< GUIDE D'IMPLÉMENTATION DES RECOMMANDATIONS DE SÉCURITÉ POUR UN SYSTÈME D'IA GÉNÉRATIVE >

GROUPE DE TRAVAIL SÉCURITÉ DE L'IA

CAMPUS CYBER 2026

< PREAMBULE >

L'Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI) a publié le 29 avril 2024 un guide concernant la sécurisation des systèmes d'IA génératifs : Recommandations de sécurité pour un système d'IA Générative. L'essor de l'intelligence artificielle nécessite d'engager des réflexions sur la manière de sécuriser ces systèmes, ce qui a motivé la rédaction de ce guide. Il s'adresse, au sein d'entités publiques et privées, aux administrateurs et utilisateurs de systèmes d'IA, RSSI, DSI, Data Scientists et Data Engineers.
Il est important de noter que ce guide ne traite pas des sujets d'agentique.

Le but de ce document est de traduire les principes de sécurité recommandés par l'ANSSI en actions concrètes, mesures et méthodes pour implémenter les recommandations du guide. Ce document aborde l'ensemble des recommandations du guide de l'ANSSI. Ces mesures doivent être appliquées en tenant compte des spécificités techniques et organisationnelles propres à chaque organisation.

Ce guide a été élaboré en s'appuyant sur :

- Le guide Recommandations de sécurité pour un système d'IA générative de l'ANSSI (ANSSI, 2024) ;
- Le retour d'expérience et les recherches du Groupe de Travail sur le Sécurité de l'IA du Campus Cyber ;
- La veille technologique sur les évolutions des technologies d'IA, les dernières recommandations de sécurité émises par les organismes internationaux (NIST, OWASP…) et par les autorités compétentes (ANSSI, ENISA…), tel que le guide *Recommandations de sécurité concernant les assistants de programmation basés sur l'IA* de l'ANSSI et la BSI (ANSSI - BSI, 2024).

< SOMMAIRE >

CYCLE DE VIE DE L'IA GÉNÉRATIVE 02

1. RECOMMANDATIONS GÉNÉRALES 04

2. RECOMMANDATIONS POUR LA PHASE D'ENTRAÎNEMENT 11

3. RECOMMANDATIONS POUR LA PHASE DE DÉPLOIEMENT..... 12

4. RECOMMANDATIONS POUR LA PHASE DE PRODUCTION 13

5. CAS PARTICULIER DE LA GÉNÉRATION DE CODE SOURCE ASSISTÉE PAR L'IA 15

6. CAS PARTICULIER DE SERVICES D'IA GRAND PUBLIC EXPOSÉS SUR INTERNET 15

7. CAS PARTICULIER DE L'UTILISATION DE SOLUTIONS D'IA GÉNÉRATIVES TIERCES 16

ANNEXES 17

GLOSSAIRE 19

BIBLIOGRAPHIE 21

REMERCIEMENTS 23

< CYCLE DE VIE DE L'IA GÉNÉRATIVE >

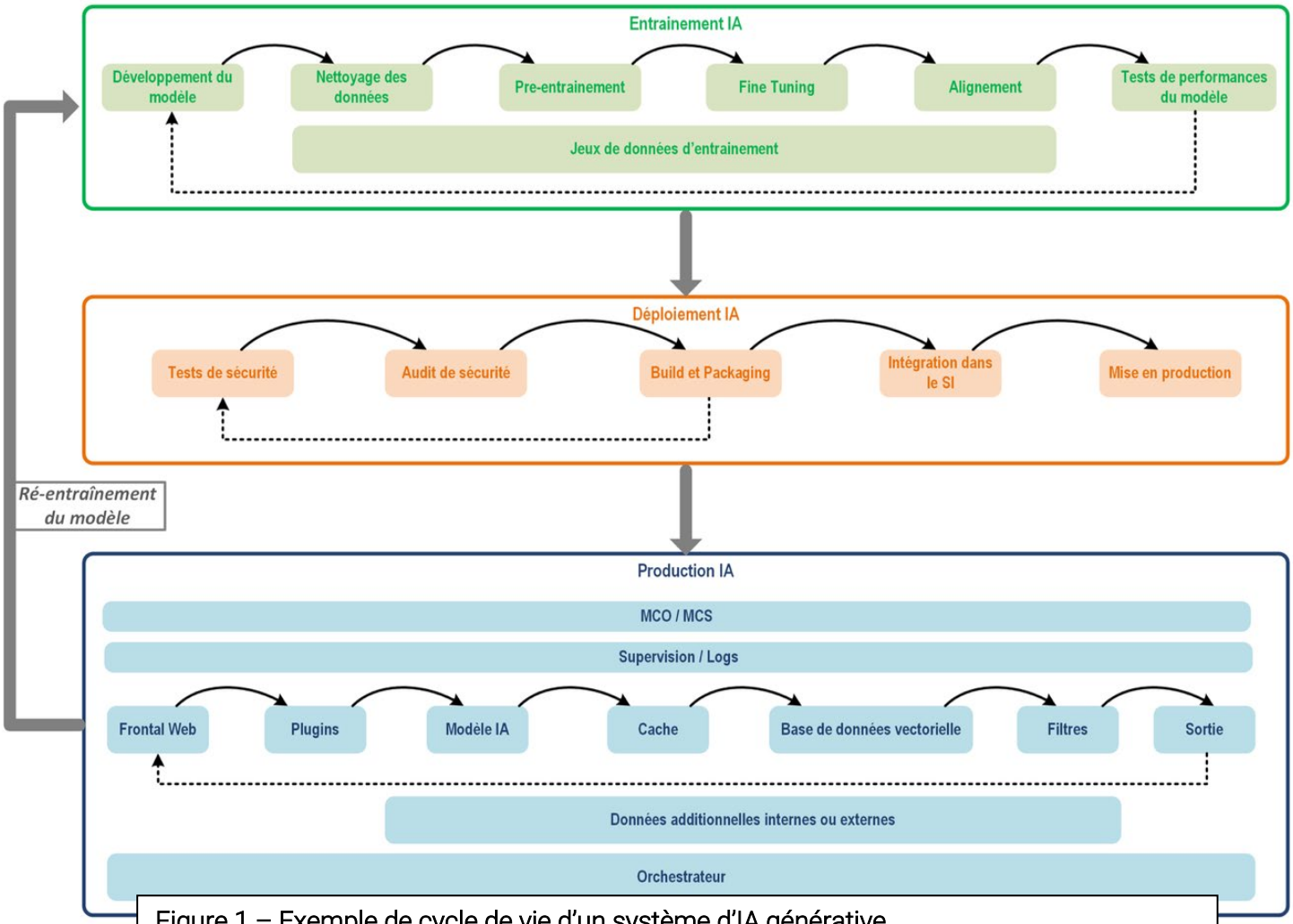


Figure 1 – Exemple de cycle de vie d'un système d'IA générative
Extrait du guide ANSSI « Recommandations de sécurité pour un système d'IA générative », 2024

