L'Intelligence Artificielle

de confiance : Les fondamentaux

THALES
Building a future we can all trust

Edito

Dans un monde en constante évolution, la capacité d'une organisation à maîtriser l'intelligence artificielle (IA) de manière responsable est devenue un impératif sociétal.

L'intelligence artificielle transforme en profondeur les économies et les modes de vie, il est de notre responsabilité collective de veiller à ce que cette révolution technologique soit guidée par des principes éthiques, transparents et durables.

Au sein de Thales, nous portons haut les exigences d'un numérique responsable tout en accompagnant des clients dont les missions requièrent les plus hauts standards en matière de performance, de résilience et de sécurité. Nous plaçons l'IA au cœur de notre stratégie en concevant des solutions durables, souveraines et dignes de confiance.

Ce document présente une vision, des engagements et des compétences de pointe dans le domaine des services numériques intégrant des composants à base d'IA. Il invite à repenser l'intelligence artificielle comme une technologie au service de l'humain, de l'environnement et de l'indépendance numérique.

Au sein de Thales, l'IA de confiance est conçue à la fois comme un cadre et une boussole pour relever les défis d'aujourd'hui tout en créant une valeur pérenne et **digne de confiance pour demain.** En collaborant avec des experts, groupes de travail spécialisés et des partenaires nationaux comme européens, Thales se positionne comme un acteur clé de l'IA de confiance au service de l'humain. En s'engageant au-delà du simple cadre réglementaire, notre démarche d'IA de confiance est un levier pour améliorer les processus métiers en définissant une responsabilité partagée entre ses différentes parties prenantes tout en favorisant une performance accrue de nos solutions et services.

THALES
Building a future we can all trust

L'Intelligence Artificielle de confiance

Kalina Genet, Edouard Roger, Clément Engel

Juin 2026

Avant-propos

Naviguer dans l'écosystème de l'IA de confiance

L'Intelligence Artificielle (IA) transforme rapidement nos sociétés et nos économies. Alors que son déploiement s'intensifie, la nécessité d'une **IA de confiance** devient un enjeu majeur pour garantir un développement qui allie responsabilité, innovation et durabilité. Cet article vise à adresser l'écosystème complexe de l'IA de confiance en proposant une **grille de lecture de l'état de l'art** en France et en Europe puis en **illustrant une mise en œuvre possible**.

Cette grille de lecture a vocation à être **transverse et accessible**, combinant des sujets s'adressant à toute personne soucieuse d'intégrer des principes d'IA de confiance dans leurs stratégies et opérations, ainsi qu'à un large éventail de professionnels — des décideurs aux experts techniques en passant par des experts cyber, des Chiefs Data/AI Officers, des experts juridiques et des spécialistes en développement durable.

L'approche proposée se veut ouverte : pour chaque thématique abordée, des approfondissements ou des déclinaisons sectorielles existent au sein de divers travaux, initiatives ou publications spécialisées. Il est à noter que le panorama proposé est évolutif et n'a pas vocation à être exhaustif.

Le Système d'IA au cœur de l'approche responsable

Pour appréhender l'IA de confiance, il est essentiel de cadrer ce que nous entendons par "IA" dans ce position paper.

Nous adoptons ici la définition proposée par l'AI Act (article 3.1)[1], qui désigne un **système d'IA** comme « *un système automatisé qui est conçu pour fonctionner à différents niveaux d'autonomie et peut faire preuve d'une capacité d'adaptation après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des entrées qu'il reçoit, la manière de générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels* ».

Nous intégrons également dans notre approche **les modèles d'IA à usage général** tels que définis dans l'AI Act (article 3.63), en tant que « *modèle d'IA, y compris lorsque ce modèle d'IA est entraîné à l'aide d'un grand nombre de données utilisant l'auto-supervision à grande échelle, qui présente une généralité significative et est capable d'exécuter de manière compétente un large éventail de tâches distinctes, indépendamment de la manière dont le modèle est mis sur le marché, et qui peut être intégré dans une variété de systèmes ou d'applications en aval, à l'exception des modèles d'IA utilisés pour des activités de recherche, de développement ou de prototypage avant leur mise sur le marché.* »

[1] Union Européenne, Règlement sur l'intelligence artificielle (AI Act), « Regulation (EU) 2024/1689 » (2024)

Notre approche systémique a donc vocation à adresser le **cycle de vie complet d'un système d'IA**, qui associe un modèle à une finalité opérationnelle donnée, depuis sa conception jusqu'à son déploiement et sa maintenance. L'IA n'est pas une fin en soi mais un outil au service d'un besoin défini.

Penser les systèmes d'IA de manière responsable soulève des défis importants dans les stratégies d'entreprise, notamment la difficulté de lier explicitement la durabilité, la compétitivité, la souveraineté et la rentabilité.

Structure du document

Nous souhaitons mettre en lumière la diversité des expertises nécessaires à l'élaboration d'une IA de confiance (juridique, environnementale, cybersécurité, technique, etc.) et de relever les travaux de référence existants sur ce sujet complexe et transverse. Nous citons donc des travaux qui peuvent aussi bien traiter des fondements d'une gouvernance éthique de l'IA que des aspects pour développer des spécifications techniques qui atténuent les menaces découlant du déploiement de l'IA.

Trois aspects majeurs sont ici déclinés :

1. **Principes fondamentaux :** Cette section vise à définir l'IA de confiance et présenter en détail les quatre piliers majeurs qui la sous-tendent : la **validité,** la **sécurité**, la **transparence** et la **responsabilité**. Ces principes sont le fruit d'une analyse approfondie de la documentation et des cadres de référence disponibles sur le marché ainsi que des travaux Thales.

2. **Panorama réglementaire et travaux de référence :** Nous dressons un état des lieux non exhaustif des initiatives et régulations clés en matière d'IA de confiance, avec un focus particulier sur les cadres législatifs européens et français, ainsi que les organismes et acteurs influents dans ce domaine. Il est à noter que de nombreuses réglementations existantes, notamment celles liées aux données personnelles (comme le RGPD) ou à la cybersécurité (comme le Cyber Resilience Act) s'appliquent indirectement à l'IA.

3. **Illustration de l'application de l'IA de confiance :** Enfin, nous souhaitons illustrer la mise en œuvre de l'IA de confiance au sein d'une entreprise comme Thales. Cette partie vise à compléter les deux premières sections, qui abordent l'IA de confiance de manière globale et holistique. La mise en œuvre de l'IA de confiance peut prendre plusieurs formes et initiatives avec des temporalités variables.

L'ensemble des sources documentaires citées dans le panorama est disponible par lien cliquable à la fin de ce document.

1. Introduction

L'intelligence artificielle transforme notre monde de par ses applications multiples et son potentiel. Cependant, cette révolution technologique ne doit pas se faire au détriment de nos valeurs et de notre planète. **L'IA de confiance devient un enjeu primordial** pour construire un avenir où **l'innovation rime avec responsabilité et durabilité.**

De nombreux concepts ont émergé pour qualifier l'IA de confiance comme l'IA responsable, IA éthique, AI for good, green AI etc… mais **qu'est-ce concrètement qu'une IA de confiance et comment la mettre en œuvre en Europe ?**

D'un point de vue juridique, le cadre réglementaire entourant l'utilisation de l'IA ne cesse d'évoluer. **Plusieurs initiatives législatives, réglementaires et normatives** à différents niveaux émergent depuis quelques années et permettent d'identifier un contour de définition de l'IA de confiance :

> **Les normes internationales :** Des organisations internationales comme l'OCDE[2] ou l'ISO[3] travaillent à l'élaboration de normes et de lignes directrices non contraignantes en matière d'IA de confiance. L'UNESCO[4] a également souligné l'importance d'une IA éthique et respectueuse des droits de l'homme.

> **Le Règlement européen sur l'intelligence artificielle (AI ACT) :** Ce règlement pionnier, entré en vigueur en 2024, vise à établir un socle de confiance pour le développement et l'utilisation de l'IA au niveau européen. Il établit un cadre réglementaire différencié selon les risques posés par les systèmes d'IA, avec des obligations spécifiques pour les systèmes à haut risque. Ce règlement sur l'IA participe au détourage de la notion responsable, notamment en imposant des obligations de transparence et de lutte contre les biais discriminatoires. Il s'agit d'une première étape essentielle pour garantir la sécurité et le respect des droits fondamentaux.

> **Les initiatives nationales :** Les gouvernements, à l'instar de la France avec sa stratégie nationale pour l'IA, mettent également l'accent sur la nécessité d'un cadre réglementaire pour le développement de l'IA. De nombreux pays, tels que les États-Unis (notamment la Californie[5]), le Canada, la Chine, la Corée du Sud, le Japon et le Royaume-Uni, développent leurs propres initiatives sur l'IA avec des approches parfois très différentes du fait notamment de divergences culturelles et/ou idéologiques. Cela se traduit par une approche « ultra-libérale » américaine tandis que l'UE et la Corée du Sud adoptent une approche plutôt « régulatrice ».

> Ce cadre réglementaire en construction met en exergue le **rôle des programmes de recherche** pour aller au-delà de ces exigences légales minimales (programmes académiques, regroupements d'entreprises, organismes nationaux, Think tanks…) et leurs initiatives pour proposer des lignes directrices, des référentiels, des standards, des méthodes et de la documentation pour faciliter l'interprétation de la loi, anticiper les enjeux de demain et faire émerger des pratiques proactives responsables et durables.

