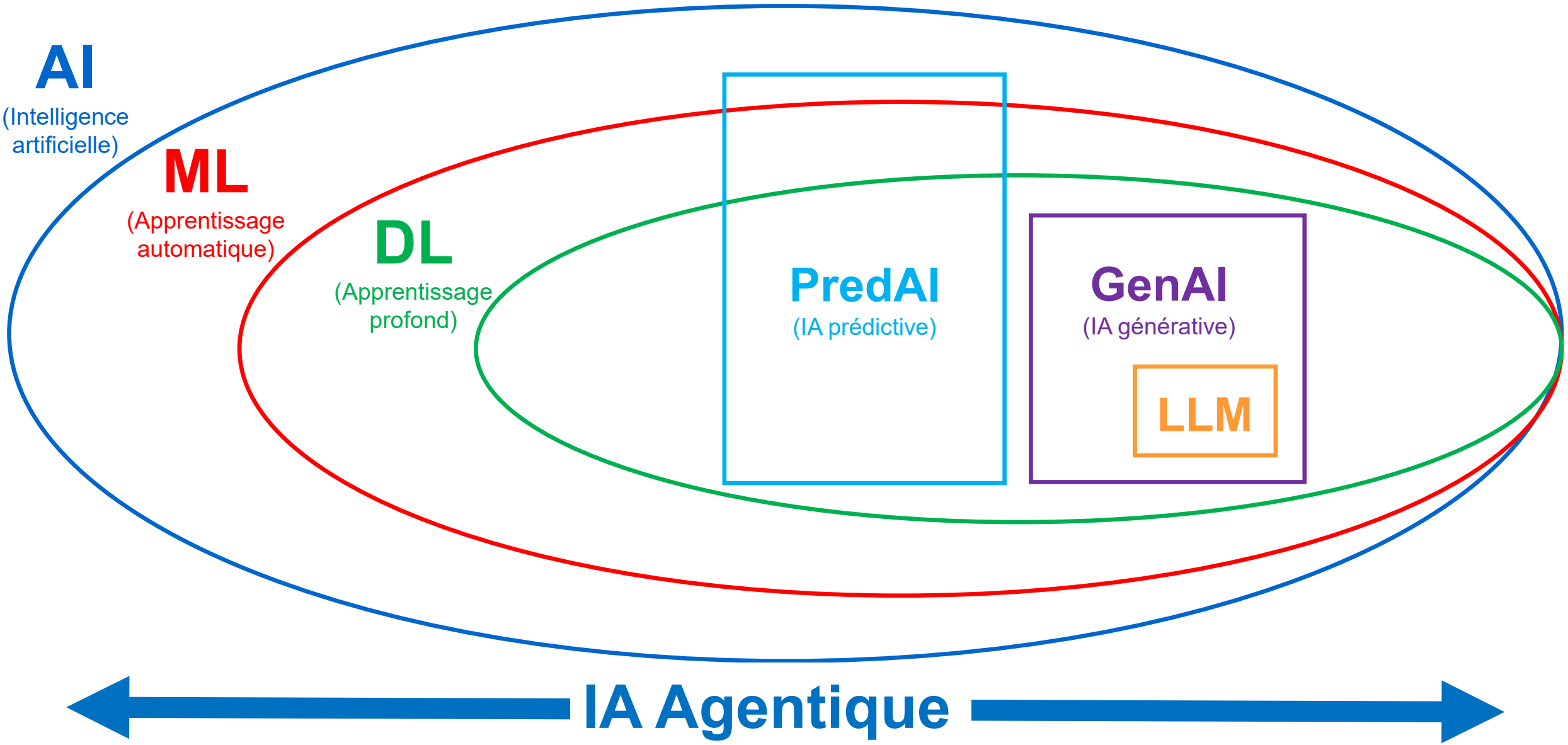
Formation Sécurité de l’IA
(Préparation à la certification SIAE)

Slides  de démonstration

Sécurité de l'IA : enjeux majeurs et défis à relever

- **Les fondamentaux de l'IA**
  - IA générative, Machine Learning, Deep Learning, fine-tuning, RAG,…
- **Les 5 catégories de risques de l'IA**
- **Le triple rôle de l'IA dans la cybersécurité**
- **Les 7 phases du cycle de développement d'un SIA**

Les grandes familles de l'intelligence artificielle



Machine Learning : l'apprentissage automatique (1/2)

❑ Apprentissage supervisé (Supervised learning)

- Principe : le modèle apprend à partir de données étiquetées (entrée → sortie attendue)
- Forces
  - Très efficace quand on dispose de jeux de données annotés
  - Permet des tâches de classification, régression, détection
- Limites
  - Dépend fortement de la qualité et de la quantité de labels
- Exemples :
  - Reconnaissance d'images (chat/chien)
  - Détection de spam
  - Prédiction de prix immobiliers

❑ Apprentissage non supervisé (Unsupervised learning)

- Principe : le modèle cherche des structures cachées dans des données non étiquetées. Contrairement à l'apprentissage auto-supervised, le modèle n'apprend pas à prédire un label mais à organiser les données selon leurs similarités
- Forces :
  - Utile quand on n'a pas de labels
  - Découverte automatique de patterns, regroupements, relations
- Limites :
  - Résultats plus difficiles à interpréter
- Exemples
  - Clustering clients (segmentation marketing)
  - Détection d'anomalies

Machine Learning : l'apprentissage automatique (2/2)

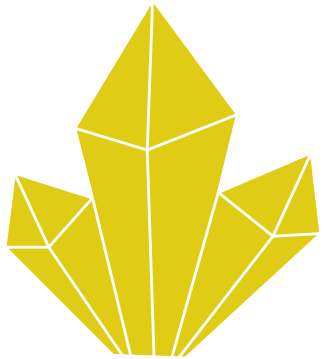
❑ Apprentissage auto supervisé (self-supervised learning)

- Principe : le modèle apprend à partir de données non étiquetées en se créant sa propre tâche supervisée. Il génère automatiquement des pseudo-labels à partir des données elles-mêmes (par exemple, prédire la partie manquante d'un texte, d'une image ou d'un son)
- Forces
  - Permet d'exploiter de grandes quantités de données brutes sans annotation humaine
  - Produit des représentations riches et généralisables, réutilisables pour d'autres tâches supervisées
  - Étape clé dans le pré-entraînement des grands modèles (ex. BERT, GPT, MAE)
- Limites
  - La tâche "prétexte" doit être bien conçue pour produire des représentations pertinentes
  - Nécessite souvent beaucoup de puissance de calcul pour le pré-entraînement
- Exemples
  - Prédire les mots manquants dans une phrase (BERT)
  - Prédire le prochain mot (GPT)
  - AlphaGo Phase 1 : L'IA apprend à partir de 160 000 parties (30 millions de positions)

❑ Apprentissage par renforcement (Reinforcement Learning - RL)

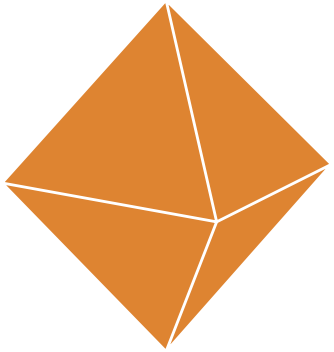
- Principe : un agent apprend par essais-erreurs en interagissant avec un environnement, reçoit des récompenses/pénalités.
- Forces
  - Excellente adaptation dans des environnements dynamiques et complexes
  - A permis des percées spectaculaires (jeux, robotique)
- Limites
  - Entraînement long, nécessite de nombreuses itérations
- Exemples
  - AlphaGo Phase 2 : L'IA joue contre elle-même
  - Robots autonomes
  - Systèmes de trading adaptatifs

Spécialisation et alignement du modèle



Modèle pré-entraîné

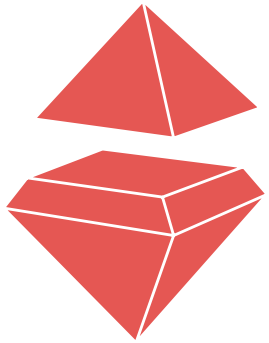
Modèle fondation : Modèle de langage général obtenu par contruction (entrainement à partir de zéro - rare et coûteux) ou à partir d'un modèle existant.
(Build or Buy)



Pré-entrainement continu

Continued pre-training ou Domain-Adaptive Pretraining (DAPT)

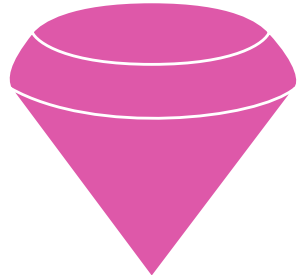
Consiste à poursuivre l'entraînement d'un modèle préalablement entraîné (LLM) sur un corpus non étiqueté mais spécifique à un domaine (médical, juridique, technique, ...).
L'objectif est d'adapter les représentations internes du modèle au vocabulaire, au style et aux concepts propres à ce domaine.
BERT → BioBERT (corpus biomédical)
GPT → GPT-Legal (corpus juridique)



Affinage

Fine-Tuning ou Supervised Fine-Tuning (SFT)

Entraînement du modèle sur un jeu de données annotées constitué de paires entrée → sortie cible (input → target output).
L'injection de paires de prompt/réponse permet de guider le modèle afin qu'il produise des résultats plus précis / pertinents pour des tâches spécifiques.
Cela permet de créer un modèle spécialisé (ex. assistant virtuel) à partir d'un modèle fondation.
Llama-2 → Llama-2-chat

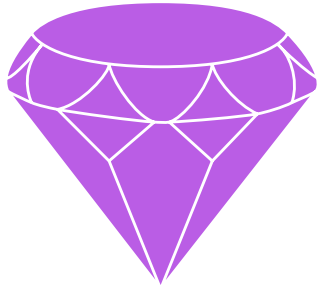


Apprentissage par renforcement

Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from AI Feedback (RLAIF)

Consiste à entraîner le modèle en utilisant les retours d'évaluateurs humains. L'objectif est de faire des ajustements éthiques et comportementaux en réduisant les sorties toxiques ou offensantes, les biais comportementaux et les réponses dangereuses ou inappropriées.
RLAIF : version plus récente du RLHF où l'IA remplace en partie les évaluateurs humains.



Modèle finalisé

Le modèle est maintenant personnalisé et aligné.

