RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

ANALYSE COMMUNE HAUT NIVEAU DES RISQUES CYBER LIÉS À L'IA

Développer la confiance dans l'IA à travers une approche par les risques cyber

PARIS AI ACTION SUMMIT – TRUST IN AI – CYBERSECURITY

Cette publication menée sous l'égide de l'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI) résulte de travaux collaboratifs avec des experts institutionnels français du domaine (CNIL, INRIA, LNE, PEReN, AMIAD) et a été cosignée dans sa version anglaise par les partenaires internationaux suivants :

- Office fédéral allemand de sécurité des technologies de l'information (BSI)
- Centre canadien pour la cybersécurité (CCCS)
- Service national de renseignement de Corée du Sud (NIS)
- Autorité des systèmes d'informations d'Estonie (RIA)
- Centre national de cybersécurité irlandais (NCSC-IE)
- Agence nationale de cybersécurité italienne (ACN)
- Maison luxembourgeoise de la cybersécurité (LHC)
- Autorité Maltaise de l'innovation numérique (MDIA)
- Autorité de sécurité nationale norvégienne (NSM)
- Centre national de cybersécurité des Pays-Bas (NCSC-NL)
- Service général de Renseignement et de Sécurité des Pays-Bas (AIVD)
- National Research Institute polonais (NASK)
- Centre national de cybersécurité du Royaume-Uni (NCSC-UK)
- Agence de cybersécurité de Singapour (CSA)
- Centre national de cybersécurité de Slovaquie (NCSC-SK)
- Government Information Security Office de Slovénie (URSIV).

Les informations contenues dans ce document sont fournies « en l'état » et n'ont vocation qu'à alimenter les échanges sur les risques et opportunités de l'intelligence artificielle.

L'ANSSI et les organisations cosignataires ne sauraient par conséquent être tenues responsables de toute perte, préjudice ou dommage de quelque nature que ce soit causé par leur utilisation. Les informations contenues dans ce document ne constituent ni n'impliquent l'approbation ou la recommandation par l'ANSSI et les organisations cosignataires d'entités, produits ou services tiers. Les liens et les références à des sites web et documents tiers sont fournis à titre purement informatif et n'impliquent ni l'approbation ni la recommandation de ces ressources par rapport à d'autres.


AMIAD
Agence ministérielle pour l'intelligence artificielle de défense


CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS


Inria


LABORATOIRE
NATIONAL
DE MÉTROLOGIE
ET D'ESSAIS
LNE


GOUVERNEMENT
Liberté
Égalité
Fraternité
Pôle d'Expertise de la
Régulation Numérique

DÉVELOPPER LA CONFIANCE DANS L'IA À TRAVERS UNE APPROCHE PAR LES RISQUES CYBER

Les agences internationales et autorités gouvernementales cosignataires du présent document soutiennent une approche par les risques afin de favoriser l'usage des systèmes d'IA de confiance et de rendre plus sûre leur chaine de valeur. Ils appellent à poursuivre les discussions au-delà du Sommet IA, en ayant une approche équilibrée prenant en considération les opportunités, les risques de l'IA et l'évolution de la menace cyber dans le cadre de l'adoption de cette technologie.

En constant développement depuis les années 1950, l'intelligence artificielle (IA) est une technologie susceptible d'influencer tous les secteurs, de la défense à l'énergie, en passant par la santé et la finance, etc. L'adoption rapide de l'IA et la dépendance accrue à ces technologies, notamment l'usage des modèles de langue (LLM)[1], devraient encourager les parties prenantes à évaluer les risques connexes, dont ceux liés à la cybersécurité.

Sans mesures adéquates, et compte tenu du fait que les utilisateurs tendent toujours à sous-estimer l'importance des risques cyber, des acteurs malveillants pourraient exploiter les vulnérabilités des technologies d'IA et à l'avenir compromettre leur usage. **Il est donc impératif de comprendre et limiter ces risques, afin de promouvoir le développement de l'IA de confiance et d'en saisir pleinement les opportunités.**

Comme tout système logiciel, les systèmes d'IA[2] présentent des vulnérabilités. Ils nécessitent par conséquent d'être sécurisés dès leur conception, en respectant en particulier les bonnes pratiques existantes en cybersécurité, que ce soit pour le développement, le déploiement, la gestion des vulnérabilités et de la chaine d'approvisionnement logiciel tant propriétaire qu'open source. Ils font face aux mêmes menaces cyber que tout autre système d'information, notamment à travers leur infrastructure d'hébergement, et leur interconnexion à d'autres systèmes augmente les risques de latéralisation. Il existe également des risques spécifiques à l'IA : le rôle central des données au sein de ces systèmes pose des défis singuliers en matière de confidentialité et d'intégrité.

