HUB FRANCE IA

En partenariat avec

WAVESTONE

NOTICE

PREMIERS PAS VERS L'IA DE CONFIANCE

Juin 2025

Editorial

Dans un monde où l'intelligence artificielle se diffuse, chaque jour, un peu plus dans notre quotidien et dont l'adoption se ramifie de notre société actuelle jusqu'à nos méthodes de travail, il devient essentiel de comprendre cette technologie, d'y recourir de manière raisonnée et d'utiliser des SIA (SIA) dignes de confiance.

L'IA de confiance est une notion complexe aux définitions multiples. Elle dépasse le simple cadre normatif, en englobant tant des enjeux techniques liés à la sécurité, à la robustesse et à la protection, que des questionnements éthiques, le tout afin de prendre en compte le déploiement de cette technologie en recourant à de la sensibilisation, de la formation tout en garantissant la transparence dans les usages.

Surtout, l'IA de Confiance suppose d'instaurer une gouvernance au sein des entités, qu'elles soient fournisseurs ou déployeurs de ce type de systèmes, qu'ils soient à haut risque ou à risque minime, soumis à la réglementation européenne, à des normes techniques ou à des principes éthiques.

Pour répondre et mieux appréhender cette notion complexe, le Hub France IA – en partenariat avec la société Wavestone – a réalisé un travail collectif auprès de notre écosystème, pour offrir un socle commun, un cadre d'analyse et une grille de lecture utiles à toutes les organisations, des plus grandes aux plus petites.

Ce livrable est un travail préliminaire, une première étape qui pourra être approfondie et enrichie à l'avenir. Il pose aussi une base pour bien appréhender la notion d'IA de Confiance et l'entrelacement entre les aspects techniques, juridiques et éthiques.

Ainsi, bâtir l'IA de Confiance, c'est avant tout refuser l'illusion de la facilité et préférer l'exigence de la protection humaine, de la transparence et de la souveraineté numérique. C'est aussi se donner les moyens de préserver notre autonomie stratégique face aux puissances technologiques mondiales, tout en restant fidèle à nos valeurs. Par ce livrable, nous espérons ainsi renforcer la réflexion collective autour du développement de SIA qui soit bénéfique à notre société.

Françoise Soulié-Fogelman
Conseiller scientifique

Bertrand Cassar
VP IA de Confiance

Table des matières

Introduction

Introduction

Ce document synthétise les premières analyses et recommandations du groupe de travail sur l'IA de Confiance du Hub France IA. Il a pour objectif d'aider les organisations à comprendre les enjeux et les premières étapes pour progresser vers le déploiement de l'IA à grande échelle de manière éthique, sûre et responsable.

Les travaux du Hub France IA

Le Hub France IA s'investit de manière proactive sur les enjeux liés à l'intelligence artificielle (IA) depuis ses débuts. Plusieurs groupes de travail ont été constitués pour explorer les différentes facettes de cette révolution technologique, dont celle de la confiance, en mobilisant un large écosystème de professionnels (experts IA, RSSI, juristes, chefs de projets, scientifiques de la donnée, responsables innovation…), issus de tous les secteurs et de tous types d'organisation (TPE, PME, start-up, grands groupes, organismes de recherche…). Ces groupes de travail ont pour objectif de faire émerger des projets et produire des livrables opérationnels, intégrant notamment la question de la confiance :

- Le groupe de travail sur l'IA générative a produit un guide qui éclaire le sujet du choix des modèles de type « *Large Language Models* » (LLM) dans les organisations[1]. Le livrable esquisse les contours de la notion de confiance algorithmique au sein d'un système ;
- Les groupes de travail sur la sécurisation de l'IA et « Banque et auditabilité » ont produit des livres blancs abordant les questions de robustesse d'un système et des critères de confiance pour une Intelligence artificielle traitant de données sensibles ;
- Le groupe de travail « Boussole de l'*AI Act* » produit des livrables visant à opérationnaliser la mise en place de l'*AI Act ou* Règlement sur l'Intelligence artificielle, RIA (en français) ;
- Le groupe de travail « Éthique » ayant défini des recommandations pour mettre en œuvre des chartes éthiques de l'IA et accompagner l'opérationnalisation des principes éthiques et de leurs exigences.

Ces aspects tant technique, éthique que juridique forment les éléments constitutifs de l'IA de confiance, selon la définition retenue par le GEHN IA.

Le groupe de travail « IA de confiance », s'intègre dans les actions du Hub France IA et complète les réflexions portées par les autres groupes de travail sur les enjeux ayant trait, sous différents angles, à l'IA de confiance.

[1] Hub France IA, Choisir un modèle d'IA générative pour son organisation, Juin 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/06/Hub-France-IA-Choisir-un-modele-IA-Generative.pdf

Dès les premiers échanges entre les membres du Hub France IA mobilisés, deux constats furent en effet largement partagés :

1. La notion d'**IA de confiance** est communément évoquée quand on parle d'un développement vertueux de l'IA, mais il n'y a **pas de consensus clair** sur ce que la notion recouvre ;
2. Le sujet englobe de **nombreux enjeux** : éthique, sécurité, formation, environnement, transparence, gouvernance, *etc.*, pour lesquels les organisations ont besoin d'être accompagnées.

Ainsi, ce document vise à clarifier les enjeux sous-jacents à l'IA de confiance, proposer des grilles de lecture ainsi qu'un mode opératoire pour faciliter la prise en main par les organisations.

Face à la complexité du sujet, le Hub France IA a pris le parti de prendre pour cible les **grandes organisations car** elles sont souvent pionnières dans l'expérimentation de l'IA à grande échelle, et confrontées à des enjeux de gouvernance complexes.

Ce travail est par ailleurs un **travail préliminaire** qui pourra être approfondi dans le futur, par exemple en explorant les problématiques spécifiques aux petites et moyennes structures, ou en approfondissant certains des thèmes abordés (gouvernance, méthodologie des sécurisations des projets, impact environnemental, *etc.*).

1. Pourquoi une IA de confiance ?

1. Pourquoi une IA de confiance ?

1.1 Les risques liés à l'utilisation de l'IA

Les risques liés à l'utilisation croissante de l'IA font de sa confiance un défi majeur

Depuis 2022, l'intelligence artificielle a fait l'objet d'une attention accrue du fait du développement d'outils no code et de la démocratisation de l'IA générative. L'irruption de modèles comme ChatGPT a en effet rendu l'IA générative accessible à tous, et a renforcé l'intérêt pour l'intelligence artificielle de manière plus générale. Depuis, l'IA est passée par plusieurs phases :

- **En 2023**, c'est la phase d'**expérimentation** sur l'IA Générative : les cas d'usage se multiplient, les directions métiers s'emparent de la technologie et les *Proofs of Concepts* (PoCs) s'enchaînent à tous les niveaux, alors que les collaborateurs expérimentent en autonomie sur les solutions publiques, mettant en évidence les risques de ces solutions (hallucinations, confidentialité, fracture numérique, *etc.*). Les IA plus « classiques » sont déjà bien implantées dans les organisations selon des niveaux de maturité variés.
- **En 2024,** l'IA générative subit le test de **la preuve de valeur** : les organisations ont cherché à distinguer les usages réellement utiles, à conserver uniquement les cas d'usage créateurs de valeur, et à stabiliser leurs briques technologiques.
- **En 2025**, les organisations mettent la priorité sur la **rationalisation**. L'objectif est d'intégrer l'IA dans les systèmes d'information, les processus métiers, en maîtrisant les risques et en optimisant les dépenses.

A mesure que l'intelligence artificielle a gagné en maturité, faisant apparaître de plus en plus de cas d'usages concrets, la réflexion concernant les risques qu'elle amène s'est accélérée. Plusieurs initiatives ont ainsi vu le jour pour les recenser. Le MIT a par exemple développé une taxonomie détaillée, identifiant plus de 1600 risques liés à l'IA, classifiés en 7 domaines et 24 sous-domaines[2] :

Domaine	Sous-domaine
Discrimination & Toxicité	Discrimination injuste et mauvaise représentation
	Exposition à du contenu toxique
	Performance inégale entre les groupes
Vie privée & Sécurité	Compromission de la vie privée par obtention, fuite ou inférence correcte d'informations sensibles
	Vulnérabilités et attaques de sécurité des SIA
Désinformation	Informations fausses ou trompeuses

[2] *MIT AI Risk Repository* : https://airisk.mit.edu

	Pollution de l'écosystème d'information et perte de la réalité consensuelle
Acteurs malveillants	Désinformation, surveillance et influence à grande échelle
	Cyberattaques, développement ou utilisation d'armes, et dommages massifs
	Fraude, escroqueries et manipulation ciblée
Interaction Homme-Machine	Sur-dépendance et utilisation dangereuse
	Perte d'autonomie et d'expertise humaine
Socio-économique & Environnemental	Centralisation du pouvoir et distribution inégale des bénéfices
	Augmentation des inégalités et déclin de la qualité de l'emploi
	Dévaluation économique et culturelle de l'effort humain
	Échec de la gouvernance
	Dommages environnementaux
Sécurité, échecs et limitations des SIA	IA poursuivant ses propres objectifs en conflit avec les objectifs ou valeurs humaines
	IA possédant des capacités dangereuses
	Manque de capacité ou de robustesse
	Manque de transparence ou d'interprétabilité
	Bien-être et droits de l'IA
	Risques multi-agents

Tableau 1: Domaines et sous-domaines de risques selon le MIT

Ces travaux ne se sont pas limités à la partie théorique. Dans le cadre du *Partenariat Mondial sur l'intelligence artificielle*, l'OCDE a ainsi recensé l'ensemble des incidents et dangers signalés dans une sélection de médias internationaux entre janvier 2016 et janvier 2024. Cette analyse mettait notamment en évidence une importante accélération à partir de janvier 2023, que l'on peut imputer à la fois au développement de l'intelligence artificielle générative, mais aussi à l'attention accrue portée envers l'IA par la presse.

