Réconcilier souveraineté et exportabilité de l'IA

avec des modèles et une gestion des données adaptés

THALES
Building a future we can all trust

THALES
Building a future we can all trust

Réconcilier souveraineté et exportabilité de l'IA

avec des modèles et une gestion des données adaptés

Auteurs :

Boussad ADDAD, Rémi BLANCHETTE, Dave COUTURE, Fabien FLACHER,

Katarzyna KAPUSTA, Juliette MATTIOLI, Gabriel RANGONI

Résumé

Dans le domaine hautement souverain de la Défense, l'intelligence artificielle a démontré sa valeur stratégique et est désormais très recherchée. Étant donnée la forte dépendance de l'IA fondée sur les données à la qualité des données d'entrainement, une collaboration étroite entre les fournisseurs de systèmes et les forces armées — qui détiennent et contrôlent les plus grands volumes de données opérationnelles — est essentielle.

Un défi central réside dans la définition des conditions permettant à cette collaboration de se dérouler de manière sûre et efficace, tant au niveau national qu'international, afin de développer des modèles d'IA performants.

Ce document présente un cadre pragmatique et technique visant à concilier la souveraineté en matière d'IA avec l'exportabilité des systèmes avancés intégrant l'IA. Il identifie les principaux leviers technologiques qui rendent l'IA souveraine opérationnellement viable. Ceux-ci incluent des mécanismes sécurisés de partage de données (tels que l'apprentissage par transfert et l'apprentissage collaboratif), des techniques de préservation de la confidentialité, ainsi que des outils de traçabilité comme le watermarking.

Un élément clé de cette approche est le développement d'une chaîne d'outils souveraine AIOps, conçue pour garantir une gouvernance robuste des données, l'intégrité des modèles et un réentraînement continu — même dans des environnements déconnectés ou sensibles.

Le document aborde également les risques de cybersécurité spécifiques à l'IA, en proposant des contre-mesures pour prévenir les fuites de données, le vol de modèles et les manipulations adverses. Enfin, il souligne l'importance des choix d'infrastructure, des modèles de déploiement et des cadres juridiques pour soutenir une coopération internationale de confiance sans compromettre la souveraineté.

1. Contexte

1.1. SOUVERAINETÉ

La souveraineté en matière d'intelligence artificielle désigne la capacité d'une nation ou d'un ensemble de nations à préserver durablement sa liberté dans la gestion de l'utilisation, du développement et de la régulation de l'IA sur son territoire. Cela inclut la préservation de l'autonomie face aux influences non sollicitées et s'étend sur plusieurs dimensions :

- **Technologie** : Maîtrise des capacités avancées en IA, tant au niveau matériel que logiciel, accès aux infrastructures critiques, et disponibilité des compétences nécessaires au niveau national, afin de réduire la dépendance étrangère.
- **Réglementation** : Établissement de cadres juridiques garantissant une utilisation responsable de l'IA, protégeant la vie privée, la sécurité et les droits fondamentaux, tout en favorisant l'innovation nationale.
- **Économie** : Autonomie dans l'exploitation commerciale et industrielle de l'IA, soutien aux champions nationaux, et participation équilibrée dans la chaîne de valeur mondiale.
- **Géopolitique** : Influence dans la définition des normes internationales et positionnement stratégique dans la compétition technologique mondiale. Cela inclut notamment les aspects culturels et linguistiques, en particulier dans les grands modèles de langage.

Rendre l'intelligence artificielle souveraine est essentiel pour maintenir un contrôle autonome sur les systèmes intégrant l'IA. D'un point de vue technologique, cela implique de se concentrer sur l'autonomie technologique, la fiabilité, la confidentialité, la résistance aux cyberattaques et la confiance. Maîtriser un maximum de ces éléments permet d'aborder les parties de la chaîne d'approvisionnement de l'IA qui présentent le plus de dépendances externes en position de force.

1.2. DEFIS D'UNE IA SOUVERAINE

La souveraineté est une préoccupation majeure dans l'industrie de Défense. La souveraineté des systèmes de défense intégrant l'intelligence artificielle englobe tous les objectifs et contraintes liés à la souveraineté des autres systèmes de Défense hautement technologiques, avec quelques spécificités supplémentaires. En général, les nations et organisations qui développent ou acquièrent ces systèmes cherchent à maximiser leur contrôle autonome à travers quatre piliers fondamentaux : l'autonomie du client, la confidentialité des informations, la fiabilité de l'IA et la performance opérationnelle.

Information confidentielle

Garantir la confidentialité dans les systèmes d'IA souveraine nécessite une approche globale couvrant l'ensemble du cycle de vie. En respectant des normes telles que l'IEEE-7000 et en mettant en œuvre des mesures de sécurité robustes, les organisations peuvent protéger les informations sensibles contre tout accès non autorisé et les menaces liées au renseignement étranger. Cela préserve non seulement la vie privée individuelle et la conformité aux contrôles à l'exportation, mais aussi les avantages stratégiques et la propriété intellectuelle. Les mécanismes de confidentialité doivent protéger les données, les modèles et le savoir-faire contre la surveillance et l'espionnage. Les systèmes d'IA souveraine de Thales sont conçus selon ces critères.

IA de confiance

Les systèmes d'IA de confiance reposent sur une approche globale qui protège l'intégrité des modèles, des données, des processus, des opérations, des institutions et de la chaîne d'approvisionnement. Ensemble, ces éléments interdépendants garantissent que les systèmes d'IA restent authentiques, non compromis et conformes à leurs objectifs initiaux, assurant ainsi la confiance et la fiabilité indispensables à leur adoption à grande échelle et à leur impact positif sur la société.

Figure 1 – Les piliers de l'IA de confiance - Thales TrUE AI framework (Transparente, Compréhensible, et Ethique)

1.3. L'IA SOUVERAINE APPELLE À DAVANTAGE DE COLLABORATION

Les systèmes d'IA se distinguent par leur capacité à évoluer dans le temps — à condition d'être régulièrement entraînés avec des ensembles de données riches, diversifiés et à jour. Cette adaptabilité dynamique offre un avantage décisif par rapport aux autres technologies. Dans la période actuelle d'innovation rapide en IA, la conception et la mise en œuvre des modèles en eux-mêmes offrent également des opportunités significatives d'amélioration continue, notamment à mesure que les capacités matérielles progressent.

Deux parties prenantes clés actionnent ce double levier de progrès : les utilisateurs, qui exploitent les systèmes dans des conditions réelles et fournissent des retours d'expérience, et les développeurs, qui conçoivent et maintiennent les modèles d'IA ainsi que les produits qui les intègrent, en s'appuyant notamment sur ces retours.

Une collaboration soutenue entre ces acteurs est essentielle pour garantir la performance des systèmes sur le long terme. Les utilisateurs apportent des données opérationnelles qui enrichissent le réentraînement des modèles, tandis que les développeurs affinent les architectures et les implémentations pour optimiser la fiabilité et l'efficacité. Cette coopération doit être rendue possible par des outils et des processus dédiés — même dans des environnements où la souveraineté est une préoccupation majeure.

Ce principe s'applique au niveau national et reste tout aussi pertinent à l'international. Les modèles d'IA bénéficient d'ensembles de données d'entraînement plus larges, issus de contextes opérationnels diversifiés à travers le monde. La qualité des modèles est fortement corrélée au volume, à la diversité et à la précision des données d'entraînement. Des techniques telles que l'apprentissage frugal, les approches neuro-symboliques et l'utilisation de données synthétiques peuvent atténuer — mais non éliminer — le besoin de données réelles. En définitive, des ensembles de données plus riches contribuent à élargir les domaines opérationnels et améliorer les performances des systèmes.

D'un point de vue économique, une distribution plus large des modèles via les exportations contribue à compenser les investissements substantiels nécessaires pour maintenir la supériorité technologique. Cependant, le partage de données opérationnelles dans des contextes de défense exige une extrême prudence de la part du pays d'origine comme du pays de destination, dans le respect strict des cadres juridiques de protection du secret et de conformité commerciale.

Notre expérience nous amène à considérer qu'avec les technologies appropriées et une approche pragmatique, il devient possible de concilier souveraineté et exportabilité.

En pratique, les parties prenantes reconnaissent les bénéfices mutuels de la collaboration, bien que leurs priorités diffèrent :

Les utilisateurs de systèmes basés sur l'IA (qu'il s'agisse des pays d'origine ou de destination) attendent généralement que :

- Un accès immédiat à des systèmes de pointe dont la fiabilité est éprouvée.
- La capacité d'améliorer les performances en utilisant leurs propres données opérationnelles.
- De solides garanties que leurs données seront protégées contre toute utilisation non autorisée — y compris dans des cadres de coopération.
- Une résilience robuste des systèmes face aux ingérences externes.

Les pays d'origine, dans le contexte de l'export, attendent également :

- Une compréhension claire — et, le cas échéant, une adaptation — des performances du système avant l'exportation.
- Des bénéfices directs ou indirects issus des améliorations tirées d'usages opérationnels diversifiés.
- L'application stricte des accords avec les pays de destination, y compris les engagements relatifs à la protection du secret et à la conformité commerciale le cas échéant, mises en œuvre et facilitées par les procédures et les outils du fournisseur.