Les principaux moteurs de la réglementation de l'IA de confiance sont :

Sécurité

LA PRISE DE CONSCIENCE DES RISQUES

Ethique

LA NECESSITE DE PROTEGER LES DROITS FONDAMENTAUX

Transparence

LA DEMANDE DE TRANSPARENCE

[2] OCDE, « Principles for Trustworthy AI » (2024)
[3] ISO/IEC 42001, « AI Management system » (2023)
[4] UNESCO, « Ethics of Artificial Intelligence » (2021)

[5] Voir l'article « Governor Newsom signs SB 53, advancing California's world-leading artificial intelligence industry » | Governor of California

2. Contexte et enjeux

Au-delà des considérations éthiques et légales qui ressortent principalement lorsqu'on évoque l'IA de confiance, un autre enjeu est devenu crucial au regard des derniers constats sur la consommation énergétique des modèles d'IA générative[6] : **la durabilité.** L'IA a en effet un impact environnemental significatif si elle n'est pas conçue et utilisée de manière responsable.

Il est donc **essentiel d'intégrer la durabilité au cœur des préoccupations liées à l'IA**. Cela passe par une réflexion sur l'empreinte environnementale des systèmes d'IA, incluant non seulement les algorithmes mais aussi les infrastructures matérielles nécessaires à leur fonctionnement (accélérateurs, centres de données, utilisation de métaux rares, etc.), le développement de solutions d'IA vertes et la promotion d'une culture de l'IA responsable au sein des entreprises et des organisations.

L'IA de confiance est une approche systémique qui ne se limite donc pas à une **conformité légale ou à une minimisation des risques potentiels mais qui implique une démarche proactive** d'information, de formation et d'engagement intégrant des principes éthiques, sociaux et environnementaux.

Mais que **recouvre exactement ce terme ?** Quels sont les cadres de référence qui **guident cette démarche ?** Et comment les organisations s'efforcent-elles de **l'intégrer dans leurs pratiques ?**

L'IA de confiance est un défi collectif et nous vous proposons ici quelques clés de lecture pour répondre à ces questions notamment :

- Les **principes composant l'IA de confiance**, et les concepts qui découlent de ces principes selon une approche holistique.
- Un **panorama non exhaustif des réglementations et des travaux disponibles** pour répondre aux exigences et aux défis associés à chacun de ces principes L'ensemble de la documentation citée est accessible à la fin du document.
- Une illustration des **initiatives et mises en œuvre** de l'IA de confiance au sein d'une organisation comme Thales.

3. Définition de l'IA de confiance

Comprendre l'IA de confiance commence par une clarification des termes dont la compréhension est essentielle pour toute initiative visant à développer une IA responsable, bénéfique et sûre. Autour de la notion d'IA de confiance gravite un **écosystème de principes, concepts et termes** connexes qui peuvent rendre sa lecture complexe. Les termes IA responsable, IA éthique ou encore IA digne de confiance sont parfois employés de manière interchangeable alors qu'ils renvoient à des dimensions bien spécifiques (cf figure 1).

- **L'IA Durable** : L'IA durable vise à concevoir l'IA comme « *une technologie de choix pour accélérer la transition écologique et mieux préserver les ressources de notre planète* »[7]. En ce sens, l'IA doit être conçue pour utiliser son potentiel au service de la transition écologique (meilleure connaissance des évolutions climatiques, optimisation des flux de ressources etc…) ainsi que dans une démarche pour améliorer son efficience environnementale (optimisation énergétique, écoconception…). Le terme de « AI for green » a émergé pour qualifier la première dimension et d'IA frugale ou « green AI » pour la seconde.
- **L'IA éthique** : L'IA éthique[8] se concentre sur les principes moraux et repose sur les valeurs sociétales qui devraient guider le développement et l'utilisation de l'IA de manière vertueuse[9] notamment avec les dimensions morales et légales de respect des droits fondamentaux et du cadre réglementaire en vigueur. Lié aux enjeux de l'IA Durable, le terme de « IA for Good » désigne l'application de l'IA pour créer un impact sociétal positif au service du bien commun.
- **L'IA transparente** : La transparence de l'IA vise à « *permettre aux individus de mieux comprendre comment sont créés les systèmes d'IA et comment ils prennent leurs décisions* »[10] en fournissant des explications compréhensibles et adaptées au contexte. L'explicabilité et la connaissance du mode de fonctionnement des modèles et de la logique des algorithmes constitue des fondamentaux de l'IA de confiance en renforçant la confiance dans les résultats, l'efficacité et l'équité des décisions.
- **L'IA sécurisée et sûre** : L'IA sécurisée et sûre vise à assurer la robustesse et la résilience du système aux conditions adverses, telles que les fuites de données, manipulation de modèles, leurre et les cyber-

---

[6] *Voir l'article « Intelligence artificielle en entreprise : quels impacts environnementaux ? » | Direction générale des Entreprises*
[7] *Source : DP_Coalition_mondiale_IAdurable.pdf | Sommet pour l'action sur l'IA*
[8] *Source : Définition « Ethical AI » - HLEG Commission européenne, « Ethics Guidelines for Trustworthy AI » (2019)*
[9] *Source : L'organisation ISO présente la nuance entre l'IA éthique et l'IA Responsable, « ISO - Bâtir une IA responsable : comment aborder le débat sur l'éthique de l'IA ? »*
[10] *Source : Qu'est-ce que la transparence de l'IA ? | IBM*

attaques. La sûreté (safety) de l'IA vise à traiter les problèmes inhérents et les conséquences imprévues, tandis que la sécurité de l'IA se concentre sur la protection des systèmes d'IA contre les menaces externes. Une IA sûre est un gage de fiabilité et constitue en ce sens un cadre structurant pour une IA de confiance.

- **L'IA responsable** : L'IA responsable (RAI) est un terme générique qui désigne les choix à faire lors de l'adoption de l'IA en matière de développement, de déploiement et de supervision des systèmes d'intelligence artificielle, de manière à ce qu'ils soient conformes aux attentes juridiques et sociétales tout en limitant au maximum les répercussions potentiellement négatives. L'IA responsable renforce la confiance dans les solutions d'IA en tenant compte de l'impact sociétal plus large des systèmes d'IA et des mesures nécessaires pour aligner ces technologies sur les valeurs des parties prenantes, les normes juridiques et les principes éthiques[11]. Ce terme renvoi à la dimension de la responsabilité opérationnelle (gouvernance, auditabilité...) et vise à garantir la mise en conformité vis-à-vis de la réglementation (AI Compliance). Pour Thales, la responsabilité c'est aussi la conformité par rapport à sa Charte éthique numérique[12]. Il s'agit de mettre en œuvre des cadres de gouvernance, des processus, des rôles, des audits, des mécanismes de redevabilité pour s'assurer que les systèmes d'IA sont développés et utilisés de manière éthique et bénéfique pour la société, en minimisant les risques[13]. La responsabilité d'un système d'IA s'exerce avant tout au niveau de ses parties prenantes, soit « *toutes les organisations et personnes impliquées dans, ou affectées par, les systèmes d'IA, directement ou indirectement* » *(comme les concepteurs, intégrateurs, utilisateurs, régulateurs…)*[14].
- **L'IA de confiance (Trustworthy AI)** : La Commission Européenne a popularisé le concept d'une IA digne de confiance notamment avec les travaux du Groupe d'experts indépendants de haut niveau en intelligence artificielle (GEHN IA)[15]. Ce concept vise à traduire les principes éthiques et durables en exigences concrètes et mesurables. Une IA digne de confiance pour la Commission Européenne est une IA à la fois sûre, transparente, éthique et sous contrôle humain[16]. Elle intègre des aspects techniques comme la robustesse, la fiabilité, et la sécurité, comme des dimensions légales de l'IA éthique avec la protection des données (RGPD par exemple) et la conformité avec le règlement sur l'intelligence artificielle.

  Thales s'inscrit dans cette vision, via son approche TRUE AI[17], en concevant l'IA de confiance par la mise en oeuvre de toutes les dimensions de l'IA (éthique, durable, sécurisée, transparente, responsable…) appliquées à un niveau d'exigence strict de souveraineté, au service de secteurs stratégiques et critiques. Une IA digne de confiance **assure la validité des systèmes, leur explicabilité, leur sécurité et leur responsabilité.**

La validité est le premier principe fondamental de tout déploiement d'IA. La validité garantit qu'un système d'IA fait ce qu'il doit faire, tout ce qu'il doit faire et seulement ce qu'il doit faire en confirmant, par la fourniture de preuves objectives, que « *les exigences relatives à une utilisation ou une application spécifique prévue ont été satisfaites*[18] ».

En partant de ce constat, nous proposons de décliner ces différentes dimensions de l'IA de confiance ci-après en positionnant les **concepts et principes clés** précédemment décrit et majoritairement cités dans les travaux.

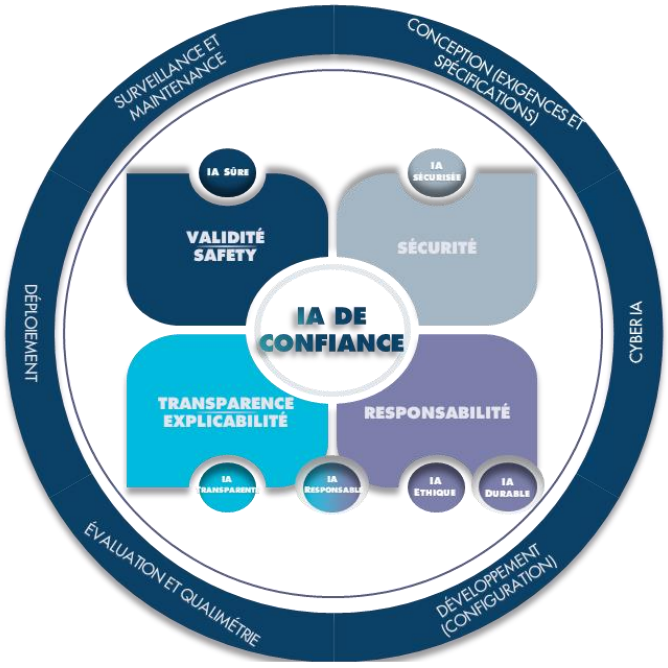


Figure 1 : L'IA de confiance et articulation avec le cycle de vie de l'IA.
Les principes de validité, de sécurité, de transparence et de responsabilité sont la base sur laquelle reposent les aspects de gouvernances, techniques et opérationnels dans l'approche de l'IA de confiance.