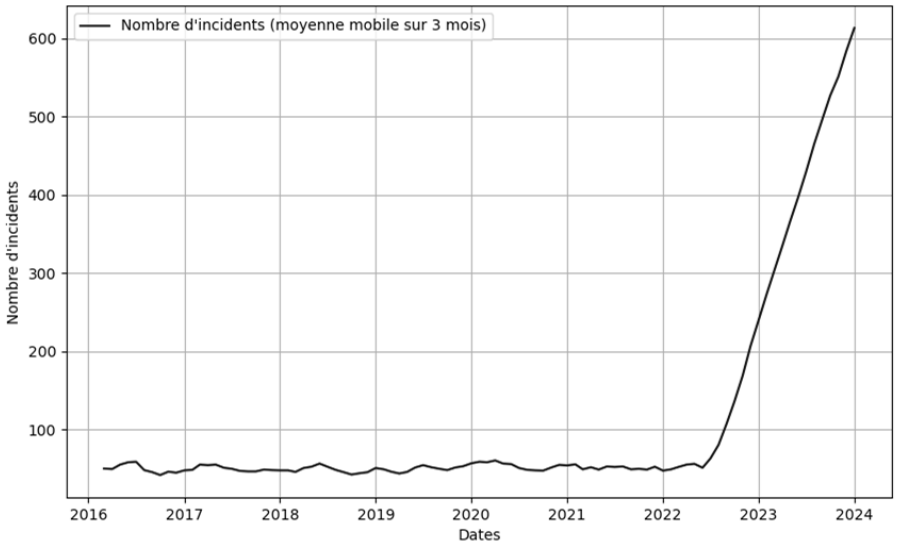

Figure 1 Nombre d'incidents rapportés par la presse entre janvier 2016 et janvier 2024 (Source : OCDE.AI Policy observatory)

Le site « *AI Incident Database* [3]», portée par l'ONG *Responsible AI Collaborative*, recense également les différents incidents rapportés par la presse.

Pour que l'IA s'installe durablement dans le quotidien des organisations et des utilisateurs, elle devra être fiable, explicable, conforme, sécurisée — et perçue comme étant de **confiance**. Ceci est indispensable pour son adoption à grande échelle et pour qu'elle impacte la société de manière positive et viendra avec l'IA responsable[4].

1.2 Les enjeux auxquels les organisations devront faire face

1.2.1 Des enjeux humains et culturels

L'introduction de l'IA dans les organisations ne saurait se réduire à une étape de modernisation technique ou d'impératif à l'innovation ; **elle inaugure un nouveau cycle d'apprentissage collectif**, où le travail humain et les capacités des machines s'interpénètrent. Loin d'un simple ajustement, ce déploiement remet en question les équilibres établis entre méthodes de fonctionnement, efficacité opérationnelle, reconnaissance des compétences et sens du travail. La coexistence de deux logiques – l'une centrée sur la course à la performance, l'autre sur le vécu professionnel, sur le terrain, **produit des tensions structurelles.**

Pour les surmonter, les organisations doivent apprendre à trouver le bon **équilibre entre la recherche de gains productivistes** théoriques avec l'introduction de SIA et les **réels besoins des collaborateurs**. C'est dans cette capacité à faire dialoguer des intérêts divergents que se jouent la durabilité des projets d'IA et l'émergence d'un vrai terreau

[3] *AI Incident Database* : https://incidentdatabase.ai/fr/
[4] Pratique de conception, de construction et de déploiement de l'IA de sorte à avoir un impact positif sur les clients et la société

fertile pour la confiance. L'IA de confiance, dans ce contexte, n'est pas une technologie en soi, mais le fruit d'un processus collectif de traduction et d'ajustement entre acteurs, finalités et outils.

Or, ce compromis ne peut émerger que si les SIA s'inscrivent dans le travail réel, là où les outils sont véritablement mis à l'épreuve. La confiance ne se construit pas dans les promesses technologiques ou les intentions de conception, mais dans l'usage quotidien, dans l'interaction directe entre l'humain et la machine. Elle ne se décrète pas dans les comités de direction ni dans les chartes réglementaires ; elle se gagne opérationnellement, dans les expériences concrètes que les collaborateurs font de ces technologies**. Cela implique que l'adoption soit pensée comme le cœur du processus de conception de l'IA de confiance. Pour cela, trois conditions sont déterminantes** :

- L'IA doit répondre à un besoin réel et perçu comme tel, car il ne peut y avoir de confiance sans utilité ;
- Les utilisateurs doivent aussi comprendre comment fonctionne la technologie pour pouvoir l'interpréter, l'interroger et la corriger le cas échéant ;
- Enfin, une formation adaptée est indispensable — non seulement pour apprendre à se servir de ces outils, mais aussi pour intégrer les principes qui les régissent. Ces démarches permettent de transformer l'IA en bras armé — un partenaire interprétable, ajustable et complémentaire aux capacités cognitives humaines.

1.2.2 Des enjeux opérationnels de fiabilité et d'utilité

La **fiabilité des SIA est cruciale pour faciliter l'adhésion de ces outils et stimuler les usages innovants au sein des organisations**. En effet, les attentes de performance, de gain de temps, de découvertes et de transformation en retour à l'usage des SIA sont fortes, qu'il s'agisse de déployer des SIA pour améliorer des processus existants ou d'inventer de nouveaux usages en renfort aux missions existantes. Or, un manque de fiabilité pourra non seulement faillir à ces objectifs, mais également entraîner des conséquences néfastes, en diminuant la confiance qu'ont les utilisateurs envers cette technologie, et les amenant potentiellement à s'en détourner.

Par ailleurs, des opportunités génératrices de valeurs pour l'organisation, par exemple commerciales, peuvent être manquées si les algorithmes biaisés ou imparfaits sont développés et/ou utilisés.

Sans oublier les ressources précieuses de temps, d'argent et d'efforts gaspillées sur des projets d'IA se révélant inutiles, dysfonctionnels, ou biaisés ou non éthiques. Citons le SIA de création de CV développé par Amazon en 2014, qui discriminait les femmes pour les postes techniques. Les équipes d'Amazon ont travaillé en vain pendant 3 ans pour essayer de corriger ce problème. [5] Il est donc essentiel pour une organisation d'être en

[5] « Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women", Roberto Iriondo, Carnegie Mellon University

mesure d'évaluer et de maintenir à l'état de l'art la fiabilité par la maîtrise des risques autour des IA qu'elle développe ou emploie, afin d'assurer la confiance autour de ses produits.

1.2.3 Des enjeux de sécurité et de transparence

Les systèmes d'intelligence artificielle représentent également un **nouveau vecteur de risques** en matière de sécurité. Parce qu'ils manipulent de grandes quantités de données et qu'ils sont parfois intégrés dans des processus critiques, ils deviennent des cibles privilégiées pour les attaquants. Un système d'IA mal protégée peut ainsi être exploitée notamment pour exfiltrer des données sensibles, accéder à des systèmes internes ou contourner des mécanismes de sécurité.

Certaines caractéristiques des IA introduisent des nouveaux défis à relever pour la sécurisation de ces systèmes. Leur comportement est souvent **non déterministe**, ce qui rend difficile la prédiction et le contrôle des résultats. De plus, ces systèmes **évoluent dans le temps** (via des mises à jour ou de l'apprentissage en continu) pouvant introduire de nouvelles failles. Enfin, un **manque de transparence** peut limiter la capacité à détecter et comprendre rapidement les anomalies ou détournements ou favoriser des mésusages posant un défi majeur en matière de sûreté.

Les vulnérabilités des SIA ne sont pas uniquement théoriques : elles se manifestent déjà dans les usages concrets :

- Les **hallucinations** (réponses incorrectes ou inventées par les modèles d'IA génératifs) peuvent induire des erreurs opérationnelles, en particulier lorsque les utilisateurs font une confiance excessive à la machine.
- Des risques de **fuites de données** sont également présents, notamment lorsqu'un modèle a été exposé à des données sensibles durant son entraînement et les restitue accidentellement dans ses réponses.
- Les SIA peuvent être une **porte d'entrée vers des infrastructures critiques** : si l'IA est intégrée à un système d'information plus large, une compromission du modèle ou de ses interactions peut permettre un mouvement latéral vers d'autres ressources. Cela est particulièrement vrai pour les IA intégrées à des assistants vocaux, des services de support client ou des outils d'automatisation métier, qui manipulent souvent des données confidentielles et/ou ont des droits étendus.

Ces vulnérabilités peuvent être activement exploitées par des attaquants à travers des techniques d'**attaques cyber spécifiques aux SIA**. Par exemple, les attaques par empoisonnement visent à injecter des données malveillantes dans le corpus d'apprentissage d'un modèle afin de fausser son comportement. L'attaque par évasion, quant à elle, consiste à fournir une entrée soigneusement manipulée pour induire une mauvaise classification par un modèle (par exemple, faire passer un stop pour une limitation de vitesse dans un SIA de conduite autonome). Une autre technique préoccupante est celle du « prompt injection » contre les SIA génératives : en

manipulant les instructions fournies au modèle, un attaquant peut le pousser à ignorer ses règles de sécurité ou à divulguer des informations restreintes.

1.2.4 Des enjeux environnementaux

Intégrer les **coûts environnementaux** dans les projets IA est une nécessité éthique pour les organisations vis-à-vis de leurs responsabilités sociétales. Le GEHN incite d'ailleurs à « la protection des individus [...] au niveau le plus élémentaire », pour que l'innovation technologique ne se fasse pas au détriment des générations futures. De nombreux impacts environnementaux peuvent en effet être envisagés :

- La consommation d'énergie des datacenters liés à l'IA a explosé depuis 2019, **inversant la trajectoire Net Zéro** des GAFAM, à l'image de toute organisation développant/entrainant des modèles ou les intégrant. Cette augmentation drastique de la consommation d'énergie interroge nos capacités à répondre durablement aux besoins énergétiques de l'IA. Alimenter les "clusters" de datacenters nécessite désormais plusieurs centaines de Mégawatt voire de Gigawatt[6].
- L'impact sur la durée de vie des équipements, car certaines personnes pourront vouloir renouveler leur smartphone pour acquérir un modèle plus récent, dopé à l'IA, et profiter de ces nouveaux usages. Cela induit un raccourcissement de la **durée de vie des équipements**.
- Le développement de l'IA est aussi consommatrice de **ressources abiotiques** cruciales et limitées, comme l'eau et le sol. Les prélèvements d'eau et la consommation d'eau augmentent avec la construction et le fonctionnement des datacenters. Selon l'AIE, la consommation d'eau liée à la seule fabrication de puces pour les centres de données atteindra 70 milliards de litres en 2030, l'équivalent de la consommation annuelle de la ville de Francfort – soit 50% de plus qu'en 2023. Dans le monde, les datacenters consomment environ 560 milliards de litres par an, un volume pouvant atteindre près de 1 200 milliards d'ici 2030.
- D'autres impacts, plus difficiles à modéliser (**déchets générés, artificialisation des sols, pressions sur les ressources minérales et la biodiversité**), sont également bien réels, de long-terme et potentiellement irréversibles (contribution au dépassement des seuils critiques des limites planétaires).

Le développement de l'IA affecte donc fortement la capacité des organisations à maintenir une trajectoire de décarbonation ambitieuse (et plus largement de transition environnementale). La prise en compte des **impacts environnementaux directs comme indirects** (selon les cas d'usages, ces derniers peuvent avoir un effet fortement démultiplicateur, avec un certain manque de transparence des principaux développeurs) doit donc se traduire dans la gouvernance et la pyramide de normes

[6] AGENCE INTERNATIONALE DE L'ENERGIE, avril 2025, Energy and AI

(ex : Référentiel général pour l'IA frugale de l'AFNOR SPEC sur l'éco-conception, green-algorithms) afin que l'adoption de l'IA se fasse en connaissance de cause.

