IMPACT AI

LES BRIEFS DE L'IA RESPONSABLE

8

Gouvernance et gestion des risques

FÉVRIER 2025

Introduction

Ce « Brief de l'IA responsable » d'Impact AI rend compte de ses travaux et propose des pratiques et des initiatives pour faire en sorte d'assurer une gestion appropriée des risques inhérents aux systèmes et modèles d'IA (générative) développés, déployés et/ou mis à disposition par les organisations.

Il s'appuie sur les retours d'expérience des membres d'Impact AI et diverses études et initiatives récentes afin d'offrir une vue d'ensemble sur la gestion des risques de l'IA (générative) et la gouvernance à assoir en la matière.

1/ Le contexte

L'Intelligence Artificielle (IA), comme toute technologie en plein essor, est source d'opportunités, mais également de nouveaux risques qui doivent être maîtrisés pour en retirer toute la valeur. La bonne gestion de ces risques est la pierre angulaire de la mise en place d'une IA digne de confiance et responsable.

Forte de ce constat, l'Union européenne (UE) a travaillé à la mise en place d'un règlement européen sur l'IA (EU AI Act), entré en vigueur le 1er août 2024 à la suite de sa publication au journal officiel de l'Union européenne. Ce règlement suit une approche « basée sur le risque » et précise les exigences relatives aux systèmes d'IA, ou encore les modèles d'IA à usage général (GPAI), en fonction de leur niveau de risque. Les exigences établies sont d'autant plus strictes que le risque de nuire à la santé, à la sécurité, et/ou aux droits fondamentaux des citoyens de l'Union européenne est élevé :

→ **Les systèmes à haut risque auront ainsi,** en fonction du cas d'utilisation, des contraintes particulières de gestion des risques et de gestion de qualité, avec la mise en place de systèmes de management afférents, orientés « produit ». Le règlement sur l'IA décrit en effet des exigences relatives à un produit avec en ligne de mire le marquage CE. Une requête de standardisation de la Commission vers le CEN-CENELEC[1] prévoit la création de plusieurs « Normes Harmonisées », valables pour toutes les industries et conférant, à la condition qu'elles soient publiées au Journal officiel de l'UE, une présomption légale de conformité aux systèmes d'IA, développés conformément à ces textes. Elles incluent une norme sur la gestion des risques des systèmes d'IA.

→ **Pour les modèles à usage général (GPAI),** les risques de l'IA sont considérés au niveau de la technologie elle-même avec, à la clé, de nouveaux risques (par exemple quant au respect de la propriété intellectuelle ou l'influence particulière sur les utilisateurs et l'impact sur l'éducation et le travail). Le règlement sur l'IA a défini pour les modèles GPAI un niveau de risque systémique, avec des obligations supplémentaires pour les fournisseurs de tels modèles. La définition et la gestion des risques relatifs à tous ces modèles GPAI sera intégrée dans un Code de pratiques en cours d'élaboration par le Bureau européen de l'IA et attendu pour mai 2025. Il n'est pas encore déterminé à ce stade si les entreprises faisant de l'affinage (fine-tuning) de modèles de fournisseurs seront également considérées comme des fournisseurs.

Dans les organisations (de toutes tailles), cette nouvelle gestion des risques s'impose et doit être anticipée dès à présent, avec une intégration dans les processus existants le cas échéant. L'inévitable évolution en fonction des rédactions en cours, des clarifications apportées, des textes précédents doit être prise en compte, au même titre que la capacité à couvrir toutes les autres législations, règlements et cadres s'appliquant.

La norme ISO/IEC 42001:2023 - Intelligence artificielle - Système de management (AIMS) propose un cadre pour aider les organisations à gérer leurs projets d'IA et permet de maximiser les avantages de l'IA, tout en renforçant la confiance de leurs clients et/ou leurs partenaires. Les normes internationales de l'ISO/IEC constituent des bonnes pratiques et des lignes directrices, que les organisations peuvent utiliser comme guide pour leur système de management d'entreprise des risques. Nous pouvons citer à ce titre également la norme ISO/IEC 23894 :2023 - Intelligence artificielle - Recommandations relatives au management du risque.