1. RECOMMANDATIONS GÉNÉRALES

R1 – INTEGRER LA SÉCURITÉ DANS TOUTES LES PHASES DU CYCLE DE VIE D'UN SYSTÈME D'IA

Mesures d'implémentation
Intégrer la sécurité dans l'ensemble des phases du cycle de vie du système d'IA, en commençant par la mise en place de Security By Design ou de Security By Default (Charter of Trust, s.d.) pour les initiatives d'IA en fonction de la sensibilité de l'initiative (ANSSI, 2017).
Mettre également en place un processus spécifique aux systèmes d'IA dans la procédure d'intégration de la sécurité dans les projets (ISP) sur l'ensemble de cycle de vie du système IA. S'assurer que les POC (*Proof Of Concept*) sont correctement évalués avant leur mise en production.

R2 – MENER UNE ANALYSE DE RISQUE SUR LES SYSTÈMES D'IA AVANT LA PHASE D'ENTRAÎNEMENT

Mesures d'implémentation
Faire une analyse de risque en appliquant la méthode EBIOS-RM (ou équivalente en fonction de la sensibilité / complexité de l'initiative), en amont de la phase d'entraînement. Adapter la méthodologie en intégrant les spécificités d'un système d'IA : ajouter un questionnaire dédié à l'IA, mettre en place un framework de gestion des risques spécifiques (NIST, 2023) (en incluant les impacts juridiques, légaux, d'image, opérationnels…). La recommandation de l'ANSSI est exhaustive dans sa mise en application et propose les mesures suivantes : • Cartographie des éléments en lien avec le système • Considération du scénario de partage de responsabilités • Considération de la protection des données Identification des impacts d'une réponse erronée (entraînant par exemple une action automatique dans un système agentique (cf. R9))

R3 – ÉVALUER LE NIVEAU DE CONFIANCE DES BIBLIOTHÈQUES ET MODULES EXTERNES UTILISÉS DANS LE SYSTÈME D'IA

Mesures d'implémentation
Définir une méthodologie claire pour établir le niveau d'acceptabilité : ↗ Cartographie • Définir un inventaire détaillé des bibliothèques et modules externes, avec leur version et leurs fonctionnalités (par exemple intégrer un AIBOM (The Linux Foundation, 2024) (OWASP, 2025) spécifique pour les composants d'IA) • Maintenir une documentation à jour avec leur utilisation, et les raisons de leur choix ↗ Évaluation de la réputation • Analyse des avis et retours de la communauté • Veille active sur les vulnérabilités connues (utilisation de notifications automatiques) ↗ Vérification des mises à jour • Mettre en place un système de suivi des versions, en s'assurant que les dernières versions contenant des correctifs soient utilisées • Utiliser des outils d'automatisation pour notifier les équipes de nouvelles versions ↗ Analyse de la documentation • S'assurer que toutes les fonctionnalités sont bien comprises

- Mettre en place toutes les recommandations de sécurité lors de l'intégration

↗ Tests de sécurité

- Utiliser des outils d'analyse statique du code

↗ Conformité aux normes

- S'assurer que les bibliothèques respectent des normes de sécurité reconnues
- Vérifier si les bibliothèques ont reçu des certifications de sécurité pertinentes

R4 – ÉVALUER LE NIVEAU DE CONFIANCE DES SOURCES DE DONNÉES EXTERNES UTILISÉES DANS LE SYSTÈME D'IA

Mesures d'implémentation
S'appuyer sur des **textes reconnus** pour avoir des pistes de bonnes pratiques sur le sujet. Plusieurs dimensions sont à prendre en compte : • Origine • Validité du format • Qualité • Pertinence • Validité statistique • Considérations éthiques (CNIL, 2024) • Présence d'instruction visant a détourner le comportement nominal du modèle Il est également important de noter que l'identification de l'origine des données peut s'avérer particulièrement complexe sur des données externes à l'organisation et notamment pour identifier des données générées par IA.

R5 – APPLIQUER LES PRINCIPES DE DEVSECOPS SUR L'ENSEMBLE DES PHASES DU PROJET

Mesures d'implémentation
Appliquer les principes du DevSecOps (ANSSI, 2024) (NIST, 2022) (NCSC-UK, 2018) (CISA, 2023) sur les systèmes d'IA (MLSecOps) et du MLOps (Dominik Kreuzberger, 2022) (NIST, 2024) (Open SSF, s.d.). La recommandation de l'ANSSI est exhaustive dans sa mise en application et propose les mesures suivantes : • Sécuriser les chaînes de CI/CD en appliquant le principe de moindre privilège pour l'accès aux outils ; • Sécuriser la gestion des secrets ; • Mettre en place des tests de sécurité statiques et dynamiques sur le code source ; • Protéger l'intégrité du code et restreindre son accès ; • Utiliser des langages de développement sécurisés.
Former les Data Scientists à l'utilisation des outils de DevSecOps.

R6 – UTILISER DES FORMATS DE MODÈLES D'IA SÉCURISÉS

Mesures d'implémentation
Utiliser des **formats sécurisé**s et à **l'état de l'art** pour stocker les modèles. La mise en place d'alertes de vulnérabilités sur les formats utilisés renforce la résilience. Une liste blanche des formats autorisés peut également être formalisée et maintenue. Il est impératif de scanner les modèles pour détecter les formats non-autorisés. Former les Data Scientists à l'utilisation de modèles sécurisés pour s'assurer de leur bonne utilisation.
Mettre en place du **fingerprinting** sur les modèles.

R7 – PRENDRE EN COMPTE LES ENJEUX DE CONFIDENTIALITÉ DES DONNÉES DÈS LA CONCEPTION DU SYSTÈME D'IA

Mesures d'implémentation
Pour les cas d'usage d'entrainements externalisés, utiliser de l'**émulation d'environnement sécurisé via virtualisation afin de durcir l'environnement tiers en accord avec les exigences de l'organisation** pour améliorer la maîtrise des données malgré un traitement externe. Dupliquer les modèles en utilisant des SLMs pour séparer la solution en un modèle entraîné sur des données sensibles et un sur les données publiques et en adaptant le modèle utilisé au contexte de l'utilisateur. **Point d'attention :** Il est important de noter que l'utilisation de SLM n'est pas adaptée à tous les cas d'usage du fait d'une potentielle baisse de performance du modèle par rapport à un LLM.
Privilégier l'utilisation de jeux de **données internes** de qualité. Redéfinir son processus de **classification de données** pour prendre en compte les jeux de données, les modèles/systèmes d'IA, et leurs relations.
Mettre en place de la confidentialité différentielle pour les données confidentielles