[11] *Source: What is responsible AI? | IBM*
[12] *Présentation de la Charte dans la partie 5.1.1 – Frameworks fondamentaux*
[13] *Scope « Responsabilité » couvert par les « Principes de l'IA » adoptés en mai 2019 par l'OCDE*
[14] *Renvoi à la notion de « stakeholders » et de « Accountability » OCDE, « Recommendation of the Council on Artificial Intelligence » (2019)*
[15] *HLEG Commission européenne, « Ethics Guidelines for Trustworthy AI »(2019)*
[16] *Source : Commission européenne, « Excellence et confiance en matière d'intelligence artificielle »*
[17] *Transparent, Reliable, Understandable, Ethical AI décrit dans le Position Paper Thales « Cycle de vie d'ingénierie IA fiable de bout en bout pour les systèmes critiques »*
[18] *Source : ISO/IEC 22989:2022, « Artificial intelligence concepts and terminology » (2022)*

4. Composants de l'IA de confiance

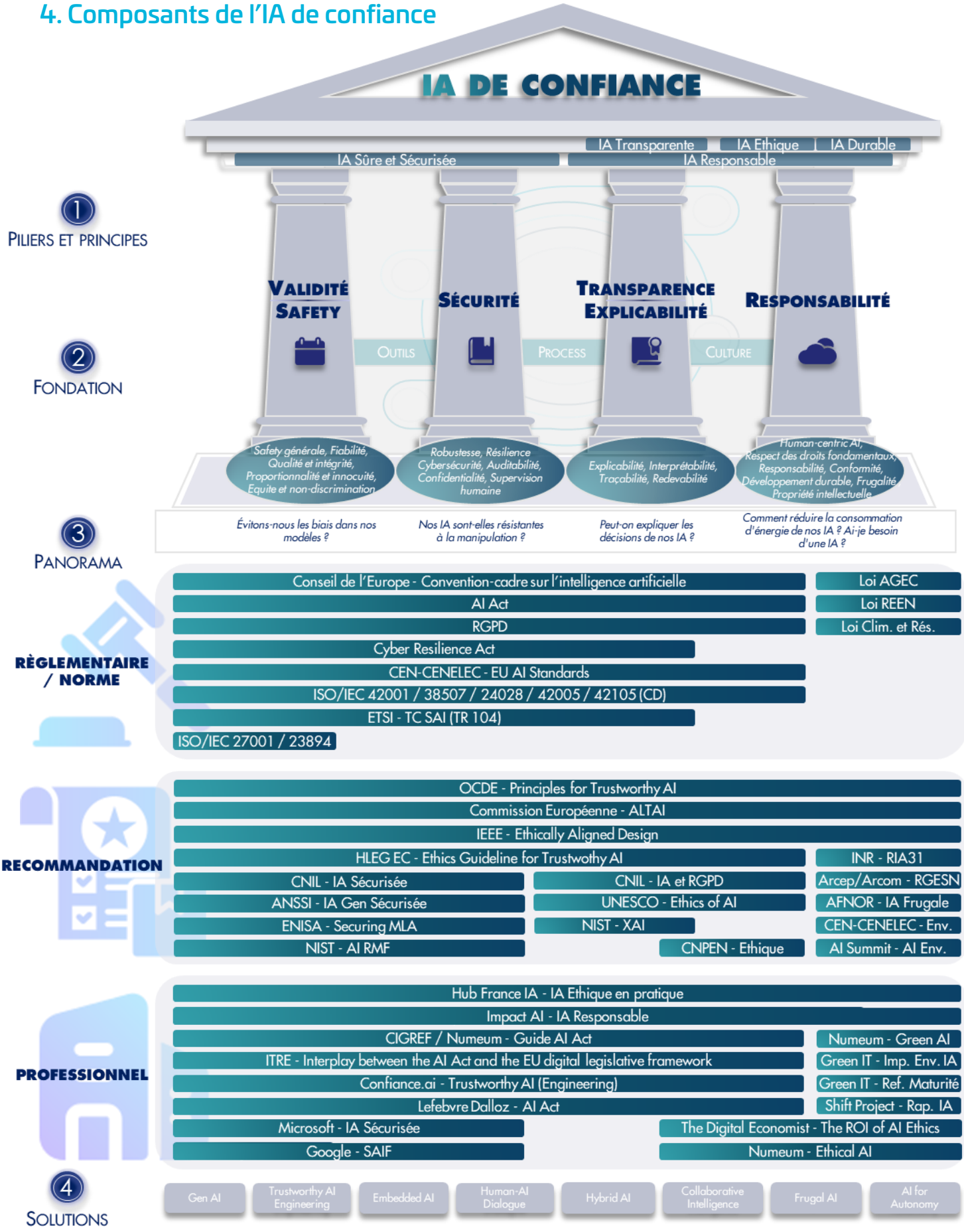


Figure 2 : Le temple de l'IA de confiance

4.1 PILIERS ET PRINCIPES DE L'IA DE CONFIANCE[19]

VALIDITE & SAFETY

Ce pilier intègre les aspects liés à l'évaluation et à la certification de la fiabilité, des performances et de l'éthique (notamment contre les biais) garantissant la validité d'un système d'IA, ainsi que les aspects de sûreté (safety) visant à prévenir les accidents et déviances internes ou conséquences néfastes liés à son utilisation.

- **Plan de sûreté (safety) général** : Assure que le système agira « comme prévu sans nuire aux êtres vivants ni à l'environnement. Cela implique notamment de réduire au minimum les conséquences imprévues et les erreurs. » En ce sens, des processus doivent être mis en place visant à clarifier et à évaluer les risques potentiels liés à l'utilisation des systèmes d'IA. « Le niveau des mesures de sécurité requises dépend de l'ampleur du risque posé par un système d'IA, qui dépend elle-même des capacités du système. Lorsqu'il est prévisible que le processus de développement ou le système lui-même présentera des risques particulièrement élevés, il est crucial que des mesures de sécurité soient élaborées et testées de manière proactive. » *[CE]*
- **Fiabilité** : « Propriété de cohérence du comportement et des résultats prévus » [*ISO/IEC]*. « Un système d'IA fiable est celui qui fonctionne correctement avec une gamme variée d'entrées et dans diverses situations » afin de l'examiner et prévenir les dommages imprévus. *[CE]*
- **Qualité et Intégrité** : Assure que les données /modèles /inputs /outputs ne sont pas altérés (absence de modifications inappropriées ou données malveillantes) et sont de haute qualité. *[CE]*
- **Proportionnalité et innocuité :** « Aucun des processus liés au cycle de vie des systèmes d'IA ne devrait aller au-delà de ce qui est nécessaire pour atteindre des buts ou objectifs légitimes ». *[UNESCO]*
- **Equité et non-discrimination** : « Les acteurs de l'IA doivent promouvoir la justice sociale, garantir l'équité et lutter contre les discriminations, tout en adoptant une approche inclusive pour s'assurer que les bénéfices de l'IA sont disponibles et accessibles à tous ». *[UNESCO]* Les principes d'équité et de non-discrimination visent à lutter contre les biais dès la conception du système d'IA (en assurant la diversité des données d'entraînement, la surveillance de ces données et des algorithmes, la transparence des décisions prises par les systèmes d'IA…).

SECURITE

Ce pilier englobe les aspects qui protègent les utilisateurs, les systèmes et les données contre des risques ou menaces d'attaques cyber et de déviances externes.

