5 recommandations pour des IA dignes de confiance

Le contexte : un impératif de confiance à créer

L'Intelligence Artificielle (IA) devrait pleinement contribuer au développement économique dynamique et durable, à la résolution de nombreux défis planétaires, sociaux, sociétaux, climatiques et démocratiques, et au progrès et au bien-être des individus.

Mais elle suscite de nombreuses inquiétudes et présente de nouveaux risques significatifs qui pourraient entacher la confiance et casser la dynamique de développement.

Il convient donc d'identifier les principes, démarches, outils et bonnes pratiques spécifiques aux IA qui faciliteront le développement et la mise à disposition d'IA susceptibles de tirer parti de ces opportunités tout en levant les freins et en maîtrisant les risques associés.

Il faut aussi identifier les conditions de leur mise en œuvre effective. Les spécificités des IA font que cette mise en œuvre apparaît difficile aujourd'hui. De nombreux principes et dispositifs nécessaires ne sont pas disponibles, sont incomplets ou difficiles d'application.

Les différentes parties prenantes de l'écosystème des IA ont néanmoins besoin d'avoir confiance dans ces outils et dispositifs 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. Cette confiance est un facteur clé à leur diffusion et adoption.

Compte tenu de l'importance de ces enjeux, plusieurs Etats et organisations nationales et internationales (OCDE, UE, …) se sont saisis du sujet et ont identifié plusieurs principes et exigences à mettre en œuvre en sus de ceux qui existent déjà pour donner confiance.

C'est dans ce contexte qu'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.

L'objectif premier du groupe a donc été de comprendre le contexte qui incite à cet impératif de confiance en précisant la nature de ce besoin, les acteurs intéressés, 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 sur les affirmations concernant les IA dignes de confiance.

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

5 recommandations pour donner confiance

Vous trouverez ci-après les 5 grandes recommandations du groupe de travail pour des IA dignes de confiance :

R1 - Clarifier le contenu des exigences attendues des IA et identifier leurs spécificités propres qui devront être pris en compte pour donner confiance notamment en ce qui concerne la transparence et le traitement des biais et de la non-discrimination

⇒ R1-1 Définir la liste et le contenu des **exigences attendues** des IA
⇒ R1-2 Identifier la manière de prendre en compte l'exigence de **transparence** notamment en
- R1-2-1 requérant **l'explicabilité « by design »**
- R1-2-2 exigeant la **traçabilité** et la **piste d'audit** des traitements

⇒ R1-3 Identifier la manière de prendre en compte l'exigence d'équité et de non-discrimination notamment en
- R1-3-1 appliquant de **bonnes pratiques de développement et d'analyse des données**
- R1-3-2 rendant les **« biais volontaires » transparents**
- R1-3-3 limitant les risques de « biais techniques » en publiant le **degré de maturité de l'apprentissage.**

R2 - Élaborer une démarche de la **gestion de la prise de risque spécifique aux IA**

⇒ R2-1 Elaborer des **modèles spécifiques de classification des risques**

R3 - Elaborer une démarche de gouvernance adaptée avec l'affectation des responsabilités associées

R4 - Identifier les bonnes pratiques spécifiques qui répondent aux exigences et risques spécifiques liés aux IA

⇒ R4-1 Définir les **tests e**t les outils associés adaptés par catégorie d'IA pour les phases d'apprentissage et opérationnelle, et les outils permettant de les réaliser. Définir aussi les tests d'algorithmes à effectuer dans le cadre des audits.

R5 - Obtenir une « certification » par un tiers de confiance pour les IA « sensibles »

⇒ R5-1 Proposer des **exemples d'affirmations** à certifier
⇒ R5-2 Elaborer des **guides d'audit** (questionnaires, check-lists, matrices, …)

R1 – Clarifier le contenu des exigences attendues des IA et identifier leurs spécificités propres qui devront être pris en compte pour donner confiance notamment en ce qui concerne la transparence et le traitement des biais et de la non-discrimination

Il existe de nombreux concepts et définitions associés aux IA. 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 les résultats. Il apparaît utile de les passer en revue et de définir un socle commun de compréhension.

Il serait particulièrement utile de :

⇒ R1-1 Définir la liste et le contenu des **exigences attendues** des IA

5 recommandations pour donner confiance

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, les différencier, les regrouper et les hiérarchiser. Ainsi, les notions d'efficacité, d'efficience, de performance, de fiabilité, de loyauté, de sincérité, de robustesse, de résilience, de durabilité, de sécurité, de transparence, d'explicabilité, de respect de la vie privée, d'équité, de non-discrimination, de neutralité, d'humanité, de conformité, de régularité, … ne sont pas clairement assimilés aujourd'hui.

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 sont mis en œuvre le sont pour répondre à ces exigences. Elles sont aussi la base des réglementations en matière d'IA et le fondement des opinions fournies dans le cadre des certifications par les tiers de confiance.

