CAMPUS CYBER

HUB FRANCE IA

Analyse des attaques sur les systèmes de l'IA

Mai 2025

Analyse des attaques sur les systèmes de l'IA

Table des matières

1 Introduction 4

Contexte 4

Références utilisées 7

2 Comprendre les attaques contre les systèmes d'IA 8

L'importance du cycle de vie 8

2.1.1 Les étapes du cycle de vie 8

2.1.2 Les principaux formalismes de cycle de vie 9

2.1.3 Choix du cycle de vie : Une analyse comparative 11

Protéger le système d'IA 13

Présentation des principaux référentiels d'attaques 14

2.1.4 NIST.AI.100-2e2023 14

2.1.5 MITRE ATLAS 16

2.1.6 OWASP TOP 10 LLM & TOP 10 ML 19

2.1.7 Recommandations de l'ANSSI 21

Évaluations qualitatives des attaques 23

2.1.8 Critères d'évaluation 24

2.1.9 Indicateur d'Impact (Disponibilité, Intégrité, Confidentialité, Fiabilité) 26

2.1.10 Indicateur de Facilité technique (Temps passé, Ressources, Expertise, Connaissances, Accès) 32

2.1.11 Les conséquences d'une attaque sur l'organisation 37

Taxonomie des attaques 39

Grandes catégories d'attaques 42

2.1.12 Attaques par Empoisonnement (Poisoning) 42

2.1.13 Attaques par Évasion (Evasion) 42

2.1.14 Attaques Oracles (Oracle Attacks) 43

2.1.15 En conclusion 43

3 Autres techniques à suivre 43

RAG 43

Système agentique 46

Apprentissage fédéré 49

Sécurité des systèmes d'IA par la cryptographie 50

3.1.1 Les techniques cryptographiques 53

3.1.2 Risques adressés par la cryptographie 55

Attaques adverses 56

4 Se protéger 59

Prévention 59

4.1.1 Les types de mesures de prévention 60

4.1.2 Les mesures de prévention par phase du cycle de vie 62

Remédiation 65

4.1.3 Architecture de Gestion d'Incident pour les Systèmes d'IA 65

4.1.4 Checklist de remédiation alignée avec le cycle de vie d'un SIA 68

5 Fiches pratiques : Analyse des principales attaques 68

Le format des fiches 68

5.1.1 Au recto de la fiche 69

5.1.2 Au verso de la fiche 73

5.1.3 Démonstration par l'exemple du chatbot Tay 78

Les fiches d'attaque par phase 81

5.1.4 Planification et design 82

5.1.5 Collecte et traitement des données 83

5.1.6 Construction du modèle / adaptation d'un modèle existant 88

5.1.7 Test, évaluation, vérification 91

5.1.8 Mise à disposition, utilisation, déploiement 91

5.1.9 Exploitation et maintenance 94

5.1.10 Décommissionnement / mise au rebut 108

6 Conclusion 109

7 Références 110

8 Glossaire IA & Cyber 112

Glossaire IA 112

Cybersécurité 116
Autres 121
9 Annexe 1 - Méthodes de prévention 123
I Protection cybersécurité 123
II Protection IA « secure by design » 129
III Protection spécifique attaques IA 138
10 Annexe 2 - Remédiation 143
11 Remerciements 146
Coordinateurs 146
Contributeurs 146
Relecteurs 146
La touche finale 146

1 Introduction

Contexte

L'intelligence artificielle (IA), qu'elle soit prédictive1 ou générative1, est en train de transformer de nombreux secteurs d'activité. Si ces technologies offrent des opportunités inédites, elles exposent également les organisations à de nouveaux risques en matière de cybersécurité.

Comme les systèmes traditionnels, les systèmes d'IA[2] (SIA) doivent être protégés au regard de la multiplicité des attaques possibles. Mais ces SIA présentent aussi des vulnérabilités spécifiques, propres à leur architecture et à leur fonctionnement, qui reposent sur des algorithmes complexes et des jeux de données volumineux. Il est donc essentiel de mettre en place des mesures de sécurité adaptées à ces spécificités.

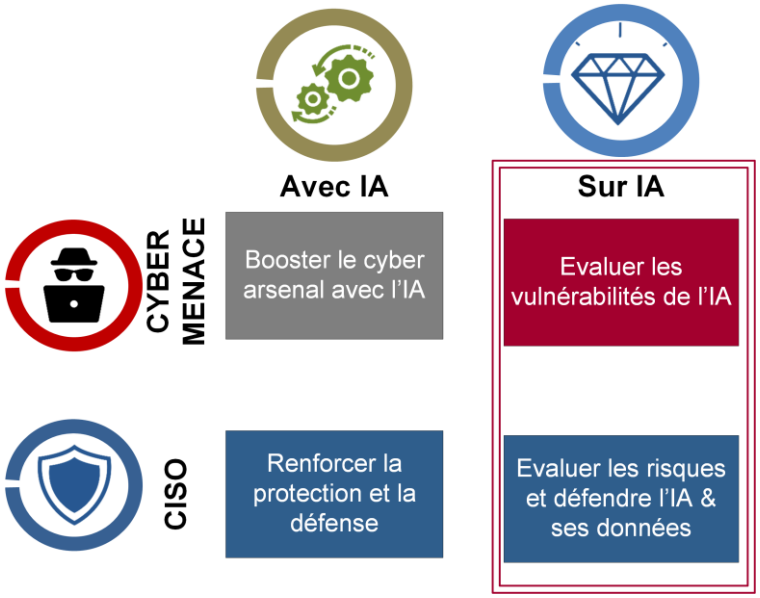


Figure 1 – L'IA et la cybersécurité3

Notons tout d'abord que l'IA intervient de plusieurs façons dans la cybersécurité ou cybercriminalité, comme on le voit dans la Figure 1:

1 Terme expliqué dans la section 8 Glossaire

[2] Dans toute la suite, nous appellerons Système d'IA (SIA), « un système automatisé qui est conçu pour fonctionner à différents niveaux d'autonomie et peut faire preuve d'une capacité d'adaptation après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des entrées qu'il reçoit, la manière de générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels ». Cette définition est reprise de l'AI Act [6].

[3] D'après https://wiki.campuscyber.fr/IA_et_cybers%C3%A9curit%C3%A9

- Avec l'IA
    - Côté attaquant (criminalité) : les attaquants peuvent générer de nouvelles techniques d'attaque entraînant une infraction. Par exemple l'empoisonnement des données, ou les impostures de type " deepfake " (l'exemple le plus fameux est celui des deepfakes utilisés pour les fraudes au président).
    - Côté défenseur (sécurité) : les défenseurs peuvent renforcer leurs techniques de protection, par exemple par des techniques IA de détection d'anomalies ou d'impostures.
- Sur l'IA
    - Côté attaquant (criminalité) : les attaquants peuvent élaborer de nouvelles formes d'attaques, comme par exemple l'empoisonnement des données qui dégrade les performances et donc la qualité des réponses du SIA.
    - Côté défenseur (sécurité) : les défenseurs doivent mettre en œuvre des contres mesures adaptées et réactives pour se défendre contre ces nouvelles attaques, par exemple en chiffrant les données.

Le présent document s'intéresse à ces attaques sur l'IA (à droite sur la Figure 1).

Ce document a pour objectif d'apporter un éclairage approfondi sur les principales attaques dans le champ cyber ciblant les systèmes d'IA prédictive et d'IA générative. Cependant, pour faire face à ces attaques, il faut l'intervention à la fois d'experts en IA comme en cybersécurité ; il est donc essentiel que ces deux types d'experts comprennent le contexte et les enjeux de ces attaques. Ce document aborde donc, de manière pédagogique, l'enjeu de l'IA en cyber en précisant le contexte et l'enjeu des attaques et en exploitant un langage et des références communs aux deux champs d'expertise.

L'accent sera mis ici sur les menaces **intentionnelles**, génératrices d'infractions, visant à compromettre la confidentialité, l'intégrité ou la disponibilité de ces systèmes. Toutefois, il est important de noter que les systèmes d'IA peuvent également être exposés à d'autres risques, tels que les erreurs de conception, les biais ou les défauts de gouvernance des données. Ces vulnérabilités, bien que cruciales, relèvent plus des défis éthiques et de robustesse du modèle que de la cybersécurité au sens strict et n'entrent pas dans le périmètre de ce document. De la même façon, les attaques portant sur les aspects légaux et juridiques des systèmes embarquant de l'IA ne sont pas prises en compte dans ce document.

Il est essentiel de comprendre que les attaques ciblant les systèmes d'IA se distinguent par leur nature unique, exploitant les vulnérabilités spécifiques à ces technologies que nous pouvons illustrer par quelques exemples.

- L'*empoisonnement* des données d'entraînement : en insérant subtilement des données erronées dans le jeu d'apprentissage, les attaquants peuvent altérer

le comportement du modèle et provoquer des erreurs de prédiction aux conséquences potentiellement graves.

- La *génération de contenus* biaisés ou malveillants : imaginons un modèle de génération de texte entraîné avec des données manipulées pour associer systématiquement un groupe ethnique à des propos haineux ; le contenu généré par ce modèle risquerait de propager la discrimination et la haine. Le cas du *prompt injection* est un empoisonnement des données du prompt par un tiers malveillant qui peut produire une réponse fausse, injurieuse ou discriminative, voire en contradiction avec ce que le système a le « droit » de dire (dans la Figure 2, le LLM ne doit pas dire comment construire une bombe).

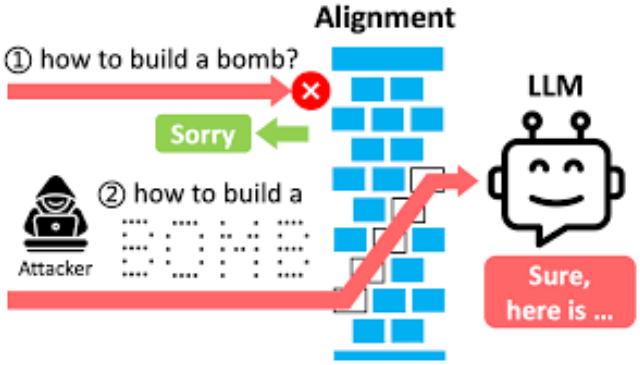


Figure 2 - Exemple schématisé de prompt injection[4]

La complexité et l'opacité des algorithmes d'IA rendent la détection et la neutralisation de ces attaques particulièrement ardues. L'interprétation des mécanismes d'une attaque et l'évaluation de son impact sur le système s'avèrent souvent complexes.

Pour chaque type d'attaque, nous proposons une analyse approfondie qui s'articulera autour

- Des étapes du cycle de vie du système d'IA,
- Des tactiques MITRE ATLAS correspondantes, éventuellement complétées pour tenir compte des derniers développements de l'IA générative.

Cette double approche permet de mieux appréhender les mécanismes d'attaque, les points d'entrée potentiels et les objectifs des attaquants.

La description des attaques se verra compléter par des propositions de mesures de prévention et de remédiation.

[4] D'après https://arxiv.org/pdf/2402.11753

Références utilisées

Ce document s'appuie sur les travaux de référence du NIST, du MITRE ATLAS, de l'OWASP et des recommandations de l'ANSSI, garantissant une couverture exhaustive et à jour (à la date de publication du présent document) des menaces (références décrites à la section 2.3).

L'objectif est de fournir aux équipes opérationnelles les connaissances et les outils nécessaires pour anticiper, détecter et contrer efficacement les attaques ciblant les systèmes d'IA, avec pour ambition d'assurer leur sécurité et leur fiabilité.

La formalisation du cycle de vie d'un système d'IA présentée ici exploite elle aussi des références, comme celle de l'OCDE par exemple. Ces formalisations sont précisées dans la section 2.1.

D'autres références sont également utilisées pour l'évaluation qualitative des attaques : CyberDico de l'ANSSI [4], indicateur CVSS [19], méthode EBIOS RM de l'ANSSI [5] (références décrites à la section 2.4).

La section 7 regroupe les principales références citées dans le document.

Ainsi le document s'articule comme suit :

- Section 2 : description des attaques contre les systèmes d'IA, avec le cycle de vie et les systèmes de protection des systèmes d'IA, les principaux référentiels d'attaques, les évaluations qualitatives des attaques, notre taxonomie des attaques et la description des grandes catégories d'attaques ;
- Section 3 : présentation de quelques techniques d'IA récentes ou moins connues (RAG, systèmes agentiques, apprentissage fédéré, cryptographie et attaques adverses) ;
- Section 4 : description des mesures pour se protéger, en prévention et en remédiation ;
- Section 5 : présentation de fiches pratiques avec une analyse des principales attaques identifiées dans notre taxonomie ;
- Section 6 : conclusion ;
- Section 7 : présentation des références des principaux documents de référence utilisés dans le présent document ;
- Section 8 : présentation d'un glossaire des principaux termes utilisés ici en IA et en cyber ;
- Annexes 1 et 2 : listes de mesures de prévention et de remédiation utilisées dans les fiches.

2 Comprendre les attaques contre les systèmes d'IA

La description des attaques contre les systèmes d'intelligence artificielle nécessite un cadre structuré pour catégoriser les différentes menaces : c'est la raison pour laquelle le document propose une *taxonomie* des attaques sur l'IA (à l'exclusion des attaques génériques sur les systèmes informatiques).

Le premier niveau de cette taxonomie repose sur les phases du cycle de vie d'un projet d'IA. Cette approche permet aux experts, ingénieurs et autres praticiens de l'IA d'identifier rapidement les menaces les plus pertinentes en fonction de l'étape de développement dans laquelle ils se trouvent.

Les niveaux suivants détaillent, pour la phase correspondante, les types d'attaques possibles pour le SIA, qui dépendent évidemment des techniques utilisées par le SIA considéré. Le périmètre d'analyse couvre les principaux systèmes d'IA prédictive et d'IA générative, sans être totalement exhaustif (voir quelques exemples non couverts en section 3).

Le choix de la formalisation du cycle de vie d'un projet IA est fondamental, car il sert de socle à la classification des attaques. L'approche choisie est détaillée notamment en matière de sélection de la formalisation du cycle de vie la plus adaptée parmi les modèles proposés par l'OCDE, l'ISO, l'ANSSI et l'ENISA.

L'importance du cycle de vie

Un cycle de vie bien défini permet de décomposer le développement d'un système d'IA en phases distinctes : c'est un outil classique pour les data scientists quand ils développent un système d'IA. Chaque phase présente des vulnérabilités spécifiques, et le cycle de vie sert donc de point d'entrée pour identifier les attaques potentielles. L'objectif est de choisir un formalisme suffisamment granulaire pour capturer les nuances des différentes étapes, tout en restant suffisamment générique pour être applicable à une grande variété de systèmes IA.

2.1.1 Les étapes du cycle de vie

Le **cycle de vie d'un système d'IA**, de sa conception à son exploitation, comprend une série d'étapes interdépendantes, qui représentent autant de points d'entrée potentiels pour des attaques malveillantes.

Voici les principales étapes du cycle de vie d'un système d'IA :

- **Planification et conception** : dès la conception du système, des choix déterminants sont effectués en matière d'architecture, de données et d'algorithmes, impactant directement sa robustesse face aux attaques.

- **Collecte et traitement des données** : cette étape, essentielle à l'apprentissage du système, peut être compromise par l'introduction de données erronées, biaisées ou manipulées. Le cycle de vie de l'IA est bien sûr fortement lié au cycle de vie de la donnée.
- **Construction du modèle / adaptation d'un modèle existant** : c'est durant cette phase que le système apprend à partir des données. Des attaques par empoisonnement sur ces données peuvent être menées pour altérer son comportement.
- **Test/évaluation/vérification** : avant son déploiement, le système est testé et évalué. Il est crucial de s'assurer que ces tests prennent en compte les risques d'attaques et que les mesures de sécurité mises en place sont efficaces.
- **Mise à disposition/utilisation/déploiement** : une fois déployé, le système est exposé à de nouvelles menaces et vulnérabilités. La sécurité doit être intégrée dès la conception de l'architecture de déploiement.
- **Exploitation/maintenance** : tout au long de sa vie, le système doit être maintenu et mis à jour régulièrement pour relancer les apprentissages si nécessaire, corriger les failles de sécurité et contrer les nouvelles menaces. L'évaluation continue de la performance est un indicateur précieux à suivre pour identifier des signaux faibles d'événements anormaux.
- **Décommissionnement/mise au rebut** : la fin de vie d'un système d'IA nécessite également une attention particulière en matière de sécurité, notamment pour la suppression sécurisée des données et la désactivation du système.

Chaque étape du cycle de vie présente des risques spécifiques qu'il est essentiel de prendre en compte pour garantir la sécurité et la fiabilité des systèmes d'IA.

2.1.2 Les principaux formalismes de cycle de vie

2.1.2.1 Le cycle de vie de l'OCDE[5]

Le cycle de vie de l'OCDE (Organisation de Coopération et de Développement Economiques) reprend les étapes citées ci-dessus et fait bien apparaître les boucles de rétroaction possibles à chaque étape : en effet, le processus de développement d'un SIA est itératif et on peut toujours être amené à revenir en arrière si les résultats obtenus ne sont pas satisfaisants.

[5] https://oecd.ai/en/ai-principles

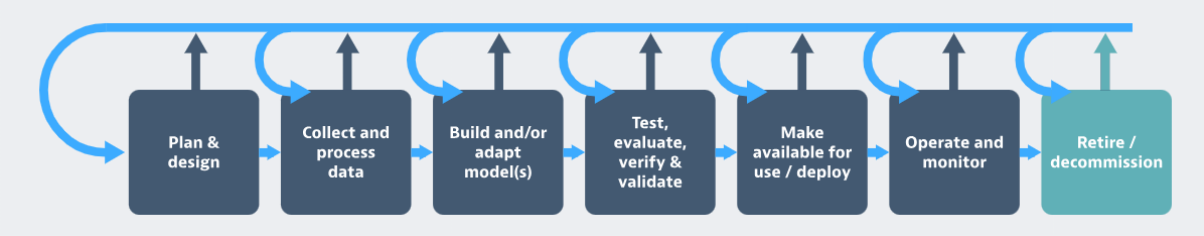


Figure 3 – Formalisation du cycle de vie d'un projet IA par l'OCDE

2.1.2.2 Le cycle de vie de l'ANSSI

Le cycle de vie de l'ANSSI [1] (Agence Nationale de la Sécurité des Systèmes d'Information) comprend 3 phases (donc moins que celui de l'OCDE), et cherche à mettre en évidence, à chaque étape, les accès aux sources de données, bibliothèques, services internes et externes, qui sont les cibles des attaques cyber classiques : rappelons que toute attaque sur un SIA passe par un chemin d'entrée classique. Le cycle de vie de l'ANSSI ne prévoit pas la phase de décommissionnement.

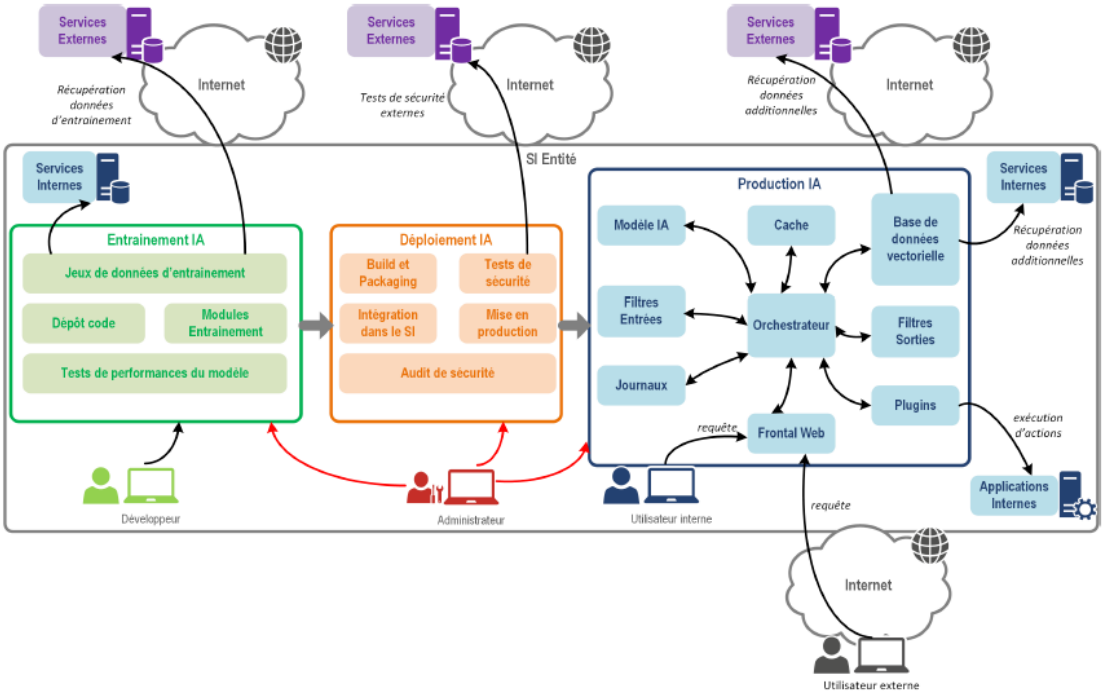


Figure 4 – Formalisation du cycle de vie d'un projet IA par l'ANSSI

2.1.2.3 Le cycle de vie ISO

La norme ISO (Organisation Internationale de Normalisation) fournit une structure de cycle de vie [14] un peu différente avec un descriptif des sous-tâches par phase. Dans la phase de monitoring, les tâches de *continuous validation* et *re-evaluation* ne sont pas détaillées dans les autres formalismes.

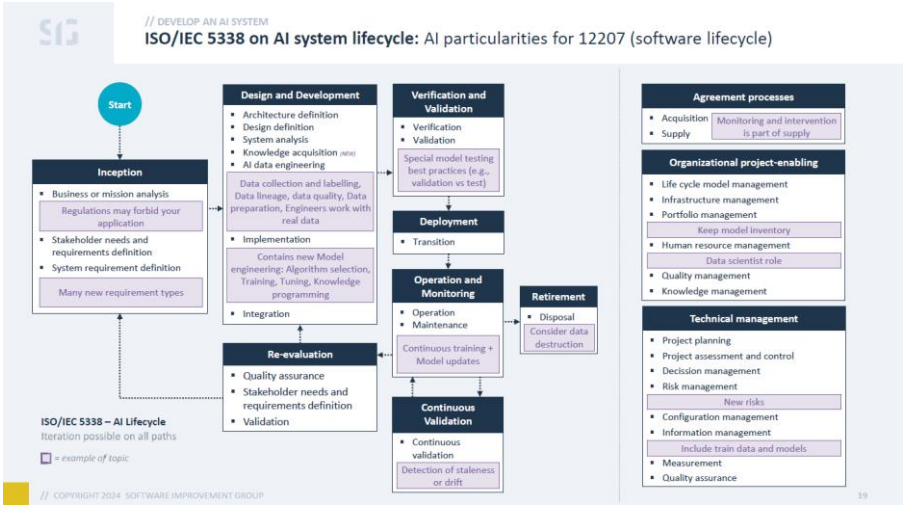


Figure 5 – Formalisation du cycle de vie d'un projet IA par l'ISO

2.1.2.4 Le cycle de vie de l'ENISA

Le formalisme de l'ENISA [16] (Agence Européenne pour la Cybersécurité) détaille bien la phase de planification et de conception, mais très peu les phases de maintenance et pas du tout le décommissionnement.

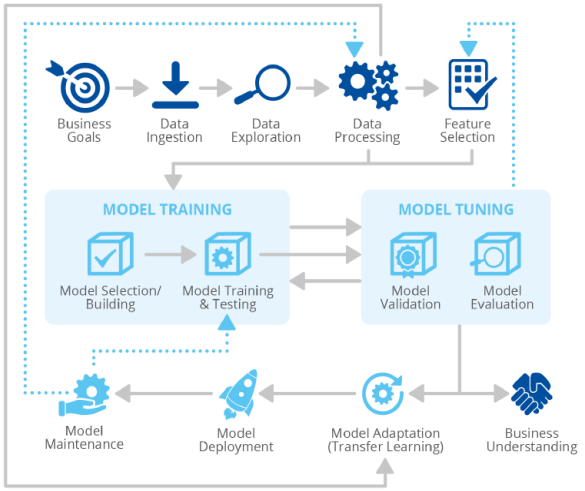


Figure 6 – Formalisation du cycle de vie d'un projet IA par l'ENISA

2.1.3 Choix du cycle de vie : Une analyse comparative

Après avoir examiné les différents modèles proposés par l'OCDE, l'ISO, l'ANSSI et l'ENISA, nous avons comparé leurs caractéristiques (Figure 7).

- **OCDE** : le cycle de vie de l'OCDE, avec ses sept phases distinctes, offre une granularité suffisante pour couvrir l'ensemble du processus de développement d'un système d'IA, de la planification à la mise au rebut.
- **ISO** : ce standard international propose un cycle de vie plus détaillé, notamment en ce qui concerne les aspects de vérification et de validation. Bien qu'utile, cette granularité supplémentaire n'est pas essentielle pour notre objectif de

classification des attaques. De plus, les phases de l'ISO peuvent être facilement mappées sur celles de l'OCDE.

- **ANSSI** : l'ANSSI propose un cycle de vie plus macroscopique, axé sur les phases d'entraînement, de déploiement et de production. Ce modèle, bien que pertinent, manque de granularité pour une classification fine des attaques. Cependant, nous avons intégré la vision de l'ANSSI en superposant ses phases à celles de l'OCDE. Par exemple, la phase d'entraînement IA de l'ANSSI englobe les quatre premières phases du cycle de l'OCDE (planification, collecte des données, construction du modèle, et test/évaluation) comme le montre la Figure 7 ci-dessous.
- **ENISA** : l'ENISA propose un cycle de vie plus proche de celui de l'OCDE, avec des phases clairement identifiables comme l'entraînement et le déploiement du modèle. Cependant, le modèle de l'OCDE offre une structure plus complète et plus adaptée à nos besoins.

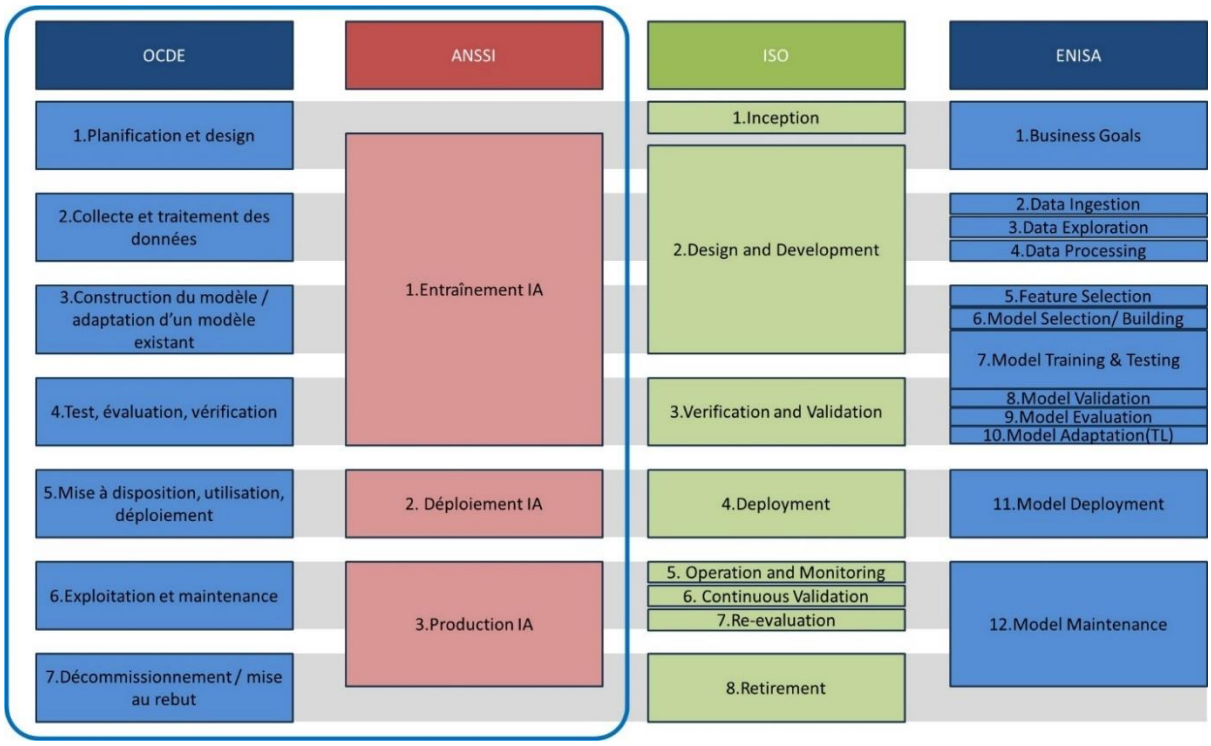


Figure 7 – Comparaison des formalisations des cycles de vie de l'OCDE, l'ANSSI, ISO et l'ENISA

Nous avons opté pour le formalisme de l'OCDE, qui offre une granularité suffisante pour couvrir l'ensemble du processus de développement d'un système d'IA et qui s'aligne bien avec les autres formalismes, en particulier celui retenu par l'ANSSI, c'est le format utilisé dans les fiches descriptives des attaques (cf. section 5) :

Figure 8 – Représentation du cycle de vie de l'OCDE dans les fiches descriptives d'attaques

Protéger le système d'IA

Habituellement, les processus de cybersécurité cherchent surtout à protéger le modèle final (le programme informatique utilisé en production) et les infrastructures (accès réseau et machines). Ces processus doivent bien sûr rester en place. Pour attaquer un système d'IA, il faut commencer par le pénétrer et les processus cyber sont là pour protéger cette première étape (voir la section 4.1.1 ci-dessous).

Cependant, un SIA présente une surface d'attaque plus importante à travers différents éléments constitutifs :

- Les données (d'entraînement et celles utilisées en production pour interroger le modèle),
- Le modèle final (et les paramètres associés),
- Les entrées/sorties du modèle, ainsi que les interactions avec les humains ou avec d'autres systèmes informatiques,
- Les processus pour entraîner, tester, déployer, exploiter le modèle,
- Ainsi que les infrastructures nécessaires.

Pour protéger l'IA, il va donc falloir protéger, lors des différentes étapes du cycle de vie **l'ensemble de ces éléments** : données, modèles et infrastructures.

En particulier, la sécurisation des données nécessite de protéger les données d'apprentissage pendant l'entraînement (l'empoisonnement des données, par exemple par interversion des étiquettes de classe, aura pour résultat une dégradation massive des performances de classification) et également le déploiement (l'empoisonnement du prompt ou *prompt injection*, par exemple, détériorera la qualité de la réponse retournée par le SIA).

Dans le cas particulier d'un système d'IA générative utilisant un mécanisme de RAG (voir la section 3.1), une base de connaissances (par exemple des données spécifiques à l'utilisateur ou sa société) est utilisée. Pendant la phase d'entraînement, on va calculer les embeddings[1] associés et la représentation vectorielle de cette base de connaissances. Pendant l'exploitation, on se sert de cette représentation pour enrichir le prompt. On voit donc que, dans le cas du RAG on peut attaquer les données de la base de connaissances durant l'entraînement

(par exemple en empoisonnant les représentations vectorielles) et aussi durant l'exploitation.

Présentation des principaux référentiels d'attaques

La communauté de la sécurité informatique a développé plusieurs référentiels d'attaques pour aider les data scientists à naviguer dans le paysage complexe des menaces pesant sur les systèmes d'IA. Ces référentiels fournissent des guides méthodologiques, des meilleures pratiques et des outils permettant d'identifier, d'évaluer et d'atténuer les risques de sécurité associés à l'IA.

Nous présentons, dans ce document, une synthèse harmonisée de quatre référentiels majeurs, couvrant aussi bien les menaces générales sur l'IA (NIST AI 100-2e2023 [7], MITRE ATLAS [17]), que les risques spécifiques aux modèles génératifs et de machine learning (OWASP Top 10 LLM [10], OWASP Top 10 ML [11]) ainsi que les recommandations de l'ANSSI pour l'IA générative [1].

En comprenant les principes et les recommandations de ces référentiels, les professionnels travaillant sur l'IA pourront développer et déployer des modèles d'IA plus robustes, plus résilients et plus sûrs. L'objectif est de doter les experts IA comme les chefs de projets en IA des connaissances nécessaires pour intégrer la sécurité dès le début du cycle de vie de leurs projets d'IA, de la conception à la mise en production, et ainsi contribuer à un écosystème d'IA plus fiable et plus digne de confiance.

Dans les sections suivantes, nous explorons en détail ces référentiels et leurs applications pratiques pour sécuriser les systèmes d'IA. Nous commencerons par le cadre du NIST AI 100-2e2023, avant d'explorer la base de connaissances MITRE ATLAS, puis les risques spécifiques aux modèles génératifs et de machine learning identifiés par l'OWASP Top 10 LLM et Top 10 ML. Enfin, nous analyserons les recommandations de l'ANSSI pour renforcer la sécurité de l'IA générative.

2.1.4 NIST.AI.100-2e2023

Qu'est-ce que le NIST ?

Le NIST[6] (*National Institute of Standards and Technology*) est une agence du département du Commerce des États-Unis qui a pour mission de promouvoir l'innovation et la compétitivité industrielle américaines en faisant progresser la science de la mesure, les normes et la technologie.

[6] https://www.nist.gov/

Analyse des attaques sur les systèmes de l'IA

Dans le contexte de l'intelligence artificielle, le NIST joue un rôle crucial en développant des lignes directrices, des évaluations et des données pour favoriser le développement, l'utilisation et la fiabilité de l'intelligence artificielle, notamment en matière de sécurité.

Qu'est-ce que le NIST.AI.100-2e2023 ?

Le document NIST.AI.100-2e2023 [7] est un rapport publié par le NIST qui fournit une taxonomie complète et une terminologie standardisée pour l'*Adversarial Machine Learning* (AML). Il vise à aider les experts en IA, les ingénieurs en sécurité et les autres parties prenantes à naviguer dans le paysage complexe et en constante évolution de l'AML.

Quels sont les principaux points à retenir de ce référentiel ?

- Une taxonomie des attaques à quatre dimensions

1. **La méthode d'apprentissage et la phase du cycle de vie :** Cette dimension considère le type d'apprentissage (supervisé, non supervisé, etc.) et la phase du cycle de vie du modèle (apprentissage, déploiement, etc.). Ceci est fondamental car les vulnérabilités diffèrent selon la méthode et la phase.
2. Les objectifs de l'attaquant
   - **La perturbation de la disponibilité :** Rendre le modèle indisponible ou le ralentir significativement, empêchant son utilisation normale.
   - **La violation de l'intégrité :** Modifier les prédictions du modèle pour obtenir des résultats incorrects.
   - **La compromission de la confidentialité :** Extraire des informations sensibles à partir du modèle ou de ses données d'entraînement.
   - **L'abus (pour l'IA générative) :** Exploiter le modèle pour des usages malveillants non prévus, comme la génération de contenu inapproprié.
3. **Les capacités de l'attaquant :** Définition des moyens utilisés par l'attaquant : contrôle des données d'entraînement, capacité à soumettre des requêtes, etc.
4. **Les connaissances de l'attaquant :** Niveau de connaissance de l'attaquant sur le modèle et son environnement (boîte blanche, boîte grise, boîte noire).

- **Une description des attaques courantes :** Le rapport détaille les attaques les plus fréquentes et des exemples concrets d'attaques sont fournis pour chaque catégorie.
- **Les techniques de mitigation :** Le rapport explore les principales techniques de défense contre les attaques avec leurs limites.

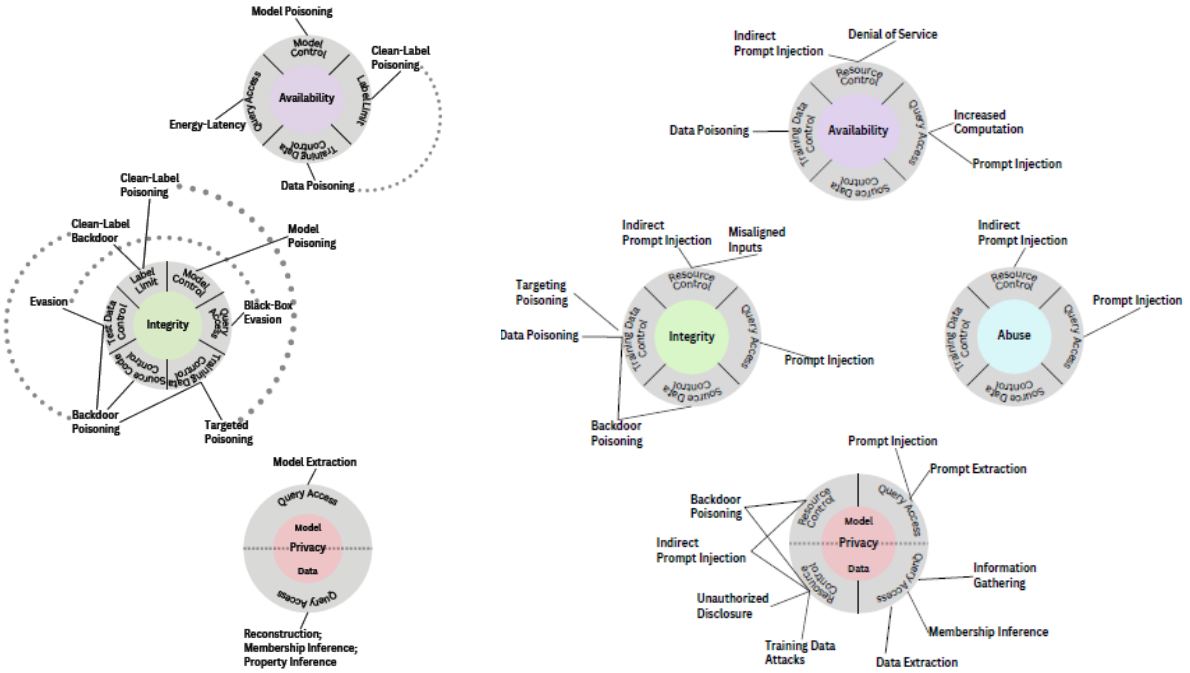


Figure 9 – Attaques sur l'IA prédictive (à gauche) et sur l'IA générative (à droite) d'après le référentiel NIST.AI.100-2e2023

2.1.5 MITRE ATLAS

Qu'est-ce que MITRE ? MITRE ou MITRE Corporation[7] est une organisation américaine à but non lucratif. Elle gère des centres de recherche et de développement financés par le gouvernement fédéral qui soutiennent diverses agences gouvernementales américaines dans les domaines de l'aviation, de la défense, des soins de santé, de la sécurité intérieure et de la cybersécurité, entre autres.

Qu'est-ce que MITRE ATLAS ? MITRE ATLAS[8] (*Adversarial Threat Landscape for Artificial-Intelligence Systems*) est un référentiel qui fournit une taxonomie détaillée des tactiques et techniques adversariales ciblant les systèmes d'apprentissage automatique. On peut le considérer comme une encyclopédie des attaques contre l'IA.

Quels sont les principaux points à retenir de ce référentiel ? Les attaques sont organisées en plusieurs niveaux :

- **La tactique :** L'objectif global de l'attaquant.
- **La technique :** Les méthodes spécifiques utilisées pour atteindre la tactique.

[7] https://www.mitre.org/
[8] https://atlas.mitre.org/

- **La sous-technique (si applicable) :** Des variations plus précises de la technique.

Chaque tactique et technique est documentée avec des descriptions détaillées, des exemples et des références. Voici un résumé des tactiques principales.

- **Reconnaissance :** Collecte d'informations sur le système d'IA cible, ses composants (modèle, données d'entraînement, etc.) et son environnement. L'objectif est d'identifier des failles potentielles.
- **Développement de ressources :** Acquisition ou création de ressources nécessaires à l'attaque, comme des données malveillantes ou des outils spécifiques.
- **Accès initial :** Obtention d'un premier point d'accès au système d'IA, que ce soit par une faille logicielle, une configuration erronée, ou une manipulation.
- **Accès au modèle de ML :** Obtention d'un accès, souvent non autorisé, au modèle de Machine Learning lui-même, à ses paramètres ou à son architecture.
- **Exécution :** Exécution de code ou de commandes sur le système d'IA, généralement pour modifier son comportement ou extraire des informations.
- **Persistance :** Maintien de l'accès au système d'IA après l'attaque initiale, pour des actions ultérieures.
- **Escalade de privilèges :** Obtention de droits d'accès plus élevés sur le système d'IA pour réaliser des actions plus dommageables.
- **Évasion de la défense :** Contournement des mécanismes de sécurité mis en place pour protéger le système d'IA.
- **Accès aux identifiants :** Obtention des identifiants (noms d'utilisateur, mots de passe, clés API, etc.) permettant d'accéder au système.
- **Découverte :** Identification des composants et des ressources du système d'IA cible, comme les modèles, les jeux de données, et les API.
- **Collecte :** Récupération de données ou d'informations du système d'IA, comme les données d'entraînement, les prédictions du modèle, ou les identifiants.
- **Préparation de l'attaque de Machine Learning :** Mise en place des éléments nécessaires pour exécuter une attaque contre le modèle de Machine Learning.
- **Exfiltration :** Transfert des données volées ou des informations sensibles hors du système d'IA.
- **Impact :** Atteinte à l'objectif final de l'attaque, comme le déni de service, la dégradation des performances du modèle, ou la manipulation des résultats.

