ainsi, les notions suivantes mériteraient d'être clarifiées :

- Transparence : traçabilité, piste d'audit, auditabilité, explicabilité, interprétabilité, jouabilité, compréhension, anonymat, secret des affaires, …
- Equité, non-discrimination : neutralité, impartialité, objectivité, loyauté, diversité, biais, contre-biais, croyances, erreurs, vérités, corrélations, causalités, …
- Fiabilité, sécurité, robustesse : sincérité, sûreté, conformité, régularité, pérennité, résilience, fragilité, précaution, vigilance, durabilité, dépendance, …
- Vie privée, dignité : humanité, vie intime, sûreté, consentement libre et éclairé à l'utilisation des données et à l'utilisation des résultats, protection des données versus protection des droits (à être informé, à être consulté, à l'oubli, à la déconnexion numérique, …)
- Efficacité, Efficience : performance, rentabilité, utilisation responsable, …
- Responsabilités : rendre compte, libre arbitre, autonomie, indépendance, consentement, imputabilité, éléments probants, intérêts et conflits d'intérêts, supervision, contrôle, …

Clarifier ces notions d'exigences est d'autant plus nécessaire qu'elles sont au cœur de l'appréciation du niveau de confiance proposé par les acteurs des IA. Les différents dispositifs et outils qui seront mis en œuvre le seront pour répondre à ces exigences. Elles sont aussi la base des réglementations

en matière d'IA. Elles sont aussi le fondement des opinions fournies dans le cadre des certifications par les tiers de confiance.

b. Elaborer une démarche efficace de développement et de mise à disposition d'IA responsables, dignes d'humanité et de confiance en portant une attention particulière sur le dispositif de gouvernance et d'affectation des responsabilités, sur celui de gestion de la prise de risque et sur les modèles spécifiques en soutien à cette démarche

Plusieurs types d'IA, de techniques d'apprentissage, d'outils, de dispositifs, … sont mis en œuvre. Plusieurs types de services et de fonctionnalités sont offerts. Plusieurs types d'acteurs sont concernés. Chacun amène sa part d'opportunités et de risques.

Pour prendre en compte la diversité de ces spécificités et permettre le respect des exigences attendues, il convient de définir une démarche

- **intégrée**, avec la mise en place de l'ensemble des dispositifs par l'ensemble des parties prenantes, qui mis bout à bout contribue aux IA responsables, dignes d'humanité et de confiance
- **holistique**, incluant l'ensemble des types de moyens, de dispositifs et de bonnes pratiques de gouvernance, de management et opérationnel qui doivent être mis en place, que ce soit des structures organisationnelles, des référentiels, des directives, des processus, des outils, des compétences, des comportements ou des ressources informationnelles
- **contextuelle**, adaptée à chaque type de situation, en fonction du type et du niveau de risques et des exigences spécifiques attendues,
- **dynamique**, évoluant en fonction de modifications fréquentes de l'ensemble de ces éléments,

Cette démarche doit être **simple**, à base de modèles, **flexible et souple**, permettant l'intégration de nouveaux éléments, et **alignée** aux principaux référentiels, favorisant une large adhésion.

Cette démarche devra prendre en compte trois sujets qui méritent une attention particulière dans l'environnement des IA :

- Elaborer une démarche de mise en place d'un **dispositif de gouvernance adapté** avec l'affectation des responsabilités associées.

  Plusieurs décisions concernant plusieurs parties prenantes vont nécessiter des arbitrages : quels engagements concernant les exigences, quelles assertions les concernant, quel niveau de transparence, quelles certifications, … Par ailleurs, qui dit exigences attendues, dit responsabilités en cas de problème ou de contestation. De nombreux acteurs sont impliqués dans la réalisation de la chaîne de valeur. Il convient donc de pouvoir imputer les responsabilités, ce qui n'est pas aisé dans un contexte d'apprentissage profond, par exemple.

- Elaborer une démarche spécifique de **gestion de la prise de risque**.