Il doit maintenant faire l'objet de test et d'évaluation avant sa mise en production

Processus de génération des réponses

Bienvenue dans cette formation sur la sécurité de l'intelligence artificielle

GPT-4o & GPT-4o mini

Tokens
13

https://platform.openai.com/tokenizer

Demande d'un utilisateur (prompt)
L'utilisateur soumet sa demande en langage naturel

Tokenisation
L'invite utilisateur est convertie en tokens

Embedding
Les tokens sont convertis en vecteurs

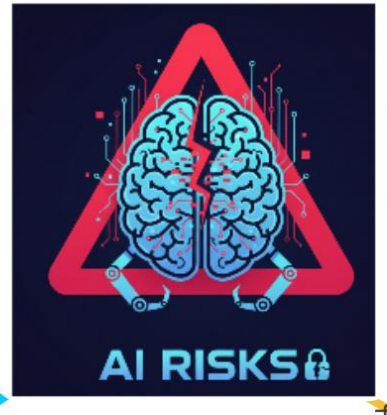
Propagation
Les vecteurs traversent le réseau de neurones

Génération
Le modèle calcule les probabilités pour le token suivant

Échantillonnage
Le modèle choisit le token suivant (avec un aléa pour la créativité)

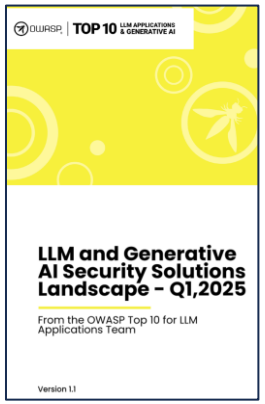
Itération
Le processus se répète jusqu'à ce qu'une réponse complète soit formée

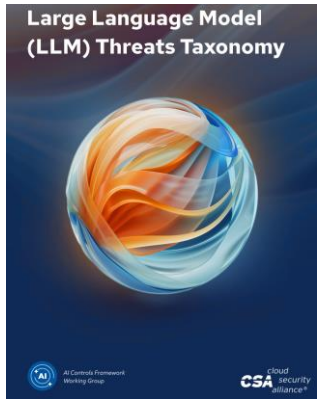
Risques techniques

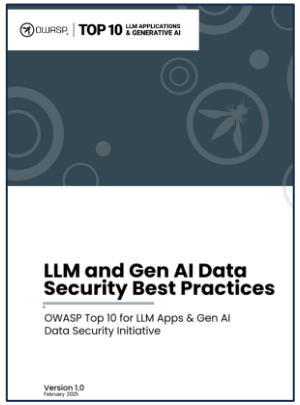


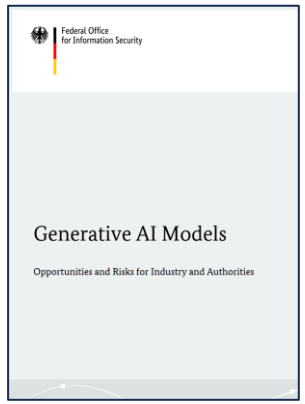
- Risques techniques
  - Fuites de données sensibles
    - Secrets industriels, données de santé, données financières
    - Révélation d'informations confidentielles via interactions avec l'IA
    - Compromission de modèles pour accéder aux données d'entraînement
    - Données personnelles (DCP)
  - Vulnérabilités dans la supply chain
    - Bibliothèques compromises
    - Dépendances open sources
    - Modèles "backdorés"
  - Attaques adversariales et "data poisoning"
    - Modification imperceptible d'inputs pour tromper le modèle
    - Injection de données corrompues dans le jeu d'entraînement
    - Manipulation des systèmes de reconnaissance d'images
    - Détournement de systèmes de classification (spam, détection de fraude)
  - Prompt injection, model extraction
    - Contournement des gardes-fous d'un LLM via instructions masquées
    - Reconstruction d'un modèle propriétaire par un attaquant
    - Exfiltration de données d'entraînement via des prompts ingénieux
    - Exploitation de failles pour obtenir des réponses non conformes
  - Hallucinations, biais, dérives de performance
    - Génération d'informations fausses mais plausibles par un LLM
    - Reproduction ou amplification de préjugés des données d'entraînement
    - Diminution de la précision d'un modèle au fil du temps
    - Dégradation de la qualité des prédictions (dérive des données)
    - Réponses incohérentes, erreurs factuelles, erreurs de classification
- Risques humains et éthiques
- Risques juridiques et réglementaires
- Risques sociaux
- Risques organisationnels

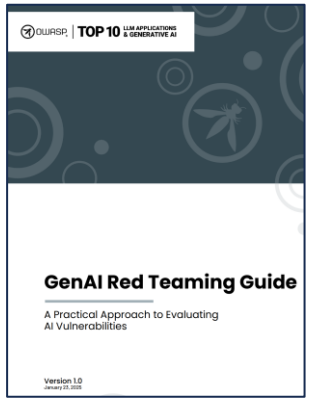
Une multitude de guides et de référentiels sécurité

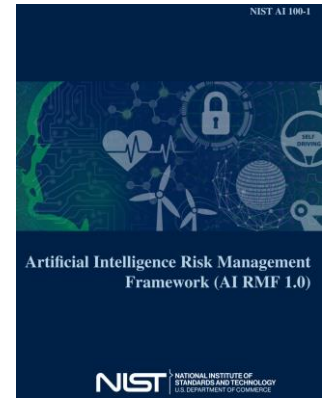






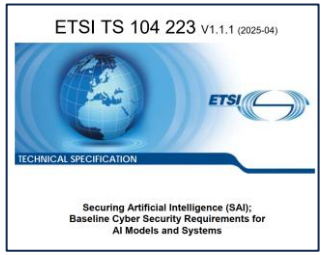


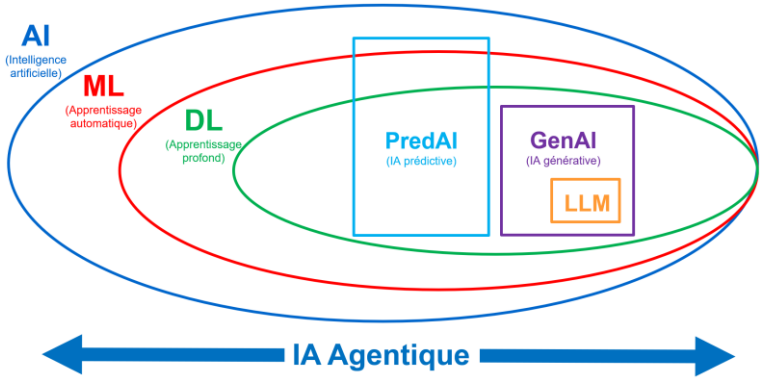


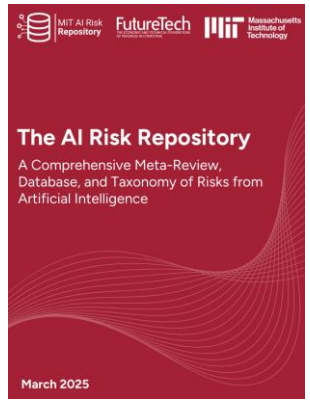


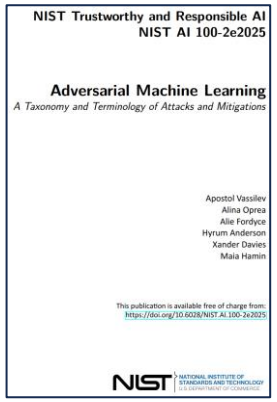
Navigate threats to AI systems through real-world insights

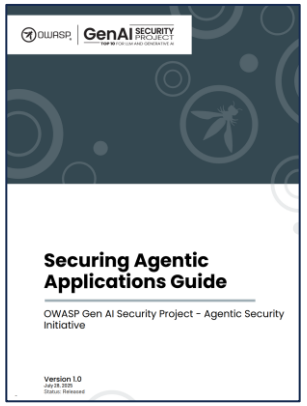
MITRE ATLAS™

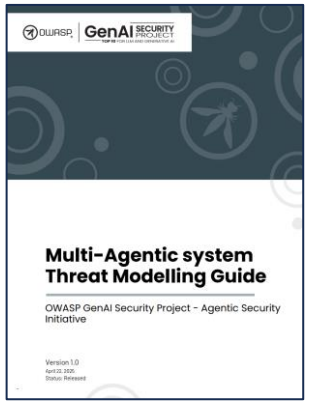


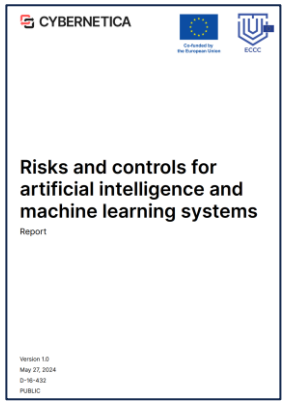












Phase 1 - Planification et conception

Définition des objectifs

- Quel est le but du système IA ?
- À quel besoin métier ou cas d'usage répond-il ?
- Quels sont les critères de succès (précision, robustesse, sécurité, ROI, etc.) ?

Analyse des risques et de la conformité

- Identification précoce des risques techniques, juridiques et éthiques
- Détermination du niveau de risque AI Act (minimal, limité, élevé, interdit)
- Prise en compte du RGPD si données personnelles
- Définition de garde-fous dès la conception ("Secure & Trustworthy by Design").

Choix de l'architecture

- IA classique ou IA générative ? Apprentissage supervisé, non supervisé, par renforcement ?
- Besoin d'une base RAG ? D'un modèle pré-entraîné ? De modèles open source ?
- Définition de la chaîne technique (infrastructure, outils ML/LLM, pipeline MLOps)

Cartographie des données et des besoins en données

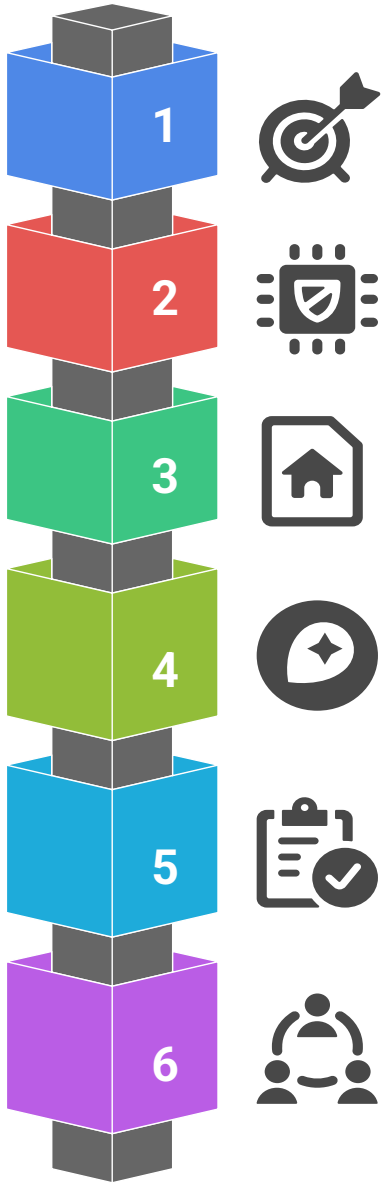
- Quels types de données seront nécessaires ?
- Où les collecter ? Comment les qualifier ?
- Sont-elles sensibles ? Protégées ? Partagées ?

