Pour une IA de confiance : enjeux juridiques, éthiques et opérationnels

fntc

Présentation de la FnTC :

La Fédération des Tiers de Confiance du numérique (FnTC) rassemble éditeurs de logiciels, prestataires de services, experts, professionnels réglementés, start up, acteurs internationaux, utilisateurs et structures institutionnelles.

Notre objectif depuis 2001 : **une digitalisation fiable et sécurisée.**

Notre méthode :

- Produire des expertises et des outils pour que les personnes et les organisations puissent au sein du monde numérique préserver leurs droits et limiter leurs risques.
- Elaborer de la doctrine, en produisant des guides, des référentiels et des labels.
- Participer à la normalisation et à la standardisation des bonnes pratiques numériques au niveau national (Afnor) et international (ISO)
- Assurer des formations universitaires, comme les Masters Droit du numérique des Universités de Corse, de La Rochelle et de Lyon, ainsi que de la formation continue.

fntc

Fédération des Tiers de Confiance du numérique

Sommaire

Introduction :

L'orientation de l'Intelligence Artificielle (IA) vers le bien commun constitue un socle essentiel de la confiance numérique, dépassant la stricte conformité réglementaire pour interroger la finalité même des technologies déployées.

Le Règlement (UE) 2024/1689 sur l'IA (AI Act ou RIA) incarne cette ambition en établissant un cadre juridique harmonisé qui promeut une IA « axée sur l'humain et digne de confiance », tout en garantissant un niveau élevé de protection de la santé, de la sécurité et des droits essentiels consacrés dans la Charte des droits fondamentaux de l'Union européenne, y compris la démocratie, l'État de droit et la protection de l'environnement (considérant 1). L'IA doit ainsi rester un outil au service des personnes, visant à améliorer le bien-être humain, et non à s'y substituer ou à porter atteinte à l'autonomie individuelle.

L'IA est définie dans ce Règlement (UE) 2024/1689 comme *« un système fondé sur des machines, conçu pour fonctionner avec différents niveaux d'autonomie et qui, pour des objectifs explicites ou implicites, infère à partir des entrées reçues comment générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions pouvant influencer des environnements physiques ou virtuels »* (considérant 12).

Le règlement met en garde contre les usages dévoyés de l'IA, notamment les pratiques de manipulation, d'exploitation ou de contrôle social, qui sont explicitement interdites car contraires aux valeurs de l'Union, telles que le respect de la dignité humaine, la liberté, l'égalité, la démocratie et l'État de droit, ainsi qu'aux droits fondamentaux comme la non-discrimination, la protection des données, la vie privée et les droits de l'enfant (considérants 27 à 29). Cela suppose que les systèmes d'IA soient conçus pour prévenir la perte du libre arbitre humain, l'homogénéisation des sociétés, et pour favoriser la diversité et l'inclusivité, conformément à l'approche fondée sur les risques et à la nécessité d'une supervision humaine effective (considérant 27).

Le cadre du Règlement IA, en interdisant certaines pratiques inacceptables (pratiques manipulatrices, notation sociale, catégorisation biométrique sur des critères sensibles), en posant des exigences strictes pour les systèmes d'IA (SIA) à haut risque, et en exigeant transparence, robustesse et contrôle humain, vise ainsi à instaurer une IA de confiance, au service des citoyens et des valeurs fondamentales de l'Union européenne.

L'IA de confiance, dans cette perspective, ne se limite pas à la performance technique : elle implique que les systèmes soient conçus, déployés et gouvernés dans le respect des valeurs fondamentales, de la transparence, de l'explicabilité, de la sécurité et de la responsabilité, tout en anticipant et limitant les usages détournés ou les biais discriminatoires.

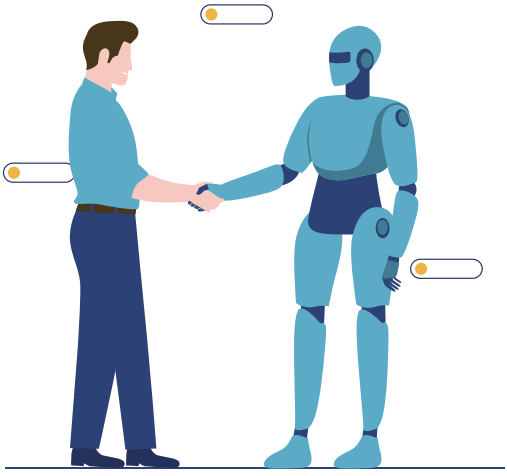
En ce sens, l'IA de confiance ne se limite pas à la performance technique, cela implique des enjeux de souveraineté et de maîtrise technologique.

Le législateur européen insiste sur la capacité d'inférence, d'apprentissage et d'adaptation de ces systèmes, qui les distingue des logiciels traditionnels, et sur la nécessité d'un cadre harmonisé pour garantir une IA axée sur l'humain et digne de confiance, assurant un haut niveau de protection de la santé, de la sécurité et des droits fondamentaux consacrés dans la Charte, y compris la démocratie, l'État de droit et la protection de l'environnement (considérants 1 et 27).

En ce sens**, une IA de confiance est une IA dont la conception, la mise en œuvre et l'utilisation respectent les exigences de transparence, de robustesse technique, de gouvernance des données, de supervision humaine, de non-discrimination et de responsabilité,** telles que posées par le Règlement IA et les lignes directrices éthiques européennes.

Elle doit **permettre à tout moment la compréhension et le contrôle de ses décisions, garantir la traçabilité de ses processus, prévenir les effets indésirables, et rester alignée sur les intérêts et les droits des personnes.** L'IA de confiance s'inscrit donc dans une démarche proactive de maîtrise du risque, d'anticipation des usages, et de respect des principes éthiques et juridiques, condition sine qua non pour instaurer une confiance durable dans l'innovation algorithmique, au service du bien commun et des valeurs européennes.