- **Robustesse** : Capacité d'un système d'IA « à maintenir son niveau de performance en toutes circonstances » [*ISO/IEC]*. La robustesse d'un système d'IA est une notion liée à la prévention des dommages, en englobant la robustesse « d'un point de vue technique ([…] selon le domaine d'application ou la phase du cycle de vie) et d'un point de vue social ([…] selon le contexte et l'environnement dans lesquels le système fonctionne) », de manière à ce que les systèmes d'IA « se comportent comme prévu tout en minimisant les dommages involontaires et inattendus, et en prévenant les dommages inacceptables. ». *[CE]*
- **Résilience** : « Capacité d'un système à récupérer rapidement une condition opérationnelle après un incident » [*ISO/IEC]*. La résilience face aux attaques des systèmes d'IA implique la protection contre les vulnérabilités (au niveau des données, modèles ou infrastructures) qui peuvent leur permettre d'être exploités « par une intention malveillante ou par l'exposition à des situations inattendues ». *[CE]*
- **Cybersécurité :** Ensemble des mesures de défense qui garantissent l'intégrité, la confidentialité, la disponibilité et la traçabilité d'un système d'IA face aux menaces cyber. Elles concernent les données d'entraînement, le développement, le fonctionnement du système, ainsi que des mesures transversales. *[CNIL]*
- **Auditabilité :** « Implique l'activation de l'évaluation des algorithmes, des données et des processus de conception … L'évaluation par des auditeurs internes et externes, et la disponibilité de tels rapports d'évaluation, peuvent contribuer à la fiabilité de la technologie ». *[CE]*
- **Confidentialité et protection des données privées** : Garantit la « confidentialité et la protection des données tout au long du cycle de vie d'un système d'IA. Cela inclut les informations fournies initialement par l'utilisateur, ainsi que les informations générées sur l'utilisateur au cours de son interaction avec le système ». *[CE]*
- **Supervision humaine (Human oversight)** : Assure qu'un système d'IA ne prend pas de décisions critiques de manière autonome et reste sous contrôle humain ou ne provoque pas d'effets indésirables. « La surveillance peut être assurée par des mécanismes de gouvernance tels qu'une approche « human-in-the-loop » (HITL), « human-on-the-loop » (HOTL) ou « human-in-command » (HIC) ». *[CE]*

[19] ***Sources Validité & Safety*** : *Commission européenne (CE), « Ethics Guidelines for Trustworthy AI » (2019); ISO/IEC 22989:2022, « Artificial intelligence concepts and terminology » (2022); UNESCO, « Ethique de l'IA » ;*
Sources Sécurité : *ISO/IEC 22989:2022, « Artificial intelligence concepts and terminology » (2022); Commission européenne (CE), « Ethics Guidelines for Trustworthy AI » (2019); CNIL « IA : Garantir la sécurité du développement d'un système d'IA | CNIL »*
Sources Transparence & Explicabilité : *OCDE, « Recommendation of the Council on Artificial Intelligence » (2019) ; ISO/IEC 22989:2022, « Artificial intelligence concepts and terminology » (2022) ; IBM, « Qu'est-ce que l'interprétabilité de l'IA ? » (2024) ; IEEE, « Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems» (2019)*
Source Responsabilité : *Commission européenne (CE), « Ethics Guidelines for Trustworthy AI » (2019); OCDE, « Recommendation of the Council on Artificial Intelligence» (2019); UNESCO, « Ethique de l'IA »; IEEE, « Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems » (2019); Union Européenne, « AI Act » (2024); OMPI, « Inventions fondées sur l'IA » (2024); CNIL, « Déclaration commune sur la gouvernance fiable des données pour l'IA » (2025); IBM, « Conformité de l'IA : définition, importance et premiers pas | IBM » ; UNESCO, « Artificial Intelligence and Sustainable Development » (2019); AFNOR, « IA frugale » (2024);*

TRANSPARENCE & EXPLICABILITE

Ce pilier s'assure que les décisions de l'IA sont compréhensibles, accessibles et qu'il est possible de retracer les processus pour des explications claires.

- **Transparence** : Vise à rendre les décisions et processus de l'IA explicables pour que les personnes touchées par un système d'IA puissent comprendre et saisir les résultats *[OCDE]*. C'est « la propriété d'un système d'IA selon laquelle les informations appropriées sur le système sont mises à la disposition des parties prenantes concernées ». *[ISO/IEC]*
- **Interprétabilité** : Alors que l'explicabilité décrit « les raisons qui conduisent un système d'IA à produire des résultats », l'interprétabilité se concentre sur « la compréhension du fonctionnement interne des modèles d'IA » pour mieux comprendre, clarifier et rendre accessible les processus de prise de décision (architecture, caractéristiques, biais...). *[IBM]*
- **Traçabilité** : En complément de la sécurité et safety, la traçabilité de « l'ensemble de données, des processus et des décisions prises au cours du cycle de vie du système d'IA » permet « une analyse des résultats du système d'IA et des réponses à la demande, adaptée au contexte et cohérente avec l'état de l'art », grâce à un suivi et une documentation technique détaillée. *[OCDE]*
- **Redevabilité :** La transparence assure la redevabilité qui implique que le système d'IA doit être « créé et exploité pour fournir une justification de toutes les décisions prises ». Cela concerne l'obligation de rendre compte techniquement, par des mécanismes de suivi, des effets causés par l'IA par le respect des exigences réglementaires et en rendant intelligible la logique sous-jacente au système d'IA en veillant à ce qu'elle fournisse une justification non ambiguë de toutes les décisions prises. *[IEEE]* La redevabilité permet de démontrer comment une décision a été prise (données utilisées, algorithme appliqué, processus de validation, etc.), alors que la responsabilité permet d'identifier l'acteur qui devra répondre de cette décision.

RESPONSABILITE

Ce dernier pilier regroupe les principes relatifs à la responsabilité des acteurs en particulier sur les aspects de conformité, d'éthique et de durabilité de l'IA. L'enjeux est de promouvoir une IA qui respecte les valeurs morales et humaines, en prenant en compte les impacts environnementaux et sociétaux, en faveur du bien-être humain et du développement durable.

- **IA centrée sur l'humain :** Approche qui met en avant le respect de la primauté de l'état de droit, des droits de l'homme et des valeurs démocratiques tout au long du cycle de vie du système d'IA incluant la liberté, la dignité et l'autonomie humaine, la protection de la vie privée et des données, la non-discrimination et l'égalité, la diversité, l'équité, la justice sociale. *[CE, OCDE]*
- **Respect des droits fondamentaux : «** Droits inscrits dans les traités de l'UE, la Charte de l'UE et le droit international des droits de l'Homme » (par référence à la dignité, aux libertés, à l'égalité et à la solidarité, aux droits des citoyens et à la justice) qui permettent de fournir des « bases pour identifier des principes et valeurs éthiques à opérationnaliser dans le contexte de l'IA ». *[CE]*
- **Responsabilité des acteurs** : Etablir la responsabilité (morale et juridique) de l'ensemble des acteurs de la chaîne de valeur de l'IA relativement au bon fonctionnement sur l'ensemble du cycle de vie des systèmes d'IA, conformément aux exigences légales et à l'état de l'art. *[OCDE, UNESCO, IEEE, AI Act]*
- **Conformité réglementaire (cadre juridique, standards et normes)** : La conformité en matière d'IA fait référence *« aux décisions et pratiques qui permettent de se conformer aux lois et réglementations régissant l'utilisation des systèmes d'IA. Ces normes comprennent les lois, les règlements et les politiques internes visant à assurer le développement responsable des modèles d'IA et de leurs algorithmes au sein des organisations ». [IBM]*
- **Développement durable** : Promeut la recherche, la collaboration multipartite, l'accès équitable à l'IA et la sensibilisation à l'éthique et à la pensée critique pour éviter les effets négatifs sur les sociétés, en facilitant l'accès à des infrastructures de données inclusives (pour éviter les biais) et de calculs durables pour aligner la croissance de l'IA avec les objectifs climatiques globaux (notamment les 17 ODD adoptés par l'ONU en 2015). Les technologies de l'IA devraient être continuellement évaluées en fonction de leur impact sur la durabilité. *[UNESCO, OCDE]*
- **Bien-être humain :** Promeut la « qualité de vie des individus et des communautés » avec l'utilisation de l'IA. *[OCDE]*
- **Frugalité :** Réduction globale, dès la conception, « des besoins en ressources matérielles et énergétiques et des impacts environnementaux associés via une redéfinition des usages ou des exigences de performance, ou encore via une réorientation des besoins du producteur du système d'IA au fournisseur du service considéré, » afin de limiter son empreinte écologique. *[AFNOR]*
- **Propriété intellectuelle** : Enjeux de protection des données d'apprentissage (œuvres protégées par le droit d'auteur), et de respect et de reconnaissance de la propriété intellectuelle des inventions créées par l'IA (données, algorithmes, modèles), intégrant l'utilisation de l'IA ou réalisées avec l'aide d'IA. *[OMPI]* L'intégration des principes de protection des données doit se faire dès la conception. *[CNIL]*

4.2 FONDATION POUR L'ADOPTION

Si les quatre piliers constituent des principes essentiels de l'IA de confiance, leur consolidation dépend intrinsèquement d'une organisation et de fondations solides qui reposent sur : **des outils adaptés, des processus bien définis et une culture d'entreprise engagée.**

De manière générale, la conception et le déploiement de l'IA doit s'accompagner d'une réflexion approfondie sur leur acceptabilité sociale. L'acceptabilité sociale, dans le contexte de l'IA, fait référence au degré auquel la technologie est perçue comme **appropriée, juste et bénéfique** par les individus et la société dans son ensemble. L'UNESCO, par exemple, dans sa Recommandation sur l'éthique de l'intelligence artificielle (2021), souligne l'importance d'une participation inclusive des parties prenantes pour garantir que les systèmes d'IA soient "acceptables par la société". D'autres travaux de recherche, notamment ceux de la Commission européenne dans son Livre Blanc sur l'intelligence artificielle (2020), mettent également en avant l'importance de la confiance et de l'adhésion du public pour un développement réussi de l'IA.

Au-delà de la faisabilité technique, adopter une démarche d'**IA de confiance repose sur le questionnement avant tout de la réelle nécessité de l'IA et de son impact**. Cela implique de questionner pour chaque nouveau projet la nécessité même de déployer une IA pour résoudre le problème posé.

Ce sujet est d'autant plus important mis en perspective avec le **paradoxe de Jevons** qui soulève un défi majeur. Ce paradoxe stipule qu'une augmentation de l'efficience dans l'utilisation d'une ressource peut paradoxalement conduire à une augmentation de sa consommation totale. Appliqué à l'IA, cela signifie que si l'IA rend certains processus plus efficaces, elle pourrait, en même temps, encourager une utilisation plus répandue et potentiellement excessive, soulevant des questions quant à sa **véritable nécessité et à son impact global sur l'environnement**, l'emploi, ou encore la consommation de données. La question de l'acceptabilité sociale et des raisons profondes de nos choix technologiques devient alors d'autant plus cruciale.