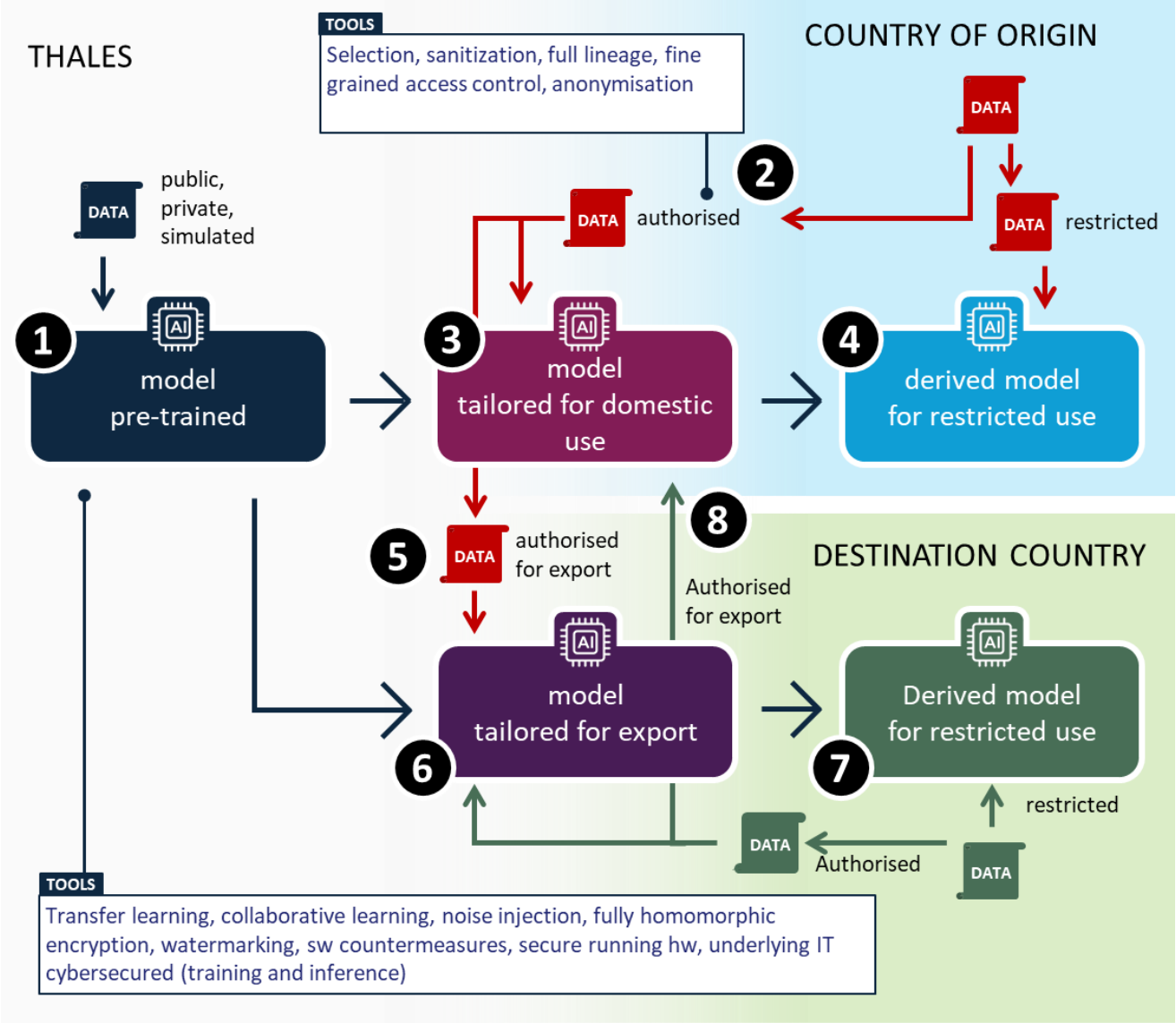
Les fournisseurs, de leur côté, attendent l'accès aux données opérationnelles nécessaires pour répondre à ces attentes — tant au niveau national qu'international.

Malgré des priorités différentes, une convergence est possible. Avec des workflows appropriés, des chaînes

d'outils opérationnels d'IA souveraine et des technologies de cybersécurité avancées, toutes les parties peuvent collaborer efficacement.

Schéma global de collaboration

Le schéma 1 ci-dessous présente un cadre de collaboration structuré conçu pour une IA souveraine et exportable.



1. **Développement initial du modèle**
   Le modèle d'IA est initialement développé et entraîné à partir de données détenues ou accessibles par le fournisseur du système (ex. : Thales).

2. **Intégration des données souveraines par le pays d'origine**
   Le pays d'origine autorise le fournisseur à améliorer le modèle en utilisant ses propres données opérationnelles, à l'exclusion possible des informations hautement sensibles. Ce processus inclut des contrôles d'accès robustes et un suivi de la traçabilité des données. Cette intégration peut intervenir dès le départ et se poursuivre tout au long du cycle de vie du modèle, en fonction des retours opérationnels.

3. **Optimisation collaborative des performances**
   Après réentraînement avec des données supplémentaires, le modèle atteint son niveau de performance collaborative maximal.

4. **Réentraînement local exclusif par le pays d'origine**
   Dans certains cas, le pays d'origine peut demander la possibilité de réentraîner le modèle avec des données qu'il ne souhaite pas partager. Le fournisseur peut fournir des outils d'entraînement IA sécurisés pour permettre ce réentraînement local exclusif.
5. **Préparation du modèle pour l'export**
   Afin de faciliter la coopération internationale, le pays d'origine peut autoriser le fournisseur à utiliser un sous-ensemble de ses données pour entraîner une version du modèle destinée à l'exportation.
6. **Collaboration avec le pays de destination**
   Comme pour l'usage domestique, le pays de destination peut collaborer avec le fournisseur pour réentraîner le modèle avec ses propres données, afin d'adapter les performances à son contexte opérationnel spécifique.
7. **Réentraînement local exclusif par le pays de destination**
   Le pays de destination peut également demander la possibilité de réentraîner le modèle avec des données qu'il ne souhaite pas partager avec le fournisseur.
8. **Contribution réciproque des données**
   Le pays de destination peut autoriser le fournisseur à intégrer une partie de ses données dans le modèle domestique, renforçant ainsi la collaboration avec le pays d'origine. Cet échange bidirectionnel maximise la performance du système.

La mise en œuvre de ces étapes successives d'entraînement et d'échanges sécurisés de données nécessite le déploiement de technologies spécifiques :

- **Chaînes d'outils collaboratives pour l'IA** : plateformes permettant la préparation, la désensibilisation, le partage et la gestion sécurisée des données entre parties prenantes.
- **Modèles d'IA sécurisés** : architectures dotées de mécanismes de traçabilité et de protection, s'appuyant sur des techniques avancées de cybersécurité et de cryptographie.
- **Processus sécurisés imposés** : outils et workflows garantissant la conformité et la sécurité tout au long du cycle de développement et d'entraînement, sans coûts excessifs.

Les chapitres suivants détaillent les aspects techniques de la mise en œuvre de ces capacités.

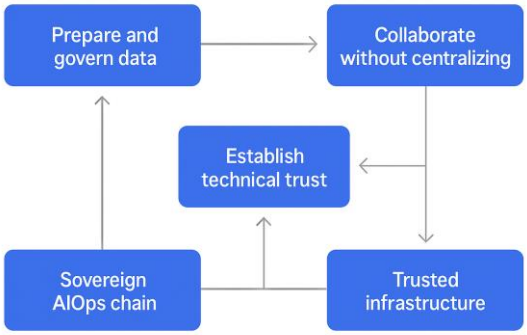
Au-delà de la dimension technique, ce modèle collaboratif transforme profondément les relations opérationnelles et commerciales entre fournisseurs et clients. Il nécessite des accords-cadres dédiés définissant les conditions d'accès, d'utilisation, de modification et de partage des données et des modèles — garantissant que les systèmes restent dans un état opérationnel équitable et optimal pour toutes les parties.

Cela peut impliquer des **accords généraux de sécurité bilatéraux ou multilatéraux** et un ensemble de licences de contrôle des exportations assorties de conditions strictes, qui doivent être anticipées et adaptées en fonction des besoins. Une préparation et une coordination importantes seront nécessaires entre le fournisseur, le pays d'origine et, ultérieurement, le pays de destination. Toute juridiction tierce ou extraterritoriale potentiellement applicable en matière de conformité commerciale devra être évaluée et prise en compte avant et pendant la coopération.

Que l'IA fasse ou non partie du projet, cette phase de préparation du cadre et son suivi tout au long des projets ne doivent pas être sous-estimés afin de garantir la fluidité et le succès de la coopération.

2. Vue d'ensemble de l'approche

La méthode proposée repose sur une logique industrielle et progressive, structurée autour de cinq piliers techniques : la préparation et la gouvernance des données, le choix des paradigmes de collaboration, les mécanismes de confiance, la chaîne AIOps souveraine et l'infrastructure contrôlée.



Framework for data collaboration in sovereign environments

Préparer et gouverner les données

La première étape consiste à atténuer les risques à la source. Avant tout partage de données, il est nécessaire de mettre en place **une chaine de nettoyage et désensibilisation** afin de conserver les informations utiles à l'apprentissage tout en supprimant les éléments sensibles — tels que les métadonnées inutiles, les indicateurs géographiques ou les valeurs aberrantes critiques. Ce processus améliore la qualité des données et réduit la dépendance à des techniques de confidentialité lourdes.

Une gouvernance solide accompagne cette préparation : la gestion des versions et le suivi de la lignée garantissent la traçabilité des transformations ; l'accès est contrôlé via des mécanismes basés sur les rôles (RBAC) et les attributs (ABAC) ; les secrets sont chiffrés et gérés de manière sécurisée. Un Feature Store versionné et chiffré assure la cohérence entre les environnements d'entraînement et de production. Le résultat est un ensemble de jeux de données coopératifs avec des niveaux de risque documentés et auditables.

Collaborer sans centraliser

La performance ne nécessite pas de centraliser les données. Deux approches privilégiées sont l'apprentissage par transfert et l'apprentissage collaboratif.

L'apprentissage par transfert consiste à partager des modèles de base entraînés sur des données non sensibles, qui sont ensuite adaptés localement à l'aide de jeux de données souverains. Cette méthode permet de mutualiser les investissements scientifiques sans déplacer les données critiques.