Il convient de noter que plusieurs dispositifs qui permettraient de répondre à ces exigences sont aujourd'hui soit incomplets soit en développement. En particulier, deux exigences qui ont de fortes spécificités IA posent problème, la transparence et, les biais et la non-discrimination, et méritent une attention particulière.

⇒ R1-2 Identifier la manière de prendre en compte l'exigence 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 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, si des incidents ont eu lieu, etc.

Elles ont besoin de réponses claires, 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 ? …

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 des autres.

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

⇒ R1-2-1 Requérir **l'explicabilité « by design »**

Il s'agit d'identifier au préalable les informations qu'il convient de communiquer et à qui, les affirmations 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 de ces IA.

⇒ R1-2-2 Exiger la **traçabilité** et la **piste d'audit** (audit trail)

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 et par qui, 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.

⇒ R1-3 Identifier la manière de prendre en compte l'exigence d'équité et de non-discrimination

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

Certaines erreurs sont ou découlent de vulnérabilités assez classiques : données saisies de manière erronée, erreurs de programmation, de traitement ou de suivi, introduction non valable d'informations fausses, cyber-attaques, …

D'autres sont plus spécifiques à l'environnement des IA. Il convient de les identifier et les qualifier, puis de trouver les dispositifs qui permettent de les appréhender, les éviter ou les corriger.

Ainsi, dans ces environnements, plusieurs erreurs de résultats découlent plutôt d'erreurs de raisonnement d'origine multiples : erreurs d'interprétation ou d'appréciation, introduction de biais cognitifs, mauvais choix des variables dans les modèles, … 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.

Plusieurs autres erreurs sont des biais techniques qui résultent de l'application de règles probabilistes, qui sont la base de l'apprentissage profond, ou de l'application de la notion d'aléa ou de hasard. Ainsi, on sait d'avance qu'il y aura x% d'erreurs.

Par ailleurs, plusieurs résultats considérés comme biaisés et discriminatoires découlent de biais volontaires (quotas, …) 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 contribuent 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'iniquités 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 ces erreurs et certains biais et ainsi fournir des résultats non discriminatoires.

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

⇒ R1-3-1 Limiter les discriminations du fait de biais par application **de bonnes pratiques de développement et d'analyse des données**

⇒ R1-3-2 Rendre les **« biais volontaires » transparents**

⇒ R1-3-3 Limiter les risques de « biais techniques » en publiant le **degré de** maturité de l'apprentissage

R2 - Élaborer une démarche de la **gestion de la prise de risque spécifique aux IA**

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

En effet, de nouveaux risques apparaissent tels 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, la non-protection des données personnelles, économiques et industrielles, la non-responsabilisation des différentes parties prenantes, …

Dans ce contexte, de nouveaux principes sont à prendre en compte : principes de minimisation des risques, de précaution, de vigilance, de proportionnalité, de non-malfaisance, d'alerte, …

Pour répondre de manière adaptée aux différentes exigences attendues, il convient d'intégrer ces spécificités à la démarche générale de gestion de la prise de risque.

Il s'agit de définir une démarche qui permette d'identifier l'ensemble des dispositifs (structures organisationnelles, référentiels, directives, processus, outils, compétences, comportements, ressources informationnelles, …) à mettre en place par l'ensemble des parties prenantes qui répondent aux attentes et risques spécifiques et qui pourront s'adapter au contexte de chaque situation et aux évolutions dans le temps.

Cette démarche se doit d'être simple, souple et flexible permettant l'intégration de nouveaux éléments, facile à mettre en œuvre, à maintenir et à comprendre, et alignée aux principaux référentiels internationaux, favorisant une large adhésion.

Pour favoriser la mise en place d'une telle démarche, il convient d'

⇒ R2-1 Élaborer des **modèles spécifiques de classification des risques**

La diversité et la complexité du paysage des IA conduisent en effet à la nécessité de disposer de modèles et classifications appropriés notamment pour gérer la prise de risque. Catégoriser est un élément clé à une meilleure compréhension des différents phénomènes complexes du monde qui nous entoure et facilite ainsi l'appréhension des risques et la mise en œuvre de solutions adaptées à chaque situation.

Les éléments suivants pourraient faire l'objet de classifications : types d'algorithmes, d'apprentissage, de raisonnements, de données, d'outils, de services, ….

Il convient néanmoins 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.

R3 - Élaborer une démarche de gouvernance adaptée avec l'affectation des responsabilités associées

5 recommandations pour donner confiance

Plusieurs décisions concernant diverses parties prenantes vont nécessiter des arbitrages : quels engagements concernant les exigences, quelles affirmations les concernant, quel niveau de transparence, quels niveaux de risques acceptables, quel niveau de confiance souhaité, 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 faut donc pouvoir imputer les responsabilités, ce qui n'est pas aisé dans un contexte d'apprentissage profond, par exemple.