1.2.5 Des enjeux réglementaires et de gouvernance

Le prise en compte des risques liés à l'IA devient dans tous les cas inévitables pour les organisations, particulièrement en Europe, car de plus en plus d'autorités de régulations s'emparent du sujet. On pense évidemment au règlement européen sur l'intelligence artificielle (RIA ou *AI Act*), qui a déjà posé un **cadre réglementaire important** concernant la conception et le déploiement de SIA. Néanmoins, d'autres textes seront également pertinents :

- Les textes législatifs sur les données personnelles, et en premier lieu le RGPD, qui reste fondamental dès lors qu'il y a traitement de données à caractère personnel, qui toucheront tout SIA manipulant de telles données ;
- Les réglementations sectorielles existantes (santé, finance, *etc.*), qui pourraient concerner également les SIA ;
- Les spécificités nationales*[7] ;
- Le droit international et les principes de diligence raisonnable.

Les organisations doivent donc dès maintenant être attentives aux cadres réglementaires et à leur évolution qui s'appliqueront aux SIA déployés en leur sein, qu'elles soient simples utilisatrices ou conceptrices de SIA.

[7] Comme en France via le Code du Travail qui impose des impératifs de dialogue social avec les représentants du personnel quant à l'introduction de nouvelles technologies.

2. Mettre en place une gouvernance autour de l'IA de confiance

2. Mettre en place une gouvernance autour de l'IA de confiance

Pour que les organisations puissent faire face à ces enjeux, elles doivent d'abord **définir la gouvernance** qui prendra ce sujet en main. Une gouvernance adaptée ne se limite pas à "superviser l'IA", elle crée les conditions de la confiance : en pilotant les risques, en alignant les parties prenantes, et en portant une vision éthique et stratégique.

2.1 Définir les principes d'une IA de confiance propres à l'organisation

2.1.1 Les définitions multiples de l'IA de confiance

La notion d'IA de confiance ne fait pas consensus. En effet, rien qu'entre les différentes régions du monde, les approches varient considérablement :

- **Aux États-Unis**, l'approche est normative et volontariste, portée par le NIST AI Risk Management Framework (AI RMF), avec une forte orientation technique et opérationnelle sur la gestion des risques ;
- **En Europe**, l'approche retenue est celle portée par le GEHN IA (Groupe d'Experts de Haut Niveau sur l'IA) mis en place par la Commission Européenne, et qui a inspiré l'*AI Act* (RIA) ;
- **En Chine,** l'approche est réglementaire, centralisée et sectorielle, guidée par une vision de contrôle étatique et de sécurité publique. La Cyberspace Administration of China (CAC) a publié plusieurs réglementations telles que la réglementation sur les algorithmes de recommandation, la Data Security Law ou encore la PIPL ;
- Les **directives provisoires sur les services d'IA générative** (2023), qui imposent un cadre strict de conformité, de transparence, de supervision humaine et d'interdiction de contenus illégaux ou jugés "désinformants".
- **Au niveau international**, la norme ISO 42001 cherche à aider les organisations à identifier et gérer les risques liés à l'IA, en intégrant des exigences sur la transparence, l'éthique, la supervision humaine et la sécurité, tandis que l'OCDE a défini en 2019 cinq grands axes pour une IA bénéfique, responsable et respectueuse des droits humains, encourageant la transparence, la robustesse, la responsabilité et le développement inclusif des SIA.

Cette diversité des approches et des conceptions de l'IA de confiance constitue un premier facteur de complexité pour les organisations internationales, même si ces textes se rejoignent sur bien des aspects.

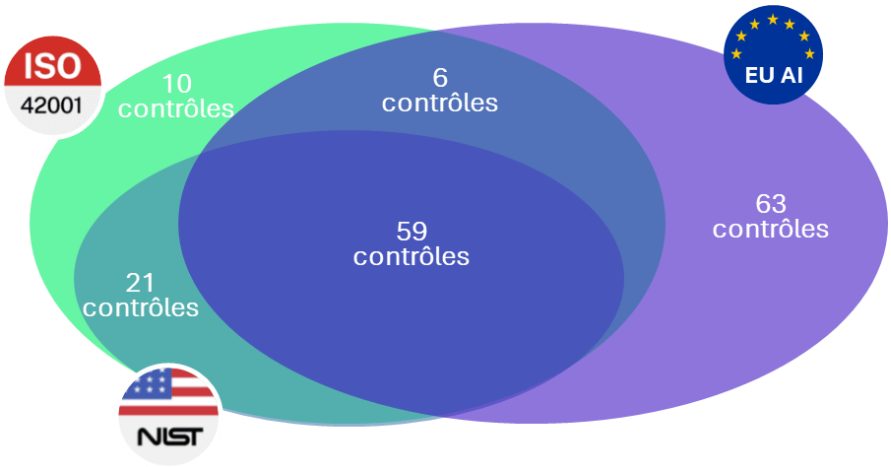

Figure 2 – Recoupement de contrôles de conformité entre 3 frameworks - ISO 42001, EU AI Act et le NIST RMF 1.0 (Source : modulos.ai)

Ces travaux du Hub France IA se sont appuyés en priorité sur une approche européenne et notamment celle du GEHN IA, sur laquelle se base la réglementation européenne (AI Act ou RIA).

2.1.2 Les lignes directrices pour une IA de Confiance du GEHN IA[8]

Le texte du GEHN IA, publié en avril 2019 par la Commission Européenne, établit les Lignes directrices en matière d'éthique pour une IA digne de confiance. Ce document propose une vision globale et structurée de l'IA de confiance, articulée autour de trois caractéristiques fondamentales qui doivent être respectées pendant toutes les étapes du cycle de vie des SIA :

- **Une IA licite**, conforme à toutes les lois et réglementations en vigueur ;
- **Une IA éthique**, respectant les principes et les valeurs fondamentales de la société ;
- **Une IA robuste,** capable de maintenir un comportement fiable, cohérent et sécurisé face à des perturbations.

Ces trois dimensions ne s'opposent pas, mais se renforcent mutuellement. Elles forment un cadre de référence permettant aux organisations de structurer leur démarche.

Pour assurer le respect de ses trois composantes, le GEHN IA identifie sept principes éthiques clés pour une IA de confiance :

1. **Action humaine et contrôle humain** : comprend les droits fondamentaux, l'action humaine et le contrôle humain ;
2. **Robustesse technique et sécurité** : comprend la résilience aux attaques et la sécurité, les plans de secours et la sécurité générale, la précision, la fiabilité et la reproductibilité ;

[8] LIGNES DIRECTRICES EN MATIERE D'ETHIQUE POUR UNE IA DIGNE DE CONFIANCE, GEHN IA, 2019.

3. **Respect de la vie privée et gouvernance des données** : comprend le respect de la vie privée, la qualité et l'intégrité des données, et le contrôle de l'accès aux données ;
4. **Transparence** : comprend la traçabilité, l'explicabilité et la communication ;
5. **Diversité, non-discrimination et équité** : comprend l'absence de biais injustes, l'accessibilité et la conception universelle, et la participation des parties prenantes ;
6. **Bien-être sociétal et environnemental** : comprend la durabilité et le respect de l'environnement, l'impact social, la société et la démocratie ;
7. **Responsabilité** : comprend l'auditabilité, la réduction au minimum des incidences négatives et la communication à leur sujet, les arbitrages et les recours.

Au-delà des principes, le GEHN IA a également développé une méthodologie d'évaluation[9] (listes d'exigences pratiques et de questions pragmatiques) que les organisations peuvent utiliser pour évaluer le niveau de confiance de leurs SIA.

Ce cadre sert aujourd'hui de fondement à de nombreuses initiatives nationales et européennes, notamment dans le cadre de l'implémentation de l'*AI Act* (ou RIA).

2.1.3 Des principes à adapter au contexte

Les trois caractéristiques de l'IA de confiance selon le GEHN IA, **licite, éthique et robuste**, ainsi que le cadre assurant la réalisation d'une IA de confiance, peuvent être abordés de manière très différente d'une organisation à l'autre, en fonction de sa culture, son utilisation de l'IA et de son mode de gouvernance. Elles se moduleront avec le temps en fonction de l'évolution des technologies, de la réglementation et de la littératie initiale des collaborateurs et de son écosystème.

En effet, la conception de la confiance n'est pas la même selon l'usage qui est fait de l'IA. Certaines organisations sont simplement **consommatrices** de systèmes (elles mettent en œuvre des IA construites par d'autres sur leurs périmètres), pendant que d'autres sont **créatrices ou intégratrices** de briques d'IA (elles construisent intégralement, ou assemblent des briques construites par d'autres pour faire des applications internes). L'article 3 de l'*AI Act*[10] définit ainsi les notions de fournisseur (« *Provider* ») et de déployeur (« *Deployer* »).

Il en va de même pour la culture et le fonctionnement (mode de gouvernance) de l'organisation qui peut être plus ou moins sensible aux notions de sûreté/sécurité permettant d'intégrer plus ou moins facilement les enjeux d'IA de confiance.

La confiance se construit, se nourrit et peut s'éroder ou se renforcer avec le temps. Les outils et processus doivent s'inscrire dans une **logique de gouvernance vivante**. La

[9] AI HLEG, 2020, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment
[10] Règlement sur l'intelligence artificielle, article 3

confiance se cultive, se mesure et s'ajuste aux évolutions technologiques, réglementaires, sociétales, environnementales et à l'utilisation des systèmes.

Il sera ainsi essentiel que les décisions et orientations stratégiques prises par l'organisation concernant l'IA de confiance soient fréquemment revues et adaptées. Ces adaptations devront se nourrir non seulement des évolutions technologiques, juridiques, et normatives, mais également de la manière dont l'IA est adoptée dans l'organisation et des impacts qu'elle a déjà sur l'organisation et ses parties prenantes internes et externes.

Ainsi, chaque organisation doit commencer par définir ses grands principes sous-tendant sa définition propre de l'IA de confiance notamment éthiques et réglementaires.

2.2 Clarifier les Rôles et responsabilités

La mise en œuvre de ces principes va réunir une multitude de parties-prenantes qu'il va falloir animer via une comitologie claire et sponsorisée à haut niveau :

- **Les équipes opérationnelles** : Data (développement), IT (déploiement), Métier (cas d'usage), Cyber et Compliance (réglementation)
- **Le soutien de niveau COMEX** : en général pris en charge par le *Chief Data Officer* ou *Chief AI Officer* rattaché au plus haut niveau de l'organisation

2.2.1 Le besoin de créer des nouveaux rôles

Aujourd'hui, les initiatives autour de l'IA de confiance sont souvent initiées et pilotées par les équipes techniques ou conformité (DSI, RSSI, DPO, direction juridique), qui intègrent bien les enjeux associés (qualité des données, robustesse, sécurité ou respect des règlementation). Si cette approche garantit un certain niveau de maîtrise, elle laisse de côté d'autres dimensions fondamentales :

- Les **enjeux éthiques** (valeurs, équité, biais, inclusion)
- Les **impacts sociaux** (acceptabilité, transformation des métiers, fracture numérique)
- L'**adoption utilisateur**, levier clé de la confiance.