L'apprentissage collaboratif permet un entraînement distribué (par exemple, apprentissage fédéré ou fragmenté), où chaque participant calcule des mises à jour locales et ne partage que les paramètres du modèle. Ces échanges sont sécurisés grâce à des techniques cryptographiques (par exemple, MPC, FHE), des contrôles d'intégrité, ainsi que des mécanismes de révocation ou de désapprentissage pour gérer l'évolution des partenariats.

Établir la confiance technique

Une collaboration efficace repose sur des garanties vérifiables. La confidentialité est assurée par le chiffrement des données au repos et en transit, des environnements d'exécution sécurisés (par exemple, TEE), ainsi que des techniques de préservation de la vie privée telles que le chiffrement homomorphe et la confidentialité différentielle.

La propriété intellectuelle et la conformité réglementaire concernant l'usage autorisé sont protégées via le watermarking des modèles et des journaux signés qui capturent le contexte d'entraînement. La robustesse des modèles est renforcée par l'entraînement adversarial, la détection de déclencheurs et des tests réguliers contre des attaques telles que l'inversion, l'extraction et l'empoisonnement. Ces mécanismes transforment la confiance d'une simple déclaration en un contrat technique auditable.

Industrialiser avec une chaîne AIOps souveraine

La souveraineté est mise en oeuvre via une chaîne d'outils AIOps qui automatise et trace chaque étape du cycle de vie de l'IA. Elle inclut :

- L'intégrité de la chaîne d'approvisionnement (par exemple, journaux signés, SBOM/AIBOM)

- L'orchestration on-prem avec des pipelines codifiés et des contrôles éthiques
- Un déploiement compatible hors ligne
- Une surveillance continue (par exemple, détection de dérive, réentraînement supervisé)

Les politiques sont traduites en code pour garantir la conformité aux contraintes opérationnelles (ODD) et activer des modes sécurisés en cas de violation. Cette usine d'IA souveraine permet la construction, l'évaluation et la maintenance de modèles coopératifs sans perte de contrôle.

Déployer sur une infrastructure de confiance

Enfin, l'infrastructure doit être alignée à la fois sur les exigences technologiques et juridictionnelles. Une architecture multi-couches — combinant matériel sécurisé, virtualisation et conteneurisation — est essentielle. Les zones de confiance et les modèles de déploiement (on-prem, hybride, orienté edge) doivent être choisis en fonction des besoins de souveraineté.

Les échanges de modèles et de données sont effectués via des packages chiffrés et signés, orchestrés par la chaîne AIOps. Cela garantit la confidentialité, la traçabilité et la résilience face aux menaces.

La méthode proposée repose sur une logique industrielle et progressive, structurée autour de cinq piliers techniques : préparation et gouvernance des données, choix des paradigmes de collaboration, mécanismes de confiance, chaîne AIOps souveraine et infrastructure contrôlée.

3. Cadre et technologies habilitantes

Les organisations et les nations peuvent hésiter à partager des données d'entraînement, craignant que cela ne compromette leur autonomie ou n'expose des informations confidentielles. Bien que le partage de modèles pré-entraînés puisse sembler une alternative plus sûre, il peut également révéler involontairement des informations sur les données sous-jacentes — sauf s'il est effectué de manière sécurisée.

Cependant, la promesse d'une performance accrue grâce à la collaboration reste une incitation forte à explorer des méthodes sécurisées pour partager les connaissances et les données en IA. Ce besoin stimule le développement de nouvelles technologies favorisant une coopération compatible avec la souveraineté.

L'apprentissage par transfert et l'apprentissage collaboratif se distinguent comme les deux paradigmes fondamentaux pour une collaboration sécurisée autour des données. Thales identifie ces approches comme essentielles pour construire des plateformes d'IA prêtes pour la souveraineté. Elles permettent à plusieurs acteurs d'entraîner conjointement des modèles sans échanger de données brutes, préservant ainsi la souveraineté tout en maintenant des standards élevés de confidentialité et de sécurité.

Ces cadres peuvent être renforcés par un ensemble de leviers de confiance, notamment :

- Désensibilisation des données : suppression ou anonymisation des éléments sensibles avant partage.
- Garanties cryptographiques : sécurisation des échanges de données et de modèles.
- Techniques d'anonymisation : protection des points de données individuels.
- Mécanismes de traçabilité : assurer la transparence et la responsabilité tout au long du cycle de vie de l'IA.

Ensemble, ces technologies constituent la colonne vertébrale d'un environnement de collaboration en IA sécurisé et souverain.

3.1. DÉSENSIBILISATION DES DONNÉES D'ENTRAÎNEMENT

Quel que soit le cadre de collaboration adopté pour l'entraînement des modèles d'IA, la première étape essentielle consiste à évaluer dans quelle mesure les données disponibles et pertinentes peuvent être partagées en toute sécurité avec des tiers — qu'il s'agisse de fournisseurs de modèles ou de partenaires à l'exportation.

Cette évaluation doit garantir la cohérence entre l'autorisation de partager des modèles entraînés et celle de partager les données d'entraînement sous-jacentes. Dans de nombreux cas, un équipement intégrant l'IA et approuvé pour l'exportation apporte plus de valeur à l'acquéreur que ses composants individuels, y compris les données brutes utilisées pour entraîner le modèle. Une fois entraîné, le modèle devient une incarnation fonctionnelle des connaissances qu'il a assimilées. Par conséquent, les données strictement utilisées pour développer un équipement autorisé à l'export peuvent souvent être considérées comme exportables également.

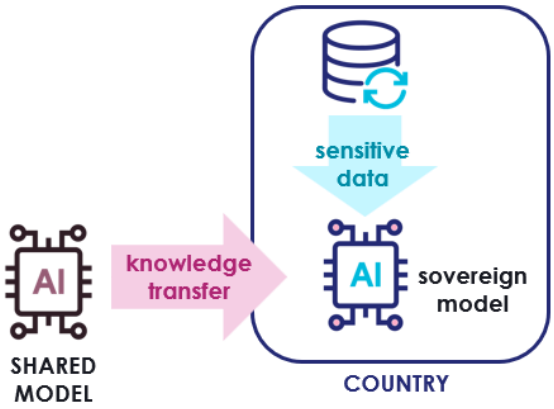
Cependant, certains jeux de données peuvent contenir des informations qui ne contribuent pas directement à la performance du modèle mais pourraient révéler des détails sensibles — tels que les caractéristiques de la chaîne de capture des données, les spécifications des capteurs ou les lieux et calendriers des missions. Ces éléments doivent être exclus ou anonymisés lors de la phase de préparation des données afin d'éviter toute exposition inutile.

Une désensibilisation des données correctement exécutée protège non seulement la confidentialité, mais réduit également le besoin de recourir à des techniques lourdes de préservation de la vie privée pendant l'entraînement. Elle aide aussi à identifier et éliminer les données fausses ou contradictoires susceptibles de dégrader la performance du modèle.

Une fois cette phase critique terminée, l'accès au jeu de données d'entraînement nettoyé doit rester strictement contrôlé — tant pendant le processus d'entraînement que lors de l'inférence. Dans certains contextes, cette protection peut aller jusqu'à limiter l'accès, même parmi les partenaires de confiance.

3.2. APPRENTISSAGE PAR TRANSFERE

L'apprentissage par transfert offre une approche équilibrée qui combine coopération internationale et souveraineté nationale. En partageant des modèles fondamentaux entraînés sur des ensembles de données publics ou anonymisés, les pays et les organisations peuvent renforcer collectivement leurs capacités en IA tout en conservant le contrôle sur leurs données locales sensibles.



Cette méthode collaborative commence par le développement de modèles de base partagés, construits à partir de données publiques non sensibles ou de jeux de données privés anonymisés. Ces modèles reflètent une expertise et des ressources mutualisées, formant une base technologique commune qui serait prohibitive à développer individuellement.

Chaque nation ou organisation peut ensuite affiner ces modèles partagés en utilisant ses propres données souveraines. Cette approche duale — tirer parti des connaissances collectives tout en préservant le contrôle local — permet d'optimiser les performances sans compromettre la confidentialité ni la conformité réglementaire. Elle crée un cercle vertueux : la coopération internationale accélère le développement des modèles fondamentaux, et tous les participants bénéficient des améliorations obtenues.

Le format utilisé pour partager les modèles joue un rôle critique. Les standards ouverts tels qu'ONNX (Open Neural Network Exchange) doivent être privilégiés, car ils permettent d'exécuter des modèles entraînés dans un framework (par exemple PyTorch ou TensorFlow) dans un autre. Cela garantit l'interopérabilité entre outils, plateformes et matériels.

En réduisant la dépendance à des fournisseurs spécifiques, ONNX renforce la souveraineté technologique et donne aux organisations un meilleur contrôle sur leur infrastructure et leurs flux de données. Sa nature ouverte et standardisée favorise un déploiement flexible — qu'il soit sur site ou dans des environnements cloud privilégiés — tout en

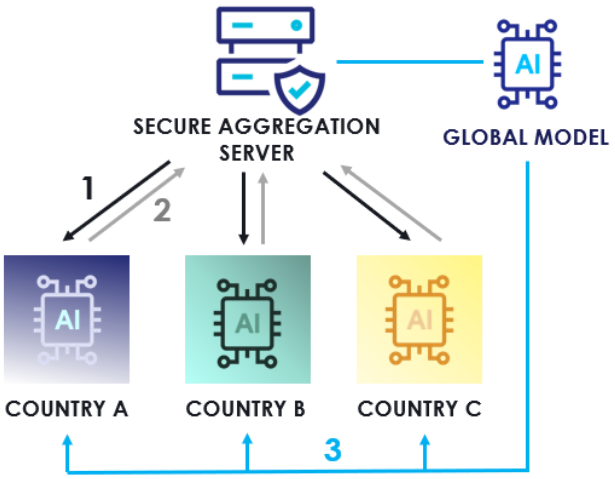
respectant les exigences de résidence des données et de conformité. Il améliore également la transparence et l'auditabilité, renforçant à la fois la sécurité et les efforts réglementaires.