Ainsi, dans le cas d'un accident d'une voiture autonome, il peut s'agir d'un problème d'outils (reconnaissance d'images, de sons, …), de modèles, de données d'apprentissage, de données opérationnelles, d'erreurs d'algorithmes de décision, de choix éthiques, de biais, d'aléas (phénomènes naturels, imprévus, niveau de maturité, taux d'erreur accepté, …), de la voiture, … Dans ce contexte, à qui imputer la responsabilité ?

Le processus de décision dépend grandement de la responsabilité qu'oblige cette décision. Mais plusieurs facteurs spécifiques aux IA conduisent à rendre encore plus difficile l'affectation des responsabilités : aspect diffus, non traçabilité des actions, non traçabilité des résultats aux sources, difficulté d'apporter des éléments probants, insécurité juridique, …

Il est souvent proposé pour traiter ce sujet que l'humain conserve en toute situation le contrôle final des décisions. Mais qu'en est-il de leur responsabilité si ces humains ne fondent pas leurs décisions sur les résultats prédictifs issus des IA plus « fiables » que les humains mais dont ils n'auraient pas la parfaite maîtrise du fait d'une certaine opacité ? Pourrait-il être tenu responsable d'une perte de chance pour le patient de guérir ?

Qui dit donc prise de décision, arbitrages, engagements pris, responsabilités, … dit dispositif de gouvernance adapté (évaluation des options de création de valeur, choix des orientations, pilotage de la performance et de la conformité au regard de ces orientations, des engagements pris et de la règlementation en vigueur, …). Dans ce contexte, il convient d'identifier la démarche et les leviers qui pourront permettre l'affectation pertinente des responsabilités aux résultats issus des IA (outils d'explicabilité, de traçage des décisions et des actions, traçage aux résultats, conservation des éléments probants, place de l'assurance, évolution statut juridique, …).

R4 – Identifier les bonnes pratiques spécifiques qui répondent aux exigences et risques spécifiques liés aux IA

De nombreux référentiels de bonnes pratiques en matière de système d'information peuvent être utilisés dans le contexte du développement et de l'utilisation des IA (COBIT 2019, ITIL V4, ISO, …). Néanmoins, il convient de les compléter pour répondre aux exigences et aux préoccupations spécifiques aux IA notamment en matière de transparence et d'explicabilité et en matière d'équité et de non-discrimination.

Plusieurs outils et techniques ont été développés ou sont en cours de développement. Mais, souvent, ils ont des limites ne répondant qu'à certaines exigences et que dans certains contextes (boîtes noires, …). Il convient donc d'identifier ces bonnes pratiques, leurs limites et le niveau de confiance qu'elles peuvent apporter à certaines assertions concernant les exigences attendues.

Dans ce contexte, la mise en place d'outils de tests de la performance et de la fiabilité des algorithmes est un élément déterminant dans la recherche de confiance. Il convient donc de

⇒ R4-1 Définir les **tests e**t les outils associés adaptés par catégorie d'IA pour les phases d'apprentissage et opérationnelle, et les outils permettant de les réaliser. Définir aussi les tests d'algorithmes à effectuer dans le cadre des audits.

Un des dispositifs privilégiés pour valider le bon fonctionnement des IA est d'effectuer des tests d'algorithmes. Mais les tester dans un contexte d'apprentissage profond peut amener son lot de difficultés. Ces algorithmes sont complexes, opaques, utilisant un nombre 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 en cours de développement susceptibles de répondre pleinement aux attentes en la matière ? Il nous paraît utile d'essayer de répondre à ces questions.

De même, les tests d'algorithmes sont largement utilisés par les auditeurs. Leur utilisation dans un contexte d'IA est-elle adaptée pour satisfaire leurs diligences ? Y-a-t-il des outils spécifiques qui faciliteraient leurs travaux ?

R5 - Obtenir une « certification » par un tiers de confiance pour les IA « sensibles »

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 (labels à définir).

Pour cela, il faut déterminer les objets à auditer, identifier les parties prenantes intéressées, identifier le type d'opinion attendue et les affirmations 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

⇒ R5-1 Proposer des **exemples d'affirmations** à certifier

Plusieurs éléments doivent être définis pour déterminer le type et l'objectif d'un audit spécifique et notamment les affirmations à certifier et le niveau de confiance attendu. Il faut s'assurer que les options choisies in fine répondent à des besoins réels et à des tarifs soutenables.

Afin de faciliter cet exercice, il serait utile d'illustrer les types d'opinion possibles à l'issue d'un audit. On pourra alors voir le type d'apport de confiance correspondant et si cela convient.

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

⇒ R5-2 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.

Janvier 2021
Rédigé par Patrick STATCHENKO