En effet, là où les systèmes d'information étaient historiquement encadrés par des architectures centralisées, des feuilles de route rationalisées et des mécanismes de contrôle bien identifiés, l'IA vient percuter ces repères par sa complexité technique, ses impacts multidimensionnels et ses évolutions constantes. La gouvernance doit donc permettre de croiser les regards et d'aligner les parties prenantes, pour une approche plus holistique du sujet.

Il est donc nécessaire de créer des nouveaux rôles autour de l'IA de confiance. Cela peut se traduire dans une organisation centralisée par les **référents IA de confiance**, garants de la création et de la bonne application du référentiel de l'organisation. Celui-ci peut notamment prendre en charge la veille réglementaire et technologique afin de pouvoir maintenant le référentiel. Dans des organisations décentralisées, il s'agira des « ***Trustworthy AI champions*** », relais au sein des différentes équipes.

Au-delà de ces nouveaux rôles, il est crucial de s'assurer que les métiers, les RH et la RSE soient intégrés à l'ensemble des étapes des projets de SIA. Cela va se faire via la création d'une gouvernance autour du **comité d'IA de confiance** qui pourra être supervisée par le DRH, garant de la transformation culturelle de l'organisation.

2.2.2 La création du comité d'IA de confiance

Cette instance regroupe en général les parties prenantes au sujet IA de confiance dans l'organisation ainsi que des experts externes (scientifiques, universitaires, juristes, philosophes). Son rôle est modulable, mais peut prendre la forme suivante :

- Orienter les décisions stratégiques qui sont prises sur l'utilisation et le développement (le cas échéant) des IA (**fonction stratégique**) ;
- Valider que les lignes directrices de la direction générales respectent bien les décisions stratégiques prises (**fonction de contrôle**) ;
- Aider à statuer sur les orientations prises pour certains cas d'usage (**fonction d'arbitre**, notamment sur les aspects éthiques).

Ce comité permet de réinsérer le traitement des aspects de l'IA de confiance qui ne sont que rarement pris en compte dans l'organisation traditionnelle (éthique, justesse de traitement, environnement) en ciblant à la fois la gestion des risques mais aussi l'accompagnement des projets en tant qu'opportunités de développement du référentiel interne. Il couvre les activités suivantes :

- **Développe** et **maintient à jour** des **politiques d'IA de confiance** et de **procédures associées** (ex : Data & IA Ethique, évaluation de SIA, procédures d'explicabilité et de contestabilité…) ;
- **Développe**, **opérationnalise** et **supervise** l'application du **référentiel d'IA de Confiance** (Gouvernance, Gestion des risques, Industrialisation, Culture et sensibilisation) ;
- Assure une **veille réglementaire**, évalue l'évolution des exigences, s'interface avec l'écosystème externe (ex : partenaires technologiques, législateurs, monde académique…) pour saisir les bonnes pratiques et les intégrer ;
- Anime un **guichet unique d'aide** sur les questions d'IA de confiance ;
- Peut éventuellement coordonner des **groupes de travail** traitant de l'IA de confiance ;
- Maintient à jour un registre des SIA à l'échelle de l'organisation.

Ce comité a vocation à s'intégrer dans les **"parcours projets",** où ils incluent des exigences sur l'IA de confiance en s'assurant que chaque projet ait à consulter les différents pôles pertinents (sécurité, data, *privacy*, conformité, RSE) sur les thèmes de l'IA de confiance qui sont sous leur responsabilité. L'adaptation de sa méthodologie projet en ce sens, quand l'organisation en a déjà un, facilitera l'adoption. Il pourra notamment intervenir dans différentes instances :

- Un **conseil IA de confiance**, au besoin : regroupant l'exécutif et le comité IA de confiance, permet d'escalader à haut niveau (risque, arbitrage, alertes, incidents) ;
- Un **comité IA de confiance**, bimestriel : rassemblant le comité IA de confiance, DPO, Conformité, Risque, Juridique, IT, Cybersécurité, présente les avancées des projets, valide les prochaines étapes dans la méthodologie IA de confiance, rapporte les éléments issus de la veille ;
- Un **comité projet IA de confiance**, mensuel : entre le comité IA de confiance et les référents, coordonne les référents, guide et assiste dans l'application du référentiel et en contrôle sa bonne application ;

- Un **suivi projet AI**, bimensuel : entre les métiers et les référents, applique la méthodologie.

Certaines organisations[11] [12] intègrent également une Convention salariée sur l'IA (en particulier générative) avec des salariés tirés au sort pour fournir des recommandations à la direction.

2.2.3 Avec un soutien au niveau du COMEX

Enfin, il est partagé qu'une implication forte du COMEX sera nécessaire pour :

- **Superviser les impacts** de la technologie à tous les niveaux de l'organisation ;
- **Arbitrer** sur la gouvernance et les décisions politiques voire stratégiques ;
- **Valider une stratégie** claire, en lien avec le comité Ethique de l'IA.

En revanche, sur la déclinaison de la responsabilité sur l'IA de confiance au sein de l'organisation, il n'y a pas de réponse unique. Plusieurs facteurs vont jouer :

1. **Le secteur d'activité :** par exemple, en finance, la direction conformité peut naturellement porter le sujet tandis que dans la santé, ce sont souvent les directions qualité et affaires réglementaires (QARA) ;
2. **Le niveau d'ambition** : certaines organisations veulent faire de l'IA de confiance un levier différenciateur sur le marché, ce qui implique une gouvernance renforcée.

Pour un soutien de la direction efficace, les dirigeants (sponsors) doivent eux-mêmes se sensibiliser et se former à l'IA de Confiance pour **légitimer les actions, les ressources à mobiliser, et l'embarquement de multiples fonctions**. En particulier, outre la transparence des fournisseurs, la prise en compte des impacts environnementaux aux différentes étapes du cycle de vie du projet IA repose sur un soutien à un certain niveau de Direction qui soit à la fois conscient de l'importance des enjeux et moteur pour la poursuite de la sobriété numérique dans un contexte de croissance des usages de l'IA.

2.3 Comment déployer cette gouvernance à l'échelle ?

Évidemment, il n'existe pas de modèle unique de gouvernance "clé en main" applicable à toutes les organisations. Chaque organisation ou institution adapte sa démarche en fonction de ses priorités, de son secteur, de sa maturité technologique, ou encore de son rôle vis-à-vis de l'IA (développeur, déployeur, ou les deux).

Le premier type de gouvernances observé sur le marché correspond à une **gouvernance décentralisée**. Dans ce modèle, la gouvernance IA s'appuie sur les

[11] MAIF, convention salariée sur l'IA.
[12] MACIF, manifeste éthique sur l'IA.

structures existantes, en **injectant les responsabilités IA au sein des différentes équipes** métier, IT, juridique, *etc*.

Ce modèle aura l'avantage de nécessiter moins de réorganisation et de s'intégrer plus facilement aux pratiques existantes, de responsabiliser les équipes locales tout en gardant une vue sur l'ensemble des traitements. A contrario, la montée en compétence sera plus lente, avec un risque d'incohérences entre les équipes ainsi qu'une moindre démultiplication par les retours d'expérience.

Le modèle de **gouvernance centralisée** avec une équipe dédiée (interne ou mixte) regroupant les compétences autour de l'IA de confiance agit comme centre de compétences et structure de pilotage transverse.

Cela accélère la montée en maturité d'une équipe d'experts, assure la cohérence de traitement et infuse une culture commune autour de l'IA de confiance. Il est, cependant, plus compliqué de s'insérer dans les processus métiers existants notamment si le lien avec le métier n'est pas maintenu et peut apporter un sentiment de manque d'efficacité si mal interfacer avec les gouvernances « locales. Les ressources de ce modèle doivent également être priorisées entre les différents projets.

La réponse peut donc se trouver dans des modèles hybrides, combinant pilotage stratégique centralisé et déploiement décentralisé. Les exemples suivants illustrent l'importance de tester, ajuster, et documenter les dispositifs mis en place.

Exemple 1 :

- Un **comité éthique stratégique** qui définit les grandes orientations, rassemblant notamment des membres du COMEX et des experts externes ;
- Un **pilotage du sujet** à quatre têtes : technique, juridique, éthique, métiers/utilisateurs (en lien avec la définition du GEHN) ;
- Un **comité opérationnel mensuel** mettant en œuvre les orientations dans les projets internes, suivis par les référents RGPD et IA de confiance, formés aux risques, de chaque branche et filiale.

Exemple 2 :

- La **définition des grandes orientations** en termes d'IA au niveau comité de direction / comité exécutif ;
- L'**arbitrage de cas complexes** par le comité éthique au niveau management ;
- Le **traitement de l'ensemble des cas d'usage** sur l'ensemble des aspects (impact social, sociétale, environnemental) par des groupes de travail opérationnels.

3. Implémenter un cadre de gestion des risques

3. Implémenter un cadre de gestion des risques

Afin de déployer des SIA respectant les sept principes clés du GEHN pour une IA de confiance, analyser les risques est fondamental pour anticiper et prendre en compte les questions éthiques pertinentes dans le contexte du projet, tout en échangeant avec les parties prenantes impliquées pendant l'ensemble des phases du cycle de vie.