3.3. APPRENTISSAGE COLLABORATIF

L'apprentissage collaboratif offre un cadre puissant pour favoriser la coopération entre nations et organisations tout en préservant la souveraineté des données. Il permet à plusieurs parties prenantes d'entraîner conjointement des modèles d'IA sur des ensembles de données distribués, sans jamais centraliser ni échanger explicitement des informations sensibles, soumises à des contrôles d'exportation ou classifiées.

L'entraînement est réalisé de manière décentralisée : chaque participant conserve son propre modèle local et ne partage que des mises à jour liées au modèle (gradients) avec un serveur agrégateur. La forme la plus répandue d'apprentissage collaboratif est l'apprentissage fédéré, où tous les participants utilisent la même architecture de modèle. Cette approche se décline en différentes configurations selon la distribution des données (horizontale ou verticale), la présence d'une entité centrale de coordination et le nombre d'appareils participants (par exemple, de nombreux dispositifs IoT sous un même propriétaire ou plusieurs serveurs appartenant à différentes organisations).

Une autre technique prometteuse est l'apprentissage fragmenté (Split Learning), qui répartit les responsabilités d'entraînement du modèle de manière à ce que seules des activations intermédiaires — et non les données brutes — soient échangées. Cette méthode est particulièrement adaptée aux scénarios où les données ou les composants du modèle doivent rester strictement séparés.



Thales renforce ces pipelines d'apprentissage collaboratif de base avec des fonctionnalités avancées telles que la signature cryptographique, le suivi de provenance et des registres de modèles immuables. Ces caractéristiques garantissent l'intégrité, l'authenticité et la traçabilité de tous les composants IA tout au long de leur cycle de vie.

Bien que la recherche continue d'affiner les garanties de sécurité de l'apprentissage fédéré, les techniques existantes offrent déjà une protection substantielle. Par exemple, des méthodes de préservation de la confidentialité peuvent être intégrées pour réduire le risque de fuite de données via des attaques spécifiques à l'IA ciblant le serveur d'agrégation central — protégeant ainsi les informations sensibles même dans des collaborations transfrontalières.

Pour gérer les droits de propriété, le watermarking appliqué au machine learning peut être utilisé. Cette technique permet à chaque participant d'intégrer une contribution vérifiable dans le modèle final, assurant une attribution transparente et favorisant une collaboration sécurisée. Combinée à des capacités de désapprentissage, elle offre également la flexibilité de révoquer des contributions si nécessaire.

Dans l'ensemble, l'apprentissage collaboratif — renforcé par des garanties cryptographiques et des mécanismes de traçabilité — donne aux organisations la possibilité d'entraîner des modèles haute performance en toute confiance, sachant que chaque étape du processus repose sur la transparence, l'auditabilité et la résilience.

3.4. MECANISME DE CONFIANCE

Pour permettre une collaboration en IA sécurisée et souveraine, la confiance doit être intégrée au niveau

technique. Cela nécessite une combinaison de protections cryptographiques, de matériel sécurisé et de mécanismes de traçabilité garantissant l'intégrité des données, l'authenticité des modèles et la protection des droits de propriété.

Cryptographie et matériel sécurisé

Les techniques cryptographiques jouent un rôle central dans la sécurisation des systèmes d'IA. Elles protègent les données et les modèles, qu'ils soient au repos ou en transit. Des méthodes avancées comme le chiffrement homomorphe et le calcul multipartite (MPC) étendent cette protection aux données en cours d'utilisation — pendant l'entraînement ou l'inférence — bien qu'elles puissent introduire une surcharge de calcul. Dans certains cas, des alternatives plus légères comme l'obfuscation peuvent être utilisées pour équilibrer performance et sécurité.

Le matériel sécurisé contribue également à la confiance. Les environnements d'exécution sécurisés (Trusted Execution Environments, TEE) offrent des zones isolées pour les calculs sensibles, les protégeant même du système d'exploitation hôte. Ces protections matérielles renforcent la souveraineté en garantissant que les opérations critiques restent confidentielles et résistantes aux manipulations.

Traçabilité et protection de la propriété

La collaboration entre organisations et nations soulève des questions importantes sur la traçabilité des contributions et la protection de la propriété intellectuelle. Les mécanismes traditionnels de gestion des droits numériques (DRM) — tels que les droits d'auteur ou les brevets — sont souvent inadaptés au contexte du machine learning, où les modèles peuvent être copiés, ajustés ou extraits via des attaques spécifiques à l'IA.

Pour y remédier, les modèles peuvent être chiffrés ou obfusqués afin d'empêcher toute réplication non autorisée. De plus, le watermarking appliqué au machine learning offre une solution adaptée : il insère des marques persistantes et vérifiables dans les modèles, qui restent détectables même après réentraînement. Ces filigranes servent de preuve de propriété et permettent d'affirmer les droits sur l'origine et l'intégrité d'un modèle.

Ensemble, ces leviers de confiance constituent la base d'un environnement de collaboration en IA sécurisé et auditable — respectant la souveraineté, protégeant les actifs sensibles, garantissant la conformité aux contrôles d'exportation et assurant la responsabilité à toutes les étapes du cycle de vie de l'IA.

4. Chaîne d'outils IA adaptée aux données critiques et à la coopération

La souveraineté en matière d'IA ne peut pas se limiter à des déclarations politiques : elle exige une infrastructure industrielle robuste. Au cœur de cette infrastructure se trouve une chaîne d'outils AIOps dédiée, qui agit comme une usine automatisée de construction et d'exécution. Elle sécurise les données, valide les chaînes d'approvisionnement et garantit la résilience opérationnelle grâce à une surveillance et un réentraînement continus.

Plus important encore, cette chaîne d'outils fournit une plateforme commune permettant aux nations alliées de consolider des efforts fragmentés en une capacité souveraine unifiée — suffisamment solide pour rivaliser avec les leaders technologiques mondiaux. Dans la course à la maîtrise de l'IA, une chaîne AIOps souveraine n'est pas un luxe : c'est une nécessité stratégique, malgré la complexité qu'elle implique.

L'IA diffère fondamentalement des vagues technologiques précédentes. Alors que l'ingénierie matérielle suit des lois physiques prévisibles et que le développement logiciel repose sur une logique déterministe, l'IA introduit de nouvelles couches de complexité. Les modèles d'apprentissage dépendent des distributions de données, de l'optimisation stochastique et de l'expérimentation itérative. Leur performance peut se dégrader avec le temps, nécessitant une surveillance constante, un réentraînement et un redéploiement.

Cette fragilité rend les systèmes d'IA fortement dépendants de leur socle opérationnel — et dangereusement exposés si ce socle repose sur des plateformes étrangères. Dans ce contexte, une chaîne AIOps n'est pas simplement une commodité technique : c'est le fondement de la souveraineté en IA.

Contrairement aux pipelines DevOps traditionnels, une chaîne AIOps répond aux défis uniques des projets IA. Elle gère la version des données, suit les expériences et les modèles, automatise la conformité aux normes éthiques et réglementaires, et prend en charge la surveillance des performances et le réentraînement.

Dans les applications de défense, la chaîne doit être adaptée aux environnements critiques et sensibles, y compris les systèmes isolés ou déconnectés. Elle doit prendre en charge le déploiement sur site, appliquer des contrôles d'accès stricts et garantir la traçabilité complète des données et des artefacts de modèles.

Les politiques de sécurité, les procédures opérationnelles et les points de contrôle de conformité sont intégrés dans des workflows automatisés, répondant non seulement aux exigences techniques mais aussi aux obligations légales, de contrôle des exportations et de souveraineté.

Bien conçue, cette chaîne d'outils offre trois capacités stratégiques :

- Sécurité et contrôle des actifs de données
- Intégrité de la chaîne d'approvisionnement IA
- Opérations souveraines continues

Elle agit à la fois comme un accélérateur de productivité et comme un socle de confiance, garantissant que les systèmes IA sont industriels, sécurisés, traçables et conformes aux standards internes et externes pour une IA éthique et fiable.

4.1. SECURITE ET CONTROLE DES DONNEES

Dans les systèmes d'IA, les données sont à la fois l'actif le plus stratégique et le plus vulnérable. Pour garantir leur protection, une chaîne AIOps souveraine doit offrir un ensemble complet de capacités permettant de sécuriser les données le long de leur cycle de vie.

Gouvernance et traçabilité des données

Un data lake local et versionné doit enregistrer l'origine, les autorisations, les licences et la traçabilité complète de tous les jeux de données. Des contrôles automatisés de qualité permettent d'éviter la contamination des licences et de créer des ensembles de données reproductibles — essentiels pour les audits et les examens liés à l'exportation. Par exemple, cela permet de prouver qu'un modèle de défense a été entraîné exclusivement sur des sources autorisées.

Gestion des secrets et sécurité matérielle

Les identifiants sensibles doivent être stockés de manière sécurisée dans des coffres-forts matériels tels que les modules de sécurité matériels (HSM) ou les modules de plateforme sécurisés (TPM). Le chiffrement par environnement, la segmentation réseau et l'analyse en temps réel réduisent les risques

de fuite d'identifiants et de mouvements latéraux dans des infrastructures sensibles.

Contrôle d'accès granulaire

L'accès aux données et aux sorties des modèles doit être régi par des modèles RBAC (Role-Based Access Control) et ABAC (Attribute-Based Access Control). Ces mécanismes garantissent que les utilisateurs ne peuvent accéder qu'aux données, exécuter des tâches ou tester des modèles dans les limites de leur niveau d'autorisation. Le système doit également évaluer la sensibilité des modèles indépendamment des données d'entraînement, en renforçant les restrictions lorsque des capacités émergentes ou des contextes de déploiement l'exigent.