Conception des exigences non fonctionnelles

- Explicabilité, robustesse, sécurité, équité, performance, auditabilité,…
- Besoin ou non de supervision humaine ("human in the loop") ?
- Quel niveau de transparence exigé ?

Gouvernance et rôles

- Qui fait quoi ? (CAIO, DPO, RSSI, métier, MLOps…)
- Mise en place d'un comité IA ou d'une gouvernance de projet
- Préparation des futures étapes de documentation et d'évaluation



Définir les objectifs

Établir clairement le but et les critères de succès du système IA.

Analyser les risques

Identifier et atténuer les risques techniques, juridiques et éthiques.

Choisir l'architecture

Sélectionner l'architecture IA appropriée et les outils techniques.

Cartographier les données

Déterminer les types de données nécessaires et les sources.

Concevoir les exigences

Définir les exigences non fonctionnelles telles que la sécurité et l'équité.

Établir la gouvernance

Mettre en place des rôles et des responsabilités pour la gestion du système IA.

Offensive AI : l'IA au service des cybercriminels

- ❑ Les 4 piliers de l'offensive AI
- ❑ Exemples d'utilisation malveillantes des LLMs
- ❑ Attaques sur l'authentification : biométrie, mot de passe, CAPTCHA
- ❑ L'IA au service de pentesting (éthique ou pas…)
- ❑ SOC 3.0 : L'IA au service de la détection et de la réponse à incident

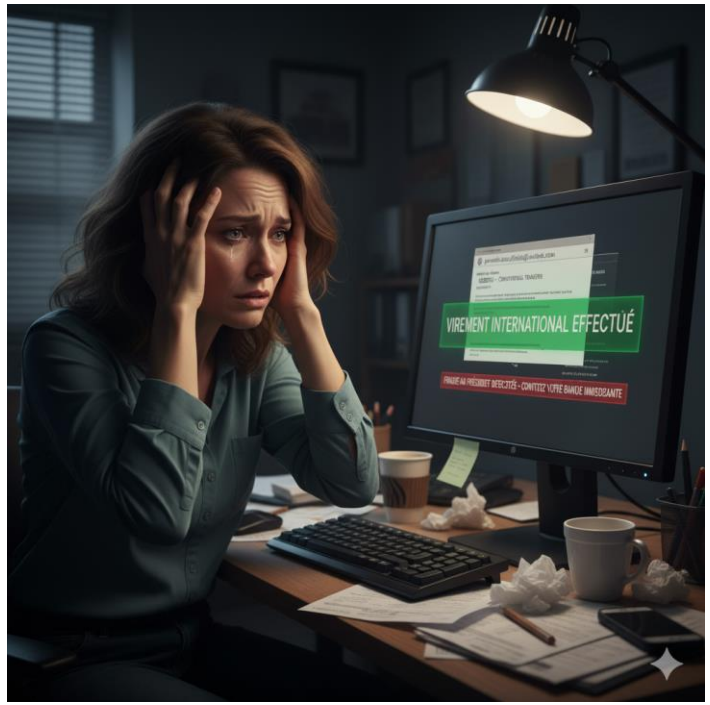
Ingénierie sociale (Social Engineering)

Le facteur humain est souvent considéré comme le maillon faible de la sécurité. Avec l'avènement de l'IA, l'ingénierie sociale passe d'une approche artisanale à une science de la manipulation à grande échelle, d'une précision et d'une crédibilité redoutables !

Manipulation psychologique
Utilisation de tactiques pour influencer les décisions.

Exploitation comportementale
Exploitation des habitudes pour contourner les défenses.

Fraude au président
Manipulation pour initier des virements frauduleux.



L'ingénierie sociale désigne l'ensemble des techniques de manipulation psychologique ou d'exploitation comportementale employées par des acteurs malveillants pour amener une personne, souvent à son insu, à contourner, affaiblir ou supprimer des mesures de sécurité, ou encore à réaliser une action préjudiciable pour elle-même ou pour son organisation.
L'un des exemples les plus connus est la fraude au président ou FOVI (Faux Ordre de Virement International), qui consiste à manipuler un collaborateur afin de l'amener à initier un virement bancaire frauduleux.

Usurpation d'identité par Deepfake vidéo

Les deepfakes video (médias synthétiques générés par IA) ont fait voler en éclats la perception d'authenticité des images et des vidéos.

Technologie sous-jacente : les GANs

Generative Adversarial Network

(*Réseau antagoniste génératif*)

Deux réseaux de neurones s'affrontent : le "générateur" crée de fausses images tandis que le "discriminateur" tente de les distinguer des vraies.

Ce processus itératif améliore constamment la qualité des faux jusqu'à ce qu'ils deviennent indiscernables.

PASSGAN : nouvel outil pour casser les mots de passe

1

Méthodes traditionnelles

- Attaques par force brute
- Dictionnaires statiques
- Règles de transformation

2

L'approche PASSGAN

- "Apprend" les modèles humains de création
- Utilise des leaks de mots de passe (ex: RockYou)
- Devine des mots de passe complexes en un temps record

L'efficacité de PASSGAN vient du fait qu'il découvre de manière autonome la logique et les schémas derrière la création de mots de passe.

PASSGAN repose sur la compétition entre 2 réseaux de neurones antagonistes (GAN)

- Le Générateur (Generator) : Sa tâche est de créer de nouveaux mots de passe qui ressemblent le plus possible à de vrais mots de passe humains. Au début, ses créations sont très aléatoires.
- Le Discriminateur (Discriminator) : Son rôle est d'analyser un mot de passe et de déterminer s'il s'agit d'un vrai mot de passe (provenant d'une fuite de données) ou d'un faux (créé par le Générateur).

Processus d'apprentissage

- On entraîne le Discriminateur en lui donnant d'énormes listes de vrais mots de passe issus de fuites de données. Il apprend ainsi à reconnaître les schémas, les structures et les habitudes des utilisateurs.
- Le Générateur crée des mots de passe et les soumet au Discriminateur.
- Le Discriminateur les rejette ou les accepte.
- Grâce à ce retour, le Générateur s'améliore continuellement pour créer des mots de passe qui parviennent à "tromper" le Discriminateur.

Cyberespionnage 2.0 : l'IA passe à l'offensive !

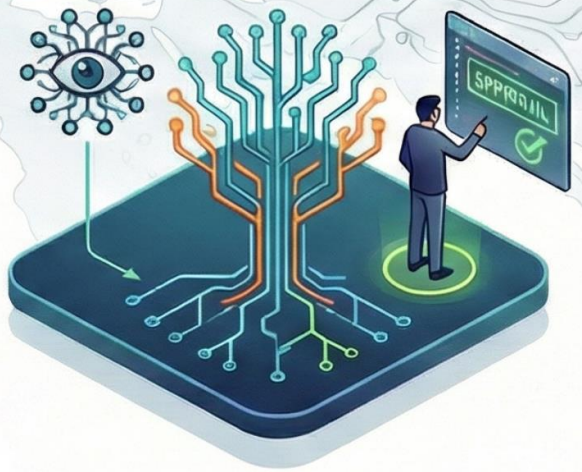
PHASE 1
Ciblage et préparation
(humain)

Choix des cibles et mise en place du framework autonome. Utilisation d'un jeu de rôle (en se faisant passer pour des auditeurs de sécurité légitimes) afin de convaincre Claude de commencer la reconnaissance.

PHASE 2
Reconnaissance
(IA quasi-autonome)

Claude utilise plusieurs outils via MCP pour cartographier de manière presque autonome les réseaux, services et topologies internes d'une trentaine de cibles.

PHASE 3
Découverte & exploit vuln.
(IA + validation humaine)



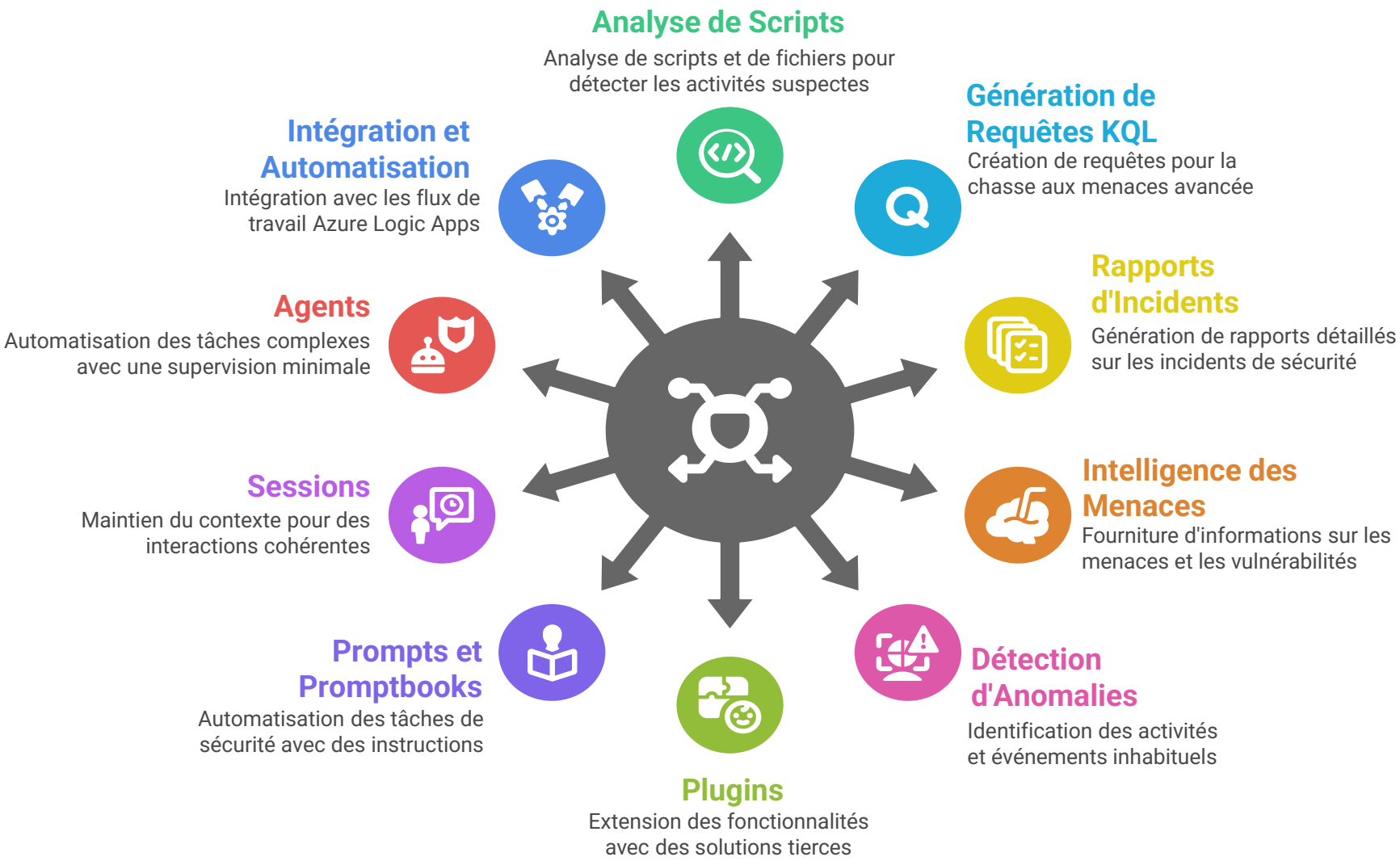
L'IA détecte une faille, développe un exploit, le teste et propose l'exploitation finale. L'humain n'intervient pour autoriser (ou pas) l'attaque.