Si les bonnes pratiques d'hygiène numérique demeurent efficaces, l'émergence d'une cyberdéfense s'appuyant sur les technologies d'IA de confiance est un enjeu important face à de nouvelles menaces en constante évolution. Bien qu'elles ne soient pas l'objet de ce document, ces solutions contribuent d'ores et déjà au renforcement des capacités de cyberdéfense et continueront de se développer à l'avenir, non seulement en matière de supervision et de détection d'intrusion, mais aussi en termes d'analyse de la menace, de réponse à incident et d'investigation numérique, d'automatisation des processus de sécurité, etc.

En parallèle, l'utilisation malveillante de l'IA est susceptible de s'étendre et de se perfectionner. L'IA réduit l'expertise requise pour mener certaines attaques et augmente leur étendue ou leur efficacité. Ces effets sont d'ores et déjà observés dans les domaines du phishing, de l'ingénierie sociale, de la recherche de vulnérabilités et du développement de codes malveillants. L'IA générative avancée pourrait rendre possible des attaques de grande échelle sur l'ensemble de la cyber-kill-chain, et ce à moindre coût. Par conséquent, il est nécessaire de suivre l'évolution de l'utilisation malveillante de l'IA.

Si le sujet des solutions d'IA défensives et offensives est déjà largement abordé dans la littérature scientifique et dans les référentiels existants et en cours

de développement, ce document met l'accent sur la cybersécurité des systèmes d'IA et constitue une analyse haut niveau des risques cyber. Il présente de premiers éléments pour permettre à chaque organisation d'apprécier les risques cyber liés aux systèmes IA et d'identifier les mesures de sécurité à mettre en œuvre pour les maitriser, en s'appuyant sur les Lignes directrices pour le développement sécurisé des systèmes d'IA, élaborées en collaboration avec plus de 20 organisations internationales et publiées conjointement en novembre 2023[3].

1. LLM : Large Language Model, modèle d'IA générative utilisé pour le traitement du langage et la génération de langage naturel.

2. Dans ce document, un système d'IA est défini comme un système logiciel qui repose sur un modèle d'IA construit par un apprentissage statistique à partir d'un jeu de données d'entraînement.

3. Cybersecurity and Infrastructure Security Agency (CISA) and National Cyber Security Centre (NCSC), Secure AI Systems Development Guidelines, November 26, 2023.

ENJEUX ET PERIMETRE D'UNE ANALYSE DE RISQUE IA

Cette analyse des risques vise les composants individuels de l'IA, mais également la sécurité des systèmes d'IA plus larges intégrant ces composants. Ce document propose une synthèse globale des risques pesant sur ces systèmes plutôt qu'une liste exhaustive de leurs vulnérabilités. L'annexe 2 contient une liste de références détaillant notamment ces vulnérabilités.

Le déploiement d'un système d'IA peut offrir de nouveaux chemins d'attaque pour les acteurs malveillants si des mesures de sécurité adéquates ne sont pas mises en place. Il convient donc de mener une analyse de risque avant tout déploiement de système d'IA, afin d'apprécier les risques cyber et d'identifier les mesures de sécurité à mettre en œuvre pour les maitriser.

La connaissance et l'étude des chaines d'approvisionnement de l'IA est essentielle pour limiter les risques liés aux vulnérabilités spécifiques portées par les fournisseurs et autres parties prenantes de tout système d'IA. Ces chaines d'approvisionnement reposent généralement sur trois piliers :

1. la capacité de calcul ;
2. les modèles d'IA et leurs dépendances logicielles associées ;
3. Data.

Chacun de ces piliers implique des acteurs parfois distincts, parfois communs, et dont la maturité en matière de cybersécurité peut varier considérablement.