Les politiques "as code" alignent les attributs des utilisateurs (rôle, habilitation, projet, environnement) avec la sensibilité des opérations et des artefacts. Cela empêche les accès non autorisés, limite les actions à fort impact par des utilisateurs sous-habilités, permet la révocation automatique ainsi qu'une traçabilité complète pour les audits ou la réponse aux incidents.

Gestion des fonctionnalités

Un Feature Store local garantit la cohérence entre les environnements hors ligne et en ligne, prend en charge le chiffrement au repos et étend les contrôles d'accès granulaire. Il évite les incohérences entre l'entraînement et la production, empêche la fuite de signaux sensibles et clarifie les politiques de propriété et de mise à jour — par exemple, en limitant les fonctionnalités liées à la santé à des rôles spécifiques.

4.2. INTÉGRITÉ DE LA CHAINE D'APPROVISIONNEMENT

Le développement de l'IA repose sur des chaînes complexes et souvent opaques — incluant des frameworks open source, des modèles préentraînés et des services tiers. Sans supervision rigoureuse, ces dépendances peuvent introduire des vulnérabilités cachées compromettant la souveraineté et la sécurité.

Une chaîne AIOps souveraine doit valider chaque composant, appliquer la conformité et empêcher toute insertion malveillante ou non autorisée dans les systèmes critiques. Cela se fait grâce à une combinaison de gestion sécurisée des artefacts et de suivi robuste des expérimentations.

Chaîne d'approvisionnement sécurisée des artefacts

Un registre local doit stocker tous les codes, modèles et exécutables avec des signatures cryptographiques et des attestations de provenance (par exemple, selon le modèle SLSA). La chaîne doit également générer des AI Bills of Materials (AI-BOM) pour établir une confiance de bout en bout. Ces mesures préviennent les attaques sur la chaîne d'approvisionnement et garantissent que les modèles et conteneurs déployés correspondent à leurs sources vérifiées.

Suivi des expérimentations et reproductibilité

Un tracker auto-hébergé (par exemple, MLflow) doit consigner toutes les exécutions d'entraînement et d'évaluation, y compris les versions de code, les paramètres, les instantanés de jeux de données, les configurations d'environnement et les artefacts résultants. Les builds déterministes sont imposés via des dépendances figées, et chaque exécution est accompagnée d'un SBOM ou AI-BOM généré. Cela permet la reproductibilité, facilite la comparaison et soutient une exploration efficace.

Registre des modèles avec provenance

Les modèles doivent être stockés avec leur contexte d'entraînement, les références de données, les rapports d'évaluation et les signatures numériques. Cela garantit une gouvernance complète du cycle de vie et permet uniquement aux modèles examinés et traçables de progresser vers le déploiement. Des approbations multipartites et des mécanismes de rollback contrôlés renforcent la confiance et la responsabilité.

Orchestration et automatisation

Des outils d'orchestration sur site (par exemple, Airflow, Argo) exécutent des pipelines codifiés avec des politiques « as code » intégrées. Ces pipelines appliquent des points de contrôle qualité, des validations et des vérifications de conformité de manière cohérente et transparente. L'automatisation élimine les erreurs manuelles et garantit que les exigences de confidentialité, de sécurité et de réglementation sont respectées à chaque étape.

4.3. OPÉRATIONS SOUVERAINES CONTINUES

Les modèles d'IA nécessitent une attention constante — ils ne peuvent pas simplement être déployés puis laissés sans surveillance. Les conditions réelles

évoluent en permanence : les environnements opérationnels changent et des événements imprévus peuvent modifier significativement les schémas de données et le comportement des systèmes. Ces changements peuvent entraîner une dégradation des performances ou une dérive des prédictions, qui doivent être détectées et corrigées rapidement.

Une chaîne AIOps robuste et souveraine est essentielle pour gérer ces défis. Elle doit prendre en charge la surveillance continue des modèles en production, permettant la détection précoce des baisses de performance et des changements contextuels. Lorsque de tels changements surviennent, le système doit faciliter un réentraînement et un redéploiement efficaces à partir de jeux de données mis à jour.

Des mécanismes de sécurité intégrés sont également cruciaux pour garantir la sûreté opérationnelle en cas d'anomalies soudaines de performance. Les fonctionnalités clés incluent :

Gestion du déploiement et de l'exécution

Les modèles doivent pouvoir être déployés sur des cibles variées — systèmes embarqués, dispositifs edge, clusters privés — avec des stratégies de déploiement et de rollback compatibles hors ligne. Cela réduit l'écart entre les environnements de laboratoire et de terrain et permet des mises à jour sécurisées sans accès Internet (par exemple, via des packages signés livrés par support physique).

Surveillance, détection de dérive et boucles de rétroaction

Des outils de télémétrie sur site (par exemple, Prometheus, OpenTelemetry) et des frameworks de détection de dérive (par exemple, Evidently, Alibi Detect) suivent les changements de données, les anomalies de prédiction et les événements critiques. Ces outils fournissent des alertes précoces, déclenchent le réentraînement sous contraintes définies et accélèrent la réponse aux incidents.

Conformité ODD et orchestration sécurisée

Les contraintes de l'Operational Design Domain (ODD) — telles que le géorepérage, les conditions environnementales, l'état des capteurs, la latence et les règles d'engagement — doivent être codifiées sous forme de politiques et appliquées à l'exécution. Cela garantit que les modèles fonctionnent dans leurs limites déclarées et active des réponses sécurisées et auditées lorsque des seuils sont atteints ou dépassés.

Ensemble, ces capacités permettent une exploitation souveraine continue des systèmes d'IA. Elles assurent que les modèles restent fiables, sécurisés et conformes tout au long de leur cycle de vie — même dans des environnements sensibles ou déconnectés. Cette colonne vertébrale opérationnelle est essentielle pour maintenir la confiance, l'agilité et la souveraineté dans les déploiements critiques.

5. Modèle d'intelligence artificielle cyber-sécurisé

La cybersécurité est une pierre angulaire de l'IA de confiance. Contrairement aux systèmes logiciels traditionnels, les solutions d'IA sont intrinsèquement basées sur les données et suivent un cycle de vie qui inclut la collecte des données, la conception de l'architecture du modèle, l'entraînement, le déploiement et souvent le réentraînement ou l'ajustement. Ce cycle de vie élargi augmente la surface d'attaque, exposant les modèles à des menaces à plusieurs étapes.

Les systèmes d'IA présentent également des vulnérabilités uniques. Au-delà des risques classiques, ils sont sensibles à des attaques spécifiques à l'IA, telles que les exemples adversariaux — des entrées conçues pour tromper le modèle — et les techniques d'inversion de modèle, qui tentent de reconstruire les données d'entraînement à partir des sorties du modèle.

Dans le contexte de la conciliation entre souveraineté et exportabilité, une préoccupation majeure est le risque d'exposition d'informations souveraines via des systèmes d'IA exportés. Ce risque stimule le développement de technologies améliorant la confidentialité, qui permettent d'exploiter les données pour l'apprentissage tout en anonymisant les échantillons individuels. Ces techniques aident à déterminer quelles parties des informations peuvent être partagées en toute sécurité au-delà des frontières.

Les autres défis critiques en cybersécurité :

- Protéger les droits de propriété sur les données et les modèles.
- Prévenir les déviations comportementales causées par des manipulations externes.

5.1. REVISITER L'ANALYSE DE SECURITE

Le modèle classique Confidentialité–Intégrité–Disponibilité (CIA) reste fondamental en sécurité de l'information. Dans le contexte de l'IA, il est souvent étendu par une quatrième dimension : Propriété, qui traite des risques de vol ou d'utilisation abusive des connaissances en IA. Bien que la confidentialité soit liée à la vie privée, les techniques de protection de la vie privée vont au-delà de la protection des données personnelles : elles visent à empêcher toute forme de fuite d'information, ce qui est parfois mal compris.

Une stratégie robuste de sécurité pour l'IA commence par l'identification des actifs à protéger — généralement les données d'entraînement, les données d'entrée et le modèle lui-même. Elle se poursuit par la construction d'un modèle de menace, qui évalue les vecteurs d'attaque potentiels et leur impact. Sur cette base, des contre-mesures appropriées sont appliquées. Elles se répartissent en deux catégories :

- **Techniques d'amélioration du modèle**, qui modifient le système d'IA lui-même et nécessitent souvent un réentraînement.
- **Techniques indépendantes du modèle**, telles que le filtrage des requêtes, qui peuvent être mises en œuvre sans modifier le modèle.

Bien que la cybersécurité de l'IA soit encore un domaine émergent, elle évolue rapidement — notamment dans le domaine de l'IA générative. Ces dernières années, plusieurs organismes de référence ont publié des recommandations, dont ETSI TC SAI, NIST, MITRE ATLAS et ANSSI. Ces efforts préparent la voie à des schémas de certification qui formaliseront la validation de la sécurité des IA et harmoniseront les métriques de cybersécurité dans l'industrie.

5.2. Éviter les fuites de données des modèles partagés

La fuite de données reste l'une des menaces les plus critiques pour la confidentialité des systèmes d'IA — surtout lorsque les modèles sont partagés entre organisations ou au-delà des frontières. Si les risques traditionnels proviennent de chaînes d'approvisionnement non sécurisées, des menaces plus subtiles et sophistiquées émergent des modèles eux-mêmes.

En effet, les modèles d'apprentissage automatique conservent souvent des traces des données sur lesquelles ils ont été entraînés. Les attaquants ayant accès à un modèle déployé peuvent exploiter cela via des techniques telles que :

- **L'inversion de modèle**, qui tente de reconstruire les échantillons d'entraînement originaux à partir des sorties du modèle.

- **L'inférence d'appartenance**, qui détermine si un point de données spécifique faisait partie de l'ensemble d'entraînement.
- **L'inférence de propriétés**, qui cherche à extraire des caractéristiques générales de l'ensemble d'entraînement.