  De nouveaux risques apparaissent : la perte de contrôle et du libre arbitre, l'enfermement algorithmique, l'opacité et la non-reproductibilité des résultats, le renforcement de certains biais, l'absence de traçabilité et d'imputabilité, la plus grande dépendance et fragilité, et la non-responsabilisation des différentes parties prenantes. De nouveaux principes sont proposés : principes de minimisation des risques, de précaution, de vigilance, de proportionnalité, de non-malfaisance, d'alerte, … Il convient d'intégrer ces spécificités à la démarche générale de gestion de la prise de risque.

- Elaborer les **modèles spécifiques** en soutien à cette démarche

Une démarche adaptée se doit d'être simple, souple et flexible, facile à mettre en œuvre, facile à maintenir et facile à comprendre. La diversité et la complexité du paysage des IA conduisent à la nécessité de pouvoir disposer de modèles et classifications appropriés qui permettent notamment d'aborder efficacement la nature des apports, des risques, des solutions, ... Catégoriser est un élément essentiel à une meilleure compréhension et appréhension des différents phénomènes complexes du monde qui nous entoure.

Les éléments suivants pourraient faire l'objet de classifications :

- Types d'apprentissage : continu, initial, …, données individuelles, de groupe, …
- Types de modèle : supervisé, non supervisé, par renforcement, …
- Types de domaines : symbolisme, connexionnisme, comportementalisme, …
- Types de raisonnements : inductif, déductif, probabiliste, par analogie, …
- Types de modes de fonctionnement : fait observable ou pas, action déterministe ou pas, environnement statique ou dynamique, variable discret ou continu, …
- Types d'algorithme : classification, régression, clustering, …
- Types d'intervention humaine
- Mode d'acquisition des variables : saisie, capteur, …
- Types de traçabilité : traitement, décisions, processus, …
- Types de données : apprentissage, tests, de production, de renforcement, …
- Types d'outils : reconnaissance langage, reconnaissance visuelle, outils de segmentation, systèmes experts, …
- Types de services : chat bots, robots, marketing prédictif, …
  types d'algorithmes, d'apprentissage, de raisonnements, de données, d'outils, de services, ….

Il convient de noter que toute classification se fait sur des critères explicites ou implicites qui vont discriminer d'une manière ou d'une autre. Il s'ensuit la nécessité d'être vigilent. L'environnement externe évolue et les classifications doivent donc évoluer aussi.

c. Identifier la manière de prendre en compte les exigences spécifiques aux IA et en particulier la transparence et, la non-discrimination et les biais. Concernant la transparence, requérir l'explicabilité « by design » et exiger la traçabilité et la piste d'audit (audit trail) des traitements. Concernant la non-discrimination et les biais, limiter les discriminations du fait de biais par application de bonnes pratiques de développement et d'analyse des données, rendre les « biais volontaires » transparents et mesurés et limiter les « biais techniques » en publiant le degré de maturité de l'apprentissage.

Nous avons identifié plusieurs types d'exigences qui permettraient de disposer d'IA responsables dignes d'humanité et de confiance. Pour satisfaire ces diverses exigences, il convient de mettre en place des dispositifs spécifiques (techniques, outils, …). Il en existe déjà certains mais compte tenu de l'émergence de cet impératif d'IA responsables, dignes d'humanité et de confiance, ces dispositifs sont aujourd'hui soit incomplets soit en développement.

Deux de ces exigences qui ont de fortes spécificités IA posent problème, la transparence et la non-discrimination, et méritent une attention particulière.

La transparence

Il existe un fort déséquilibre quant à l'accès aux informations concernant les IA par les différentes parties prenantes. Les IA apparaissent très souvent comme étant opaques notamment pour les personnes qui, in fine, sont concernées par les décisions qui sont prises à leur égard. Elles ne savent

pas si elles sont directement concernées par ces décisions, si oui comment les décisions sont prises, si ces décisions sont pertinentes, fiables, équitables, non discriminantes et conformes à leurs droits, si elles peuvent s'y opposer et comment, qui est responsable en cas de problèmes, si leurs données personnelles sont protégées et ont été utilisées à mauvais escient sans leur consentement exprès, si des incidents ont eu lieu, etc.