Le règlement européen sur l'IA ("AI Act" ou "RIA") du 13 juin 2024 est entré en vigueur le 1er août 2024. Il pose principalement des obligations applicables aux systèmes d'IA « à haut risque » (les plus sensibles) et aux modèles d'IA à usage général (modèles de fondation de type LLM), ainsi que des obligations de transparence pour les systèmes en interaction avec les personnes (ex. : chatbot). La plupart de ces obligations entreront en application le **2 août 2026.**



1

1. Le cadre juridique de l'IA

1.1. Le règlement européen sur l'IA

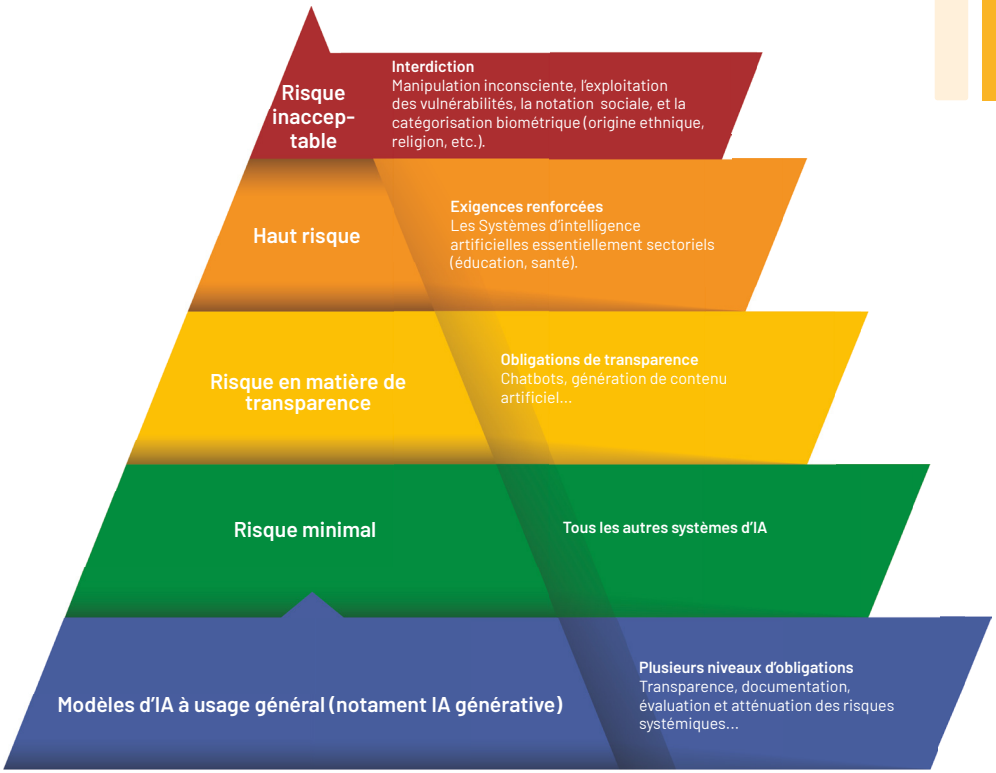
Avec l'entrée en vigueur du Règlement sur l'IA, l'Union européenne franchit une étape majeure vers une régulation fondée sur la confiance, en imposant un haut niveau de protection des droits fondamentaux dans le développement et le déploiement des systèmes d'intelligence artificielle à haut risque.

Le texte introduit notamment une logique d'évaluation de l'impact de ces dispositifs d'IA, et s'inscrit dans une démarche d'anticipation des risques centrée sur les personnes, au regard des niveaux de risque des dispositifs d'intelligence artificielle.

Le RIA prévoit 4 catégories pour classer les SIA en fonction de leur niveau de risque :

- SIA interdits (Art. 5) : l'utilisation de ces systèmes n'est pas autorisée, le risque associé à leur utilisation étant inacceptable ;
- SIA à haut risque (Art. 6) : il s'agit du niveau de risque pour lequel le RIA prévoit le plus d'exigences, en matière de documentation, de qualité des données, de performance, de sécurité, de mise en place d'un système de gestion et d'atténuation des risques et présence d'un contrôle humain dans le système.
- SIA avec obligations particulières de transparence (Art. 50) : exigences en matière de transparence pour l'utilisateur.
- SIA à risque minime : pas d'obligations.