Ces attaques peuvent compromettre des informations sensibles même si le modèle a été ajusté avec de nouvelles données avant d'être partagé. En 2023, l'équipe Friendly Hackers de Thales a démontré ce risque en reconstruisant des données confidentielles lors du défi CAID — soulignant l'importance de protections robustes.

Pour atténuer ces risques, plusieurs stratégies peuvent être employées :

- **Confidentialité différentielle** : introduit du bruit contrôlé dans les entrées, sorties ou algorithmes d'entraînement, rendant statistiquement impossible de déterminer si un point de données spécifique a été utilisé. Bien qu'efficace, elle peut affecter la performance du modèle et doit être appliquée après d'autres mesures, comme la désensibilisation des données.
- **Désapprentissage automatique** (Machine Unlearning) : approche émergente permettant de supprimer l'influence de données spécifiques d'un modèle entraîné sans nécessiter un réentraînement complet. Elle soutient le « droit à l'oubli » et est particulièrement utile lorsque des données deviennent obsolètes ou lorsque des modèles doivent être exportés sans conserver d'informations sensibles.

Actuellement, la vérification qu'un modèle ne fuit pas d'informations indésirables repose principalement sur les garanties de confidentialité différentielle et les tests d'intrusion empiriques. Comme la quantification des niveaux de confidentialité reste un défi de recherche, il est essentiel d'exclure les données hautement sensibles — telles que les valeurs aberrantes — dès la phase de prétraitement.

5.3. PARTAGER DES MODELES TOUT EN LES GARDANT SECRETS

Certaines organisations ou nations peuvent souhaiter rendre des modèles d'IA disponibles pour des partenaires externes tout en conservant la confidentialité de leur architecture interne, de leurs paramètres ou des données d'entrée/sortie. Inversement, les destinataires peuvent vouloir utiliser ces modèles sans révéler leurs propres requêtes ou données d'entrée.
Garantir la confidentialité des modèles partagés représente un défi distinct de la protection des données d'entraînement. L'objectif est de permettre l'exécution des modèles sur des dispositifs non fiables sans exposer leur structure interne ni les données qu'ils traitent. Ce concept est similaire à la protection du code logiciel pendant son exécution.

Une solution théorique est le **chiffrement entièrement homomorphe (FHE)**, qui permet d'effectuer des calculs directement sur des données chiffrées. Bien que des avancées récentes aient amélioré son applicabilité à l'IA, le FHE reste très coûteux en calcul et convient actuellement mieux à des modèles plus petits et moins complexes.

Comme alternative plus pratique, **les techniques d'obfuscation** offrent un niveau de protection modéré à un coût computationnel inférieur, bien qu'elles garantissent une sécurité plus faible. Une autre option viable est l'utilisation des **environnements d'exécution sécurisés (TEE)** — des enclaves matérielles qui isolent les calculs sensibles. Des technologies telles que Intel SGX, AMD SEV et ARM TrustZone permettent aux modèles de s'exécuter en toute sécurité, même sur des plateformes potentiellement non fiables.

Bien que les TEE soient plus matures que les implémentations FHE, elles présentent des compromis, notamment une dépendance vis-à-vis de fournisseurs matériels spécifiques et une vulnérabilité aux attaques par canaux auxiliaires. Néanmoins, elles constituent un outil précieux pour renforcer la souveraineté et la confidentialité dans les déploiements transfrontaliers d'IA.

5.4. AFFRONTER LES ATTAQUES IA SUR L'INTÉGRITÉ

Garantir l'intégrité des modèles d'IA est essentiel, en particulier dans les contextes souverains et de défense. Parmi les menaces les plus préoccupantes figurent les attaques par empoisonnement de données et les attaques par portes dérobées, qui peuvent compromettre le comportement et la fiabilité du modèle.

- **Empoisonnement de données** : consiste à manipuler délibérément les ensembles d'entraînement afin d'amener le modèle à se comporter de manière incorrecte ou imprévisible.

- **Attaques par portes dérobées** (backdoor) : introduisent des déclencheurs cachés lors de l'entraînement qui, lorsqu'ils sont activés par des entrées spécifiques, provoquent des sorties malveillantes ou non prévues. Ces portes dérobées peuvent se propager via l'apprentissage par transfert ou collaboratif, ce qui les rend particulièrement dangereuses lorsque les modèles sont réutilisés à partir de sources externes.

La stratégie de mitigation la plus efficace consiste à sécuriser la chaîne d'approvisionnement des données dès le départ. L'échantillonnage des données à partir d'environnements contrôlés réduit considérablement le risque de manipulation. Des contre-mesures supplémentaires peuvent être appliquées lors de l'inférence, notamment :

- **Désensibilisation des données**, qui filtre les entrées potentiellement nuisibles.
- **Détection des déclencheurs**, qui identifie et neutralise les schémas d'activation des portes dérobées.

Ces techniques contribuent à préserver l'intégrité des modèles et à renforcer la confiance — en particulier lorsque les modèles sont partagés ou adaptés au-delà des frontières souveraines.

5.5. ASSURER LA ROBUSTESSE DES MODELES DEPLOYÉS

Les réseaux neuronaux sont intrinsèquement vulnérables aux **exemples adversariaux** — des entrées subtilement modifiées pour induire en erreur le modèle tout en paraissant inchangées pour un observateur humain. Cette menace est particulièrement prononcée dans les applications de vision par ordinateur, où de légères altérations de pixels peuvent amener un modèle à mal classer une image — par exemple, interpréter un char comme une ambulance en raison de perturbations imperceptibles.

Ces vulnérabilités exposent les modèles déployés à des attaques à distance, surtout lorsque les adversaires disposent d'informations sur l'architecture ou les faiblesses du modèle. Cela constitue un défi majeur pour la souveraineté et la confiance, notamment dans les contextes de défense et entre alliés.

Pour atténuer ces risques, plusieurs contre-mesures peuvent être mises en œuvre :

- **Entraînement adversarial** : renforce la robustesse du modèle en l'exposant à des échantillons adversariaux durant la phase d'entraînement.
- **Prétraitement des entrées** : applique des filtres (par exemple, filtrage passe-bas) pour éliminer le bruit adversarial des entrées sans nécessiter de réentraînement du modèle.
- **Techniques d'ensembles** : combine plusieurs modèles ou algorithmes pour améliorer la fiabilité prédictive et réduire la vulnérabilité aux attaques ciblées.

Un point critique concernant les exemples adversariaux est leur **transférabilité** — une attaque réussie sur un modèle fonctionne souvent sur d'autres, même s'ils diffèrent par leur architecture. Cela souligne l'importance de la **vérification de la robustesse**, en particulier lors de l'intégration de modèles tiers dans des systèmes souverains.

5.6. SÉCURISER L'APPRENTISSAGE COLLABORATIF

Même lorsque les données brutes ne sont jamais échangées explicitement, les cadres d'apprentissage collaboratif — en particulier **l'apprentissage fédéré** (Federated Learning, FL) — doivent être renforcés par des techniques de protection de la vie privée afin de minimiser le risque de fuite d'informations via le modèle final.

Dans les configurations FL centralisées, une vulnérabilité clé réside dans le **serveur d'agrégation**, qui orchestre l'entraînement et collecte les mises à jour des participants. Ces mises à jour peuvent révéler involontairement des informations sensibles, ce qui fait du serveur une cible privilégiée pour les attaquants.

Pour atténuer ce risque, des méthodes cryptographiques avancées telles que le **chiffrement entièrement homomorphe** (FHE) ou le **calcul multipartite** (MPC) peuvent être appliquées lors de l'agrégation. Ces techniques garantissent que les contributions restent confidentielles, même si le serveur n'est pas fiable. Bien que le FHE introduise une surcharge de performance, il est bien adapté aux tâches d'agrégation, qui sont moins lourdes que l'entraînement complet ou l'inférence.

Cependant, le chiffrement seul ne supprime pas la possibilité de fuites de confidentialité à partir du modèle final. Par conséquent, des protections supplémentaires doivent être mises en œuvre pendant

l'entraînement, telles que l'injection de bruit pour anonymiser les contributions individuelles. Ces techniques de préservation de la vie privée aident à protéger les données sensibles — même dans des collaborations transfrontalières.

Thales a développé **SaferLearn**, un cadre flexible pour l'apprentissage collaboratif sécurisé. Il prend en charge plusieurs protocoles d'apprentissage et intègre des mécanismes de sécurité en couches. SaferLearn permet l'exploitation de données sensibles distribuées sans nécessiter d'entité centrale de confiance ni une architecture de modèle unifiée entre les participants.

Le cadre inclut également des outils pour :

- Mesurer la qualité des contributions individuelles.
- Détecter les tentatives d'empoisonnement dès les premières étapes de l'entraînement.

Dans les contextes sensibles à la souveraineté, SaferLearn fournit une base robuste pour l'amélioration continue des modèles exportés — en exploitant les données des utilisateurs de manière sécurisée et responsable.

5.7. PROTÉGER LES DROITS DE PROPRIÉTÉ DES FOURNISSEURS DE MODELES ET DE DONNÉES

La protection de la propriété intellectuelle des fournisseurs de modèles et de données est essentielle dans les environnements collaboratifs et orientés vers l'exportation. Deux principaux types d'attaques visant le vol de modèles menacent ces droits :

Extraction directe : les attaquants exploitent des vulnérabilités du système pour accéder aux paramètres du modèle, soit par des attaques par canaux auxiliaires, soit en les récupérant directement depuis la mémoire de l'appareil.

Entraînement de modèle substitut : les attaquants interrogent le modèle cible et utilisent les réponses comme étiquettes pour entraîner une réplique, clonant ainsi son comportement sans accéder à sa structure interne.