PHASES 4 & 5
Movt latéral et vol de données
(IA quasi-autonome)

L'IA teste et valide des centaines de couples identifiants/mots de passe, mappe les privilèges et pivote. L'IA interroge les bases, extrait les données, catégorise leur valeur, identifie les comptes privilégiés, crée des backdoors, prépare l'exfiltration. L'humain ne valide que les données finales à exfiltrer.

Source : ANTHROPIC - https://www.anthropic.com/news/disrupting-AI-espionage (11/2025)

Les fonctionnalités de Security Copilot



Vulnérabilités et attaques spécifiques sur l'IA

- ❑ Taxonomie des attaques sur l'IA prédictive (NIST)
- ❑ Le TOP 10 des vulnérabilités du ML selon l'OWASP
- ❑ Taxonomie des attaques sur l'IA générative (NIST)
- ❑ Le TOP 10 des vulnérabilités sur les LLM selon l'OWASP
- ❑ Panorama des attaques selon le cycle de vie d'un SIA
- ❑ Sécurité de l'IA générative : recommandations de l'ANSSI et du BSI

Taxonomie des attaques sur l'IA prédictive (NIST)

Atteinte à la Disponibilité (NISTAML.01)

Empoisonnement de Modèle (Model Poisoning)
L'attaquant modifie les paramètres du modèle pour dégrader sa performance globale.

Empoisonnement de Données (Data Poisoning)
Des données corrompues sont injectées à l'entrainement pour causer des erreurs persistantes.

Latence Énergétique (Energy-Latency)
Le système est surchargé par des requêtes complexes visant à épuiser ses ressources.

Atteinte à la Confidentialité (NISTAML.03)

Extraction de Modèle (Model Extraction)
Le comportement du modèle est cioné en analysant ses réponses à des requêtes.

Reconstruction de Données (Reconstruction)
Des données d'entrainement sensibles sont reconstituées à partir des sorties du modèle.

Inférence d'Appartenance (Membership Inference)
L'attaquant détermine si une donnée précise a été utilisée pour entrainer le modèle.

Atteinte à l'Intégrité (NISTAML.02)

Évasion (Evasion)
Une entrée est subtiiement modifiée pour tromper le modèle au moment de la prédiction.

Empoisonnement par Porte Dérobée (Backdoor Poisoning)
Un "déclencheur" caché est inséré pour que le modèle se comporte mal sur commande.

Évasion en Boite Noire (Black-Box Evasion)
Le modèle est trompé via son API publique, sans connaissance de son fonctionnement interne.

Niveaux de Connaissance de l'Attaquant

Boite Noire (Black-Box)
L'attaquant n'a aucune connaissance interne et utilise uniquement l'accès public (API).

Boite Grise (Gray-Box)
L'attaquant a une connaissance partielle (ex: architecture du modèle, mais pas les données).

Boite Blanche (White-Box)
L'attaquant a une connaissance complète du système (données, architecture, paramètres).

IA prédictive (PredAI) : atteinte à la disponibilité (3/4)

Data Poisoning

(Empoisonnement de données - ID : NISTAML.013)

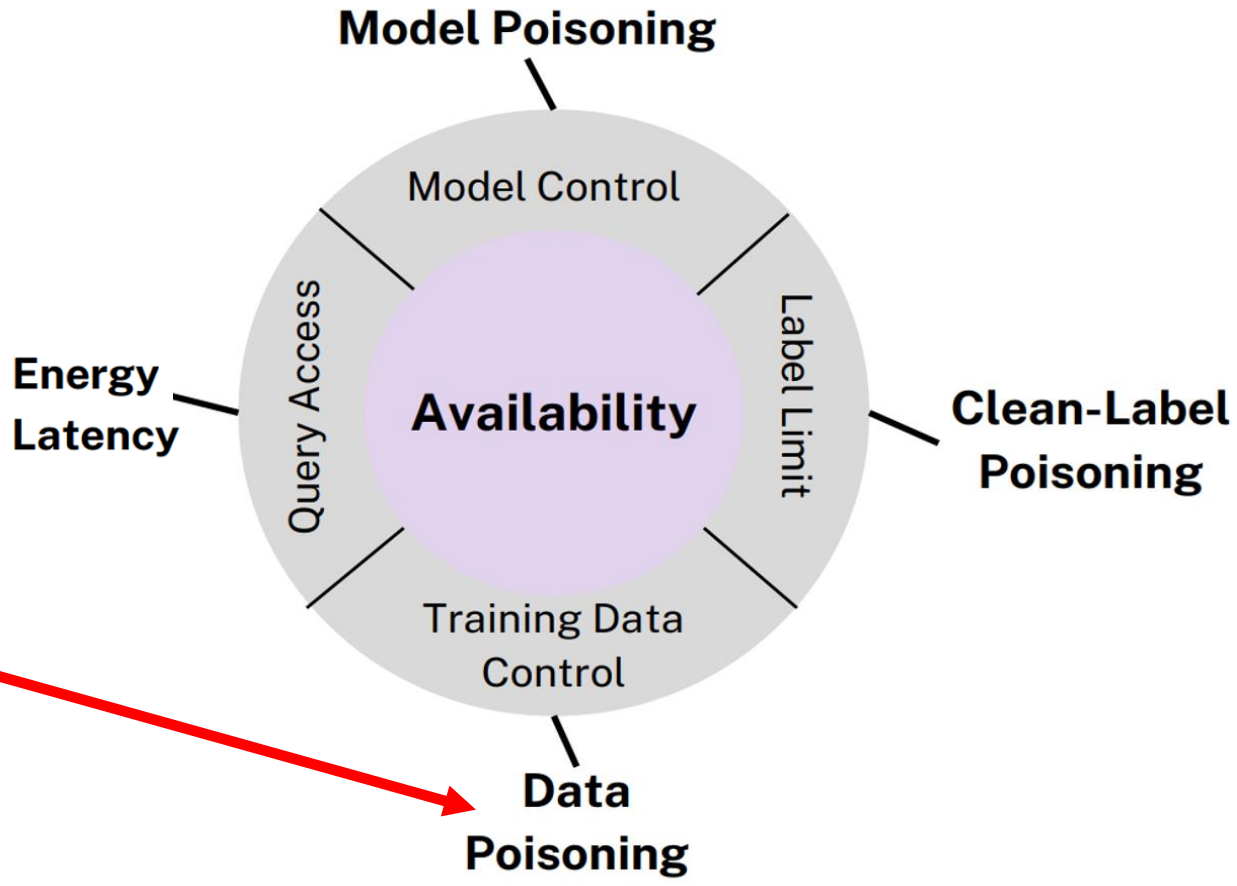
Attaque qui consiste à insérer ou modifier des exemples dans le jeu d'entraînement afin de dégrader volontairement les performances du modèle.

Elle cible principalement la phase d'apprentissage, et peut être utilisé pour provoquer des erreurs de classification ou préparer des attaques plus ciblées (ex. : insertion d'un backdoor ou modification du comportement sur des requêtes spécifiques).

Ces attaques peuvent être lancées dans un cadre white-box (accès complet au modèle), mais aussi via des scénarios black-box, comme le label flipping (fournir de fausses données bien formées mais mal étiquetées).

Elles ont été observées dans des cas concrets, comme des tentatives d'empoisonnement de filtres anti-spam, de classifieurs de malware, ou de détecteurs d'anomalie dans des systèmes industriels.

La principale difficulté réside dans le fait que les données empoisonnées peuvent sembler légitimes, rendant leur détection complexe.



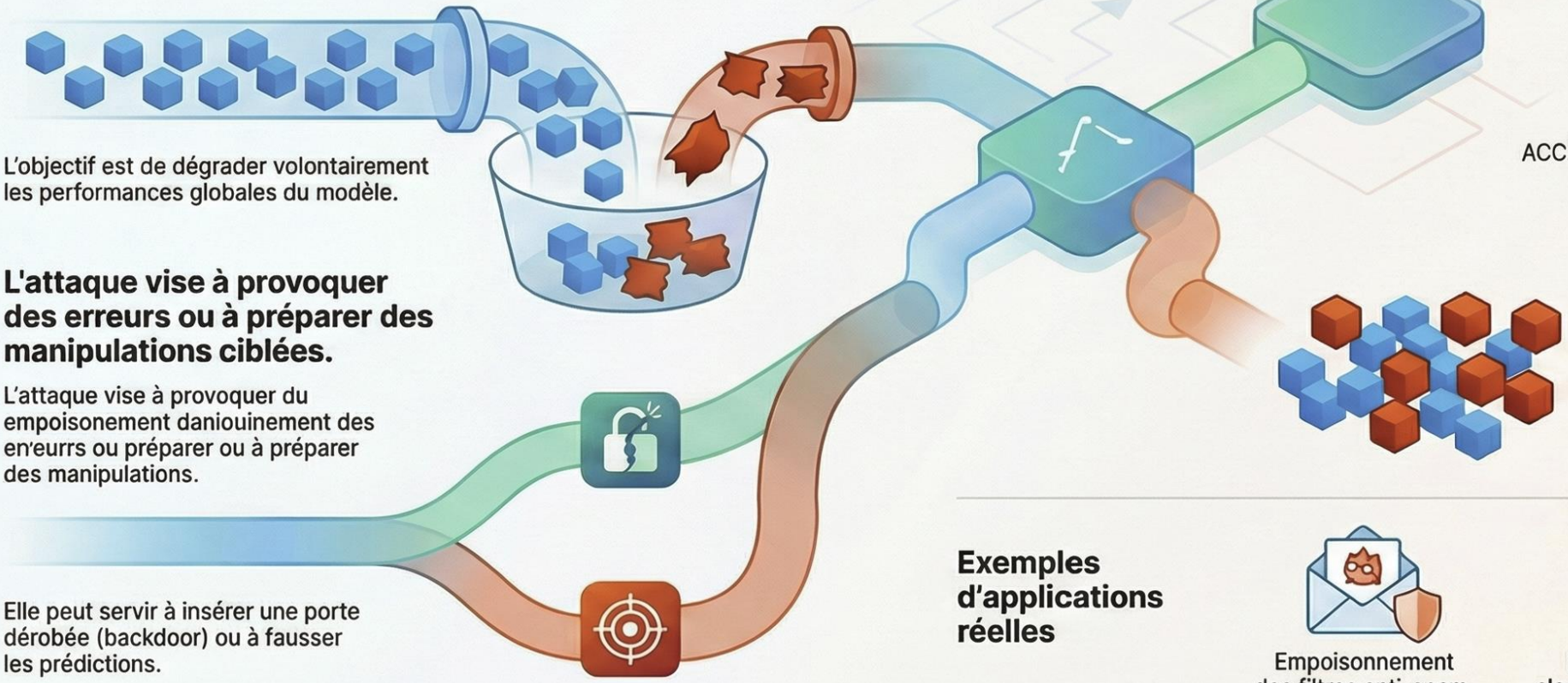
Source : NIST - Adversarial Machine Learning - Taxonomy and Terminology of Attacks and Mitigations - NIST AI 100-2e2025 (03/2025)