Elles ont besoin de réponses claires à ces interrogations, sans lesquelles elles n'auront pas confiance dans ces IA et seront donc réticentes à les adopter. Mais leur fournir de manière transparente ces informations n'est pas aisé. Par exemple, quelles informations fournir, à qui et dans quelles circonstances ? Y-a-il des contraintes techniques ou légales pour ne pas les fournir ? Seront-elles compréhensibles et probantes ? Des outils existent-ils pour faciliter cette transparence ? Quelles sont leurs limites ? Qui doit rendre des comptes et à qui ? …

Dans le contexte des IA, souvent considérées comme des boîtes noires, qui font peur, ce sujet de préoccupation qu'est la transparence est un élément essentiel dans l'obtention d'IA responsables, dignes d'humanité et de confiance et mérite d'être étudié plus en détail. Il est nécessaire de trouver le juste équilibre entre le droit et le besoin de savoir des uns et le droit et le besoin de ne pas divulguer ces informations des autres.

Pour répondre à cette situation, deux points particuliers doivent être mis en place :

⇒ Requérir **l'explicabilité « by design ».**

Il s'agit d'identifier au préalable les informations qu'il convient de communiquer et à qui, les assertions des parties prenantes qu'elles souhaitent mettre en avant, le niveau d'exigence attendue, les dispositifs qui permettront de les satisfaire et les éléments probants correspondants. Ce n'est qu'en les ayant définis au préalable qu'ils pourront être intégrés efficacement lors du développement des IA.

⇒ Exiger la **traçabilité** et la **piste d'audit** (audit trail) des traitements.

Lorsque des décisions sont prises, une traçabilité suffisante des opérations effectuées et une piste d'audit qui permet de retracer toutes les actions qui ont été effectuées sur ces opérations y compris le traitement des incidents sont nécessaires. Il faut pouvoir les imputer aux différents acteurs et ainsi affecter des responsabilités spécifiques. Il faut aussi pouvoir auditer les conditions de mise en œuvre et les résultats et ainsi pouvoir fournir une opinion circonstanciée sur le respect des exigences attendues.

La non-discrimination

Les erreurs issues des IA peuvent provenir de plusieurs sources : les données et leur interprétation, les règles et variables utilisées dans ces règles, l'interprétation des résultats.

Certaines erreurs sont assez classiques : données saisies de manière erronée, erreur de programmation, erreur dans le traitement, erreur de suivi, etc…

Certaines découlent de vulnérabilités de sécurité qui aussi sont assez classiques : introduction non valable d'informations fausses, cyber-attaques, …

Mais, d'autres erreurs sont plus spécifiques à l'environnement des IA. Il convient donc d'une part de les identifier et les qualifier, et d'autre part trouver les dispositifs qui permettent de les appréhender, les éviter ou les corriger.

Certains résultats erronés découlent d'erreurs de raisonnement. Ces erreurs de raisonnement sont multiples : erreurs d'interprétation ou d'appréciation, introduction de biais cognitifs, mauvais choix ou absence de choix de variables essentiels, … Ces erreurs peuvent avoir des impacts sur l'équité des résultats notamment en introduisant des discriminations qui peuvent être en opposition avec la loi et les réglementation ou inacceptables par la société.

D'autres erreurs résultent de l'application de règles probabilistes, qui est la base de l'apprentissage profond. Elles peuvent aussi découler de l'application de la notion d'aléa ou de hasard. Des erreurs peuvent aussi résulter de la manière dont ces règles probabilistes sont appliquées avec l'introduction de biais cognitifs. Les décisions qui découlent de cet apprentissage peuvent être biaisées de manière inconsciente.

Plusieurs résultats considérés comme biaisés et discriminatoires découlent de biais volontaires qui ne sont donc pas des erreurs.

On pourrait s'attendre à ce que les IA permettent de proposer des solutions neutres et objectives non pourvues de biais cognitifs et donc sans discriminations. Mais, en réalité, de nombreux éléments peuvent contribuer à fausser les résultats : modèles biaisés, données incomplètes, non représentatives ou reflétant les biais présents dans la société, ... Aussi, les IA peuvent renforcer la prise en compte de biais cognitifs ou d'inéquités précédemment circonscrits en les généralisant souvent de manière opaque ou non intelligible.