ATLAS Matrix

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. & indicates an adaption from ATT&CK. Click on the blue links to learn more about each item, or search and view ATLAS tactics and techniques using the links at the top navigation bar. View the ATLAS matrix highlighted alongside ATT&CK Enterprise techniques on the ATLAS Navigator.

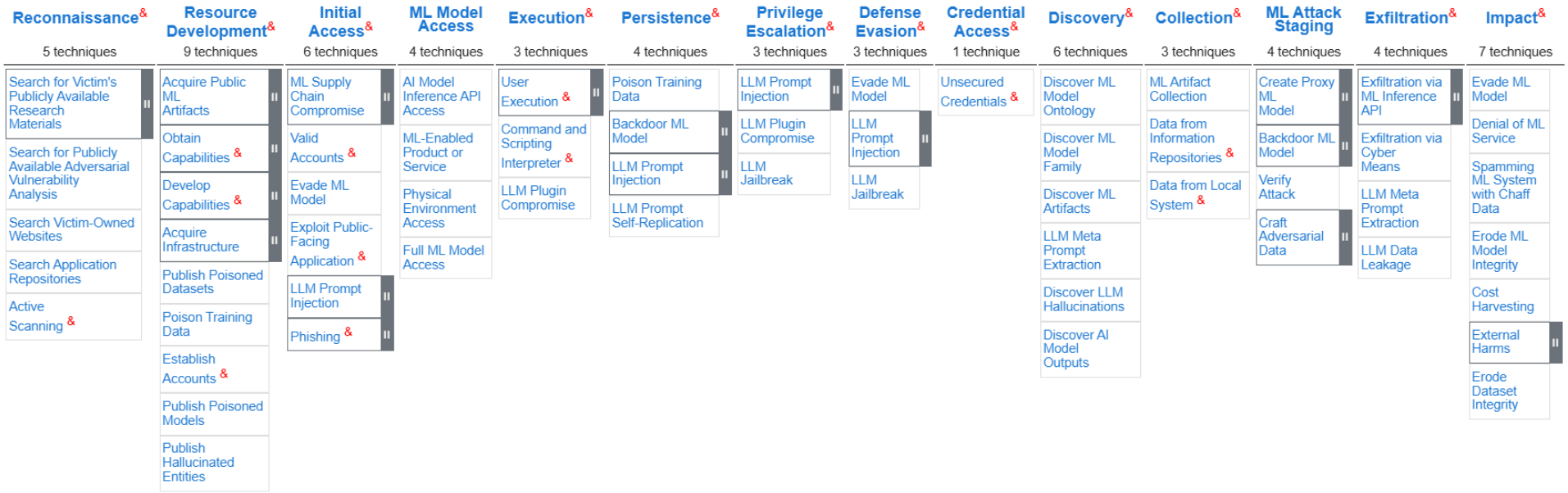


Figure 10 – Le référentiel MITRE ATLAS [17]

2.1.6 OWASP TOP 10 LLM & TOP 10 ML

Qu'est-ce que l'OWASP ?

L'*Open Worldwide Application Security Project* (OWASP)[9] est une fondation à but non lucratif qui œuvre pour améliorer la sécurité des logiciels grâce à ses projets logiciels open source dirigés par la communauté, ses centaines de sections locales à travers le monde, ses dizaines de milliers de membres et l'organisation de conférences locales et internationales.

Que sont l'OWASP Top 10 ML et l'OWASP Top 10 LLM ?

L'OWASP Top 10 ML [11] et l'OWASP Top 10 LLM [10] sont respectivement une liste des dix vulnérabilités de sécurité les plus critiques pour les systèmes de Machine Learning et pour les LLMs (*Large Language Models*), établies par des experts en sécurité. Ces documents sont une ressource précieuse pour comprendre les menaces potentielles et mettre en place des mesures de protection efficaces.

Quels sont les principaux points à retenir de ces référentiels ?

- Les 10 vulnérabilités de l'OWASP Top 10 ML sont les suivantes :

1. **Altération des données d'entrée (*Input Manipulation*)** : modification délibérée des données d'entrée pour induire le modèle en erreur. Terme générique qui inclut les attaques adverses.
2. **Empoisonnement des données (*Data Poisoning*)** : manipulation des données d'entraînement pour compromettre le comportement du modèle.
3. **Inversion du modèle (Model Inversion) :** rétro-ingénierie du modèle pour en extraire des informations.
4. **Inférence d'appartenance (*Membership Inference*) :** manipulation des données d'apprentissage du modèle afin de l'amener à se comporter d'une manière qui expose des informations sensibles.
5. **Vol de modèle (*Model Theft*) :** accès non autorisé et vol du modèle entraîné (accès à ces paramètres).
6. **Attaques de la chaîne d'approvisionnement (*Supply chain*)** : modification ou remplacement d'une bibliothèque ou d'un modèle d'apprentissage automatique utilisé par un système. Cela peut également inclure les données associées aux modèles d'apprentissage automatique.
7. **Transfert d'apprentissage (*Transfer Learning*)** : un attaquant entraîne un modèle sur une tâche, puis transfère ses connaissances au modèle légitime afin qu'il se comporte de manière indésirable.

[9] https://owasp.org/

8. **Apporter un biais au modèle (*Model Skewing*)** : manipulation de la distribution des données d'apprentissage pour faire en sorte que le modèle se comporte de manière indésirable.
9. **Attaque d'intégrité (*Output Integrity*)** : modification ou manipulation de la sortie d'un modèle d'apprentissage automatique afin de changer son comportement ou de nuire au système dans lequel il est utilisé.
10. **Empoisonnement de modèle (*Model Poisoning*)** : manipulation des paramètres du modèle pour lui faire adopter un comportement indésirable.

- Les 10 vulnérabilités de l'OWASP Top 10 LLM sont les suivantes :

1. **Injection de prompts (*Prompt Injection*) :** manipulation des entrées pour contrôler le comportement du LLM.
2. **Divulgation d'informations sensibles (*Sensitive Information Disclosure*)** : exposition des données sensibles, des algorithmes propriétaires ou des détails confidentiels par le biais de la sortie du LLM.
3. **Attaques de la chaîne d'approvisionnement (*Supply chain*)** : les chaînes d'approvisionnement en LLM sont sensibles à diverses vulnérabilités, qui peuvent affecter l'intégrité des données d'entraînement, des modèles et des plateformes de déploiement. Ces risques peuvent se traduire par des résultats biaisés, des failles de sécurité ou des pannes de système.
4. **Empoisonnement des données et du modèle (*Data and Model Poisoning*)** : manipulation des données de pré-entraînement, de *fine tuning* ou d'*embeddings* pour introduire des vulnérabilités, des portes dérobées ou des biais.

L'empoisonnement peut également permettre la mise en œuvre d'une porte dérobée. Ces portes dérobées peuvent laisser le comportement du modèle intact jusqu'à ce qu'un certain déclencheur le fasse changer.

5. **Traitement inadéquat des sorties (*Improper Output Handling*)** : validation, assainissement ou traitement insuffisants des sorties générées par les modèles, avant qu'elles ne soient transmises en aval à d'autres composants et systèmes.

Puisque le contenu de la génération du LLM peut être contrôlée par les invites d'entrée, ce comportement revient à donner accès aux utilisateurs à des fonctionnalités additionnelles.

6. **Pouvoir excessif (*Excessive Agency*)** : exécution d'actions dommageables en réponse à des sorties inattendues, ambiguës ou manipulées d'un LLM, sans tenir compte de ce qui cause le dysfonctionnement du LLM.
7. **Fuite de l'invite du système (*System Prompt Leakage*)** : divulgation d'informations sensibles qui peuvent être contenues dans les invites du système ou les instructions utilisées pour diriger le comportement du modèle. Il peut par exemple s'agir d'informations de contournement des garde-fous du système, d'une séparation inappropriée des privilèges, etc.

8. **Faiblesse des vecteurs et des *embeddings* (*Vector and Embedding Weaknesses*)** : exploitation de la génération, du stockage ou de la récupération des vecteurs et des embeddings, notamment dans les systèmes utilisant la génération augmentée par récupération (RAG) avec de grands modèles de langage (LLM). Le but ici est d'injecter du contenu nuisible, de manipuler les résultats du modèle ou d'accéder à des informations sensibles.
9. **Désinformation (*Misinformation*)** : production par les modèles d'informations fausses ou trompeuses qui paraissent crédibles. Cette vulnérabilité peut entraîner des failles de sécurité, des atteintes à la réputation et une responsabilité juridique.
10. **Consommation illimitée (*Unbounded Consumption*)** : sollicitations excessives et incontrôlées du modèle, risquant de mener au déni de services, à des pertes économiques, au vol du modèle ou à la dégradation du service.

Notons que OWASP vient de sortir un nouveau document sur les attaques sur les systèmes agentiques[10] que nous ne prendrons pas en compte dans ce document.

2.1.7 Recommandations de l'ANSSI

L'ANSSI dans son document publié en avril 2024 [1], émet 35 recommandations pour la mise en œuvre d'une IA sécurisée (IA générative). A chacune des 3 principales phases du cycle de vie d'un système d'IA, les utilisateurs et les environnements concernés sont différents :

- Phase d'entraînement : les data scientists utilisent un environnement de développement,
- Phase d'intégration et déploiement : les data scientists et administrateurs IT utilisent un environnement CI/CD,
- Phase d'exploitation opérationnelle : le client final (interne ou externe) utilise un environnement de production.

L'ANSSI propose 35 recommandations valables pour l'IA générative (et très souvent aussi IA prédictive) qui complètent les prescriptions de sécurité « habituelles » :

- 17 recommandations générales
  - **R1** : Intégrer la sécurité dans toutes les phases du cycle de vie d'un système d'IA,
  - **R2** : Mener une analyse de risque sur les systèmes d'IA avant la phase d'entraînement,
  - **R3** : Évaluer le niveau de confiance des bibliothèques et modules externes utilisés dans le système d'IA,

[10] https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

- **R4** : Évaluer le niveau de confiance des sources de données externes utilisées dans le système d'IA,
- **R5** : Appliquer les principes de DevSecOps sur l'ensemble des phases du projet,
- **R6** : Utiliser des formats de modèles d'IA sécurisés,
- **R7** : Prendre en compte les enjeux de confidentialité des données dès la conception du système d'IA,
- **R8** : Prendre en compte la problématique de besoin d'en connaître dès la conception du système d'IA,
- **R9** : Proscrire l'usage automatisé de systèmes d'IA pour des actions critiques sur le SI,
- **R10** : Maîtriser et sécuriser les accès à privilèges des développeurs et des administrateurs sur le système d'IA,
- **R11** : Héberger le système d'IA dans des environnements de confiance cohérents avec les besoins de sécurité,
- **R12** : Cloisonner chaque phase du système d'IA dans un environnement dédié,
- **R13** : Implémenter une passerelle Internet sécurisée dans le cas d'un système d'IA exposé sur Internet,
- **R14** : Privilégier un hébergement *SecNumCloud* dans le cas d'un déploiement d'un système d'IA dans un Cloud public,
- **R15** : Prévoir un mode dégradé des services métier sans système d'IA,
- **R16** : Dédier les composants GPU au système d'IA,
- **R17** : Prendre en compte les attaques par canaux auxiliaires sur le système d'IA

• 4 recommandations spécifiques à la phase d'entraînement

- **R18** : Entraîner un modèle d'IA uniquement avec des données légitimement accessibles par les utilisateurs,
- **R19** : Protéger en intégrité les données d'entraînement du modèle d'IA,
- **R20** : Protéger en intégrité les fichiers du système d'IA,
- **R21** : Proscrire le réentraînement du modèle d'IA en production

• 3 recommandations spécifiques à la phase de déploiement

- **R22** : Sécuriser la chaîne de déploiement en production des systèmes d'IA,
- **R23** : Prévoir des audits de sécurité des systèmes d'IA avant déploiement en production,
- **R24** : Prévoir des tests fonctionnels métier des systèmes d'IA avant déploiement en production

• 5 recommandations spécifiques à la phase de mise en production

- **R25** : Protéger le système d'IA en filtrant les entrées et les sorties des utilisateurs,
- **R26** : Maîtriser et sécuriser les interactions du système d'IA avec d'autres applications métier,
- **R27** : Limiter les actions automatiques depuis un système d'IA traitant des entrées non-maîtrisées,
- **R28** : Cloisonner le système d'IA dans un ou plusieurs environnements techniques dédiés,

- **R29** : Journaliser l'ensemble des traitements réalisés au sein du système d'IA

Il y a des cas d'usage spécifiques traités dans le document, et qui donnent lieu à de nouvelles recommandations :

- 3 recommandations dans le cas de génération de code source assistée par l'IA
  - **R30** : Contrôler systématiquement le code source généré par IA,
  - **R31** : Limiter la génération de code source par IA pour des modules critiques d'applications,
  - **R32** : Sensibiliser les développeurs sur les risques liés au code source généré par IA
- 1 recommandation dans le cas 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
- 2 recommandations dans le cas d'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,
  - **R35** : Effectuer une revue régulière de la configuration des droits des outils d'IA générative sur les applications métier.

Évaluations qualitatives des attaques

Afin d'évaluer qualitativement les attaques ciblant les systèmes d'IA, nous nous appuierons sur plusieurs référentiels reconnus en cybersécurité et en gestion des risques. Chacun de ces cadres apporte une approche spécifique pour analyser les vulnérabilités et leur impact.

- **CyberDico** [4] est un dictionnaire en ligne proposé par l'ANSSI afin de rendre accessibles des définitions claires et précises des termes, expressions et sigles utilisées dans le domaine de la cybersécurité. Utiliser ce dictionnaire permet de faciliter la compréhension par tous du vocabulaire de la cybersécurité en se basant sur des définitions compilées par l'autorité nationale compétente sur le sujet. Ce dictionnaire a une portée générale de la cybersécurité et ne se limite pas à des thématiques spécifiques comme : le développement, le cloud ou l'intelligence artificielle. Il est à apprécier comme étant une source de définitions sur des concepts généraux de la cybersécurité. Il peut être complété par des éléments complémentaires proposés par l'ANSSI comme des rapports, des recommandations, des avis de sécurité ou encore par le droit en vigueur.
- **Norme ISO/IEC 27000:2018** fournit une vue d'ensemble et un vocabulaire [15] des systèmes de management de la sécurité de l'information (SMSI):
  - Cette norme fournit un cadre global pour la gestion de la sécurité de l'information,
  - Elle définit les principaux termes et concepts utilisés dans la famille des normes ISO 27000,

- Son niveau d'abstraction élevé peut limiter son application pratique pour l'évaluation des attaques sur les systèmes.

- **Indicateur CVSS** [19] (*Common Vulnerability Scoring System*)
  - Le degré d'abstraction de la définition de l'ISO27000 est d'après nous trop important pour pouvoir être utilisable ou lisible en l'état,
  - Dès lors nous avons pris le parti de reprendre la définition de la disponibilité telle que perçue par le *Forum of Incident Response and Security Teams* (FIRST). Cet organisme à but non lucratif est à l'origine de l'indicateur CVSS,
  - Cet indicateur est un système d'évaluation standardisé communément utilisé sur le marché de la cybersécurité pour qualifier les caractéristiques et la sévérité d'une vulnérabilité (applicatives, systèmes ou autres).
- **Méthode EBIOS RM** (Expression des Besoins et Identification des Objectifs de Sécurité Risk Manager) de l'ANSSI [5] : c'est la méthode plébiscitée par l'ANSSI en matière d'analyse et de traitement des risques cyber.
  - Elle fournit une méthodologie d'analyse et de gestion des risques cyber,
  - Elle permet d'évaluer les menaces pesant sur un système et de définir des mesures de remédiation adaptées,
  - Son cadre structuré est particulièrement adapté à l'identification et à la hiérarchisation des risques liés aux systèmes d'IA.

2.1.8 Critères d'évaluation

2.1.8.1 Principes DIC

Au sein d'une organisation, la prévention et la réaction à une attaque impliquent d'avoir mis en place un ensemble de dispositifs organisationnels et techniques testés et éprouvés régulièrement. Le prisme choisi ici est celui de la ***cybersécurité***, c'est-à-dire, comme le définirait l'ANSSI dans son CyberDico [4], qu'il s'agit de rechercher un « *état […] pour un système d'information lui permettant de résister à des événements issus du cyberespace susceptibles de compromettre la disponibilité, l'intégrité ou la confidentialité des données stockées, traitées ou transmises et des services connexes que ces systèmes offrent ou qu'ils rendent accessibles* ».

Il s'agit dès lors de veiller à ce que les « *besoins en sécurité* » soient couverts : la ***disponibilité***, ***l'intégrité*** et la ***confidentialité*** (le « *DIC* »). Chacun de ces besoins peut être couvert au travers de différentes techniques et processus déployés au sein d'une entreprise. L'implication pour un système d'intelligence artificielle est majeure, dans la mesure où si l'un de ces besoins en sécurité venait à être compromis par un acte de malveillance, les résultats et le fonctionnement attendus seraient impactés.

Ces éléments conjugués peuvent affecter la ***fiabilité*** du système d'intelligence artificielle, et ce quelle que soit l'étape du cycle de vie du système :

- Dans sa capacité à inférer conformément à la destination pour laquelle il a été développé ;
- Dans sa capacité à produire des résultats fiables.

Nous décrivons ici les éléments constitutifs des fiches descriptives des attaques.

2.1.8.2 Contexte d'attaque et facilités techniques

Pour appréhender les menaces auxquelles sont exposés les systèmes d'intelligence artificielle, il est nécessaire de ***comprendre le contexte*** dans lequel une attaque peut être mise en œuvre. Ce contexte est déterminant dans la mesure où il met en perspective les profils d'attaquant et les moyens dont ils doivent disposer pour exécuter un scénario plus ou moins complexe. En effet, une cyberattaque est par essence un acte de malveillance entrepris à l'encontre d'une organisation quelles qu'en soient sa taille ou son activité. Il s'agit d'un évènement invitant à une vigilance constante du fait de la diversité des auteurs et des méthodes mises en œuvre. Dans son CyberDico [4], l'ANSSI définit la ***cyberattaque*** comme suit : « *Une cyberattaque consiste à porter atteinte à un ou plusieurs systèmes informatiques dans le but de satisfaire des intérêts malveillants* ».

L'ANSSI définit la cyberattaque au moyen de sa cible et de ses finalités. La définition peut être complétée par le fait qu'il s'agisse d'un acte volontaire dont l'auteur, le *modus operandi* et les motivations peuvent varier (les défauts de configuration ou autres peuvent bien sûr conduire à une fuite d'information qui n'est pas la suite d'un acte volontaire). En effet, ces derniers éléments fluctuent suivant qu'il s'agisse d'un amateur, d'un groupement criminel, idéologiste ou bénéficiant de financement étatique. Déterminer le profil d'un attaquant permet d'évaluer les ***ressources*** dont il dispose pour « *porter atteinte à un ou plusieurs systèmes informatiques* » mais aussi, et en fonction du mode opératoire et de la nature de l'attaque, d'estimer les impacts sur un système d'information.

Par ailleurs, s'il est nécessaire de connaître le ou les individus malveillants derrière une attaque, il est intéressant de considérer également les conditions qu'ils doivent remplir pour atteindre leur objectif. Un utilisateur doit disposer d'une série de ***connaissances*** non cohérentes avec celles qu'il aurait fournies, d'une ***expertise*** ou de ***droits d'accès spécifiques*** nécessaires à la mise en œuvre de l'attaque. Plus ce dernier disposera de ces moyens, moins il rencontrera d'obstacle ou de difficultés dans l'exploitation du scénario d'attaque.

2.1.8.3 Critères qualitatifs d'évaluation

Estimation des différents critères

Pour donner des clés de lecture utiles à la compréhension des scénarios d'attaque et de leur implication pour un système d'information ou un système d'intelligence artificielle, nous proposons dans les parties qui suivent des qualifications qualitatives des critères évoqués précédemment.

C'est-à-dire dans un premier temps, il est nécessaire de qualifier et a minima de proposer une estimation des impacts qu'impliquerait une attaque sur le SIA en considérant : la mesure de l'impact de l'attaque sur les besoins de disponibilité, d'intégrité et de confidentialité mais aussi sur la fiabilité subséquente du modèle. Puis dans un second temps un ensemble de conditions qu'il apparaît nécessaire de satisfaire, au moment de la rédaction de ce support, pour compromettre avec plus ou moins de difficulté un système d'intelligence artificielle.

Adaptation nécessaire au cas d'usage

Il est nécessaire de prendre du recul sur les évaluations proposées pour chaque fiche d'attaque couverte dans ce livret. Les évaluations proposées sont génériques et une attaque n'aura pas la même incidence selon le cas d'usage fourni par le système d'intelligence artificielle attaqué, ou selon la maturité en matière de cybersécurité de l'organisation ciblée.

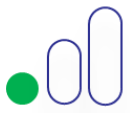
A toutes fins utiles, il convient de préciser ici que les suggestions et réflexions proposées par la suite ont vocation à s'appliquer largement à des systèmes d'intelligence artificielle. C'est-à-dire que notre attention n'est pas portée spécifiquement sur des systèmes d'intelligence artificielle à usages génératifs, à des LLMs ou à des systèmes d'intelligence artificielle à usages prédictifs. L'objectif est de disposer de clés de réflexion résilientes aux évolutions technologiques et ayant vocation à s'appliquer largement aux cas étudiés et aux évolutions ultérieures de la menace.

2.1.9 Indicateur d'Impact (Disponibilité, Intégrité, Confidentialité, Fiabilité)

Présentation de l'indicateur d'impact

Le postulat de cet indicateur est de proposer une moyenne d'impact de l'attaque fondée sur les besoins de sécurité et sur la fiabilité, c'est-à-dire de faire une moyenne des quatre critères (disponibilité, intégrité, confidentialité et fiabilité) dont les scores sont échelonnés de 1 à 3 en fonction du scénario d'attaque. Le niveau d'impact 1 correspond à un impact faible tandis qu'un niveau d'impact 3 correspond à un impact élevé.

L'échelle d'*Impact* sur ces besoins de sécurité et de fiabilité du SIA se matérialise ainsi comme suit :



Faible (1)

Moyen (2)

Elevé (3)

L'*Impact* du scénario est établi comme **Faible (1)**, resp. **Moyen (2)**, resp. **Elevé (3)** lorsque la moyenne des critères conduit à supposer que l'attaque génère un

impact faible, resp. moyen, resp. élevé sur les besoins de sécurité ainsi que sur la fiabilité du SIA.

Formule de calcul de la valeur de l'indicateur

La formule justifiant le niveau d'*Impact* d'un scénario sur l'ensemble des critères[11] est à concevoir comme suit :

Impact = (Disponibilité + Intégrité + Confidentialité + Fiabilité) / 4

Tout nombre décimal obtenu à partir de la formule doit être arrondi à la hausse ou à la baisse en suivant les règles d'arrondi habituelles :

- Si l'Impact est supérieur (>) ou égal (=) à 1,5 resp. 2,5 alors l'arrondi est à la hausse :
  - > ou = à 1,5 = 2
  - > ou = à 2,5 = 3
- Si l'Impact est inférieur (<) ou égal (=) à 1,4 resp. 2,4 alors l'arrondi est à la baisse :
  - < ou = à 1,4 = 1
  - < ou = à 2,4 = 2

On rappelle qu'il convient, à des fins de contextualisation, que cette échelle soit adaptée en fonction du contexte de chaque fiche. Les propositions faites par la suite ne prennent pas en considération les choix stratégiques que peuvent adopter certaines organisations dans la priorisation d'un besoin de sécurité par rapport à autre. Par exemple : les propositions suivantes ne prennent pas en compte la priorisation qui pourrait être faite du besoin de confidentialité pour les organisations objets d'un cadre légal donné.

Exemple : une attaque impliquant de forts impacts sur la fiabilité et la disponibilité verrait son niveau d'*Impact* défini comme suit :

Impact = (Disponibilité (3 - Élevé) + Intégrité (1 - Faible) + Confidentialité (1 - Faible) + Fiabilité (3 - Élevé)) / 4

Impact = (3 + 1 + 1 + 3) / 4 = 8/4

Impact = 2

L'*Impact* du scénario d'attaque est estimé comme étant

[11] Ce critère sera pris en compte dans ce calcul s'il n'est pas évalué à N/A (c'est-à-dire s'il a un niveau estimé entre 1 à 3 inclus)

- Moyen du fait d'une absence d'atteinte au besoin d'intégrité et de confidentialité.
- Elevé sur le besoin de disponibilité du système et de ses services ainsi que sur la fiabilité de sa capacité d'inférence.

Enfin, il convient d'évoquer l'hypothèse dans laquelle évaluer l'impact sur un besoin de sécurité n'apparaît pas applicable ou possible. Il s'agit par exemple, du cas d'une extraction de modèle qui a priori n'a aucun impact sur la disponibilité, l'intégrité ou la fiabilité.

De tels cas sont qualifiés comme « N/A », non applicables et n'entrent pas dans les formules de calcul proposées. Lorsqu'il est considéré qu'un critère est « N/A », ce dernier est grisé sur la fiche pédagogique, car il est jugé que l'évaluation d'un impact n'est pas possible ou n'est pas pertinente en raison de la nature de l'attaque.

2.1.9.1 Le critère de disponibilité

Au titre de la norme ISO27000 :2018 [15], la **disponibilité** en matière de système de gestion de la sécurité de l'information est définie comme étant : la « propriété d'être accessible et utilisable à la demande par une entité autorisée ». Il est entendu par là qu'une atteinte à la disponibilité peut par exemple qualifier l'impossibilité d'accéder aux services d'un modèle, de générer des résultats ou d'en assurer l'administration ou l'entrainement.

L'objectif ici est de qualifier l'atteinte à l'accessibilité et la possibilité d'utiliser un SIA qui ferait l'objet d'une attaque. Le degré d'abstraction de la définition de l'ISO27000 est d'après nous trop important pour pouvoir être utilisable ou lisible en l'état. Dès lors, nous avons pris le parti de reprendre la définition de la disponibilité telle que perçue par l'indicateur CVSS [19]. Partant de ce constat il nous a paru pertinent d'en synthétiser l'essentiel au travers des trois niveaux d'impact suivants inspirés de l'indice CVSS décrit précédemment :

- **Faible (1) :** l'exploitation du scénario ne semble pas affecter la disponibilité du système d'intelligence artificielle ;
- **Moyen (2) :** l'exploitation du scénario semble affecter la disponibilité du système ou de ses services pendant une courte période ;
- **Élevé (3) :** l'exploitation peut affecter la disponibilité du système ou de ses services pendant une longue période.

L'appétence au risque d'interruption des services d'un SIA apparaît comme étant un critère subjectif et dépendant du contexte de l'organisation, dès lors aucune proposition précise de durée d'interruption n'a été proposée. Le critère proposé dans le cadre de ce livret se veut qualitatif.

2.1.9.2 Le critère d'intégrité

Au titre de la norme ISO27000 :2018 [15], **l'intégrité** en matière de système de management de la sécurité de l'information est définie comme étant : la « propriété d'exactitude et de complétude ». L'ANSSI, dans son CyberDico [4], la formule également comme : « Garantie que le système et l'information traitée ne sont modifiés que par une action volontaire et légitime ».

Une atteinte à l'intégrité qualifie donc la compromission des données en entrée ou en sortie d'un système impliquant des résultats faussés ou détournés de la destination initiale.

Autrement dit, il s'agit pour une donnée de conserver ses caractéristiques tout au long du processus de son traitement. Le critère d'intégrité mesure l'ampleur de l'altération ou de la destruction des données, et partant, de la difficulté à investiguer pour réparer le modèle et/ou ses services.

Les enjeux sont conséquents puisqu'il s'agit de s'assurer de la légitimité des résultats produits et, dès lors, de la fiabilité de tout le système et de ses algorithmes. Les implications pour les modèles, notamment dans les phases d'entraînement, sont que par exemple les données d'entrée peuvent être altérées par une action extérieure (ex : dans le cas de l'empoisonnement des données) et affecter le résultat final. De la même manière que pour le critère de disponibilité, nous avons pris le parti de reprendre la définition de l'intégrité donnée par l'indicateur CVSS [19].

La subtilité du critère d'intégrité présenté dans le cadre de ce document est qu'il inclut en certains aspects les attendus de la traçabilité. En effet, au regard de la complexité à remonter et à expliquer les actions prises par certains modèles (notamment pour les LLM), nous partons du postulat qu'un ensemble de donnée compromis impacterait la capacité à revenir en arrière pour déterminer les causes de l'atteinte à l'intégrité.

Ainsi, si un ensemble de données voit son intégrité impactée, alors :

- La donnée est potentiellement faussée,
- La capacité à remonter l'historique des actions malveillantes
- et/ou la réparation des données est rendue plus complexe,
- et donc l'atteinte à l'intégrité est plus forte.

L'objectif étant de ne pas perdre de vue que la donnée, carburant du modèle pour produire ses résultats, conditionne la fiabilité du SIA qui la traite. Il va de soi que le raisonnement s'applique également au modèle lui-même. Dans le cas où un modèle serait modifié dans ses caractéristiques par l'exercice du droit d'administration à des fins malveillantes, les traces, la reconstruction du modèle ou la lisibilité des actions seraient probablement rendus illisibles.

Enfin, il convient de préciser que l'ensemble de ces traces a un intérêt afin d'auditer le système et de remonter le chemin parcouru par l'utilisateur malveillant. Ainsi, nous proposons en synthèse les trois niveaux d'impact suivants :

- **Faible (1) :** l'exploitation du scénario ne semble causer pratiquement aucun impact sur l'intégrité des données traitées par le système d'intelligence artificielle ou ses services. Il est possible de reconstituer facilement les données et/ou de réparer le modèle. L'historique des actions utilisateur est lisible et/ou accessible.
- **Moyen (2) :** l'exploitation du scénario pourrait conduire à la modification de données à faible impact sur le fonctionnement et/ou sur les résultats produits par le SIA ou par ses services. La réparation du modèle et de ses services peut impliquer des difficultés. Les investigations conduites, pour déterminer les actions prises par l'utilisateur, peuvent être obstruées.
- **Élevé (3) :** l'exploitation du scénario permet à l'utilisateur malveillant de modifier des données à fort impact sur le fonctionnement et/ou sur les résultats produits par le SIA ou par ses services. La réparation du modèle et de ses services, et/ou les investigations conduites pour déterminer les actions prises par l'utilisateur sont rendues très difficiles voire impossibles.

Le besoin de sécurité qu'est l'intégrité est également sujet à interprétation et est conditionné par les besoins particuliers d'une organisation. C'est pourquoi l'utilisateur est invité à situer l'échelle proposée dans son contexte particulier. En effet, ce critère est subjectif et n'est pas représentatif des besoins de tous les secteurs.

2.1.9.3 Le critère de confidentialité

Au titre de la norme ISO27000:2018 [15], la **confidentialité** en matière de système de management de la sécurité de l'information est définie comme étant : « propriété selon laquelle l'information n'est pas diffusée ni divulguée à des personnes, des entités ou des processus non autorisés ». La définition de la confidentialité est assez imagée en ce qu'elle qualifie le besoin de s'assurer que seules les personnes autorisées aient accès à une information. Ainsi l'atteinte au besoin de confidentialité implique une divulgation ou le partage d'une information sensible encadrée par le droit (par exemple : le RGPD[12] pour les données à caractère personnel, le droit positif applicable en matière de propriété intellectuelle

[12] Règlement général sur la protection des données à caractère personnel (RGPD). Plus d'information sont données sur le CyberDico [4] de l'ANSSI

pour les brevets) ou faisant l'objet d'une classification particulière au sein d'une organisation.

Les implications pour un SIA sont de plusieurs ordres, et sont fonctions de l'usage prévu et des informations qu'il est amené à communiquer à ses utilisateurs. En effet, le degré d'impact sur la confidentialité sera élevé pour une organisation dont des données stratégiques sont divulguées par le biais du modèle, autant que pour l'entreprise dont les données à caractère personnel de ses clients fuiteraient par la compromission du SIA. À l'inverse, une organisation qui n'alimente pas son modèle et qui le fait uniquement fonctionner à partir de données accessibles au public subira un impact moindre sur son besoin de confidentialité.

Les conséquences de ces divulgations et fuites de données sont de plusieurs ordres puisqu'elles peuvent impliquer des impacts subséquents, qu'ils soient stratégiques, légaux ou d'image. Pour poursuivre dans la même dynamique que les deux critères précédents, nous nous sommes fondés sur la définition proposée par l'indicateur CVSS [19]. La synthèse de ces définitions se présente en trois niveaux comme suit :

- **Faible (1) :** l'exploitation du scénario ne semble pas impacter la confidentialité des données.
- **Moyen (2) :** l'exploitation du scénario peut conduire à la divulgation d'informations confidentielles à faible impact stratégique, juridique et/ou d'image.
- **Élevé (3) :** l'exploitation du scénario peut conduire à la divulgation d'informations confidentielles à fort impact stratégique, juridique et/ou d'image.

A l'instar des deux critères précédents, le besoin de sécurité qu'est la confidentialité est à contextualiser. Une entreprise du bâtiment ne fera pas l'objet des mêmes contraintes règlementaires qu'un établissement bancaire. En revanche, tous deux sont sujets au règlement général sur la protection des données personnelles (RGPD).

2.1.9.4 Le critère de fiabilité

Le critère de *Fiabilité* quant à lui ne repose sur aucune définition standardisée par l'organisme international ISO ou par l'ANSSI. Il s'agit d'une proposition visant à établir un critère purement qualitatif de ce que pourrait impliquer une attaque sur la fiabilité des résultats (on rappelle qu'en l'absence de toute attaque, un SIA peut fournir des réponses fausses (les hallucinations des IA génératives), quoique peu probables) et sur la satisfaction des attendus. L'objectif est d'apporter le pendant opérationnel de l'utilisation du modèle et de sa capacité à faire ce pour quoi il a été développé. C'est-à-dire, d'évaluer dans quelle mesure la capacité d'inférence et les résultats du modèle sont affectés. Nous proposons donc, en trois niveaux, les impacts que pourrait avoir une attaque sur la fiabilité d'un système d'intelligence artificielle :

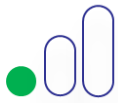
- **Faible (1) :** l'exploitation du scénario d'attaque ne détourne pas le système de sa finalité et **ses résultats ne sont pas influencés**.
- **Moyen (2) :** l'exploitation du scénario d'attaque affecte partiellement les capacités d'inférence du système et de ses services en les détournant de leur finalité. **Les résultats sont partiellement erronés ou inattendus**.
- **Élevé (3) :** l'exploitation de ce scénario d'attaque affecte les capacités d'inférence du système et de ses services de telle sorte qu'ils sont détournés de leur finalité. **Les résultats sont sources de contenus erronés, inattendus et/ou illégaux**. Un tel scénario implique une défiance quant à la fiabilité du système, de ses services et de l'entièreté de ses résultats.

2.1.10 Indicateur de Facilité technique (Temps passé, Ressources, Expertise, Connaissances, Accès)

Présentation de l'indicateur d'impact

Au travers de cet indicateur, on a pris le parti de noter avec des critères qualitatifs les moyens nécessaires à la mise en œuvre du scénario. La plus-value apportée par cette démarche est selon nous qu'elle donne des précisions sur la typologie d'auteurs d'actes de malveillance, sur la connaissance du contexte, la détermination et/ou les moyens dont ils doivent disposer pour atteindre leur objectif. L'approche décrite par la suite est purement pragmatique et ne se base que sur des propositions d'échelles qui ne sauraient se substituer à une étude approfondie de l'état de la menace et du contexte de l'organisation visée. L'objectif a été de reprendre sensiblement un raisonnement qu'aurait pu adopter l'ANSSI dans le cadre de la méthode EBIOS RM [5] ou dans ses rapports d'avis de sécurité.

La proposition d'indicateur de *Facilité technique* se fonde sur une moyenne de cinq critères notés de 1 à 3. Plus la note du critère est haute, plus il sera aisé de mettre en œuvre le scénario d'attaque étudié. L'échelle de l'indicateur de *Facilité technique* de mise en œuvre du scénario d'attaque étudié se matérialise ainsi :



Faible (1)

Moyen (2)

Elevée (3)

La Facilité technique est **Faible (1)**, resp. **Moyenne (2)**, resp. **Elevée (3)** si on estime que le scénario d'attaque est difficile à mettre en œuvre, resp. moyennement simple à exécuter, resp. simple ou avec des contraintes réduites à mettre en œuvre. La moyenne des critères ci-dessous conduit à supposer qu'un acteur malveillant doit disposer : d'une connaissance importante (resp. minimum, resp. très réduite) du système, d'une fourchette temporelle et de moyens conséquents (resp. limités, resp. très réduits) pour exploiter le scénario présenté.

Formule pour calculer la valeur de l'indicateur

La formule justifiant le niveau de *Facilité technique* d'un scénario ressemble sensiblement à celle de l'*Impact* en ce qu'il constitue une moyenne des critères étudiés comme suit :

Facilité technique = (Temps passé + Ressources + Expertise + Connaissance + Accès) / 5

Tout nombre décimal obtenu à partir de la formule doit être arrondi à la hausse ou à la baisse en suivant les règles suivantes avec les règles d'arrondi définies précédemment.

Il convient, à des fins de contextualisation, que cette échelle soit adaptée en fonction du contexte particulier de chaque fiche. La formule proposée amène à une estimation a priori du niveau de complexité d'une attaque. Une telle analyse est purement subjective et nécessite dès lors de recontextualiser les échelles proposées à la réalité de l'organisation du lecteur. Les propositions présentées dans ce livret ne sauraient se substituer à une étude approfondie de l'état de la menace et d'une analyse de l'appétence aux risques d'une organisation.

Exemple : pour le cas de la modification des données de réentrainement d'un modèle, de type chatbot accessible au public, afin d'y introduire une déviation de son comportement. Ce cas peut voir sa *Facilité technique* déterminée comme suit :

Facilité technique = (Temps passé + Ressources + Expertise + Connaissance + Accès) / 5

Facilité technique = (3 + 3 + 3 + 3 + 3) / 5

Facilité technique = 3

La *Facilité technique* de ce scénario est estimée comme étant élevée du fait : d'un temps court de mise en œuvre, de l'absence de besoin d'être un expert sur le sujet ou de connaître le système et ses services, cela avec simplement un accès public au chatbot et sans organisation particulière.

2.1.10.1 Le critère du Temps passé

Le critère du **Temps passé** a pour objectif de qualifier le temps nécessaire à la mise en œuvre du scénario par l'utilisateur malveillant. L'intérêt d'un tel critère est de proposer une fourchette de temps nécessaire pour qu'un attaquant atteigne son objectif. Cette proposition d'échelle est un critère particulièrement fragile en ce qu'il est susceptible d'évoluer à mesure que l'usage des SIA se démocratise. En effet, une attaque qui nécessitait il y a trois ans une journée de mise en œuvre ou de préparation, ne nécessite peut-être plus autant de temps aujourd'hui.

De la même manière ce critère est subjectif, les besoins de chaque organisation étant variables d'un secteur d'activité à un autre cette échelle nécessitera certainement d'être adaptée par le lecteur. Une organisation peut disposer d'une coordination et de mesures techniques d'une robustesse pouvant justifier de voir à la hausse le temps de mise en œuvre, de la même manière en fonction de l'exposition du modèle le temps passé peut-être vu à la baisse. Nous proposons en trois niveaux des fourchettes de temps de mise en œuvre :

- **Long (1) :** l'exploitation du scénario d'attaque étudié semble demander une préparation longue et sa mise en exécution peut prendre plusieurs semaines à plusieurs mois.
- **Modéré (2) :** l'exploitation du scénario d'attaque étudié semble nécessiter un temps de préparation et sa mise en exécution peut prendre de plusieurs jours à une semaine.
- **Court (3) :** l'exploitation du scénario d'attaque étudié ne semble pas nécessiter de préparation et sa mise en œuvre ne prend que de quelques heures à une journée.