Règlement IA – Approche par les risques



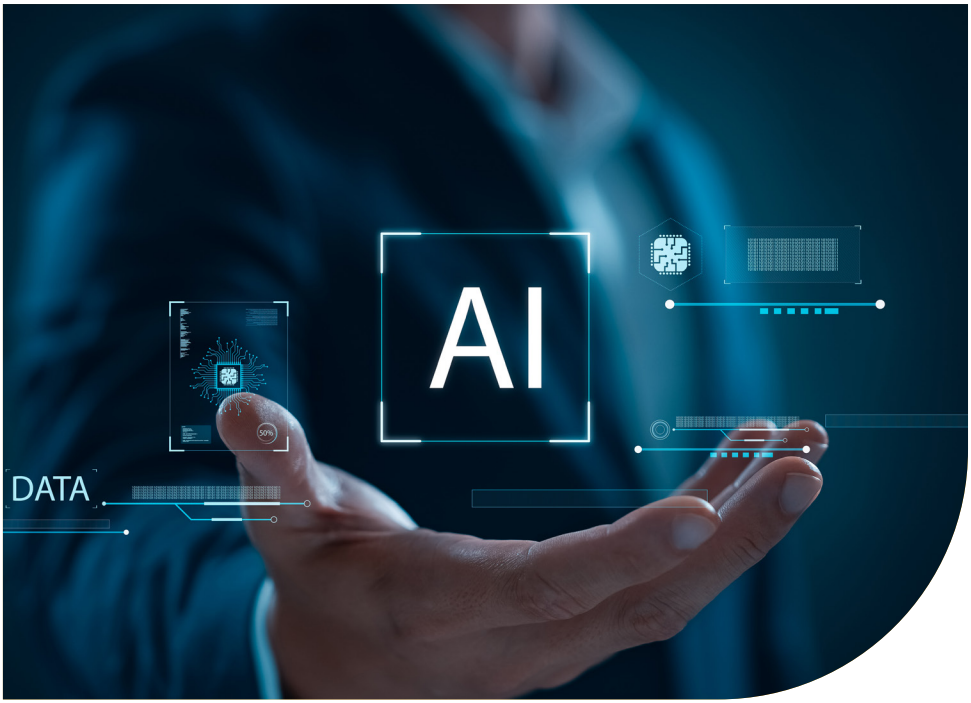
1.2. Articulation entre le Règlement IA et le RGPD

Le Règlement européen sur l'IA présente des points communs avec l'analyse d'impact relative à la protection des données prévue à l'article 35 du Réglement général de protection des données (RGPD) : toutes deux relèvent d'une approche ex ante, doivent être actualisées en cas de modification substantielle, et reposent sur une évaluation des risques fondée sur les droits. Mais leur périmètre diffère : le RGPD est centré avant tout sur la protection des données personnelles, tandis que le RIA couvre l'ensemble des droits fondamentaux consacrés par la Charte de l'UE – y compris la dignité, la liberté d'expression, la non-discrimination, les droits des travailleurs, ou encore la protection des enfants et de l'environnement (cf. article 1er et considérant 48 du RIA).

Si le RIA et le RGPD reposent sur un principe d'approche par les risques, toutefois ces textes diffèrent sur les points suivants :

	RGPD	RIA
Appréhension des risques	Critères attachés aux risques (nature des données, finalité poursuivie)	Catégorisation des Systèmes d'Intelligence artificielle (SIA) en fonction de leur niveau de risque (les pratiques interdites, les systèmes d'IA à haut risque [atteinte à la sécurité physique des personnes ou à leurs droits fondamentaux] et ceux à risque limité) et liste pour les IA à haut risque (annexe III)
Définition des acteurs	Définition liée au cycle de vie des données (fournisseur RT/ST et déployeur RT)	Définition liée au cycle de fabrication et de mise sur le marché des technologies (fournisseurs/déployeurs)
Obligations prévues	Gouvernance des données et obligations de sécurité, périmètre des exigences plus large	Gouvernance interne de la conformité des produits, fondée sur la documentation technique, la gestion des risques et de la qualité et la traçabilité et l'information des déployeurs
Transparence	Vis-à-vis de la personne concernée (utilisateurs finaux)	Dans la relation entre professionnels
Droits reconnus aux utilisateurs finaux	Droit à l'information (art. 13 §2 f), traçabilité (art. 14)	Droit à l'explication du SIA (attaché au droit à l'information), droit à la traçabilité des sources
Analyse d'impact	Sur les risques liés à la disponibilité, l'intégrité et la confidentialité des données	Sur les droits fondamentaux (dignité, liberté, égalité, justice…)
Finalité des traitements de données personnelles	Détermination de la finalité (art. 5 §1 b - CNIL): Finalité définie dès la phase de développement i.e finalité déterminée en amont du projet, compréhensible (explicite) et en lien avec l'activité de l'organisme (légitime) => même finalité pour le déploiement (unique) Finalité non définie dès la phase de développement (ex. SIA à usage général) => conditions cumulatives à respecter telles que l'identification des fonctionnalités et des capacités techniquement envisageables	Transparence requise sur la finalité initiale de la collecte des données. (Considérant 67 du RIA et Art. 10 §2 -b sur la gouvernance des données) Respect du principe de limitation des finalités (Considérant 94 : l'utilisation de systèmes d'IA à des fins d'identification biométrique de nature répressive) Un modèle d'IA peut répondre à diverses finalités tant pour une utilisation directe que pour une intégration dans d'autres systèmes d'IA (Art. 3-66 : définition SIA)

	RGPD	RIA
Principe de licéité/ base légale des traitements de données personnelles	Consentement : mise en œuvre difficile (retrait) Respect d'une obligation légale : le RIA ne constitue pas cette base légale pour la mise en œuvre des SIA (cf. dérogation art. 22 2 du RGPD non applicable) sauf pour le développement de certains systèmes d'IA d'intérêt public dans le cadre du bac à sable réglementaire de l'IA (sous réserve du respect des conditions de l'art. 59 RIA) La poursuite d'un intérêt légitime : conditions à respecter (Legitimate Impact assesment)	Aucune référence du RIA – Le principe de licéité posé par le RGPD s'applique dès lors que le SIA traite de données à caractère personnel



Les critères de l'IA de confiance

2

2.1. Transparence

La transparence désigne l'idée que toutes les parties prenantes affectées par le résultat d'un système d'IA devraient pleinement comprendre le fonctionnement interne de ce système, depuis son développement, son entraînement et son déploiement jusqu'aux facteurs qui influencent ses décisions.

Ainsi c'est le cycle de vie de l'IA qui doit être pris en compte pour la transparence, des données d'entraînement jusqu'au résultat issu des données fournies à l'algorithme. La réglementation européenne intègre les chaînes de responsabilité, dans le cas où plusieurs acteurs interviennent.