Il convient donc d'étudier plus en détail la problématique concernant ces types d'erreurs, les biais et les discriminations associées plus spécifiques aux IA. Il conviendrait aussi a priori de mettre en place les dispositifs qui permettront d'éviter les erreurs et certains biais et ainsi fournir des résultats équitables ne favorisant pas les discriminations.

A cet effet, quelques dispositifs particuliers doivent être mis en œuvre

- ⇒ Limiter les discriminations du fait de biais par application **de bonnes pratiques de développement et d'analyse des données**
- ⇒ Rendre les **« biais volontaires » transparents** et mesurés
- ⇒ Limiter les « biais techniques » en publiant le **degré de maturité** de l'apprentissage.

d. Présenter les dispositifs de tests pour des IA responsables, dignes d'humanité et de confiance ?

Un des dispositifs privilégiés pour valider le bon fonctionnement des IA est d'effectuer des tests d'algorithmes. Mais tester par exemple des algorithmes d'apprentissage profond peut amener son lot de difficultés. Ces algorithmes ne sont pas linéaires, ne sont pas stables, sont souvent complexes et opaques, et peuvent utiliser un nombre très significatif de variables. Les tests traditionnels des programmes informatiques ne sont pas toujours adaptés à ce contexte. Il est donc nécessaire d'utiliser des types de tests spécifiques. Quels sont-ils ? Quels niveaux de fiabilité procurent-t-ils ? Quels sont leurs limites ? Y-a-t-il suffisamment d'éléments probants ? Quels sont les techniques ou outils en cours de développement qui seraient susceptibles de répondre pleinement aux attentes en la matière ? Il nous paraît utile d'essayer de répondre à ces questions.

Par ailleurs, les tests sont aussi un dispositif largement utilisé par les auditeurs. L'utilisation de tels tests dans un contexte d'IA est-elle encore adaptée ? Il serait donc utile d'identifier comment les auditeurs pourraient continuer d'utiliser les tests pour satisfaire leurs diligences. Y-a-t-il des outils spécifiques qui faciliteraient leurs travaux ?

e. Obtenir une « certification » par un tiers de confiance pour les IA « sensibles », proposer des exemples d'assertions qui pourraient être certifiées et élaborer des guides d'audit (questionnaires, check-lists, matrices, …)

Il faut dans un premier temps identifier les situations qui conduiront à la nécessité d'obtenir une certification. Cela pourra dépendre de l'importance de l'impact des IA concernées. Il pourra s'agir de secteurs d'activité critiques (santé, transports, défense, système judiciaire, …) et de situations critiques (produisant des effets juridiques, recrutement, reconnaissance faciale, …). Cette intervention pourrait alors être considérée comme une mission d'intérêt général pour le compte de l'ensemble des parties prenantes.

Dans d'autres cas sensibles pour les parties prenantes, on pourra contractualiser les types d'intervention et d'opinion permettant d'obtenir le niveau de confiance convenu entre elles.

Pour cela, il faut déterminer les objets à auditer, identifier les parties prenantes intéressées, identifier le type d'opinion attendue et les assertions sur lesquelles l'opinion portera. En résumé, quel référentiel de bonnes pratiques et quel référentiel d'audit. Ces choix permettront de définir la nature et le niveau des moyens à mettre en œuvre qui doivent être techniquement faisables, financièrement soutenables et cohérents avec la création de valeur attendue.

Pour faciliter le travail des auditeurs, il serait utile de

⇒ Proposer des **exemples d'assertions** qui pourraient être certifiées.

Plusieurs éléments doivent être définis pour déterminer le type et l'objectif d'un audit spécifique et notamment les assertions à certifier. Plusieurs types d'assertions pourront alors être validés pour un certain niveau de confiance et pour un certain niveau d'efforts. Il convient de s'assurer que les options qui seront choisies in fine répondront à des besoins réels à des tarifs soutenables.

Afin de faciliter cet exercice, nous pensons utile d'illustrer les types d'opinion possibles qui résulterait d'un audit. Chacun pourra alors voir le type d'apport de confiance qu'engendrerait ces types d'opinion et si cela pourrait les satisfaire.