Ces questionnements et le coût initial d'une IA de confiance, incluant la prise en compte des critères éthiques et environnementaux en phase de conception, la validation rigoureuse des modèles, la transparence et la gouvernance, peut sembler élevé et peser sur la compétitivité. Cependant, ce coût doit être perçu comme un investissement stratégique, d'autant plus pour les systèmes critiques, afin d'assurer **la viabilité à long terme dans écosystème technologique de plus en plus axé sur la responsabilité et l'éthique.**

Des Outils au service de l'humain et de la responsabilité

Une démarche d'IA de confiance gagne considérablement en efficacité lorsqu'elle s'appuie sur une boîte à outils technologiques appropriée. Cela inclut des outils pour protéger et sécuriser les algorithmes comme des plateformes permettant la traçabilité des données et des modèles ; des solutions pour l'explicabilité de l'IA facilitant la compréhension des décisions algorithmiques ; des méthodes d'analyse du cycle de vie (ACV) d'écoconception ; ou encore des instruments d'audit pour détecter et corriger les biais potentiels. Ces outils ne sont pas de simples compléments ; ils sont des catalyseurs qui permettent de passer des intentions aux actions concrètes, offrant les moyens techniques de mesurer, de contrôler et d'améliorer la responsabilité des systèmes d'IA.

Des Processus intégrés

L'IA de confiance doit être intégrée dans le tissu même des processus de développement et de déploiement de chaque solution d'IA ainsi que dans sa gouvernance. Il s'agit d'ancrer des rôles spécifiques, des revues régulières de conformités réglementaires, de respect des droits fondamentaux, de considérations éthiques et de sécurités dès la conception, d'intégrer des évaluations d'impact systématiques, et de définir des boucles de retours pour une amélioration continue. En inscrivant ces exigences dans les étapes clés du cycle de vie de l'IA, on garantit que la responsabilité n'est pas une option, mais une exigence opérationnelle à chaque niveau.

Une Culture d'entreprise responsable

Enfin, la transformation la plus profonde s'opère au niveau humain. La construction d'une IA de confiance est avant tout une affaire de culture d'entreprise orientée sur la gouvernance, la définition des rôles et l'acceptabilité. Cela implique de sensibiliser, de former et de responsabiliser l'ensemble des équipes (des développeurs aux décideurs) aux enjeux éthiques, sociétaux et environnementaux de l'IA. C'est en favorisant une culture axée sur la diligence, la redevabilité et la curiosité critique que l'on encourage chacun à remettre en question, à être proactif, à anticiper les risques et à s'engager activement dans la conception de systèmes d'IA qui non seulement servent les objectifs de l'entreprise, mais aussi une utilisation plus vertueuse de l'IA. Une culture forte est le ciment qui lie les processus et les outils, transformant les principes abstraits en pratiques quotidiennes et durables.

L'IA de confiance est une approche essentielle pour assurer que l'IA soit utilisée de **manière bénéfique pour tous en restant au service de l'humain.**

4.3 PANORAMA DE LA DOCUMENTATION

Une **variété d'acteurs** alimente les réflexions et législations autours des différents principes de l'IA de confiance : centres de recherche, associations d'entreprises, communautés professionnelles, autorités administratives, organisations internationales ou gouvernementales… Pour chacun des piliers et des notions associées, il est possible de retrouver **des référentiels, guides, ou fiches pratiques / méthodologiques** pour répondre au mieux aux enjeux soulevés.

Ce document présente une partie de la documentation identifiée sur **trois niveaux** :

Il existe différents cadres (base légale, règlements, normes ou standards) pour répondre aux exigences réglementaires ou respecter les normes recommandées par un organisme de normalisation. La réglementation et la normalisation sont deux aspects considérés comme complémentaires.

Travaux réalisés par des autorités administratives, missions ministérielles, organisations internationales ou gouvernementales sous forme de référentiels, guides, fiches pratiques et recommandations. L'objectif est généralement de fournir des outils, un cadre et des bonnes pratiques pour faciliter l'application de la loi ou répondre aux enjeux de l'IA de confiance.

Travaux réalisés par des associations de professionnels ou d'entreprises, des Think tanks, des communautés professionnelles ou entreprise. Il s'agit généralement de guides, de documents méthodologiques ou de recherches sur les thématiques d'une IA de confiance, axés sur l'applicabilité de la loi ou d'un principe, la spécificité d'un secteur d'activité ou un outil.

L'ensemble de la documentation citée est disponible à la fin du document (7. Ressources). Pour chaque thématique évoquée, des travaux plus approfondis ou adaptés à certains secteurs d'activités sont disponibles au sein des groupes de travail, organisations ou institutions cités. Nous encourageons ainsi le lecteur à explorer les **ressources complémentaires** offertes par ces acteurs clés selon ses besoins ou intérêts.

4.4 SOLUTIONS

Le déploiement d'une IA de confiance est un défi **multidimensionnel** en passant par la gouvernance jusqu'à la technologie adéquate. Le panorama proposé met en lumière les aspects de gouvernance, réglementaires et certaines dimensions techniques concernant la sécurité, mais il est nécessaire de souligner l'importance de la technologie pour avoir une vison **complète du déploiement d'un système d'IA.**

La sélection adaptée de technologies d'IA en apprentissage est centrale dans l'approche de l'IA de confiance, afin de répondre **aux besoins et contraintes appliquées à des secteurs d'activité ou métiers spécifiques.**

Les technologies d'IA permettent de répondre à des dimensions spécifiques de l'IA de confiance et au sein de Thales des technologies différanciantes ont été identifiées en particulier pour répondre aux **besoins spécifiques des systèmes critiques** qui nécessitent de prendre en compte des contraintes de sécurité et de sûreté adaptées au niveau de criticité.

Ces technologies sont présentées dans l'article "IA pour les systèmes critiques" (2024)[20] (voir figure 3) et visent à apporter des **réponses concrètes aux enjeux de l'IA.**

Pour mettre en œuvre une IA responsable et digne confiance des technologies tels que l'IA hybride permettent de combiner les forces des méthodes connexionnistes en y intégrant la logique et le raisonnement de l'IA symbolique pour garantir la transparence, l'interprétabilité et la robustesse ; l'IA frugale, une approche qui permet de minimiser l'empreinte écologique et les coûts de déploiement ; et l'IA auto-explicable, essentielle pour bâtir la confiance en rendant les décisions des systèmes d'IA compréhensibles et justifiables par les humains.

La mise en œuvre de l'IA de confiance peut donc prendre **plusieurs formes et temporalités selon les besoins**, les enjeux, l'organisation, les secteurs d'activité et les contextes.

La conviction de Thales est que, sous réserve d'être parfaitement maîtrisée, l'IA apporte un support à l'humain pour qu'il puisse prendre des décisions à la fois meilleures et plus rapides afin d'apréhender **l'IA comme un complément de l'intelligence humaine.**

Thales se concentre sur le renouvellement de ses solutions à un niveau toujours croissant de performance, de sécurité et de durabilité. Cette démarche prend forme en partie au travers la mise en œuvre de **l'IA de confiance et d'initiatives variées pour concevoir une IA fiable dès sa conception, depuis les données et les algorithmes jusqu'à la validation.**

En savoir plus :

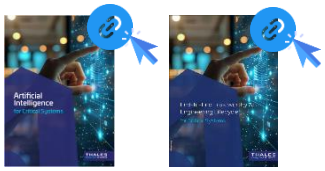


Figure 3 : Les technologies différenciantes en IA

[20] *Juliette MATTIOLI et Christophe MEYER, « IA dans les systèmes critiques», (2024)*

5. Initiatives Thales pour mettre en œuvre une IA digne de confiance

Au sein de Thales, l'IA de confiance est **un engagement profond** qui ne se limite pas à la conformité réglementaire mais qui guide depuis plusieurs années sa stratégie d'innovation. L'IA de confiance couvre des aspects très spécifiques (comme la sécurité, la transparence, l'éthique, la durabilité…) qui ont pu être mis en place à des temporalités différentes en s'appuyant sur des expertises variées avec une coordination adaptée. Outre l'expertise de ses acteurs et leur coordination, le déploiement de l'IA de confiance relève de nombreux facteurs comme les convictions sur le numérique et l'IA, la stratégie commerciale et de développement, la nature des produits et services, l'organisation interne, le secteur d'activité, la réglementation applicable etc…

Au sein du groupe Thales, les réflexions sur l'éthique et la gouvernance numérique s'ancrent depuis le début de l'utilisation des technologies et de l'IA dans ses projets. En 2019, les fondations de l'IA de confiance au sein du groupe Thales sont posées et déployées avec **la démarche TRUE AI** suivie en 2022 de la **Charte éthique numérique** donnant les lignes directrices d'un numérique responsable et de confiance et s'appliquant à l'ensemble des activités du Groupe, en cohérence avec notre raison d'être. La démarche TRUE est articulée autour de quatre piliers pour mettre en œuvre une IA de confiance : Transparence, fiabilité, explicabilité et éthique. Ce cadre vise à assurer l'intégration dès la conception des systèmes d'IA des exigences **fortes de transparence, de compréhension du fonctionnement des modèles d'IA et de respect des principes éthiques.**

Cette responsabilité s'exprime dans toutes ses dimensions (technologique, sociétale et environnementale) afin d'assurer que l'IA conçue soit non seulement performante et fiable, mais également pérenne sur le long terme.