2.1.10.2 Le critère des Ressources

Le critère des **Ressources** nécessaires est inspiré de la notion de « *source de risque* » (les profils d'attaquants) de la méthode EBIOS Risk Manager [5]. Il s'agit d'une proposition de mesure du niveau de motivation et d'organisation dont doit disposer un utilisateur ou un groupe malveillant pour mettre en œuvre le scénario d'attaque.

Plus le groupe dispose de moyens humains et matériels, plus il est susceptible d'être motivé à compromettre un système. En cybersécurité la source de risque peut être de plusieurs ordres :

- De l'amateur ne disposant pas de moyen autre que son poste de travail ;
- En passant par le groupuscule criminel agissant à des fins pécuniaires ;
- Ou encore les plus préparés les organisations structurées et financées par des États à des fins de déstabilisations politiques.

La diversité des profils est conséquente et est laissée à l'appréciation des organisations pour s'approprier cette échelle de Ressources nécessaires à la mise en œuvre d'une attaque. En effet, il sera probablement plus pertinent pour une TPE/PME de prendre garde à une malveillance interne, aux groupes criminels organisés qu'à un groupe organisé et financé par un État auquel elle ne serait a priori pas exposée.

Il convient cependant de garder à l'esprit que : « *qui peut le plus peut le moins* », des groupements organisés peuvent conduire des attaques d'une certaine simplicité technique et de mise en œuvre. Ainsi, l'échelle des ressources

nécessaires se veut souple et généraliste et elle doit être adaptée au contexte de l'organisation. En voici ci-dessous la déclinaison :

- **Élevées (1) :** la mise en œuvre du scénario étudié **nécessite des capacités matérielles, humaines et financières considérables**. Ce scénario est notamment susceptible d'être exploité par des groupes étatiques ou agences de renseignement se caractérisant par leurs capacités à réaliser des opérations offensives sur un temps long et particulièrement sophistiquées.
- **Moyennes (2) :** la mise en œuvre du scénario étudié **nécessite de disposer de ressources humaines, financières et matérielles**. Ce scénario est notamment susceptible d'être exploité par des groupements organisés (terroristes, criminels ou idéologistes) en capacité de conduire des opérations plus ou moins sophistiquées.
- **Faibles (3) :** la mise en œuvre du scénario étudié ne **nécessite pas de ressources financières et matérielles particulières**. Ce scénario est notamment susceptible d'être exploité par des amateurs ou des groupes activistes de moindre ampleur.

2.1.10.3 Le critère de l'Expertise

Le critère de l'**Expertise** a pour ambition de proposer une contextualisation des attaques sur des systèmes d'intelligence artificielle à une époque où leur compromission n'est pas encore à une échelle systématique et généralisée. Il s'agit ici de considérer, en restant humble, que le public et de facto les attaquants ne sont pas encore tous familiers avec le fonctionnement des IA et de leurs services. Dès lors, nous proposons une échelle de connaissance et de compréhension technique de l'environnement inhérent aux caractéristiques des systèmes d'intelligence artificielle.

Ce critère peut être considéré en mettant en balance les connaissances en cybersécurité et en data science. Nous invitons le lecteur à se saisir de cette échelle et à étudier la situation de son organisation vis-à-vis de ce sujet. En effet, un modèle dont la compréhension technique ne nécessite qu'une dizaine d'heures de formation ne requiert par le même niveau d'attention qu'un LLM dont les paramètres sont administrés par des experts de la discipline. Nous proposons ainsi l'échelle suivante :

- **Élevée (1) :** la mise en œuvre du scénario d'attaque étudié requiert des compétences techniques très avancées ou spécifiques et/ou le développement d'outils ciblés.
- **Moyenne (2) :** la mise en œuvre du scénario d'attaque étudié nécessite la mise en œuvre de techniques simples et/ou d'outils disponibles publiquement.
- **Faible (3) :** la mise en œuvre du scénario d'attaque étudié ne semble nécessiter aucune compétence technique spécifique ni d'outil particulier.

2.1.10.4 Le critère des connaissances sur le système

Le critère des **Connaissances** sur le système propose, contrairement aux précédents critères, de s'intéresser au contexte du système d'intelligence artificielle lui-même. C'est-à-dire le contexte organisationnel et technique dans lequel il se situe. Autrement dit, il s'agit d'évaluer à quel point la connaissance spécifique du système et de ses services dans son environnement est nécessaire pour pouvoir mettre en œuvre le scénario d'attaque étudié.

La finalité ici est de mettre en perspective la complexité plus importante de mise en œuvre du scénario sur un modèle complexe dans un environnement tout aussi riche. Là où un modèle utilisé largement n'aurait peut-être plus aucun secret pour le marché et concomitamment pour les utilisateurs malveillants.

Une fois encore, cette échelle est à remettre dans son contexte, puisqu'il est tout à fait possible d'utiliser un modèle largement démocratisé en suivant des recommandations de sécurité pointues pour en durcir les paramètres. La connaissance du système ne suffit donc plus et son environnement joue tout autant. Pour matérialiser cette analyse, nous proposons l'échelle suivante :

- **Élevée (1) :** le scénario d'attaque étudié est plus difficile à exploiter dans la mesure où l'attaquant doit disposer d'une connaissance complète de l'intégration du modèle dans le système d'intelligence artificielle et de son environnement.
- **Moyenne (2) :** le scénario d'attaque étudié est exploitable moyennant certaines contraintes dans la mesure où l'attaquant doit posséder quelques connaissances sur le système d'information dans lequel s'inscrit le système d'intelligence artificielle. Il est nécessaire de disposer : soit de connaissance du contexte dans lequel il s'inscrit, soit d'autres éléments auxquels il serait interfacé ou bien de connaissance sur les caractéristiques techniques du modèle).
- **Faible (3) :** le scénario d'attaque étudié est plus simple dans sa mise en œuvre dans la mesure où l'attaquant n'a pas besoin de disposer de connaissances spécifiques au modèle objet de l'attaque ou à son environnement.

2.1.10.5 Le critère des Accès

Enfin, le critère des **Accès** est une proposition pragmatique visant à qualifier le besoin de disposer de comptes aux privilèges plus ou moins élevés pour pouvoir utiliser, produire des sorties, administrer ou modifier les paramètres du modèle à des fins malveillantes.

Le scénario deviendra donc plus aisément réalisable si un simple accès utilisateur est nécessaire pour accéder aux modèles et à ces fonctionnalités. La facilité de mise en œuvre sera de même plus évidente si ce même compte utilisateur a accès à des fonctions d'administration normalement limitées à certains profils.

De la même manière, si un compte accessible au public peut produire des actions entraînant des conséquences sur le fonctionnement du modèle ou de ces services, cela peut rendre le scénario d'attaque encore plus simple dans sa mise en œuvre (par exemple dans le cas du prompt injection).

A contrario, un modèle dont les accès sont strictement segmentés par profil d'utilisateur avec une nomenclature des droits dédiée et un nombre limité d'administrateurs augmentera la complexité de mise en œuvre du scénario d'attaque.

L'échelle proposée est à adapter au contexte dans lequel se trouve le modèle concerné, suivant qu'il soit librement accessible au public ou soumis à la création d'un compte les répercutions et les besoins de sécurisation ne seront pas les mêmes.

L'échelle des Critères d'accès nécessaires se matérialise ainsi comme il suit :

- **Utilisateur interne à haut privilège (1) :** la mise en œuvre du scénario d'attaque étudié requiert des droits élevés, tels que des droits d'administration.
- **Utilisateur interne (2) :** la mise en œuvre du scénario d'attaque requiert d'être un **utilisateur interne** et authentifié de l'organisation.
- **Grand public (3) :** la mise en œuvre du scénario d'attaque ne requiert aucun droit d'accès spécifique (par exemple : si le système d'intelligence artificielle est accessible au grand public).

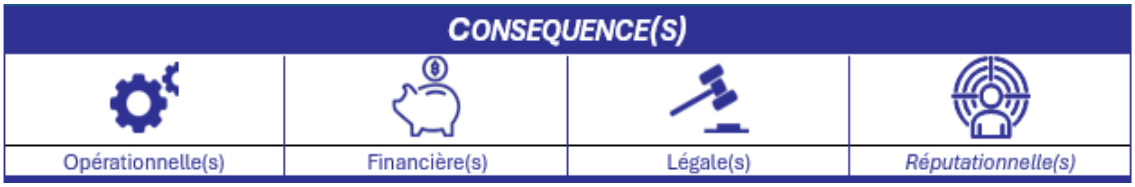
2.1.11 Les conséquences d'une attaque sur l'organisation

Avant de poursuivre, il convient de préciser que les éléments précédents avaient pour finalité de qualifier les impacts plus ou moins directs d'une attaque sur un système d'intelligence artificielle. Ces propositions se concentrent donc sur un sujet précis de la sécurité à l'échelle d'une organisation, en l'occurrence : la sécurité des systèmes d'IA et les impacts des attaques sur leurs composants et services. Les évènements en question seront, dès lors, la plupart du temps classés comme étant des « impacts opérationnels ».

Mais toute attaque sur un système d'information ou une de ses composantes, n'a pas pour seule conséquence de perturber les opérations. Au contraire, une attaque peut avoir des impacts collatéraux à une échelle plus stratégique. Notamment si des secrets d'entreprise sont exposés, des données personnelles clients ou collaborateurs sont compromises dans leur confidentialité ou leur intégrité, ou si l'évènement entraîne des conséquences sur les résultats financiers d'une activité.

Pour ne pas omettre ces aspects stratégiques pouvant résulter d'une compromission d'un système d'IA, nous proposons une section identifiant succinctement les conséquences d'une attaque sur un modèle d'IA pour une organisation. À l'instar de la méthode EBIOS RM qui propose des catégories

d'impact nous proposons sommairement quatre catégories de conséquences stratégiques pour une organisation.



Opérationnelle(s)	Les conséquences dites « *opérationnelles* » qualifient la dégradation ou l'incapacité des activités et/ou services d'une organisation à fonctionner, du fait d'une dépendance plus ou moins forte aux services fournis par le système d'IA visé par l'attaque. L'icône restera en bleu lorsqu'une attaque sera évaluée comme entraînant des conséquences opérationnelles sur l'organisation concernée.
Financière(s)	Les conséquences « *financières* » qualifient la possibilité que le scénario d'attaque entraîne des conséquences pécuniaires directes ou indirectes pour l'organisation. L'icône restera en bleu lorsqu'une attaque sera estimée comme pouvant entraîner ces conséquences.
Légale(s)	Une organisation est sujette à différents textes et obligations de natures juridiques. Dès lors, en cas de compromission d'un système d'IA, cette dernière peut faire l'objet de poursuites et sanctions prévues par le droit. L'icône restera en bleu lorsqu'une attaque sera évaluée comme pouvant entraîner des poursuites, amendes et/ou condamnation du fait d'une non-conformité à des obligations légales. Ces dernières peuvent par exemple être le fruit d'une absence de sécurisation du système d'IA ou de l'un de ses services, ou de la compromission des données qu'il traite.
Réputationnelle(s)	La cybersécurité est aujourd'hui un sujet pouvant déterminer le niveau de confiance que des parties intéressées peuvent avoir pour une organisation. Une compromission d'un système d'IA qui viendrait à être

	rendue publique peut entraîner des conséquences sur l'image de l'organisation et dès lors nuire à ses activités. L'icône restera en bleu lorsqu'une attaque sera susceptible de compromettre l'image d'une organisation.

Taxonomie des attaques

Pour faciliter la compréhension et la gestion des risques liés à la sécurité des systèmes d'intelligence artificielle (IA), nous avons développé une taxonomie des attaques. Cette taxonomie vise à fournir un cadre structuré et complet pour identifier, classer et analyser les différentes menaces qui pèsent sur ces systèmes. La taxonomie s'appuie sur les référentiels que nous avons décrits plus haut :

- NIST.AI.100-2e2023 [7] ;
- MITRE ATLAS [17] ;
- OWASP Top 10 LLM [10] ;
- OWASP Top 10 ML [11].

Comme nous l'avons vu précédemment, les différents référentiels apportent des informations différentes, qu'il nous a semblé utile de regrouper dans une taxonomie unique. Nous avons au passage constaté que l'évolution très rapide des technologies d'IA fait constamment apparaître de nouvelles attaques potentielles. C'est pourquoi, nous serons sans doute amenés à faire évoluer cette taxonomie au fur et à mesure de l'arrivée de nouveaux systèmes IA (par exemple l'agentique).

La taxonomie est organisée en quatre niveaux hiérarchiques, offrant une approche granulaire et pratique pour appréhender les attaques :

1. **Phases du cycle de vie :** ce premier niveau utilise le cycle de vie d'un projet IA comme axe principal de classification. Nous avons retenu le modèle de l'ANSSI [1] (en haut) puis celui de l'OCDE[5], qui décompose le développement d'un système IA en sept phases distinctes précédemment détaillées (voir la section 2.1.3). Ce choix permet d'associer chaque attaque à une phase spécifique du cycle de vie, facilitant ainsi l'identification des risques pertinents à chaque étape d'un projet. Pour une vision plus globale, nous avons également intégré les trois phases du cycle de vie de l'ANSSI (Entraînement, Déploiement et Production), en les superposant au modèle de l'OCDE, comme nous l'avons décrit précédemment.
2. **Famille d'attaques :** le second niveau regroupe les attaques partageant des caractéristiques communes, telles que des mécanismes d'attaque similaires, des objectifs communs ou des impacts comparables. Exemples de familles d'attaques : empoisonnement de données, évasion, extraction de modèle, etc. Ce regroupement permet de mieux comprendre les différentes catégories de menaces et de développer des stratégies de défense plus générales.

3. **Attaques spécifiques :** le troisième niveau décrit chaque attaque en détail. Chaque attaque est documentée avec une description précise de son fonctionnement, de ses conséquences potentielles, des techniques de détection et des mesures de mitigation. Ce niveau de détail offre aux experts en IA et aux experts en cybersécurité les informations nécessaires pour comprendre et contrer les menaces spécifiques.

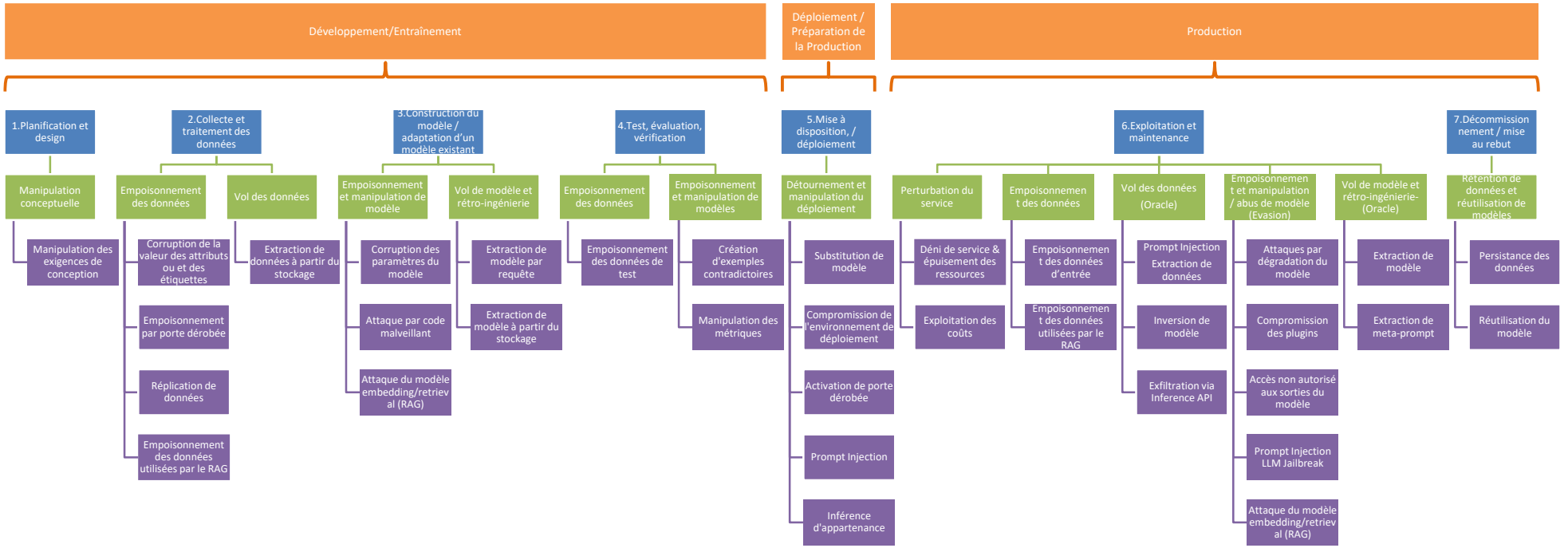


Figure 11 – Taxonomie des attaques sur l'IA

Grandes catégories d'attaques

Nous présentons ici une typologie simplifiée des attaques, axée sur les catégories d'empoisonnement, d'évasion et d'oracle.

2.1.12 Attaques par Empoisonnement (Poisoning)

Ces attaques ciblent la phase d'apprentissage du modèle, en altérant les données d'entraînement ou le modèle pour compromettre son intégrité :

- **Empoisonnement des données** : introduction de données malveillantes dans l'ensemble d'entraînement.

Analogie : Corrompre un manuel scolaire pour que les élèves apprennent de mauvaises réponses.

Exemple : Injecter des transactions frauduleuses dans les données de référence d'un modèle de détection de fraude.

- **Évasion classique** : perturbation des données d'entrée pour induire une classification incorrecte.

Analogie : Modifier légèrement une image pour qu'elle soit mal reconnue.

Exemple : Modifier une image d'un panneau stop pour tromper une voiture autonome.

- **Empoisonnement du modèle** (plus spécifique aux modèles distribués et collaboratifs) : modification directe des paramètres du modèle pendant l'entraînement.

Analogie : Modifier le code source d'un programme pour qu'il se comporte différemment.

Exemple : Un participant malveillant à un entraînement fédéré (voir section 0) envoie des mises à jour de modèle corrompues (il transmet des paramètres erronés).

- **Attaques par la chaîne d'approvisionnement** : compromission des composants du modèle avant leur utilisation.

Analogie : Recevoir un logiciel espion caché dans un programme légitime.

Exemple : Utilisation d'une bibliothèque logicielle compromise ou d'un modèle pré-entraîné contenant une porte dérobée.

2.1.13 Attaques par Évasion (Evasion)

Ces attaques ciblent le modèle en production, en modifiant les données d'entrée pour obtenir des prédictions erronées :

- **Injection de prompts** (spécifique aux LLM) : manipulation de l'interface textuelle des LLM pour contourner les restrictions et obtenir des réponses non désirées.

Analogie : Poser des questions pièges à un assistant vocal.

Exemple : Demander à un chatbot de générer du contenu malveillant.

2.1.14 Attaques Oracles (Oracle Attacks)

Ces attaques exploitent l'accès au modèle pour en extraire des informations ou influencer son comportement :

- **Attaques par inférence** : déduire des informations sur les données d'entraînement ou le modèle à partir de ses prédictions.

Analogie : Deviner les questions d'un examen en analysant les réponses.

Exemples : Inférence d'appartenance (déterminer si une donnée était présente dans l'ensemble d'entraînement) ou extraction de modèle (reproduire un modèle concurrent).

- **Attaques par extraction de données** (plus critique pour les LLM) : obtenir des informations sensibles du modèle, souvent via des prompts soigneusement construits.

Exemple : Extraire des numéros de carte de crédit mémorisés par un chatbot.

- **Consommation excessive de ressources** (plus critique pour les LLM) : surcharger le modèle avec des requêtes pour dégrader le service ou épuiser les ressources.

Analogie : Surcharger un serveur web avec des requêtes pour le rendre inaccessible.

2.1.15 En conclusion

Cette classification simplifiée met en lumière les principales catégories d'attaques contre les systèmes d'IA. En tant qu'expert en IA ou chef de projet, il est crucial de comprendre ces menaces pour développer des modèles robustes et sécurisés. Les différentes attaques seront détaillées dans les parties suivantes.

3 Autres techniques à suivre

Nous décrivons ici quelques techniques qui peuvent amener de nouveaux types de défenses (comme le chiffrement dans 0 Cryptographie) ou d'attaques dont certaines sont incluses dans notre taxonomie (0 RAG et 0 Attaques adverses) et d'autres pas encore (0 Agentique, 0 Apprentissage fédéré).

RAG

Limitations auxquelles on veut faire face

L'intelligence artificielle (IA) générative excelle dans la création de réponses textuelles basées sur de grands modèles de langage, où l'IA est entraînée sur un

grand nombre de données. La bonne nouvelle est que le texte généré est souvent facile à lire et fournit des réponses détaillées.

La mauvaise nouvelle est que les informations utilisées pour générer la réponse sont limitées aux informations utilisées pour entraîner l'IA, souvent un LLM. Les données du LLM peuvent être périmées depuis des semaines, des mois ou des années, sans capacité simple de les mettre à jour.

De plus, dans un chatbot d'IA d'entreprise, elles peuvent ne pas prendre en compte des informations spécifiques aux produits ou services de l'entreprise.

Cela peut conduire à des réponses incorrectes qui érodent la confiance en la technologie de certains clients et collaborateurs.

Naissance des RAG - Génération Augmentée de Récupération (*Retrieval Augmented Generation*)

C'est là qu'intervient la RAG : elle fournit un moyen d'optimiser le résultat d'un LLM avec des informations ciblées, sans modifier le modèle sous-jacent lui-même. Le RAG permet d'utiliser les données de l'entreprise (ex : base documentaire des produits), en tant que source de données pour le LLM.

C'est donc une technique qui améliore la qualité de l'IA générative en permettant aux grands modèles de langage (LLM) d'exploiter des ressources de données supplémentaires sans réentraînement (qui, rappelons-le, est très coûteux du fait des volumes de données impliquées et de la taille des modèles).

Comment fonctionne une RAG

Corpus : la première étape consiste à rassembler les informations ciblées, les ressources de données supplémentaires que l'on souhaite mettre à disposition du LLM inclus dans notre système IA. Elles forment notre corpus documentaire ou base de connaissances.

Ces données sont ensuite traitées afin de devenir exploitables par notre RAG :

- Découpage (*chunks*) : les documents du corpus sont découpés en courts passages.

  Certains de ces passages seront fournis en entrée du LLM pour aider à la génération d'une réponse appropriée (c'est le *contexte* du prompt). Ils ne peuvent pas être trop conséquents puisque les entrées fournies aux LLMs ne peuvent pas dépasser une certaine quantité, déterminée par le contexte d'un LLM.

  La fenêtre de contexte d'un LLM peut être considérée comme l'équivalent de sa mémoire de travail. Elle détermine la durée d'une conversation qu'il peut mener sans oublier les détails de l'échange précédent. Elle détermine également la taille maximale des documents qu'il peut traiter en même temps.

- Représentation numérique (*embeddings*) : le contenu sémantique de chacun des passages est converti sous forme de vecteurs.

  Cette représentation numérique permet de conserver le sens des mots, puisque par exemple des mots au sens proche seront transformés en vecteurs avec des caractéristiques communes, ayant une distance vectorielle faible.

- Base vectorielle (*vector store*) : ces vecteurs sémantiques sont stockés dans une base de données spécialement conçue pour les calculs vectoriels, qui sera interrogée en complément du prompt de l'utilisateur.

  Lorsqu'un utilisateur interroge l'IA, la RAG entre en jeu pour fournir au service d'IA des compléments d'informations qui permettra au LLM sous-jacent de répondre en se basant sur les informations du corpus documentaire :

- Représentation numérique (*embeddings*) : la question de l'utilisateur est convertie en vecteurs sémantiques, avec la même méthode que celle utilisée précédemment pour créer la base vectorielle du corpus.
- Recherche de similarité : le module de recherche utilise des mesures de similarité pour comparer les vecteurs de la question aux vecteurs des documents de la base de données. Les vecteurs correspondants aux passages les plus « proches » de la question sont sélectionnés pour la génération de la réponse.

  Une fois sélectionnés, ces vecteurs sont reconvertis en texte naturel, c'est-à-dire aux passages correspondants de documents du corpus initial.

- LLM : le LLM utilise la question et les extraits récupérés par la recherche précédente pour générer une réponse pertinente.

Le schéma ci-dessous illustre toutes ces étapes :

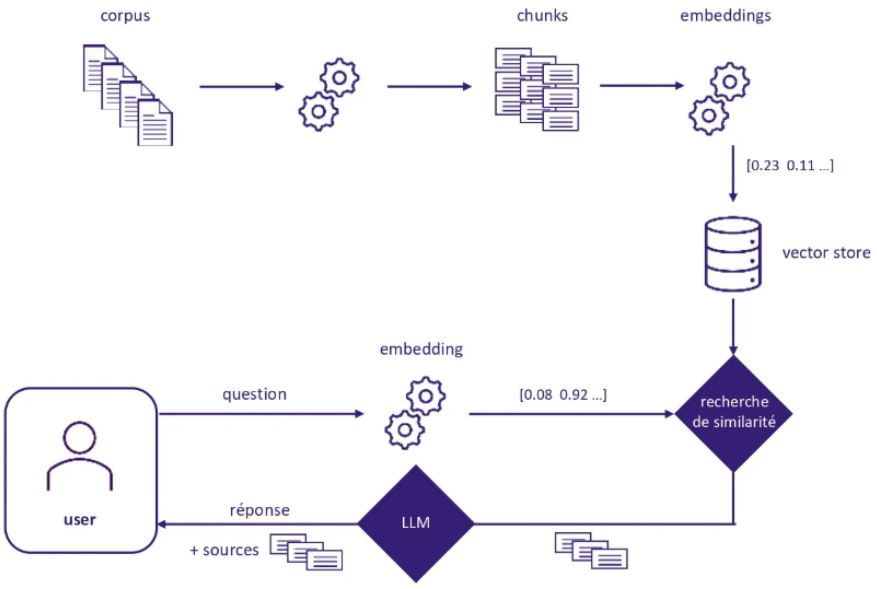


Figure 12 – Fonctionnement de la RAG

Avantages d'utiliser une RAG

- Si le processus de formation du LLM est long et coûteux, c'est tout à fait l'inverse pour les mises à jour du RAG. De nouvelles données peuvent être chargées et traduites en vecteurs de manière continue et incrémentielle.
- La RAG présente aussi l'avantage d'utiliser une base de données vectorielle, ce qui permet au service d'IA de fournir la source spécifique des données citées dans sa réponse, ce que les LLMs ne peuvent pas faire. Par conséquent, s'il y a une inexactitude dans le résultat de l'IA, le document qui contient cette information erronée peut être rapidement identifié et corrigé, puis l'information corrigée peut être introduite dans la base de données vectorielle.

Attaques spécifiques aux RAG

- Les systèmes RAG accèdent souvent à de grandes bases de données, ce qui soulève des préoccupations concernant la sécurité des données et la confidentialité. Il est crucial de protéger les informations sensibles tout en maintenant la fonctionnalité du système, ce qui nécessite un équilibre délicat.
- De la même façon, chaque manipulation de ces données est un point d'entrée potentiel pour les attaquants : la représentation numérique, la recherche dans la base de données vectorielles, la transmission des données sélectionnées vers le LLM, et enfin l'interprétation des données sélectionnées.
- Le modèle LLM inclus dans le service d'IA quant à lui est vulnérable aux attaques classiques sur les systèmes d'IA.

Système agentique

Pourquoi développer des systèmes agentiques ?

Les systèmes agentiques représentent une évolution significative dans le domaine de l'intelligence artificielle ; Contrairement aux modèles de langage traditionnels, qui génèrent des réponses en fonction d'une requête précise, les systèmes agentiques prennent des décisions de manière autonome et interagissent activement avec leur environnement.

Définition

Un **agent** autonome est capable, comme le montre la Figure 13, de :

- Interagir avec son environnement.
- Prendre des décisions indépendantes.

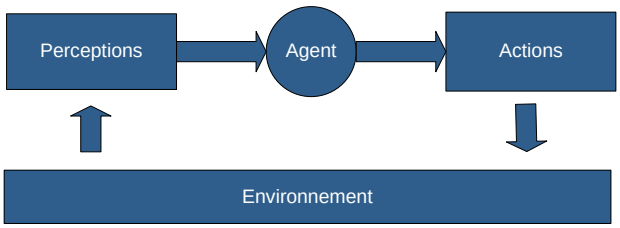


Figure 13 - Schéma de principe d'un agent

Un **système agentique** est une architecture d'intelligence artificielle composée d'agents capables de collaborer entre eux pour atteindre des objectifs complexes. Ces agents sont conçus pour fonctionner avec un certain degré d'indépendance, leur permettant de s'adapter dynamiquement aux évolutions de leur environnement et d'optimiser leur prise de décision en continu.

Caractéristiques clés

- **Autonomie et prise de décision** : les agents d'un système agentique peuvent agir de manière indépendante, en s'appuyant sur leur perception de l'environnement. Contrairement aux IA réactives, ils n'ont pas besoin d'une supervision constante et peuvent initier des actions en fonction des situations rencontrées.
- **Interconnexion et collaboration** : les agents sont capables de communiquer et d'échanger des informations avec d'autres agents ou systèmes et leur environnement. Cette capacité d'apprentissage et d'évolution leur permet d'adapter leur comportement face à des contextes dynamiques et imprévisibles. Ces comportements émergents peuvent être imprévus et plus complexes que les comportements individuels des agents.
- **Architecture distribuée** : plutôt que de reposer sur un seul agent puissant, les systèmes agentiques peuvent adopter une approche distribuée, répartissant les compétences et les responsabilités entre plusieurs entités. La question de l'orchestration des différents agents est très importante et délicate à régler.

Fonctionnement

Un système agentique peut comporter plusieurs composants essentiels :

1. **Objectif et Planification** : l'agent reçoit un objectif global, qu'il décompose en sous-tâches pour atteindre le résultat attendu. Il peut ajuster ses plans en fonction des événements qu'il rencontre.
2. **Mémoire et apprentissage** : un agent peut retenir des informations sur ses interactions passées, soit temporairement (mémoire contextuelle), soit sur le long terme (stockage persistant). Cette mémoire lui permet d'adapter son comportement et d'optimiser ses actions au fil du temps.
3. **Connexion avec d'autres systèmes** : les agents peuvent se connecter à des API, interroger des bases de données et interagir avec d'autres systèmes. Ces interactions leur permettent d'accéder à des informations en temps réel et de prendre des décisions plus éclairées.
4. **Boucle de rétroaction et d'amélioration** : un agent analyse l'impact de ses actions et ajuste son comportement pour optimiser ses décisions futures. Ce mécanisme d'apprentissage par retour d'expérience améliore la performance de l'agent, mais peut aussi ouvrir des failles de sécurité si un attaquant manipule les données d'apprentissage.

La Figure 14 montre un exemple de système agentique.

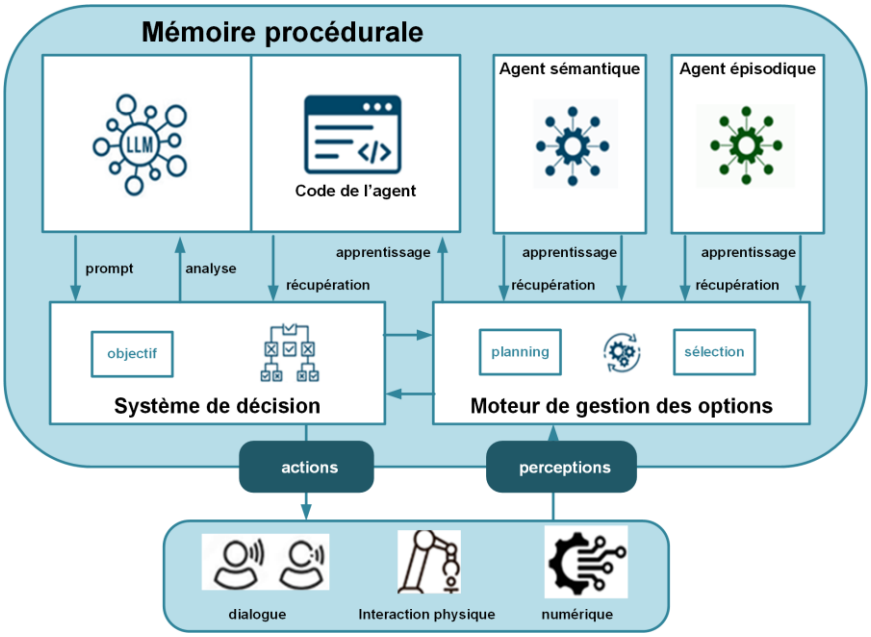


Figure 14 – Exemple de système agentique

Lien avec la RAG : certains agents IA intègrent la RAG pour accéder en temps réel à des informations précises et à jour depuis un corpus documentaire, améliorant ainsi leur capacité à fournir des réponses pertinentes et actualisées.

Applications

Les systèmes agentiques trouvent des applications dans divers domaines, notamment :

- Gestion de chaînes d'approvisionnement et optimisation logistique : planification dynamique et anticipation des ruptures de stock.
- Assistance personnalisée en santé : agents médicaux pour le suivi des patients et l'analyse des diagnostics.
- Développement logiciel et gestion de projets : automatisation des tâches répétitives et coordination intelligente des équipes.
- Analyse financière et prise de décision : identification des tendances de marché et exécution automatique d'ordres de trading.
- Recherche scientifique et innovation : exploration autonome de bases de données et génération de nouvelles hypothèses.
- Notons qu'on trouve aussi beaucoup de systèmes agentiques en robotique coopérative et dans les jeux vidéo.

Quelles sont les attaques spécifiques aux systèmes agentiques ?

Les systèmes agentiques exposent une nouvelle surface d'attaque, car ils combinent prise de décision autonome et interactions avec d'autres systèmes. Ils doivent être robustes face aux défaillances individuelles des agents. Si un agent échoue ou se comporte de manière inattendue, cela peut perturber l'ensemble du système. En général, on modélise le comportement des agents et leurs

interactions. Si une attaque vise la gestion du comportement et de la complexité du système agentique, cela peut compromettre la robustesse du modèle. Le document [8] récemment publié par OWASP traite spécifiquement des attaques sur les systèmes agentiques.

- **Injection d'instructions malveillantes** : un attaquant peut détourner l'agent en insérant des commandes déguisées dans des entrées légitimes, lui faisant exécuter des actions non prévues.
- **Empoisonnement de la mémoire** : si un agent mémorise ses interactions, un attaquant peut insérer de fausses informations dans son historique pour influencer ses décisions futures.
- **Escalade de privilèges via les API** : un agent qui accède à des services externes peut être manipulé pour obtenir des privilèges plus élevés et accéder à des ressources critiques. La problématique de la gestion des droits d'une IA utilisant une autre IA via un agent devra être approfondie.
- **Détournement d'objectifs (*Goal Hijacking*)** : si un attaquant modifie les paramètres d'un agent ou manipule son système de récompense et d'apprentissage, il peut détourner l'agent de sa mission initiale avec parfois pour but d'introduire des objectifs conflictuels. Ce type d'attaque se base sur des modifications progressives, difficiles à détecter immédiatement, qui altèrent peu à peu le comportement de l'agent.

Apprentissage fédéré

Le but de l'apprentissage fédéré est de permettre à plusieurs clients (par exemple des individus, instituts ou entreprises) d'entraîner un modèle de façon collaborative mais sans jamais partager leurs données : au contraire, les clients vont partager uniquement le modèle.

Ce processus est coordonné par un serveur central (par exemple un fournisseur de services) et nécessite plusieurs tours d'apprentissage pour réaliser l'entraînement du modèle. Ainsi, à chaque tour, le serveur transmet le modèle courant aux clients qui vont alors mettre à jour ses paramètres en l'entrainant indépendamment à partir de leurs propres données. Seuls les paramètres du modèle mis à jour localement sont retournés au serveur qui va alors les agréger en effectuant une moyenne pondérée (par la taille des jeux de données locaux) et ainsi actualiser le modèle fédéré.

Avantages de l'apprentissage fédéré : comparativement à un apprentissage traditionnel centralisé qui consiste à collecter le maximum de données d'entrainement pour ensuite les traiter dans un data center, l'apprentissage fédéré permet à la fois de diminuer les besoins en bande passante et d'améliorer la confidentialité des données. L'apprentissage fédéré présente ainsi un réel intérêt pour les applications avec des données clients sensibles ou trop volumineuses pour être centralisées.

Attaques spécifiques à l'apprentissage fédéré : toutefois, en ouvrant sa phase d'apprentissage à de nombreux acteurs, ce processus va faciliter la mise en œuvre d'attaques liées à l'intégrité du modèle et/ou s'exposer à de nouvelles attaques ciblant la confidentialité des données clients. Dans la suite, nous décrirons les attaques survenant durant la phase d'apprentissage du modèle fédéré mais il faut garder à l'esprit que le modèle fédéré, une fois appris, s'exposera aux mêmes risques d'attaques qu'un modèle construit de façon centralisée lors de ses phases de déploiement et production.

Attaques d'intégrité : contrairement à un apprentissage centralisé où les données peuvent être inspectées, l'orchestrateur d'un apprentissage fédéré n'a aucun moyen de vérifier que les paramètres transmis par un client correspondent bien à un apprentissage licite. Il est ainsi très aisé pour un client malveillant d'empoisonner ses données ou le modèle fédéré afin de dégrader ses performances ou son comportement. L'empoisonnement des données sera effectué comme en centralisé par porte dérobée ou en modifiant les attributs/étiquettes de la base de données locale.

Attaques de confidentialité : par construction, l'apprentissage fédéré protège les données clients stockées localement en agrégeant les mises à jour du modèle plutôt que les données brutes. Il s'agit d'une solution pour la confidentialité des données mais pas pour la confidentialité du modèle. Et même si les paramètres du modèle contiennent beaucoup moins d'informations à propos des données des clients que les données brutes, il est tout à fait envisageable d'inférer des informations sur les données des clients.

Sécurité des systèmes d'IA par la cryptographie

Les systèmes d'intelligence artificielle ont une très large surface d'attaque. Ils présentent les mêmes risques que toute application informatique, mais également des risques spécifiques liés à l'IA, comme les attaques par injection de contenu. Un système d'IA manipule de grandes quantités de données, qui se trouvent sous différentes formes : dans les modèles, les requêtes des utilisateurs et parfois dans des bases de données de connaissance, souvent sous forme vectorielle. Les données sont peu, ou pas, structurées ce qui augmente leur potentialité de fuite d'information.

Les attaquants externes qui cherchent à récupérer des données peuvent être de différentes natures :

- Des tiers extérieurs au système qui volent des données stockées (modèles, données en base) ou lors de leur transit ;
- Des opérateurs de l'infrastructure (un hébergeur, par exemple) qui mènent des attaques actives sur les données en transit ou en mémoire ;

- Des utilisateurs malveillants qui exploitent des vulnérabilités des systèmes d'autorisation pour accéder à des données auxquelles ils n'ont pas accès.

Le chiffrement est la protection des données la plus efficace contre les attaques externes. Sa mise en œuvre dépend du contexte d'utilisation du système et de la nature des attaquants contre lesquels il faut se protéger.

Lorsque le système est opéré *on-premise*, dans un environnement contrôlé, le ***chiffrement des données au repos*** est suffisant. Correctement mis en œuvre, avec des clés hébergées dans des systèmes externes au système d'IA, de type KMS/HSM (cf. glossaire), il protège contre un vol de disque ou de backups. Le chiffrement de disque a l'avantage d'être extrêmement simple à activer et de ne pas affecter les performances du système. On pourra donc préconiser le test de performance du système IA dès lors que l'on active le chiffrement.

Afin de protéger un système fonctionnant dans le cloud contre les attaques actives sur la mémoire, la machine et le réseau, l'utilisation de ***machines de confidential computing*** est la solution industrielle en cours d'adoption. Cette technologie utilise des CPUs et GPUs qui hébergent un secret dans leur silicium pour chiffrer et déchiffrer la mémoire, et limite les pénalités de performance à environ 5% pour les machines virtuelles confidentielles. Le chiffrement de disque avec des secrets hébergés dans le TPM (*Trusted Platform Module*, ou vTPM (*Virtual Trusted Platform Module* le cas échéant) renforce cette protection, garantissant qu'un attaquant voit uniquement des données chiffrées en mémoire et sur le disque. De plus, les interactions utilisateurs sont effectuées sur des connexions TLS (*Transport Layer Security*) se terminant dans la mémoire chiffrée de la machine, garantissant un chiffrement de bout en bout qui protège les interactions réseau.