Pour contrer ces menaces, plusieurs stratégies de mitigation peuvent être déployées :

- Sécurisation des hyperparamètres et paramètres du modèle.
- Détection des schémas de requêtes anormales pendant l'exploitation du modèle.
- Obfuscation des sorties d'inférence pour limiter l'exposition d'informations.
- Mise en œuvre de protocoles robustes de gestion de la propriété intellectuelle (IP).

Une technique particulièrement efficace est le **watermarking appliqué au machine learning**. Inspirée du filigrane multimédia, cette méthode insère une signature secrète dans le modèle lors de l'entraînement. Cette signature peut ensuite être révélée par le propriétaire du modèle pour vérifier la paternité. Les filigranes peuvent prendre la forme de modifications subtiles des paramètres ou de déclencheurs comportementaux spécifiques activés par des entrées conçues.

La vérification de la propriété repose sur la démonstration de la connaissance du filigrane intégré — soit en l'extrayant à l'aide d'une clé secrète, soit en soumettant une entrée prédéfinie qui déclenche une réponse unique.

Dans les contextes d'apprentissage collaboratif, le watermarking peut être adapté pour suivre les contributions individuelles. Chaque participant peut intégrer une marque distincte dans le modèle partagé, fournissant une preuve de participation. Si nécessaire, ces contributions peuvent être supprimées grâce à des techniques de **désapprentissage**, permettant une gestion flexible et granulaire des actifs intellectuels partagés.

Cette approche est particulièrement précieuse lorsque la dynamique de collaboration évolue dans le temps. Thales développe et promeut des solutions de watermarking adaptées au développement d'IA multipartite sécurisé et transparent — renforçant la confiance et la responsabilité au-delà des frontières souveraines.

6. Infrastructure IT et mise en œuvre

La construction de systèmes d'IA compatibles avec la souveraineté nécessite des choix stratégiques en matière d'infrastructure. Pour permettre un déploiement et une collaboration sécurisés au-delà des frontières — sans compromettre les données nationales ni l'autonomie — les organisations doivent adopter une approche centrée sur l'infrastructure. L'architecture sous-jacente, des centres de données aux réseaux, détermine qui contrôle réellement les données et les modèles.

L'emplacement des données n'est pas un simple détail technique — c'est un levier stratégique qui influence la confiance, la conformité et l'éligibilité pour les missions sensibles ou soumises à des contrôles à l'exportation. Sans une infrastructure alignée sur les exigences de souveraineté, les nations risquent des violations réglementaires et peuvent manquer des opportunités de coopération internationale.

Conscients de cela, les gouvernements et alliances mettent en place des partenariats technologiques de confiance et intègrent des garanties de souveraineté numérique. Chez Thales, nous considérons l'investissement dans une infrastructure d'IA prête pour la souveraineté comme fondamental pour une collaboration mondiale respectant les lois nationales et l'autonomie stratégique.

Une base architecturale robuste et flexible doit prendre en charge :

- Des modèles de déploiement diversifiés.
- Des pipelines d'apprentissage fédéré et collaboratif.
- Des échanges sécurisés de données et de modèles.
- Une surveillance et une gouvernance efficaces.

Une gouvernance forte et des protocoles clairs de réponse aux incidents sont essentiels pour garantir la conformité et la responsabilité. Avec la bonne infrastructure, les organisations acquièrent la confiance nécessaire pour déployer des solutions d'IA qui respectent les exigences de souveraineté et de contrôle des exportations — libérant l'innovation sans compromis.

6.1. MATÉRIEL

La **base architecturale** d'une plateforme d'IA compatible avec la souveraineté doit être multi-couches et conçue avec soin pour répondre à la fois aux exigences de performance technique et de contrôle juridictionnel.

Elle commence par des composants matériels robustes — serveurs, CPU, GPU/TPU et infrastructures réseau — sur lesquels s'ajoutent des couches de virtualisation créant des environnements informatiques sécurisés et isolés. La containerisation et les frameworks d'orchestration (par ex. Kubernetes) gèrent les charges de travail IA de manière cohérente entre les modèles de déploiement. Au sommet de la pile, une couche AIOps rationalise l'entraînement des modèles, leur déploiement et la surveillance en temps réel, garantissant l'excellence opérationnelle.

Cette architecture modulaire permet d'appliquer des politiques à chaque couche — des hyperviseurs aux maillages de services — en contrôlant efficacement les flux de données transfrontaliers. Les zones de confiance doivent être alignées sur les frontières de souveraineté, permettant la création d'environnements spécifiques au projet, à l'entreprise, au partenaire, au niveau national ou international, reflétant le niveau de confiance et de supervision juridique requis. Avec des frontières de souveraineté clairement définies, les organisations peuvent garantir que les données et actifs sensibles restent pleinement protégés.

Le choix stratégique des composants matériels est essentiel. Opter pour des CPU, GPU et TPU conformes aux contrôles à l'exportation et aux exigences de résidence des données aide à maintenir la souveraineté. L'exécution sécurisée et la résilience matérielle doivent être prioritaires pour garantir l'intégrité et la sécurité des systèmes d'IA.

Les nations leaders adoptent de plus en plus **les environnements d'exécution sécurisés** (TEE) — des enclaves protégées intégrées aux CPU ou aux architectures système — pour protéger les modèles et données sensibles pendant l'exécution. Des technologies comme Intel SGX, AMD SEV et ARM TrustZone créent des zones mémoire isolées et chiffrées où le code peut s'exécuter en toute sécurité, même vis-à-vis du système d'exploitation hôte. Cela permet de déployer des modèles sur des serveurs étrangers ou non fiables sans exposer leurs paramètres internes.

Des protections similaires apparaissent dans les accélérateurs spécifiques à l'IA : certains GPU prennent désormais en charge la mémoire chiffrée ou l'isolation au niveau de l'hyperviseur, étendant les capacités de calcul confidentiel aux matériels IA haute performance.

Dans certains contextes, opter pour une exécution uniquement sur CPU peut être un choix souverain. Bien que l'entraînement avancé des IA nécessite généralement une accélération GPU/TPU, ces composants sont souvent soumis à des restrictions d'exportation et à des risques liés à la chaîne d'approvisionnement. Pour des applications moins critiques ou moins gourmandes en ressources, les environnements basés sur CPU offrent une alternative viable — réduisant la dépendance étrangère tout en maintenant le contrôle opérationnel.

En définitive, les décisions matérielles doivent être guidées par une stratégie de segmentation des risques : GPU haut de gamme pour les projets approuvés et surveillés, CPU ou composants hérités pour les autres. Cette approche équilibre performance, souveraineté et conformité dans des scénarios opérationnels variés.

6.2. MODELE DE DÉPLOIEMENT

Les modèles de déploiement sont au cœur de la construction de plateformes d'IA conformes aux exigences de souveraineté. Chaque modèle offre des capacités distinctes tout en présentant ses propres défis.

On-Premises / Cloud Souverain

Ce modèle garantit que les charges de travail IA restent entièrement à l'intérieur des frontières nationales, en s'appuyant sur des centres de données gouvernementaux ou des clouds souverains certifiés. Il assure la conformité avec les réglementations nationales et les mandats de souveraineté des données. Cette approche offre un contrôle et une sécurité maximaux — une exigence essentielle pour les domaines sensibles tels que le gouvernement, la défense et les infrastructures critiques. Cependant, elle s'accompagne souvent d'une complexité opérationnelle accrue et d'une évolutivité limitée.

Déploiement Hybride

Le modèle hybride combine l'infrastructure nationale avec des ressources cloud mondiales sélectionnées. Les données sensibles sont traitées localement, tandis que les tâches moins critiques ou gourmandes en calcul sont déléguées à des partenaires cloud étrangers de confiance. Les gouvernements collaborent fréquemment avec des fournisseurs majeurs pour établir des clouds internes qui associent contrôle local et élasticité. Grâce à une partition stricte des données, des protections techniques robustes et une intégration avec des régions approuvées, ce modèle garantit que les données protégées restent sécurisées. Il équilibre efficacité opérationnelle et souveraineté en exploitant de manière sélective les capacités des clouds mondiaux.

Déploiement Centré sur l'Edge

Lorsque les solutions cloud centralisées ou partenaires ne sont pas viables, le **edge computing** devient une alternative stratégique. Ce modèle décentralise le traitement IA vers des points locaux — tels que des dispositifs IoT, des unités de terrain ou des stations de base 5G — rapprochant le calcul des lieux où les données sont générées. Il prend en charge une latence ultra-faible et respecte les politiques « les données restent locales », permettant une réactivité en temps réel et une continuité opérationnelle même dans des environnements déconnectés. Bien que la gestion des mises à jour sur des nœuds distribués pose des défis, des piles logicielles conteneurisées et orchestrées avec des mises à jour OTA (over-the-air) offrent une solution efficace. En transmettant uniquement des **informations synthétisées** plutôt que des données brutes, ce modèle renforce la souveraineté et complète les stratégies de cloud souverain.

6.3. GOUVERNANCE

Une gouvernance efficace est une pierre angulaire pour une collaboration en IA sécurisée et souveraine. L'utilisation de modèles de conteneurs permet la mise en œuvre de contrôles fins d'adhésion et de retrait, ainsi que de mécanismes robustes pour faire respecter le « droit à l'oubli ». Ces capacités donnent aux contributeurs la possibilité de retirer leurs données ou leurs modèles à tout moment, garantissant leur suppression complète et vérifiable des environnements collaboratifs.

Des fonctionnalités avancées telles que la signature cryptographique, le suivi de provenance, les registres de modèles immuables assurent une attribution transparente des contributions et établissent une chaîne de responsabilité auditable tout au long du cycle de développement de l'IA.

En adoptant une approche Policy-as-Code, les organisations peuvent automatiser l'application des règles de participation, traduisant sans friction les accords juridiques et les cadres de partenariat en contrôles techniques exécutables.