R8 – PRENDRE EN COMPTE LA PROBLÉMATIQUE DE BESOIN D'EN CONNAÎTRE DÈS LA CONCEPTION DU SYSTÈME D'IA

Mesures d'implémentation
Définir précisément les **stratégies de réentraînement**, et ce dès la conception du système d'IA, afin d'éviter le réentraînement avec des données plus sensibles. Ces problématiques doivent être considérées dans les processus **d'intégration de la sécurité dans les projets**. Redéfinir son processus de **classification de données** pour prendre en compte les jeux de données, les modèles/systèmes d'IA, et leurs relations.
Mettre en place une politique de **gestion des droits** basée sur la confidentialité du modèle et des données qu'il utilise.
Dupliquer les modèles en utilisant des SLMs pour séparer la solution en un modèle entraîné sur des données sensibles et un autre sur les données publiques, en adaptant le modèle utilisé au contexte de l'utilisateur.
Selon le contexte, prendre en compte les niveaux de classification de la Défense Nationale et appliquer les mesures de sécurité nécessaires.

R9 – PROSCRIRE L'USAGE AUTOMATISÉ DE SYSTÈMES D'IA POUR DES ACTIONS CRITIQUES SUR LE SI

Mesures d'implémentation
Effectuer un **inventaire** des actions critiques sur le SI. Identifier celles déjà réalisées par des systèmes d'IA (par exemple des actions qui portent un niveau de risque « High » au sens de l'IA Act ou de l'organisation). Mettre en place des **campagnes de sensibilisation** régulières pour les développeurs. Il peut être pertinent de collecter leurs retours afin de créer des cas d'usage réalistes. Intégrer une **intervention humaine** dans les processus critiques (approche « human-in-the-loop »).
Mesure dégradée : mettre en place un comité de haut niveau pour la validation

R10 – MAÎTRISER ET SÉCURISER LES ACCÈS À PRIVILÈGES DES DÉVELOPPEURS ET DES ADMINISTRATEURS SUR LE SYSTÈME D'IA

Mesures d'implémentation
Plusieurs étapes à suivre pour l'administration sécurisée (ANSSI, 2021): • Identifier les composants à sécuriser; • Identifier les **acteurs principaux** qui interagissent sur les composants ; • Identifier les **activités** sur les composants, et les opérations à privilèges ; • Identifier les **rôles** selon les acteurs identifiés, formaliser les droits et habilitations selon les activités et opérations ; • Accorder aux développeurs uniquement les **permissions nécessaires** pour accomplir leurs tâches ; • Tester le bris de glace avant la mise à disposition des droits et habilitations sur la solution ; • Remonter les logs vers le SOC pour analyser les comportements suspects ; • Revoir périodiquement les droits et habilitations.

R11 – HÉBERGER LE SYSTÈME D'IA DANS DES ENVIRONNEMENTS DE CONFIANCE COHÉRENTS AVEC LES BESOINS DE SÉCURITÉ

Mesures d'implémentation
Effectuer des **analyses de risques** liés aux environnements d'hébergement, pour chacune des phases du cycle de vie, puis faire une étude comparative entre les besoins de sécurité et les **coûts**. Selon le niveau de confidentialité, **chiffrer les données** en transit et au repos avec des règles d'accès strictes, et des mécanismes à l'état de l'art. Adapter l'hébergement des composants du SIA (présentation, application, modèle, données) en fonction des risques. Dans les cas d'hébergement via des services nuagiques, il est important de considérer les enjeux liés à la protection des données en lien avec les lois extraterritoriales et d'utiliser les services de sécurisation proposés par les fournisseurs de services nuagiques.

R12 – CLOISONNER CHAQUE PHASE DU SYSTÈME D'IA DANS UN ENVIRONNEMENT DÉDIÉ

Mesures d'implémentation

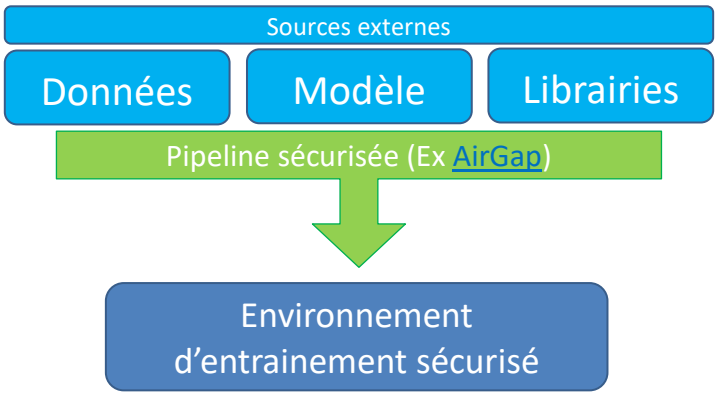
Cloisonnement à appliquer sur différents contextes :

Cloisonnement réseau

Utilisation de solutions de **segmentation réseau** lors du **réentrainement**

Ex : Solutions du type AirGap :

- Jumpbox « jetable »
- Killswitch
- Microsegmentation
- Discovery des assets



Cloisonnement stockage

L'ensemble des contributeurs rencontrent des difficultés pour mettre en place la **ségrégation physique du stockage** dans leur organisation.

- Mise en place d'une **ségrégation logique** dans le cas d'un coût trop élevé pour mettre en place la **ségrégation physique**
- Confidentialité différentielle

Cloisonnement système

Les **coûts des GPU** (utilisées dans tous les environnement) rendent le cloisonnement système **impossible** au sein d'un groupe

L'ensemble des contributeurs rencontrent des difficultés pour **accéder aux GPU** dans leur organisation.

Confidentialité différentielle : utilisation de solution permettant de garantir la **confidentialité** des données en ajoutant du **bruit contrôlé** dans le dataset

- ✓ Mesure dérogatoire pour assurer un minimum de confidentialité des données
- **Librairie Pysyft & Solution Sarus**
- ✓ Assignation de temps de calculs GPU sur le Cloud

Cloisonnement des comptes et des secrets

Quels sont les **différents rôles** des utilisateurs / administrateurs (Data scientists, security officier, DPO, etc…) dans vos organisations ?