Cela implique que les autorités de régulation soient en mesure, sous certaines conditions, d'examiner le code source, les données de fonctionnement et les critères de décision des algorithmes, dans le cadre des IA à haut risque.

Pour ces systèmes d'IA, le principe de transparence devra s'appliquer sur plusieurs éléments :

→| La transparence des données :

Cette disposition vise à s'assurer de l'exactitude et de la licéité des données utilisées par le système d'intelligence artificielle, et à lutter contre la discrimination qui pourrait résulter de biais.

L'utilisation des données doit en premier lieu répondre aux principes de transparence du RGPD notamment lors de leur utilisation pour l'apprentissage de l'algorithme, même si elles ne sont pas identifiables dans les données de sortie. Au titre de l'article 12 du RGPD précisé aux articles 13 et 14, le responsable de traitement doit informer la personne concernée des données ou catégories de données à caractère personnel concernées, ainsi que de l'identité du responsable du traitement et des finalités du traitement auquel sont destinées les données à caractère personnel ainsi que la base juridique du traitement. Des exceptions existent en cas de collecte indirecte.

Le Règlement européen sur l'IA va au-delà et introduit une exception aux règles de protection des données, dans un objectif de correction des biais. Elle prévoit que fournisseurs de systèmes d'IA à haut risque puissent traiter **aux fins de la détection et de la correction des biais, des catégories particulières de données à caractère personnel.**

Une documentation devra être établie par le fournisseur de modèle d'IA à usage général, et tenue à disposition des autorités de régulation.

La transparence des algorithmes

Le Règlement européen sur l'IA prévoit des obligations de transparence des algorithmes tant pour les IA à usage général, que pour les IA à haut risque.

Néanmoins, le RIA prend en compte les enjeux liés à la propriété intellectuelle. Ainsi les autorités de surveillance et la Commission européenne sont soumises au respect de la confidentialité des informations transmises et de ces règles.

Par ailleurs, des obligations moindres de transparence devront également être transmises, via une documentation technique moins détaillée sur l'algorithme aux "déployeurs", définis dans le Considérant 13 comme *"une personne physique ou morale, une autorité publique, une agence ou un autre organisme utilisant un système d'IA sous son autorité, sauf si le système d'IA est utilisé dans le cadre d'une activité personnelle non professionnelle"* .

Deux obligations pèsent sur les fournisseurs :

1. Elaborer et tenir à jour la documentation technique du **modèle de l'IA pour les autorités de contrôle.**

2. Elaborer et tenir à jour les informations et la **documentation pour les fournisseurs d'IA** qui souhaitent intégrer le modèle dans leur système d'IA, en tenant compte des conditions de confidentialité.

Le Code de bonnes pratiques de l'IA* prévoit de faire figurer les informations suivantes :

- Information générales (nom du fournisseur, famille de modèle, mise sur le marché…)
- Propriétés du modèle (architecture, modalités des entrées et des sorties…)
- Méthodes de distribution et de licence (canaux de distribution, licences…)
- Utilisation du modèle (politique d'utilisation pour des usages acceptables, usages envisagés. Types et nature des systèmes d'IA dans lesquels le modèle d'IA à usage général peut être intégré
- Processus d'entrainement (description générale des principales étapes du processus d'entraînement, spécification technique, paramètres…)
- Information sur les données d'entrainement, de test et de validation (type, **nombre de points de données**, comment les données ont été obtenues, les mesures liées aux données personnelles…)
- Ressources informatiques (hardware)
- Consommation énergétique.
- Mesures additionnelles pour les fournisseurs d'IA à risque systémique (évaluation, adaptation du modèle, architecture, test)

*Ces différents éléments sont détaillés dans le "Code de bonnes pratiques de l'IA à usage général", publié par la Commission européenne.

→| La transparence vis-à-vis des utilisateurs finaux

La législation européenne sur l'IA introduit également des obligations de transparence, dans les cas suivants :

- Interaction avec une IA :

Les personnes physiques concernées doivent être informées qu'elles interagissent avec un système d'IA, sauf si cela ressort clairement du contexte d'utilisation. Ainsi une information doit être faite pour l'utilisation des chat-bots utilisant de l'IA.

- Résultat produit par une IA :

Les fournisseurs de systèmes d'IA, y compris de systèmes d'IA à usage général, qui génèrent des contenus de synthèse de type audio, image, vidéo ou texte, veillent à ce que les sorties des systèmes d'IA soient marquées dans un format lisible par machine et identifiables comme ayant été générées ou manipulées par une IA.

En pratique, cela implique pour ces fournisseurs de mettre en œuvre des mécanismes techniques d'identification fiables, robustes et efficaces pour assurer leur traçabilité tout au long de leur cycle de vie. L'AI Office publiera les guidelines correspondantes en juin 2026.

Des exceptions existent en cas d'utilisation de l'IA pour la prévention ou la détection des infractions pénales, d'enquêtes ou de poursuites en la matière, ou lorsque le contenu généré par l'IA a fait l'objet d'un processus d'examen humain ou de contrôle éditorial et lorsqu'une personne physique ou morale assume la responsabilité éditoriale de la publication du contenu.

2.2. Explicabilité et auditabilité

L'IA de confiance repose sur sa capacité à être comprise, auditée et tracée tout au long de son cycle de vie. Cela implique des mécanismes d'explicabilité des algorithmes, une documentation rigoureuse des décisions prises par l'IA, et une preuve de continuité assurant la traçabilité des actions entreprises par le système d'IA. Pour garantir cette explicabilité et auditabilité, plusieurs axes doivent être privilégiés :

- **Une documentation des algorithmes et des modèles**

Les modèles IA doivent être documentés de manière exhaustive, incluant la méthodologie de conception, les jeux de données utilisés, les transformations appliquées et les résultats attendus. Chaque mise à jour doit être consignée dans un registre accessible, permettant d'assurer la traçabilité des versions et des modifications apportées. L'enjeu sera ici d'avoir la capacité de pouvoir démontrer à tout moment les fondements d'une décision prise par une IA. Cela rejoint les principes de gestion documentaire où chaque version doit être conservée et accessible en cas d'audit ou de litige.