En effet, comme nous l'avons vu précédemment (voir chapitre 1, **Pourquoi une IA de confiance ?**)

), les SIA présentent de nombreux risques qui peuvent survenir à chaque étape du cycle de vie des SIA. Voici quelques exemples :

ETAPE	EXEMPLES DE RISQUES
IDEATION	• Cas d'utilisation interdit par la réglementation. • Promettre des résultats irréalistes ou de sous-estimer les limites techniques et éthiques de l'IA • Biais issus de la formulation du problème, pouvant introduire des discriminations dès l'origine
CONCEPTION	• Choix inadapté des données : biaisées, non représentatives ou non conformes à la réglementation, de mauvaise qualité • Choix inadapté de la solution technique (trop complexe ou non robuste)
DEVELOPPEMENT	• Introduction de vulnérabilités exploitables (ex. : injection de code, accès non autorisé aux données, empoisonnement des données) • Mauvaise documentation ou incomplète
VERIFICATION & VALIDATION	• Tests insuffisants ou non représentatifs, masquant des défaillances en production. • Mauvaise définition des indicateurs de performance, rendant difficile l'évaluation de la pertinence ou de la robustesse du système
EXPLOITATION & SURVEILLANCE	• Dégradation des performances dans le temps • Réapprentissage sur des données corrompues ou non contrôlées, introduisant de nouveaux biais ou vulnérabilités • Absence de contrôle humain sur les décisions critiques, ce qui peut amplifier les erreurs ou empêcher la détection rapide d'incidents • Mauvaise utilisation du système

Tableau 2 : Exemples de risques pouvant survenir aux différentes étapes du cycle de vie d'un SIA

De plus, les risques peuvent résider sur différentes couches technologiques, par exemple :

COUCHE TECHNOLOGIQUE	EXEMPLES DE RISQUES
Flux, transformation et stockage des données	• Propagation de biais • Qualité des données • Sécurité • Fuite de données • Accès non autorisés
Infrastructure	• Violations de la confidentialité des données • Perte et corruption de données • Consommation énergétique élevée • Empreinte carbone significative • Menaces d'attaques réseau • Accès non autorisé
SIA	• Amplification des biais • Manque d'explicabilité • Dérive du modèle • Risque de qualité et d'intégrité des données • Problèmes de propriété intellectuelle • Consommation d'énergie élevée • Hallucination • Manipulation des prompts • Intégrité RAG • Biais dans la génération de contenu • Transparence algorithmique • Surveillance inadéquate • Surdépendance aux décisions

Tableau 3 : Exemples de risques pouvant résider sur les différentes couches technologiques

Il est ainsi nécessaire d'adopter une démarche de détection, qualification et mitigation des risques tout le long du cycle de vie des SIA, en prenant en compte l'ensemble des sources de risques possibles et des acteurs.

Le processus se fait généralement en 3 temps, de façon itérative à chaque étape du cycle de vie :

1. ***Risk screening* / préqualification** : l'objectif est d'identifier rapidement (avec un questionnaire court) le niveau de risques posés par le SIA vis-à-vis des régulations mais également des principes définis par l'organisation
2. ***Risk assessments* / qualification des risques (quanti / quali)** : l'objectif est d'évaluer de façon plus détaillée et exhaustive l'ensemble des risques associés au SIA et sur chaque étape de son cycle de vie.
3. **Plan de mitigation et contrôles** : l'objectif est de définir le plan de mitigation et de contrôles des risques identifiés, sur chaque étape du cycle de vie du SIA.

Mais avant de commencer ce processus, il convient de recenser les SIA présents dans l'organisation.

3.1 Recensement des SIA

3.1.1 Alimenter un registre des SIA

Les risques liés à l'intelligence artificielle doivent être pilotés sur les systèmes existants (déjà en production), ceux en cours de développement ou de déploiement, mais également dès l'idéation. Pour cela, il y a deux actions à mener de front :

1. **Réaliser l'inventaire de tous les SIA existants dans l'organisation**, qu'ils aient été développés en interne ou intégrés à des solutions du marché, qu'ils soient déjà en production ou en cours de déploiement.

   Ce travail de recensement devra se faire auprès de tous les métiers (RH, Marketing, ventes, Operations, Production, IT, Finance, Service Client, R&D, Administration, *etc.*) au moyen de questionnaires et entretiens ciblés avec les responsables de chaque équipe.

2. **Intégrer à la gouvernance** (définie au chapitre **Erreur ! Source du renvoi i ntrouvable.**) tous les nouveaux projets utilisant de l'IA.

Ces deux actions permettront d'alimenter un registre des SIA, point d'entrée pour le suivi global des risques. Les éléments essentiels à documenter dans ce registre incluent :

- **Les sources des données :** origines des jeux d'entraînement, méthodes de collecte et de nettoyage ;
- **Les architectures techniques :** *frameworks*, bibliothèques et infrastructures cloud utilisés ;
- **Les cas d'usage métier :** objectifs déclarés, bénéfices attendus et parties prenantes concernées ;
- **Le contexte de déploiement ou d'usage :** secteurs d'application, populations impactées et intégration avec d'autres systèmes ;
- **Analyse préalable des risques et de la conformité réglementaire** : statut vis-à-vis du Règlement européen sur l'intelligence artificielle (ex. : classification des risques) et audits tiers.

Le Hub France IA, dans le cadre du Groupe de Travail « Boussole de l'*AI Act* », a ainsi proposé un feuillet récapitulatif pour présenter comment identifier et filtrer les projets d'IA au sein de son organisation, et d'en pré-identifier les risques, pour alimenter un tel registre[13].

3.1.2 Ne pas oublier les fournisseurs tiers de SIA

Afin de réaliser un inventaire exhaustif des systèmes d'IA, il est essentiel de ne pas oublier les **systèmes fournis par des tiers** (éditeurs, prestataires, intégrateurs). En effet, ces solutions embarquent de plus en plus fréquemment des composants d'IA, parfois invisibles pour l'utilisateur final, et qui pourrait exposer l'entreprise à des **risques juridiques, éthiques et opérationnels**. Par ailleurs, ces fournisseurs doivent être en mesure de fournir les documents de conformité requis (classification, documentation technique, évaluation des risques, *etc.*).

3.1.3 Attention au « Shadow AI »

Ces travaux ont fait émerger un constat partagé par de nombreuses organisations : l'IA entre dans les projets de manière souvent informelle, partielle, voire clandestine. Ce phénomène, le « *Shadow AI* », fait écho à celui du « *shadow IT* » observé depuis quelques années. Il se manifeste souvent par :

- Une utilisation des collaborateurs d'outils généralistes du marché non autorisés par l'organisation ;
- Des modules / options IA ajoutés par des fournisseurs sans modification des contrats existants / analyse de risque préalable ;
- Des petits contrats SaaS, souscrits par les métiers en dehors des processus Achats de l'organisation ;
- Des expérimentations internes exploitées et/ou mises en production hors des radars de l'IT.

Ce « Shadow AI » pose plusieurs risques majeurs :

- **Perte de contrôle sur les données :** données sensibles, confidentielles ou personnelles pouvant être partagées avec des solutions non validées par l'organisation ;
- **Risques et menaces sécuritaires ;**
- **Usages inefficients, contre-productifs voire délétère :** certaines IA sont utilisées à contre-emploi (par exemple l'IA générative utilisée comme moteur de recherche), le même SaaS est déployé par plusieurs entités d'une organisation sans coordination ;
- **Problèmes de propriétés intellectuelles.**

[13] GT BAIA, *Définition d'un projet*, Hub France IA, 2025_03_Feuillet_Definition_Projet_AIAct_HFIA_final-1.pdf

Il est donc crucial pour les organisations d'identifier ces systèmes d'IA déployés de façon « *shadow* » et de les intégrer au registre des SIA. Ainsi, les organisations ne devront pas se reposer uniquement sur le département IT pour identifier les SIA déployés, mais bien interroger l'ensemble des équipes métiers.

3.1.4 Prérequis

Pour assurer le succès de ce travail de recensement des systèmes d'IA, il est nécessaire d'avoir au préalable :

- **Posé une définition claire de l'intelligence artificielle**
- **Sensibilisé les équipes sur les risques** posés les IA
- **Communiqué sur la gouvernance** et les différents chantiers en cours.

En effet, les membres de l'organisation collaboreront d'autant plus facilement qu'ils en comprennent les raisons, les tenants et aboutissants (voir le chapitre 4 sur la formation et la communication).

3.2 Préqualification (*Risk Screening*)

Tous les systèmes d'IA identifiés ne nécessiteront pas de lancer des évaluations détaillées des risques. Ces évaluations peuvent être couteuses en ressources pour les équipes juridiques, DPO, sécurité, et le nombre de SIA identifiés peut être très important. Il convient donc de préqualifier ces SIA pour identifier rapidement (avec un questionnaire court) le niveau de risques posés par les SIA vis-à-vis des réglementations mais également des principes définis par l'organisation.

L'organisation va donc devoir commencer par construire ce questionnaire de préqualification, qui aura pour objectifs de détecter, avec un nombre minimum de questions [14] :

- **Le niveau de risques selon la classification de l'*AI Act*[15]**
- **L'utilisation de données à caractères personnels et le niveau de risques pour les droits et libertés des personnes concernées** (Article 35 du RGPD) (un risque élevé demandera la mise en œuvre d'une analyse d'impact relative à la protection des données – DPIA)
- **Le niveau de risques liés à la cybersécurité**
- Toute autre classification de risques importants pour l'organisation – par exemple, l'impact environnemental ou encore l'impact opérationnel en cas de mal fonction

Le résultat de cette préqualification permettra :

[14] Sur ce questionnaire, la tendance se trouve entre 10 et 30 questions.
[15] Hub France IA, septembre 2024, Fiche de Définitions clés sur l'*AI Act*, Livre blanc « A toolbox for managining risks of AI systems »

1. **De pouvoir évaluer ces niveaux de risques vis-à-vis du business case** – par exemple, en cas de niveaux de risques élevés et de bénéfices attendus relativement faibles, il sera probablement préférable de ne pas lancer le projet
2. **D'identifier dès le démarrage les actions de conformité** qui seront à mener, et notamment les évaluations détaillées de la part des fonctions expertes telles que juridique, sécurité, DPO

3.3 Identification et Evaluations des risques

Là où le questionnaire de préqualification vise à se faire une idée des niveaux de risques liés au cas d'usage, les évaluations des risques visent à être exhaustives dans l'identification et la qualification des risques. Plusieurs évaluations pourront être menées, et ce par différents acteurs :

- **Une analyse d'impact relative à la protection des données (AIPD)** par le DPO et son équipe,
- **Une analyse des risques en cybersécurité** par le SISO et son équipe,
- **Une analyse des risques sur les différentes étapes du cycle de vie du SIA et sur les différentes couches technologiques et organisationnelles**, par les équipes projet métiers et techniques.

Pour réaliser cette dernière (c.), l'organisation devra mettre en place une méthodologie qui doit :

- **S'adapter au positionnement de l'organisation vis-à-vis du SIA évalué** : simple utilisateur, intégrateur de modules IA dans ses produits, créateur de SIA ou de modèles ;
- **Tenir compte du contexte** dans lequel sera utilisé le système ainsi que son usage prévu ;
- **Permettre d'identifier les potentielles conséquences** pour l'organisation, les individus ou la société si le risque devait se matérialiser ; ainsi que la probabilité que celui-ci se matérialise ;
- **Permettre une évaluation qualitative et quantitative des risques** – cette dernière se faisant au moyen d'un ensemble d'outils qu'il faudra sélectionner, installer sur la plateforme hébergeant le SIA, et paramétrer ;
- **Aboutir à une priorisation** des risques pour traitement.

Pour cela, l'organisation pourra s'inspirer des ressources suivantes :

- Le questionnaire ALTAI[16] définit par le GEHN IA

[16] Commission Européenne, Juillet 2020, Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

- La bibliothèque de risques constituée par le MIT[17] que vous avons évoqué précédemment (voir Tableau 1 : Domaines et sous-domaines de risques selon le MIT)
- Les domaines de risques présentés dans le rapport « *AI Safety Report* » rédigé dans le cadre de l'*AI Action Summit* de Paris (Janvier 2025)
- Le catalogue d'outils et de métriques pour une IA de confiance[18], mis à disposition par l'OCDE.