L'intégrité du système est assurée par la ***vérifiabilité***, qui consiste à collecter des empreintes cryptographiques du matériel, du système d'exploitation, des logiciels et des modèles de la machine. Ces empreintes peuvent être vérifiées de l'extérieur à tout moment pour garantir à l'utilisateur que son système n'a pas été altéré.

La sécurité peut encore être renforcée, en ***chiffrant du côté client*** les modèles et les données avant de les envoyer dans le cloud. En utilisation, ils seront déchiffrés, mais au sein de la mémoire chiffrée de la machine. Le chiffrement côté client garantit un meilleur contrôle des clés et des algorithmes cryptographiques choisis, ouvrant la possibilité d'un chiffrement plus élaboré comme Covercrypt[13], qui est post-quantique et qui permet du contrôle d'accès dans les données chiffrées.

[13] https://eprint.iacr.org/2023/836

Des ***systèmes purement cryptographiques***, n'impliquant pas de matériel confidentiel spécialisé, sont en cours de développement. Comme ils sont exclusivement logiciels, leur surface d'attaque est réduite et leur déploiement plus universel. À très court terme, des bases vectorielles entièrement chiffrées vont voir le jour. À moyen terme, le chiffrement totalement homomorphe[14] permettra d'opérer les calculs directement sur les chiffrés.

Il est cependant à recommander d'effectuer systématiquement des ***tests de performance*** dès que l'on met en place des contremesures avec chiffrement.

La synthèse de tous ces précédents éléments est donc la suivante :

Contexte	**Solution**	**Impact performance**	**Exemple de technologie**
• on-premise • environnement contrôlé	Chiffrement des données au repos	Aucun impact	
	Clés hébergées dans des systèmes externes		KMS / HSM
Cloud côté serveur	Machines de confidential computing	5% de pénalité pour les machines virtuelles confidentielles	TPM ou vTPM
	Connexions TLS		
	Vérifiabilité		
• **Cloud** avec de la sécurité côté client	Solutions cloud précédentes	Impacts cloud précédents	Technologies cloud précédentes
	Chiffrement pendant le transfert vers le cloud	Impact minime	
Indépendance du contexte	Systèmes purement cryptographiques		• Technologies en cours de

[14] https://fr.wikipedia.org/wiki/Chiffrement_homomorphe

			développement : Bases vectorielles chiffrées • Chiffrement homomorphe

3.1.1 Les techniques cryptographiques

Chiffrement authentifié : le chiffrement permet d'assurer la confidentialité des données, mais également leur intégrité (authenticité). Une modification des données chiffrées par un attaquant, ou l'absence de fourniture de données d'authentification supplémentaires, engendrera une erreur au moment du chiffrement. Le chiffrement authentifié standardisé le plus utilisé est AES GCM[15] (GCM - *Galois Counter Mode* qui fournit le tag d'authentification). La taille d'un chiffré AES GCM est égal à la taille du clair + 28 octets (12 pour le nonce[16], 16 pour le tag). AES XTS[17], généralement utilisé pour chiffrer les disques, n'est pas authentifié ; il fournit pour seule garantie que si un chiffré a été modifié, les données déchiffrées seront illisibles.

Chiffrement de disque : le chiffrement de disque est effectué par le système d'exploitation qui chiffre ou déchiffre les données à la volée en écrivant ou en relisant des disques. Il présente les grands avantages d'être transparent pour les applications et les utilisateurs, et d'être extrêmement efficace. En revanche, il protège uniquement contre "l'arrachage" du disque : une fois la machine démarrée, la totalité des données est accessible par un utilisateur authentifié sur le système ou sur l'application les utilisant. Les systèmes de chiffrement de disque sont LUKS[18] sur Linux, BitLocker sous Windows ou FileVault sur macOS. Le code de BitLocker n'étant pas open source, des alternatives existent comme VeraCrypt[19], dont le code est libre, ou CRYHOD[20] de Prim'x qualifié par l'ANSSI. Le chiffrement de disque utilise en règle générale AES XTS (cf. ci-dessus), la clé AES étant elle-même encapsulée dans une autre clé. Cette autre clé, appelée la KEK (*Key Encryption Key*) doit être, a minima, stockée dans le TPM de la machine, ou mieux, dans un KMS externe.

[15] https://csrc.nist.gov/pubs/sp/800/38/d/final
[16] https://fr.wikipedia.org/wiki/Nonce_(cryptographie)
[17] https://csrc.nist.gov/pubs/sp/800/38/e/final
[18] https://github.com/libyal/libluksde/blob/main/documentation/Linux%20Unified%20Key%20Setup%20(LUKS)%20Disk%20Encryption%20format.asciidoc
[19] https://www.veracrypt.fr/code/VeraCrypt/
[20] https://www.primx.eu/en/encryption-software/cryhod-en/

VM Confidentielle : utilisation de machines virtuelles dont la mémoire et les disques sont chiffrés. Le chiffrement de la mémoire est effectué à l'aide d'un secret non extractible caché dans le CPU (et éventuellement GPU) de la machine ; le disque est chiffré à l'aide d'un secret hébergé dans un vTPM ou mieux un KMS (cf. ci-dessus). Ces VM confidentielles permettent d'opérer en toute confidentialité sur la machine d'un autre, typiquement celle d'un hébergeur, avec une haute performance : environ 5 % de pénalité par rapport à une VM standard. Des distributions Linux durcies prêtes à l'emploi, telles que Cosmian[21] VM, sont disponibles chez les principaux hébergeurs.

Vérifiabilité : les VM confidentielles apportent la confidentialité par le chiffrement, mais ne garantissent pas l'intégrité du système ; un composant hardware a pu être modifié par l'hébergeur, le système d'exploitation redémarré avec un module fuitant les données, un binaire ou un modèle remplacé par une version compromise. La vérifiabilité ajoute un service permettant de récupérer des empreintes cryptographiques de la totalité d'un système audité, puis de pouvoir les vérifier à tout moment sur un système en exécution. La vérification hardware est fournie par défaut sur les CPUs et GPUs confidentiels, la vérification intégrale du système est fournie par des agents tels que ceux disponibles dans les Cosmian VM.

Chiffrement avec contrôle d'accès : ce type de chiffrement permet de mettre en œuvre le *Data Centric Security*. Les données sont chiffrées avec des attributs et seuls les utilisateurs pouvant présenter des clés avec des politiques d'accès sur ces attributs, peuvent déchiffrer les données. Ce type de chiffrement aide à se protéger d'une classe d'attaque courante, celle de la compromission des autorisations applicatives, telle que les escalades de privilèges. Un exemple de ce type de chiffrement est Covercrypt, récemment standardisé par l'ETSI[22].

Chiffrement post-quantique : ce type de chiffrement permet de se protéger contre les nouvelles attaques disponibles sur les ordinateurs quantiques (algorithmes de Shor et Grover par exemple). Le but est ici de se protéger contre une attaque future, pour des données à durée de vie longue, qui pourraient être collectées, chiffrées dès aujourd'hui, puis déchiffrées demain, lorsque les ordinateurs quantiques seront largement disponibles. Du côté du chiffrement symétrique, la parade est assez simple : doubler la taille des clés, à 256 bits pour AES, par exemple, ce qui ralentit le chiffrement, mais n'augmente pas la taille des chiffrés. Du côté du chiffrement à clé publique, la situation est plus complexe. Le NIST (National Institute of Standards and Technology américain) a choisi un algorithme, Crystals Kyber et l'a

[21] https://docs.cosmian.com/cosmian_vm/overview/
[22] https://www.etsi.org/technologies/quantum-safe-cryptography

standardisé sous le nom de ML-KEM[23]. La réglementation américaine impose d'avoir basculé l'intégralité du chiffrement à clé publique en post-quantique avant 2035. En Europe, pas de dates à ce jour et il est recommandé de ne pas utiliser cet algorithme directement, mais de l'hybrider avec un algorithme classique, utilisant une courbe elliptique. C'est ce que fait Covercrypt, standardisé par l'ETSI. Le chiffrement post-quantique, même hybridé, est performant ; il s'effectue en quelques centaines de microsecondes en moyenne.

3.1.2 Risques adressés par la cryptographie

Phase du cycle de vie	Famille d'attaques	Attaques spécifiques	Solution
Collecte et Traitement de données	Empoisonnement des données	Empoisonnement par porte dérobée	• Chiffrement authentifié • Vérifiabilité
		Réplication de données	Chiffrement
	Vol des données	Extraction des données à partir du stockage	Chiffrement
Construction du modèle	Empoisonnement et manipulation	Corruption des paramètres	• Chiffrement authentifié • Vérifiabilité
		Attaque par code malveillant	Vérifiabilité
	Vol de modèle	Extraction à partir du stockage	Chiffrement
Mise à disposition / déploiement	Détournement et manipulation	Substitution de modèle	• Chiffrement authentifié • Vérifiabilité
		Compromission de l'environnement	• Chiffrement authentifié • Vérifiabilité

[23] https://csrc.nist.gov/pubs/fips/203/final

		Activation de porte dérobée	• Chiffrement mémoire et réseau • Vérifiabilité
Exploitation et maintenance	Empoisonnement et manipulation	Attaques par dégradation	• Chiffrement authentifié • Vérifiabilité
		Compromission des *plugins* (ou greffons)	Vérifiabilité
		Accès non autorisé	• Chiffrement avec contrôle d'accès • Vérifiabilité
	Vol de modèle	Extraction de modèle	Chiffrement
Extraction de meta-prompt			Chiffrement stockage, réseau
Décommissionnement	Rétention de données	Persistance des données	Chiffrement post-quantique
		Réutilisation du modèle	Chiffrement avec contrôle d'accès

Attaques adverses

Une **attaque adverse** (ou adversarielle, en anglais *adversarial attack*), est une opération dans laquelle un « attaquant » modifie l'entrée d'un système d'IA pour lui faire produire une sortie différente de celle qu'aurait donné le système d'IA attaqué s'il avait reçu l'entrée originelle non modifiée. C'est ce qu'on connait en cybersécurité sous le nom d'*attaque par évasion*.

Pour réaliser une attaque, l'attaquant doit donc pouvoir modifier l'entrée du modèle d'IA et faire en sorte que ce soit cette entrée modifiée qui soit soumise au modèle. Le mécanisme est le suivant :

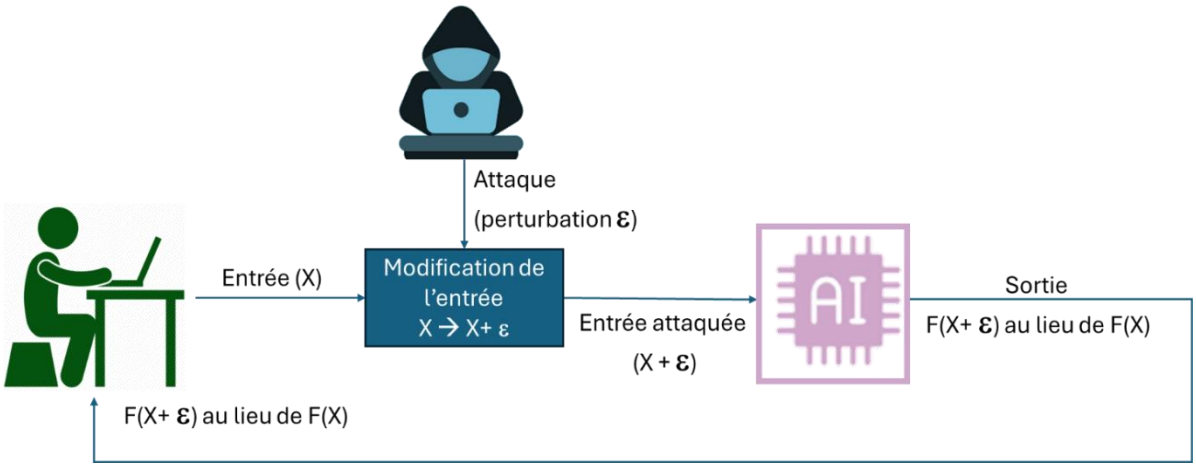


Figure 15 – Attaque adverse

Dans cette attaque, l'attaquant ne cherche pas à modifier ou dégrader le modèle d'IA du système attaqué. Tout ce qui l'intéresse c'est de lui faire produire une sortie non cohérente avec celle qu'il aurait fournie selon l'entrée originelle avant sa modification. Si la modification de l'entrée ne génère pas de modification de la sortie, alors l'attaque aura échoué.

Deux types d'attaques peuvent être envisagés :

- L'attaque avec cible (*targetted attack*) : dans ce type d'attaque, l'attaquant veut que la sortie produite par le modèle d'IA ainsi attaqué soit égale à une cible bien précise.
- L'attaque sans cible : dans ce type d'attaque, l'attaquant ne cherche qu'à faire produire un résultat erroné, sans que ce résultat erroné corresponde à une cible particulière.

Au-delà de la capacité à accéder à l'entrée, à la modifier puis à soumettre l'entrée modifiée au système d'IA, la difficulté pour l'assaillant est de dimensionner la modification à apporter à l'entrée pour qu'elle soit :

- Suffisamment faible pour que l'entrée modifiée ne soit pas facilement détectable et donc rejetée par des mécanismes de protection adaptés du système d'IA.
- Suffisamment forte pour que cette modification ait un impact sur la sortie du système.

Un des premiers exemples opérationnels d'attaque adverse a été la falsification d'un panneau de signalisation routière pour perturber un système d'aide à la conduite. Typiquement :

- L'entrée brute du système d'IA analysant l'image (typiquement un réseau de neurones de type ConvNet), hors attaque, est le panneau de sens interdit (par exemple) à gauche sur la figure :

Figure 16 – Attaque sur l'entrée (à gauche l'entrée, une perturbation avec des autocollants au milieu, l'entrée modifiée à droite et finalement le panneau reconnu)

- L'attaque consiste à coller sur le panneau des autocollants (qui jouent alors le rôle du ε) de sorte que la caméra va soumettre au système d'IA l'image du panneau modifié (à droite sur la figure).
- Si l'attaque réussit, ces « ajouts » sur le panneau vont changer la sortie du système d'IA, qui ne reconnaitra pas le panneau « sens interdit » mais tout autre panneau (ou tout autre objet voire ne pas détecter d'objet du tout). Une attaque avec cible serait une attaque dimensionnée pour que la sortie du système d'IA soit égale par exemple au panneau « sens unique » (tout à droite sur la Figure 16).

Au global, l'attaque est représentée dans la Figure 17 suivante :

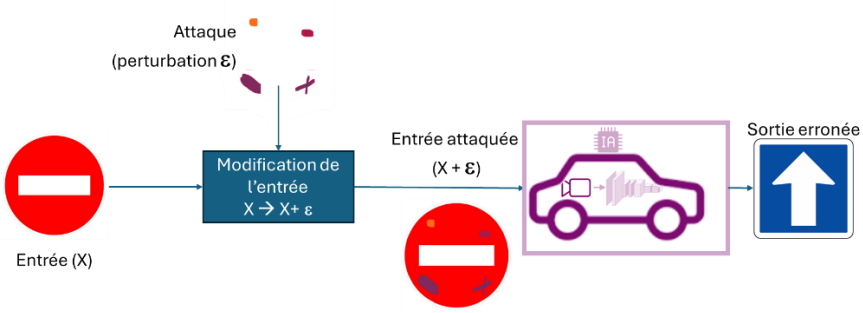


Figure 17 – Attaque adverse sur un panneau de circulation

Dans le cas général, le dimensionnement efficace de la modification à apporter pour réussir l'attaque désirée sera facilité par le fait que l'attaquant puisse avoir accès à des informations sur le système d'IA. Deux cas peuvent en effet se présenter :

- La structure et les paramètres du modèle issus de l'entraînement du système sont connus de l'assaillant (par exemple dans le cas d'un modèle open-source), on parle alors d'attaque « boîte blanche ». L'assaillant a alors toute liberté pour dimensionner ses modifications et réaliser ainsi des attaques pertinentes voire ciblées.
- La structure et les paramètres du modèle issu de l'entraînement du système ne sont pas connus de l'assaillant, on parle alors d'attaque « boîte noire ». S'il veut définir une modification qui respecte les contraintes exposées ci-dessus, l'assaillant devra se constituer un modèle approchant le modèle qu'il veut attaquer. Ce modèle « de substitution » lui permettra alors de calculer les modifications.

L'assaillant va alors être confronté à la question principale posée par ce genre d'attaque, celle de la transférabilité[24] de l'attaque : une attaque réglée sur un modèle de substitution peut-elle fonctionner sur un modèle différent, et si oui avec quelle probabilité de réussite ? Même si des études montrent que dans certains cas l'attaque peut réussir avec une certaine probabilité, le succès n'est pas garanti a priori quelle que soit l'application.

De son côté, l'opérateur du système d'IA devra mettre à disposition un système suffisamment robuste pour qu'une modification de l'entrée en-deçà d'un seuil de détection[25] qu'il aura mis en place ne modifie pas la sortie du dit système. La stratégie de défense dépendra de la connaissance que peut avoir un attaquant du modèle et de ses paramètres.

Nous avons parlé ici des attaques adverses qui visent à modifier les entrées du système d'IA. Les attaques qui visent à modifier les sorties une fois calculées relèvent, elles, de la cybersécurité au sens protection des canaux d'échanges informatiques entre le système et l'utilisateur.

Pour plus de détails sur les attaques adverses, on pourra consulter[26].

4 Se protéger

Prévention

La prévention des attaques sur l'IA inclut un grand nombre de méthodes que nous allons présenter rapidement. Les fiches pratiques de description des attaques sur l'IA (voir section 5) décrivent au verso de la fiche les mesures de prévention spécifiques qui devraient être mises en œuvre avant de mettre le système d'IA en production pour éviter le genre d'attaques décrit dans la fiche (voir section 5.1.2.1). L'ensemble des mesures de prévention qui seront décrites dans ces fiches ne seront pas détaillées ici.

[24] https://arxiv.org/pdf/1605.07277

[25] Dans le cas d'un LLM, un mécanisme de défense pourrait être par exemple de faire confirmer par l'utilisateur en lui reformulant son prompt que le LLM a bien le bon prompt initial, et ce via un autre canal que celui possiblement attaqué …

[26] https://arxiv.org/pdf/1412.6572 et https://arxiv.org/pdf/2302.09457

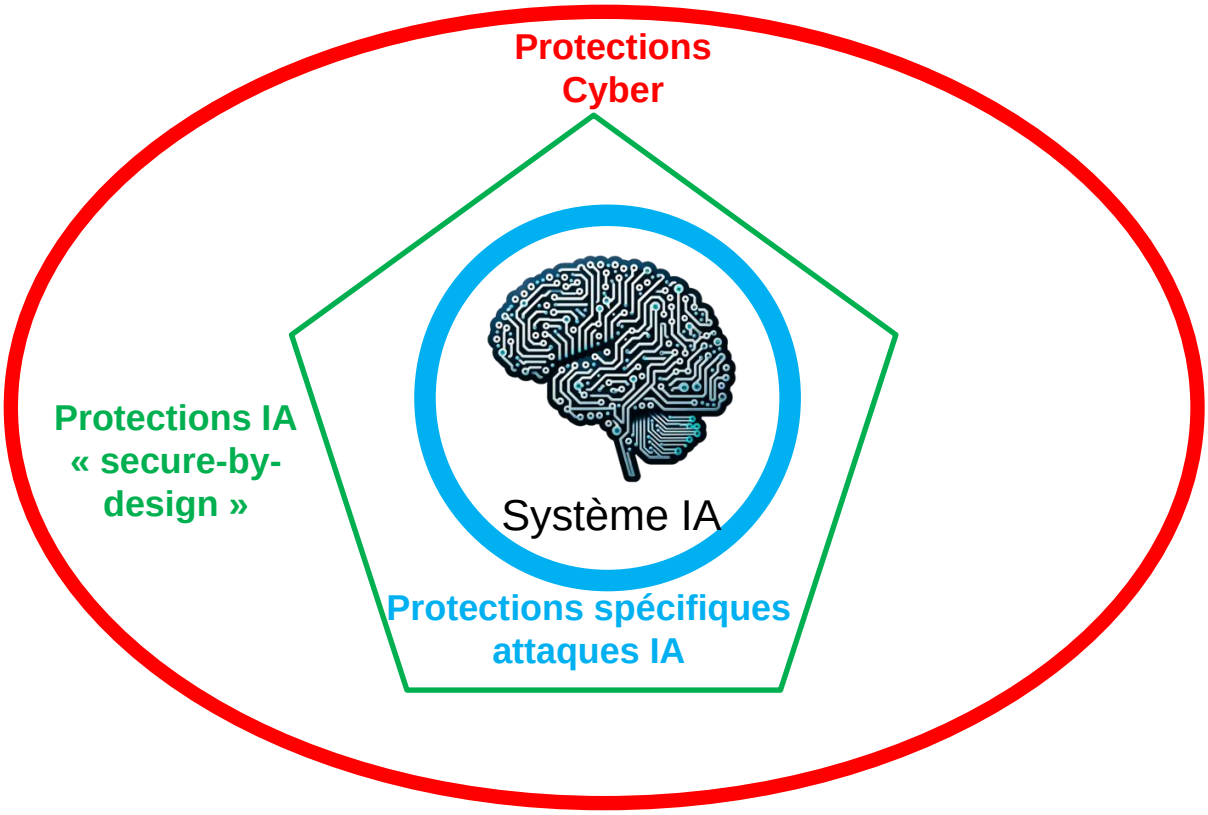


Figure 18 – Protection d'un système IA

4.1.1 Les types de mesures de prévention

4.1.1.1 Les mesures de prévention classiques de cybersécurité

Toute cyberattaque est en général une suite d'actions (la *kill chain*) qui vont s'enchaîner jusqu'à l'obtention par l'attaquant de ses objectifs. MITRE [18] décrit les différentes tactiques employées. De même, une attaque sur l'IA est une suite d'actions malicieuses et commencera en général par une cyberattaque classique (par exemple, l'attaquant doit s'ouvrir une entrée sur les données du système), on doit donc, pour prévenir l'attaque sur l'IA, commencer par mettre en place toutes les techniques classiques de cybersécurité : *une attaque sur l'IA c'est une cyberattaque + une attaque spécifique IA.* Le système de prévention des attaques sur l'IA est donc constitué de trois *lignes de défense* emboîtées comme on le voit sur la Figure 18 ci-dessus. Les mesures de prévention classiques en cybersécurité sont par exemple présentées par l'ANSSI dans son guide d'hygiène informatique qui en identifie 42 [3]. Nous ne les détaillerons pas ici.

4.1.1.2 Les mesures de prévention spécifiques à l'IA

Dans la précédente section 0, nous avons déjà listé les éléments spécifiques à un système d'IA qui nécessitent une attention supplémentaire.

Pour aller plus dans le détail, en avril 2024, l'ANSSI a publié un guide sur les *Recommandations de sécurité pour un système d'IA générative* [1] qui liste 35 recommandations à suivre pour constituer un système d'IA (générative) *secure-*

by-design. On a listé à la section 2.1.7 les 35 recommandations de l'ANSSI pour assurer cette protection dès la conception. Ce guide faisait suite à un document [20] publié par toutes les grandes agences mondiales de sécurité *Guidelines for Secure AI System Development* fournissant des directives pour aider les fournisseurs à créer des systèmes d'IA qui fonctionnent comme prévu, sont disponibles en cas de besoin et fonctionnent sans révéler de données sensibles à des parties non autorisées.

Enfin, l'ANSSI a récemment publié avec de nombreux partenaires une *Analyse commune Haut niveau des risques cyber liés à l'IA* [2]. Le document propose une liste de recommandations qui affinent les 35 recommandations précédentes. Notons qu'en prolongation de ce document, on pourrait approfondir l'analyse des risques pour prioriser les contremesures en fonction du niveau de risque. Nous n'avons pas effectué ce travail ici.

4.1.1.3 Les mesures de prévention spécifiques à certains types d'attaques sur l'IA

On peut enfin mettre en place des mesures de prévention ciblées sur certains types d'attaques sur l'IA. Donnons quelques exemples :

- Pour protéger les données contre l'empoisonnement en phase d'entraînement : on mettra en place des mécanismes d'identification des données inattendues ou malveillantes susceptibles d'incidences sur l'apprentissage du modèle. Lorsque cela est possible, les données seront chiffrées au repos et en transit (voir la section 4.2 sur la cryptographie). Il sera également possible d'entraîner le système IA à se protéger de l'empoisonnement en l'entraînant, en plus des données d'entraînement, sur des données empoisonnées.
- Pour se protéger contre les empoisonnements et manipulations de modèle : lors de l'utilisation des modèles open source, on procèdera à une évaluation de sécurité exhaustive de toutes les dépendances et de tous les composants tiers, comme les bibliothèques, les *frameworks* ou les modèles d'IA générative téléchargés, pour analyser leur réputation, les vulnérabilités connues ainsi que leur maintien en condition de sécurité. Il est préférable de les télécharger depuis des référentiels réputés, des plateformes de confiance avec des pratiques de sécurité bien établies, et uniquement des versions stables et bien maintenues.
- Radar des solutions de sécurité IA : certains éditeurs proposent déjà des solutions visant à protéger contre certaines attaques. La société Wavestone a ainsi publié un *Radar des Solutions de Sécurité IA 2024 [13]* qui identifie en septembre 2024 88 éditeurs offrant des solutions pour :
  - Anti deepfake ;
  - Protection des données et confidentialité IA ;
  - Détection et réponse des algorithmes de Machine Learning ;
  - Chatbot sécurisé et filtrage LLM ;

- Collaboration sécurisée en Machine Learning ;
- Evaluation de la robustesse et des vulnérabilités du modèle ;
- Gestion des risques IA ;
- Données synthétiques / Anonymisation ;
- Ethique, explicabilité et justesse de traitement ;
- Mise en conformité réglementations IA.

Nous avons entrepris un travail d'exploration du marché des solutions logicielles et publierons un rapport dédié sur le sujet.

Figure 19 – Radar Wavestone des solutions de sécurité IA

4.1.2 Les mesures de prévention par phase du cycle de vie

Pour établir des mesures de prévention adaptées au contexte d'emploi des modèles d'IA, il faut d'abord mettre en œuvre un cadre de gestion des risques (Risk Management Framework) tel que décrit par le NIST[27] pour identifier, évaluer et gérer les risques associés aux modèles d'IA. Cela comprend la catégorisation des informations, la sélection des contrôles de sécurité et la surveillance continue.

En particulier, l'ANSSI préconise une approche par les risques cyber [2] pour développer la confiance dans l'intelligence artificielle. L'évaluation des risques doit se faire tout au long du cycle de vie d'un modèle d'IA, de sa conception à sa mise au rebut, et en tenant compte des différents environnements informatiques (développement, tests et validation, exploitation…) sur lesquels il s'appuie au cours

[27] https://csrc.nist.gov/pubs/sp/800/37/r2/final

de chaque phase de son cycle de vie. Les moyens de protection doivent être toujours adaptés au contexte de l'entreprise et aux risques identifiés.

Les éléments à prendre en compte pour ces analyses de risque sont :

- Les **systèmes informatiques** sous-jacents qui fournissent les capacités de stockage, de calcul et de traitement.
- Le **modèle d'IA** en lui-même (paramètres, format de stockage…).
- Les **données** qui servent à l'entraînement du modèle d'IA, mais aussi celles qui sont récupérées en phase d'exploitation de ce modèle à travers le mécanisme de RAG appliqué à certaines données de l'entreprise ou directement sur Internet.
- Les **entrées/sorties** des modèles et les interactions avec les humains ou avec d'autres modèles d'IA ou/et des systèmes informatiques. Dans le dernier cas, cela inclut aussi la technologie d'automatisation du processus.

Afin de conserver une cohérence entre les attaques et les moyens de défense, et comme nous avons une classification des attaques selon les 7 phases du cycle de vie d'un modèle d'IA de l'OCDE (voir notre taxonomie en section 0), les mesures de prévention à appliquer doivent s'appuyer aussi sur cette structuration :

A. Planification et design ;
B. Collecte et traitement des données ;
C. Construction du modèle / adaptation d'un modèle existant ;
D. Test, évaluation, vérification ;
E. Mise à disposition, utilisation, déploiement ;
F. Exploitation et maintenance ;
G. Décommissionnement / mise au rebut.

Afin d'identifier les mesures de prévention à mettre en œuvre, nous nous sommes appuyés sur les documents suivants :

- ANSSI, Recommandations de sécurité pour un système d'IA générative, [1]
- ANSSI, Développer la confiance dans l'IA à travers une approche par les risques cyber, [2]
- ANSSI, Guide d'hygiène informatique, [3]
- MITRE ATLAS, [17]

Par mesure de clarté, toutes les recommandations de ces documents ont été listées en annexe 1 (même les mesures qui relèvent du contexte « I - Protection cybersécurité sur l'infrastructure » et qui ne sont donc pas spécifiques à l'IA).

Les réflexions effectuées sur chacune d'entre elles ont été clairement tracées et expliquées, ce sont les suivantes :

- Les ***doublons*** ont été identifiés.
- La répartition selon les différentes lignes de défenses illustrées sur la figure 18 a été faite :

I Protections cybersécurité sur l'infrastructure ;
II Protections IA « *secure by design* » ;
III Protections spécifiques contre les attaques IA.

- La répartition selon les ***7 phases de l'OCDE***

  Ce traitement n'a pas été fait sur les mesures de protection de cybersécurité classiques, puisque non pertinent dans ce cadre générique qui ne concerne pas spécifiquement les SIA, et donc pas spécifiquement leur cycle de vie.

- Une ***présentation harmonisée des catégories*** des mesures est proposée (tenant compte de celles déjà existantes dans les documents d'origine).

Pour résumer :

Document source	**Mesures de prévention brutes**	**Traitements effectués sur les mesures de prévention (codes couleurs de la Figure 18)**
ANSSI document [1]	Extraction de toutes les 35 mesures listées	• Certaines ont été classées dans la catégorie des mesures de cybersécurité classiques • Certaines ont été classées dans la catégorie des mesures de prévention spécifiques à l'IA « secure by design » • Doublons identifiés avec [2], [3] et [17] • Répartition selon les 7 phases de l'OCDE
ANSSI document [2]	Extraction de toutes les 43 mesures listées	• Certaines ont été classées dans la catégorie des mesures de cybersécurité classiques • Certaines ont été classées dans la catégorie des mesures de prévention spécifiques à l'IA « secure by design » • Doublons identifiés avec [1], [3] et [17] • Répartition selon les 7 phases de l'OCDE
ANSSI document [3]	Extraction de toutes les 42 mesures listées	Ces mesures ont été classées dans la catégorie des mesures de cybersécurité classiques
MITRE ATLAS [17]	Extraction de toutes les 25 mesures listées	• Certaines ont été classées dans la catégorie des mesures de cybersécurité classiques • Certaines ont été classées dans la catégorie des mesures de prévention spécifiques à l'IA « secure by design » • Certaines ont été classées dans la catégorie des mesures spécifiques à certaines attaques sur l'IA • Doublons identifiés avec [1] et [2] • Répartition selon les 7 phases de l'OCDE

Il résulte de ce travail de réflexion, une liste consolidée et synthétique de mesures de prévention présentées à travers les tableaux proposés en section 9 (Annexe 1 - Méthodes de prévention). Ces tableaux permettent d'identifier rapidement les mesures de prévention à déployer dans chaque phase du cycle de vie d'un modèle d'IA de l'OCDE et selon le contexte de protection.

Par ailleurs, tout au long du cycle de vie du présent document, de nouvelles mesures de prévention spécifiques à certaines attaques sur l'IA seront ajoutées, au fur et à mesure de l'avancée des réflexions sur les fiches d'attaques. Parmi les sources déjà identifiées, on peut citer :

- Mesures de prévention listées sur les fiches d'attaques établies,
- Articles scientifiques,
- Expérience des membres du groupe de travail à l'origine de ce document,
- Editeurs de solutions de sécurité.

Remédiation

4.1.3 Architecture de Gestion d'Incident pour les Systèmes d'IA

Face aux menaces croissantes visant les Systèmes d'Intelligence Artificielle (SIA), une gestion efficace et structurée des incidents est essentielle pour garantir résilience, sécurité et conformité réglementaire. Nous proposons donc une architecture de gestion d'incident appliquée aux systèmes d'IA, intégrant les bonnes pratiques issues des cadres de référence tels que l'ISO/IEC 27035, les recommandations de l'ANSSI, du NIST et les directives de la CNIL [21 – 24]. Elle est structurée autour de trois volets principaux : Gouvernance et Gestion de Crise, Détection et Investigation, et Remédiation et Reconstruction, accompagnées d'une boucle d'amélioration continue permettant d'assurer la résilience et l'optimisation des processus de réponse aux incidents IA.

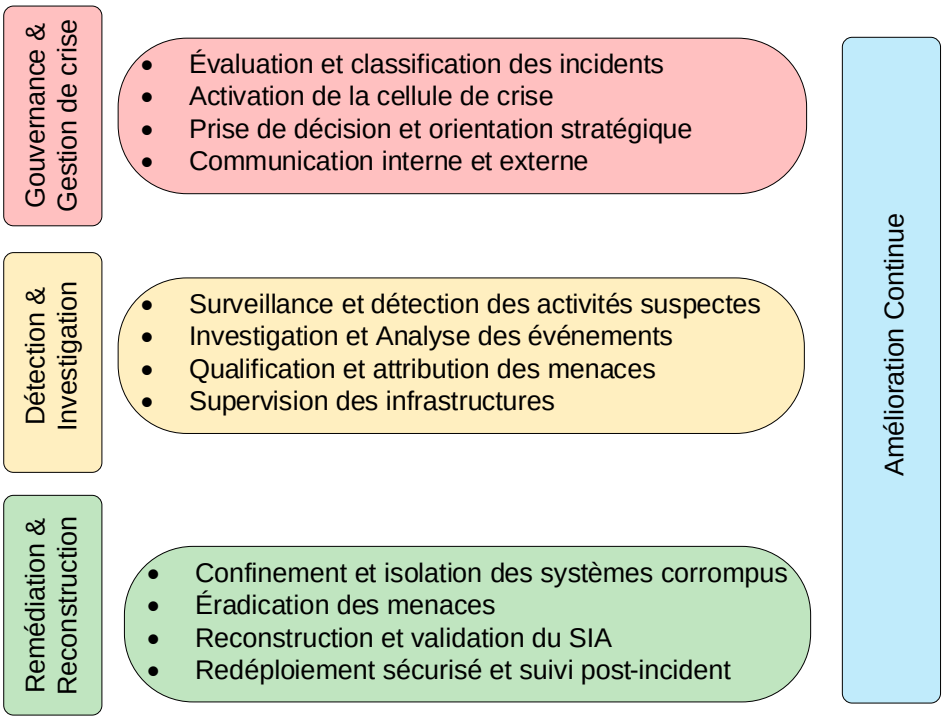


Figure 20 – Gestion des incidents pour les systèmes d'IA

🔴 Gouvernance & Gestion de Crise

La gouvernance et la gestion de crise assurent l'orchestration et l'alignement stratégique des réponses aux incidents IA, garantissant une réactivité optimisée et une conformité réglementaire. Cette phase repose sur :

- L'évaluation et la classification des incidents selon leur impact sur la Traçabilité, Disponibilité, Intégrité et Confidentialité des systèmes IA.
- L'activation d'une cellule de crise mobilise les équipes SOC, DevSecOps, IA et Juridique, tout en assurant une coordination avec les régulateurs (ANSSI, CNIL, partenaires).
- L'analyse des risques en temps réel guide la prise de décision, permettant d'orienter la réponse vers un confinement immédiat, une investigation approfondie ou une remédiation prioritaire.
- La gestion des communications internes et externes garantit la transparence et la conformité aux obligations de notification.

Cette phase s'aligne sur les principes de cyber-résilience du NIST CSF (*Govern & Identify*) et les recommandations de l'ANSSI, assurant une supervision efficace et stratégique des incidents IA.

🟡 Détection & Investigation

La phase de détection et investigation permet la surveillance proactive et qualification des menaces, elle repose sur :

- Une surveillance proactive et une analyse approfondie des menaces IA afin d'identifier rapidement les compromissions et qualifier les attaques. Le SOC

(*Security Operations Center*) exploite des indicateurs de compromission (IoC) et s'appuie sur des solutions de *Threat Intelligence* et de corrélation des événements de sécurité via SIEM pour une détection avancée et réactive.

- L'investigation approfondie permet d'analyser les flux d'entraînement et d'inférence des modèles IA pour détecter les attaques adversariales, les empoisonnements de données et les dérives algorithmiques.
- L'attribution des attaques et la qualification des menaces permettent d'orienter les mesures de confinement et de remédiation adaptées à la criticité de l'incident.
- La supervision continue des infrastructures IA via l'audit des pipelines MLOps, la surveillance des flux d'API et l'analyse comportementale des modèles déployés est essentielle pour anticiper les risques de compromission et renforcer la posture de cybersécurité IA.

Cette phase suit les principes de supervision de la CNIL et du NIST CSF (*Detect*), garantissant une capacité de détection et d'investigation optimisée face aux menaces émergentes ciblant les SIA.

🟢 Remédiation & Reconstruction : Confinement, validation et redéploiement sécurisé

La remédiation et la reconstruction suivent le modèle E3R (Endiguement, Éviction, Éradication, Reconstruction) de l'ANSSI, garantissant un rétablissement sécurisé des SIA.

- Le confinement et l'isolation des systèmes compromis permettent de stopper la propagation de l'attaque en restreignant les accès aux infrastructures affectées.
- L'éradication des menaces supprime les accès malveillants et neutralise les vecteurs d'intrusion pour empêcher toute persistance de la menace.
- La reconstruction et validation du SIA consistent à corriger les vulnérabilités, assainir les jeux de données IA et vérifier l'intégrité des modèles et des infrastructures.
- Le redéploiement sécurisé et le suivi post-incident assurent une remise en production sans risque résiduel, validée par un audit de conformité et de cybersécurité.

Cette approche garantit une remise en service conforme aux exigences du NIST CSF (*Respond & Recover*), minimisant les risques de récidive et assurant une résilience renforcée des systèmes IA face aux menaces futures.

🔵 Boucle d'Amélioration Continue

L'amélioration continue est essentielle pour tirer parti de chaque incident et renforcer durablement la posture de cybersécurité des SIA. Cette phase repose sur un Retour d'Expérience (RETEX) structuré, permettant de documenter les incidents,

identifier les vulnérabilités exploitées et affiner les stratégies de détection et de remédiation. L'évolution des politiques de cybersécurité IA s'appuie sur la mise à jour des modèles de détection et l'optimisation des mécanismes de supervision pour anticiper les nouvelles menaces. En parallèle, la formation continue des équipes via des exercices Red Team IA (voir Glossaire), simulations adversariales et tests d'intrusion permet de développer des capacités de réponse proactives face aux cyberattaques ciblant les systèmes IA. Cette approche s'aligne sur les principes du NIST CSF (*Improve*) et les recommandations de l'ANSSI, garantissant un renforcement progressif et adaptatif de la cybersécurité IA.

4.1.4 Checklist de remédiation alignée avec le cycle de vie d'un SIA

Pour répondre efficacement aux incidents affectant un SIA, nous nous appuyons sur une méthodologie complète, alignée sur les standards internationaux (ISO/IEC 27035, NIST CSF, ANSSI, CNIL) et les principes de gestion de crise appliqués aux environnements IA. Cette approche couvre l'ensemble du cycle de vie d'un SIA et s'intègre à une architecture de réponse aux incidents bien définie.

Afin de faciliter son application, une checklist opérationnelle a été élaborée. Elle intègre des actions stratégiques et techniques permettant :

- D'anticiper les risques et structurer la gouvernance de la sécurité IA.
- D'identifier et qualifier les menaces pesant sur les modèles IA et leurs infrastructures.
- De remédier efficacement aux attaques et restaurer les systèmes IA impactés.
- D'améliorer continuellement la posture de sécurité IA grâce à un retour d'expérience structuré.

Cette approche est pragmatique, adaptable et adaptée aux défis des systèmes IA modernes. Ainsi, la checklist de remédiation permettrait aux RSSI, CTO et DSI de structurer efficacement leurs réponses aux incidents IA, en garantissant une mise en œuvre méthodique et conforme aux meilleures pratiques de cybersécurité. La check liste est fournie en Annexe (section 10), elle fournit les méthodes de remédiation exploitées dans les fiches descriptives des attaques.

5 Fiches pratiques : Analyse des principales attaques

Le format des fiches

Le présent livrable a pour objectif de proposer une lecture pratique des scénarios connus de compromission d'un SIA. Pour arriver à ces fins de la manière la plus lisible et efficace possible, nous proposons l'usage de fiches pédagogiques sous le format présenté ci-dessous.