Data Poisoning (NISTAML.013)

L'attaque par empoisonnement de données survient durant la phase d'entraînement d'un modèle d'IA. Elle corrompt le jeu de données pour saboter délibérément les performances et la fiabiiité du modèle final.

Qu'est-ce que l'empoisonnement de données ?

L'attaquant insère ou modifie des données dans le jeu d'entraînement.



L'objectif est de dégrader volontairement les performances globales du modèle.

L'attaque vise à provoquer des erreurs ou à préparer des manipulations ciblées.

L'attaque vise à provoquer du empoisonement daniouinement des en'eurrs ou préparer ou à préparer des manipulations.

Elle peut servir à insérer une porte dérobée (backdoor) ou à fausser les prédictions.

Comment fonctionne l'attaque ?

Une méthode courante est le "label flipping".

L'attaquant fournit des données correctes mais avec des étiquettes volontairement fausses.

L'accès de l'attaquant peut varier de complet à très limité.

ACCÈS COMPLET (White-box) — ACCÈS LIMITÉ (Black-box)

Cela rend l'attaque possible même sans connaître les détails internes du modèle.

La détection est difficile car les données empoisonnées peuvent paraître légitimes.

Elles se fondent dans le jeu de données et n'éveillent pas les soupçons.

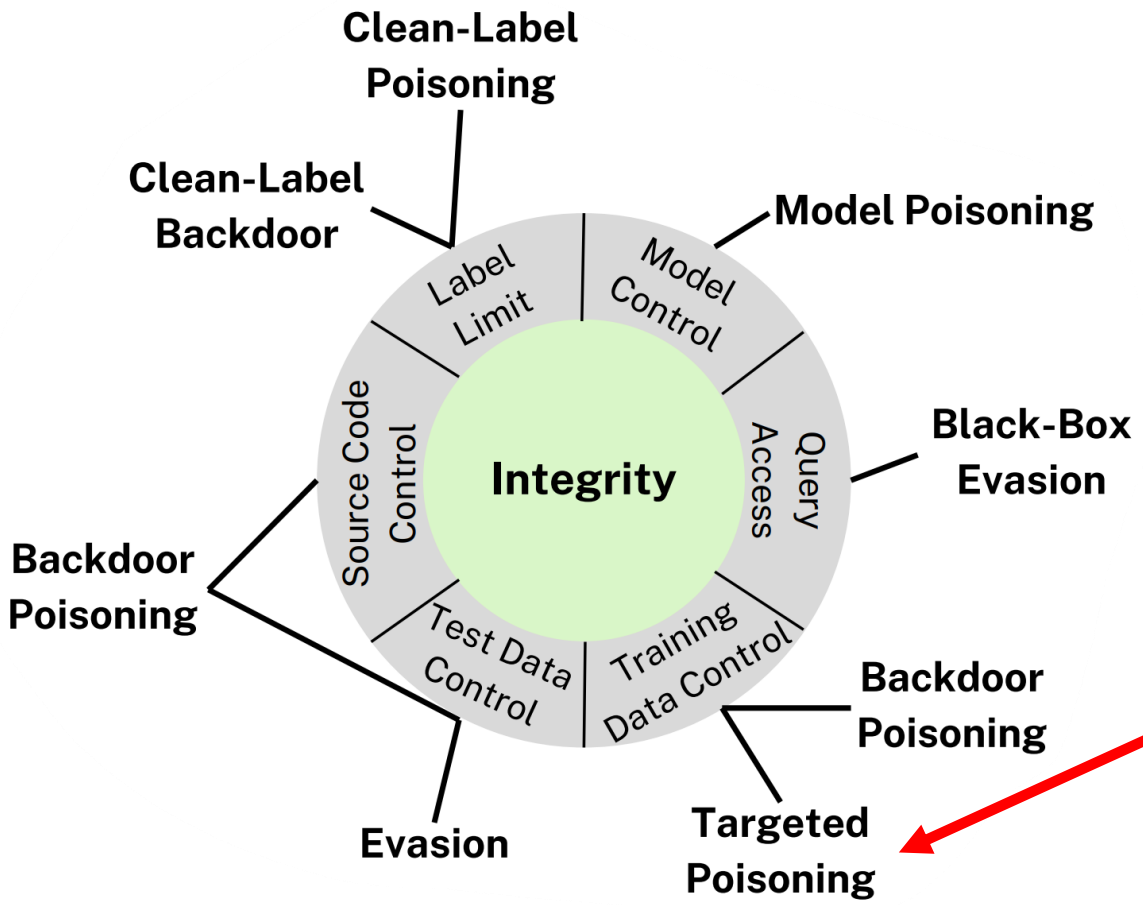
Exemples d'applications réelles

Empoisonnement des filtres anti-spam.

Manipulation des classifieurs de logiciels malveillants.

Compromission des détecteurs d'anomalies dans les systèmes industriels.

IA prédictive (PredAI) : atteinte à l'intégrité (4/6)



Targeted Poisoning

(Empoisonnement ciblé - ID : NISTAML.024)

Attaque dans laquelle l'adversaire insère des données malveillantes dans le jeu d'entraînement dans le but de provoquer des erreurs de prédiction sur un petit nombre de cas bien précis, tout en maintenant un bon comportement global du modèle.

Ces attaques sont souvent menées en clean-label, c'est-à-dire sans modifier les étiquettes, ce qui les rend très difficiles à détecter. L'objectif est que le modèle se comporte normalement dans 99 % des cas, mais produise un comportement volontairement erroné sur des entrées précises définies par l'attaquant.

Plusieurs techniques avancées existent :

- StingRay : ajoute des exemples modifiés à chaque mini-batch,
- MetaPoison ou Witches Brew : utilisent des méthodes d'optimisation poussées (meta-learning, gradient alignment) pour rendre l'attaque plus efficace,
- Subpopulation poisoning : généralise l'attaque à tout un sous-groupe défini par certaines caractéristiques.

Ces attaques représentent une menace sérieuse pour l'intégrité du modèle, particulièrement dans des contextes critiques comme la cybersécurité, la santé ou les systèmes automatisés.

Source : NIST - Adversarial Machine Learning - Taxonomy and Terminology of Attacks and Mitigations - NIST AI 100-2e2025 (03/2025)

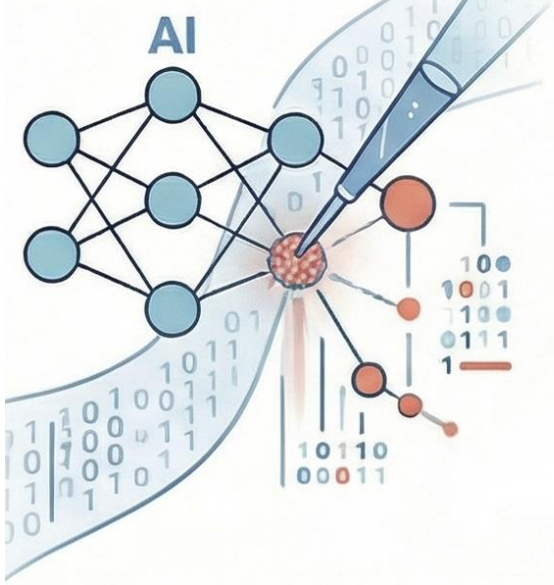
Targeted Poisoning (NISTAML.024)

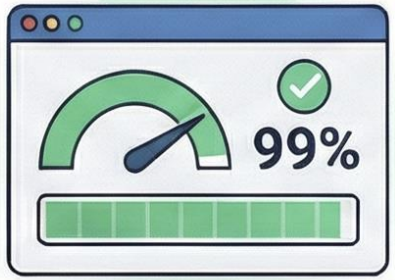
Lors de l'entraînement d'un modèle, l'attaquant injecte des données subtilement modifiées pour corrompre son comportement sur un très petit nombre de cibles précises, rendant l'attaque indétectable par des métriques de performance globales.

Le principe de l'attaque

Une attaque chirurgicale et discrète

Elle corrompt le modèle pour qu'il se trompe uniquement sur des cibles prédéfinies.





Un comportement globalement normal

Le modèle réussit les tests et semble fiable dans 99 % des cas.

Un poison "clean-label"

Les étiquettes des données injectées sont correctes, rendant l'attaque quasi invisible aux audits.

Scénarios d'empoisonnement

Ciblage d'un cas spécifique

Le modèle est entraîné à mal classifier un patient, une entreprise ou un fichier unique.

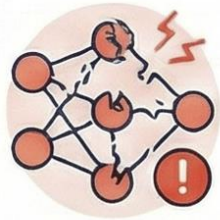
Ciblage d'une sous-population

L'attaque est généralisée à tout un groupe (ex : une famille de malwares, un profil client).