Pour les SIA achetés à des tiers, il convient également d'ajouter les évaluations indépendantes des fournisseurs (biais, sécurité) et de s'assurer de la présence de clauses contractuelles sur la transparence des modèles, le traitement des données et la maintenance.

Une fois les différentes analyses menées et les risques identifiés et évalués, il faut déterminer les plans d'actions afin de supprimer ou mitiger ces risques.

3.4 Plan d'actions de mitigations des risques et contrôles

A partir des risques identifiés, évalués et priorisés lors de la phase d'évaluations des risques, l'organisation devra déterminer les actions à mener pour supprimer ou mitiger les risques, ainsi que les actions de contrôles à mettre en œuvre.

Afin d'accompagner les équipes en charge de définir le plan de mitigations, l'organisation pourra se constituer une bibliothèque de contrôles, en s'inspirant par exemple de ceux fournis par le standard ISO/IEC 42001 ou le Framework AI RMF 1.0 du NIST.

3.5 Processus de gestion des risques

Comme nous l'avons vu précédemment, les risques peuvent apparaitre aux différentes étapes du cycle de vie d'un SIA (voir Tableau 2 : Exemples de risques pouvant survenir aux différentes étapes du cycle de vie d'un). Il est donc nécessaire de refaire les analyses décrites en 3.3 et de revoir les plans d'actions de mitigations (3.4) à chaque étape :

[17] MIT, MIT AI Risk Repository
[18] OECD, Catalogue of Tools and Metrics for Trustworthy AI

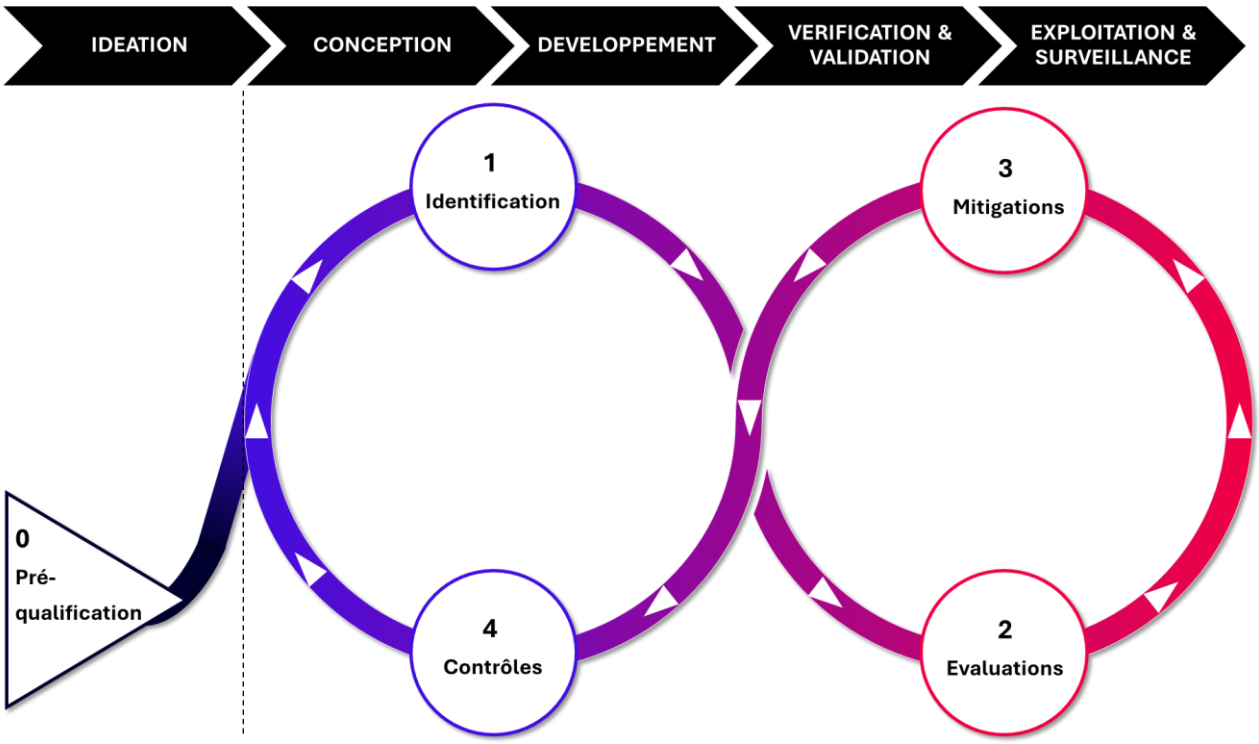

Figure 3 Cycle de vie des SIA et processus de gestion des risques

Ainsi, là où la préqualification interviendra seulement au moment de l'identification du projet ou du système, les phases d'identification et évaluation des risques, de mitigations et de contrôles devront se répéter à chacune des différentes étapes du cycle de vie.

4. Bâtir une culture autour de l'IA de confiance

4. Bâtir une culture autour de l'IA de confiance

4.1 Sensibiliser et former les collaborateurs est un impératif

4.1.1 Limiter les risques

Nous l'évoquions dans les chapitres précédents : **la sensibilisation aux risques posés par l'utilisation de systèmes IA auprès de l'ensemble des collaborateurs est fondamental** pour les organisations. Cela permet de limiter les risques posés par l'utilisation d'outils non autorisés par l'organisation ou par une mauvaise utilisation des outils autorisés, l'un comme l'autre pouvant entrainer :

- **Des problèmes de sécurité**, comme des fuites de données personnelles ou confidentielles, ou encore la création d'une faille de sécurité ;
- **Des problèmes opérationnels**, liés à une confiance excessive dans les résultats de l'outil (entrainant une absence de vérification ou d'esprit critique), ou à une utilisation hors du contexte dans lequel l'outil était prévu ;
- **Des coûts financiers et environnementaux**, lorsque des SIA coûteux en puissance de calculs sont sollicités à mauvais escient.

À cet égard, **l'article 4 de *l'AI Act* impose, depuis février 2025, que toutes les organisations qui déploient ou utilisent des systèmes d'IA doivent développer la littératie numérique autour de l'IA et de ses risques** auprès de leurs employés, afin de leur garantir une maîtrise suffisante.

Des programmes de formation doivent donc être déployées sur deux aspects essentiels :

- **La formation au fonctionnement et aux possibilités offertes par l'IA dans la réalisation du travail de vos équipes**. Cela doit permettre de s'assurer qu'elles exploitent l'IA à son plein potentiel, mais surtout qu'elles comprennent bien son intérêt pour elles, afin de garantir une utilisation avec discernement et confiance de l'IA.
- **La formation aux risques, aux limites et aux conséquences parfois inattendues** de l'utilisation de l'IA, en s'appuyant sur des exemples réels, pour qu'elles puissent détecter au plus vite les cas à risque et pour démystifier les craintes ou les promesses sans fondement autour de l'IA.

4.1.2 Limiter les fractures internes

Mais l'IA ne transforme pas uniquement des processus opérationnels, **elle reconfigure en profondeur les rapports au travail, à la décision, et à la connaissance**. Dans un contexte où les niveaux de maturité sur ces sujets sont hétérogènes, et où les équipes fonctionnent encore trop souvent en silos, **le risque est grand de voir émerger des fractures internes** : entre les métiers et la tech, les experts et les opérationnels, les

jeunes générations et les plus expérimentées, les enthousiastes et les inquiets. Créer les conditions d'une adoption généralisée suppose donc de bâtir un socle de compréhension commun à tous les niveaux de l'organisation. Cela passe par une communication et une acculturation progressive et accessible, qui outille chacun selon ses besoins mais aligne l'ensemble autour de principes partagés — transparence, responsabilité, éthique, sécurité. C'est à ce prix que les organisations peuvent instaurer une véritable culture de la confiance, capable de résister aux incertitudes et d'accompagner durablement le changement.

4.1.3 Développer les compétences techniques

Pour que cette culture de la confiance se déploie, il est essentiel de s'attaquer à un défi de taille : les compétences techniques. **Les profils véritablement capables de maîtriser l'ensemble des dimensions de l'IA de confiance — qu'il s'agisse de sécurité, d'éthique, d'explicabilité ou de conformité — restent aujourd'hui extrêmement rares**. Le savoir est souvent dispersé entre silos techniques, métiers ou profils réglementaires, freinant la création d'une vision globale et cohérente. Faute de coordination et de responsabilisation claire, ces manques nourrissent des tensions organisationnelles : les rôles se chevauchent, les décisions tardent, et la gouvernance devient incertaine. Les enjeux autour de la compétence sont donc multiples : d'une part, concernant leur exhaustivité, car toutes les briques de la confiance (sécurité, transparence, éthique, conformité, robustesse, *etc.*) doivent être couvertes ; d'autre part, la mise à jour des connaissances, car la technologie IA évolue rapidement, exigeant une veille continue et une montée en compétence agile.

4.1.4 Rassurer ou cadrer les collaborateurs

Mais au-delà des compétences techniques, il existe également un enjeu culturel et symbolique majeur. **L'IA — et tout particulièrement l'IA générative — suscite des perceptions ambivalentes** : certains collaborateurs y voient une menace pour leur emploi, d'autres un outil d'émancipation professionnelle ou d'augmentation de leurs capacités. Ce fossé de perception traduit une hétérogénéité dans les niveaux de maturité et une anxiété liée à l'inconnu, inhérente à toute introduction d'une nouvelle technologie. Il en résulte des comportements contrastés, allant du rejet pur et simple à l'usage de « *shadow AI* » discuté plus haut. Seule une communication et une acculturation larges, structurées et adaptées aux réalités du terrain permettra de transformer ces représentations floues ou anxiogènes en leviers d'appropriation et de discernement.

4.1.5 Un rôle central pour les DRH

Dans cette dynamique de transformation, les Directions des Ressources Humaines jouent un rôle central. En tant que garantes de la transformation culturelle, elles sont en première ligne pour **structurer des dispositifs d'apprentissage continu, faire évoluer les postures, et soutenir l'alignement stratégique des équipes**. Leur mission dépasse

la seule formation technique : il s'agit aussi de prévenir les usages à risque (notamment via la sensibilisation aux bonnes pratiques de sécurité), d'améliorer l'expérience collaborateur (en donnant des repères clairs sur l'usage de l'IA au quotidien), de rassurer face aux craintes de remplacement (en promouvant une logique de complémentarité humain-machine), ou encore de retisser un lien intergénérationnel autour d'un usage collectif et responsable de l'IA. En développant une culture commune autour de l'IA, les DRH peuvent faire émerger un état d'esprit tourné vers l'apprentissage, l'innovation et la coopération — conditions sine qua non d'une confiance active et durable.