- **Des preuves de continuité**

La continuité des systèmes IA doit être garantie par des mécanismes de conservation des preuves numériques (comme l'archivage des versions de modèles et logs des actions critiques).

Il s'agit ainsi de s'assurer que chaque décision automatisée soit archivée de manière à pouvoir être réexaminée en cas de litige ou de demande d'audit.

Une approche "document management" doit également intégrer des éléments de traçabilité algorithmique : chaque décision doit être associée à un jeu de données et à une version d'algorithme spécifique.

- **Une explicabilité des décisions**

Les modèles IA doivent être accompagnés d'outils d'explication des décisions permettant de retracer le raisonnement de l'IA. Cette démarche est essentielle dans des secteurs à haut risque (RH, médical, énergie, financier) où les décisions peuvent avoir des impacts majeurs sur les individus. Les solutions d'explicabilité doivent également être adaptées aux profils des utilisateurs : des visualisations simplifiées pour les métiers, des rapports techniques pour les data scientists. Cela pose la nécessité de conserver des journaux pour potentiellement pouvoir répondre à des demandes d'explicabilité ou des contestations. Dans le cadre du Règlement européen sur l'IA, ces outils permettent également de démontrer la conformité des modèles en cas d'audit par les autorités compétentes.

2.3. Fiabilité et robustesse

La robustesse est citée dans l'Annexe IV du Règlement européen sur l'IA comme un élément déterminant.

La validation de la fiabilité des SIA devra passer par des exercices de tests rigoureux (matrice de confusion, validation croisée, injection de données trompeuses ...) avant le déploiement des systèmes.

Données de tests, tests automatisés, tests de non-régression, inventaire de la configuration de tests, logiciel, librairies tierces, résultats des tests, devront être archivés / séquestrés chez une autorité tiers indépendante.

Malgré ces différentes mesures mises en place, les phénomènes d'hallucinations restent un défi majeur et demanderont une vigilance constante.

2.4. Souveraineté

La souveraineté numérique peut se définir comme la capacité à maîtriser l'ensemble des technologies sur toute la chaine de valeur de la donnée et à identifier son degré de dépendance à des technologies extra-européennes.

Les dispositifs d'IA sont d'autant plus au cœur des enjeux de souveraineté qu'ils exploitent des données pouvant être sensibles (personnelles, économiques...). Au-delà des moteurs d'IA, il convient également de s'assurer que les données sont hébergées dans des environnements souverains, non soumis aux règles d'extraterritorialité.

Ainsi, la souveraineté et l'identification de son niveau de dépendance sont nécessaires pour garantir la sécurité des données. Plus largement, l'intégralité de ce qui compose l'IA doit être sous contrôle : les investissements en recherche & développement, le matériel, données et algorithmes, réseaux, logiciels, hébergement...

De nombreuses initiatives sont actuellement en cours visant à définir les critères de dépendance, via l'Observatoire du Numérique et le CSF Numérique de confiance.

2.5. Sobriété

La question de la soutenabilité environnementale de l'intelligence artificielle s'impose désormais comme une priorité. Il convient ainsi de s'interroger de manière systématique sur la pertinence de recourir à l'IA.

Lorsque le recours à l'IA s'avère nécessaire, la sobriété doit guider chaque étape de sa mise en œuvre : concevoir des systèmes économes avec une puissance de calcul et un volume de données utilisées adaptés aux besoins.

Cela nécessite également la mise en place d'une gouvernance interne, et d'interroger ses fournisseurs et prestataires sur leurs engagements en matière de politique environnementale.

Adopter une approche sobre de l'intelligence artificielle contribue directement à renforcer la confiance des utilisateurs dans ces technologies.

2.6. Supervision humaine

L'intégration du principe de la supervision humaine est une pierre angulaire du Règlement européen sur l'IA et constitue une exigence fondamentale pour garantir une intelligence artificielle digne de confiance. Il s'agit d'une finalité essentielle permettant d'assurer que l'IA ne se substitue jamais à la responsabilité et à l'autonomie humaines et demeure un outil au service des personnes.

La supervision humaine vise à limiter la délégation de pouvoir à la technologie et à préserver le libre arbitre.

Ainsi, il exige que les systèmes d'IA à haut risque (santé, justice, éducation, emploi, accès aux services essentiels), restent sous contrôle humain effectif, avec des mécanismes de supervision, de validation et de recours.

Sur le plan pragmatique, le contrôle humain effectif sur les sorties (outputs) des systèmes d'IA est indispensable pour éviter que des erreurs d'appréciation, des biais ou des hallucinations produisent des conséquences dommageables, parfois irréversibles, permettant ainsi à l'utilisateur de valider ou corriger le résultat algorithmique.

2.7. Finalité et Notice d'utilisation

Le Règlement européen sur l'IA indique que le fournisseur d'un SIA s'engage à transmettre une notice d'utilisation. Cette notice joue un rôle central pour garantir la transparence, la sécurité et la conformité des systèmes d'IA, en particulier des systèmes d'IA à haut risque. En imposant cette obligation, le RIA permet aux déployeurs de faire un choix éclairé et contribue à la responsabilisation des fournisseurs.