- ➔ **Environnements d'expérimentation : pas de restriction des rôles mais interdiction des données métiers**
- ➔ **Environnements de production, de préproduction et de développement : Rôles classiques et restriction de l'assignation des rôles administrateur**

Mise en place d'une **gouvernance des accès utilisateurs et administrateurs** par environnement en fonction des rôles.
Utilisation de **secrets spécifiques** à chaque environnement

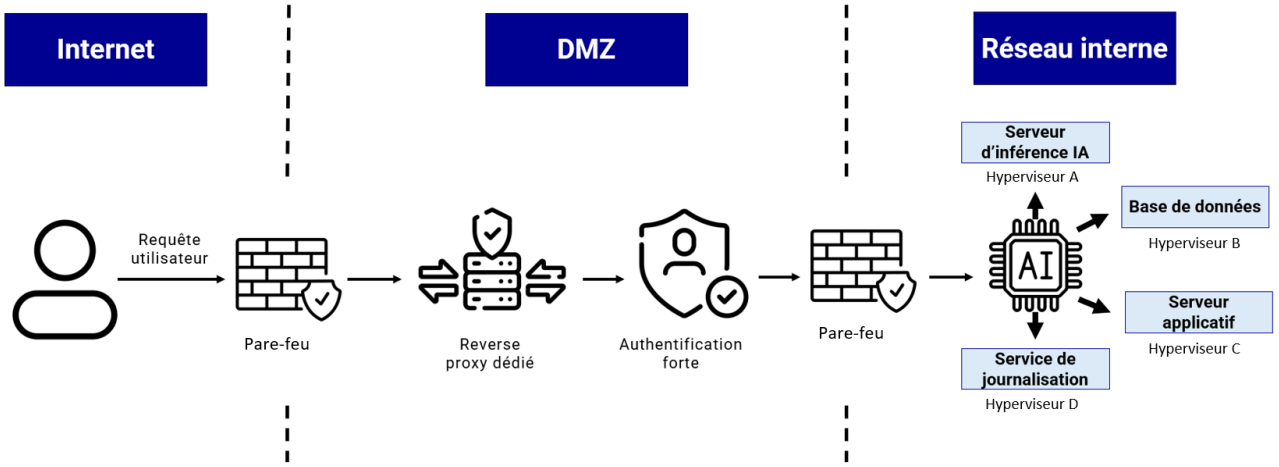
R13 – IMPLÉMENTER UNE PASSERELLE INTERNET SÉCURISÉE DANS LE CAS D'UN SYSTÈME D'IA EXPOSÉ SUR INTERNET

Mesures d'implémentation

La recommandation est exhaustive dans sa mise en application, ainsi que le guide de l'ANSSI dédié (ANSSI, 2020) :

- Dédier un reverse-proxy aux systèmes d'IA ;
- Appliquer un découpage réseau via une DMZ, et mettre en place un pare-feu ;
- Ne pas utiliser un annuaire interne, mais utiliser des mécanismes d'authentification externes ;
- Ne pas mutualiser sur un même hyperviseur ces différentes fonctions.

Exemple d'architecture :



R14 – PRIVILÉGIER UN HÉBERGEMENT SECNUMCLOUD DANS LE CAS D'UN DÉPLOIEMENT D'UN SYSTÈME D'IA DANS UN CLOUD PUBLIC

Mesures d'implémentation
Choisir, si possible, une offre de confiance qualifiée **SecNumCloud** dans le cas d'un hébergement sur un Cloud public.
Mesure dégradée : Sinon, adopter une approche multi-Cloud hybride en fonction des niveaux de sensibilité des cas d'usage. Cependant, il est nécessaire **d'évaluer la maturité cyber** des fournisseurs et la sécurité des données (données business, logs, métadonnées, données d'inférence, …).

R15 – PRÉVOIR UN MODE DÉGRADÉ DES SERVICES MÉTIER SANS SYSTÈME D'IA

Mesures d'implémentation
En fonction de la criticité du système d'IA, intégrer la mise en place d'un mode dégradé **dès la conception**. Mettre en place des **tests périodiques** de simulation d'indisponibilité. Tester les systèmes lors d'exercices de gestion de crise. Intégrer le modèle d'IA dans le périmètre des **BIA**. Prévoir un moyen de **désactiver** le modèle en cas d'incident. Faire valider formellement au Métier le mécanisme et les conditions de mise en place des mesures dégradées.

R16 - DÉDIER LES COMPOSANTS GPU AU SYSTÈME D'IA

Mesures d'implémentation
En fonction du niveau de confidentialité, utiliser des **GPU dédiés** par cas d'usage de système d'IA.
Mesure dégradée : Dédier, au minimum **virtuellement**, les ressources physiques par cas d'usage de système d'IA et/ou mutualiser les ressources par modèles à niveau de confidentialité identique.

R17 - PRENDRE EN COMPTE LES ATTAQUES PAR CANAUX AUXILIAIRES SUR LE SYSTÈME D'IA

Mesures d'implémentation
Mettre en place des mesures de **détection** et de **mesures** pour répondre efficacement aux attaques par canaux auxiliaires, pour compenser leur complexité de prévention.
Introduire des **opérations factices** afin de perturber l'analyse du modèle par un attaquant.
Renseigner une **longueur aléatoire** dans les réponses afin de masquer la longueur réelle des jetons de réponse. **Grouper** les jetons de réponse plutôt que de les envoyer individuellement.
Point d'attention : Il est important de noter que les mesures proposées pour cette R17 restent complexes et coûteuses à mettre en œuvre.

2. RECOMMANDATIONS POUR LA PHASE D'ENTRAÎNEMENT

R18 - ENTRAÎNER UN MODÈLE D'IA UNIQUEMENT AVEC DES DONNÉES LÉGITIMEMENT ACCESSIBLES PAR LES UTILISATEURS

Mesures d'implémentation
Etudier **l'éligibilité** des jeux de données pour l'entrainement par rapport au besoin d'en connaître des utilisateurs finaux du système IA. Revoir la méthode de **classification des données** sur l'ensemble des données utilisées en entrainement. Mener une **analyse de risque** pour s'assurer de la protection des données. Appliquer au maximum **l'anonymisation** ou la **pseudonymisation** des données. Vérifier la **licence** en cas d'utilisation de jeux de données externes à l'organisation.

R19 - PROTÉGER EN INTÉGRITÉ LES DONNÉES D'ENTRAÎNEMENT DU MODÈLE D'IA

Mesures d'implémentation
Mettre en place un système de **hash** afin de vérifier l'intégrité des données en amont de sa prise en charge. Adopter une **gestion des droits** stricte et clairement définie.
Vérifier les **signatures** et **empreintes**.
Mettre en place des **contrôles d'intégrité** des fichiers et du **versionnage des modèles et des jeux de données** via les plateformes d'analyse de données.

R20 – PROTÉGER EN INTÉGRITÉ LES FICHIERS DU SYSTÈME D'IA

Mesures d'implémentation
Appliquer les pratiques **DevSecOps** (ANSSI, 2024) (sécurisation des fichiers, gestion des droits clairement définie). Utiliser des **plateformes d'analyse de données** (DSPM) (Varonis, 2024) pour effectuer un travail de supervision et de détection. Mettre en place des solutions de blocage de modification de fichiers.
Mettre en place des **contrôles d'intégrité** des fichiers et du **versionnage des modèles et des jeux de données** via les plateformes d'analyse de données.