C'est à cette double condition — former et faire comprendre — que l'organisation pourra pleinement embarquer ses collaborateurs dans la transformation.

4.2 Communication et engagement autour de l'IA

Former sans communiquer, c'est prendre le risque d'instruire sans embarquer. Pour que la montée en compétences irrigue réellement la culture de l'organisation, elle doit s'accompagner d'un effort de communication structuré, régulier et incarné. Faire saisir les objectifs, valoriser les initiatives, partager les retours d'expérience, rendre visibles les réussites comme les écueils, autant d'actions qui contribuent à normaliser l'usage responsable de l'IA dans les pratiques quotidiennes.

4.2.1 Les chartes et politiques

Les premiers outils clés pour structurer la communication autour de l'IA de confiance sont la création de chartes et de politiques internes. Ces documents définissent les principes directeurs et les engagements de l'organisation en matière d'IA, établissant ainsi un cadre normatif essentiel pour guider les pratiques et assurer une gouvernance responsable. Toutefois, ce cadre doit être régulièrement mis à jour pour rester pertinent face à l'évolution rapide des technologies et des attentes sociétales.

La **politique d'IA d'une organisation** est un document stratégique qui met en œuvre les principes directeurs. Elle précise les actions concrètes, les processus internes et les responsabilités des différents acteurs (DSI, équipes juridiques, DRH) pour garantir une gestion responsable et éthique de l'IA. Cette politique est construite de manière collaborative pour répondre aux préoccupations de toutes les parties prenantes : dirigeants, employés, experts techniques, services juridiques, ressources humaines et clients. Elle doit être régulièrement révisée pour assurer son adéquation avec les évolutions légales et technologiques.

La **charte éthique de l'IA** formalise, quant à elle, les engagements spécifiques de l'organisation en matière de transparence, de sécurité, de respect des droits des utilisateurs et d'impact social, définissant les cas d'utilisation de l'IA, les objectifs visés et les limites des technologies déployées. Elle garantit une utilisation responsable et conforme aux normes éthiques strictes, souvent inspirée de cadres internationaux

comme le GEHN IA précédemment cité, tout en étant adaptée aux spécificités de l'organisation. La sémantique peut varier pour ce genre de documents, marquant la volonté de certaines organisations de se distinguer par un engagement plus ou moins incarné, au travers de « Manifeste » par exemple. Ces chartes ou autres manifestes marquent une volonté de poser un cadre structurant et évolutif, intégrant des valeurs spécifiques comme la solidarité, l'équité ou l'engagement sociétal. En tant que garant de la culture d'une organisation et donc de ses pratiques de travail, le Groupe de travail considère que ce sont aux DRH d'être en figure de proue de cet exercice, aux côtés évidemment des DSI et de la Direction juridique/conformité. Certaines organisations vont même jusqu'à encourager leurs collaborateurs à prendre part, voire proposer, leurs propres recommandations sur le sujet[19].

Quelle que soit la terminologie, la mise en place de ces chartes et politiques doit être accompagnée d'une communication régulière et incarnée. Ces documents ne doivent pas rester théoriques, mais être appliqués concrètement. La communication devient un levier pour instaurer une culture partagée où l'IA est utilisée de manière responsable, transparente et éthique. La DRH et les responsables de la communication interne doivent s'assurer de leur diffusion, compréhension et incarnation à tous les niveaux. En impliquant les parties prenantes, y compris les managers, les experts techniques et les équipes RH, l'organisation renforce la confiance et favorise un usage respectueux et sécurisé de l'IA.

4.2.2 La nécessaire communication entre équipe pour briser les silos

À mesure que l'intelligence artificielle évolue, une collaboration efficace tout au long des cycles de vie des projets demeure un défi majeur pour les équipes d'IA. A ce titre, 20 % des leaders dans ce domaine considèrent la collaboration comme leur besoin le plus important non satisfait[20], soulignant ainsi que construire des équipes d'IA cohérentes est tout aussi essentiel que développer l'IA elle-même.

La collaboration en IA est mise à mal par les silos d'équipes, les environnements de travail en mutation, les objectifs mal alignés et les demandes commerciales croissantes.

Pour les équipes intervenant sur la conception de SIA, ces défis se manifestent dans quatre domaines clés :

- **Fragmentation** : éviter les outils, workflows et processus disjoints rendant difficile pour les équipes de fonctionner comme une unité cohérente ;

[19] MAIF, Novembre 2024, Propositions de la Convention salariée sur l'intelligence artificielle générative
[20] DataRobot, *The Unmet AI Needs Survey*, 22 octobre 2024: https://www.datarobot.com/newsroom/press/survey-reveals-only-34-of-ai-professionals-feel-fully-equipped-to-meet-business-goals

- **Complexité de coordination** : aligner les équipes transversales sur les priorités, les délais et les dépendances devient d'autant plus difficile à mesure que le nombre de projets s'accroit ;
- **Communication incohérente** : prévenir les lacunes dans la communication qui peuvent entrainer des opportunités manquées, des redondances, des retouches et des confusions sur l'état des projets et les responsabilités ;
- **Intégrité des modèles :** garantir la précision, l'équité et la sécurité des modèles nécessite des transferts fluides et une surveillance constante, mais les équipes déconnectées manquent souvent de la responsabilité partagée ou des outils d'observabilité nécessaires pour la maintenir.

Au-delà des difficultés propres aux équipes techniques, l'incompréhension entre les décideurs (COMEX ou autres comités éthiques, stratégiques, *etc.*) et les équipes techniques (développeurs, *data scientist*, *product owners*, *etc.*) demeure un problème majeur.

Comment faire en sorte que les équipes techniques expliquent de manière compréhensive les enjeux, opportunités, limites ou risques techniques d'un projet ou d'un outil IA aux décideurs ? Comment aligner les développements sur les priorités des décideurs ? Comment apporter le bon niveau de compréhension technique à un décideur pour qu'il prenne les bonnes décisions et saisisse tous les risques des IA développées par ses équipes ? Autant de questions que seules une communication efficace et un niveau de transparence maximal au sein d'une organisation pourront régler.

Pour réduire les silos organisationnels, il faut parler la même langue et utiliser les mêmes référentiels. Cela peut se traduire par :

- La mise en place d'outils visuels : *dashboards* unifiés montrant l'état des modèles, risques et métriques métier
- Des formations croisées : ateliers techniques pour les décideurs et sessions stratégiques pour les *data scientists*
- De processus hybrides : intégrer des représentants métier dans les comités d'éthique IA et vice versa
- Une documentation adaptative : versions simplifiées pour les dirigeants (résultats, risques) et techniques (architecture, tests)

4.3 Les 5 règles à retenir

Les 5 règles à retenir pour accompagner la montée en compétence sur l'IA de confiance.

Dans la continuité de l'instauration d'une culture de confiance et de la mise en place de chartes et de politiques claires, la montée en compétences sur l'IA devient un élément clé pour garantir l'efficacité et la durabilité de cette transformation. Pour que

l'adoption de n'importe quel SIA se fasse de manière durable, un plan structuré de montée en compétences est indispensable.

Il doit couvrir plusieurs niveaux : d'abord, une sensibilisation aux risques liés à l'IA, afin de responsabiliser les créateurs et les utilisateurs face aux enjeux éthiques, sécuritaires et sociaux, tout en intégrant des actions concrètes pour éviter les dérives. Ensuite, il est crucial de faire monter en compétences les différents acteurs de la chaîne de création et de déploiement des IA — des équipes techniques aux experts en gestion des risques, en passant par ceux en charge de la méthodologie — pour assurer une maîtrise complète des défis associés à l'IA, du développement à l'exploitation. Enfin, une attention particulière doit être portée à la sensibilisation et à la formation du top management. Leur soutien est primordial pour assurer la cohérence de la démarche et l'intégration des valeurs de l'IA de confiance dans les décisions stratégiques. L'objectif est de faire de chaque acteur un maillon fort de la chaîne de gouvernance de l'IA, en leur fournissant les connaissances nécessaires pour prendre des décisions éclairées, dans le respect des principes éthiques et des exigences de sécurité.

Pour accompagner cette montée en compétences, nous proposons 5 règles concrètes pour démarrer de façon pragmatique. Au-delà de ces 5 règles, vous pouvez également vous inspirer du Bureau européen de l'IA, récemment créé, qui a publié un référentiel vivant de pratiques de montée en compétences[21], sur la base de retours d'expériences de plusieurs organisations.

4.3.1 Règle n°1 : adopter une approche différenciée selon l'exposition à l'IA

L'impact de l'IA est hétérogène en fonction des secteurs d'activités, des emplois et des tâches à accomplir. Certains métiers seront naturellement « augmentés », tandis que d'autres subiront de profondes mutations, voire disparaîtront.

Au-delà d'un socle minimal transverse de connaissances à transmettre (réglementation, fonctionnement, *etc.*) à tous les collaborateurs, une approche différenciée par métier doit également être prise. Pour répondre à ces enjeux, il est essentiel d'élaborer une analyse d'impacts vis-à-vis des emplois selon leur degré d'exposition à l'IA et ainsi adopter une stratégie différenciée :

- Métiers peu exposés : sensibilisation aux principes fondamentaux de l'IA pour développer une culture générale et prévenir les usages inadaptés ;
- Métiers augmentés : formation approfondie sur l'intégration des SIA dans les processus de travail afin d'améliorer la productivité et l'efficacité ;
- Métiers en transformation : requalification et acquisition de nouvelles compétences pour s'adapter aux évolutions du poste ;

[21] EU AI Office, Février 2025, Living repository to foster learning and exchange on AI literacy

- Métiers émergents : création de formations dédiées pour répondre aux nouveaux besoins.

4.3.2 Règle n°2 : augmenter la fréquence d'apprentissage

L'accélération des avancées technologiques impose une révision complète des modalités d'apprentissage. Selon une étude de l'OCE, alors qu'en 1987, une compétence acquise était pertinente pendant 30 ans, elle n'est désormais plus valide que 2 ans selon l'Organisation Internationale du Travail. Le Forum Économique Mondial estime même que 44% des compétences actuelles seront obsolètes d'ici cinq ans. Face à cette obsolescence accélérée des savoir-faire, il est indispensable de :

1. **Instaurer une formation en continu** plutôt que des formations ponctuelles, permettant aux collaborateurs d'évoluer avec la technologie plutôt que de subir ses transformations ;
2. **Miser sur des formats agiles** via des modules courts, du *micro-learning*, des simulations interactives et des mises en situation concrètes ;
3. **Intégrer l'IA directement dans l'ingénierie pédagogique** : utiliser l'IA pour personnaliser les parcours d'apprentissage en fonction des besoins et des compétences de chaque collaborateur.

Cette approche garantit une adaptation proactive aux mutations du marché, tout en renforçant la confiance des utilisateurs dans ces outils.