Cette notice d'utilisation comprend :

- L'identité et coordonnées du fournisseur du système d'IA et éventuellement de son mandataire,
- La destination du système d'IA
- Le niveau d'exactitude des résultats y compris les informations relevant de la robustesse et de la cybersécurité
- Les circonstances connues ou prévisibles liées à l'utilisation du système d'IA susceptibles d'entraîner des risques pour la santé et la sécurité ou pour les droits fondamentaux
- Les capacités et caractéristiques techniques du système d'IA à fournir des informations pertinentes pour expliquer ses sorties
- La performance du système d'IA en ce qui concerne des personnes ou groupes de personnes spécifiques à l'égard desquels le système est destiné à être utilisé
- Les spécifications relatives aux données d'entrée, ou toute autre information pertinente concernant les jeux de données d'entraînement, de validation et de test utilisés
- Les informations permettant aux déployeurs d'interpréter les sorties du système d'IA à haut risque et de les utiliser de manière appropriée
- Les modifications du système d'IA à haut risque et de sa performance qui ont été prédéterminées par le fournisseur au moment de l'évaluation initiale de la conformité
- Les mesures de contrôle humain
- Les mesures techniques mises en place pour faciliter l'interprétation des sorties des systèmes d'IA à haut risque par les déployeurs
- Les ressources informatiques et matérielles nécessaires
- La durée de vie attendue du système d'IA à haut risque
- Toutes les mesures de maintenance et de suivi, y compris leur fréquence, nécessaires pour assurer le bon fonctionnement de ce système d'IA, notamment en ce qui concerne les mises à jour logicielles
- Une description des mécanismes compris dans le système d'IA à haut risque qui permet aux déployeurs de collecter, stocker et interpréter correctement les journaux.

2.8. Non-discrimination et respect de la personne humaine

L'IA ne doit pas devenir un instrument de standardisation des comportements ou des valeurs, mais au contraire respecter la diversité des individus et des cultures. Le maintien de la diversité, la préservation de la dignité humaine et la non-discrimination constituent ainsi des critères fondamentaux pour une IA de confiance.

Le Considérant 27 du RIA indique ainsi que "les systèmes d'IA sont développés et utilisés comme un outil au service des personnes, qui respecte la dignité humaine et l'autonomie de l'individu".

Ainsi, mesurer la capacité d'un système d'IA à respecter et favoriser les différences individuelles et culturelles devient une exigence centrale.

2.9. Éthique et équité

L'éthique et l'équité représentent deux piliers fondamentaux et complémentaires dans la construction d'une IA de confiance. Elles constituent non seulement une exigence morale et légale, mais aussi une condition indispensable à l'acceptabilité et à la confiance du grand public envers l'IA.

L'éthique recouvre l'ensemble des valeurs et principes moraux guidant la conception et l'usage des systèmes d'intelligence artificielle : elle exige la transparence des algorithmes, le respect strict de la vie privée, une prise en compte des biais des données d'entrainement et une attention particulière aux conséquences sociales, humaines et environnementales de ces technologies.

L'équité impose aux systèmes d'IA de garantir une impartialité rigoureuse, assurant que les résultats produits ne génèrent aucune forme de discrimination ou d'injustice envers un individu ou un groupe particulier et impose notamment une vigilance particulière quant à la sélection des données utilisées, en veillant à leur diversité et à leur représentativité.

Le Considérant 27 du RIA reprend ainsi les sept principes éthiques du GEHN IA pour que "*l'IA soit digne de confiance et saine sur le plan éthique" : "action humaine et contrôle humain; robustesse technique et sécurité; respect de la vie privée et gouvernance des données; transparence; diversité, non-discrimination et équité; bien-être sociétal et environnemental; et responsabilité*".

Assurer l'éthique et l'équité ne relève pas uniquement d'un défi technique, mais suppose un dialogue structuré entre ingénieurs, juristes, sociologues, experts des données, et représentants des utilisateurs finaux. Ce dialogue interdisciplinaire, associé à un cadre normatif clair et partagé, est indispensable.

2.10. Sécurité

Pour un système d'intelligence artificielle, il s'agira de signifier les incidents et vulnérabilités en matière de cybersécurité conformément aux exigences de la directive NIS2. Une politique de divulgation coordonnée de vulnérabilité doit être mise en œuvre par l'entreprise.

Par ailleurs, chaque SIA devrait disposer d'une politique de sécurité des systèmes d'information (PSSI), et l'entreprise devrait être dotée d'un plan de continuité d'activité (PCA) et d'un plan de reprise d'activité (PRA).

La norme ISO 27001 recense les mesures permettant de sécuriser un système d'intelligence artificielle.

2..11. Responsabilité

Le Règlement européen sur l'intelligence artificielle (RIA) ne traite pas directement de la question de la responsabilité juridique en cas de dysfonctionnement d'un système d'IA. Les États membres n'ont pas réussi à trouver un accord sur le sujet.
En l'absence de dispositions spécifiques à cet égard, la répartition des responsabilités entre le producteur, le fournisseur ou l'utilisateur du système d'IA reste soumise aux législations nationales. En France, c'est :

- Le régime de la responsabilité du fait des produits défectueux, tel que défini aux articles 1245 à 1245-17 du Code civil, mais ces dispositions ne s'appliquent pas en matière d'IA car non applicables aux biens incorporels. Au niveau européen, la Directive (UE) 2024/493 du Parlement européen et du Conseil du 21 février 2024 relative à la responsabilité du fait des produits défectueux abroge et remplace la Directive 85/374/CEE. Ce texte doit être transposé en droit français au plus tard le 9 décembre 2026 et s'appliquera aux solutions d'IA.
- Le droit commun de la responsabilité civile délictuelle :
- Les articles 1240 et suivants du code civil.
- Plus spécifiquement, la responsabilité du fait des choses, telle que définie à l'article 1242-1 du Code Civil : *"On est responsable non seulement du dommage que l'on cause par son propre fait, mais encore de celui qui est causé par le fait des personnes dont on doit répondre, ou des choses que l'on a sous sa garde"*. Il peut aussi s'appliquer dans le cadre d'une intelligence artificielle.