Risques déterministes et risques non déterministes

La gestion des risques déterministes repose principalement sur le respect des principes fondamentaux de sécurité pour garantir une bonne hygiène en matière de cybersécurité. Les systèmes d'IA, y compris ceux d'IA générative, sont exposés aux mêmes risques cyber que les logiciels

1. Ensemble, le CEN et le CENELEC fournissent une plate-forme pour l'élaboration de normes européennes et d'autres spécifications techniques dans un large éventail de secteurs, en veillant également à ce que les normes correspondent à toute législation pertinente de l'UE.

traditionnels. Il est donc crucial de mettre en place des mesures de sécurité robustes pour protéger ces systèmes contre les menaces courantes telles que les attaques par déni de service, les intrusions, et les logiciels malveillants. En suivant les meilleures pratiques en matière de cybersécurité, les organisations peuvent réduire significativement les risques déterministes associés à l'IA.
Les risques non déterministes sont spécifiques aux systèmes d'IA et aux contenus qu'ils génèrent le cas échéant. On y retrouve des risques déjà mentionnés dans les précédents « Briefs de l'IA responsable », tels que les biais algorithmiques, les erreurs que l'IA peut commettre ou les hallucinations ou fabrications, c.à.d. des contenus inventés dans un récit globalement cohérent, ou encore le risque de mauvaise interprétation d'un résultat par l'utilisateur. On peut aussi mentionner la possibilité de chercher à tromper, troubler ou bien de contraindre une IA. Ces attaques sont chaotiques, difficiles à identifier et ne peuvent pas être stoppées de façon déterministe. Ce type de risques est en soi nouveau et de tels risques ne peuvent pas être « étiquetés » comme des vulnérabilités traditionnelles : on peut seulement les rendre plus ou moins probables. C'est ce que doivent assurer une gestion et une gouvernance appropriées en la matière.

Les retours d'expériences des membres d'Impact AI

Un certain nombre d'entreprises et d'organisations membres d'Impact AI ont mis en œuvre des solutions concrètes afin de gérer les risques de leurs systèmes et modèles d'IA et permettre une gouvernance à l'échelle.

AXA : élaboration d'un cadre d'évaluation des risques de façon collaborative

Les risques liés à l'intelligence artificielle dépassent ceux traditionnellement associés aux systèmes non-IA et peuvent se manifester sous diverses formes (ex : biais dans les algorithmes, manque de transparence dans les processus de prise de décision, vulnérabilités en matière de cybersécurité, etc.).
Pour assurer une gouvernance appropriée tout au long du cycle de vie des systèmes d'IA, un cadre d'évaluation des risques IA a été élaboré chez AXA, grâce à un effort collaboratif entre les entités et les équipes du Groupe. Ce cadre remplit deux objectifs clés : accompagner la gestion des projets et produits IA tout en assurant un cadre cohérent d'évaluation, d'atténuation ou de remédiation des risques. Il se compose de deux éléments clés :

→ **Un lexique des risques IA,** qui définit un vocabulaire commun à l'ensemble du groupe et propose des méthodes pour identifier et remédier à ces risques.

→ **Une matrice d'évaluation des risques IA,** qui détaille les bonnes pratiques et directives du Groupe en matière d'évaluation des risques liés à l'IA.

Étant donné la nature transversale de l'IA, ces travaux ont été intégrés aux normes, politiques et cadres existants (ex : sécurité, protection des données, architecture, etc.) afin d'éviter les doublons et d'optimiser les processus. De plus, le secteur de l'assurance étant déjà familiarisé avec la gestion des risques, il bénéficie d'une gouvernance solide en place depuis de nombreuses années. Ainsi, le cadre présenté s'inscrit logiquement dans celui, plus large, de la gestion des risques opérationnels, tout en étant en adéquation avec les principes d'IA responsable établis par AXA.

Crédit Agricole : le rôle central du DataLab Groupe

Le DataLab du Groupe Crédit Agricole est son pôle de référence pour la conception interne de solutions Data & IA. Créé en 2016, il a opéré plusieurs transformations successives afin, tout en conservant un ADN R&D, de se doter des capacités pour développer en interne des solutions Data & IA, innovantes, nativement industrielles, de

confiance et responsables. L'ensemble des projets mis en œuvre suivent une méthode exigeante ayant fait l'objet d'une certification LNE (axée sur les thématiques IA de confiance), et d'une labélisation RSE (LabelIA Labs, IA responsable et de confiance). Cette méthode inclue une taxonomie des risques IA à prendre en compte pour les projets, incluant les risques techniques, mais aussi les risques non techniques, ainsi que des exemples de mesures d'atténuation qui permettent de réduire le risque. Ces mesures d'atténuation doivent être vues comme un ensemble permettant d'aboutir à un système robuste, et non individuellement, car aucune de ces mesures n'est fiable à 100 %.