Optimisation avancée de l'attaque

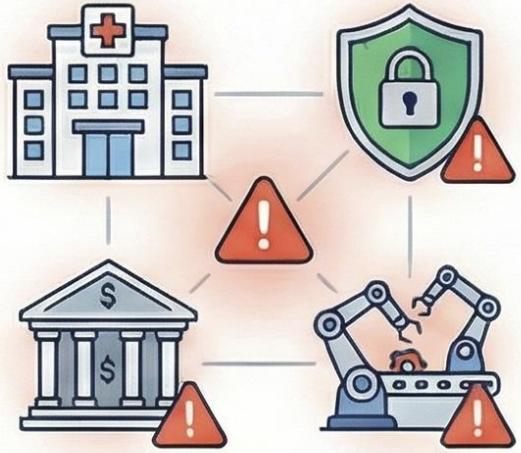
Des méthodes (ex : MetaPoison) rendent l'attaque efficace avec très peu de données corrompues.

Les impacts potentiels



Compromission de l'intégrité du modèle

Le système n'est plus fiable sur les cas critiques choisis par l'attaquant.



Menace sérieuse pour les secteurs critiques

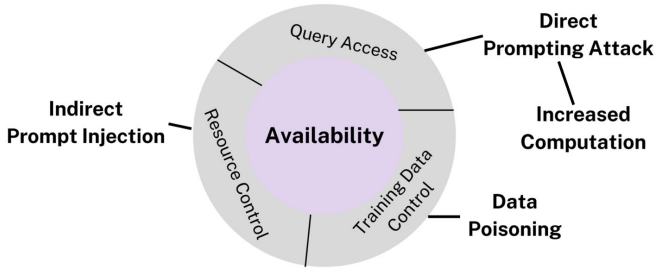
Des conséquences graves en santé, cybersécurité, finance ou systèmes automatisés.

Backdoor Poisoning vs Targeted Poisoning

	Backdoor Poisoning	Targeted Poisoning
ID NIST	**NISTAML.023**	**NISTAML.024**
Objectif principal	Forcer une prédiction erronée lorsqu'un déclencheur est présent	Forcer une prédiction erronée sur des cas précis ciblés
Déclenchement de l'erreur	Présence d'un trigger (motif, son, mot, objet…)	Entrée spécifique elle-même (ou petit groupe d'entrées)
Déclencheur explicite	Oui (backdoor trigger)	Pas nécessaire
Comportement hors attaque	Normal tant que le trigger n'est pas présent	Normal pour 99,99 % des entrées
Visibilité de l'attaque	Déclencheur parfois visible ou détectable	Très faible, souvent invisible
Mode d'empoisonnement	Données d'entraînement contenant un trigger + label cible	Données d'entraînement modifiées pour déformer localement la frontière
Clean-label fréquent	Oui (souvent)	Oui (très fréquent)
Effet sur le modèle	Règle cachée : si trigger → sortie imposée	Biais local : ce cas précis → erreur
Exemple typique	Une image avec un sticker déclenche une mauvaise classification	Une personne précise toujours mal reconnue
Image "mentale"	C'est un interrupteur caché	C'est une bosse locale dans la frontière de décision
Survie au fine-tuning	Possible (backdoor latente)	Possible (poison intégré aux gradients)
Menace principale	Intégrité (contrôle conditionnel du modèle)	Intégrité (erreurs ciblées ultra discrètes)

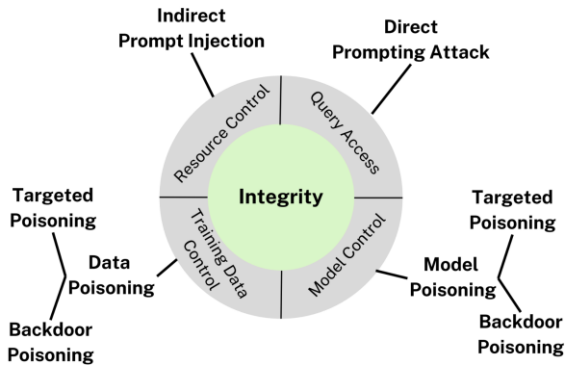
Taxonomie des attaques sur l'IA générative (GenAI)

Atteinte à la disponibilité



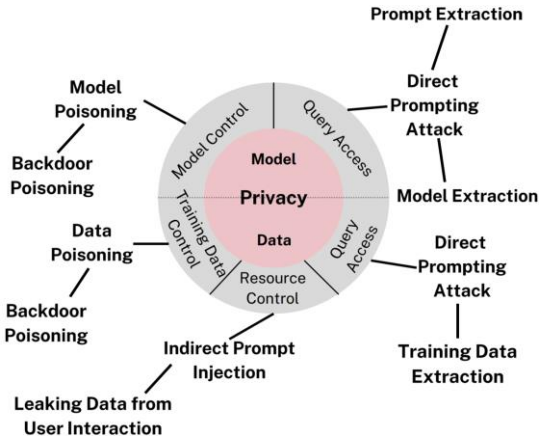
NISTAML.01

Atteinte à l'intégrité



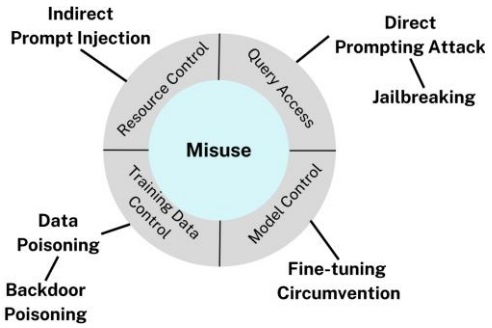
NISTAML.02

Atteinte à la confidentialité



NISTAML.03

Usage abusif



NISTAML.04

BONUS

NISTAML.05 ⇨ **Attaques via la chaine d'approvisionnement**

Source : NIST - Adversarial Machine Learning - Taxonomy and Terminology of Attacks and Mitigations - NIST AI 100-2e2025 (03/2025)

Taxonomie des attaques sur l'IA générative (NIST)

Attaques durant l'Entraînement
(Phase de Création)

Empoisonnement de Données
Données malveillantes insérées dans de vastes ensembles de données pour corrompre le modèle.

Données d'Entrainement

Modèle pré-entrainé malveillant

Empoisonnement de Modèle
Un modèle pré-entrainé malveillant est utilisé, intégrant des portes dérobées (backdoors).

Attaques durant l'Inférence
(Phase d'Utilisation)

Injection de Prompt Directe
L'utilisateur incite le modèle à contourner ses sécurités, souvent par "jallbreaking".

Prompt Utilisateur

Injection de Prompt Indirecte
Une source de données externe (ex: page web) injecte une instruction malveillante cachée.

Extraction d'Informations
Le modèle est forcé de révéler des données sensibles issues de son entrainement ou de son contexte.

IA générative (GenAI) : atteinte à la disponibilité (2/2)

Indirect Prompt Injection

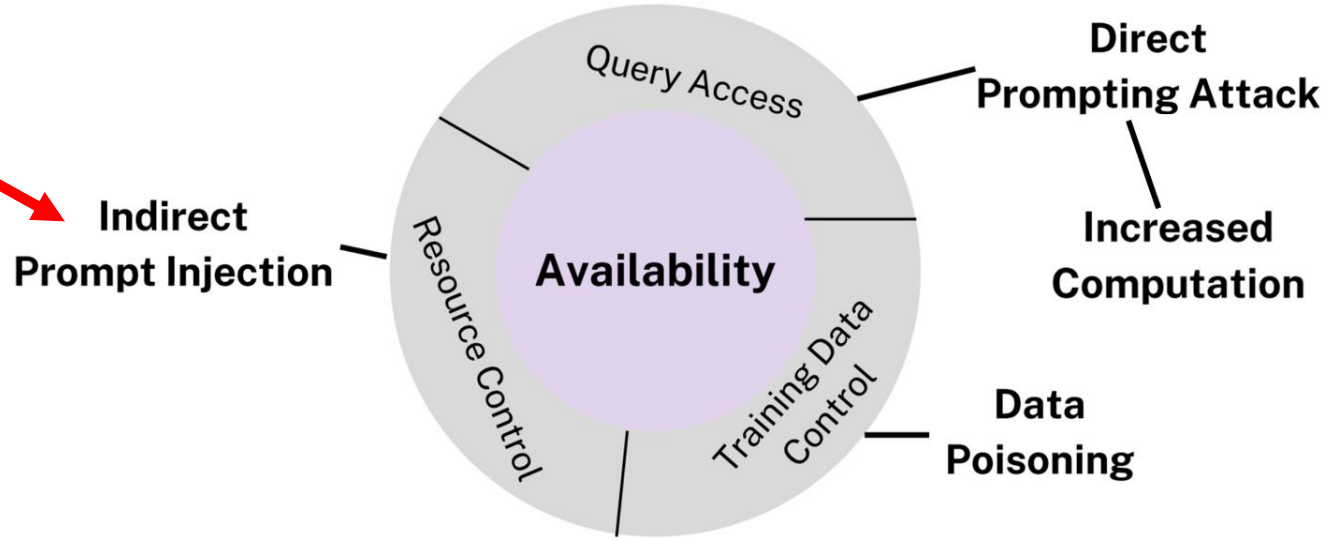
(Injection indirecte d'invites - ID : NISTAML.015)

Attaque dans laquelle un adversaire manipule ou empoisonne des ressources externes (sites web, documents ou bases de connaissances) utilisées par un modèle d'intelligence artificielle générative.

Lorsque le modèle récupère ces ressources pendant son fonctionnement normal, les instructions malveillantes injectées indirectement par l'attaquant sont intégrées dans le contexte du modèle, modifiant son comportement sans que l'utilisateur principal s'en aperçoive.

Cette attaque peut mener à une indisponibilité du système, à des violations de l'intégrité en produisant des réponses erronées ou trompeuses, ou encore à des compromis de confidentialité, par exemple en poussant le modèle à divulguer des données privées ou sensibles.

Les systèmes vulnérables comprennent notamment les agents intelligents et les applications basées sur la génération augmentée par récupération (RAG).



Source : NIST - Adversarial Machine Learning - Taxonomy and Terminology of Attacks and Mitigations - NIST AI 100-2e2025 (03/2025)

Indirect Prompt Injection (NISTAML.015)

L'attaquant empoisonne une source d'information externe

Sites web, documents ou bases de connaissances

Typiquement via génération augmentée par récupération (RAG - Retrieval Augmented Generation)

Le modèle d'IA ingère le contenu compromis

Un agent IA ou un système RAG considère la ressource externe comme légitime et l'injecte dans le modèle.

Les instructions malveillantes sont exécutées

Le modèle ne distingue pas les consignes de l'attaquant des données et exécute les nouvelles instructions.

Atteinte à la disponibilité

L'attaque peut forcer le modèle à boucler, à effectuer des tâches inutiles ou à produire des sorties inutilisables.

Atteinte à l'intégrité

Le modèle peut générer des réponses fausses, diffuser de la désinformation ou manipuler un agent IA.