Deux types d'audit pourraient être illustrés : un contexte assez simple, type Parcoursup, et un contexte plus complexe, type voiture autonome.

⇒ Elaborer des **guides d'audit** (questionnaires, check-lists, matrices, …)

Ils pourront aider les auditeurs à mener leurs missions dans les meilleures conditions compte tenu des difficultés auxquelles ils seront confrontés liées aux spécificités des IA.

8. De nombreuses contributions complémentaires par d'autres organisations sont en cours et pourront contribuer à traiter ces problématiques

De nombreuses organisations se sont emparées de ces problématiques, identifiées comme essentielles au développement d'IA responsables, dignes d'humanité et de confiance. Elles ont décidé d'apporter leurs contributions pour faire progresser leur traitement. Certains de leurs travaux ont été publiés très récemment et plusieurs autres sont prévus d'ici 2021. Nous espérons donc qu'ils viendront compléter la panoplie de dispositifs permettant l'offre d'IA responsables, dignes d'humanité et de confiance.

On pourrait ainsi citer plusieurs initiatives telles que :

- Les travaux techniques du **Centre Commun de Recherche de l'Union Européenne** publiés en 2020 dont l'objet était de créer les ponts entre les directives et les solutions techniques actuelles ou en cours de développement, relatifs à la robustesse et à l'explicabilité de l'IA. L'objectif était aussi de mettre en avant des outils de certification de systèmes d'IA

- Le lancement en juin 2020 par une quinzaine de pays dont la France du **Partenariat Mondial sur l'IA** qui doit s'employer à jeter les ponts entre la théorie et la pratique axés notamment sur l'utilisation responsable des IA et la gouvernance des données

- La mission d'un **comité d'éthique sur le numérique du Comité Consultatif National d'Ethique**, qui doit aborder, d'ici début 2021, dans le domaine des sciences de la vie et de la santé, les thèmes comme la transparence sur le traitement des données collectées, les responsabilités partagées entre constructeur, assureur et utilisateur, la transparence et l'explicabilité du fonctionnement de ces algorithmes. Il doit également mettre en place les moyens nécessaires à l'information et la prise de décision individuelle et collective.

- Le mandat de l'**UNESCO** pour élaborer un instrument normatif mondial d'ici 2021 notamment sur les problèmes de diversité culturelle ou du genre pour lutter contre les biais algorithmiques discriminatoires

- La création d'un **laboratoire commun EDF, Total et Thales** en 2020 effectuant des travaux de recherche relatifs au développement d'IA dites de confiance, suffisamment sûres pour servir dans des systèmes critiques tels que centrales nucléaires, voitures autonomes ou tours de contrôle

- Les travaux de l'**Institut Montaigne** en 2020 sur le contrôle des biais dans les algorithmes dont l'objectif était d'y apporter des solutions concrètes

- La mise en lumière des risques de discriminations de l'IA du **Défenseur des Droits en partenariat avec la CNIL** en mai 2020. Leurs recommandations incluent le développement des études de mesure et les méthodologies de prévention des biais, le renforcement des obligations en matière d'information, de transparence et d'explicabilité des algorithmes et la réalisation des études d'impact pour anticiper les effets discriminatoires des algorithmes

- La production d'un rapport en avril 2020 **« Toward trustworthy AI develpment : Mechanisms for supporting verifiable claims »** par un groupe international pluridisciplinaire regroupant plusieurs dizaines de spécialistes de l'industrie de l'IA, du monde universitaire et de la société civile. Il couvre les sujets tels que assertions vérifiables, audit par des tiers de confiance, pistes d'audit, explicabilité et tests. Il identifie de nombreuses techniques disponibles ou en cours d'études

qu’elles soient institutionnelles, logiciels ou matériels. Il met en avant les avantages, inconvénients, risques et coûts, les limites, les contextes où cela marche ou pas (informatique traditionnelle, apprentissage machine, apprentissage profond, …), les thèmes et sous-thèmes traités. Il mentionne aussi les différents types de tests, leur efficacité et leurs limites.