R21 – PROSCRIRE LE RÉ-ENTRAÎNEMENT DU MODÈLE D'IA EN PRODUCTION

Mesures d'implémentation
Effectuer le réentraînement du modèle **hors de l'environnement de production** en répétant les phases du cycle de vie du système d'IA. Séparer les environnements via de la conteneurisation. *Voir R12 – Cloisonner chaque phase du système d'IA dans un environnement dédié.* Le réentrainement doit systématiquement passer par une phase de validation/supervision humaine avant le passage en production.

3. RECOMMANDATIONS POUR LA PHASE DE DÉPLOIEMENT

R22 - SÉCURISER LA CHAÎNE DE DÉPLOIEMENT EN PRODUCTION DES SYSTÈMES D'IA

Mesures d'implémentation
Intégrer la chaîne de déploiement dans les processus standards de mise en production. Les recommandations de l'ANSSI relatives à **l'administration sécurisée** des SI doivent être appliquées (ANSSI, 2021). Appliquer également les principes du **DevSecOps** (ANSSI, 2024) (gérer les secrets de manière sécurisée, séparer les infrastructures de CI/CD, réinstancier régulièrement l'infrastructure de CI/CD, prévoir des tests de sécurité automatisés).

R23 – PRÉVOIR DES AUDITS DE SÉCURITÉ DES SYSTÈMES D'IA AVANT DÉPLOIEMENT EN PRODUCTION

Mesures d'implémentation
Effectuer des **audits de sécurité** sur tous les aspects du système d'IA, et ce avant sa mise en production. La recommandation de l'ANSSI est force de proposition dans sa mise en application et propose les mesures suivantes : • Tests d'intrusion sur tous les composants du système ; • Tests de sécurité sur le code source ; • Tests spécifiques aux vulnérabilités d'un système d'IA ; • Tests manuels via des auditeurs (auditeurs certifiés PASSI recommandés selon le niveau de criticité).

R24 – PRÉVOIR DES TESTS FONCTIONNELS MÉTIER DES SYSTÈMES D'IA AVANT DÉPLOIEMENT

Mesures d'implémentation
Mettre en place plusieurs catégories de **tests** : • Tests de robustesse • Test satisfaction métier • Tests de qualité de réponses • Tests automatisés *Un point d'attention sur la définition des métriques d'évaluation de qualité.*

4. RECOMMANDATIONS POUR LA PHASE DE PRODUCTION

R25 – PROTÉGER LE SYSTÈME D'IA EN FILTRANT LES ENTRÉES ET LES SORTIES DES UTILISATEURS

Mesures d'implémentation
Mettre en place un système de **filtrage des requêtes/réponses** afin de contrôler : ↗ En entrée : • Les requêtes, unitaires ou en série, à but malveillant ; • Les requêtes jugées non légitimes ; • Les comportements atypiques ; • Les requêtes trop longues et/ou trop fréquentes. ↗ En sortie : • L'extraction d'informations internes ; • L'extraction d'informations sensibles ; • Les réponses à contenu toxique ou à charge malveillante ; • Les réponses trop longues.

R26 – MAÎTRISER ET SÉCURISER LES INTERACTIONS DU SYSTÈME D'IA AVEC D'AUTRES APPLICATIONS MÉTIER

Mesures d'implémentation
Intégrer les interactions du système d'IA avec les applications métier dans l'analyse de risque du système d'IA. Mettre en place des **règles d'interaction** des applications selon le niveau de risque identifié. • La recommandation de l'ANSSI est exhaustive dans sa mise en application et propose les mesures suivantes :Filtrage réseau, chiffrement et authentification (ANSSI, 2020) ; • Utilisation de protocoles d'identification sécurisés (ANSSI, 2020) ; • Contrôle des autorisations d'accès à la ressource ; • Journalisation. Effectuer une **revue des accès et habilitations** du compte de service du système d'IA de manière régulière.

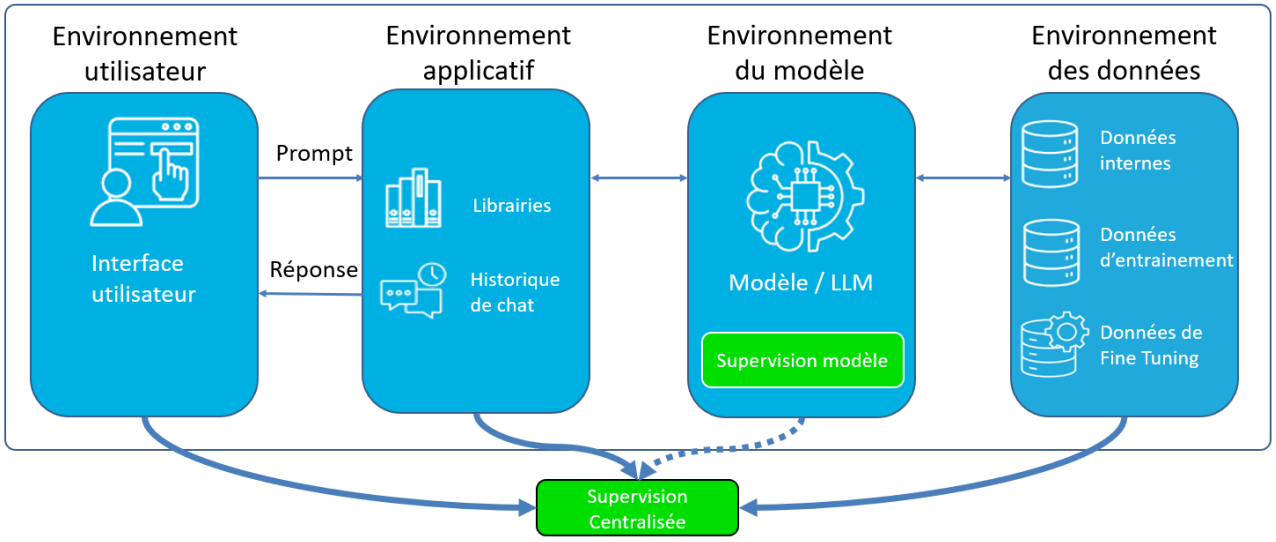
R27 – LIMITER LES ACTIONS AUTOMATIQUES DEPUIS UN SYSTÈME D'IA TRAITANT DES ENTRÉES NON-MAÎTRISEES

Mesures d'implémentation
Limiter toute action automatisée par IA, du fait de la difficulté à en maîtriser les données d'entrées. En fonction de l'impact et des risques, valider **par une interaction humaine** systématiquement chaque action automatique effectuée par un système d'IA utilisant des données non maitrisées. **Limiter les droits** du compte de service de l'IA.