PRINCIPAUX RISQUES ET SCENARIOS D'ATTAQUE

L'infrastructure socle sur laquelle repose un système d'IA conserve les vulnérabilités inhérentes à toute architecture informatique. Un système d'IA peut aussi être compromis à différentes étapes de son cycle de vie, depuis la collecte des données brutes jusqu'à la phase d'inférence. Les attaques associées peuvent généralement se répartir selon trois grandes familles :

- **empoisonnement :** altération des données d'entraînement ou du modèle affectant la réponse du système d'IA à toutes les entrées ou à une entrée spécifique ;
- **extraction :** reconstruction ou récupération de données confidentielles du système d'IA ou du modèle après la phase d'apprentissage (paramètres et configuration du modèle, données d'entrainement) ;
- **évasion :** altération des données d'entrée afin de modifier le fonctionnement attendu du système d'IA.

De telles attaques peuvent aboutir au dysfonctionnement d'un système d'IA (risques de disponibilité et d'intégrité), ce qui compromettrait la fiabilité des décisions ou processus automatisés, voire au vol ou à la divulgation de données sensibles (risques de confidentialité).

L'opacité des systèmes d'IA modernes représente un défi supplémentaire qui doit impérativement être pris en compte par les organisations. Le niveau d'explicabilité de ces systèmes varie considérablement selon la nature du modèle sous-jacent : certains fonctionnent comme des « boîtes noires », ce qui peut complexifier l'interprétation et la justification de leurs décisions. Ce manque de transparence rend plus difficile l'identification et l'investigation d'incidents potentiels, compliquant ainsi les efforts de sécurisation.

Les principaux scénarios de risques impliquant un système d'IA sont les suivants :

- **Compromission de l'infrastructure d'hébergement et d'administration des systèmes d'IA :** des acteurs malveillants peuvent compromettre la confidentialité, l'intégrité et la disponibilité d'un système d'IA en exploitant un large éventail de vulnérabilités courantes, qu'elles soient d'ordre technique, organisationnel ou humain. La compromission de l'infrastructure d'hébergement d'un système d'IA représente un vecteur d'attaque particulièrement plausible et critique, et doit à ce titre être prise en compte pendant tout le cycle de vie du système d'IA.

- **Compromission de la chaine d'approvisionnement :** un attaquant peut exploiter une vulnérabilité sur une des parties prenantes de la chaine d'approvisionnement[4]. Par exemple, des bibliothèques open source sont souvent utilisées dans le développement de systèmes d'IA, et souvent intégrées dans des frameworks plus larges. Une attaque sur ces bibliothèques pourrait compromettre l'ensemble du système d'IA.

- **Latéralisation via les interconnexions entre les systèmes d'IA et les autres systèmes :** les systèmes d'IA sont souvent interconnectés à d'autres SI pour permettre une communication fluide et une intégration efficace des données. Cependant, ces interconnexions peuvent poser de nouveaux risques. A titre d'exemple, des attaques comme l'injection indirecte de prompt, qui exploitent les LLM en introduisant des instructions malveillantes via des sources externes contrôlées par un attaquant, peuvent être utilisées pour extraire des informations sensibles ou exécuter des commandes malveillantes à distance. Ce risque est d'autant plus important en cas d'interconnexion avec des SI industriels, ces derniers permettant une interaction directe avec le monde physique.

- **Lacunes humaines et organisationnelles :** un manque de formation peut entraîner une dépendance excessive à l'automatisation et une capacité insuffisante à détecter les comportements anormaux des systèmes d'IA. En outre, l'usage de shadow IA[5] au sein des organisations est susceptible d'exacerber cette perte de maîtrise et d'augmenter certains risques : fuite de données confidentielles, violation de réglementations, atteinte à la réputation en nuisant à l'image de l'organisation, etc. A long termen l'utilisation intensive et prolongée de l'IA pourrait entrainer un risque de dépendance technologique et, dans le cas d'une éventuelle défaillance, ces technologies ne pourraient alors probablement plus être remplacées par une action humaine. Ce risque est prépondérant lorsque des systèmes d'IA sont impliqués dans des activités critiques (ex. environnements industriels), notamment en cas d'automatisation forte des processus métier.