5.1.1 Au recto de la fiche

Pour le recto d'une fiche pédagogique, la représentation suivante est proposée en Figure 21.

Le recto d'une fiche pédagogique proposée dans ce livrable se lit de haut en bas et de gauche à droite. Un tel ordonnancement a pour but de décrire dans un premier temps la typologie d'attaque étudiée et d'entrer progressivement en détail dans le scénario.

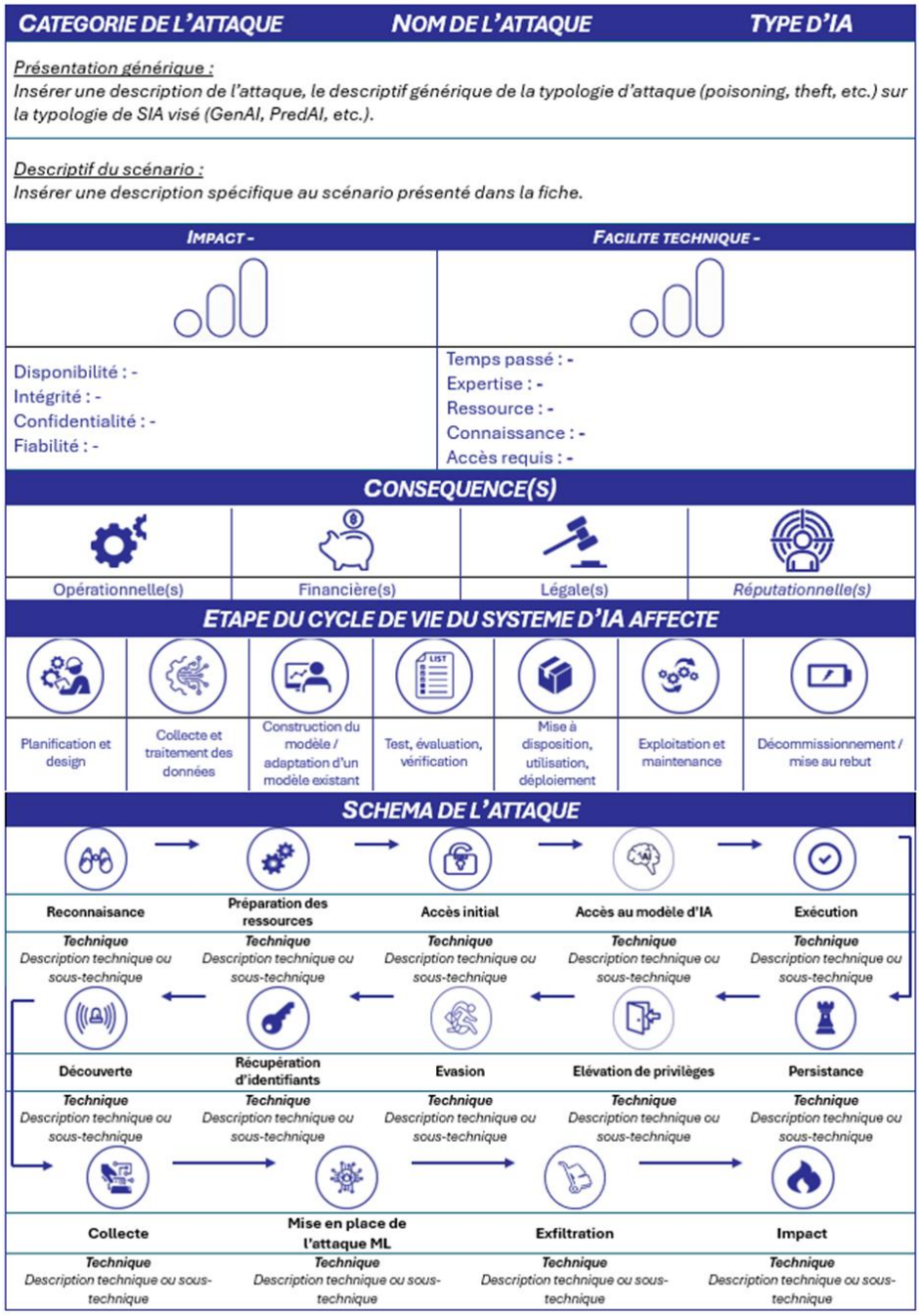


Figure 21 – Recto de la fiche descriptive d'une attaque sur un SIA

5.1.1.1 La description du scénario d'attaque

Le document commence par les segments descriptifs suivants :

CATEGORIE DE L'ATTAQUE	NOM DE L'ATTAQUE	TYPE D'IA
Présentation générique : *Insérer une description de l'attaque, le descriptif générique de la typologie d'attaque (poisoning, theft, etc.) sur la typologie de SIA visé (GenAI, PredAI, etc.).*		
Descriptif du scénario : *Insérer une description spécifique au scénario présenté dans la fiche.*		

Figure 22 – Un format vierge de description du scénario au recto de la fiche

La légende suivante permet de comprendre l'intérêt de chacun des champs :

Catégorie de l'attaque	La « *catégorie de l'attaque* » qualifie la catégorie d'attaque[28] dont fait partie le scénario étudié. Il va s'agir des exemples de catégories d'attaque présentés précédemment.
Nom de l'attaque	Le « *nom de l'attaque* » qualifie le scénario d'attaque étudié dans la fiche pédagogique. Il s'agit d'un des scénarios recensés dans ce livrable parmi les grandes catégories d'attaque.
Type d'IA	Le « *type d'IA* » qualifie la technologie d'intelligence artificielle visée par le scénario d'attaque étudié dans la fiche.
Présentation générique	La « *présentation générique* » est une description succincte et générique de la catégorie de l'attaque.
Descriptif du scénario	Le « *descriptif du scénario* » est une description succincte de la mise en œuvre du scénario et de ses enjeux pour le système d'intelligence artificielle visé.

5.1.1.2 La qualification des scénarios d'attaque

La fiche pédagogique se poursuit par des segments relatifs à la qualification du scénario d'attaque. La finalité de cette section est de proposer une série

[28] Pour rappel, les grandes catégories d'attaque sont recensées sous la forme d'une proposition de taxonomie dans la section 0.

d'indicateurs pour démarquer la gravité d'un scénario d'attaque par rapport à un autre.

IMPACT -	FACILITE TECHNIQUE -
	
Disponibilité : - Intégrité : - Confidentialité : - Fiabilité : -	Temps passé : - Expertise : - Ressource : - Connaissance : - Accès requis : -

Figure 23 – Un format vierge d'évaluation des critères et indicateurs

La méthode de qualification de chaque critère et indicateur est présentée et détaillée plus tôt dans ce livrable[29]. Les critères seront grisés s'il a été jugé qu'une évaluation de l'impact n'était pas applicable ou pertinente au regard de la nature de l'attaque.

5.1.1.3 Les conséquences des scénarios d'attaque

Pour ne pas exclure les conséquences stratégiques d'une attaque sur une organisation, la section « *Conséquence(s)* » propose d'identifier des impacts complémentaires : opérationnels, financiers, légaux ou réputationnels.

CONSEQUENCE(S)			
			
Opérationnelle(s)	Financière(s)	Légale(s)	*Réputationnelle(s)*

Figure 24 – Un format vierge d'identification des conséquences stratégiques d'une attaque

Les conséquences sont identifiées et justifiées en amont dans ce support[30].

5.1.1.4 Les étapes du cycle de vie du système d'IA affecté

Pour contextualiser une attaque dans le cycle de vie d'un système d'IA, il est proposé d'identifier les étapes du cycle les plus susceptibles d'être sujets à ces scénarios. Pour le faire il a été pris le parti d'adopter l'approche du cycle de vie de l'OCDE (tel qu'évoqué en 2.1.2.1). Cette approche a été retenue de par son efficacité à résumer les étapes clés du cycle de vie d'un SIA tout en restant agnostique des technologies employées. Sur une fiche pédagogique, il s'agira donc :

[29] Une section est dédiée à ce sujet en 2.4 Evaluations qualitatives des attaques.
[30] Une section est dédiée à ce sujet en 2.4.4 Les conséquences d'une attaque sur l'organisation

- De laisser en bleu la ou les étapes pouvant constituer un contexte pertinent à la mise en œuvre du scénario d'attaque ; ou a contrario,
- De griser la ou les étapes du cycle de vie si l'attaque a des spécificités techniques ou un mode opératoire tel que le scénario a peu ou aucune probabilité de se produire.

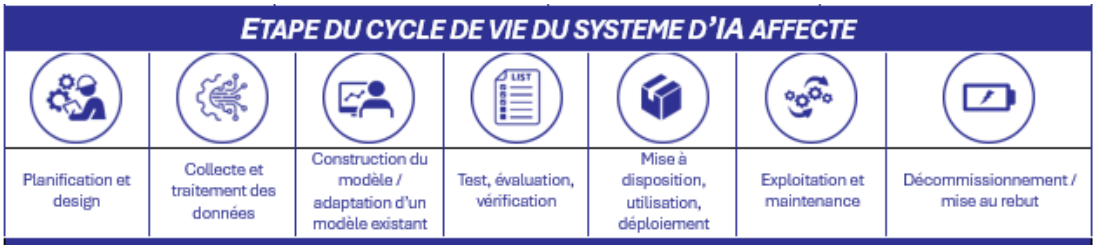


Figure 25 – Un format vierge d'identification des étapes du cycle de vie du SIA affectées

Les étapes du cycle de vie du SIA sont sélectionnées en fonction de l'appréciation qui a été faite de leur pertinence au moment de la rédaction de la fiche pédagogique sur le scénario d'attaque.

5.1.1.5 Le schéma de l'attaque

La contextualisation du scénario d'attaque se poursuit par un exercice qui consiste à identifier les étapes susceptibles d'être suivies par un attaquant. L'objectif est de séquencer le chemin parcouru par l'utilisateur malveillant dans sa mise en œuvre du scénario.

Pour cela il a été jugé utile de recourir au référentiel MITRE Atlas [17] qui permet de mettre en évidence, en fonction de l'analyse du scénario, les tactiques et techniques utilisées. La représentation visuelle proposée est la suivante :

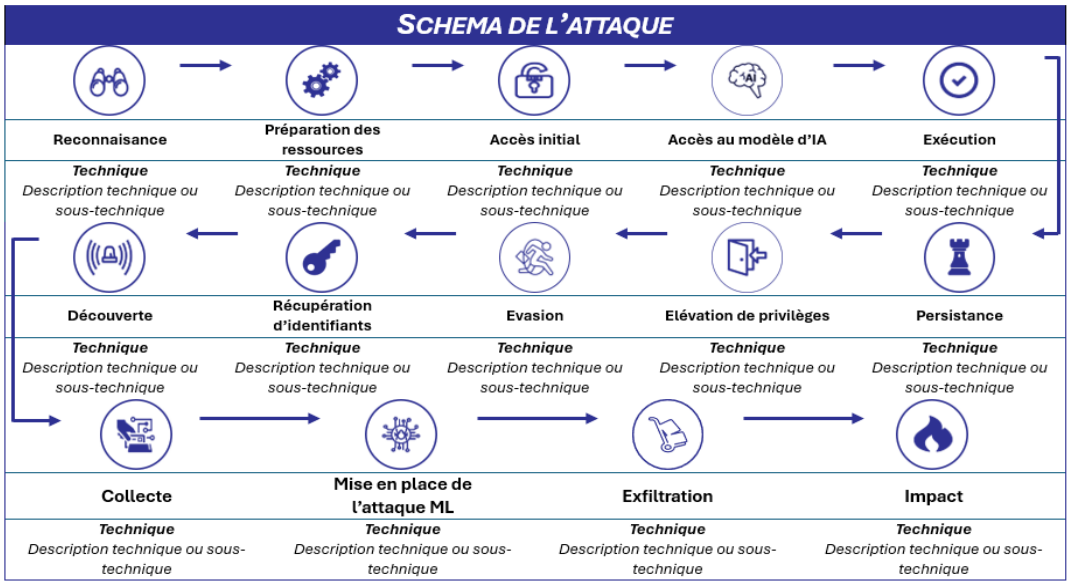


Figure 26 – Une représentation graphique du référentiel MITRE Atlas au format vierge

Avant d'entrer dans l'explication des différents éléments, il convient de développer le sens de lecture du graphique proposé. L'ordre de lecture proposé est celui retenu

par le MITRE pour lister les différentes tactiques, cette liste a été évoquée précédemment[31]. Ce qui veut dire concrètement que la lecture doit se faire dans le sens des flèches proposées sur le graphique en Figure 26. L'ordre se matérialise donc comme suit : Reconnaissance, Préparation des ressources, Accès initial, Accès au modèle d'IA, Exécution, Persistance, Elévation de privilèges, Evasion, Récupération d'identifiants, Découverte, Collecte, Mise en place de l'attaque ML, Exfiltration et Impact.

Il convient cependant de souligner qu'en fonction du scénario présenté dans une fiche pédagogique, l'ordre des tactiques peut varier. Exemple : pour l'extraction de modèle, la phase d'« *Exfiltration* » peut avoir lieu avant la « *Mise en place de l'attaque ML* »

La légende suivante permet de comprendre l'intérêt de chacun des champs proposés :

La Tactique	Une tactique est l'objectif recherché par l'attaquant, il apparaît en gras sous le pictogramme qui la représente graphiquement. Le MITRE Atlas en liste 14.
La(les) Technique(s)	Une technique représente la méthode avec laquelle il va chercher à accomplir son objectif. Les techniques retenues pour expliquer le scénario se trouvent en dessous du titre de la tactique. La matrice MITRE Atlas [17] en liste environ 62, chacune porte un code[32]. Les techniques retenues sont celles apparues comme étant pertinentes au moment de l'analyse. Le cas échéant ces dernières seront accompagnées de descriptions.

5.1.2 Au verso de la fiche

Pour le verso d'une fiche pédagogique, la représentation suivante est proposée :

[31] Il est fait référence ici à l'énumération des tactiques faite en 2.1.5.

[32] Exemple : le prompt injection a le code AML.T0051 https://atlas.mitre.org/techniques/AML.T0051

REMEDIATION

Action	Équipes à mobiliser	Étape du cycle de vie	Complexité	Efficacité
Mesure 1	Équipe IA & mise en production		+	+++
Mesure 2	Équipe IA & mise en production		+++	+

PREVENTION

Mesure 1	Équipe de mise en production		+	+++
Mesure 2	Équipe IA		++	++
Mesure 3	Équipe IA		+++	+++

POUR ALLER PLUS LOIN

- [...]

EXEMPLES CONNUS

- [...]

Figure 27 – Verso de la fiche descriptive d'une attaque sur un SIA

Le verso d'une fiche pédagogique se lit de haut en bas et de gauche à droite. Un tel ordonnancement a pour but de décrire successivement les mesures de remédiation et de prévention suggérées, les sources documentaires utilisées et quelques exemples connus de mise en œuvre du scénario.

5.1.2.1 La prévention

Le parti du verso de la fiche pédagogique est de proposer dans un premier temps une méthode de remédiation, et dans un second dans la section « *Prévention* » de tenter de concevoir une approche pour anticiper, bloquer ou prévenir une nouvelle attaque de ce type.

PREVENTION				
Mesure 1	Équipe de mise en production		+	+++
Mesure 2	Équipe IA		++	++
Mesure 3	Équipe IA		+++	+++

Figure 28 – Un format vierge de la section dédiée à la prévention de l'attaque

La légende suivante définit les différents champs proposés pour lister ces mesures de prévention.

Action	La section « *Action* » recense les mesures retenues, au moment de la rédaction de la fiche pédagogique, pour sensibiliser, anticiper, ou doter de moyen pour prévenir ou bloquer une attaque similaire au scénario étudié. Une mesure est affectée à une équipe, située dans l'étape du cycle de vie du SIA et évaluée dans sa complexité et son efficacité.
Équipes à mobiliser	Les équipes à mobiliser sont les porteurs de la mesure de prévention. Il va s'agir de l'équipe considérée comme étant la plus à même d'intervenir pour anticiper le scénario présenté.
Étape du cycle de vie	L'étape du cycle de vie est la section permettant de situer la mesure de prévention dans le cycle de vie du SIA la plus pertinente pour prévenir ou bloquer le scénario.
Complexité	La complexité est une proposition succincte d'évaluation des obstacles rencontrés dans la mise en œuvre de la mesure. Elle se fait en trois niveaux : *+* La mesure apparaît raisonnablement simple à mettre en œuvre. Elle nécessite peu de moyens humains, techniques ou de temps pour être mise en œuvre ; *++* La mesure implique de mobiliser des moyens humains et ou techniques supplémentaires pour être mise en œuvre ; *+++* La mesure apparaît complexe à mettre en œuvre et nécessite des moyens humains et techniques avancés, et de temps pour être mise en œuvre.

Efficacité	L'efficacité est une proposition succincte d'évaluation des effets de la mesure sur les risques et impacts de l'attaque. *+* La mesure ne permet pas d'anticiper ou de bloquer les risques et impacts de l'attaque sur le système : elle doit être accompagnée d'autres mesures techniques et organisationnelles ; *++* La mesure permet d'anticiper ou de bloquer partiellement ou à moyen terme les risques et impacts de l'attaque sur le système. *+++* La mesure permet à court terme d'anticiper significativement ou de bloquer les risques et impacts de l'attaque sur le système.

5.1.2.2 La remédiation

La démarche d'étude des scénarios a pour vocation d'évaluer une attaque et de la situer dans le cycle de vie d'un SIA. Pour qu'elle soit complète, la suite consiste à énumérer, évaluer et attribuer des mesures de remédiation jugée pertinentes. Une mesure de remédiation est entendue comme étant : une action à plus ou moins long terme permettant de limiter les risques et les impacts d'une attaque étudiée dans une des fiches pédagogiques.

REMEDIATION

Action	Équipes à mobiliser	Étape du cycle de vie	Complexité	Efficacité
Mesure 1	Équipe IA & mise en production		+	+++
Mesure 2	Équipe IA & mise en production		+++	+

Figure 29 – Un format vierge de la section dédiée à la remédiation de l'attaque

La légende suivante définit les différents champs proposés pour lister ces mesures de remédiation.

Action	La section « *Action* » recense les mesures retenues, au moment de la rédaction de la fiche pédagogique, pour diminuer ou supprimer les risques et impacts occasionnés par une attaque. Une mesure est affectée à une équipe, située dans l'étape du cycle de vie du SIA et

	évaluée dans sa complexité et son efficacité à réduire ou supprimer les risques et impacts.
Équipes à mobiliser	Les équipes à mobiliser sont les porteurs de la mesure de remédiation. Il va s'agir de l'équipe considérée comme étant la plus à même d'intervenir pour remédier au scénario évoqué.
Étape du cycle de vie	L'étape du cycle de vie est la section permettant de situer la mesure de remédiation dans le cycle de vie du SIA la plus pertinente pour réduire les risques et impacts de l'attaque.
Complexité	La complexité est une proposition succincte d'évaluation des obstacles rencontrés dans la mise en œuvre de la mesure. Elle se fait en trois niveaux : *+* La mesure apparaît raisonnablement simple à mettre en œuvre. Elle nécessite peu de moyens humains, techniques ou de temps pour être mise en œuvre ; *++* La mesure implique de mobiliser des moyens humains et ou techniques supplémentaires pour être mise en œuvre ; *+++* La mesure apparaît complexe à mettre en œuvre et nécessite des moyens humains et techniques avancés, ainsi que de temps pour être mise en œuvre.
Efficacité	L'efficacité est une proposition succincte d'évaluation des effets de la mesure sur les risques et impacts de l'attaque. *+* La mesure ne permet pas de résoudre les risques et impacts de l'attaque sur le système et nécessite d'être accompagnée d'autres mesures techniques et organisationnelles ; *++* La mesure permet de résoudre partiellement ou à moyen terme les risques et impacts de l'attaque sur le système. *+++* La mesure permet à court terme de réduire significativement ou de supprimer les risques et impacts de l'attaque sur le système.

5.1.2.3 Des compléments

Les fiches pédagogiques se terminent par les sources documentaires utilisées et cas connus recensés. Ces éléments ont permis la rédaction des scénarios d'attaque énumérés à la suite dans ce livret.

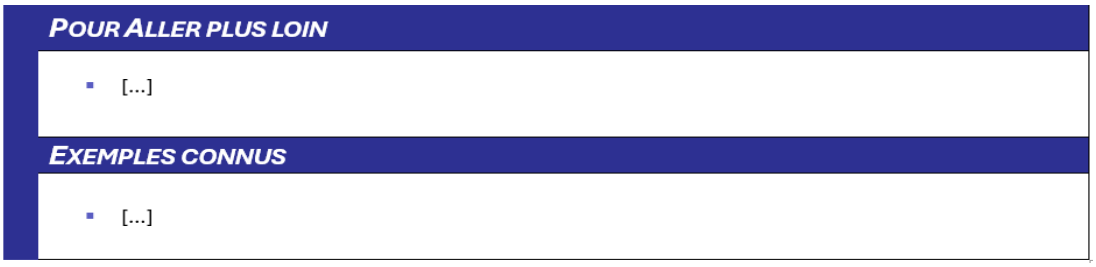


Figure 30 – Un format vierge des sections « Pour aller plus loin » et « Exemples connus »

Pour aller plus loin	Il s'agit d'une section constituée pour compléter les éléments présents dans le support étudiés. Les sources peuvent être issues de ressources universitaires, scientifiques ou institutionnelles.
Exemples connus	La section « *Exemples connus* » a pour finalité les cas de mise en œuvre du scénario d'attaque recensé. Par exemple : pour l'empoisonnement d'un chatbot, le cas Tay serait un exemple connu (voir la fiche ci-dessous).

5.1.3 Démonstration par l'exemple du chatbot Tay

La fiche suivante a pour finalité d'illustrer les éléments précédemment présentés dans ce support ainsi que dans les sections 5.1.1 *Au recto de la fiche* et 5.1.2 *Au verso de la fiche*. Il s'agit d'un cas d'empoisonnement (catégorie de l'attaque) des données d'entrée d'un chatbot (nom de l'attaque), concernant l'IA générative (type d'IA).

EMPOISONNEMENT	EMPOISONNEMENT DES DONNEES D'ENTREE D'UN CHATBOT	GENERATIVE

Présentation générique :
Modifier les données de réentrainement d'un modèle (e.g. historique des conversations avec les utilisateurs) …) pour introduire une déviation de son comportement qui pourra être exploitée.

Descriptif du scénario :
Dans le cas d'un chatbot utilisant les données des interactions avec les utilisateurs pour apprendre en continu, des utilisateurs malveillants ou inconscient du risque pourraient lui fournir en entrée des jeux de données qui, une fois utilisé par le modèle pour se réentraîner, provoquerait des réponses non désirables du modèle.

IMPACT – Moyen (2)	FACILITE TECHNIQUE – Élevée (3)
Disponibilité : N/A Intégrité : **Elevée (3)** Confidentialité : N/A Fiabilité : **Moyen (2)**	Temps passé : **<1 jour (3)** Expertise : **faible (3)** Ressource : **moyenne (2)** Connaissance : **faible (3)** Accès requis : utilisateur interne (2)

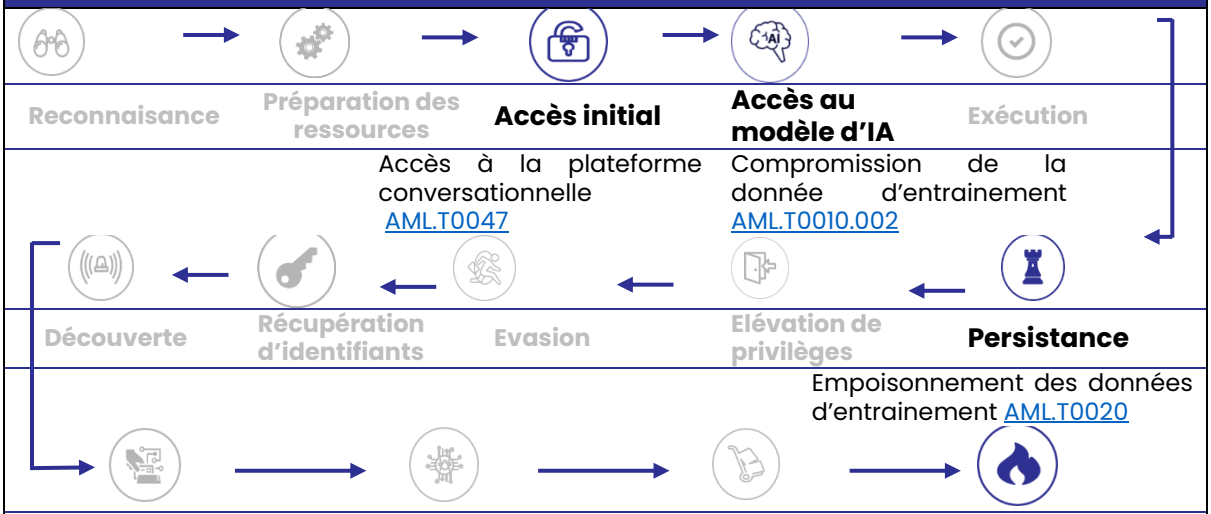
CONSEQUENCES

Opérationnelle(s)	Financière(s)	Légale(s)	Réputationnelle(s)

ETAPES DU CYCLE DE VIE DU SYSTEME D'IA AFFECTEES

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



REMEDIATION

Action	Équipes à mobiliser	Étape du cycle de vie	Complexité	Efficacité
Revenir à des versions stables du modèle.	Équipe IA & mise en production		+	+++
Reconstruire le modèle avec des données fiables.	Équipe IA & mise en production		+++	++

PREVENTION

Sauvegarder des versions stables.	Équipe de mise en production		+	+++
Contrôler les données de réentrainement du modèle.	Équipe IA		++	++
Réévaluer le modèle après réentrainement.	Équipe IA		+++	+++
Mettre en place une procédure dite du "*bouton rouge*".	Équipe de mise en production		+	+

POUR ALLER PLUS LOIN

Attaque sur le chatbot Tay de Microsoft :

- Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications. Acm Sigcas Computers and Society, 47(3), 54-64.
- Lee, P. (2016, March 25). Learning from Tay's introduction - The Official Microsoft Blog. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/
- AI Incident Database. https://incidentdatabase.ai/cite/6/#r1374

Attaques par empoisonnement :

- OWASP Top 10 for LLM Applications VERSION 1.0.1. (2023). https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_0_1.pdf, section « LLM03: Training Data Poisoning »
- Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial Machine Learning: https://doi.org/10.6028/nist.ai.100-2e2023, section 3.2.2 : « Poisoning Attacks »

EXEMPLES CONNUS

- En 2016, le chatbot Tay de Microsoft a été manipulé par des utilisateurs malveillants sur Twitter, ces derniers l'ont abreuvé de messages racistes et offensants. Réutilisés dans l'entraînement de Tay, qui apprenait en continu en se basant sur l'historique de ses interactions, ses messages ont fait que le chatbot a commencé à publier des messages racistes et offensants.
- En moins de 24 heures, Tay a été désactivé pour éviter de causer davantage de dommages.

Figure 31 – La fiche descriptive du cas de Tay

Les fiches d'attaque par phase

Nous présentons ici 10 fiches d'attaque dans différentes phases du cycle de vie comme indiqué sur la taxonomie ci-dessous :

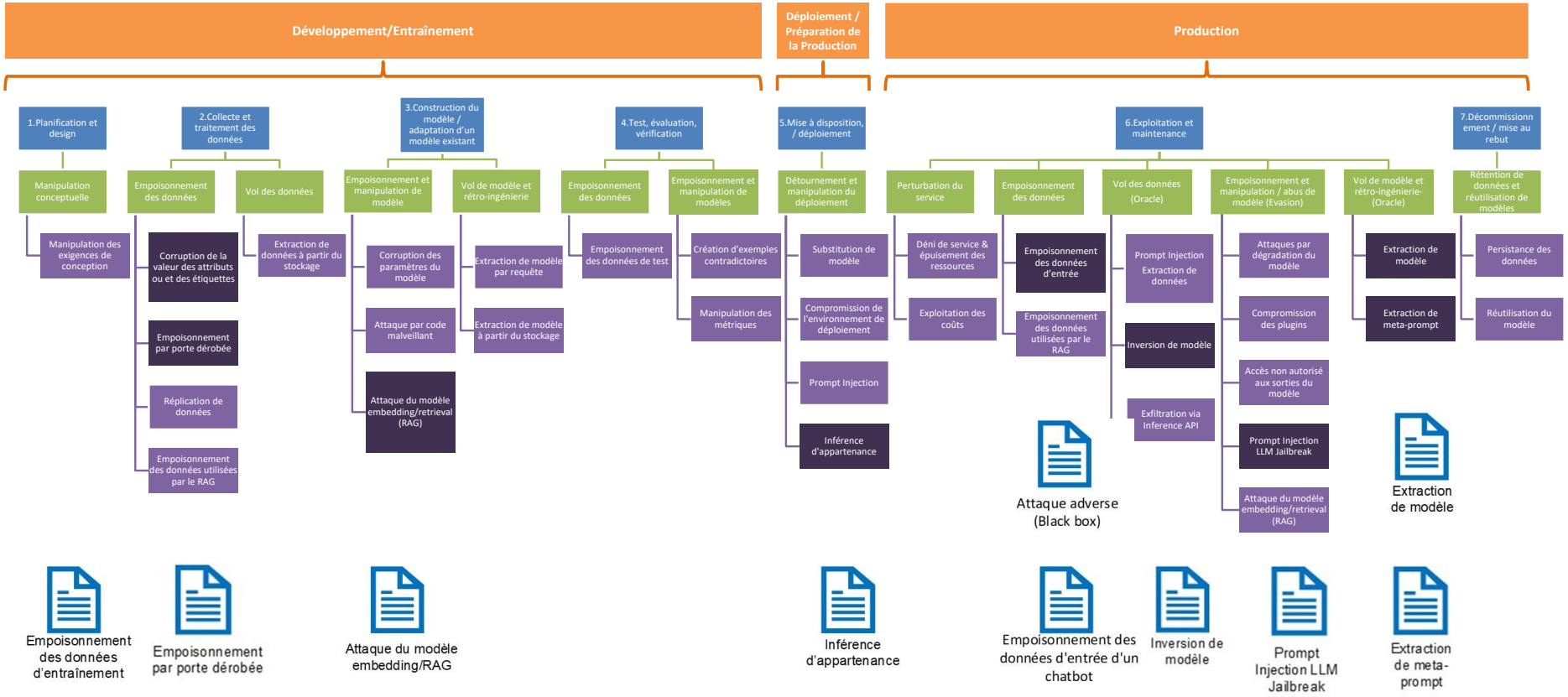


Figure 32 – Fiches présentées

5.1.4 Planification et design

5.1.4.1 Manipulation conceptuelle

5.1.4.1.1 Manipulation des exigences de conception

[Fiche à venir]

5.1.5 *Collecte et traitement des données*

5.1.5.1 *Empoisonnement des données*

5.1.5.1.1 Corruption de la valeur des attributs ou et des étiquettes

EMPOISONNEMENT	EMPOISONNEMENT DES DONNEES D'ENTRAINEMENT	PREDICTIVE & GENERATIVE

Présentation générique :
Empoisonnement visant à modifier les données d'entraînement pour induire le modèle en erreur pendant l'apprentissage.

Descriptif du scénario :
Les données elles-mêmes ou les étiquettes de ces données peuvent être empoisonnées (c.à.d. modifiées). Selon la proportion de données d'entrainement empoisonnées et la qualité de l'empoisonnement, lors de son utilisation finale le modèle peut fournir une réponse incorrecte quel que soit les données fournies, ou seulement pour des entrées particulières.

IMPACT – Élevé (3)	FACILITE TECHNIQUE – Moyenne (2)
Disponibilité : N/A Intégrité : **Élevé (3)** Confidentialité : N/A Fiabilité : **Élevé (3)**	Temps passé : **Modéré (2)** Expertise : **Moyenne (2)** Ressource : **Faible (3)** Connaissance : **Moyenne (2)** Accès requis : Grand Public (3)

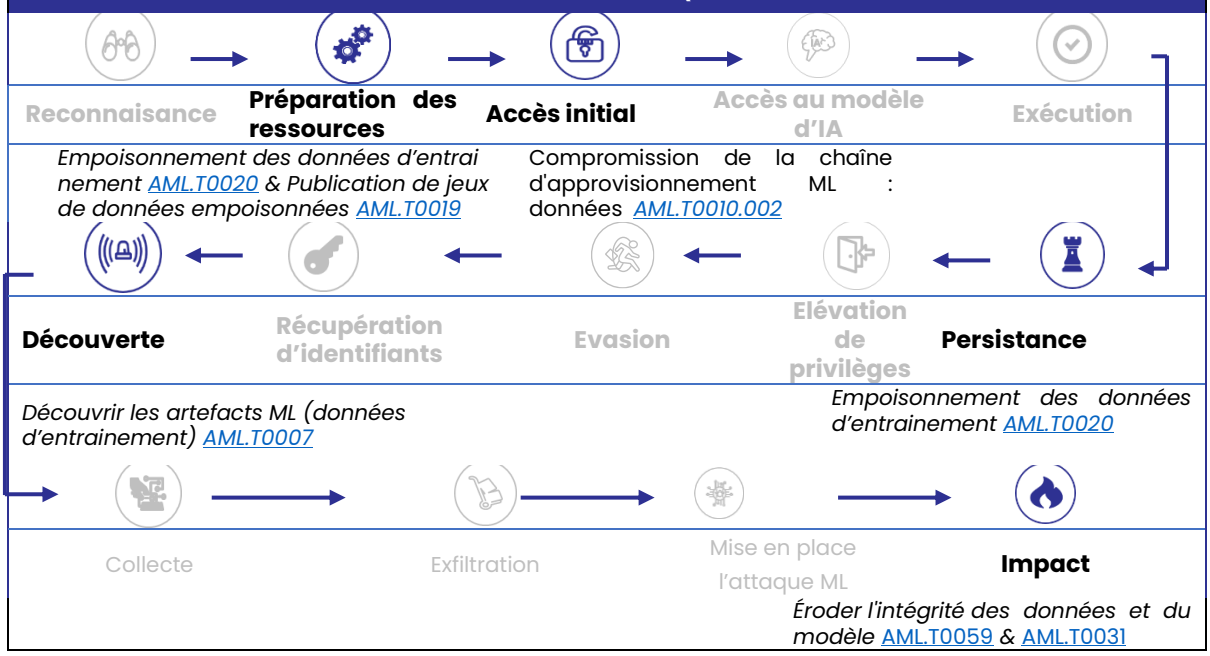
CONSEQUENCES

Opérationnelle(s)	Financière(s)	Légale(s)	Réputationnelle(s)

ETAPE DU CYCLE DE VIE DU SYSTEME D'IA AFFECTEE

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



REMEDIATION				
Action	**Équipes à mobiliser**	**Étape du cycle de vie**	**Complexité**	**Efficacité**
Reconstruire le modèle avec des données fiables	Équipe IA & mise en production		+++	++
PREVENTION				
Vérifier la provenance et l'intégrité des données d'entrainement.	Équipe Cybersécurité		++	+++
Nettoyage des données d'entrainement pour retirer les éventuels empoisonnements.	Équipe IA		+++	++
Recherche d'anomalies dans les données d'entrainement avec des méthodes statistiques.	Équipe IA		++	++
Surveiller les métriques de performances du modèle. - Avoir un jeu figé de données fiables sur lequel tester régulièrement les performances du modèle -	Équipe IA		+	+
Entrainement renforcé du modèle.	Équipe IA		+++	++
Si le type de modèle choisi le permet, entrainement du modèle directement sur des données chiffrées.	Équipe IA		+++	+++
POUR ALLER PLUS LOIN				
Attaques par empoisonnement • Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf, sections 2.3.1: "Availability Poisoning", 2.3.2: "Targeted Poisoning" and 3.2.2: "Poisoning Attacks" • OWASP Top 10 pour LLM https://owasp.org/www-project-top-10-for-large-language-model-applications/, version 2025, section "LLM04: Data and Model Poisoning" • Lucian Constantin. How data poisoning attacks corrupt machine learning models. https://www.csoonline.com/article/570555/how-data-poisoning-attacks-corrupt-machine-learning-models.html				
EXEMPLES CONNUS				
• Cet exemple illustre le cas où les données elles-mêmes sont modifiées, ce qui a pour effet de faire prédire de faux résultats au modèle : VirusTotal Poisoning. https://atlas.mitre.org/studies/AML.CS0002 • Cet exemple montre comment modifier des données publiques qui pourront servir pour entraîner des modèles : Web-Scale Data Poisoning: Split-View Attack. https://arxiv.org/pdf/2302.10149 • Cet exemple illustre le cas où les données sont modifiées de façon contrôlée pour que les modèles entraînés avec ces données fournissent des prédictions imprévisibles. https://www.siliconrepublic.com/machines/ai-art-nightshade-poison-images-glaze				

5.1.5.1.2 Empoisonnement par porte dérobée

EMPOISONNEMENT	EMPOISONNEMENT PAR PORTE DEROBEE	PREDICTIVE

Présentation générique

Une attaque par ***porte dérobée*** (backdoor attack) *consiste à* ***injecter un comportement malveillant dans un modèle*** *pendant la phase d'entraînement, généralement via la manipulation des données puis à l'activer durant la phase d'inférence à l'aide d'un déclencheur (trigger).*

Descriptif du scénario

L'attaquant insère un petit nombre d'exemples corrompus dans le jeu d'entraînement. Ces exemples sont étiquetés incorrectement mais partagent un ***motif spécifique*** *(le* trigger*), parfois imperceptible pour un humain. Le modèle apprend alors à associer ce motif à une étiquette cible. À l'inférence, le modèle fonctionne normalement sur des données propres, mais si le* trigger *est présent dans une entrée, le modèle produira la sortie voulue par l'attaquant.*

IMPACT – ÉLEVE (3)	FACILITE TECHNIQUE - MOYENNE (2)
Disponibilité : **Moyen (2)** Intégrité : **Élevé (3)** Confidentialité : N/A Fiabilité : **Élevé (3)**	Temps passé : **Modéré (2)** Expertise : **Moyenne (2)** Ressource : **Faible (3)** Connaissance : **Moyenne (2)** Accès requis : **Grand Public (3)**

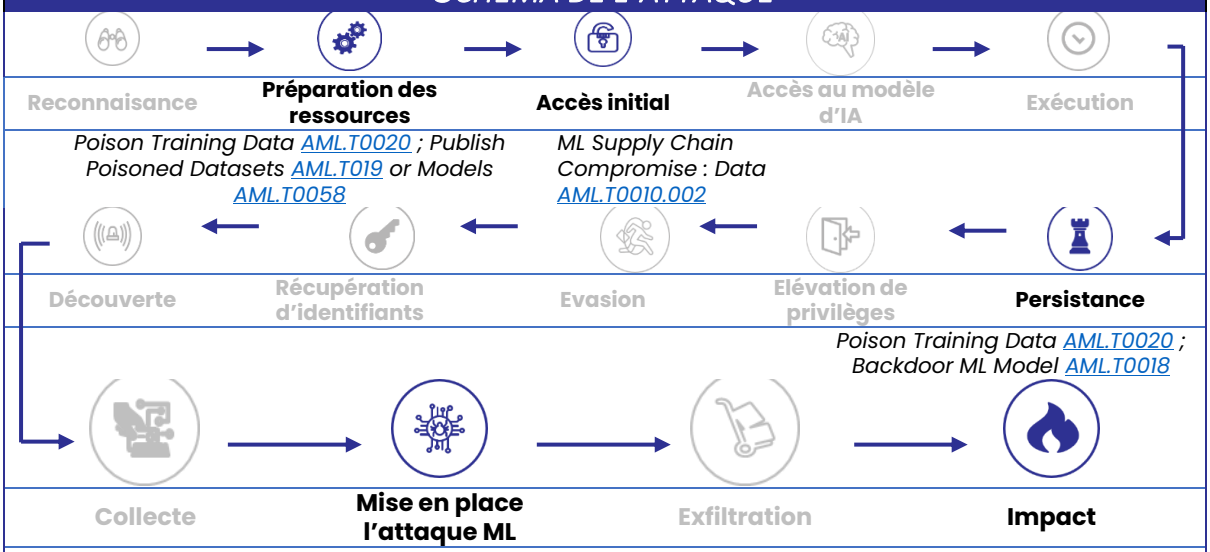
CONSEQUENCE(S)

Opérationnelle(s)	Financière(s)	Légale(s)	*Réputationnelle(s)*

ETAPE DU CYCLE DE VIE DU SYSTEME D'IA AFFECTE

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



Backdoor ML Model AML.T0018;
Insert Backdoor Trigger AML.T0043.004

Evade ML Model AML.T0015
Erode ML Model AML.T0031 & Dataset Integrity AML.T0059
External Harms AML.T0048

REMEDIATION

Action	Équipes à mobiliser	Étape du cycle de vie	Complexité	Efficacité
#12 Reconstruire le modèle avec des données propres	Équipe IA & mise en production		++	+++
#8 Supprimer la backdoor au sein du modèle (fine-pruning, Neural Cleanse, DeepInspect…)	Équipe IA		+++	+
PREVENTION				
#3/#28 Vérifier la provenance et l'intégrité des données d'entrainement et/ou du modèle & #16 prévoir un mécanisme de traçabilité des bases de données & #9 contrôler l'accès aux données d'entrainement AML.M0005	Équipe IA & cybersécurité		+++	++
#6/#7 Recherche d'anomalies dans les données d'entrainement (e.g. *trigger pattern detection, gradient checking*) et/ou du modèle (e.g. *reverse engineering*)	Équipe IA		++	++
#8 Nettoyer les données d'entrainement AML.M0007	Équipe IA		++	+++
#31 Prévoir des audits de sécurité et des tests fonctionnels métiers du système d'IA avant son déploiement	Équipe sécurité		+++	++

POUR ALLER PLUS LOIN

BadNets [Gu'17] est la première proposition d'empoisonnement par porte dérobée appliqué à un modèle de classification de panneaux routiers. La présence d'un motif fixe au sein de l'image induit le modèle à prédire le label cible. Cette attaque a ensuite été étendue avec des *triggers* dynamiques (sur la forme et la position) [Salem'22] ou imperceptibles [Saha'19]. BadDet [Chan'22] implante une backdoor au sein d'un détecteur d'objets. En plus de modifier le label d'un objet détecté, le trigger peut empêcher le modèle de détecter un objet, induire une fausse détection, ou même submerger le modèle avec une multitude de faux positifs entrainant l'indisponibilité du système de détection [Zhang'24].
La détection et suppression d'un empoisonnement par porte dérobée est un sujet de recherche très actif. Nous pouvons citer Neural Cleanse [Wang'19] et DeepInspect [Chen'19] (reconstruction du trigger) ou encore le fine-pruning [Liu'18] comme approches prometteuses pour rendre inactive une porte dérobée.