La gouvernance nécessaire à une IA de confiance

3

La mise en place d'une gouvernance centrale et efficace pour les IA de confiance est un enjeu stratégique. Elle implique, au sein des organisations, différentes directions : juridiques, techniques, métiers et repose sur des piliers fondamentaux tels que la formation des équipes, la structuration des données, et la standardisation des processus IA.

Elle permet de standardiser les outils et méthodes, et de mutualiser les ressources (modèles, données, bonnes pratiques).

Ces méthodologies rigoureuses doivent s'appliquer aux projets internes et aux projets externes, permettant ainsi de cadrer les initiatives internes et d'éviter les projets non conformes aux réglementations.

3.1. Formation interne

L'article 4, entré en vigueur le 2 février 2025, oblige toute organisation d'être en capacité de maîtriser leur système d'intelligence artificielle : "*Les fournisseurs et les déployeurs de systèmes d'IA prennent des mesures pour garantir, dans toute la mesure du possible, un niveau suffisant de maîtrise de l'IA pour leur personnel et les autres personnes s'occupant du fonctionnement et de l'utilisation des systèmes d'IA pour leur compte, en prenant en considération leurs connaissances techniques, leur expérience, leur éducation et leur formation*".

Les équipes doivent être formées non seulement à l'utilisation des modèles IA, mais également à leur explicabilité et leur auditabilité. Une collaboration avec les data scientists peut également être mise en place pour garantir la qualité et la traçabilité des données.

Une gouvernance des formations pourra aussi inclure une sensibilisation aux risques de biais algorithmiques ainsi que des méthodes de détection de ces derniers, et à la gestion des cas de non-conformité.

3.2. Structuration des données

La gouvernance des données est un levier central pour le déploiement des IA de confiance : traçabilité des jeux de données, gestion des métadonnées, et standardisation des formats de données. De même, l'élaboration de politiques documentaires, procédures qualité ou de gouvernance, permettant de consigner les décisions IA et les versions des modèles est essentielle pour garantir la traçabilité des données et des modèles.

Des référentiels communs pourront ainsi en découler et être établis pour assurer une cohérence dans le suivi des données entre les différents départements.

3.3. Standardisation des processus IA

Pour garantir la conformité des systèmes IA, il est essentiel de mettre en place des standards de documentation, d'audit et de vérification des modèles IA. L'inclusion d'un cycle de vie documenté pour chaque projet IA permet de s'assurer que chaque phase (de la conception à la mise en production) soit alignée avec les exigences du RIA. L'articulation avec les métiers tels que celui de DPO, de RSSI, ou encore le records management, permet également d'assurer une gestion intégrée des preuves numériques, incluant le versionnage, la sécurité, l'archivage électronique et la réversibilité des modèles IA.

La norme ISO 42001 "System management de la qualité appliquée à l'Intelligence artificielle" permet un accompagnement global, à la fois juridique, organisationnel et sécuritaire, sur la mise en place et l'utilisation d'un système d'intelligence artificielle.

3.4. Implication des tiers et partenaires

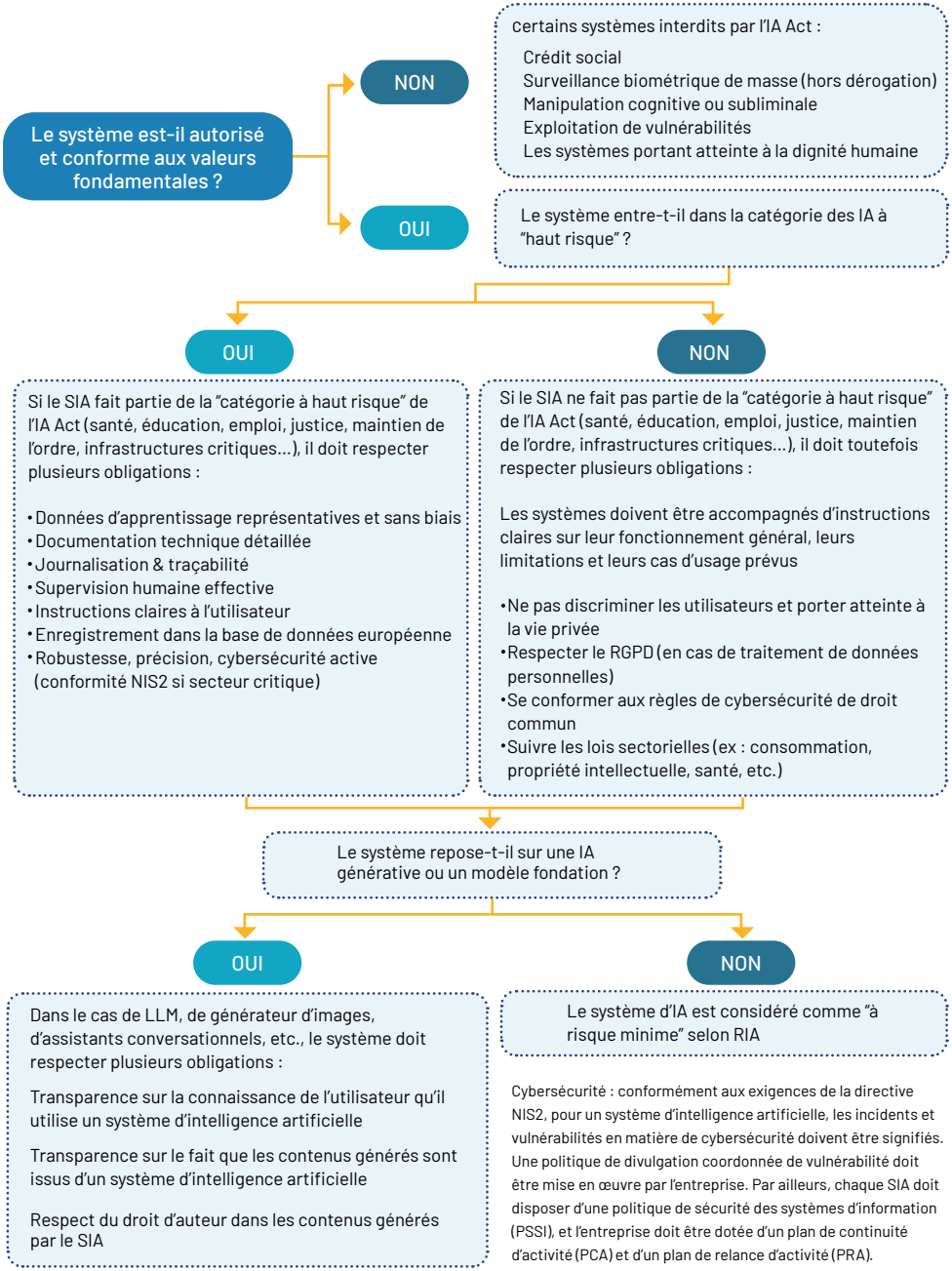
Un aspect clé de la gouvernance consistera aussi à s'assurer que les fournisseurs de solutions IA respectent les mêmes exigences de transparence, d'auditabilité et de continuité que celles définies en interne. Cela inclut des clauses contractuelles sur la traçabilité des modèles, l'accès aux audits et la documentation des décisions automatisées, et doit intégrer des points de contrôle régulier.