R28 – CLOISONNER LE SYSTÈME D'IA DANS UN OU PLUSIEURS ENVIRONNEMENTS TECHNIQUES DÉDIÉS

Mesures d'implémentation

En fonction du cas d'usage du système d'IA et de sa criticité d'un point de vue métier, **dédier** les serveurs utilisés pour l'hébergement du système d'IA : les isoler **physiquement**, sinon **logiquement** des autres serveurs applicatifs. Mettre en place une architecture **4 tiers** :



Cloisonner les machines virtuelles au niveau des droits (via des Cloud Policies si l'hébergeur est Cloud).
Mesure dégradée : **Conteneuriser** afin de limiter les accès physiques, une solution qui permet également d'assurer la résilience du système. Dans ce cas, il est recommandé d'isoler le modèle via une machine virtuelle dédiée et de l'appeler via une passerelle API.

R29 – JOURNALISER L'ENSEMBLE DES TRAITEMENTS RÉALISÉS AU SEIN DU SYSTÈME D'IA

Mesures d'implémentation

Mettre en place un système de **journalisation** (ANSSI, 2022), dès le début du projet (Intégration By Design)., en accord avec les exigences issues de l'AI Act (AI Act, 2024).
La recommandation de l'ANSSI est exhaustive dans sa mise en application et propose de journaliser les données suivantes :

- Requêtes des utilisateurs (respecter la protection des données personnelles (CNIL, 2022)) ;
- Traitements réalisés sur les requêtes avant envoi au modèle ;
- Appel à des plugins ;
- Appel à des données additionnelles ;
- Traitements réalisés sur les réponses par les filtres ;
- Réponses données aux utilisateurs ;
- L'avis de l'utilisateur.

5. CAS PARTICULIER DE LA GÉNÉRATION DE CODE SOURCE ASSISTÉE PAR L'IA

R30 – CONTRÔLER SYSTÉMATIQUEMENT LE CODE SOURCE GÉNÉRÉ PAR IA

Mesures d'implémentation
Vérifier tout code source généré par IA, de manière **manuelle et/ou automatique** (SAST, DAST, … (OWASP, 2024)) par un passage dans un pipeline de sécurisation des codes des applications. Utiliser si possible des solutions spécifiquement pensées pour le contrôle de code généré par IA incluant les vérifications relatives à la propriété intellectuelle et les fonctionnalités cachées. Dans le cas de revue de code à l'aide d'outil d'IA, s'assurer que le modèle utilisé par l'outil de revue de code est bien **distinct de celui utilisé pour la génération de code**.

R31 – LIMITER LA GÉNÉRATION DE CODE SOURCE PAR IA POUR DES MODULES CRITIQUES D'APPLICATIONS

Mesures d'implémentation
Former les équipes de développeurs sur ces problématiques. Faire valider l'obtention de « licences d'IA » par le RSSI : blocage des licences pour les développeurs travaillant sur des projets critiques. L'obtention de la « licence d'IA » est conditionnée à la validation de la formation.

R32 – SENSIBILISER LES DÉVELOPPEURS SUR LES RISQUES LIÉS AU CODE SOURCE GÉNÉRÉ PAR IA

Mesures d'implémentation
Effectuer des **campagnes de sensibilisation** des développeurs et managers, réalisées et/ou dispensées par des équipes expertes (CERT, prestataire de confiance...). Inclure les équipes de Cybersécurité et de Data Scientist pour faciliter les échanges et la compréhension.

6. CAS PARTICULIER DE SERVICES D'IA GRAND PUBLIC EXPOSÉS SUR INTERNET

R33 – DURCIR LES MESURES DE SÉCURITÉ POUR DES SERVICES D'IA GRAND PUBLIC EXPOSÉS SUR INTERNET

Mesures d'implémentation
La recommandation de l'ANSSI est exhaustive dans sa mise en application et propose les mesures suivantes : • Entraîner et/ou raffiner le modèle uniquement sur des données publiques ; • Exiger une authentification des utilisateurs ; • Analyser les requêtes utilisateur ; • Contrôler et valider les réponses avant envoi, notamment afin d'empêcher la fuite de données confidentielles ; • Protéger la confidentialité des données utilisateurs avec une politique de rétention claire respectant le RGPD (CNIL, 2024) ; • Mettre en place des mesures contre les attaques par force brute et les ddos (ANSSI, 2015) ; • Sécuriser le service web en frontal (ANSSI, 2021).

7. CAS PARTICULIER DE L'UTILISATION DE SOLUTIONS D'IA GÉNÉRATIVE TIERCES

R34 – PROSCRIRE L'UTILISATION D'OUTILS D'IA GÉNÉRATIVE SUR INTERNET POUR UN USAGE PROFESSIONNEL IMPLIQUANT DES DONNÉES SENSIBLES

Mesures d'implémentation
Bloquer l'accès aux outils d'IA en ligne via un système de filtrage interne (ex : filtrage des sites accédés ou pare-feu interne). Proposer des **sensibilisations** aux usages de données sensibles avec l'IA générative et au Shadow AI. Proposer un modèle similaire en interne. Promouvoir son usage via une meilleure gestion des problématiques métier

R35 – EFFECTUER UNE REVUE RÉGULIÈRE DE LA CONFIGURATION DES DROITS DES OUTILS D'IA GÉNÉRATIVE SUR LES APPLICATIONS MÉTIERS

Mesures d'implémentation
Mettre en place des **contrôles IAM de niveau 1 et 2**, avec revue régulière via des campagnes de re-certification. **Journaliser** les accès et la revue des accès. Réduire au maximum la surface d'attaque, et mettre en place une surveillance accrue sur le périmètre résultant. *Considérer les systèmes d'IA générative comme des utilisateurs à privilèges.*
Cas d'usage simple : dialogue entre un utilisateur et une IA (plusieurs options) • Limiter les droits et habilitations du compte de service de l'IA. • Transférer les droits de l'utilisateur au système d'IA. • Approche hybride des deux options ci-dessus.
Cas d'usage complexe : dialogue entre deux IA Si le dialogue intervient après un prompt utilisateur, transmettre le token d'accès de l'utilisateur et l'utiliser pour déterminer les habilitations des IA.