Quelques exemples concrets de travaux menés par le DataLab Groupe trouvant des applications concrètes dans nos projets :

→ Avant les LLM, les IA « classiques » portaient déjà des risques, et les études ont, par exemple, couvert l'efficacité des attaques adverses sur les images documentaires et également l'efficacité des entraînements adverses comme technique d'atténuation, ainsi que les attaques de reconstruction sur des modèles d'extraction d'information ;

→ Dans le cadre de la Chaire IA de confiance et responsable menée avec le laboratoire LIX de l'école polytechnique, des travaux, relatifs à la mémorisation des données par les LLM, sont en cours ;

→ Concernant les risques liés aux IA génératives, le Crédit Agricole a comparé des solutions open source d'évaluation et de détection des risques liés aux LLM et conclu que ces solutions n'étaient aujourd'hui pas assez matures/performantes pour permettre une détection fiable et qu'un red teaming manuel était nécessaire pour exposer les failles de nos systèmes et les patcher par la suite ;

→ Des travaux autour de l'atténuation de certains risques connues des LLM, tels que p. ex. pour lutter contre :

- L'exfiltration de données des mesures comme l'entraînement adverse, la confidentialité différentielle et l'anonymisation, etc.
- Les injections d'invites (prompts) des mesures comme le filtrage de l'entrant, la re-tokenisation, la recherche par similarité parmi des prompts d'attaques connus, etc.

Ces travaux d'avant-garde menés par le DataLab Groupe servent l'ensemble du Groupe CA et ont permis, en particulier, d'alimenter la taxonomie des risques IA du Groupe CA, les scénarii d'impact, méthodes de détections et d'atténuation, qui sont partie intégrante du cadre normatif IA Groupe publié en octobre 2024, reprenant les exigences du Règlement européen sur l'IA, ainsi que les engagements volontaires du Groupe en faveur d'une IA de confiance et responsable.

« Maîtriser les risques de l'IA, c'est garantir que chaque avancée technologique renforce notre sécurité et notre confiance, sans compromettre les fondations solides que nous avons bâties » explique Aldrick Zappellini, Directeur Data & IA Groupe et Chief Data Officer du Groupe Crédit Agricole.

Microsoft France : Gouverner, Cartographier, Mesurer et Gérer

Microsoft France souligne l'utilisation de son Standard d'IA responsable pour formaliser un ensemble d'exigences en matière d'IA générative, qui s'inscrivent dans un cycle de développement de l'IA responsable, telles que couvertes dans son rapport annuel inaugural de transparence l'IA responsable. Ces exigences s'alignent sur la série d'axiomes ou de fonctions essentielles du cadre de gestion des risques liés à l'IA (AI RMF 1.0) du NIST, c'est-à-dire : Gouverner, Cartographier, Mesurer et Gérer, dans le but de réduire les risques liés à l'IA générative et les préjudices qui y sont associés.

La cartographie et la priorisation des risques est une première étape critique – et itérative – pour aller ensuite vers la mesure et la gestion systématiques (de la prévalence) des risques associés à l'IA (générative). Cette dernière éclaire en effet toutes les décisions qui en découlent. Dans l'approche globale adoptée, les risques sont identifiés via :

→ Des analyses d'impact de l'IA responsable : ces analyses, comme exigé par le Standard d'IA responsable à chaque jalon majeur, s'avèrent

précieuses afin de s'assurer de l'exploration en profondeur du système d'IA envisagé, et ce, dès les toutes premières étapes de la conception. Elles mettent en effet en exergue, pour chaque cas d'utilisation prévue et prévisible et vis-à-vis de chaque partie prenante directe ou indirecte, les risques et les préjudices potentiels associés, ainsi que les mesures d'atténuation pour y remédier, Cf. modèle d'analyse d'impact de l'IA responsable de Microsoft et son guide compagnon – Le projet au stade approbation de norme ISO/IEC FDIS 42005 Analyse d'impact d'un système d'IA (projet final) propose une alternative dans le cadre du corpus ISO/IEC 4200x.