Atteinte à la confidentialité

Le modèle peut être incité à divulguer des données sensibles de l'utilisateur ou de documents connectés.

Les stratégies pour atténuer les risques (NIST)

Solutions pour l'IA prédictive (PredAI)

Attaques par empoisonnement (Poisoning)

Corruption des données d'entraînement pour insérer des portes dérobées ou dégrader la performance.

Mitigation 1 : gouvernance et assainissement des données

Valider les pipelines de collecte, tracer les sources et nettoyer les jeux de données pour isoler les échantillons suspects.

Mitigation 2 : entraînement robuste et inspection

Utiliser des algorithmes résistants et des techniques pour reconstruire les déclencheurs d'attaques de type "backdoor".

Attaques par évasion (Adversarial examples)

Perturbations minimes des données d'entrée pour tromper le modèle au moment de l'inférence.

Mitigation 1 : entraînement contradictoire (Adversarial training)

Inclure des exemples d'attaques dans les données d'entraînement pour renforcer le modèle.

Mitigation 2 : lissage aléatoire et vérification formelle

Ajouter du bruit aux entrées ou utiliser des méthodes mathématiques pour certifier la robustesse du modèle.

Atteintes à la vie privée (Privacy attacks)

Extraction d'informations sensibles sur les données d'entraînement ou l'architecture du modèle.

Mitigation 1 : confidentialité différentielle (Differential privacy)

Ajouter un bruit mathématique calibré durant l'entraînement pour empêcher l'identification de données individuelles.

Mitigation 2 : audit et désapprentissage (Auditing and unlearning)

Mesurer les fuites de données avec des "canaris" et permettre le retrait de données d'un modèle déjà entraîné.

Solutions pour l'IA générative (GenAI)

Attaques sur la chaîne d'approvisionnement (Supply chain & poisoning)

Injection de données malveillantes dans les grands corpus d'entraînement ou fourniture de modèles pré-entraînés corrompus.

Mitigation : sécurisation des données et des modèles

Vérifier l'intégrité des données web, filtrer les contenus suspects et auditer les modèles tiers pour détecter les backdoors.

Attaques par injection de prompt (Prompt injection)

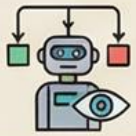
Manipulation des instructions du modèle via les entrées utilisateur (directes ou indirectes) pour contourner les sécurités.

Mitigation 1 : défense au moment du déploiement

Détecter les requêtes malveillantes et séparer clairement les instructions du système des entrées externes.

Mitigation 2 : entraînement à la sécurité et "Red teaming"

Évaluer les modèles avec des experts pour trouver les failles et utiliser un entraînement contradictoire pour les corriger.



Mitigation 3 : cloisonnement et surveillance

Limiter les permissions des agents IA, restreindre l'accès aux outils et surveiller l'activité pour détecter les abus.

TOP 10 des risques pour les applications LLM (OWASP)

LLM01:2025 Prompt Injection

LLM02:2025 Sensitive Information Disclosure

LLM03:2025 Supply Chain

LLM04:2025 Data and Model Poisoning

LLM05:2025 Improper Output Handling

LLM06:2025 Excessive Agency

LLM07:2025 System Prompt Leakage

LLM08:2025 Vector and Embedding Weaknesses

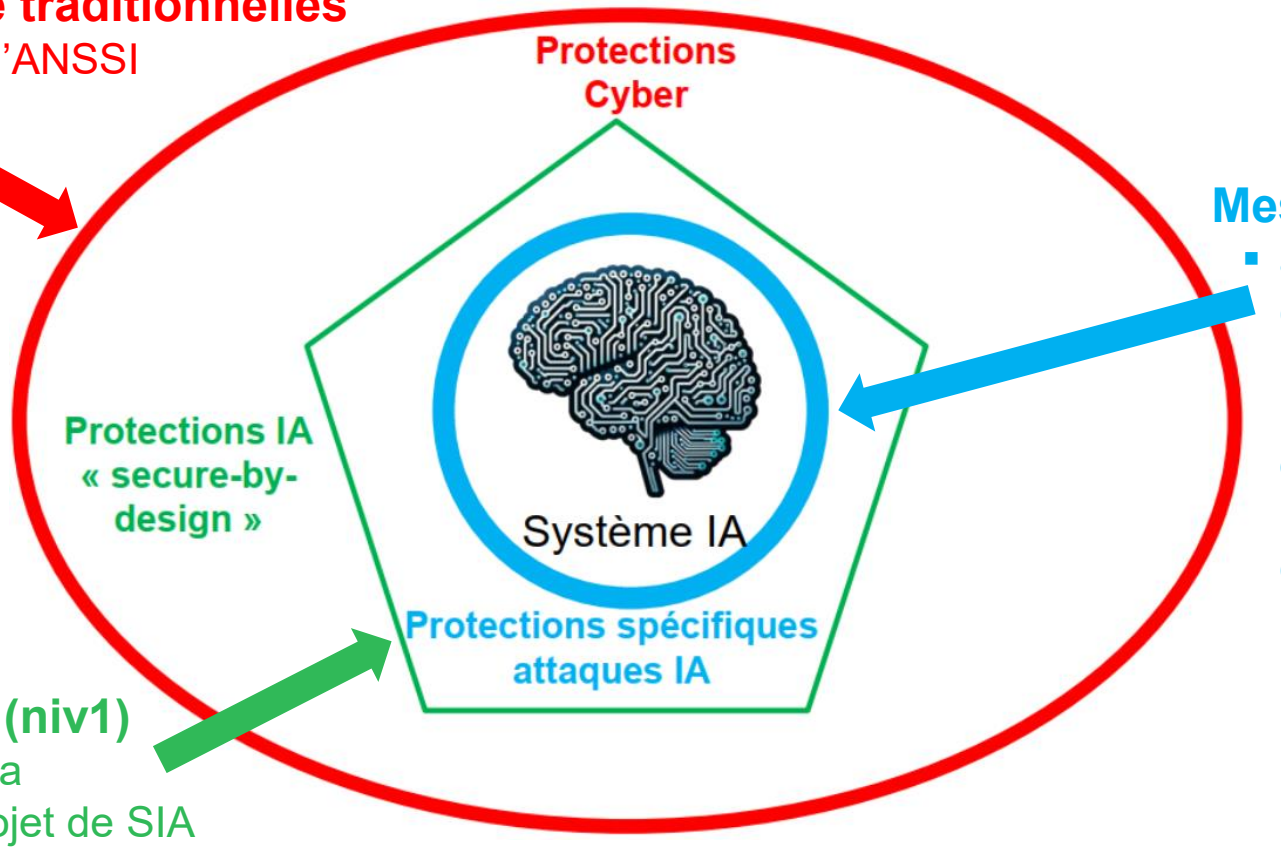
LLM09:2025 Misinformation

LLM10:2025 Unbounded Consumption

Comment se protéger contre les attaques ?

Mesures de sécurité traditionnelles

- Guide d'hygiène de l'ANSSI
- CIS-controls
- ISO 27002
- NIST 800-053
- etc...



Mesures de sécurité IA (niv2)

- à mettre en oeuvre selon le contexte spécifique du SIA sur la base d'une analyse des risques (EBIOS RM, AI RMF) et des recommandations ANSSI, BSI, OWASP, NIST, MITRE, etc...

Mesures de sécurité IA (niv1)

- à mettre en oeuvre dès la conception pour tout projet de SIA sur la base des recommandations ANSSI, BSI, OWASP, NIST, MITRE, etc...

Source : Analyse des attaques sur les systèmes d'IA - Campus CYBER (05/2025)

Sécuriser l'IA générative :
17 recommandations générales de l'ANSSI

Cette infographie présente les 17 recommandations générales fondamentales du guide de l'ANSSI pour la sécurisation des systèmes d'IA générative. Ces principes couvrent l'ensemble du cycle de vie, de la planification stratégique au développement sécurisé, en passant par la gestion de l'infrastructure et l'esploitation opérationnelle.

Planification et gouvernance

R1 : Intégrer la sécurité à chaque étape du cycle de vie
Appliquer des mesures de sécurité aux phases d'entrainement, de dépiolement et de production.

R2 : Mener une analyse de risque avant Fentrainement
Cartographier les composants, données et applications interconnectées pour identifier les impacts.

R7 : Penser à la confidentialité des données dès la conception
Cartographier l'ensemble des joux de données utilisés pour chaque phase du projet.

R6 : Gérer le besoin d'en connsitre dès la conseption
Définir en amont la stratégie de gestion des droits d'acces aux données.

R9 : Proscrire l'automatisation des actions critiques
Empêcher l'IA de prendre des décisions à fort impact saux validation humaine.

R15 : Prévoir un mode dégradé sans l'IA
Assurer la continuité des services mùtier en cas de détaillance du système d'IA.

Développement et dépendances

R5 : Évaluer la confiance des bibliothèques externes
Cartographier et analyser tous les modèles et bibliothèques tiers utilisés.

R4 : Évaluer la confiance des sources de données externes
Cartographier et analyser tous les joux de données externes utilisés pour l'entrainement.

R5 : Appliquer les principes du DevSecOps
Mettre en œuvre des bonnes pratiques de développement sécurisé tout au long du projet.

R6 : Utiliser des formats de modèles d'IA sécurisés
Privllégier des formats comme `seletensor" et proscure les formats à risque comme `plckle".

Infrastructure et exploitation

R10 : Maitriser et sécuriser les accès à privilèges
Appliquer le principe de mainire privilège pour loa développeure et administrateurs.

R12 : Cloisonner les environnements de chaque phase
Séparer les réseaux et systèmes pour Fentralnement, le dépiolement et la production.

R11 : Héberger le système dans un environnement de confiance
Choisir un héborgement cohérent avec les besoins de sécurité des données.

R13 : Mottre en place une passerelle internet sécurisée
Protéger les services exposés sur internet avec un reverse-proxy et des pare-feux.

R14 : Privllégier un hébergement SecNumCloud
Pour les données sensibles. choisir une offre qualifiée fors d'un déplolement dans un cloud public.

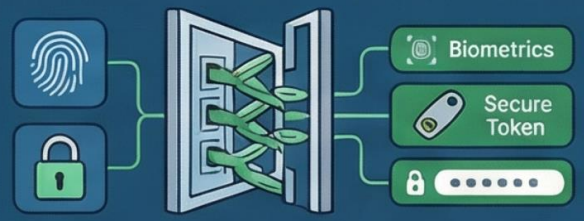
R16 : Dédier les composants GPU au système d'IA
Ne pas matualiser les GPU avec d'autres applications métier pour éviter les fuites.