< ANNEXES >

< MAPPING CONTROLE DU NIVEAU DE CONFIANCE / STANDARDS ET NORMES >

Sujet	Contrôle	Standards/Normes associés
Source des Données	Tracer l'historique des données, y compris l'origine et les transformations	ISO/IEC 25012
	Assurer une documentation complète de la collecte et du traitement des données	ISO/IEC 25012
	Évaluer la crédibilité et la réputation de la source des données	ISO/IEC 25012
	Vérifier la transparence de la collecte et du traitement des données	ISO/IEC 25012
Format des Données	Assurer la cohérence du format des données	ISO/IEC 11179, W3C Standards
	Vérifier le respect des normes pertinentes de format	ISO/IEC 11179, W3C Standards
Qualité des Données	Évaluer l'exactitude des données	ISO/IEC 25012, NIST Big Data Interoperability Framework Volume 5
	Évaluer l'exhaustivité des données	ISO/IEC 11179, ISO/IEC 25012
	Évaluer la conformité des données aux normes en vigueur	
	Évaluer la clarté des données	
	Purifier les jeux de données	
	Vérifier les données manquantes ou incomplètes	ISO/IEC 25012, NIST Big Data Interoperability Framework Volume 5
	Assurer que les données sont à jour	ISO/IEC 25012
	Identifier et quantifier le bruit et les erreurs	ISO/IEC 25012
Intégrité des Données	Vérifier que les données n'ont pas été falsifiées ou altérées	ISO/IEC 27001, ISO/IEC 25012
	Vérifier la cohérence entre les ensembles de données	ISO/IEC 27001, ISO/IEC 25012
	Assurer des mécanismes pour gérer les données redondantes	ISO/IEC 27001, ISO/IEC 25012
Biais des Données	Évaluer la représentation des données	IEEE 7003, EU Ethics Guidelines for Trustworthy AI
	Identifier les biais potentiels dans les données	IEEE 7003, EU Ethics Guidelines for Trustworthy AI
	Assurer la diversité des données	IEEE 7003, EU Ethics Guidelines for Trustworthy AI
Considérations Éthiques	Vérifier le consentement pour la collecte de données	ISO/IEC 27018, EU GDPR, IEEE Ethically Aligned Design
	Assurer que les données sont conformes aux lois sur la vie privée	ISO/IEC 27018, EU GDPR, IEEE Ethically Aligned Design
	Vérifier l'anonymisation correcte des données personnelles	ISO/IEC 27018, EU GDPR, IEEE Ethically Aligned Design
Sécurité des Données	Vérifier la conformité aux normes de sécurité	ISO/IEC 27001, NIST Cybersecurity Framework
	Appliquer les mécanismes de sécurité pertinents avant utilisation de la donnée en production (sanboxing etc…)	
Évolutivité et Performance	Évaluer l'évolutivité de la base de données	ISO/IEC 25010, NIST Big Data Interoperability Framework Volume 6

	Évaluer la performance et l'efficacité de la récupération des données	ISO/IEC 25010, NIST Big Data Interoperability Framework Volume 6
Problèmes de Licence et Juridiques	Vérifier les accords de licence et les conditions d'utilisation	EU GDPR
	Assurer la conformité aux réglementations légales	EU GDPR
Interopérabilité	Évaluer la compatibilité avec d'autres ensembles de données et systèmes	ISO/IEC 19941, NIST Big Data Interoperability Framework Volume 4
	Évaluer la disponibilité et l'utilisabilité des API	ISO/IEC 19941, NIST Big Data Interoperability Framework Volume 4
Contexte et Métadonnées	Comprendre le contexte de la collecte de données	ISO/IEC 11179, Dublin Core Metadata Initiative (DCMI)
	Assurer l'inclusion complète des métadonnées	ISO/IEC 11179, Dublin Core Metadata Initiative (DCMI)
Gouvernance des Données	Mettre en place une/des politiques de gouvernance des données	ISO/IEC 38500, COBIT
	Intégrer les bases de données externes dans le cadre de la gouvernance des données en place	ISO/IEC 38500, COBIT

< GLOSSAIRE >

Système d'IA Ensemble des composants techniques d'une application reposant sur un modèle d'IA (son implémentation, ses services frontaux, ses bases de données, etc.).

Modèle d'IA Réseau de neurones et ses paramètres (poids, biais).

Mlops Ensemble de pratiques visant à déployer, maintenir et superviser des systèmes d'IA, de manière fiable, automatisée et efficace.

Devsecops Ensemble de pratiques visant à inclure la sécurité dans le processus de développement et de déploiement d'applications.

Human-in-the-loop Approche intégrant une intervention et une expertise humaine au cours du cycle de vie d'un système d'IA.

Sec By Design Principe d'ingénierie logicielle visant à placer la sécurité dans les applications au cœur leur conception.

Proof Of Concept (POC) Démonstration de faisabilité d'un projet effectuée au début du processus de développement d'un projet.

DSPM La gestion de la posture de sécurité des données (ou DSPM en anglais) offre une visibilité sur l'endroit où se trouvent les données sensibles, les personnes qui ont accès à ces données, la façon dont elles ont été utilisées et le niveau de sécurité du dépôt de données ou de l'application

JumpBox Serveur intermédiaire sécurisé qui sert de point d'entrée unique et contrôlé pour permettre aux utilisateurs d'accéder à différents actifs (machine, réseaux ou applications).

KillSwitch Dispositif de sécurité permettant de neutraliser instantanément un système.

Microsegmentation	Technique permettant de diviser un réseau en segments distincts au niveau de la couche applicative afin d'accroître la sécurité et de réduire l'impact d'une violation.
Asset discovery	Ou repérage des actifs informatiques. Processus d'identification et de documentation de tous les appareils, logiciels et systèmes au sein d'un réseau ou d'une organisation.

< BIBLIOGRAPHIE >

ANSSI - BSI. (2024). Récupéré sur https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/ANSSI_BSI_AI_Coding_Assistants.pdf?__blob=publicationFile&v=7

ANSSI - BSI. (2025). *Principes d'architecture Zero Trust pour les LLM* . Récupéré sur BSI: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/ANSSI-BSI-joint-releases/LLM-based_Systems_Zero_Trust.html

ANSSI. (2015, mars). *Comprendre et anticiper les attaques DDoS*. Récupéré sur https://cyber.gouv.fr/publications/comprendre-et-anticiper-les-attaques-ddos

ANSSI. (2017, septembre). *Guide d'hygiène informatique : renforcer la sécurité de son système d'information en 42 mesures.*, Guide ANSSI-GP-042 v2.0. Récupéré sur https://cyber.gouv.fr/publications/guide-dhygiene-informatique

ANSSI. (2020, mars). *Recommandations de sécurité relatives à TLS*, Guide ANSSI-PA-035 v1.2. Récupéré sur https://cyber.gouv.fr/publications/recommandations-de-securite-relatives-tls

ANSSI. (2020, septembre). *Recommandations pour la sécurisation de la mise en œuvre du protocole OpenID Connect*, Guide ANSSI-PA-080 v1.0. Récupéré sur https://cyber.gouv.fr/publications/recommandations-pour-la-securisation-de-la-mise-en-oeuvre-du-protocole-openid-connect