- **Dysfonctionnement dans les réponses apportées par un système d'IA :** un attaquant pourrait compromettre une base de données utilisée pour l'entraînement d'un modèle d'IA afin de provoquer des réponses erronées une fois celui-ci en production. Ce type d'attaque requiert un effort certain de la part de l'attaquant, car les développeurs de modèles d'IA intègrent déjà des pratiques qui tendent à améliorer la résilience des systèmes face aux empoisonnements volontaires et malveillants des données d'entraînement. Cependant, il peut se montrer particulièrement dangereux s'il est, par exemple, employé pour catégoriser des données telles que des images utilisées dans les domaines de la santé ou de la sécurité physique.

4. Bibliothèques logicielles, fournisseurs de modèles pré-entrainés, prestataires de services, etc.

5. Le « shadow IA » se définit par l'utilisation de solutions d'IA générative grand public sans l'approbation ou la supervision des services informatiques de l'organisation.

PRINCIPALES RECOMMANDATIONS POUR LES UTILISATEURS, LES FOURNISSEURS ET LES DEVELOPPEURS DE SYSTEMES D'IA

Lorsque que le déploiement d'un système d'IA est envisagé, la sensibilité du cas d'utilisation doit être évaluée et prise en compte. La complexité, la maturité cyber, l'auditabilité, et l'explicabilité, du système d'IA doivent en effet correspondre aux exigences portées par le cas d'utilisation en matière de cybersécurité et de confidentialité des données.

Dans le cadre du développement, du déploiement ou de l'utilisation d'une solution d'IA, les recommandations suivantes constituent un ensemble de bonnes pratiques à destination des utilisateurs, fournisseurs et développeurs :

- **Le niveau d'autonomie du système d'IA doit être ajusté en fonction de l'analyse de risque, du besoin métier, de la criticité des actions entreprises.** La validation humaine doit être intégrée dès que nécessaire dans ce processus. Cela aidera par ailleurs à traiter les risques qui peuvent affecter la fiabilité du système d'IA ;

- **Une cartographie de la chaine d'approvisionnement** doit être réalisée, en intégrant à la fois les **composants d'IA et les autres composants matériels et logiciels**, ainsi que les jeux de données (nature, approvisionnement et traitement – notamment pour limiter les risques d'empoisonnement et évaluer l'impact des risques liés à l'extraction) ;

- **Les interconnexions entre le système d'IA et tout autre système d'information doivent être recensées et limitées aux besoins** liés au cas d'utilisation, afin de réduire autant que possible les chemins d'attaque ;

- **Les systèmes d'IA doivent être supervisés et maintenus en continu,** pour s'assurer qu'ils fonctionnent comme prévu, sans biais ni vulnérabilité qui pourraient avoir un impact sur la cybersécurité, atténuant ainsi les risques liés à la nature de «boîte noire» de certains systèmes d'IA ;

- **Un processus de veille technologique et réglementaire** doit être mis en place afin d'anticiper les évolutions du contexte, **d'identifier les nouvelles menaces,** et d'adapter les stratégies pour prendre en compte les enjeux futurs ;

- Il est indispensable de **former et sensibiliser** aux enjeux et aux risques de l'intelligence artificielle, **en interne** et notamment au niveau des directions afin de garantir que les décisions d'implémentation sont éclairées.

Pour les utilisateurs, fournisseurs et développeurs de systèmes d'IA, une check-list des mesures recommandées est proposée en annexe.

RECOMMANDATIONS A L'ATTENTION DES DECIDEURS

En tenant compte des contextes régionaux et nationaux, les décideurs devraient avoir pour objectif de :

- **Soutenir la recherche sur de nouvelles méthodes pour atténuer ces risques,** notamment dans le domaine de l'adversarial machine learning (attaques spécifiques à l'IA et les défenses contre ces attaques), les technologies de protection des données et les cas d'usage offensifs émergents de l'IA.

- **Promouvoir le développement des évaluations de sécurité et des capacités de certification,** sur la base de standards communs afin de renforcer la confiance dans les modèles d'IA, les applications, les données et les infrastructures.