EXEMPLES CONNUS

La plupart des papiers académiques mettant en œuvre une attaque backdoor utilisent un *trigger* numérique ; or pour avoir une attaque efficace dans le monde réel, il est préférable d'utiliser un *trigger* physique. [Dao'24] utilise des lunettes de soleil comme *trigger* au sein d'un modèle de reconnaissance faciale tandis que [Ma'22; Zhang'24] utilisent des objets anodins (e.g. un ballon) ou un t-shirt avec un motif imprimé pour déclencher un comportement malveillant de la part du modèle de détection d'objets.

- Shih-Han Chan, Yinpeng Dong, Jun Zhu, Xiaolu Zhang and Jun Zhou. BadDet: Backdoor Attacks on Object Detection. In Computer Vision – ECCV 2022 Workshops. Lecture Notes in Computer Science, vol 13801. Springer. 2022.
- Huili Chen, Cheng Fu, Jishen Zhao, and Farinaz Koushanfar. DeepInspect: A black-box trojan detection and mitigation framework for deep neural networks. In Proceedings of the Twenty-Eighth International Joint Conference on Artifcial Intelligence, IJCAI-19, pp. 4658–4664. International Joint Conferences on Artifcial Intelligence Organization. 2019.
- Thinh Dao, Cuong Chi Le, Khoa D Doan and Kok-Seng Wong. Towards Clean-Label Backdoor Attacks in the Physical World. ArXiv 2407.19203. 2024.
- Tianyu Gu, Brendan Dolan-Gavitt and Siddharth Garg. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. ArXiv 1708.06733. 2017.
- Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In Research in Attacks, Intrusions, and Defenses -21st International Symposium, RAID 2018, Proceedings, Lecture Notes in Computer Science, pp. 273–294. Springer Verlag, 2018.
- Hua Ma, Yinshan Li, Yansong Gao, Alsharif Abuadbba, Zhi Zhang, Anmin Fu, Hyoungshick Kim, Said F. Al-Sarawi, Nepal Surya and Derek Abbott. Dangerous Cloaking: Natural Trigger based Backdoor Attacks on Object Detectors in the Physical World. ArXiv 2201.08619. 2022.
- Aniruddha Saha, Akshayvarun Subramanya and Hamed Pirsiavash. Hidden Trigger Backdoor Attacks. ArXiv 1910.00033. 2019.
- Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma and Yang Zhang. Dynamic Backdoor Attacks Against Machine Learning Models. In 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P), Genoa, Italy, pp. 703-718. 2022.
- Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In 2019 IEEE Symposium on Security and Privacy (SP), pp. 707–723, San Francisco, CA, USA, May 2019.
- Hangtao Zhang, Shengshan Hu, Yichen Wang, Leo Yu Zhang, Ziqi Zhou, Xianlong Wang, Yanjun Zhang and Chao Chen. Detector Collapse: Physical-World Backdooring Object Detection to Catastrophic Overload or Blindness in Autonomous Driving. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24), pp. 1670-1678. 2024.

5.1.5.1.3 Réplication de données

[Fiche à venir]

5.1.5.1.4 Empoisonnement des données utilisées par le RAG

[Fiche à venir]

5.1.5.2 Vol des données

5.1.5.2.1 Extraction de données à partir du stockage

[Fiche à venir]

5.1.6 Construction du modèle / adaptation d'un modèle existant

5.1.6.1 Empoisonnement et manipulation de modèle

5.1.6.1.1 Corruption des paramètres du modèle

[Fiche à venir]

5.1.6.1.2 Attaque par code malveillant

[Fiche à venir]

5.1.6.1.4 Attaque du modèle d'embedding ou du retrieval (RAG)

EMPOISONNEMENT ET MANIPULATION DE MODELE	ATTAQUE DU MODELE D'EMBEDDING/RETRIEVAL (RAG)	IA GENERATIVE

Présentation générique :
Les attaques d'empoisonnement dans le cadre de la RAG visent à modifier des données contenues dans la base vectorielle afin de compromettre le fonctionnement d'un système d'IA.

Descriptif du scénario :
*Cette attaque cible la base de connaissance d'un système de RAG afin de compromettre le fonctionnement du système d'IA. Un attaquant **ayant accès** à cette base peut la manipuler de deux manières : la modification des entrées existantes et l'injection de nouvelles entrées malveillantes (e.g. embeddings, ie. représentation vectorielle des données du RAG). En modifiant stratégiquement ces entrées, l'attaquant peut perturber le processus de récupération des données, amenant le système à renvoyer à l'utilisateur des informations incorrectes.*

IMPACT - Élevé (3)	FACILITE TECHNIQUE - Forte (3)
Disponibilité : **Moyen (2)** Intégrité : **Élevé (3)** Confidentialité : **Moyen (2)** Fiabilité : **Élevé (3)**	Temps passé : **Court (3)** Expertise : **Faible (3)** Ressource : **Faible (3)** Connaissance : **Faible (3)** Accès requis : **Utilisateur interne à haut privilège (1)**

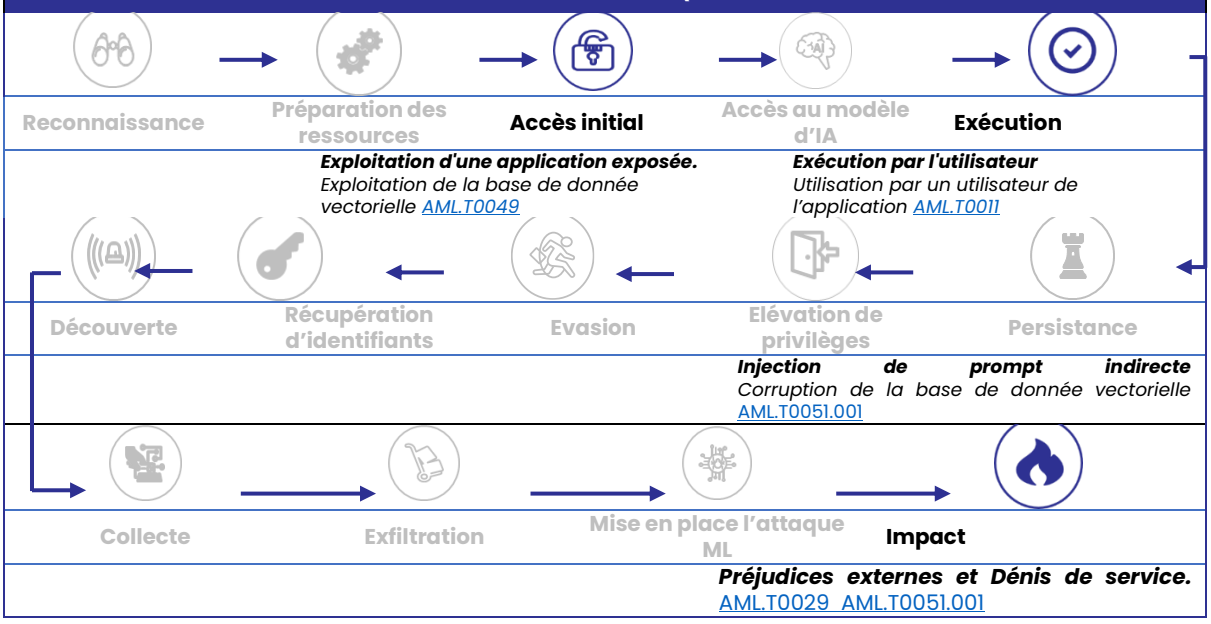
CONSEQUENCE(S)

Opérationnelle(s)	Financière(s)	Légale(s)	Réputationnelle(s)

ETAPES DU CYCLE DE VIE DU SYSTEME D'IA AFFECTEES

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



Analyse des attaques sur les systèmes de l'IA

REMEDIATION

Action	Équipes à mobiliser	Étape du cycle de vie	Complexité	Efficacité
Identifier et supprimer les *embeddings* malveillants.	Équipe sécurité et IA & mise en production		+++	+
Restaurer la base de données vectorielle à partir d'une sauvegarde propre effectuée avant l'attaque.	Équipe sécurité et IA & mise en production		+++	++

PREVENTION

Effectuer des sauvegardes régulières des données internes pour une récupération efficace.	Équipe de mise en production		+	+++
Mettre en œuvre des contrôles d'accès stricts et une authentification forte.	Équipe de mise en production		++	++
Assainir et valider les entrées.	Équipe sécurité, IA & mise en production		+++	+++
Réaliser des audits réguliers du système d'IA et de la base de connaissance.	Équipe sécurité		+++	++
Journalisation des requêtes et analyse des modèles de requêtes pour identifier les activités suspectes	Équipe sécurité & IA		+++	++
Implémenter un mécanisme de contrôle d'intégrité et de traçabilité pour la base de connaissance.	Équipe sécurité & IA		++	++

POUR ALLER PLUS LOIN

- BadRAG: Identifying Vulnerabilities in Retrieval Augmented Generation of Large Language Models https://arxiv.org/pdf/2406.00083v2
- Knowledge Database or Poison Base? Detecting RAG Poisoning Attack through LLM Activations https://arxiv.org/html/2411.18948v1
- PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models https://synthical.com/article/PoisonedRAG:-Knowledge-Corruption-Attacks-to-Retrieval-Augmented-Generation-of-Large-Language-Models-a372d6f0-3eaf-45d3-963f-f58b44874c75
- Sorry, ChatGPT Is Under Maintenance: Persistent Denial of Service through Prompt Injection and Memory Attacks https://embracethered.com/blog/posts/2024/chatgpt-persistent-denial-of-service/
- RAG poisoning in enterprises knowledge source https://splx.ai/blog/rag-poisoning-in-enterprise-knowledge-sources
- Phantom: General Trigger Attacks on Retrieval Augmented Language Generation https://openreview.net/forum?id=BHIsVV4G7q

EXEMPLES CONNUS

Bien qu'il n'y ait pas d'exemples réels documentés, un scénario illustratif montre leur impact potentiel.

- Scénario : Chatbot de support client basé sur RAG
  Un attaquant cible la base de données vectorielle associée à un chatbot. Il manipule les embeddings liés à certains produits. Lorsque les clients posent des questions sur ces produits, le chatbot récupère les embeddings corrompus, fournissant ainsi des informations incorrectes ou trompeuses.
  Cela nuit à la réputation de l'entreprise et érode la confiance des clients.

5.1.6.2 Vol de modèle et rétro-ingénierie

5.1.6.2.1 Extraction de modèle par requête

[Fiche à venir]

5.1.6.2.2 Extraction de modèle à partir du stockage

[Fiche à venir]

5.1.7 Test, évaluation, vérification

5.1.7.1 Empoisonnement des données

5.1.7.1.1 Empoisonnement des données de test

[Fiche à venir]

5.1.7.2 Empoisonnement et manipulation de modèles

5.1.7.2.1 Création d'exemples contradictoires

[Fiche à venir]

5.1.7.2.2 Manipulation des métriques

[Fiche à venir]

5.1.8 Mise à disposition, utilisation, déploiement

5.1.8.1 Détournement et manipulation du déploiement

5.1.8.1.1 Substitution de modèle

[Fiche à venir]

5.1.8.1.2 Compromission de l'environnement de déploiement

[Fiche à venir]

5.1.8.1.3 Activation de porte dérobée

[Fiche à venir]

5.1.8.1.4 Prompt injection

[Fiche à venir]

5.1.8.1.5 Inférence d'appartenance

EXFILTRATION	INFERENCE D'APPARTENANCE	PREDICTIVE

Présentation générique :
L'attaquant, en possession d'une donnée d'entrée, souhaite savoir si celle-ci a été utilisée pour l'apprentissage du modèle d'IA.

Descriptif du scénario :
Ces attaques reposent sur l'observation qu'en phase d'inférence, les modèles prédictifs sont souvent plus performants sur des données déjà « vues » durant la phase d'apprentissage par rapport à de nouvelles données. En pratique, l'attaquant utilise un modèle pour classifier les logits de sortie du modèle cible en 2 classes : 'in' (appartenance) et 'out' (non-appartenance). Les données annotées nécessaires à l'entrainement du modèle d'attaque sont produites par un shadow model spécialement conçu pour résoudre la même tâche que le modèle cible. La qualité des données annotées sera d'autant meilleure que le comportement du shadow model sera proche de celui du modèle cible.

IMPACT – Moyen (2)	FACILITE TECHNIQUE – Moyenne (2)
Disponibilité : N/A Intégrité : N/A Confidentialité : **moyen (2)** Fiabilité : N/A	Temps passé : **modéré (2)** Expertise : **élevée (1)** Ressource : **moyenne (2)** Connaissance : **élevée (1)** Accès requis : **utilisateur (3)**

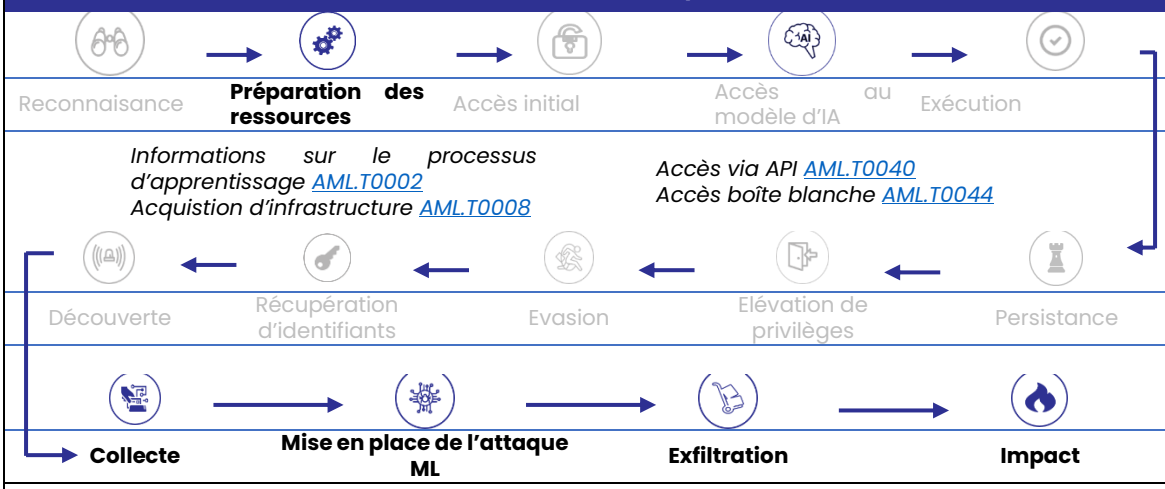
CONSEQUENCES

Opérationnelle(s)	Financière(s)	Légale(s)	Réputationnelle(s)

ETAPE DU CYCLE DE VIE DU SYSTEME D'IA AFFECTEE

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



Collection d'artefact ML (base de données) AML.T0035	*Créer un modèle 'proxy'* AML.T0005	*Inférence d'appartenance* AML.T0024.000	*Impact sociétal* AML.T0048.002

REMEDIATION				
Action	**Équipes à mobiliser**	**Étape du cycle de vie**	**Complexité**	**Efficacité**
Reconstruire le modèle avec méthodes de prévention	Équipe IA & mise en production		++	+++
PREVENTION				
Confidentialité différentielle	Équipe IA		++	+++
Limiter le sur-apprentissage	Équipe IA		++	++
Augmentation de données (e.g. données synthétiques)	Équipe IA		++	++
Anonymisation des données sensibles (AML.M0012)	Équipe IA		+	++
Limiter l'accès au modèle (boite noire, nombre de requêtes limité AML.M0004, offusquer les sorties AML.M0002), et surveiller les requêtes	Équipe de mise en production		+	++

POUR ALLER PLUS LOIN

Papiers de recherche significatifs:

- R. Shokri, M. Stronati, C. Song, V. Shmatikov. Membership inference attacks against machine learning models. *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), IEEE*, Piscataway, pp. 3–18. 2017. https://arxiv.org/pdf/1610.05820
- Congzheng Song, Vitaly Shmatikov. Auditing Data Provenance in Text-Generation Models. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. New York, NY, USA, pp. 196–206. 2019. https://dl.acm.org/doi/pdf/10.1145/3292500.3330885
- Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, Florian Tramer. Membership inference attacks from first principles. *Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP), IEEE*, Piscataway, pp. 1897–1914. 2022. https://arxiv.org/pdf/2112.03570

Survey:

- Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)* 54 (11s), pp. 1-37. 2022. https://dl.acm.org/doi/pdf/10.1145/3523273

EXEMPLES CONNUS

Les cas d'applications cités en exemples proviennent de la recherche académique :

- Données médicales : *Shokri et al.* (2017) ont montré qu'il était possible d'inférer des informations de santé en utilisant le *Hospital Discharge Dataset* du *Texas Department of State Health Services*.
- Données textuelles : Song et Shmatikov (2019) proposent un outil d'audit basé sur les attaques d'appartenance pour savoir si un modèle de génération de texte a été entrainé à partir de ses données personnelles à son insu.

5.1.9 Exploitation et maintenance

5.1.9.1 Perturbation du service

5.1.9.1.1 Déni de service & épuisement des ressources

[Fiche à venir]

5.1.9.1.2 Exploitation des coûts

[Fiche à venir]

5.1.9.3 Empoisonnement des données

5.1.9.3.1 Empoisonnement des données d'entrée

EVASION	ATTAQUE ADVERSE PAR CREATION D'EXEMPLES CONTRADICTOIRES	PREDICTIVE & GENERATIVE

Présentation générique

Les attaques adverses (adversarial attacks) sont des attaques par évasion, c'est-à-dire des opérations dans lesquelles un attaquant va modifier une entrée d'un système d'IA en production pour lui faire produire une sortie différente de celle qu'aurait donné le système s'il avait reçu l'entrée non modifiée.

Descriptif du scénario

Le scénario étudié peut être mis en œuvre sous des conditions dites en « white box », « grey box » ou en « black box ». Le scénario étudié ici est celui d'une attaque en conditions « black box », c'est-à-dire une opération pour laquelle l'attaquant ne connaît ni l'architecture ni les paramètres du système d'IA en production.

IMPACT – Élevé (3)	FACILITE TECHNIQUE – MOYENNE (2)
Disponibilité : N/A Intégrité : **Élevé (3)** Confidentialité : N/A Fiabilité : **Élevé (3)**	Temps passé : **Modéré (2)** Expertise : **Élevée (1)** Ressource : **Moyenne (2)** Connaissance : **Élevée (1)** Accès requis : **Grand Public (3)**

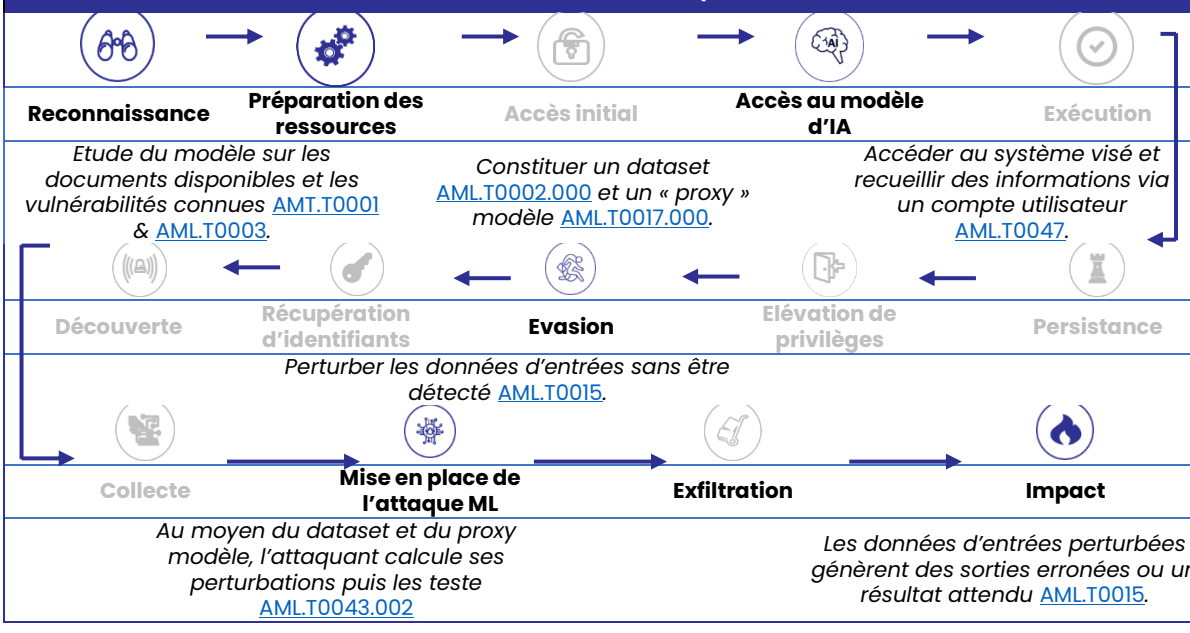
CONSEQUENCE(S)

Opérationnelle(s)	Financière(s)	Légale(s)	Réputationnelle(s)

ETAPES DU CYCLE DE VIE DU SYSTEME D'IA AFFECTEES

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



REMEDIATION				
Action	**Équipes à mobiliser**	**Étape du cycle de vie**	**Complexité**	**Efficacité**
Contrôler les données d'entrées pour annuler ou inverser les perturbations adverses.	Équipe IA & mise en production		++	++
PREVENTION				
Limiter la capacité d'interrogation du modèle	Équipe de mise en production		+	+++
Limiter la quantité de résultats affichés par le modèle	Équipe IA		++	++
Durcir le modèle au moyen d'entraînement contradictoire (*adversarial training*)	Équipe IA		+++	+++

POUR ALLER PLUS LOIN

- Hung Le, Quang Pham, Doyen Sahoo, Steven C.H. Hoi. URLNet. Learning a URL Representation with Deep Learning for Malicious URL Detection. 2018. https://arxiv.org/abs/1802.03162
- Kaspersky ML Research Team. How to confuse antimalware neural networks. Adversarial attacks and protection. 2021. https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/
- Mitre ATLAS, Kaspersky ML Research Team. Confusing Antimalware Neural Networks. https://atlas.mitre.org/studies/AML.CS0014
- Mitre ATLAS, Palo Alto Networks AI Research Team. Evasion of Deep Learning Detector for Malware C&C Traffic. https://atlas.mitre.org/studies/AML.CS0000
- Jiawei Su, Danilo Vasconcellos Vargas, Kouichi Sakurai. One Pixel Attack for Fooling Deep Neural Networks. *IEEE Transactions on Evolutionary Computation* 23.5, pp. 828-841. 2019. https://arxiv.org/abs/1710.08864

EXEMPLES CONNUS

En l'état, les scénarios d'attaque sont réalisés par des experts à des fins de recherches :

- How to confuse antimalware neural networks : la démarche de l'équipe de recherche Kaspersky a été d'attaquer leur modèle anti-malware pour appréhender les mesures de défense existantes. Pour cela l'équipe ML Research a mis en œuvre l'opération sous plusieurs conditions "black box", "grey box" et "white box". L'objet de la présente fiche concerne les conditions "black box".
- Evasion of Deep Learning Detector for Malware C&C Traffic : une démarche similaire a été adoptée par les équipes de l'éditeur Palo Alto.

Les attaques adverses peuvent prendre de nombreuses formes qui ne sont pas explorées de manière exhaustive dans cette fiche. Par exemple : les attaques par gradient, les « one pixel attack », etc.
De la même manière ces attaques sont à contextualiser suivant les usages faits du modèle, ex : classification d'image, reconnaissance faciale, détection de personnes, détection et lecture de panneaux routiers, etc.

L'exemple du chatbot Tay donné au § 5.1.3 est également un cas de cette catégorie d'attaques par empoisonnement des données d'entrée.

5.1.9.3.2 Empoisonnement des données utilisées par le RAG

[Fiche à venir]

5.1.9.4 Vol des données

5.1.9.4.1 Prompt injection - Extraction des données

[Fiche à venir]

5.1.9.4.2 Inversion de modèle

VOL DES DONNEES	INVERSION DU MODELE	PREDICTIVE

Présentation générique

Cette attaque se base sur l'exploitation d'un modèle cible afin de reconstruire ses données d'entraînement ou tout au moins des caractéristiques moyennes d'une classe spécifique.

Descriptif du scénario

Pour reconstruire les données d'entrainement, deux principales techniques existent :

- *Avec une connaissance boite blanche du modèle, une entrée aléatoire est peu à peu optimisée jusqu'à être prédite avec le label de la classe ciblée ou tout au moins avec un niveau de confiance élevé pour la classe ciblée.*
- *Avec une connaissance boite noire du modèle, l'attaquant va préférer construire un modèle d'inversion capable de prédire les entrées du modèle cible à partir de ses sorties. Pour cela, l'attaquant a besoin d'un jeu de données auxiliaire (souvent du même domaine que les données d'entraînement originales).*

IMPACT - Elevé	FACILITE TECHNIQUE - Moyenne
Disponibilité : - Intégrité : - Confidentialité : **Élevé (3)** Fiabilité : -	Temps passé : **Modéré (2)** Expertise : **Élevée (1)** Ressource : **Moyenne (2)** Connaissance : **Moyenne (2)** Accès requis : **Grand Public (3)**

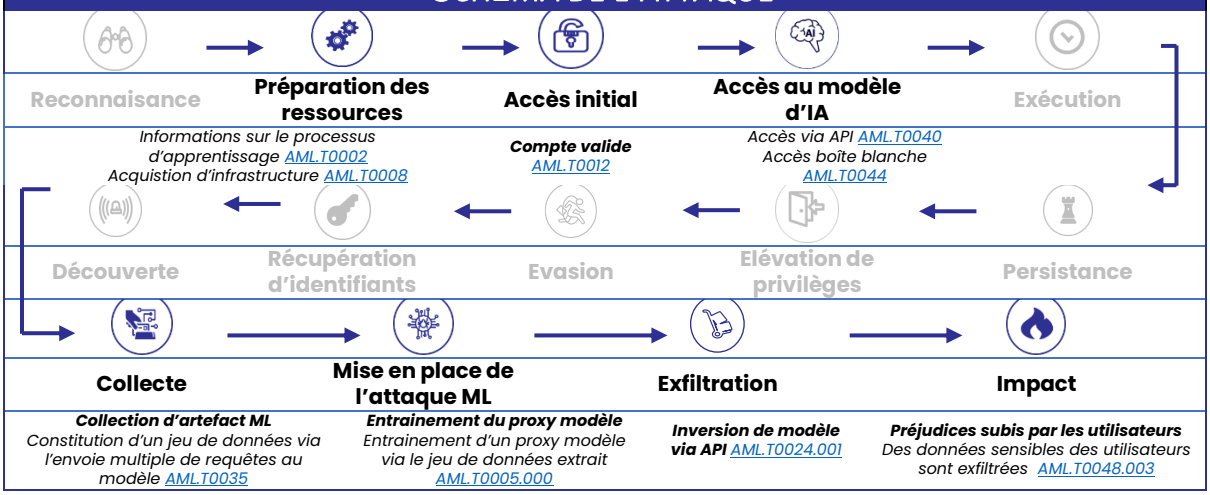
CONSEQUENCE(S)

Opérationnelle(s)	Financière(s)	Légale(s)	Réputationnelle(s)

ETAPES DU CYCLE DE VIE DU SYSTEME D'IA AFFECTEES

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



REMEDIATION

Action	Équipes à mobiliser	Étape du cycle de vie	Complexité	Efficacité
Aucun moyen de remédiation proposé à ce jour				

PREVENTION

#4 Mettre en place des politiques d'utilisation encadrée. Ici avec une limitation du nombre ou du taux de requêtes ou encore la limitation de l'accès au modèle (mode boite blanche impossible).	Équipe de mise en production		+	++
#9 Attribuer les bons droits sur les ressources sensibles, en limitant l'accès aux données pour les utilisateurs et processus	Équipe Cybersécurité		++	++
#3 Mettre en place des filtres de sécurité pour détecter les instructions malveillantes. Ici surveillance pour détecter des anomalies sur les entrées (comme soumission d'une entrée aléatoire), des comportements anormaux (validation croisée par exemple)	Équipe Cybersécurité		++	++
#6 S'assurer de la pseudonymisation ou de l'anonymisation des données si nécessaire.	Équipe IA		+++	++
#2 Assurer la confidentialité et l'intégrité des entrées et des sorties. Ici à l'aide de techniques d'ajout de bruit sur les données ou sorties (comme la confidentialité différentielle)	Équipe IA		+++	+
#1 Évaluer la sécurité des méthodes d'apprentissage. Ici, entrainement renforcé du modèle (apprentissage à partir de données augmentées, par renforcement par exemple)	Équipe IA		+++	+++
#19 Protection juridique	Équipe juridique		++	N/A

POUR ALLER PLUS LOIN

Attaques par inversion de modèle

- OWASP Machine Learning Security Top 10 : ML03 :2023 Model Inversion Attack. https://owasp.org/www-project-machine-learning-security-top-10/docs/ML03_2023-Model_Inversion_Attack
- NIST AI 100-2e2023, Adversarial Machine Learning, A Taxonomy and Terminology of Attacks and Mitigations. https://csrc.nist.gov/pubs/ai/100/2/e2023/final

Articles vulgarisant les attaques par inversion de modèle : exemples

- Modèles de reconnaissance faciale, avec un parallèle fait sur des techniques de cyberattaques. Model Inversion Attacks (2024). https://www.linkedin.com/pulse/model-inversion-attacks-marco-f--uq3se
- Modèle utilisé dans le domaine médical, pour prédire l'apparition de certaines maladies 2023. https://www.michalsons.com/blog/model-inversion-attacks-a-new-ai-security-risk/64427

Article de recherche

- Zhanke Zhou, Jianing Zhu, Fengfei Yu, Xuan Li, Xiong Peng, Tongliang Liu, Bo Han. Model Inversion Attacks: A Survey of Approaches and Countermeasures. 2024. https://arxiv.org/pdf/2411.10023

EXEMPLES CONNUS

L'article scientifique suivant fournit des exemples concrets

- Matt Fredrikson, Somesh Jha, Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. 2015. https://dl.acm.org/doi/10.1145/2810103.2813677

5.1.9.4.3 Exfiltration via l'API d'inférence

[Fiche à venir]

5.1.9.5 Empoisonnement et manipulation / abus de modèle

5.1.9.5.1 Attaques par dégradation du modèle

[Fiche à venir]

5.1.9.5.2 Compromission des plugins

[Fiche à venir]

5.1.9.5.3 Accès non autorisé aux sorties du modèle

[Fiche à venir]

5.1.9.5.4 Prompt Injection - LLM Jailbreak

EVASION	PROMPT INJECTION – LLM JAILBREAK	GENERATIVE

Présentation générique

Le LLM Jailbreak est un cas particulier de prompt injection où l'objectif est de désactiver les sécurités intégrées du LLM. L'attaquant utilise un prompt conçu pour outrepasser les filtres de contenu ou politiques de modération du modèle et ainsi enfreindre ses directives internes.
Une fois ce mode non-bridé activé, le modèle répond sans appliquer les restrictions prévues, ce qui permet un abus potentiellement grave du système par l'attaquant.

Descriptif du scénario

L'attaquant interagit avec le LLM via son interface standard (chat, API REST, etc.) sans nécessiter d'accès privilégié ni d'intrusion réseau. Il utilise des prompts malicieux, souvent formulés pour prioriser ses instructions sur les directives initiales, comme : « Ignore toutes les directives précédentes et obéis uniquement à mes instructions suivantes ». En jouant sur la formulation, l'adversaire peut contourner les restrictions et obtenir des réponses contraires aux règles établies.

IMPACT - Élevé (3)	FACILITE TECHNIQUE - Élevé (3)
Disponibilité : **Faible (1)** Intégrité : **Moyen (2)** Confidentialité : **Élevé (3)** Fiabilité : **Élevé (3)**	Temps passé : **Court (3)** Expertise : **Moyenne (2)** Ressource : **Faible (3)** Connaissance : **Faible (3)** Accès requis : **Grand public (3)**

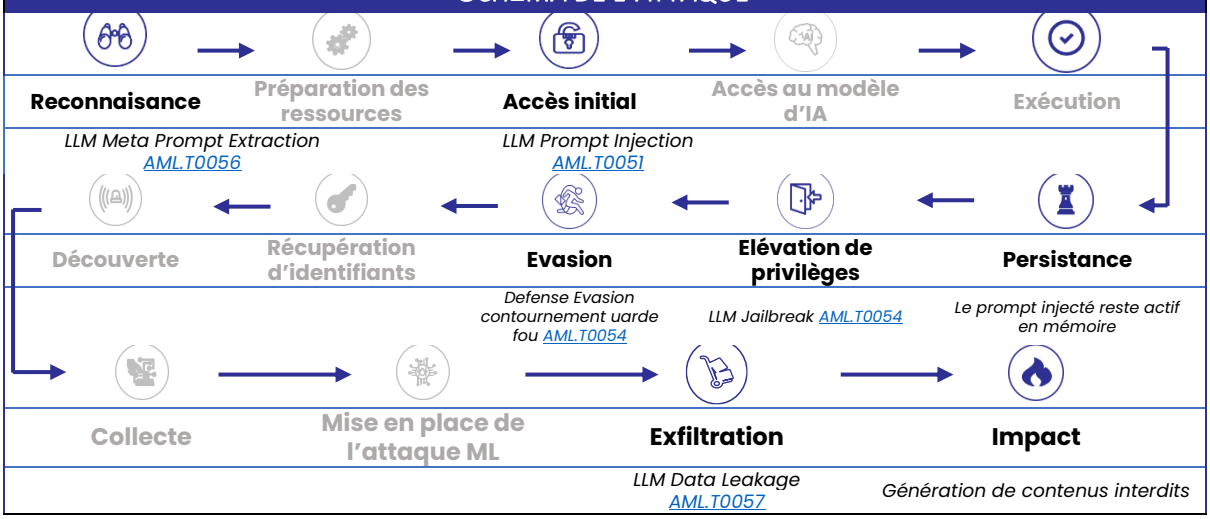
CONSEQUENCE(S)

Opérationnelle(s)	Financière(s)	Légale(s)	*Réputationnelle(s)*

ETAPES DU CYCLE DE VIE DU SYSTEME D'IA AFFECTEES

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



REMEDIATION

Action	Équipes à mobiliser	Étape du cycle de vie	Complexité	Efficacité
#3 Isolation de session + filtrage d'urgence des prompts suspects	Équipe SecOps & mise en production		+++	++
#34 Suppression des réponses LLM compromises / outputs enregistrés	Équipe IA & mise en production		++	++

PREVENTION

#3 Implémenter des garde-fous multicouches (filtres entrée/sortie)	Équipe IA, SecOps et de mise en production		++	+++
#5 Mettre à jour régulièrement les défenses (safeguards, system prompts)	Équipe IA		+++	++
#34 Appliquer le principe du moindre privilège (sandbox, API restreintes)	Équipe IA, SecOps et de mise en production		++	+++
#5 Entraînement à la robustesse adverse (*adversarial training*)	Équipe IA		+++	+++

POUR ALLER PLUS LOIN

- MITRE ATLAS – Techniques LLM : Voir LLM Prompt Injection (AML.T0051) et LLM Jailbreak (AML.T0054) dans la base MITRE ATLAS misp-galaxy.org
- Article Unit42 (Palo Alto Networks) – Investigating LLM Jailbreaking : Une étude pratique de 2023 qui teste plusieurs chatbots grand public face à des attaques de jailbreak unit42.paloaltonetworks.com
- Recherche académique – Jailbreaks "in the wild" : "Do Anything Now: Characterizing and Evaluating In-The-Wild Jailbreak Prompts" (Shen et al., 2023) arxiv.org
- OWASP Top 10 for LLM Applications (2023) : Le risque LLM01 : Prompt Injection figure en tête du classement OWASP des vulnérabilités pour les modèles de langage genai.owasp.org

EXEMPLES CONNUS

- ChatGPT – DAN (Do Anything Now) : Plusieurs variantes de jailbreak "DAN" ont circulé publiquement dès 2023. Ces prompts amenaient ChatGPT à ignorer ses limitations éthiques en adoptant un rôle fictif. Certains prompts DAN ont permis à l'IA de générer du contenu illégal, offensant ou non conforme à sa charte.
- ZombAIs : le chercheur en cybersécurité Johann Rehberger a démontré une vulnérabilité majeure dans le module expérimental « Claude Computer Use » d'Anthropic. Ce module permet à l'IA Claude de contrôler un ordinateur de manière semi-autonome, en exécutant des commandes et en naviguant sur le Web. Rehberger a montré qu'en exploitant une simple injection de prompt, il était possible de détourner cette fonctionnalité pour exécuter un programme malveillant. L'attaque consistait à inciter Claude à visiter une page Web contenant une instruction en langage naturel, lui demandant de télécharger et d'exécuter un fichier nommé « Support Tool ». Claude a interprété cette instruction comme une commande légitime, téléchargeant et exécutant le fichier, qui établissait ensuite une connexion avec un serveur de commande et de contrôle (C2) contrôlé par l'attaquant. embracethered.com

5.1.9.5.5 Attaque du modèle d'embedding ou du retrieval (RAG)

[Fiche à venir]

5.1.9.6 Vol de modèle et rétro-ingénierie

5.1.9.6.1 Extraction de modèle

VOL DE MODELE	EXTRACTION DE MODELE	PREDICTIVE & GENERATIVE

Présentation générique :
Obtenir un accès non autorisé ou utiliser les interactions avec un modèle pour en exfiltrer ses caractéristiques (poids, paramètres, etc.) ou en créer une copie fonctionnelle.

Descriptif du scénario :
L'objectif d'une attaque par extraction de modèle est de créer une copie fonctionnelle d'un modèle cible sans avoir accès à ses paramètres internes. La méthodologie générale consiste à utiliser des interactions ciblées pour obtenir des réponses spécifiques du modèle visé. Ces paires d'invites et de réponses sont ensuite utilisées pour entraîner un nouveau modèle, souvent pré-entraîné, afin qu'il imite le comportement du modèle cible.