→ Des revues de la (cyber)sécurité et de la protection de la vie privée : son cycle de développement sécurité (Secure Development Lifecycle ou SDL) d'usage obligatoire a été mis à jour – et continue à l'être dans le cadre de l'initiative pluriannuelle « La sécurité avant tout » (Secure Future Initiative ou SFI) –, pour intégrer les étapes de gouvernance de son Standard d'IA responsable. À ce titre, des processus d'identification et d'analyse des risques, tels que la modélisation des menaces des systèmes d'IA et de leurs dépendances, permettent une compréhension globale et unifiée des risques et des mesures d'atténuation pour les systèmes envisagés.

→ Des pratiques de red teaming et de tests de résistance ont été mises au point par l'équipe AI Red Team (AIRT). Cela consiste à sonder et tenter de « casser » le système (et/ou le ou les modèles utilisés) dans des conditions réelles d'utilisation et de mieux comprendre comment les risques identifiés se manifestent, de mettre à jour les angles morts et d'identifier ainsi de façon pro-active de nouveaux risques qui n'auraient pas été anticipés au départ, de définir des stratégies de défense et de valider ultérieurement la pertinence des mesures d'atténuation envisagées. Des années de red teaming ont permis d'acquérir une connaissance inestimable des stratégies les plus efficaces et d'en tirer huit leçons présentées dans le livre blanc Lessons from Red Teaming 100 Generative AI Products.

Comme il n'y a pas de « ligne d'arrivée » – il s'agit d'un voyage et non d'une destination –, plus qu'une formalité, l'approche retenue est un processus structuré et itératif qui responsabilise. Les risques évoluent, tout comme les stratégies d'atténuation, et c'est pourquoi il est impératif de revoir l'évaluation des risques d'un système (ou de l'une de ses fonctionnalités) chaque fois que nécessaire tout au long de son cycle de vie et à tout le moins toutes les années. Les évaluations ainsi faites, au travers des analyses, revues et pratiques précédentes sont un guide vivant qui s'adapte au paysage en évolution constante pour une IA digne de confiance.

Les bonnes pratiques identifiées

Selon l'avis de nos membres et plus globalement de l'industrie, le cadre de gestion des risques liés à l'IA (AI RMF 1.0) du NIST constitue, à date, l'un des cadres les plus respectés. Ce travail solide élaboré au travers d'un processus consensuel et transparent bénéficie, en effet, des années d'expérience du NIST dans le domaine de la cybersécurité et de la sûreté, où des cadres et des normes similaires ont joué un rôle essentiel.
Ce cadre gratuit, volontaire et flexible à destination des organisations, s'articule autour d'une série d'axiomes ou de fonctions de base qui vise à :

1. Gouverner. Systématiser et organiser les activités au sein de l'ensemble de l'organisation :
> Établir une culture de gestion des risques au sein de l'organisation.
> S'aligner sur les principes, les politiques et les priorités stratégiques.
> Pour les systèmes d'IA considérés

2. Cartographier. Disposer d'une compréhension approfondie quant aux risques du système d'IA :
> **Identifier** et classer par ordre de priorité les risques et leurs préjudices potentiels qui pourraient résulter de ce système d'IA en procédant à des analyses, des tests d'exploration et de résistance (red teaming) itératifs.
> **Identifier** les mesures d'atténuation pour le traitement de ces risques.

3. Mesurer. Mesurer les risques et leurs impacts:
> **Évaluer les risques** ainsi retenus et leur niveau de prévalence en matière de fréquence et de gravité en établissant des mesures claires.
>**Mettre en place le test**, l'évaluation, la vérification et la validation (TEVV), à la fois vis-à-vis de ces risques et de leurs mesures d'atténuation, en créant des jeux de tests de mesure et en effectuant itérativement des mesures systématiques manuels et automatisés :

- La mesure manuelle est utile pour i) mesurer les progrès accomplis à l'égard d'un petit nombre de risques prioritaires afin d'atténuer des préjudices spécifiques, ii) définir et rapporter des métriques jusqu'à ce que la mesure automatisée soit suffisamment fiable pour être utilisée seule et iii) vérifier ponctuellement ou périodiquement la qualité de la mesure automatique.
- La mesure automatisée est utile pour i) mesurer à grande échelle avec une couverture accrue afin de fournir des résultats plus complets et ii) mesurer en continu, afin de surveiller toute régression au fil de l'évolution du système, de son utilisation et des mesures d'atténuation.