Initiatives grand public

Pour qu'une intelligence artificielle soit véritablement de confiance, il ne suffit pas qu'elle soit techniquement performante ou réglementairement conforme : il faut aussi que le public puisse se l'approprier en pleine connaissance de cause. Aujourd'hui encore, l'IA reste trop souvent perçue comme une technologie complexe, opaque, voire menaçante, suscitant autant de fascination que d'appréhension. Afin de surmonter ces perceptions, un effort collectif de pédagogie et de sensibilisation s'avère indispensable. Il s'agit de dépasser les idées reçues, d'éclairer les réels enjeux sociétaux et éthiques de l'IA, mais aussi de rendre accessibles ses principes de fonctionnement, ses usages concrets et ses limites intrinsèques.

Le rôle des pouvoirs publics est essentiel pour impulser ces actions à grande échelle, tel que cela a été initié avec les Café IA, le plan national Osez l'IA, mais il doit impérativement être complété par la mobilisation active d'entreprises, d'institutions académiques, de collectivités locales et d'associations. Ce large partenariat est la clé d'une pédagogie réussie, permettant à chacun de se forger une opinion éclairée et critique, à partir d'informations fiables et transparentes. En donnant ainsi la parole aux utilisateurs finaux, on favorise l'appropriation active et responsable de ces technologies.

ARBRE DE DECISION
QUE FAUT-IL METTRE EN PLACE POUR UTILISER UN SYSTEME D'INTELLIGENCE ARTIFICIELLE ?



Cybersécurité : conformément aux exigences de la directive NIS2, pour un système d'intelligence artificielle, les incidents et vulnérabilités en matière de cybersécurité doivent être signifiés. Une politique de divulgation coordonnée de vulnérabilité doit être mise en œuvre par l'entreprise. Par ailleurs, chaque SIA doit disposer d'une politique de sécurité des systèmes d'information (PSSI), et l'entreprise doit être dotée d'un plan de continuité d'activité (PCA) et d'un plan de relance d'activité (PRA).

CONCLUSION

L'intelligence artificielle est une technologie profondément transformatrice, aux impacts majeurs sur nos économies, nos sociétés et nos vies quotidiennes. À ce titre, elle ne peut rester sans encadrement. Les réflexions autour des critères, des moyens et de la gouvernance d'une intelligence artificielle de confiance ne font que commencer. Si les textes réglementaires apportent des premiers jalons, de nombreux aspects restent à préciser, notamment en matière de responsabilité, de transparence ou d'éthique. Pour garantir l'adoption d'une technologie maîtrisée et bénéfique à tous, il apparait indispensable de l'accompagner par des règles claires, des standards partagés et une gouvernance transparente. **La création d'un label d'IA de confiance constituerait ainsi une avancée majeure pour établir un cadre commun**, protéger les utilisateurs et valoriser les acteurs engagés dans une démarche responsable.

La mise en place de standards et de règles communes pour promouvoir une IA de confiance ne pourra réussir que par la mobilisation conjointe de l'ensemble des parties prenantes : pouvoirs publics, entreprises, monde académique, société civile. C'est à ce prix que l'intelligence artificielle pourrait devenir non seulement une source d'innovation mais également un vecteur de progrès collectif et de confiance.

COMITÉ DE RÉDACTION :

- Cabinet Caprioli & Associés
- Groupe Enso
- Conseil national de l'ordre des experts-comptables
- Docaposte
- Doxallia
- Orano
- Signaturit
- Société générale
- Tessi

*La FnTC remercie le **Dr. Maxime Derian** pour sa précieuse contribution*

fntc
Fédération des Tiers de Confiance du Numérique

5, impasse Gomboust
75001 Paris
infos@fntc-numerique.com
fntc-numerique.com

NOVEMBRE 2025