4.3.3 Règle n°3 : renforcer et prioriser l'apprentissage des compétences cognitives

L'IA ne possède ni intuition, ni créativité, ni intelligence émotionnelle. Ces compétences humaines deviennent donc un atout majeur à cultiver, d'autant plus dans un contexte où l'on est plus susceptible d'être remplacé par un collègue qui sait utiliser l'IA que par une IA elle-même.

Une IA de confiance passe ainsi par le fait de savoir interagir avec elle au travers de nos aptitudes cognitives humaines. Pour conserver cette valeur ajoutée et garantir une utilisation éthique et pertinente de ces systèmes, il est essentiel de renforcer des compétences transversales comme :

1. **L'esprit critique et le discernement** : savoir analyser les résultats fournis par l'IA et ne pas les prendre pour des vérités absolues ;
2. **La créativité et l'innovation** : imaginer des usages inédits de l'IA pour créer de la valeur ;
3. **L'intelligence sociale et émotionnelle** : maintenir la qualité des interactions humaines dans un monde de plus en plus automatisé ;
4. **Le sens éthique et la responsabilité** : comprendre les implications des décisions assistées par IA et garantir une utilisation alignée avec les valeurs de l'organisation.

Ces compétences permettent aux collaborateurs de travailler en complémentarité avec l'IA et non en opposition, instaurant ainsi une relation de confiance et de maîtrise des outils.

4.3.4 Règle n°4 : soigner vos experts en les formant à la confiance et d'éthique

L'IA de confiance repose en grande partie sur ceux qui conçoivent les systèmes. Son développement et son intégration dans les organisations doivent être pensés en amont par des experts capables de concevoir des systèmes fiables, durables et éthiques.

Il est donc primordial de **former des rôles clés au sein des organisations** (*data scientists*, *data engineer*, ingénieurs IA, designers UX, *etc.*) à des notions essentielles :

- **Écoconception et impact environnemental** : développer des IA plus sobres en énergie et limiter l'empreinte carbone des modèles d'apprentissage ;
- **Réduction des biais algorithmiques** : sensibiliser aux biais présents dans les données et aux bonnes pratiques de conception pour garantir des décisions justes et inclusives ;
- **Transparence et explicabilité** : construire des systèmes intelligibles pour les utilisateurs finaux afin de faciliter leur adoption et renforcer la confiance.

4.3.5 Règle n°5 : cultiver l'humilité face à une technologie encore en construction

L'IA est encore en développement, et il est essentiel que les organisations fassent preuve d'humilité. Les erreurs font partie du processus d'adoption, et tout ne peut pas être accompli d'un coup. L'IA évolue rapidement, et chaque organisation doit rester flexible et prête à ajuster ses stratégies face aux nouvelles technologies.

Il est crucial de reconnaître les limites actuelles de l'IA, notamment en matière de compréhension contextuelle, d'empathie et de créativité. Une IA de confiance se construit sur la reconnaissance de ces limitations, en évitant des attentes irréalistes.

L'humilité permet d'accepter les imperfections de l'IA et d'adopter une approche proactive pour résoudre les problèmes. Cela favorise l'amélioration continue des SIA, les rendant ainsi plus fiables et éthiques dans leur déploiement.

5. Comment commencer ?

5. Comment commencer ?

Le déploiement d'une IA de confiance au sein d'une organisation nécessite une approche structurée et réfléchie.

5.1 Evaluez la situation actuelle

La première étape consiste à réaliser les états des lieux suivants :

- **La gouvernance de l'IA** (rôles et responsabilités, comitologie, modèle opérationnel),
- **Les méthodologies et bonnes pratiques** suivies par les équipes tout le long du cycle de vie des SIA,
- **Le niveau de littératie en IA** de l'ensemble des collaborateurs et des acteurs des projets basés sur l'Intelligence artificielle.

Là où les deux premiers états des lieux consistent à interroger les acteurs liés à l'achat, au développement ou au déploiement des SIA dans l'organisation, évaluer le niveau de littératie en IA requiert d'interroger l'ensemble des collaborateurs. Pour ce faire, nous recommandons de créer un formulaire en ligne avec des questions de deux typologies différentes :

- **Des questions mesurant le ressenti des personnes**, comme « *De 1 à 10, comment évalueriez-vous votre niveau de compréhension de l'Intelligence artificielle ?* »
- **Des questions « cas pratiques », pour mesurer les réelles connaissances**, comme « *Parmi les outils suivants, lesquels contiennent de l'IA ?* »

De nombreux exemples de questionnaires existent dans la littérature scientifique, comme les questionnaires MAILS, AILS, SNAIL or SAIL4ALL.

Ces états des lieux vous permettront ensuite d'identifier les écarts par rapport à vos principes directeurs et aux obligations de l'*AI Act* pour les prioriser puis décliner en chantiers alimentant votre feuille de route.

5.2 Posez la gouvernance et définir les principes

Il est ensuite nécessaire de mettre en place une gouvernance efficace, à plusieurs niveaux, pour piloter vos risques de manière proactive. Cela implique de :

- Établir des **structures de gouvernance claires pour la surveillance de l'IA** (p. ex., un conseil de gouvernance de l'IA)
- Définir et attribuer **les rôles et responsabilités concernant la législation européenne sur l'IA et les pratiques d'IA responsable**, tels que : Responsable de l'IA, Responsable/Fonction de conformité de l'IA, Propriétaires du modèle, Responsables des risques, sans oublier de :
    - **Clarifier la responsabilité** des équipes commerciales, technologiques, juridiques et de conformité

    - **Assurer la séparation des tâches** (p. ex., entre les développeurs et les évaluateurs)
- Partager une **définition claire de l'IA**
- Mettre à jour ou créer des **politiques spécifiques à l'IA** intégrant les obligations réglementaires sur l'IA (par exemple, qualité des données, documentation, surveillance humaine) et définissant les principes d'IA responsable pour l'organisation
- **Définir et partager des procédures d'évaluation, de mitigation et de contrôles des risques**
- **Définir une stratégie de communication, sensibilisation et formation** à destination de l'ensemble des collaborateurs.

Une telle gouvernance garantit que l'utilisation de l'IA reste alignée avec les valeurs de l'organisation et les attentes de la société.

5.3 Inventoriez les SIA de l'organisation et évaluez leurs risques

Pour que la gouvernance soit efficace, encore faut-il maitriser ce qu'on doit gouverner : il est crucial de **démarrer au plus tôt l'inventaire des systèmes d'IA présents dans l'organisation** (voir chapitre 3.1).

Il convient donc de **définir dès à présent une démarche** pour collecter les informations nécessaires auprès de l'ensemble de l'organisation, et en prêtant attention aux SIA déployés en « *shadow AI* » ainsi que ceux intégrés à des outils tiers.

Cette démarche pourra intégrer **le questionnaire de préqualification des risques**. Ainsi, dès la collecte des informations pour réaliser l'inventaire des SIA dans votre organisation, vous pourrez rapidement vous faire une idée du volume de SIA à haut risque selon la classification de l'*AI Act*.

5.4 Sensibilisez et Formez vos équipes

Nous ne le répéterons pas assez : la sensibilisation aux risques et la formation à l'Intelligence artificielle est cruciale pour les organisations souhaitant construire des IA de confiance (voir chapitre 4). **La stratégie d'acculturation et de formation doit être posée dès maintenant**, en priorisant :

- **La définition de l'IA et des systèmes d'IA** selon l'organisation
- **Les risques liés à l'utilisation** des systèmes d'IA.

5.5 Intégrez pleinement l'IA de confiance dans le pilotage de vos projets

Intégrer les principes de l'IA de confiance dans le pilotage projet est une condition indispensable et une exigence réglementaire de l'*AI Act* au déploiement de l'usage de

l'IA dans une organisation. **Pour assurer un développement et un déploiement responsables de l'IA, il est crucial de mettre en place des processus rigoureux & systématiques** afin de maintenir un haut niveau de qualité et de fiabilité des SIA tout au long de leur cycle de vie :

- Intégrez dès maintenant à votre cycle de vie de développement des systèmes d'IA des critères de validation des projets reprenant les principes de l'IA de confiance que vous avez adoptés.
- Appuyez-vous sur les instances de contrôle et de validation de vos projets pour documenter tous les projets d'IA et les ajouter à votre inventaire de systèmes d'IA, avec les informations pertinentes (*cf*. 5.3).
- Assurez-vous que les porteurs de projet impliquant de l'intelligence artificielle dans votre organisation soient informés de leurs responsabilités dans l'implémentation de l'IA de confiance et que les équipes sur lesquelles ils peuvent s'appuyer (juridique, data science, sécurité, RSE, RH) aient des points de contact clairement identifiés.
- Documentez et tracez toutes les décisions et actions liées à l'IA, non seulement en interne, mais aussi vis-à-vis des actionnaires et des clients. Cette transparence renforce la confiance des parties prenantes, démontre l'engagement de l'organisation envers une utilisation éthique de l'IA, et facilite la conformité avec les réglementations émergentes telles que l'*AI Act* européen.

Une fois ce cadre essentiel posé et la première phase d'apprentissage passée, il est alors possible de passer à l'étape suivante : l'industrialisation et l'automatisation des tests et des contrôles, notamment sur le plan technique. En effet, la mise en place d'une IA de confiance à l'échelle ne peut se passer, pour être durable, de processus outillés, automatisés et standardisés, pour assurer un contrôle tout au long du cycle de vie.

6. Remerciements

Remerciements

Le Hub France IA remercie l'ensemble des participants au groupe de travail IAG, et tout particulièrement les contributeurs de ce livrable.

Les pilotes :

- **Thomas Argheria**, Wavestone
- **Gérôme Billois**, Wavestone
- **Martin D'Acremont**, Wavestone
- **Jordan Fleurier**, Hub France IA
- **Cyril Nicolotto**, Hub France IA

Les contributeurs :

- **Seif Benayed**, HEDI
- **Matthieu Camus**, Privacy Impact
- **Cécilia Damon**, ASNR
- **Hugo Enderlin**, Wavestone
- **Laurent Gardes**, SNCF
- **Anthony Guinot**, Wavestone
- **Alina Holcroft**, Ethiqais
- **Arnault Ioualalen**, Numalis
- **Isaure Ladonne**, Wavestone
- **Marie Langé**, Wavestone
- **Francesca Martini**, La Poste
- **Stanislas Renondin**, Giskard
- **Avi Suissa**, MAP - Monitoring And Protection

Validation :

- **Françoise Soulié-Fogelman**, Hub France IA
- **Bertrand Cassar**, La Poste

La touche finale :

- **Mélanie Arnould**, Hub France IA

En partenariat avec

WAVESTONE

NOTICE

PREMIERS PAS VERS L'IA DE CONFIANCE

Juin 2025

HUB FRANCE IA