La mise en œuvre d'une démarche d'IA de confiance implique la **mobilisation d'acteurs variés** aussi bien techniques, que juridiques, cyber ou spécialiste produit/service en se déclinant sur plusieurs niveaux entre les outils, les processus et la culture d'entreprise. Pour piloter cette démarche et assurer la gouvernance de l'IA, il est essentiel de s'appuyer sur ces acteurs avec une cellule transverse qui assure la **mise en conformité des systèmes d'IA aux standards européens et internationaux**.

L'IA de confiance implique de nombreux principes comme décrit précédemment comme la sécurité, la sureté et la transparence. Dans l'ambition **d'élaborer des IA de confiance**, dès 2018 Thales a contribué aux travaux structurants du projet **DEEL** (DEpendable Explainable Learning), visant à développer des briques technologiques d'intelligence artificielle fiables, robustes, explicables et certifiables, appliquées à des systèmes critiques. Ces travaux scientifiques, qui ont ressemblés des chercheurs Franco-Québécois, alimentent la démarche d'IA de confiance en traitant les défis liés à la robustesse, l'explicabilité, la quantification de l'incertitude, la détection des erreurs de distribution et les biais industriels. Dans cette continuité, Thales devient en 2021 un des membres fondateurs du programme **Confiance.ai** qui vise à « *faire de la France l'un des leaders de l'IA industrielle et responsable en développant un cadre méthodologique et technologique souverain, ouvert, interopérable et durable, afin d'accélérer l'intégration d'une IA de confiance dans les industries stratégiques*. »[21]. Une communauté, dédiée à l'accélération du déploiement d'une IA responsable et de confiance dans les systèmes industriels, qui réunit acteurs industriels et académiques majeurs. Elle a pour objectif de **développer des méthodes, des outils et des standards** permettant d'accompagner la recherche et l'ingénierie dans le développement de systèmes **d'IA robustes, sécurisés, explicables et maîtrisés tout au long de leur cycle de vie** via une approche par use case. À l'issue de Confiance.ai en 2024, the **European Trustworthy AI Association** a été créée afin de poursuivre l'animation de la communauté de l'ingénierie de l'IA de confiance et de pérenniser les résultats, les méthodes et outils. Dans la continuité directe de ces travaux, le programme de recherche **CSIA (Confiance dans les systèmes d'IA)** élargit le périmètre méthodologique. Là où **Confiance.ai** s'est principalement concentré sur les modèles d'apprentissage automatique, **CSIA** complète et adapte les méthodes d'ingénierie pour intégrer l'IA hybride (combinant apprentissage et raisonnement symbolique) ainsi que l'IA générative. Il s'agit d'aligner architectures, processus de validation et cadres de maîtrise des risques avec ces nouveaux paradigmes, tout en conservant les exigences propres aux systèmes critiques.

À travers ces initiatives, Thales s'inscrit pleinement dans un écosystème national et collaboratif dédié à l'IA de confiance appliquée aux systèmes critiques, tout en tenant compte du cadre réglementaire européen de l'AI Act.

Thales présente en 2024 **son dispositif d'accélération de l'intégration de l'intelligence artificielle appliquée aux systèmes critiques**, en particulier dans le secteur de la défense tout en capitalisant sur son expertise technique et sa maitrise de l'ensemble des technologies clés de l'IA. La création de **cortAIx** a pour ambition de

[21] *Confiance.ai : Home - Confiance.ai*

rassembler les capacités IA du groupe dans le domaine de la recherche, des capteurs et des systèmes critiques pour assurer une application robuste et responsable de l'IA. Ce programme d'accélération s'applique en particulier au monde de la Défense qui implique la prise en compte de contraintes fortes et spécifiques, telles que la cybersécurité, l'embarquabilité et la frugalité, liées aux environnements critiques[22].

Le groupe Thales se positionne ainsi parmi **les leaders européens en matière de dépôts de brevets** liés à l'intelligence artificielle appliquée aux systèmes critiques et intègre déjà l'IA dans plus d'une centaine de ses produits et services. Si ceci se traduit souvent par des avancées en sécurité, transparence et éthique, Thales accorde également un fort intérêt **dans l'engagement environnemental,** à travers la participation à des programmes et le développement de solutions pour adresser par exemple l'aviation durable (Green aviation)[23], conciliant ainsi innovation technologique et durabilité dans des secteurs critiques tels **que l'aéronautique et le spatial.**

Parallèlement, l'implication au sein d'instances européennes ou d'organisations professionnelles, permet également au groupe Thales de contribuer à **l'élaboration de référentiels adaptés aux systèmes critiques et aux réalités industrielles,** tout en intégrant les enjeux de sécurité et de durabilité au cœur des réflexions.

Au sein de Thales, l'IA est conçue comme un levier pour relever les défis d'aujourd'hui tout en créant une valeur pérenne et **digne de confiance pour demain.**

5.1 Frameworks et dynamiques collaboratives de Thales

L'Intelligence Artificielle se décline au sein de Thales sur une grande diversité de sujets, allant de la recherche fondamentale aux applications concrètes au service du civil et de la défense. L'élaboration d'une IA de confiance s'appuie sur des process, outils ainsi qu'une culture adaptée qui se manifeste au travers de frameworks fondamentaux pour le numérique ainsi que des groupes de travail spécialisés.

5.1.1 – Frameworks fondamentaux

Les frameworks au niveau du groupe permettent d'établir un ensemble de lignes directrices sur la confiance et la responsabilité numériques qui est essentiel pour Thales afin d'articuler ses priorités d'innovation avec ses engagements pour construire un monde plus sûr, plus vert et plus inclusif. Ces frameworks sont le reflet de développements d'un grand nombre de processus et de réflexions sur l'éthique et la gouvernance numérique initiées dès l'utilisation de nouvelles technologies dans les projets Thales.

- *Charte Éthique Numérique*

Contexte et objectifs

Définir des lignes directrices pour un numérique responsable, sûr, et durable, en respectant les enjeux éthiques, environnementaux, et sociétaux. La Charte est une déclaration publique de l'engagement de Thales à utiliser la technologie numérique de manière responsable, éthique et respectueuse de l'environnement.

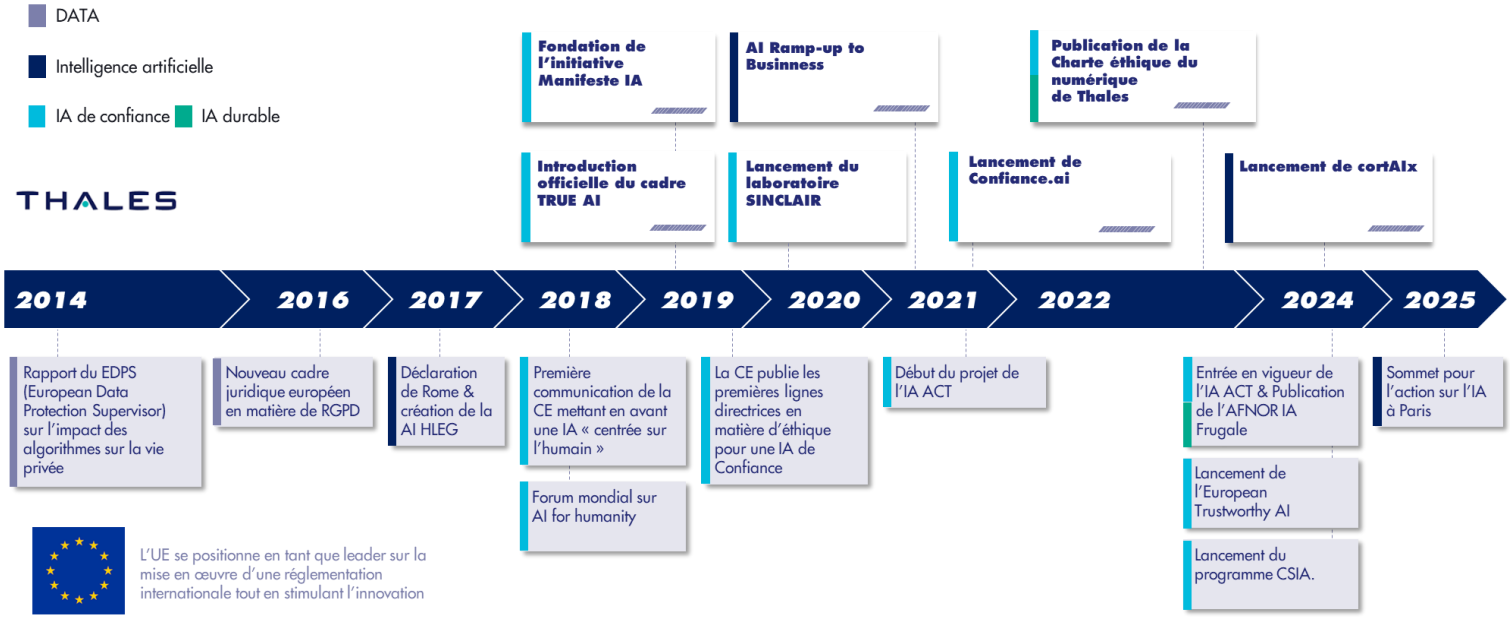


Figure 4 : Réglementation européenne de l'IA et IA #Thales

[22] *Voir article : Thales accélère dans l'IA pour la défense | Thales Group*

[23] *Voir article : Les enjeux techniques et technologiques sont indissociables des enjeux environnementaux | Thales Group et ce document*

Cibles et portée

Cette initiative s'applique à l'ensemble des produits, solutions et services de Thales, et s'adresse à tous : collaborateurs, clients, partenaires et citoyens concernés par les enjeux du numérique.