- **Promouvoir des bonnes pratiques de cybersécurité afin d'assurer le déploiement et l'hébergement sécurisé des systèmes d'IA,** en établissant des lignes directrices claires, en valorisant les réglementations existantes et applicables à ces systèmes et en partageant des retours d'expérience, permettant ainsi aux organisations d'éviter les erreurs courantes et d'optimiser l'intégration de l'IA dans leurs opérations.

- **Favoriser le dialogue des gouvernances cyber et IA** en renforçant notamment la coordination entre les agences de cybersécurité et celles spécialisées en intelligence artificielle (les AI Safety Institutes, par exemple). Cela leur permettra de définir leurs périmètres de responsabilité et ainsi de favoriser une meilleure prise en compte des enjeux cyber des systèmes d'IA. Cette collaboration facilitera le partage d'informations cruciales sur les menaces émergentes et l'harmonisation des efforts pour protéger les systèmes critiques.

- **Poursuivre les travaux au-delà du Sommet IA** en assurant notamment un suivi de l'évolution de la menace qui pèse sur les systèmes d'IA, et en invitant à une réflexion commune à l'échelle internationale pour l'identification de lignes directrices afin de mieux sécuriser l'ensemble de la chaine de valeur et ainsi favoriser une IA de confiance.

→ **ANNEXE I**

RECOMMANDATIONS POUR L'IMPLÉMENTATION SÉCURISÉE D'UN SYSTÈME D'IA

Ces recommandations offrent une approche globale sans prétendre à l'exhaustivité. Des référentiels complémentaires sont fournis en annexe 2 pour approfondir la démarche.

1. Les questions à se poser :

• Ai-je bien défini et documenté la/les finalité(s), explicite(s) et légitime(s) de mon système et ce dès la phase de développement si possible ?

• Ai-je bien intégré les aspects réglementaires dans mes réflexions ? Me suis-je bien assuré que le traitement envisagé par mon système d'IA est bien conforme au cadre légal et réglementaire en vigueur ?

• Qui a accès au système d'IA durant les différentes phases de son cycle de vie ?

• Est-ce que le principe du moindre privilège est appliqué afin de garantir la sécurité et l'intégrité du système d'IA ?

• Quelle est la chaine de dépendance du système d'IA ?

• Quelle est la réputation de mes fournisseurs et quelle est leur santé financière ?

• Est-ce que mes fournisseurs respectent les normes de sécurité, que ce soit des fournisseurs de données ou des fournisseurs de composants logiciels ?

• Est-ce qu'il est nécessaire d'implémenter une solution cloud ? Ai-je réalisé une analyse de risque globale des conséquences associées ? (sur la protection des données, etc.).

• Ai-je bien dans ma convention de service avec un prestataire pouvant manipuler mes données une clause de réversibilité ? Est-ce que la réversibilité est techniquement (moyens ou débit de transfert de données) et chronologiquement réalisable ?

• Quels sont les impacts de l'utilisation de l'IA sur le métier ? Un dysfonctionnement de l'IA peut-il mettre en danger mon organisation ?

• Y a-t-il un socle de sécurité prévu à chaque étape du cycle de vie du système d'IA (guides et référentiels de bonnes pratiques, cartographie, etc.) ?

• Mes modèles d'IA doivent-ils être protégés en confidentialité ? Est-ce qu'ils représentent une valeur importante pour mon organisation ?

• Ai-je bien intégré des séries de mesures pour intégrer dès la conception les principes de protection des données personnelles (privacy by design), cela tant pour les données et métadonnées que pour le(s) modèle(s) du système d'IA ?

2. Check-list des mesures recommandées

Recommandations générales :

☐ limiter l'usage automatisé de système d'IA pour des actions critiques sur d'autres SI ;

☐ veiller à ce que l'IA soit intégrée de manière réfléchie et appropriée dans les processus critiques et prévoir des garde-fous ;

☐ réaliser une analyse de risque dédiée en intégrant l'ensemble du contexte de l'organisation (Par exemple, l'impact d'une défaillance d'un système d'IA devrait être évalué à l'échelle de l'ensemble de l'organisation) ;

- ☐ étudier la sécurité de chaque étape du cycle de vie du système d'IA (de la collecte de données d'entrainement à la phase de décommissionnement en passant par la phase d'inférence) ;
- ☐ réaliser une analyse d'impact sur la protection des données si nécessaire ;
- ☐ identifier, suivre et protéger les composants nécessaires au modèle d'IA.