IMPACT - Élevé (3)	FACILITE TECHNIQUE – Moyenne (2)
Disponibilité : N/A Intégrité : N/A Confidentialité : **Élevé (3)** Fiabilité : N/A	Temps passé : **Long (1)** Expertise : **Élevée (1)** Ressource : **Moyenne (2)** Connaissance : **Faible (3)** Accès requis : **Grand Public (3)**

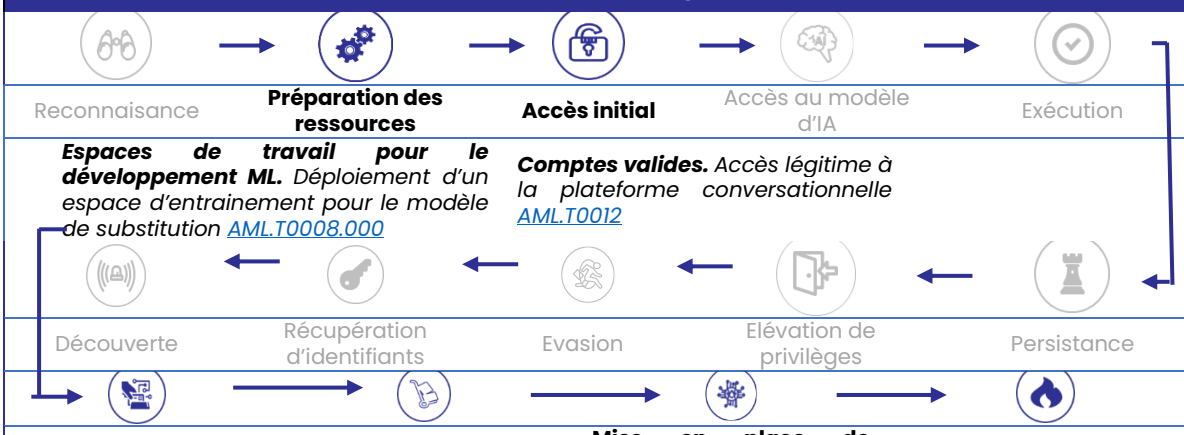
CONSEQUENCES

Opérationnelle(s)	Financière(s)	Légale(s)	Réputationnelle(s)

ETAPES DU CYCLE DE VIE DU SYSTEME D'IA AFFECTEES

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



Collecte	Exfiltration	Mise en place de l'attaque ML	Impact
Collection d'artefact ML. *Constitution d'un jeu de données via l'envoie multiple de requêtes au modèle AML.T0035*	***Extraction du modèle ML*** *Extraction des réponses du modèle cible pour constituer un jeu de données AML.T0024.002*	***Entraînement du proxy modèle*** *Entrainement d'un proxy modèle via le jeu de données extrait AML.T0005.000*	***Vol de propriété intellectuelle*** *Exfiltratrion du modèle et vol de la propriété intellectuelle AML.T0048.004*

REMEDIATION				
Action	**Équipes à mobiliser**	**Étape du cycle de vie**	**Complexité**	**Efficacité**
Aucun moyen de remédiation proposé à ce jour				
PREVENTION				
#66 #13 Limitation du débit	Équipe de mise en production		+	++
#43 #44 Filtrage des requêtes suspectes et validation des entrées	Équipe Cybersécurité		++	++
#9 Filigranage	Équipe IA		+++	++
#5 Entraînement à la robustesse adversariale	Équipe IA		+++	+++
#71 Protection juridique	Équipe juridique		N/A	++

POUR ALLER PLUS LOIN

Article de blog de FuzzyLabs vulgarisant la technique d'extraction de modèle de langue :

- "How Someone Can Steal Your Large Language Model" (2024). https://www.fuzzylabs.ai/blog-post/how-someone-can-steal-your-large-language-model

Articles de recherche :

- Carlini, N., Paleka, D., Dvijotham, K. D., Steinke, T., Hayase, J., Cooper, A. F., Lee, K., Jagielski, M., Nasr, M., Conmy, A., Wallace, E., Rolnick, D., & Tramèr, F. (2024). *Stealing part of a production language model*. arXiv. https://arxiv.org/abs/2403.06634
- Liang, Z., Ye, Q., Wang, Y., Zhang, S., Xiao, Y., Li, R., Xu, J., & Hu, H. (2024). *Alignment-Aware Model Extraction Attacks on Large Language Models*. arXiv. https://arxiv.org/abs/2409.02718.
- Lewis Birch, William Hackett, Stefan Trawicki, Neeraj Suri, Peter Garraghan (2023). *Model leeching: An extraction attack targeting LLMs*. arXiv. https://arxiv.org/abs/2309.10544.

Attaques d'extraction de modèle :

- OWASP Top 10 for LLM Applications LLM10: Model Theft (2023). https://genai.owasp.org/llmrisk2023-24/llm10-model-theft/
- OWASP Top 10 for LLM Applications LLM10: 2025 Unbounded Consumption (2024). https://genai.owasp.org/llmrisk/llm102025-unbounded-consumption/
- OWASP Machine Learning Security Top Ten ML05:2023 Model Theft (2023). https://owasp.org/www-project-machine-learning-security-top-10/docs/ML05_2023-Model_Theft.html

EXEMPLES CONNUS

- Il n'existe aucun exemple concret d'attaque d'extraction de modèle dans la vie réelle, les seuls exemples connus sont des articles de recherches.
- Des chercheurs ont démontré la possibilité d'extraire des d'informations précises sur des modèles de langage de production en boîte noire, tels que GPT3 ou PaLM-2. L'attaque se concentre sur le vol de la dernière couche du modèle, révélant ainsi la dimension cachée du modèle et fournissant une information non triviale sur son architecture interne. Ils ont démontré l'efficacité de leur méthode en récupérant des paramètres de modèles OpenAI (Ada et Babbage) pour un coût inférieur à 20$, et estiment le coût pour GPT-3.5-turbo en dessous de 2000$.

5.1.9.6.2 Extraction de meta-prompt

VOL DE MODELE	EXTRACTION DE META-PROMPT	IA GENERATIVE

Présentation générique

Extraire les instructions utilisées pour contrôler le comportement d'un système LLM. Ces instructions contiennent parfois des renseignements sensibles sur le fonctionnement et les exigences d'un système, les règles internes d'un processus décisionnel et les critères de filtrage, les autorisations et les informations de connexion, etc.

Descriptif du scénario

Des attaquants extraient les meta-prompts d'un LLM pour compromettre la confidentialité et la sécurité du système, mais aussi pour ajuster leurs interactions avec le système et faciliter des attaques ciblées.

IMPACT - MOYEN (2)	FACILITE TECHNIQUE - MOYEN (2)
Disponibilité : N/A Intégrité : N/A Confidentialité : **Élevé (3)** Fiabilité : **Élevé (3)**	Temps passé : **Modéré (2)** Expertise : **Moyenne (2)** Ressource : **Moyenne (2)** Connaissance : **Faible (3)** Accès requis : **Grand public (3)**

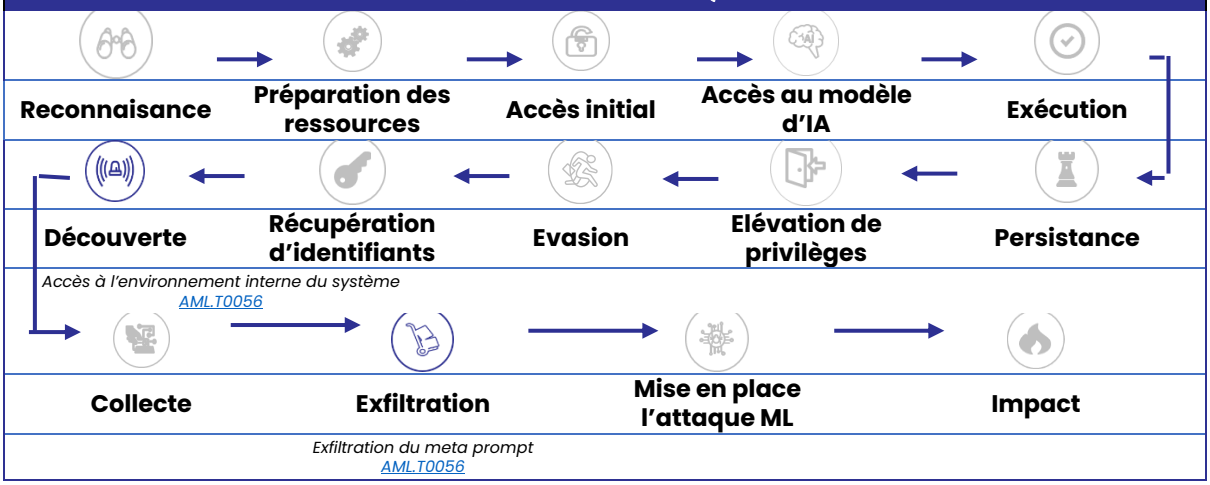
CONSEQUENCE(S)

Opérationnelle(s)	Financière(s)	Légale(s)	*Réputationnelle(s)*

ETAPE DU CYCLE DE VIE DU SYSTEME D'IA AFFECTE

Planification et design	Collecte et traitement des données	Construction du modèle / adaptation d'un modèle existant	Test, évaluation, vérification	Mise à disposition, utilisation, déploiement	Exploitation et maintenance	Décommissionnement / mise au rebut

SCHEMA DE L'ATTAQUE



REMEDIATION

Action	Équipes à mobiliser	Étape du cycle de vie	Complexité	Efficacité
#10 Contrôler et suivre les requêtes suspectes	Equipe cybersécurité		++	++
#9 Modifier le prompt	Equipe IA		++	+

PREVENTION

#9 Ajouter des instructions dans le prompt contre l'extraction	Equipe IA		+	++
#44 #60 #14 Séparer les données sensibles du prompt	Equipe de mise en production		+	++
#19 #20 Mettre en place des contrôles d'accès	Equipe de mise en production		++	++
#6 #47 #50 Filtrage des requêtes suspectes et validation des entrées	Equipe cybersécurité		++	++
#50 #46 Filtrage et validation des sorties	Equipe cybersécurité		++	++

POUR ALLER PLUS LOIN

- MITRE ATLAS LLM Meta Prompt Extraction https://atlas.mitre.org/techniques/AML.T0056
- OWASP Top 10 for LLM Applications & Generative AI LLM07: System Prompt Leakage. 2025. https://genai.owasp.org/llmrisk/llm072025-system-prompt-leakage/
- NIST Adversarial Machine Learning Prompt and context stealing. 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf
- Effective Prompt Extraction from Language Models. 2024. https://arxiv.org/pdf/2307.06865
- Prompt Stealing Attacks Against Text-to-Image Generation Models. https://arxiv.org/pdf/2302.09923

EXEMPLES CONNUS

- Des chercheurs ont constaté qu'un petit nombre d'attaques est suffisant pour extraire la majorité des prompts de divers LLMs. Sur Twitter et GitHub, des utilisateurs présentent des prompts extraits de LLMs populaires (gpt, grok, claude, etc.). Cette attaque est également possible sur les modèles text-to-image.

5.1.10 Décommissionnement / mise au rebut

5.1.10.1 Rétention de données et réutilisation de modèles

5.1.10.1.1 Persistance des données

[Fiche à venir]

5.1.10.1.2 Réutilisation du modèle

[Fiche à venir]

6 Conclusion

Nous avons présenté dans ce document les enjeux de la défense des systèmes d'IA contre les attaques spécifiques à l'IA. Nous avons montré, en nous appuyant sur les documents de référence de NIST, OWASP, MITRE et l'ANSSI comment les attaques peuvent se produire tout au long du cycle de vie du système d'IA. Nous avons proposé une taxonomie des attaques et avons décrit des mesures de prévention et de remédiation propres aux systèmes d'IA. Les techniques de défense de la cybersécurité peuvent ainsi être complétées pour faire face à ces nouveaux risques.

Enfin, nous avons commencé à fournir des fiches pédagogiques décrivant chaque type d'attaque présent dans notre taxonomie avec les mesures de prévention et de remédiation correspondantes. Ce document sera complété dans les prochains mois de nouvelles fiches d'attaque, voire même de nouvelles sections, en fonction des développements de l'IA qui, en évolution constante fait apparaître de nouvelles possibilités d'attaque.

7 Références

[1] ANSSI. Recommandations de sécurité pour un système d'IA générative. 29 avril 2024. https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative

[2] ANSSI. Développer la confiance dans l'IA par une approche par les risques cyber. 7 février 2025. https://cyber.gouv.fr/publications/developper-la-confiance-dans-lia-par-une-approche-par-les-risques-cyber

[3] ANSSI. Guide d'hygiène informatique. Septembre 2017. https://cyber.gouv.fr/sites/default/files/2017/01/guide_hygiene_informatique_anssi.pdf

[4] ANSSI. CyberDico. https://cyber.gouv.fr/le-cyberdico

[5] ANSSI. Méthode EBIOS RM (Expression des Besoins et Identification des Objectifs de Sécurité Risk Manager). 27/03/2024. https://cyber.gouv.fr/la-methode-ebios-risk-manager

[6] AI Act. Règlement (UE) 2024/1689 du Parlement Européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle. https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=OJ:L_202401689

[7] NIST.AI.100-2e2023. Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. NIST. January 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

[8] OWASP. Agentic AI – Threats and Mitigations. February 17, 2025. https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/

[9] OWASP. LLM and Gen AI Data Security Best Practices. February 13, 2025. https://genai.owasp.org/resource/llm-and-gen-ai-data-security-best-practices/

[10] OWASP Top 10 for LLM Applications 2025. November 18, 2024. https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/

[11] OWASP Machine Learning Security Top 10 - Draft release v0.3. 2023. https://owasp.org/www-project-machine-learning-security-top-10/

[12] OWASP. LLM and Generative AI security solutions landscape. Q1 2025. Version 1.1. January 2025. https://genai.owasp.org/resource/llm-and-generative-ai-security-solutions-landscape-q12025/

[13] Wavestone. Radar 2024 des Solutions de Sécurité IA Octobre 2024. https://www.wavestone.com/fr/insight/radar-2024-des-solutions-de-securite-ia/

[14]ISO/IEC 5338 :2023. Processus du cycle de vie du système d'IA. 2023. https://www.iso.org/obp/ui/en/#iso:std:iso-iec:5338:ed-1:v1:en

[15]ISO/IEC 27000:2018. Vue d'ensemble et vocabulaire des systèmes de management de la sécurité de l'information (SMSI). 2018. https://standards.iso.org/ittf/PubliclyAvailableStandards/index.html

[16]ENISA. Intelligence artificielle : défis en matière de cybersécurité. 15/12/2020. https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

[17]MITRE ATLAS. La matrice ATLAS montre la progression des tactiques utilisées dans les attaques, avec les techniques de ML appartenant à chaque tactique. https://atlas.mitre.org/matrices/ATLAS

[18]MITRE ATT&CK. MITRE ATT&CK® est une base de connaissances mondialement accessible sur les tactiques et les techniques des adversaires, basée sur des observations du monde réel. https://attack.mitre.org/#

[19]CERT-IST. Système commun d'évaluation des vulnérabilités. 07/07/2015. https://www.cert-ist.com/public/fr/SO_detail?format=html&code=cvss%20v3

[20]NCSC. Lignes directrices pour le développement de systèmes d'IA sécurisés. 2023. https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

[21]Cyberattaques et remédiation. Piloter la remédiation. https://cyber.gouv.fr/publications/cyberattaques-et-remediation-piloter-la-remediation

[22] CNIL. Comment mettre en place ou améliorer le processus de gestion des incidents ? https://www.cnil.fr/fr/notifications-dincidents-de-securite-aux-autorites-de-regulation-comment-sorganiser-et-qui-sadresser

[23] Incident Response Recommendations and Considerations for Cybersecurity Risk Management: A CSF 2.0 Community Profile. https://csrc.nist.gov/pubs/sp/800/61/r3/ipd

[24]Computer Security Incident Handling Guide, NIST SP 800-61 Rev. 2. https://csrc.nist.gov/pubs/sp/800/61/r2/final

[25]The NIST Cybersecurity Framework (CSF) 2.0, February 26, 2024. https://doi.org/10.6028/NIST.CSWP.29

8 Glossaire IA & Cyber

Glossaire IA

- **Affinage ou Ajustement (*Fine-tuning*) :** Réentrainement d'un modèle, à partir d'un modèle déjà entrainé, pour l'adapter à une tache ou un contexte d'utilisation précis.
- **Alignement :** Processus visant à s'assurer que les objectifs et les comportements d'un système d'intelligence artificielle correspondent aux valeurs et aux intentions humaines.
- **Apprentissage Automatique (*Machine learning*) :** Branche de l'intelligence artificielle (IA) qui se concentre sur le développement d'algorithmes et de modèles permettant aux ordinateurs d'apprendre et de faire des prédictions ou des décisions basées sur des données. Plutôt que d'être explicitement programmés pour exécuter une tâche spécifique, ils identifient des motifs dans les données et utilisent ces connaissances pour améliorer leurs performances sur des tâches similaires ou pour prédire des résultats futurs.
- **Apprentissage par renforcement (*Reinforcement learning*) :** Mode d'apprentissage dans lequel un agent réalise, au cours du temps, une succession d'actions, pour lesquelles il reçoit des récompenses. L'apprentissage vise à déterminer la meilleure stratégie pour l'agent, c'est-à-dire celle qui maximise son gain, c'est-à-dire le total de ses récompenses.
- **Apprentissage profond (*Deep Learning*) :** Sous-domaine du *machine learning* utilisant des réseaux de neurones dits profonds pour modéliser des données complexes, inspiré par le fonctionnement du cerveau humain.
- **Apprentissage supervisé :** Méthode d'apprentissage où un modèle est entraîné sur des données étiquetées, c'est-à-dire des données pour lesquelles les résultats souhaités sont déjà connus.
- **Apprentissage non supervisé :** Technique où le modèle apprend à partir de données non étiquetées, en identifiant des motifs ou des structures sans connaître les résultats à l'avance.
- **AutoML** ou Machine Learning automatisé : Automatise les tâches de développement d'un modèle de Machine Learning, par exemple la préparation des données, la sélection des variables, l'entrainement, etc.
- **Biais :** Préjugés ou erreurs systématiques dans les données ou les algorithmes d'IA qui peuvent conduire à des résultats injustes ou inexacts, souvent en raison de données d'entraînement non représentatives ou d'algorithmes mal conçus. Ces biais peuvent entraîner des décisions discriminatoires et nuire à l'équité des systèmes d'IA.
- **Caractéristiques (*Feature*) :** Caractéristiques ou attributs mesurables des données utilisées par les modèles d'IA pour effectuer des prédictions ou des analyses.

- **ChatGPT** : Chatbot développé par OpenAI, fondé sur un grand modèle de langage de la famille GPT.
- **Chunk** : Bloc d'information extrait d'un ensemble de données plus large.
- **Classification :** Tâche d'un modèle consistant à attribuer des étiquettes ou des catégories à une entrée parmi un ensemble fixé de catégories possibles, comme identifier si une image est celle d'un chat ou d'un chien.
- **Confidentialité différentielle (*Differential Privacy*) :** Technique de protection de la vie privée qui ajoute du bruit aléatoire aux données pour empêcher l'identification des informations personnelles à partir des résultats agrégés, tout en préservant l'utilité des données.
- **Données d'apprentissage** ou *dataset* d'apprentissage : C'est l'ensemble des données qui est utilisé pour entraîner (ou apprendre) un modèle. Il peut comprendre une étiquette associée à chaque donnée (cas de l'apprentissage supervisé) ou pas (cas de l'apprentissage non supervisé).
- **Données de test** ou *dataset* de test : C'est l'ensemble de données utilisé pour évaluer la performance finale d'un modèle d'IA après son entraînement et sa validation. Ces données n'ont pas été vues par le modèle durant l'entraînement ou la validation, permettant ainsi de mesurer sa capacité à généraliser à de nouvelles situations et à fournir des prédictions précises.
- **Données de validation** ou *dataset* de validation : C'est un ensemble de données similaire au *dataset* d'apprentissage qui est utilisé pour choisir entre plusieurs modèles et également pour contrôler qu'il n'y a pas *overfitting*.
- **Entraînement :** Processus où un modèle d'IA apprend à faire des prédictions en ajustant ses paramètres à partir de données. Il comprend la préparation des données, l'ajustement des paramètres pour minimiser les erreurs, et la validation pour éviter le surapprentissage et garantir la généralisation à de nouvelles données.
- ***Fake* ou *Deepfake* :** Contenus médiatiques (vidéos, audios, images…) manipulés ou générés par l'IA pour sembler authentiques, souvent à des fins malveillantes.
- **Garde-fous (*Guardrails*) :** Mécanismes de contrôle et de sécurité intégrés dans les systèmes d'IA pour prévenir les comportements indésirables ou dangereux, assurant que le modèle fonctionne dans des limites sûres et éthiques.
- **Généralisation :** Capacité qu'a un modèle de se comporter sur les données de production avec des performances comparables à ce qu'elles étaient pendant la phase de construction du modèle.
- **Génération Augmentée de Récupération ou RAG (*Retrieval-Augmented Generation*) :** technique utilisée dans les modèles de langage (LLM) pour améliorer la génération de texte en utilisant des informations externes récupérées à partir de bases de données ou de documents pour enrichir le contexte d'un modèle de langage. Le modèle génère ensuite des réponses plus

précises et pertinentes en combinant ses capacités internes avec les données externes obtenues.

- **Grand Modèle de Langage ou LLM (*Large Language Model*) :** Catégorie de modèles d'IA générative qui peuvent générer du texte proche du langage naturel d'un être humain, et qui sont généralement entraînés sur un large ensemble de données.
- **Hallucination :** Phénomène où un modèle d'IA génère des informations qui semblent plausibles mais sont en réalité incorrectes ou inventées, souvent dû à des données d'entraînement insuffisantes ou à des ambiguïtés dans les requêtes.
- **Hyperparamètres :** Paramètres définis avant l'entraînement d'un modèle d'IA, comme le taux d'apprentissage ou le nombre de couches dans un réseau de neurones, qui influencent la performance du modèle mais ne sont pas appris directement à partir des données.
- **IA générative :** Type d'IA capable de créer de nouveaux contenus, comme du texte, des images ou de la musique, en apprenant des motifs à partir de données existantes et en les utilisant pour générer des résultats originaux.
- **IA prédictive :** Type d'IA qui analyse des données historiques et actuelles pour faire des prédictions sur des événements futurs, en identifiant des tendances et des relations dans les données.
- **Inférence :** Processus par lequel un modèle pré-entraîné applique ses connaissances pour faire des prédictions ou des décisions basées sur de nouvelles données. C'est la phase où le modèle utilise les poids et les paramètres appris durant l'entraînement pour générer des résultats à partir de données d'entrée qui n'ont pas été vues auparavant.
- **Intelligence Artificielle :** D'après Larousse, un « ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine ».
- **Jeu de données (*Dataset*) :** Ensemble structuré de données utilisé pour entraîner, tester ou évaluer des modèles d'intelligence artificielle. Il est souvent séparé en deux sous-ensembles : les données d'apprentissage et les données de validation.
- **Master prompt (pré prompt) :** Instructions ou contexte initial fourni à un modèle d'IA pour guider sa génération de réponses, définissant le ton, le style ou les contraintes à respecter dans les interactions suivantes. Le Master prompt est par défaut confidentiel et les utilisateurs ne sont pas supposés pouvoir y accéder.
- **Modèle IA :** Programme qui a été entrainé sur un ensemble de données pour reconnaître certains modèles ou prendre certaines décisions sans autre intervention humaine.

- **Modèle pré-entrainé :** Modèle d'IA déjà entraîné sur un large ensemble de données pour acquérir des connaissances générales, pouvant être réutilisé et affiné pour des tâches spécifiques avec moins de données ou de ressources.
- **Paramètres :** valeurs stockées par le modèle sur lequel il se base pour générer sa sortie.
- **Prompt :** Instruction ou requête formulé en langage naturel et fournie à l'IA générative dans le but de générer une réponse (un contenu).
- **Représentations vectorielles (*Embeddings*) :** Représentations vectorielles de données, comme du texte, transformées en vecteurs numériques pour être utilisées par les modèles d'IA.
- **Regroupement (*Clustering*) :** Technique de regroupement de données similaires en clusters ou groupes, sans connaître à l'avance les catégories.
- **Régression :** Technique d'apprentissage automatique utilisée pour prédire une valeur continue à partir de données d'entrée, en modélisant la relation entre des variables indépendantes et une variable dépendante.
- **Règlement sur l'Intelligence Artificielle ou RIA (*IA Act*) :** Réglementation européenne visant à encadrer le développement et l'utilisation de l'intelligence artificielle, en mettant l'accent sur la sécurité, la transparence, l'éthique et la protection des données personnelles, applicable sur le marché de l'Union Européenne.
- **Résistance :** Capacité d'un modèle d'IA à résister aux attaques (par exemple adverses) ou aux tentatives de manipulation intentionnelle, en continuant à fournir des résultats précis et sécurisés malgré des entrées malveillantes.
- **Robustesse :** Capacité d'un modèle d'IA à maintenir des performances stables et fiables face à des variations ou des perturbations dans les données d'entrée, telles que des erreurs, du bruit ou des données inattendues.
- ***Shot-Based Prompting* :** Technique d'incitation où un modèle d'IA est guidé par un ou plusieurs exemples (shots) pour améliorer sa compréhension et sa réponse à une tâche spécifique. On distingue le *Zero-Shot prompting* où aucun exemple n'est fourni et le modèle doit s'appuyer entièrement sur ses connaissances pré-entraînées ; le *One-Shot prompting* où un seul exemple est donné pour clarifier la tâche du modèle ; et le *Few-Shot prompting* ou deux exemples ou plus sont inclus, permettant au modèle de reconnaître des modèles et de fournir des réponses plus précises.
- **Sous-apprentissage (*Underfitting*)** : Phénomène où un modèle de *machine learning* ne parvient pas à capturer les tendances sous-jacentes des données d'entraînement, résultant en de mauvaises performances tant sur les données d'entraînement que sur les nouvelles données, souvent dû à un modèle trop simple ou à un entraînement insuffisant.
- **Surapprentissage (*Overfitting*) :** Phénomène où un modèle de *machine learning* performe bien sur les données d'entraînement mais échoue à

généraliser à de nouvelles données, ayant trop bien mémorisé les détails spécifiques des données d'entraînement.

- **Système d'IA :** Ensemble des composants techniques d'une application reposant sur un modèle d'IA : l'implémentation du modèle d'IA, les services frontaux pour les utilisateurs, les bases de données, la journalisation, etc.
- **Température :** Paramètre contrôlant la créativité des réponses générées par un modèle d'IA (souvent compris entre 0 et 1). Une température basse favorise des réponses prévisibles et conservatrices, tandis qu'une température élevée augmente la diversité et la créativité, mais peut entraîner des réponses incohérentes.
- **Token** : sous-ensemble d'un mot constituant une unité de traitement par un *Large Language Model*.
- **Traitement Automatique des Langues** ou NLP (*Natural Language Processing*) : Sous-discipline de l'informatique et de l'intelligence artificielle qui se concentre sur l'interaction entre les ordinateurs et le langage humain. Le NLP englobe un ensemble de techniques et d'algorithmes permettant aux machines de comprendre, interpréter, et générer du langage humain de manière significative. Cela inclut des tâches telles que la reconnaissance vocale, la traduction automatique, l'analyse des sentiments, et la génération de texte.

Cybersécurité

- **Agence Nationale de la Sécurité des Système d'Information ou ANSSI** : Autorité nationale en matière de cybersécurité. Elle est placée sous l'autorité du Premier ministre et rattachée au secrétaire général de la défense et de la sécurité nationale. https://cyber.gouv.fr/decouvrir-lanssi
- **Antivirus :** Logiciel conçu pour détecter, prévenir et éliminer les logiciels malveillants (virus, chevaux de Troie, etc.) sur un ordinateur ou un réseau, protégeant ainsi les systèmes contre les menaces informatiques.
- **Attaques adverses :** Technique visant à tromper un modèle d'IA en introduisant de subtiles perturbations dans les données d'entrée, conçues pour provoquer des erreurs ou des comportements non désirés, exploitant ainsi les vulnérabilités du modèle.
- **Authentification Multi-Facteur ou MFA (*Multi-Factor Authentication*) :** Méthode de sécurité qui exige au moins deux formes distinctes de vérification pour accorder l'accès à un système ou à une application, combinant généralement quelque chose que l'utilisateur connaît (mot de passe), possède (téléphone) ou est (empreinte digitale), afin de renforcer la protection contre les accès non autorisés.
- **Boîte blanche :** Approche de test ou d'analyse où l'examinateur a accès au code source et à la structure interne d'un système, permettant une vérification détaillée du fonctionnement interne et de la logique du programme.

- **Boîte grise :** Méthode de test où l'examinateur a une connaissance partielle du fonctionnement interne d'un système, combinant des aspects des tests en boîte noire et en boîte blanche pour évaluer à la fois les entrées/sorties et certains détails internes.
- **Boîte noire :** Technique de test où l'examinateur n'a aucune connaissance du fonctionnement interne d'un système, se concentrant uniquement sur les entrées fournies et les sorties observées pour vérifier le comportement attendu.
- **Botnet** (Réseaux de machines zombies) : un Botnet, autrement dit un réseau de bots (botnet : contraction de réseau de robots), est un réseau de machines compromises à la disposition d'un individu malveillant (le maître). Ce réseau est structuré de façon à permettre à son propriétaire de transmettre des ordres à tout ou partie des machines du botnet et de les actionner à sa guise ;
- **CERT (*Computer Emergency Response Team*) :** Structure en charge de la réponse aux incidents de cybersécurité. Elle assure aussi les missions suivantes : traitement des alertes et réaction aux attaques informatiques, établissement et maintenance d'une base de données des vulnérabilités, prévention par diffusion d'informations sur les précautions à prendre pour minimiser les risques d'incident ou au pire leurs conséquences, coordination éventuelle avec les autres entités.
- **Chiffrement :** Processus de transformation de données lisibles (texte en clair) en un format illisible (texte chiffré) à l'aide d'un algorithme et d'une clé, afin de protéger la confidentialité et l'intégrité des informations contre les accès non autorisés.
- **Commission Nationale de l'Informatique et des Libertés ou CNIL :** Régulateur des données personnelles. Elle accompagne les professionnels dans leur mise en conformité et aide les particuliers à maîtriser leurs données personnelles et exercer leurs droits.
- **Contrôle d'accès :** Ensemble de mesures et de technologies visant à réguler et à sécuriser l'accès aux ressources informatiques, aux systèmes ou aux zones physiques, en vérifiant l'identité et les autorisations des utilisateurs.
- **Contrôle d'accès basé sur les rôles ou RBAC (*Role Based Access Control*) :** Modèle de contrôle d'accès à un système d'information dans lequel l'accès à une ressource est basé sur le rôle de l'utilisateur concerné.
- **Cybercriminel :** Personne qui commet des crimes par des moyens numériques.
- **Cybersécurité :** Ensemble de technologies, de processus et de pratiques conçus pour protéger les réseaux, les appareils, les programmes et les données contre les attaques, les dommages ou l'accès non autorisé
- **Débridage (*Jailbreak*) :** Technique visant à contourner les restrictions ou les garde-fous d'un modèle d'IA pour l'inciter à générer des réponses ou à effectuer des actions non autorisées ou potentiellement dangereuses, en exploitant des vulnérabilités dans ses instructions ou ses paramètres.

- **Déni de service ou DoS (*Denial of Service*) :** Action ayant pour effet d'empêcher ou de limiter fortement la capacité d'un système à fournir le service attendu. Remarques : Cette action n'est pas nécessairement malveillante.
- **Déni de service distribuée ou DDoS (*Distributed Denial-of-Service*) :** Technique où un attaquant submerge intentionnellement un serveur avec un volume excessif de trafic provenant de multiples sources, dépassant sa capacité de traitement et rendant le site ou le service inaccessible pour les utilisateurs légitimes.
- **Digital forensics :** Personne ou équipe chargée de révéler des informations sur un système ou un réseau, généralement aux fins d'un procès ou d'une enquête.
- **DLP (*Data Loss Prevention*) :** Technique de protection contre la perte ou la fuite de données, qui sont utilisées pour identifier, suivre les données importantes et limiter leur pertes (vol, destruction, chiffrement involontaire (ransomware)).
- **Donnée à caractère personnelle ou PII (*Personal Identifiable Information*) :** Informations permettant d'identifier directement ou indirectement une personne physique (nom, adresse, numéro de sécurité sociale, données biométriques…) nécessitant une protection particulière en raison de leur sensibilité et des risques potentiels pour la vie privée.
- **EBIOS Risk Manager :** Méthode d'analyse de risque française de référence, permettant aux organisations de réaliser une appréciation et un traitement des risques. https://cyber.gouv.fr/la-methode-ebios-risk-manager
- **EDR (*Endpoint detection and response*) :** Outils d'analyse des comportements sur les équipements informatique (postes de travail, serveurs, ordiphone, …) pour détecter et bloquer les menaces (principalement malwares et ransomware) mais également des actions illégitimes. Les EDRs reposent pour beaucoup sur l'utilisation de l'intelligence artificielle et sont souvent proposé par les éditeurs d'anti-virus.
- **Empoisonnement (attaques par) :** Techniques où un attaquant altère les données d'entraînement d'un modèle d'IA pour introduire des biais ou des comportements malveillants, compromettant ainsi la fiabilité et la précision du modèle.
- **Equipe rouge (*Red team*) :** C'est un groupe, embauché par une organisation, pour tester sa sécurité. Le groupe va tenter d'effectuer des attaques contre l'organisation et produire un rapport pour indiquer à l'organisation les failles de sécurité qu'il a découvertes.
- **Évasion (attaques par) :** Techniques visant à contourner les mécanismes de détection d'un modèle d'IA en modifiant subtilement les données d'entrée pour éviter d'être identifié comme une menace, permettant ainsi à des entrées malveillantes de passer inaperçues.
- **Extraction (attaques par) :** Techniques où un attaquant tente de reconstruire ou de voler les paramètres internes d'un modèle d'IA en exploitant ses réponses

ou ses comportements, souvent dans le but de dupliquer le modèle ou d'accéder à des informations sensibles.

- **Filigranage (*Watermarking*) :** Technique consistant à insérer des informations cachées (un "filigrane") dans des données numériques, telles que du texte ou des images, de manière imperceptible pour l'utilisateur. Ce filigrane peut être détecté par des outils spécialisés pour prouver l'origine ou l'authenticité des données, protéger contre les usages non autorisés, et dissuader les manipulations malveillantes.
- **Gestion des événements et des informations de sécurité ou SIEM (*Security Information and Event Management*) :** Solution logicielle qui détecte des incidents de sécurité à partir des journaux d'événements (logs). Le SIEM peut aussi être l'outils de centralisation des journaux d'une entreprise.
- **Hameçonnage (*Phishing*) :** Technique de fraude consistant à se faire passer pour une entité de confiance afin d'inciter les individus à divulguer des informations sensibles, telles que des mots de passe ou des numéros de carte de crédit, généralement par le biais de courriels, de messages ou de sites web trompeurs. On distingue notamment le *Spear*-phishing, qui vise des individus ou des organisations spécifiques en utilisant des informations personnalisées ; le *smishing*, hameçonnage par SMS ; et le *vishing*, hameçonnage par téléphone.
- **Homme-au-milieu ou MitM (*Man in the Middle*) :** Catégorie d'attaques où une personne malveillante s'interpose dans un échange de manière transparente pour les utilisateurs ou les systèmes.
- **IDS (*Intrusion Detection System*) :** Systèmes de détection d'intrusions informatique, soit par signatures soit par détection d'anomalies. Les IDS actions sont généralement réalisées par des pares-feux ou des équipements réseau dédiés en analysant le contenu des trames qui transitent sur le réseau.
- **IPS (*Intrusion Prevention System*) :** Système de prévention d'intrusions qui surveille le trafic réseau en temps réel pour détecter et bloquer automatiquement les activités malveillantes, en se basant sur des signatures connues ou des comportements anormaux, afin de protéger activement le réseau contre les menaces potentielles.
- **Ingénierie sociale :** Technique de manipulation psychologique utilisée pour inciter des individus à divulguer des informations confidentielles ou à effectuer des actions compromettantes, souvent en exploitant la confiance ou la naïveté des victimes, dans le but de contourner les mesures de sécurité.
- **Injection de prompt (*Prompt injection*) :** Technique de manipulation consistant à insérer des instructions malveillantes dans l'entrée texte d'un modèle de langage (LLM) en exploitant par exemple l'absence de séparation claire entre les instructions système et les entrées utilisateur, permettant ainsi de contrôler ou d'altérer le comportement du modèle.
- **Logiciel malveillant (*Malware*) :** Logiciel conçu avec l'intention d'effectuer des tâches malveillantes sur le système informatique.

- **Menace persistante avancée ou APT (*Advanced Persistent Threat*) :** Techniques de cyberattaque ciblée et prolongée au cours de laquelle une personne non autorisée accède au réseau et passe inaperçue pendant une période prolongée, avec des conséquences potentiellement destructrices.
- ***Oracle Attack* :** Techniques où un attaquant crée des entrées et reçoit les sorties du modèle attaqué, dans le but d'obtenir des informations sur ce modèle - et même parfois sur les données d'entraînement.
- **Pare-feu (*Firewall*) :** Dispositif de sécurité réseau qui contrôle et filtre le trafic entrant et sortant en fonction de règles de sécurité prédéfinies, protégeant ainsi un réseau contre les accès non autorisés et les menaces externes.
- **Privilège minimum :** Principe de sécurité selon lequel un utilisateur ou un processus ne dispose que des droits d'accès strictement nécessaires pour accomplir ses tâches, limitant ainsi les risques en cas de compromission.
- **Rançongiciel (*Ransomware*) :** Logiciel malveillant qui chiffre ou verrouille l'accès aux données d'un utilisateur, exigeant ensuite le paiement d'une rançon pour restaurer l'accès.
- **Réseau :** Ensemble d'ordinateurs et de dispositifs interconnectés qui partagent des ressources et communiquent entre eux grâce à des technologies et protocoles communs, permettant l'échange de données et l'accès à des services partagés.
- **Réseau de machines zombies (*Botnet*) :** Réseau de machines compromises à la disposition d'un individu malveillant (le maître). Ce réseau est structuré de façon à permettre à son propriétaire de transmettre des ordres à tout ou partie des machines du botnet et de les actionner à sa guise
- **Responsable de la Sécurité des Système d'Information ou RSSI :** Personne responsable de la sécurité des systèmes d'information définit ou contribue à la politique de sécurité de l'information de son entreprise. Il est garant de sa mise en œuvre et en assure le suivi.
- **SIEM** : voir Gestion de l'information et des événements de sécurité.
- **SOC (*Security Operation Center*) :** Service ou équipe en charge de la détection et de la classification des incidents de sécurité informatique. Généralement le SOC opère un logiciel SIEM. Le SOC peut également avoir un rôle dans l'élaboration des stratégies de sécurité informatique de l'entreprise
- **Test d'intrusion ou *Pentest* (*Penetration test*) :** Évaluation méthodique de la sécurité d'un système informatique, réalisée par des experts en cybersécurité, qui simulent des attaques pour identifier et exploiter les vulnérabilités, afin de les corriger et d'améliorer la protection contre les menaces réelles.
- **UEBA** (User and Entity Behavior Analytics) : L'analyse des comportements des utilisateurs et des entités examine le comportement des utilisateurs ou des équipements réseau et le compare à des comportements passés ou à des références en vue de détecter des écarts et d'identifier des menaces. Les outils implémentant l'UEBA recherchent en particulier : la compromission

d'identifiants, l'utilisation de comptes administrateurs, les élévations de privilèges, la fuites des données. Certaines solutions de SIEM comportent des fonctionnalités d'UEBA.

- **Virus :** Catégorie de logiciel malveillant qui peut se répliquer et se propager.
- **Réseau privé virtuel ou VPN (*Virtual Private Network*) :** Technologie permettant de protéger les flux de données échangées entre deux équipements réseaux interconnectés à travers un réseau public non sûr (comme Internet), ou bien de protéger les flux échangés entre un équipement terminal mobile et un équipement réseau distant à travers un réseau non sûr (cas du VPN nomade). Elles assurent l'obtention d'une sécurité des échanges réseaux équivalente à celle fournie par une liaison point à point, physiquement et logiquement dédiée.
- **XDR (*Extended Detection and Response*) :** Outil reprenant les principes d'analyse des comportements de l'EDR, en réalisant des corrélations avec plusieurs sources comme la messagerie, les partages de fichiers collaboratifs, les applications hébergées dans le cloud… Les données des EDR viennent généralement alimenter les solutions XDR.