Description

Thales définit dix engagements en matière de confiance et de responsabilité numériques, structurés autour de quatre axes prioritaires :

- **Maintenir** le contrôle par l'humain, et la transparence des systèmes d'IA.
- **Renforcer** la sûreté et la sécurité des solutions.
- **Favoriser** un monde plus respectueux de l'environnement.
- **Placer** l'humain au centre des technologies afin de favoriser un monde inclusif et équitable.

Résultats

- Amélioration de la sécurité et résilience des solutions numériques.
- Réduction de l'impact environnemental.
- Protection accrue des données personnelles.
- Contribution à la lutte contre le changement climatique.
- Adoption de pratiques éthiques et responsables.

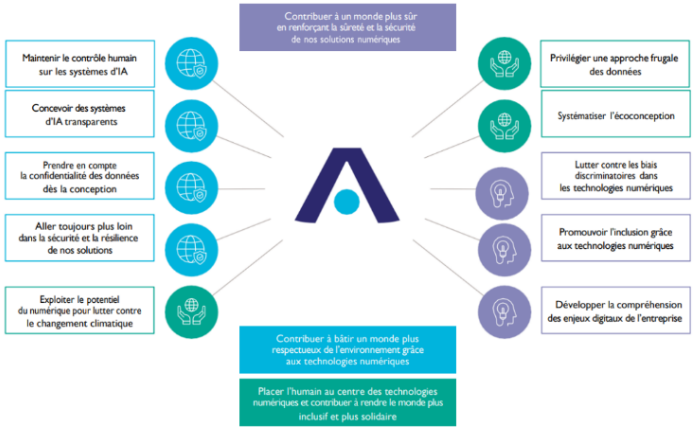


Figure 5 : Les 10 principes de la Charte éthique du Numérique de Thales

◇ TRUE AI (Transparent, Reliable, Understandable, Ethical)

Contexte et objectifs

Pour accompagner la transformation numérique et l'essor de l'intelligence artificielle, Thales a développé l'approche TRUE AI (Transparent, Reliable, Understandable, Ethical) en fournissant des produits et services responsables, alignés sur les engagements du groupe en matière de responsabilité numérique et d'éthique. L'objectif est de renforcer la confiance des utilisateurs et des partenaires, condition essentielle à l'adoption des technologies sensibles.

Cibles et portée

- Utilisateurs finaux,
- Clients,
- Partenaires,
- Régulateurs.

Ayant une portée nationale et internationale, TRUE AI est intégrée dans la Charte éthique numérique de Thales et est conforme aux cadres réglementaires (ex. RGPD, AI ACT).

Description

TRUE AI repose sur quatre principes majeurs de l'IA de confiance appliqués à toutes les solutions du groupe :

- **Transparent** : Règles de conception et de déploiement claires, respect de la confidentialité et du consentement.
- **Reliable** : Garantie de fiabilité pour les systèmes critiques au sens « Mission, Safety and Business Critical ».
- **Understandable** : Explicabilité des usages et des résultats pour garantir l'acceptabilité.
- **Ethical** : Conformité réglementaire, respect des standards, promotion de la non-discrimination et de l'égalité.

Résultats

- **Diffusion** d'un cadre clair pour un usage responsable de l'IA et des technologies sensibles.
- **Renforcement** de la confiance et de l'acceptabilité des solutions Thales.
- **Positionnement** du groupe comme partenaire fiable et responsable.

Figure 6 : Thales TRUE AI – Transparent, Reliable, Understandable, Ethical

5.1.2 – Exemples de contributions et collaborations nationales et européennes

Thales joue un rôle moteur dans l'écosystème national et européen de l'IA en contribuant à des ***groupes de travail et projets collaboratifs****, afin de décliner les travaux en* ***méthodologies ou standards appliqués aux enjeux industriels ou civils****. Ces initiatives, menées en partenariat avec des industriels, des laboratoires de recherche et des institutions, s'inscrivent dans une dynamique de coopération nationale et européenne visant à renforcer la souveraineté technologique et à garantir le développement d'une IA de confiance.*

◇ *Confiance.ai, Euro CSIA et European Trustworthy AI Association (#Safety #Securité #Transparence #Responsabilité)*

Contexte et objectifs

Programme de R&D (2021–2024) issu du Grand Défi « Sécuriser, fiabiliser et certifier des systèmes fondés sur l'intelligence artificielle », réunissant 13 industriels et académiques français dont Thales en tant que membre fondateur. Ce collectif a pour objectif de développer un socle méthodologique et technologique permettant l'industrialisation d'IA fiables et certifiables, en cohérence avec l'AI Act, afin de répondre aux enjeux de robustesse, d'explicabilité et de certification des systèmes critiques.

Cibles et portée

Le programme cible des industries stratégiques telles que l'aéronautique, la défense, l'automobile ou encore l'énergie, avec une application directe à des cas d'usage concrets.

Description

Les travaux sont structurés autour de cinq axes de recherche — fiabilité, explicabilité, ingénierie des données, confiance by design et IA embarquée — coordonnés par l'IRT SystemX et déployés à travers des cas industriels impliquant Renault, Valeo, Thales et d'autres acteurs.

Résultats

- Élaboration d'une méthodologie outillée et d'un catalogue open source de composants ;
- Premières applications concrètes, notamment chez Thales pour l'explicabilité des algorithmes de détection d'objets.

Conclusion et perspectives

La capitalisation des travaux de Confiance.ai se formalise à travers la création de la European Trustworthy AI Association et d'un groupe de travail CSIA, prolongeant le programme en étendant la méthodologie, afin de diffuser et pérenniser les résultats, de contribuer aux standards et de préparer les prochaines évolutions (IA générative, cybersécurité, hybridation IA).

◇ *WG 114 Artificial Intelligence (#Safety #Transparence)*

Contexte et objectifs

Créé par l'EUROCAE, le WG-114 réunit des industriels tels qu'Airbus, Thales ou Collins, ainsi que des autorités comme l'EASA et la FAA, et des laboratoires, afin de cadrer l'usage de l'IA en aviation, un secteur où la certification est complexe du fait du caractère non-déterministe de ces technologies. L'objectif est d'élaborer des guides et normes de référence pour l'intégration, l'évaluation et la certification de l'IA dans les systèmes aéronautiques.

Cibles et portée

Systèmes embarqués (avions, drones) et systèmes au sol (ATM/ANS), avec une portée européenne et internationale.

Description

Groupe piloté par Airbus, Thales et Collins Aerospace. Travaux organisés en sous-groupes (applications, gestion des données ML, vérification, sécurité, facteurs humains).

Résultats

- Le Statement of Concerns (ER-022, 2021) a permis de formaliser les principales préoccupations liées à l'usage de l'IA en aviation.
- La Taxonomy in AI (ER-027, 2024) défini un cadre structuré pour catégoriser les différents types d'IA.
- Les Use Cases Considerations (attendu en 2026) guide l'évaluation des applications concrètes de l'IA dans le secteur aéronautique.
- Les Process Standard for Development and Certification/Approval of aeronautical Products implementing AI (ED-324, attendu pour 2026)

Conclusion et perspectives

Le WG-114 constitue une initiative clé pour bâtir un cadre normatif robuste autour de l'IA aéronautique. La publication de l'ED-324 sera déterminante pour consolider la confiance et l'acceptabilité réglementaire de l'IA en Europe.

5.1.3 Références de publications Thales

1. Rapport conjoint Thales & CESIN – L'IA & les RSSI

Date : Mars 2025

Participants : Thales, CESIN

Description : Livre blanc restituant une étude auprès des RSSI membres du CESIN, explorant l'impact de l'IA sur leur métier et les enjeux de gouvernance associés. Il propose défis et recommandations pour l'intégration de l'IA tout au long du cycle de vie des projets.

2. Livre blanc TAID – Trustworthiness for AI in Defence

Date : Mai 2025

Participants : Pôle d'Excellence Cyber, EDA, AMIAD, Thales, Capgemini, industriels européens, experts académiques et institutionnels

Description : Document élaboré par le groupe de travail TAID, visant à analyser et recommander les principaux aspects d'un développement de systèmes d'IA responsables, éthiques et fiables pour la défense européenne.

3. Projet DEEL – Dependable & Explainable Learning

Date : Mars 2021

Participants : IRT Saint-Exupéry, ANITI (Toulouse), IVADO et CRIAQ (Montréal), industriels (aéronautique, spatial, automobile) comme Thales, Airbus…

Description : Consortium de 60 experts (chercheurs, data scientists, ingénieurs en sécurité et certification) répartis entre Toulouse et le Québec.

Objectif : Développement des méthodes garantissant fiabilité et explicabilité de l'IA pour les systèmes critiques industriels.

4. Ouvrage collectif – IA de confiance dans la défense

Date : Juin 2025

Participants : Pôle d'Excellence Cyber, AMIAD, Thales, Capgemini, industriels européens, experts académiques et institutionnels

Description : Publication collective analysant les enjeux stratégiques, techniques, éthiques et juridiques de l'IA de confiance, au service de la souveraineté numérique et militaire.

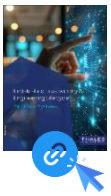
5. Position Paper – IA des systèmes critiques

Date : Novembre 2024

Participants : Thales

Description : Document présentant la vision de l'IA pour les systèmes critiques comme appui maîtrisé à la décision humaine, intégrant des exigences élevées de sécurité, de robustesse, de gouvernance et de conformité réglementaire.