ANSSI. (2020, juin). *Recommandations relatives à l'interconnexion d'un SI à Internet*, Guide ANSSI-PA-066 v3.0. Récupéré sur https://cyber.gouv.fr/publications/recommandations-relatives-linterconnexion-dun-si-internet

ANSSI. (2021, mai). *Recommandations relatives à l'administration sécurisée des SI*, Guide ANSSI-PA-022 v3.0. Récupéré sur https://cyber.gouv.fr/publications/recommandations-relatives-ladministration-securisee-des-si

ANSSI. (2021, avril). *Sécuriser un site web*, Guide ANSSI-PA-009 v2.0. Récupéré sur https://cyber.gouv.fr/publications/securiser-un-site-web

ANSSI. (2022, janvier). *Recommandations de sécurité pour l'architecture d'un système de journalisation*, Guide ANSSI-PA-012 v1.0. Récupéré sur https://cyber.gouv.fr/publications/recommandations-de-securite-pour-larchitecture-dun-systeme-de-journalisation

ANSSI. (2024, février). *Les Essentiels - DevSecOps*, Version 1.0. Récupéré sur https://cyber.gouv.fr/publications/devsecops

ANSSI. (2024, avril). *Recommandations de sécurité pour un système d'IA générative*, Guide ANSSI-PA-102 v1.0. Récupéré sur https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative

ANSSI. (2025, mars). *Prestataires de services d'informatique en nuage (SecNumCloud)*. Récupéré sur https://cyber.gouv.fr/prestataires-de-services-dinformatique-en-nuage-secnumcloud

Charter of Trust. (s.d.). *Security By Default*. Récupéré sur Chater of Trust: https://www.charteroftrust.com/security-by-default/

CISA. (2023, juin). *Defending Continuous Integration/Continuous*. Récupéré sur https://media.defense.gov/2023/Jun/28/2003249466/-1/-1/0/CSI_DEFENDING_CI_CD_ENVIRONMENTS.PDF

CNIL. (2022, avril). *IA : comment être en conformité avec le RGPD ?* Récupéré sur https://www.cnil.fr/fr/intelligence-artificielle/ia-comment-etre-en-conformite-avec-le-rgpd

CNIL. (2024, avril). *IA : Tenir compte de la protection des données dans la collecte et la gestion des données*. Récupéré sur https://www.cnil.fr/fr/tenir-compte-de-la-protection-des-donnees-dans-la-collecte-et-la-gestion-des-donnees

Dominik Kreuzberger, N. K. (2022, mai). *Machine Learning Operations (MLOps): Overview, Definition, and Architecture*. Récupéré sur https://arxiv.org/abs/2205.02302

Dong Chen, A. D. (2024, décembre). *Protecting Confidentiality, Privacy and Integrity in Collaborative Learning*. Récupéré sur https://arxiv.org/abs/2412.08534

Han Wang, H. S. (2020, juillet). *SCARF: Detecting Side-Channel Attacks at Real-time using Low-level Hardware Features*. Récupéré sur https://ieeexplore.ieee.org/document/9159708

Huili Chen, B. D. (2018, avril). *DeepMarks: A Digital Fingerprinting Framework for Deep Neural Networks*. Récupéré sur https://arxiv.org/abs/1804.03648

Microsoft. (2024, juillet). *Github - Responsible AI Toolbox*. Récupéré sur https://github.com/microsoft/responsible-ai-toolbox

NCSC-UK. (2018, novembre). *Secure development and deployment guidance*. Récupéré sur https://www.ncsc.gov.uk/collection/developers-collection

NIST. (2022, février). *Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities*. Récupéré sur https://csrc.nist.gov/pubs/sp/800/218/final

NIST. (2023, janvier). *AI Risk Management Framework*, Version 1.0. Récupéré sur https://www.nist.gov/itl/ai-risk-management-framework

NIST. (2024, juillet). *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile*. Récupéré sur https://csrc.nist.gov/pubs/sp/800/218/a/final

NIST. (2025, mars). *Source Code Security Analyzers*. Récupéré sur https://www.nist.gov/itl/ssd/software-quality-group/source-code-security-analyzers

Open SSF. (s.d.). *Visualizing Secure MLOps*. Récupéré sur Open SSF: https://openssf.org/resources/visualizing-secure-mlops-mlsecops-a-practical-guide-for-building-robust-ai-ml-pipeline-security/

OWASP. (2024). *Dynamic Application Security Testing*. Récupéré sur OWASP: https://owasp.org/www-project-devsecops-guideline/latest/02b-Dynamic-Application-Security-Testing

OWASP. (2025). Récupéré sur OWASP AIBOM: https://owaspaibom.org/

OWASP. (s.d.). *Source Code Analysis Tools*. Récupéré sur https://owasp.org/www-community/Source_Code_Analysis_Tools

Protect AI. (2024, octobre). *Github - AI exploits*. Récupéré sur https://github.com/protectai/ai-exploits

The Linux Foundation. (2024). *Implementing an AI BOM*. Récupéré sur SPDX: https://spdx.dev/implementing-an-ai-bom/

Trusted AI. (2025, janvier). *Github - Adversarial Robustness Toolbox*. Récupéré sur https://github.com/Trusted-AI/adversarial-robustness-toolbox

Varonis. (2024). *What is DSPM*. Récupéré sur Varonis: https://www.varonis.com/fr/blog/what-is-dspm

< REMERCIEMENTS >

Coordination du Groupe de Travail :

- Maxime de Jabrun, Headmind Partners
- Sophie Ravitchandirane et Stephan Cohen, BNP Paribas

Merci aux contributeurs du Groupe de travail :

- Luca Bordonaro, Capgemini
- Yannick Busca, BPCE
- Simon Cavey, EDF
- Olivier Denti, Capgemini
- Léo Deville, BPCE
- Michel Dubois, La Poste
- Mathieu Ferrandez, I-tracing
- Vincent Krafft, Headmind Partners
- Vasco Gomes, Eviden
- Guillaume Gouhier, AMIAD
- Hervé Léon, Euroclear
- Maxime Payraudeau, Headmind Partners
- Eric Savignac, Airbus D&S
- Françoise Soulié, Hub France IA
- Arthur Tondereau, Advens
- Amal Zili, Renault

CAMPUS CYBER

POUR EN SAVOIR PLUS : **WIKI.CAMPUSCYBER.FR**
ADRESSE MAIL DE CONTACT : **COMMUNAUTES@CAMPUSCYBER.FR**
5 - 7 RUE BELLINI 92800, PUTEAUX

CAMPUS CYBER 2025 © - GUIDE IMPLEMENTATION IA GENERATIVE

Ce projet a été financé par le gouvernement dans le cadre du Programme d'investissements d'avenir.