Recommandation pour l'infrastructure et l'architecture :

- ☐ définir les modalités d'utilisation du système d'IA et encadrer son intégration dans le processus décisionnel, en particulier en cas d'automatisation ;
- ☐ appliquer les mesures spécifiques aux environnements cloud si concerné en tenant compte des réglementations applicables et des politiques organisationnelles ;
- ☐ appliquer les recommandations relatives à l'infogérance si concerné ;
- ☐ appliquer les recommandations d'administration sécurisée sur le système d'IA ;
- ☐ mettre en place un système de contrôle d'accès pour les composants critiques du système d'IA.

Avoir un plan de déploiement

- ☐ concevoir l'architecture en prévoyant le passage à l'échelle (phase d'inférence) de manière à ce qu'il n'y ait pas de dégradation du niveau de sécurité ;
- ☐ appliquer les principes de DevSecOps sur l'ensemble des phases du projet ;
- ☐ concevoir le système d'IA en adoptant une approche privacy by design permettant de satisfaire les impératifs de protection des données tout au long du cycle de vie :
  - o prendre en compte les enjeux de confidentialités des données ;
  - o s'assurer de la pseudonymisation ou de l'anonymisation des données si nécessaire ;
  - o prendre en compte la problématique du besoin d'en connaitre dès la conception de système d'IA.

Être vigilant aux ressources utilisées

- ☐ utiliser des formats sécurisés pour le stockage et la distribution des modèles d'IA ;
- ☐ mettre en place des mécanismes de vérification de l'intégrité des fichiers de modèles avant leur chargement ;
- ☐ évaluer le niveau de confiance des bibliothèques et des modules externes utilisés dans les systèmes d'IA ;
- ☐ s'assurer de la qualité et évaluer le niveau de confiance des données externes utilisées dans le système d'IA ;
- ☐ s'assurer de la traçabilité des actions réalisées sur le système d'IA ;
- ☐ s'assurer que la collecte des données a été réalisée de façon loyale et éthique, pour celles utilisées tant pour le développement que pour l'utilisation du système.

Sécuriser et durcir le processus d'apprentissage

- ☐ adopter une politique stricte relative aux accès aux données par le système d'IA, dont particulièrement les données sensibles ;
- ☐ sécuriser le stockage des données d'entrainement ;
- ☐ évaluer la sécurité des méthodes d'apprentissage et de réapprentissage utilisées ;
- ☐ mettre en œuvre des mesures sur les données,

métadonnées, annotations et caractéristiques extraites mais aussi le(s) modèle(s) du système d'IA dont :

- nettoyer les données ;
- identifier les données pertinentes et strictement nécessaires (en termes de volume, catégories, granularité, typologie, etc.) ;
- pseudonymiser ou anonymiser les données si nécessaire.

Fiabiliser l'application

- ☐ implémenter une authentification multi-facteurs pour les tâches d'administration sur le système d'IA ;
- ☐ assurer la confidentialité et l'intégrité des entrées et des sorties ;
- ☐ mettre en place des filtres de sécurité pour détecter les instructions malveillantes ;
- ☐ veiller à tenir l'ensemble des données, métadonnées et annotations à jour, exactes (pour éviter la dérive de ces dernières notamment) ;
- ☐ Effectuer une évaluation continue de la précision et de la performance du modèle.

Penser une stratégie organisationnelle

- ☐ documenter les choix de conception ;
- ☐ superviser le fonctionnement du système d'IA ;
- ☐ identifier les personnes clés et encadrer le recours à des sous-traitants ;
- ☐ mettre en œuvre une stratégie de gestion de risque ;
- ☐ prévoir un mode dégradé des services métier sans système d'IA ;
- ☐ mettre en place des politiques d'utilisation encadrées de l'IA générative (selon la sensibilité de l'organisation) ;
- ☐ mettre en place un processus de veille des vulnérabilités spécifiques aux systèmes d'IA ;
- ☐ se tenir informé des évolutions techniques qui permettraient de limiter l'usage de données personnelles par exemple ;
- ☐ mettre en œuvre un système de gestion des données ;
- ☐ mettre en œuvre des méthodes sécurisées de suppression des données ;
- ☐ documenter les jeux de données produits afin de :
  - faciliter l'utilisation de la base de données ;
  - faciliter le suivi des données dans le temps jusqu'à leur suppression ou leur anonymisation ;
  - réduire les risques d'une utilisation imprévue des données ;