6. Position Paper – Cycle de vie d'ingénierie IA fiable de bout en bout pour les systèmes critiques



Date : Janvier 2026

Participants : Thales

Description : Ce document propose une vision de l'IA dans les systèmes critiques, où fiabilité, sécurité et transparence sont indispensables. Il montre comment Thales conçoit ses solutions de bout en bout, de la donnée à l'algorithme, jusqu'à la validation et l'usage opérationnel tout en suivant la chaîne de confiance industrielle.

6. Conclusion et perspectives

L'IA de confiance ne se limite pas à une exigence de conformité ; elle est une **invitation à repenser nos approches d'innovation et de collaboration.** Elle représente une **démarche proactive et systémique** intégrant des principes éthiques, sociaux et environnementaux au cœur de son développement et de son utilisation.

L'enjeu n'est pas de définir seul un modèle, mais de contribuer à une démarche collective, évolutive et inclusive, afin de construire une **l'IA éthique et responsable.** Cela se traduit par l'adaptation de nos modes de gouvernance, le renforcement du dialogue entre acteurs publics, privés et académiques, et en définissant **la responsabilité des parties prenantes** tout le long du cycle de vie des systèmes d'IA. Clarifier la responsabilité, la redevabilité et l'imputabilité des acteurs lors du développement et du déploiement des systèmes d'IA, est le cœur de l'IA de confiance, afin que les parties prenantes et les utilisateurs comprennent leurs droits et obligations.

La mise en œuvre de nouvelles régulations et l'accélération des usages de l'IA appellent à une **vigilance et réflexion continue pour poursuivre la construction d'IA de confiance**, afin de répondre aux challenges liés aux évolutions technologiques et sociétales. L'émergence d'agents IA, capables d'exécuter des tâches complexes de manière autonome et de dialoguer avec d'autres systèmes, représente un exemple de ces nouveaux challenges.

Ces systèmes exigent, en effet, une supervision renforcée afin de garantir que leurs actions demeurent alignées **avec l'intention humaine et les valeurs sociétales.** Ce contexte met d'autant plus en exergue **la question de la responsabilité :** celle des concepteurs dans la définition du besoin, des limites et des usages possibles, et celle des utilisateurs dans leur intérêt et la maîtrise des conditions d'emploi.

7. A propos de Thales

Aujourd'hui, Thales rassemble plus de 800 experts en IA au sein de son organisation interne dédiée, cortAIx, créée en 2024. Ces experts sont répartis en trois équipes principales :

- **Labs** : axés sur la recherche fondamentale et le développement de modèles.
- **Factory** : responsables des chaînes d'outils, de la cybersécurité et des infrastructures critiques.
- **Sensors** : intégrant l'IA dans les systèmes embarqués et edge.

Ensemble, ces équipes garantissent que les capacités d'IA sont non seulement à la pointe de la technologie, mais aussi opérationnellement viables, sécurisées et conformes aux exigences de souveraineté. Les équipes cortAIx de Thales sont stratégiquement réparties entre la France, le Royaume-Uni, le Canada, Singapour, l'Allemagne et les Émirats arabes unis, reflétant l'empreinte mondiale de l'entreprise et son engagement à soutenir la souveraineté régionale grâce à une expertise locale.

cortAIx
Artificial Intelligence by THALES

8. Ressources

Niveau Reglementaire / Normes

Conseil de l'Europe – Convention-cadre sur l'IA	Lien
European AI Act, *2024*	Lien
RGPD, *2016*	Lien
European Cyber Resilience Act, 2024	Lien
CEN-CENELEC JTC 21 - EU AI Act Standards, *2025*	Lien
ISO/IEC 42001 - AI Management system, *2023*	Lien
ISO/IEC 38507 - Governance implications of the use of artificial intelligence by organizations,*2022*	Lien
ISO/IEC TR 24028 - Overview of trustworthiness in artificial intelligence, *2020*	Lien
ISO/IEC 42005 - AI system impact assessment, *2025*	Lien
ISO/IEC CD 42105 - Guidance for human oversight of AI systems (under development)	Lien
ETSI - Technical Committee (TC) Securing Artificial Intelligence (SAI) TR 104, *2024-2025*	Lien
ISO/IEC 27001 - Information security management systems - Requirements, *2022*	Lien
ISO/IEC 23894 – AI Guidance on risk management, *2023*	Lien
Loi AGEC, *2020*	Lien
Loi REEN, *2021*	Lien
Loi Climat et Résilience, *2021*	Lien

Niveau recommandations

OCDE - Principles for Trustworthy AI, *2024*	Lien
Commission Européenne – Assessment List for Trustworthy Artificial Intelligence (ALTAI), *2021*	Lien
IEEE - Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, *2019*	Lien
HLEG UE - Ethics Guidelines for Trustworthy AI, *2019*	Lien
CNIL - Garantir la sécurité du développement d'un système d'IA, *2024*	Lien
ANSSI - Recommandations de sécurité pour un système d'IA générative, *2024*	Lien
ENISA - Securing Machine Learning Algorithms, *2021*	Lien
NIST - AI Risk Management Framework, *2023*	Lien
CNIL - IA et RGPD, *2025*	Lien
CNIL - Les fiches pratiques IA, *2024*	Lien
UNESCO - Ethics of Artificial Intelligence, *2021*	Lien
NIST - Four Principles of Explainable Artificial Intelligence, *2021*	Lien
CNPEN – Avis n°7 : Systèmes d'intelligence artificielle générative : enjeux d'éthique, *2023*	Lien
INR - RIA31 Guide IA Ethique et Responsable, *2024*	Lien
Arcep / Arcom / ADEME - Référentiel général d'écoconception de services numériques (RGESN), *2024*	Lien
AFNOR / Ecolab - Spec 2314 - IA Frugale, *2024*	Lien
CEN-CENELEC - Standards in support of trustworthy environmental claims, *2025*	Lien
Paris AI Action Summit - Standardization for AI Environmental Sustainability, *2025*	Lien

Niveau Professionnel

Hub France IA - Livre Blanc IA éthique en pratique (2e version), *2024*	Lien
Hub France IA / Wavestone – Notice Premier pas vers l'IA de confiance, 2025	Lien
Impact AI - Les briefs de l'IA Responsable, *2024*	Lien
CIGREF / Numeum - Guide de mise en œuvre de l'AI Act, *2025*	Lien
Confiance.ai - Livre Blanc « Towards the engineering of trustworthy AI applications for critical systems », *2022 et 2024*	Lien
ITRE (European Parliament) - Interplay between the AI Act and the EU digital legislative framework, *2025*	Lien
Lefebvre Dalloz - AI ACT, le nouveau cadre juridique européen de l'IA, *2025*	Lien
Microsoft - Livre blanc « Accelerate AI transformation with strong security », *2022*	Lien
Microsoft - IA sécurisée : processus pour sécuriser l'IA, *2025*	Lien
Microsoft – Responsible AI Impact Assessment Template, *2022*	Lien
Google – Secure AI Framework, 2023	Lien
The Digital Economist – The ROI of AI Ethics, *2025*	Lien
Numeum - Ethical AI (livret pédagogique), *2024*	Lien
Numeum / G9+ / CIGREF / Planet Tech Care / Hub France IA – AI for Green & Green AI, *2024*	Lien
Green IT - Impacts environnementaux et sanitaires de l'IA dans le monde en 2025, *2025*	Lien
Green IT - Référentiel de maturité/ 74 bonnes pratiques pour un numérique plus responsable, *2022*	Lien
The Shift Project - Intelligence artificielle, données, calculs : quelles infrastructures dans un monde décarboné, *2025*	Lien

9. Remerciements

La réalisation de cet article, résultat d'une exploration approfondie des enjeux et des perspectives de l'Intelligence Artificielle, n'aurait pas été possible sans la contribution et l'engagement de la communauté d'experts Thales.

Nous souhaitons exprimer notre gratitude aux professionnels, chercheurs, et experts qui ont accepté de partager leur savoir et leur vision, contribuant ainsi de manière significative à l'élaboration de cet ouvrage. Leur expertise a permis d'offrir une perspective polyvalente et nuancée sur un sujet en constante évolution.

Nous remercions tout particulièrement nos experts Thales :

- Pour leurs expertises en IA et leurs contributions transverses :

Thomas **DELAVALLADE**
Benoit **HUYOT**
Nicolas **MUSEUX**

- Pour les analyses sur les aspects de Sécurité, Safety et Cyber de l'IA :

Olivier **BETTAN**
Joseph **MACHROUH**
Marie **KEBALO**

- Pour les éclairages sur les aspects de robustesse, de transparence, d'explicabilité et de confiance dans l'IA :

Patricia **BESSON**
Florence **DE GRANCEY**
Nick **ERNEST**

- Pour les contributions sur les aspects éthiques de l'IA :

Benoit **HUYOT**
Pierre-Etienne **LEGOUX**

- Pour les connaissances sur les aspects de durabilité de l'IA :

Remi **BARRIERE**
Thomas **HEUSSE**
Aurore **TUAL**
François **MORIAMEZ**
Marc **HEUDE**
Pierre **GIRARD**

- Pour les éclairages sur les aspects juridiques :

Victor **CART**

Leur précision et leur rigueur ont été essentielles pour assurer l'intégrité et l'actualité des informations présentées.

Nos remerciements s'adressent également à notre leader de practice conseil Data & AI Jean-Paul **LEROUX** ainsi que, pour leur support apporté à ce projet, à :

Hélène **BRINGER**

Juliette **MATTIOLI**

Juliette **ROUILLOUX-SICRE**

Et David **SADEK**

THALES

Building a future we can all trust

4, rue de la Verrerie 92190
Meudon FRANCE

Tél. + 33(0)1 57 77 80 00

www.thalesgroup.com