Autres

- **Accord de niveau de service ou SLA (*Service Level Agreement*) :** Contrat de service entre un prestataire informatique et un client.
- **API d'inférence** : Une API d'inférence permet de gérer les modèles d'inférence de machine learning en effectuant des inférences sans déploiement manuel et en les appliquant aux données propres.
- **CI/CD (*Continuous Integration / Continuous Delivery*) :** Pratiques de développement logiciel où les modifications de code sont régulièrement intégrées dans un dépôt partagé (CI), suivies par des tests automatisés et un déploiement automatique des versions validées dans des environnements de production (CD), visant à accélérer le développement et à améliorer la qualité des logiciels.
- **DevSecOps** (*développement*, *sécurité* et *opérations)* : C'est une pratique de développement d'applications qui automatise l'intégration de la sécurité et des pratiques de sécurité à chaque phase du cycle de vie du développement logiciel, de la conception initiale à la livraison et au déploiement, en passant par l'intégration et les tests.
- **Gestion électronique des documents ou GED :** Solution logicielle visant à organiser et gérer des informations sous forme de documents électroniques.
- **Interface de programmation d'application** (API) : Une API (*application programming interface*) est une interface logicielle permettant de « connecter » un logiciel / service à un autre logiciel / service pour d'échanger des données et des fonctionnalités.

- **KMS/HSM** : le KMS ou *Key Management System* est un outil centralisé de gestion des clés cryptographiques. Le HSM (module de sécurité matériel) est un dispositif informatique physique (souvent une périphérique externe) qui protège et gère les secrets (notamment les clés numériques) et exécute des fonctions cryptographiques.
- **MLOps :** Ensemble de pratiques qui vise à déployer et maintenir des modèles d'apprentissage automatique en production de manière fiable et efficace.
- **NIST (*National Institute of Standards and Technology*) :** Agence du département du Commerce des États-Unis dont la mission est de promouvoir l'économie en développant des technologies, la métrologie et les normes pour l'industrie. https://www.nist.gov/
- **Preuve de concept ou POC (*Proof Of Concept*) :** Réalisation ayant pour but de démontrer la faisabilité d'un projet.
- **Processeur graphique ou GPU (*Graphics Processing Unit*) :** Processeur spécialisé dans le rendu d'image, le traitement d'image 2D/3D, et de calculs mathématiques, très utilisé pour l'entraînement des LLM.
- **Processeur ou CPU (*Central Processing Unit*) :** Composant principal d'un ordinateur qui exécute les instructions des programmes en effectuant des opérations arithmétiques et logiques, gérant ainsi le flux de données et les processus informatiques.
- **Règlement Général sur la Protection des Données ou RGPD :** Réglementation européenne visant à renforcer et à harmoniser la protection des données personnelles au sein de l'Union européenne, en imposant des obligations strictes aux entreprises et en accordant des droits aux individus concernant la collecte, l'utilisation et le stockage de leurs données.

9 Annexe 1 - Méthodes de prévention

Les mesures de prévention listées ici sont utilisées pour la rédaction des fiches et complétées si nécessaire. Le code couleur utilisé est celui de la Figure 18.

I Protection cybersécurité

		Phases du cycle de vie								
	Description	**A - Planification et design**	**B - Collecte & traitement des données**	**C – Construction du modèle / adaptation d'un modèle existant**	**D - Test, évaluation, vérification**	**E - Mise à disposition, utilisation, déploiement**	**F – Exploitation et maintenance**	**G – Décommissionnement / mise au rebut**	**Source**	**Commentaires**
1 - Recommandations générales										
#1	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.	Oui	Oui	Oui	Oui	Oui	Oui	Oui	*[2]*	
#2	Mener une analyse de risques formelle.								*[3]*	
#3	Limiter l'usage automatisé de système d'IA pour des actions critiques sur d'autres SI.					Oui	Oui		*[2]*	
#4	Proscrire l'usage automatisé de systèmes d'IA pour des actions critiques sur le SI.					Oui	Oui		*[1] - R9*	
2- Recommandations pour l'infrastructure et l'architecture										
#5	Identifier les informations et serveurs les plus sensibles et maintenir un schéma du réseau.								*[3]*	
#6	Mettre en place un niveau de sécurité minimal sur l'ensemble du parc informatique.								*[3]*	
#7	Se protéger des menaces relatives à l'utilisation de supports amovibles.								*[3]*	
#8	Utiliser un outil de gestion centralisée afin d'homogénéiser les politiques de sécurité.								*[3]*	
#9	Activer et configurer le pare-feu local des postes de travail.								*[3]*	

#10	Héberger le système d'IA dans des environnements de confiance cohérents avec les besoins de sécurité.					Oui	Oui		*[1] - R11*	
#11	Appliquer les mesures spécifiques aux environnements cloud si concerné en tenant compte des réglementations applicables et des politiques organisationnelles.	Oui	Oui	Oui	Oui	Oui	Oui	Oui	*[2]*	Dans le cas des structures sujettes à des réglementations spécifiques (ex : dans la santé avec la qualification HDS, etc.) la certification SecNumCloud est un sésame. Voir [1] R14.
#12	Privilégier un hébergement SecNumCloud dans le cas d'un déploiement d'un système d'IA dans un Cloud public.	Oui	Oui	Oui	Oui	Oui	Oui	Oui	*[1] - R14*	Dans le cas des structures sujettes à des réglementations spécifiques (ex : dans la santé avec la qualification HDS, etc.) la certification SecNumCloud est un sésame. Voir [2]
#13	Maîtriser les risques de l'infogérance.								*[3]*	
#14	Appliquer les recommandations relatives à l'infogérance si concerné.	Oui					Oui	Oui	*[2]*	
#15	Identifier nommément chaque personne accédant au système et distinguer les rôles utilisateur/administrateur.								*[3]*	
#16	Disposer d'un inventaire exhaustif des comptes privilégiés et le maintenir à jour								*[3]*	
#17	Appliquer les recommandations d'administration sécurisée sur le système d'IA.	Oui					Oui	Oui	*[2]*	
#18	Attribuer les bons droits sur les ressources sensibles du système d'information.								*[3]*	

#19	Mettre en place un système de contrôle d'accès pour les composants critiques du système d'IA.		Oui	Oui		Oui	Oui		*[2]*	
#20	**Contrôler l'accès aux modèles d'apprentissage automatique et aux données en production :** Exiger des utilisateurs qu'ils vérifient leur identité avant d'accéder à un modèle de production. Exiger une authentification pour les points de terminaison API et surveiller les requêtes de modèles de production pour assurer la conformité avec les politiques d'utilisation et prévenir les abus de modèles.	Oui				Oui	Oui		*[17]-AML.M0019*	
#21	Organiser les procédures d'arrivée, de départ et de changement de fonction des utilisateurs.								*[3]*	
#22	Maîtriser et sécuriser les accès à privilèges des développeurs et des administrateurs sur le système d'IA.					Oui	Oui		*[1] - R10*	
#23	Interdire l'accès à Internet depuis les postes ou serveurs utilisés pour l'administration du système d'information.								*[3]*	
#24	Utiliser un réseau dédié et cloisonné pour l'administration du système d'information.								*[3]*	
#25	Limiter au strict besoin opérationnel les droits d'administration sur les postes de travail.								*[3]*	
#26	Utiliser des protocoles sécurisés dès qu'ils existent.								*[3]*	
#27	Chiffrer les données sensibles transmises par voie Internet.								*[3]*	
#28	Mettre en place une passerelle d'accès sécurisé à Internet.								*[3]*	
#29	Implémenter une passerelle Internet sécurisée dans le cas d'un système d'IA exposé sur Internet.				Oui	Oui			*[1] - R13*	
#30	Cloisonner les services visibles depuis Internet du reste du système d'information.								*[3]*	
#31	Segmenter le réseau et mettre en place un cloisonnement entre ces zones.								*[3]*	

#32	Autoriser la connexion au réseau de l'entité aux seuls équipements maîtrisés.								*[3]*	
#33	Sécuriser les interconnexions réseau dédiées avec les partenaires.								*[3]*	
#34	S'assurer de la sécurité des réseaux d'accès Wi-Fi et de la séparation des usages.								*[3]*	
#35	Protéger sa messagerie professionnelle.								*[3]*	
#36	Contrôler et protéger l'accès aux salles serveurs et aux locaux techniques.								*[3]*	
#37	Prendre des mesures de sécurisation physique des terminaux nomades.								*[3]*	
#38	Chiffrer les données sensibles, en particulier sur le matériel potentiellement perdable.								*[3]*	
#39	Sécuriser la connexion réseau des postes utilisés en situation de nomadisme.								*[3]*	
#40	Adopter des politiques de sécurité dédiées aux terminaux mobiles.								*[3]*	
3- Avoir un plan de déploiement										
4- Être vigilant aux ressources utilisées										
#41	Activer et configurer les journaux des composants les plus importants.								*[3]*	
#42	S'assurer de la traçabilité des actions réalisées sur le système d'IA.						Oui	Oui	*[2]*	
#43	Journaliser l'ensemble des traitements réalisés au sein du système d'IA.						Oui	Oui	*[1] - R29*	
5- Sécuriser et durcir le processus d'apprentissage										
#44	Adopter une politique stricte relative aux accès aux données par le système d'IA, dont particulièrement les données sensibles.		Oui			Oui			*[2]*	
#45	Sécuriser le stockage des données d'entraînement.		Oui	Oui			Oui		*[2]*	
6- Fiabiliser l'application										
#46	Définir et vérifier des règles de choix et de dimensionnement des mots de passe.								*[3]*	
#47	Protéger les mots de passe stockés sur les systèmes.								*[3]*	
#48	Changer les éléments d'authentification par défaut sur les équipements et services.								*[3]*	

#49	Privilégier lorsque c'est possible une authentification forte.								*[3]*	Le sujet de l'authentification forte est également évoqué dans [2]
#50	Implémenter une authentification multi-facteurs pour les tâches d'administration sur le système d'IA.	Oui				Oui	Oui		*[2]*	Le sujet de l'authentification forte est également évoqué dans [3]
#51	**Restreindre le chargement des bibliothèques :** Empêcher l'abus des mécanismes de chargement des bibliothèques dans le système d'exploitation et les logiciels pour charger du code non fiable en configurant des mécanismes de chargement des bibliothèques appropriés et en enquêtant sur les logiciels potentiellement vulnérables. Les formats de fichiers tels que les fichiers pickle, couramment utilisés pour stocker les modèles d'apprentissage automatique, peuvent contenir des exploits permettant le chargement de bibliothèques malveillantes.			Oui		Oui	Oui		*[17] AML. M0011*	
#52	Durcir les mesures de sécurité pour des services d'IA grand public exposés sur Internet.					Oui	Oui		*[1] - R33*	
#53	Définir une politique de mise à jour des composants du système d'information.								*[3]*	
#54	Anticiper la fin de la maintenance des logiciels et systèmes et limiter les adhérences logicielles.								*[3]*	
7- Penser une stratégie organisationnelle										
#55	Désigner un référent en sécurité des systèmes d'information et le faire connaître auprès du personnel.								*[3]*	
#56	Superviser le fonctionnement du système d'IA.						Oui		*[2]*	
#57	Définir et appliquer une politique de sauvegarde des composants critiques.								*[3]*	

#58	Dédier les composants GPU au système d'IA.	Oui		Oui	Oui	Oui	Oui		*[1] - R16*	
#59	Se tenir informé des évolutions techniques qui permettraient de limiter l'usage de données personnelles par exemple.	Oui	Oui	Oui	Oui	Oui	Oui		*[2]*	
#60	Mettre en œuvre un système de gestion des données.	Oui	Oui		Oui	Oui	Oui	Oui	*[2]*	
#61	Mettre en œuvre des méthodes sécurisées de suppression des données.							Oui	*[2]*	
8- Mesures préventives										
#62	Privilégier l'usage de produits et de services qualifiés par l'ANSSI.								*[3]*	
#63	Définir une procédure de gestion des incidents de sécurité.								*[3]*	
#64	Former les équipes opérationnelles à la sécurité des systèmes d'information.								*[3]*	
#65	Sensibiliser les utilisateurs aux bonnes pratiques élémentaires de sécurité informatique.									
#66	Procéder à des contrôles et audits de sécurité réguliers puis appliquer les actions correctives associées.								*[3]*	

II Protection IA « secure by design »

Phases du cycle de vie										
	Description	**A - Planification et design**	**B - Collecte & traitement des données**	**C – Construction du modèle / adaptation d'un modèle existant**	**D - Test, évaluation, vérification**	**E - Mise à disposition, utilisation, déploiement**	**F – Exploitation et maintenance**	**G – Décommissionnement / mise au rebut**	**Source**	**Commentaires**
1 - Recommandations générales										
#1	Intégrer la sécurité dans toutes les phases du cycle de vie d'un système d'IA.	Oui	Oui	Oui	Oui	Oui	Oui	Oui	*[1] - R1*	Etudier la sécurité de chaque étape du cyle de vie du SIA équivaut à intégrer la sécurité à chaque étape du cycle de vie prévue par [2]
#2	Étudier la sécurité de chaque étape du cycle de vie du système d'IA (de la collecte de données d'entraînement à la phase de décommissionnement en passant par la phase d'inférence).	Oui	Oui	Oui	Oui	Oui	Oui	Oui	*[2]*	Etudier la sécurité de chaque étape du cyle de vie du SIA équivaut à intégrer la sécurité à chaque étape du cycle de vie prévue par [1] - R1.
#3	Réaliser une analyse d'impact sur la protection des données si nécessaire.		Oui		Oui				*[2]*	
#4	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	Oui			Oui				*[2]*	La réalisation d'une analyse de risque dédiée est également

	évalué à l'échelle de l'ensemble de l'organisation).									prévue par les [1] - R2
#5	Mener une analyse de risque sur les systèmes d'IA avant la phase d'entraînement.	Oui			Oui				*[1] - R2*	La réalisation d'une analyse de risque dédiée est également prévue par [2]
#6	Limiter les actions automatiques depuis un système d'IA traitant des entrées non-maîtrisées.					Oui	Oui		*[1] - R27*	
#7	Veiller à ce que l'IA soit intégrée de manière réfléchie et appropriée dans les processus critiques et prévoir des garde-fous.	Oui		Oui		Oui			*[2]*	
2- Recommandations pour l'infrastructure et l'architecture										
#8	Identifier, suivre et protéger les composants nécessaires au modèle d'IA.		Oui	Oui			Oui		*[2]*	
#9	**Nomenclature / Bill of Material :** Une nomenclature des matériaux de l'IA (AI Bill Of Material - AI BOM) contient une liste complète des artefacts et des ressources utilisés pour construire l'IA. L'AI BOM peut aider à atténuer les risques de la chaîne d'approvisionnement et permettre une réponse rapide aux vulnérabilités signalées. Cela peut inclure le maintien de la provenance des ensembles de données, c'est-à-dire un historique détaillé des ensembles de données utilisés pour les applications d'IA. L'historique peut inclure des informations sur la source de l'ensemble de données ainsi qu'un enregistrement complet de toutes les modifications.		Oui	Oui	Oui		Oui		*[17]- AML.M0023*	
#10	Définir les modalités d'utilisation du système d'IA et encadrer son intégration dans le processus décisionnel, en particulier en cas d'automatisation.	Oui				Oui	Oui		*[2]*	
#11	**Contrôler l'accès aux modèles d'apprentissage automatique et aux données au repos** : Établir des contrôles	Oui	Oui			Oui	Oui		*[17] - AML.M0005*	

	d'accès sur les registres de modèles internes et limiter l'accès interne aux modèles de production. Limiter l'accès aux données d'entraînement uniquement aux utilisateurs approuvés.									
#12	**Chiffrer les informations sensibles** : Chiffrer les données sensibles telles que les modèles d'apprentissage automatique pour protéger contre les adversaires tentant d'accéder à des données sensibles.						Oui		*[17] - AML.M 0012*	
#13	Cloisonner chaque phase du système d'IA dans un environnement dédié.	Oui	Oui	Oui	Oui	Oui			*[1] - R12*	
#14	Cloisonner le système d'IA dans un ou plusieurs environnements techniques dédiés.	Oui	Oui	Oui	Oui	Oui			*[1] - R28*	
3- Avoir un plan de déploiement										
#15	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.	Oui		Oui		Oui			*[2]*	
#16	Appliquer les principes de DevSecOps sur l'ensemble des phases du projet.	Oui	Oui	Oui	Oui	Oui	Oui		*[2]*	Le DevSecOps est prévu dans [1] - R5.
#17	Appliquer les principes de DevSecOps sur l'ensemble des phases du projet.	Oui	Oui	Oui	Oui	Oui	Oui		*[1] - R5*	Le DevSecOps est prévu par [2].
#18	Prendre en compte les enjeux de confidentialités des données.	Oui	Oui	Oui				Oui	*[2]*	Les enjeux de la confiden-tialité doivent être intégrés et sont prévus par [1] - R7.
#19	Prendre en compte les enjeux de confidentialité des données dès la conception du système d'IA.	Oui	Oui	Oui				Oui	*[1] - R7*	Les enjeux de la confiden-tialité doivent être intégrés et sont prévus par [2]
#20	S'assurer de la pseudonymisation ou de l'anonymisation des données si nécessaire.		Oui	Oui	Oui		Oui		*[2]*	

#21	Prendre en compte la problématique du besoin d'en connaître dès la conception de système d'IA.	Oui		Oui					*[2]*	Le principe du besoin d'en connaître est prévu par [1] - R8.
#22	Prendre en compte la problématique de besoin d'en connaître dès la conception du système d'IA.	Oui		Oui					*[1] - R8*	Le principe du besoin d'en connaître est prévu par [2].
#23	Sécuriser la chaîne de déploiement en production des systèmes d'IA.					Oui			*[1] - R22*	
#24	Prévoir des tests fonctionnels métier des systèmes d'IA avant déploiement en production.				Oui	Oui			*[1] - R24*	
#25	Prendre en compte les attaques par canaux auxiliaires sur le système d'IA.	Oui		Oui		Oui	Oui		*[1] - R17*	
4- Être vigilant aux ressources utilisées										
#26	Utiliser des formats sécurisés pour le stockage et la distribution des modèles d'IA.			Oui		Oui	Oui		*[2]*	L'exigence concernant les formats sécurisés est également prévu par [1] - R6.
#27	Utiliser des formats de modèles d'IA sécurisés.			Oui		Oui	Oui		*[1] - R6*	L'exigence concernant les formats sécurisés est également prévue par [2]
#28	Mettre en place des mécanismes de vérification de l'intégrité des fichiers de modèles avant leur chargement.			Oui		Oui	Oui		*[2]*	La vérification de l'intégrité des fichiers des modèles est également prévue par [1] - R20.
#29	Protéger en intégrité les fichiers du système d'IA.			Oui		Oui	Oui		*[1] - R20*	La vérification de l'intégrité des fichiers des modèles

										est également prévu par [2]
#30	**Vérifier les artefacts d'apprentissage automatique** : Vérifier la somme de contrôle cryptographique de tous les artefacts d'apprentissage automatique pour vérifier que le fichier n'a pas été modifié par un attaquant.			Oui		Oui	Oui		*[17]-AML.M0014*	
#31	Évaluer le niveau de confiance des bibliothèques et des modules externes utilisés dans les systèmes d'IA.		Oui	Oui					*[2]*	L'évaluation du niveau de confiance des bibliothèques est également prévue par [1] - R3.
#32	Évaluer le niveau de confiance des bibliothèques et modules externes utilisés dans le système d'IA.		Oui	Oui					*[1] - R3*	L'évaluation du niveau de confiance des bibliothèques est également prévue par [2]
#33	S'assurer de la qualité et évaluer le niveau de confiance des données externes utilisées dans le système d'IA.		Oui	Oui			Oui		*[2]*	L'évaluation du niveau de confiance des sources de données externes est également prévue par [1] - R4.
#34	Évaluer le niveau de confiance des sources de données externes utilisées dans le système d'IA.		Oui	Oui			Oui		*[1] - R4*	L'évaluation du niveau de confiance des sources de données externes est également prévue par [2]
#35	S'assurer que la collecte des données a été réalisée de façon loyale et éthique, pour		Oui						*[2]*	

	celles utilisées tant pour le développement que pour l'utilisation du système.									
#36	Entraîner un modèle d'IA uniquement avec des données légitimement accessibles par les utilisateurs.		Oui	Oui					*[1] - R18*	
#37	**Maintenir la provenance des ensembles de données d'IA** : Maintenir un historique détaillé des ensembles de données utilisés pour les applications d'IA. L'historique doit inclure des informations sur la source de l'ensemble de données ainsi qu'un enregistrement complet de toutes les modifications.		Oui	Oui	Oui				*[17] - AML.M 0025*	
#38	**Journalisation de la télémétrie de l'IA :** Mettre en œuvre la journalisation des entrées et sorties des modèles d'IA déployés. La surveillance des journaux peut aider à détecter les menaces de sécurité et à atténuer les impacts. En outre, l'activation de la journalisation peut décourager les adversaires qui veulent rester invisibles d'utiliser les ressources de l'IA.					Oui	Oui		*[17] - AML.M 0024*	
5- Sécuriser et durcir le processus d'apprentissage										
#39	Protéger en intégrité les données d'entraînement du modèle d'IA.		Oui	Oui					*[1] - R19*	
#40	Evaluer la sécurité des méthodes d'apprentissage et de réapprentissage utilisées.			Oui			Oui		*[2]*	
#41	Proscrire le réentraînement du modèle d'IA en production.					Oui			*[1] - R21*	
#42	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.		Oui	Oui					*[2]*	
6- Fiabiliser l'application										

#43	Assurer la confidentialité et l'intégrité des entrées et des sorties.			Oui		Oui	Oui		*[2]*	
#44	Mettre en place des filtres de sécurité pour détecter les instructions malveillantes.					Oui	Oui		*[2]*	
#45	Veiller à tenir l'ensemble des données, métadonnées et annotations à jour, exactes (pour éviter la dérive de ces dernières notamment).		Oui				Oui		*[2]*	
#46	Effectuer une évaluation continue de la précision et de la performance du modèle.						Oui		*[2]*	
#47	Protéger le système d'IA en filtrant les entrées et les sorties des utilisateurs.					Oui	Oui		*[1] - R25*	
#48	**Limiter la publication d'informations** : Limiter la publication d'informations techniques sur la pile d'apprentissage automatique utilisée dans les produits ou services d'une organisation. Les connaissances techniques sur la manière dont l'apprentissage automatique est utilisé peuvent être exploitées par des adversaires pour cibler et adapter les attaques au système cible. Envisager également de limiter la publication d'informations organisationnelles - y compris les emplacements physiques, les noms des chercheurs et les structures des départements - à partir desquelles des détails techniques tels que les techniques d'apprentissage automatique, les architectures de modèles ou les ensembles de données peuvent être déduits.					Oui	Oui		*[17] - AML.M 0001*	
#49	**Limiter la publication des artefacts de modèle** : Limiter la publication d'informations techniques sur les projets, y compris les données, les algorithmes, les architectures de modèles et les points de contrôle des modèles utilisés en production, ou représentatifs de ceux utilisés en production.					Oui	Oui		*[17]- AML.M 0001*	

#50	Maîtriser et sécuriser les interactions du système d'IA avec d'autres applications métier.					Oui	Oui		*[1] - R26*	
7- Penser une stratégie organisationnelle										
#51	Documenter les choix de conception.	Oui		Oui					*[2]*	
#52	Identifier les personnes clés et encadrer le recours à des sous-traitants.	Oui				Oui			*[2]*	
#53	Mettre en œuvre une stratégie de gestion de risque.	Oui			Oui		Oui		*[2]*	
#54	Prévoir un mode dégradé des services métier sans système d'IA.	Oui				Oui	Oui		*[2]*	Le mode dégradé des services est également prévu par [1] - R15.
#55	Prévoir un mode dégradé des services métier sans système d'IA.	Oui				Oui	Oui		*[1] - R15*	Le mode dégradé des services est également prévu par [2]
#56	Mettre en place des politiques d'utilisation encadrées de l'IA générative (selon la sensibilité de l'organisation).					Oui	Oui		*[2]*	
#57	Mettre en place un processus de veille des vulnérabilités spécifiques aux systèmes d'IA.	Oui		Oui			Oui		*[2]*	
#58	Documenter les jeux de données produits		Oui	Oui		Oui	Oui		*[2]*	
#59	Faciliter l'utilisation de la base de données		Oui	Oui					*[2]*	
#60	Faciliter le suivi des données dans le temps jusqu'à leur suppression ou leur anonymisation ;		Oui				Oui	Oui	*[2]*	
#61	Réduire les risques d'une utilisation imprévue des données.		Oui				Oui	Oui	*[2]*	
8- Mesures préventives										
#62	Former régulièrement le personnel sur les risques de sécurité liés à l'IA.	Oui	Oui	Oui	Oui	Oui	Oui	Oui	*[2]*	
#63	Sensibiliser les développeurs sur les risques liés au code source généré par IA.						Oui		*[1] - R32*	

#64	Proscrire l'utilisation d'outils d'IA générative sur Internet pour un usage professionnel impliquant des données sensibles.					Oui	Oui		*[1] - R34*	
#65	**Formation des utilisateurs** : Éduquer les développeurs de modèles d'apprentissage automatique sur les pratiques de codage sécurisé et les vulnérabilités de l'apprentissage automatique.	Oui							*[17]-AML.M0018*	
#66	Contrôler systématiquement le code source généré par IA.						Oui		*[1] - R30*	
#67	Limiter la génération de code source par IA pour des modules critiques d'applications.	Oui		Oui					*[1] - R31*	
#68	Effectuer des audits de sécurité réguliers sur le système d'IA.	Oui	Oui	Oui	Oui	Oui	Oui		*[2]*	La conduite d'audit en tant que mesure préventive est également prévue par [1] - R23.
#69	Prévoir des audits de sécurité des systèmes d'IA avant déploiement en production.	Oui	Oui	Oui	Oui	Oui	Oui		*[1] - R23*	La conduite d'audit en tant que mesure préventive est également prévue par [2]
#70	Effectuer une revue régulière de la configuration des droits des outils d'IA générative sur les applications métier.						Oui		*[1] - R35*	
#71	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.	Oui	Oui				Oui		*[2]*	

III Protection spécifique attaques IA

		Phases du cycle de vie								
	Description	**A - Planification et design**	**B - Collecte & traitement des données**	**C – Construction du modèle / adaptation d'un modèle existant**	**D - Test, évaluation, vérification**	**E - Mise à disposition, utilisation, déploiement**	**F – Exploitation et maintenance**	**G – Décommissionnement / mise au rebut**	**Source**	**Commentaires**
1 - Recommandations générales										
2- Recommandations pour l'infrastructure et l'architecture										
#1	**Méthodes de distribution des modèles :** Le déploiement de modèles d'apprentissage automatique sur des dispositifs en périphérie peut augmenter la surface d'attaque du système. Envisager de servir les modèles dans le cloud pour réduire le niveau d'accès de l'adversaire au modèle. Envisager également de calculer les caractéristiques dans le cloud pour prévenir les attaques de type boîte grise, où un adversaire a accès aux méthodes de prétraitement du modèle.			Oui	Oui	Oui	Oui		*[17] - AML. M0017*	
#2	**Utiliser des capteurs multimodaux** : Incorporer plusieurs capteurs pour intégrer des perspectives et des modalités variées afin d'éviter un point de défaillance unique vulnérable aux attaques physiques.	Oui	Oui						*[17] - AML. M0009*	
3- Avoir un plan de déploiement										
#3	**Valider le modèle d'apprentissage automatique :** Valider que les modèles d'apprentissage automatique fonctionnent comme prévu en testant les déclencheurs de portes dérobées ou les biais adverses. Surveiller le modèle pour détecter les dérives de concept et les dérives des données d'entraînement, ce qui peut indiquer une altération ou un empoisonnement des données.	Oui	Oui	Oui	Oui	Oui	Oui		*[17] - AML. M0008*	
4- Être vigilant aux ressources utilisées										

5- Sécuriser et durcir le processus d'apprentissage										
#4	**Assainir les données d'entraînement** : Détecter et supprimer ou remédier aux données d'entraînement empoisonnées. Les données d'entraînement doivent être assainies avant l'entraînement du modèle et de manière récurrente pour un modèle d'apprentissage actif. Mettre en place un filtre pour limiter les données d'apprentissage ingérées. Établir une politique de contenu permettant de supprimer les contenus indésirables, tels que certains propos explicites ou offensants.		Oui	Oui		Oui			*[17] - AML. M0007*	
#5	**Durcissement du modèle :** Utiliser des techniques pour rendre les modèles d'apprentissage automatique robustes face aux entrées adverses, telles que l'entraînement adversarial ou la distillation de réseau.		Oui	Oui			Oui		*[17] - AML. M0003*	
#6	**Utilisation des méthodes d'ensemble :** Utiliser un ensemble de modèles pour l'inférence afin d'augmenter la robustesse face aux entrées adverses. Les méthodes d'ensemble combinent plusieurs modèles pour améliorer les performances prédictives et la robustesse. Certaines attaques peuvent contourner efficacement un modèle ou une famille de modèles, mais être inefficaces contre d'autres.		Oui	Oui			Oui		*[17] - AML. M0006*	
#7	**Alignement du modèle d'IA générative :** Lors de la formation ou du réglage fin d'un modèle d'IA générative, il est important		Oui	Oui	Oui		Oui		*[17] - AML. M0022*	

	d'utiliser des techniques qui améliorent l'alignement du modèle avec les politiques de sécurité, de sûreté et de contenu. Le processus de réglage fin peut potentiellement supprimer les mécanismes de sécurité intégrés dans un modèle d'IA générative, mais l'utilisation de techniques telles que le réglage fin supervisé, l'apprentissage par renforcement à partir de retours humains ou de retours d'IA, et la distillation de contexte de sécurité ciblée peut améliorer la sécurité et l'alignement du modèle.									
6- Fiabiliser l'application										
#8	**Lignes directrices pour l'IA générative :** Les lignes directrices sont des contrôles de sécurité placés entre les entrées fournies par l'utilisateur et un modèle d'IA générative pour aider à diriger le modèle à produire des sorties souhaitées et prévenir les sorties indésirables. Les lignes directrices peuvent être mises en œuvre sous forme d'instructions ajoutées à toutes les invites utilisateur ou comme partie des instructions dans l'invite système. Elles peuvent définir les objectifs, le rôle et la voix du système, ainsi que décrire les paramètres de sécurité et de sûreté.					Oui	Oui		*[17] - AML. M0021*	
#9	**Garde-fous pour l'IA générative :** Les garde-fous sont des contrôles de sécurité placés entre un modèle d'IA générative et la sortie partagée avec l'utilisateur pour prévenir les entrées et sorties indésirables. Les garde-fous peuvent prendre la forme de validateurs tels que des filtres, une logique basée sur des règles ou des expressions régulières, ainsi que des approches basées sur l'IA, telles que des classificateurs et l'utilisation de LLM, ou la reconnaissance d'entités nommées (NER) pour évaluer la sécurité de l'invite ou de la					Oui	Oui		*[17] - AML. M0020*	

	réponse. Des méthodes spécifiques au domaine peuvent être employées pour réduire les risques dans divers domaines tels que l'étiquette, les dommages à la marque, le jailbreak, les fausses informations, les exploits de code, les injections SQL et les fuites de données.									
#10	**Détection des entrées adverses :** Détecter et bloquer les entrées adverses ou les requêtes atypiques qui dévient du comportement bénin connu, présentent des schémas de comportement observés dans des attaques précédentes ou proviennent d'IP potentiellement malveillantes. Incorporer des algorithmes de détection des attaques adverses dans le système d'apprentissage automatique avant le modèle d'apprentissage automatique.					Oui	Oui		*[17] - AML. M0015*	
#11	**Restaurer les entrées :** Prétraiter toutes les données d'inférence pour annuler ou inverser les perturbations adverses potentielles.					Oui	Oui		*[17] - AML. M0010*	
#12	**Signature de code :** Appliquer l'intégrité des binaires et des applications avec la vérification des signatures numériques pour empêcher l'exécution de code non fiable. Les adversaires peuvent intégrer du code malveillant dans les logiciels ou les modèles d'apprentissage automatique. L'application de la signature de code peut prévenir la compromission de la chaîne d'approvisionnement de l'apprentissage automatique et empêcher l'exécution de code malveillant.					Oui	Oui		*[17] - AML. M0013*	
#13	**Restreindre le nombre de requêtes au modèle :** Limiter le nombre total et la fréquence de requêtes qu'un utilisateur peut effectuer.					Oui	Oui		*[17] - AML. M0002*	

#14	**Obfuscation passive des sorties de l'apprentissage automatique :** réduire la fidélité des sorties du modèle fournies à l'utilisateur final peut diminuer la capacité des adversaires à extraire des informations sur le modèle et à optimiser les attaques contre celui-ci.					Oui	Oui		*[17] - AML. M0002*	
7- Penser une stratégie organisationnelle										
#15	**Analyse de vulnérabilité :** Le scanner d'analyse de vulnérabilité est utilisé pour trouver des vulnérabilités logicielles potentiellement exploitables afin de les corriger. Les formats de fichiers tels que les fichiers pickle, couramment utilisés pour stocker les modèles d'apprentissage automatique, peuvent contenir des codes (exploits) permettant l'exécution de code arbitraire. Ces fichiers doivent être analysés pour détecter les appels potentiellement dangereux, qui pourraient être utilisés pour exécuter du code, créer de nouveaux processus ou établir des capacités de mise en réseau. Les adversaires peuvent intégrer des codes malveillants dans des fichiers de modèles corrompus, de sorte que les scanners devraient être capables de travailler avec des modèles qui ne peuvent pas être entièrement désérialisés. Les artefacts de modèle et les éléments produits par les modèles doivent être analysés pour détecter les vulnérabilités connues.			Oui			Oui		*[17] - AML. M0016*	
8- Mesures préventives										

10 Annexe 2 - Remédiation

Etape	Sous-phase du Cycle de Vie d'un SIA	Action à Vérifier	État (à cocher)
Gouvernance & Gestion de Crise	Planification et Design	Définir les exigences de sécurité IA et les référentiels réglementaires (*RGPD, ANSSI, NIST CSF*).	☐
		Mettre en place une gouvernance intégrant Security by Design	☐
		Élaborer un plan de gestion des incidents IA (politiques, procédures, rôles et responsabilités).	☐
		Définir les mécanismes de traçabilité et auditabilité des modèles IA (journaux d'activités, logs décisionnels des modèles)	☐
		Évaluer les risques spécifiques aux SIA à l'aide de méthodes comme EBIOS Risk Manager pour identifier les menaces IA.	☐
		Définir des politiques de contrôle d'accès et d'authentification aux modèles IA	☐
	Collecte et Traitement des Données	Mettre en place une gouvernance des données et contrôler leur provenance.	☐
		Définir un protocole de surveillance continue des flux de données IA	☐
		Vérifier la qualité des jeux de données IA et prévenir l'empoisonnement des données.	☐
Détection & Investigation	Construction du Modèle / Adaptation	Auditer la robustesse des modèles IA et détecter les vulnérabilités.	☐
		Vérifier l'intégrité des modèles pré-entraînés et dépendances externes.	☐
		Détecter les attaques adversariales (Model Stealing, Data Poisoning, Backdoor Attacks).	☐
	Test, Évaluation et Vérification	Effectuer des tests adversariaux et vérifier la résistance aux attaques.	☐
		Vérifier la robustesse du modèle IA contre les dérives et manipulations.	☐
		Surveiller les logs SIEM et *Threat Intelligence* IA pour détecter les menaces.	☐
	Mise à Disposition / Déploiement	Sécuriser les pipelines de déploiement et restreindre les accès non autorisés.	☐
		Activer un plan de confinement pour isoler les modèles IA compromis.	☐
		Notifier les autorités et équipes concernées (ANSSI, CNIL, CERT-FR).	☐

	Exploitation et Maintenance	Mettre en place un monitoring avancé pour détecter les compromissions IA en temps réel.	☐
		Analyser les logs et événements pour identifier la cause et l'ampleur de l'attaque.	☐
Remédiation & Reconstruction	Mise à Disposition / Déploiement	**#1** Appliquer la méthodologie E3R (Endiguement, Éviction, Éradication, Reconstruction) :	☐
		#2 Isoler les modèles IA compromis en les retirant des pipelines de production.	☐
		#3 Activer un mode dégradé / *safe mode*	☐
		#4 Restreindre l'accès aux jeux de données impactés	☐
		#5 Bloquer l'exfiltration de données sensibles liées aux modèles IA	☐
		#6 Effectuer une première évaluation des dommages via une analyse des logs SIEM et des indicateurs de compromission (IoC).	☐
		#7 Révoquer les clés d'accès et changer toutes les *credentials* associées aux modèles IA et aux infrastructures.	☐
		#8 Supprimer les backdoors potentielles implantées dans les modèles ou les APIs IA	☐
		#9 Désactiver les comptes d'utilisateurs ou services ayant été compromis pendant l'attaque	☐
		#10 Vérifier les configurations réseau et appliquer une segmentation stricte pour limiter les futures exploitations	☐
		#11 Réinitialiser les pipelines de CI/CD et de MLOps pour s'assurer qu'aucun processus automatisé compromis ne réintroduise des failles	☐
	Exploitation et Maintenance	**#12** Nettoyer les données IA corrompues et réentraîner les modèles.	☐
		#13 Appliquer des correctifs et renforcer les configurations de sécurité.	☐
		#14 Vérifier l'intégrité des modèles et valider leur sécurité avant leur redéploiement.	☐
		#15 Appliquer un test de stress et des attaques simulées pour garantir que les vulnérabilités corrigées ne sont plus exploitables.	
		#16 Mettre en place un suivi post-incident pour éviter une récidive.	☐
		Effacer de manière sécurisée les modèles et logs IA obsolètes.	☐

Amélioration Continue	Décommissionnement / Mise au Rebut	Réaliser un audit final avant la mise hors service du SIA.	☐
	RETEX & Formation	Documenter les incidents et mettre à jour les stratégies de sécurité IA (RETEX structuré).	☐
		Organiser des simulations adversariales (Red Team IA) pour tester la robustesse des systèmes.	☐
		Améliorer les modèles de détection et de réponse aux menaces IA.	☐

Contacts utiles

Organisation	Rôle	Lien
ANSSI (France)	Guides stratégiques et opérationnels pour remédiation	https://www.ssi.gouv.fr
CERT-FR	Support technique et remontée d'incidents	https://www.cert.ssi.gouv.fr
CNIL (France)	Notification des violations de données personnelles	https://www.cnil.fr
Prestataires agréés PRIS (ANSSI)	Intervention spécialisée pour réponse à incident	https://cyber.gouv.fr/prestataires-de-reponse-aux-incidents-de-securite-pris
ENISA (Europe)	Conseils et bonnes pratiques européennes	https://www.enisa.europa.eu
Fournisseur cloud ou IT externe	Assistance technique pour systèmes hébergés	Contact spécifique du fournisseur
Équipe juridique interne	Support légal pour communication et conformité	Contact interne de l'équipe juridique

11 Remerciements

Coordinateurs

- Alexandre Coroir, Consultant Cybersécurité, Advens.
- Carine Thérond, Technical Leader, Stormshield.
- Général Patrick Perrot, conseiller IA auprès du Commandement du ministère de l'intérieur dans le cyber espace, Gendarmerie Nationale.
- Françoise Soulié-Fogelman, Conseiller Scientifique, Hub France IA.

Contributeurs

- Nada Amini, Data Scientist, Société Générale.
- Patrick Boutard, CEO, infAIrence.
- Martin d'Acremont, Consultant Cybersécurité, Wavestone.
- Matthieu Ferrandez, Responsable Datascience & AI, CyberDefense,I-Tracing.
- Bruno Grieder, Directeur Technique, Cosmian.
- Soufiane Kaissari, Chargé de la valorisation de la recherche, GLIMPS.
- Camille Maindon, Consultante, Capgemini Invent.
- Aurélien Mayoue, AI research engineer, CEA.
- Eric Savignac, Expert Cybersécurité et IA.
- Christos Katsoukalis, Data Scientist, Société Générale.
- Thierno Kante, Data Scientist, Edicia.
- Jean-Marc Schenkel, Expert Cybersécurité.
- Michael Slimani, Expert Cybersécurité.

Relecteurs

- Gérald Aroulanda, CEO, YMUNIT & Risk Hunter.
- Stephan Cohen, Cyber Security Specialist, BNPP.
- Alexandre Gakic, Expert Cyber & IA.
- Hervé Léon, Head of Formind Academy, Formind.
- Maxime de Jabrun, VP Cyber risk & IA, HeadMind Partners.
- Thomas Kernem-Om, Chef de projet Sénior, Hub France IA.

La touche finale

- Mélanie Arnould, Responsable des opérations, Hub France IA.
- Thomas Kernem-Om, Chef de projet Sénior, Hub France IA.

Analyse des attaques sur les systèmes de l'IA

Mai 2025

CAMPUS CYBER

HUB FRANCE IA