4. Gérer. Mettre en œuvre des pratiques pour atténuer les risques :
> **Exécuter les politiques établis** dans Gouverner, avec les pratiques et l'outillage identifiés pour le traitement de ces risques. Il s'agit pour cela de définir une approche itérative à plusieurs niveaux qui comprend l'expérimentation et la mesure.

> Nous constatons que la plupart des systèmes de production nécessitent un plan d'atténuation qui comprend quatre niveaux d'atténuation pour les risques cartographiés initialement : i) le modèle lui-même, ii) le système de sureté, iii) l'ancrage (grounding) et le message système, et iv) l'expérience utilisateur (UX). Dans cette approche de « défense en profondeur » :

- Les deux premiers niveaux sont généralement propres à la plateforme, où les mesures d'atténuation intégrées sont communes ou à disposition à de nombreux systèmes. Elles sont intégrées dans le modèle lui-même, p. ex. avec des techniques des techniques comme l'apprentissage par renforcement à partir du retour humain (RLHF) et l'affinage (fine-tuning) dans les modèles de base ou bien pour le second dans la plateforme qui le propose.
- Les deux niveaux suivants dépendent de l'objectif et de la conception du système considéré ; ce qui signifie que la mise en œuvre des mesures d'atténuation peut varier considérablement d'un système à un autre.

> **Répéter les mesures** en termes TEVV pour tester l'efficacité après avoir mis en œuvre des mesures d'atténuation.
> **Surveiller et améliorer en continu.**
Un tel cadre permet de guider tout au long dans les processus i) d'établissement des politiques et des pratiques, ii) d'identification des risques et de compréhension des contextes et des impacts, iii) d'évaluation et de suivi des risques en matière de mesures, et iv) de réponse et de gestion adéquates de ces risques. À cette fin, un guide et des profils sont proposés en tant que ressources supplémentaires pour une application à des besoins spécifiques comme p. ex. le profil d'IA générative à destination de ce cadre AI RMF 1.0 conçu quant aux risques uniques associés à l'IA générative, et à la fois vis-à-vis de cas d'utilisation spécifiques et d'activités trans-sectorielles.
Ce cadre peut être utilisé en conjonction avec la norme internationale 42001 :

- Les organisations qui mettent en œuvre ce cadre AI RMF 1.0 peuvent ainsi se référer aux contrôles de l'AIMS pour obtenir des conseils spécifiques.
- À l'inverse, les organisations qui mettent en œuvre l'AIMS peuvent choisir d'utiliser l'AI RMF comme cadre de gestion des risques de choix.

Les défis et questions en suspens

Les membres d'Impact AI font valoir les points d'attention suivants :
>Les risques et l'impact de l'IA générative sont évolutifs et dépendent des usages.
>La gouvernance des entreprises devra prendre en compte ce sujet et prévoir une capacité d'adaptation notamment après déploiement.
>La régulation et les « Normes Harmonisées » européennes sont encore en construction sur le sujet de la gestion des risques en 2025. Les entreprises peuvent cependant déjà se préparer à partir des normes, standards et cadres existants comme ceux cités précédemment et/ou lister dans la section suivante et doivent surtout construire cette gestion en cohérence avec leurs pratiques existantes et par rapport à leur rôle, leurs usages de l'IA et leur tolérance aux risques :

- La gestion des risques ne concerne pas seulement les équipes Sécurité et Protection des données (privacy).
- Les entreprises doivent des cadres et une culture de gestion de risque dans les équipes en interaction avec des systèmes d'IA.

Pour aller plus loin

• Atlas des risques de l'IA proposé par IBM
• AI Risk repository recensé par le MIT
• Risk management logic of the AI Act and related standards
• ISO/IEC 42001:2023 - Intelligence artificielle - Système de management
• ISO/IEC 23894:2023 - Intelligence artificielle – Recommandations relatives au management du risque
• NIST AI 100-1 Artificial Intelligence Risk Management Framework (AI RMF 1.0)
• NIST AI 600-1 Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile

Remerciements
à Philippe Beraud (Microsoft), Emilie Sirvent Hien (Orange)
et Matthieu Capron (Crédit Agricole) pour leur engagement
et leur contribution essentielle à ces Briefs de l'IA responsable.
à tous les participants aux ateliers qui ont bien voulu
partager leurs expériences et cas d'usage.

Retrouvez-nous sur

www.impact-ai.fr

pour rejoindre nos initiatives et retrouver nos divers travaux !


IMPACT AI