Ce cadre de gouvernance complet garantit que chaque initiative collaborative en IA reste conforme aux politiques applicables, respecte l'autonomie et les obligations légales de toutes les parties prenantes, et facilite une reconnaissance transparente et une gestion des contributions individuelles. En intégrant ces pratiques, les organisations peuvent construire des systèmes d'IA collaboratifs sécurisés, conformes et pleinement responsables par conception.

6.4. ÉCHANGES DE DONNÉES ET DE MODELES

Les mécanismes sécurisés d'échange de données et de modèles sont fondamentaux pour la construction de plateformes d'IA compatibles avec la souveraineté — en particulier dans les environnements d'apprentissage fédéré et collaboratif. Garantir la confiance lors de l'échange d'ensembles de données et de modèles nécessite des mesures de sécurité robustes de bout en bout.

Chaque ensemble de données ou modèle, y compris ses poids, est encapsulé et chiffré à l'aide d'algorithmes de chiffrement symétrique puissants tels que AES-256. Pour garantir l'authenticité et l'intégrité, le package est signé numériquement avec la clé privée de l'expéditeur (par exemple RSA ou ECC). Cela assure que seuls les destinataires autorisés disposant des clés de déchiffrement appropriées peuvent accéder au contenu, tandis que toute altération pendant la transmission est immédiatement détectable via la vérification de la signature.

Le package chiffré inclut une signature cryptographique, permettant aux destinataires de confirmer l'origine et l'intégrité des données ou du modèle. Cette approche garantit que, même lorsqu'ils sont transmis via des réseaux non fiables ou des infrastructures cloud tierces, la confidentialité et la provenance des actifs restent intactes.

En intégrant ces pratiques de sécurité au cœur des workflows collaboratifs en IA, les organisations peuvent instaurer la confiance, préserver la souveraineté des données et réduire les risques — jetant ainsi les bases d'écosystèmes d'IA sécurisés, conformes et résilients.

6.5. SURVEILLANCE ET GESTION DES INCIDENTS

Une surveillance robuste, une réponse efficace aux incidents et une gouvernance solide sont des piliers fondamentaux d'une plateforme d'IA compatible avec la souveraineté — en particulier lors du déploiement des modèles en production. Ces capacités garantissent l'intégrité opérationnelle, la sécurité et la conformité réglementaire, permettant des solutions d'IA fiables pour les partenaires et les clients.

Même avec des contrôles préventifs solides, des défis tels que la dérive des modèles, l'empoisonnement des données ou les attaques adversariales peuvent encore survenir. Cela rend la surveillance en temps réel du comportement et des performances des modèles essentielle comme ligne de défense critique. Une observation continue permet aux organisations de détecter les changements dans les distributions de données, l'apparition de biais ou des activités anormales pouvant signaler des menaces de sécurité. En s'appuyant sur des plateformes cloud sécurisées et des systèmes AIOps — tels que ceux proposés par Thales — les organisations peuvent accéder à des tableaux de bord automatisés et à des alertes proactives conçues pour identifier des risques comme les tentatives d'extraction de modèles ou les fuites potentielles de données.

La détection rapide doit être associée à une réponse rapide et coordonnée. L'intégration des plateformes d'IA avec le Security Operations Center (SOC) global est essentielle. Les systèmes d'IA doivent émettre des événements de sécurité directement dans l'infrastructure SIEM (Security Information and Event Management) de l'organisation, permettant une visibilité et une réponse centralisées. Les protections matérielles, y compris les modules de sécurité matériels (HSM) et les modules de plateforme sécurisés (TPM), renforcent la sécurité en empêchant l'extraction non autorisée des modèles et en garantissant l'intégrité des données. Ces composants peuvent également générer des signaux de sécurité au niveau matériel pour alerter le SOC en cas de violation.

Une exploitation efficace de l'IA nécessite également des plans de réponse aux incidents (IR) bien définis, adaptés aux complexités des frontières juridictionnelles. Ces plans doivent clairement décrire les protocoles de notification, le partage d'informations autorisé et les chemins d'escalade entre différents cadres légaux. Dans les environnements fédérés impliquant plusieurs parties

prenantes et juridictions, la gestion des incidents devient encore plus complexe. L'accès transfrontalier aux données peut entrer en conflit avec les lois de souveraineté, ce qui rend nécessaire des analyses forensiques localisées et des mini-SOC régionaux. Une gouvernance claire sur les transferts de responsabilité et la coordination des réponses conjointes garantit la conformité et la continuité opérationnelle.

En définitive, permettre une collaboration en IA compatible avec la souveraineté exige une infrastructure résiliente, des modèles de déploiement flexibles et des capacités avancées d'apprentissage fédéré.

7. Cadres juridiques et politiques pour protéger la souveraineté

Le paysage mondial de la gouvernance de l'IA évolue rapidement, avec des variations régionales significatives reflétant des priorités culturelles, économiques et politiques différentes. Bien que toutes les régions cherchent à équilibrer innovation et déploiement responsable, leurs approches divergent en termes de structure et d'accent.

Europe : L'UE est en tête avec l'AI Act, un cadre réglementaire complet basé sur les risques, qui entrera en vigueur en août 2026. Il établit des obligations claires en fonction du niveau de risque des systèmes d'IA, fixant une référence mondiale pour une IA responsable.

États-Unis : Approche plus décentralisée et sectorielle, axée sur des normes de sécurité et des lignes directrices volontaires, tout en favorisant l'innovation par la collaboration public-privé.

Canada : Projet de loi AI and Data Act, qui adopte un modèle basé sur les risques, soutenu par des codes de pratique volontaires et des investissements fédéraux dans la R&D en IA.

Royaume-Uni : Maintient un modèle réglementaire flexible et sectoriel, bien que des discussions soient en cours pour évoluer vers une structure plus unifiée.

Asie : Paysage réglementaire diversifié :

Chine : Met l'accent sur la transparence, la responsabilité algorithmique et la protection des données.

Singapour et Japon : Développent des cadres réglementaires structurés favorisant l'innovation tout en garantissant une utilisation éthique de l'IA.

Corée du Sud : La AI Framework Act, en vigueur depuis janvier 2026, s'aligne étroitement sur les principes européens, combinant supervision réglementaire et soutien fort à l'innovation industrielle.

Malgré ces différences, des principes communs émergent : transparence, équité, protection des données, responsabilité et gestion des risques. Ces

valeurs partagées visent à garantir que le déploiement de l'IA s'aligne sur le bien-être sociétal et les intérêts nationaux.

Coopération mondiale et souveraineté de l'IA : trouver l'équilibre

Une gouvernance efficace de l'IA nécessite une action coordonnée entre les parties prenantes pour façonner l'avenir de l'IA de manière à soutenir la souveraineté, la sécurité nationale, la compétitivité économique et le bien-être sociétal. Cependant, la poursuite unilatérale de la souveraineté par des États individuels risque de fragmenter les normes mondiales et de compromettre la coopération internationale. Les politiques protectionnistes peuvent freiner le commerce bénéfique, limiter la collaboration en recherche et fausser la concurrence équitable. Ainsi, la coopération internationale est essentielle pour garantir l'interopérabilité des normes techniques et l'alignement des cadres de gouvernance.

Les organisations multilatérales — G7, OCDE, UNESCO, NIST, ISO, Conseil de l'Europe — travaillent activement à établir des principes communs autour de la sécurité, la transparence, l'éthique et la sûreté. Le débat sur les normes mondiales de l'IA est devenu un enjeu géopolitique majeur, où le défi consiste à équilibrer la souveraineté nationale avec la nécessité d'une coordination et d'une harmonisation globales.

La souveraineté numérique elle-même ne se limite pas au contrôle des données — elle inclut la supervision réglementaire des plateformes numériques et la capacité à façonner les écosystèmes technologiques domestiques. Des économies majeures comme l'Union européenne affirment leur souveraineté numérique via des cadres réglementaires complets tels que l'AI Act, qui harmonise les standards éthiques et juridiques entre États membres.

Perspectives internationales et initiatives clés

Les initiatives transnationales se multiplient pour relever les défis communs de la gouvernance de l'IA :

Directives mondiales de l'UNESCO sur l'éthique de l'IA, visant à établir des principes universellement acceptés.

Collaboration de l'UE avec des ONG comme l'IEEE pour promouvoir un développement responsable de l'IA.

AI Alliance, lancée en 2023 par IBM et Meta, regroupant aujourd'hui 140 membres dans 23 pays, pour promouvoir une innovation responsable fondée sur l'intégrité scientifique, la confiance, la diversité, la sécurité et la compétitivité.

Impact AI, un « think-and-do tank » français, réunissant des acteurs pour favoriser des pratiques éthiques et inclusives en IA.

Thales, en tant que membre fondateur de l'Association européenne pour une IA de confiance, joue un rôle clé en aidant les industriels à maîtriser les technologies critiques de l'IA, renforçant ainsi l'autonomie technologique, la compétitivité et la souveraineté.

Ces efforts collectifs soulignent l'importance de bâtir une base commune pour la gouvernance de l'IA — respectant les intérêts nationaux tout en favorisant la confiance mondiale, l'innovation et l'alignement éthique.

8. A propos de Thales

Thales est directement impliqué dans les défis et opportunités décrits dans ce document à travers ses activités variées. Depuis plusieurs années, l'intelligence artificielle est intégrée dans l'ensemble de son portefeuille de produits — en particulier dans les systèmes critiques et les solutions de défense déployées ou destinées à être déployées dans de nombreux pays à travers le monde. Cette expérience opérationnelle a permis aux équipes de recherche et d'ingénierie de Thales de développer des composants, des outils et des méthodologies IA adaptés à la fois aux contextes souverains et internationaux.

Certaines de ces innovations ont été mentionnées dans les sections précédentes pour aider l'écosystème à mieux comprendre les complexités de l'IA souveraine et les solutions pratiques qui peuvent être mises en œuvre.

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 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 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

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