Mesures préventives

- ☐ former régulièrement le personnel sur les risques de sécurité liés à l'IA ;
- ☐ effectuer des audits de sécurité réguliers sur le système d'IA ;
- ☐ anticiper au maximum les problématiques potentiellement associées à l'exercice des droits (sur la propriété intellectuelle et la protection des données par exemples) sur les données d'entraînement ou sur le modèle lui-même.

→ ANNEXE 2

REFERENCES

Développements des systèmes d'IA

• AIVD. AI systems: develop them securely. 2023. Available at: AI-systems: develop them securely | Publication | AIVD

• G7. Hiroshima Process International, Code of Conduct for Organizations Developing Advanced AI Systems. 2023. Available at: 100573473.pdf

• G7. Hiroshima Process International, Guiding Principles for Organizations Developing Advanced AI Systems. 2023. Available at: 100573471.pdf

• NCSC-UK, CISA. Joint Guidelines for Secure AI System Development. 2023. Available at: Guidelines for secure AI system development - NCSC.GOV.UK

Cas d'utilisation

• ANSSI. Security recommendations for a generative AI system. 2024. Available at: https://cyber.gouv.fr/en/publications/security-recommendations-generative-ai-system

BSI, ANSSI. AI Coding Assistants. 2024. Available at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/ANSSI_BSI_AI_Coding_Assistants.pdf?__blob=publicationFile&v=7

• BSI. Generative AI Models: Opportunities and Risks for Industry and Authorities. 2025. Available at: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Generative_AI_Models.pdf?__blob=publicationFile&v=6

Vulnérabilités et sécurité de l'IA

• CSA Singapore. Guidelines and Companion Guide on Securing AI Systems. 2024. Available at: Guidelines and Companion Guide on Securing AI Systems | Cyber Security Agency of Singapore

• CSA Singapore, Resaro. Securing AI: A Collective Responsibility. 2024. Available at: Discussion Paper on Securing Artificial Intelligence (AI): A Collective Responsibility | Cyber Security Agency of Singapore

• Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J. and Sanchez Martin, J.I.. Cybersecurity of Artificial Intelligence in the AI Act. 2023. Available at: JRC Publications Repository - Cybersecurity of Artificial Intelligence in the AI Act

• Kamm L. (Cybernetica AS), Pillmann H. (RIA). Risks and controls for artificial intelligence and machine learning systems. 2024. Available at: Risks-and-controls-for-artificial-intelligence-and-machine-learning-systems.pdf

• MITRE ATLAS: MITRE ATLAS™

• OWASP AI Exchange: https://owaspai.org/ - includes several publications

• Vassilev, A., Oprea, A., Fordyce, A. and Andersen, H. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. 2024. Available at: Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations | NIST

Gestion des risques

• NIST AI Risk Management Framework: AI Risk Management Framework | NIST

• OECD. "Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI". 2023. Available at: Advancing accountability in AI | OECD

Terminologie

• ISO/IEC 22989:2022 - «Information Technology - Artificial Intelligence - Artificial Intelligence Concepts and Terminology». 2022. Available at: https://www.iso.org/standard/74296.html

• OECD. "Defining AI incidents and related terms". 2024. Available at: Defining AI incidents and related terms | OECD

Exemples de réglementation

• EU Artificial Intelligence Act
Regulation - EU - 2024/1689 - EN - EUR-Lex

• EU Cyber Resilience Act
Regulation - 2024/2847 - EN - EUR-Lex

Version 1.0 – Février 2025 – ISSN en cours

AGENCE NATIONALE DE LA SÉCURITÉ DES SYSTÈMES D'INFORMATION
ANSSI — 51, boulevard de la Tour-Maubourg — 75 700 PARIS 07 SP
www.cyber.gouv.fr