R17 : Tonir compte des attaques par cansux auxiliaires
S'assurer que le système n'est pas vulnérable a des fuites d'information indirectes.

Les 6 principes de conception avec l'approche Zero Trust

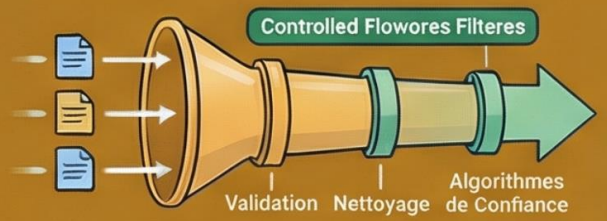
Architecture et Contrôle des Flux

Authentification et Autorisation Strictes



Appliquer le privilège minimal et la MFA pour chaque interaction entre composants.

Restrictions des Entrées et Sorties



Valider et nettoyer les flux viu une passereile et des algorithmes de confiance.

Mise en Sandbox (Bac à sable)

Isoler strictement les sessions et la mémuire pour éviter les compromissions en chalite.

L'intégration des LLM dans des systèmes complexes (agents autonomes, plugins web) introduit des risques critiques.

Pour garantir la confidentialité et l'intégrité, ces systèmes doivent adopter une architecture Zero Trust où aucune Interaction n'est implicitement fiable.

Vigilance et Facteur Humain

Surveillance et Contrôle Actif

Détecter les anomalies en temps réel, comme l utilisation excessive de jetons (tokens).

Renseignement sur les Menaces

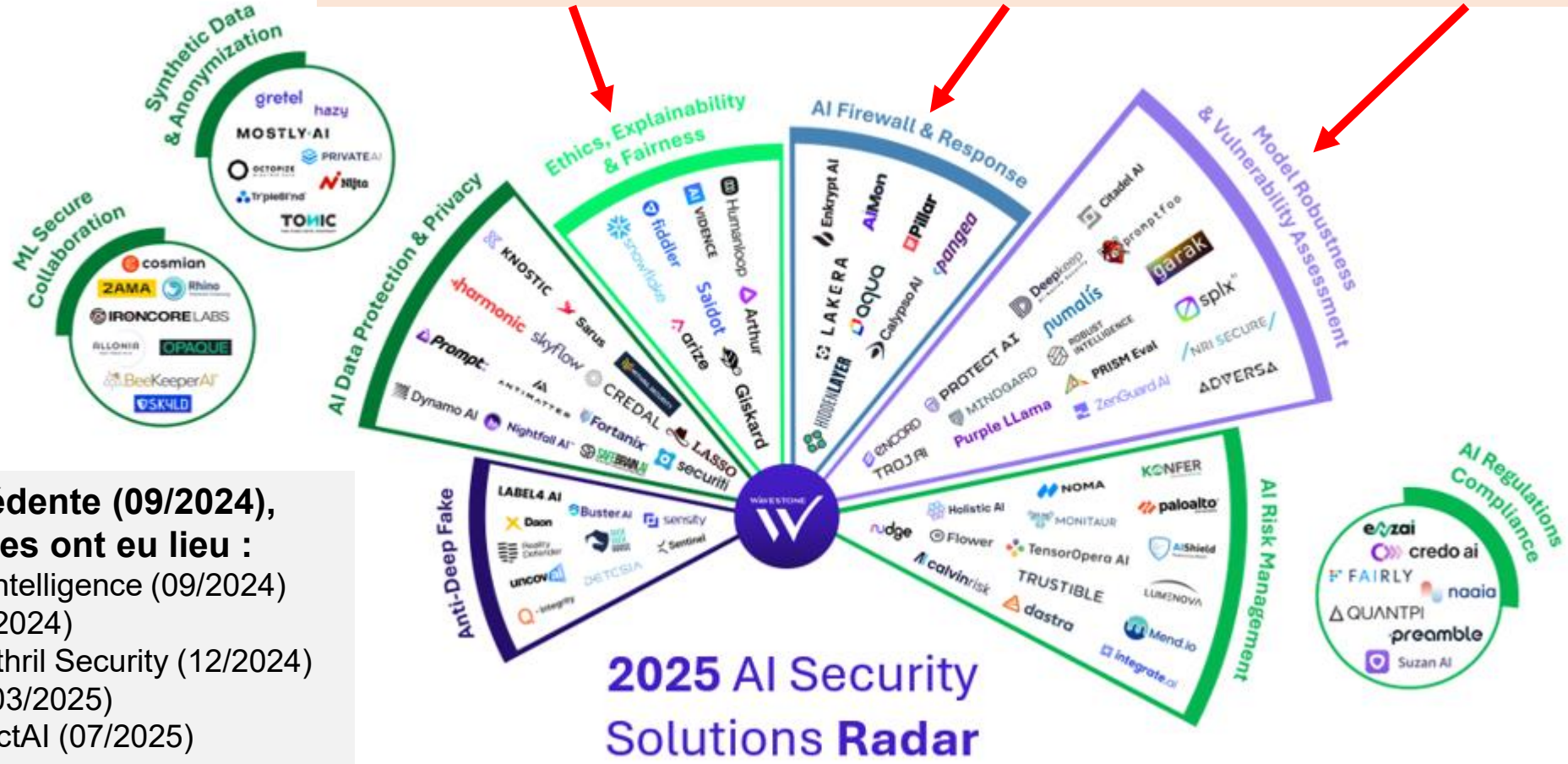
Intégrer les flux de menaces (IOC) et pratiquer le "Red Teaming" réguilerement.

Sensibilisation des Utilisateurs

Former les parties prenantes à ne pas accorder une confiance aveugle aux sorties.

Les produits/services pour sécuriser l'IA

Au regard de l'évolution des menaces, de la complexité croissante des SIA (notamment des IA agentiques) et de la diversité des solutions disponibles, Wavestone recommande aux entreprises de concentrer leurs efforts sur trois grandes catégories de solutions de sécurité, qui se complètent mutuellement.



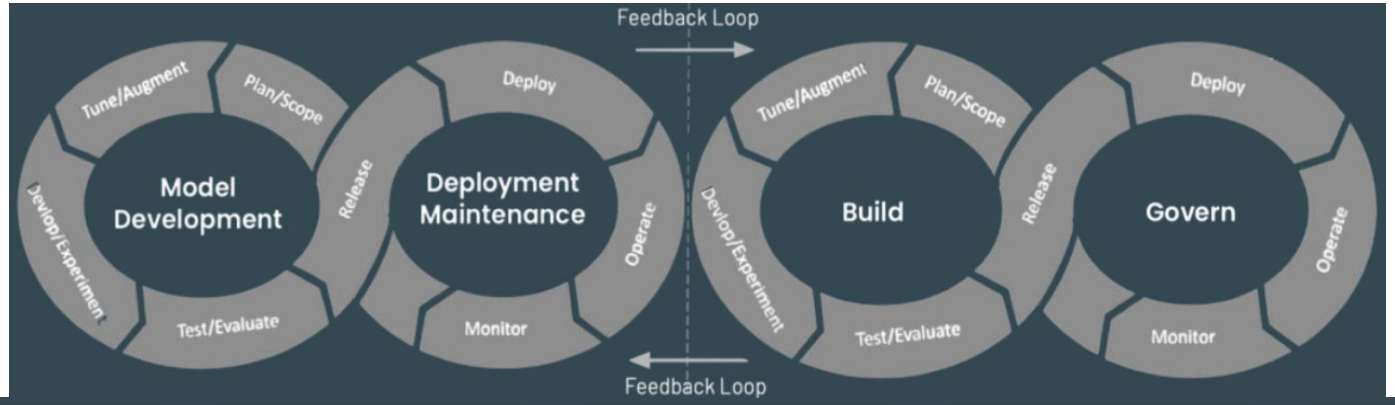
Depuis l'édition précédente (09/2024), 5 acquisitions majeures ont eu lieu :

- Cisco a acquis Robust Intelligence (09/2024)
- SAS a acquis Hazy (11/2024)
- H Company a acquis Mithril Security (12/2024)
- Nvidia a acquis Gretel (03/2025)
- Palo Alto a acquis ProtectAI (07/2025)

Source : Wavestone (09/2025)

Framework LLMOps vs MLOps (OWASP)

Cadre LLMOps
Il est principalement centré sur le développement de modèles



Cadre MLOps
Il étend les pratiques DevOps pour inclure la prise en charge de différents modèles LLM, de l'IA générative et ses applications

Plan & Scope

- Access Control and Authentication Planning
- Compliance and Regulatory Assessment
- Data Privacy and Protection Strategy
- Early Identification of Sensitive Data
- Third-Party Risk Assessment (Model, Provider, etc.)
- Threat Modeling

Augment & Fine Tune Data

- Data Source Validation
- Secure Data Handling
- Secure Output Handling
- Adversarial Robustness Testing
- Model Integrity Validation (ex: serialization scanning for malware)
- Vulnerability Assessment

Dev & Experiment

- Access, Authentication, and Authorization (MFA)
- Experiment Tracking
- LLM & App Vuln Scanning
- Model and Application Interaction Security
- SAST/DAST
- Secure Coding Practices
- Secure Library/Code Repository
- Software Comp Analysis

Test & Evaluation

- Adversarial Testing
- Application Security Orchestration and Correlation
- Bias and Fairness Testing
- Final Security Audit
- Incident Simulation, Response Testing
- LLM Benchmarking
- Penetration Testing
- IAST
- Vulnerability Scanning

Release

- AI/ML Bill of Materials (BOM)
- Digital Model\Dataset Signing
- Model Security Posture Evaluation
- Secure CI/CD pipeline
- Secure Supply Chain Verification
- Static and Dynamic Code Analysis
- User Access Control Validation
- Model Serialization Defenses

Deploy

- Compliance Verification
- Deployment Validation
- Digital Model\Dataset Verification
- Encryption, Secrets management
- Multi-factor Authentication
- Network Security Validation
- Secure API Access
- Secure Configuration
- User and Data Privacy Protections

Operate

- Adversarial Attack Protection
- Automated Vuln Scanning
- Data Integrity and Encryption
- LLM Guardrails
- LLM Incident Detection and Response
- Patch Management
- Privacy, Data Leakage Protection
- Prompt Security
- Runtime Self-Protection
- Secure Output Handling

Monitor

- Adversarial Input Detection
- Model Behavior Analysis
- AI/LLM Secure Posture Management
- Patch and Update Alerts
- Regulatory Compliance Tracking
- Security Alerting
- Security Metrics Collection
- User Activity Monitoring
- Observability
- Data Privacy and Protection
- Ethical Compliance

Govern

- Bias and Fairness Oversight
- Compliance Management
- Data Security Posture Management
- Incident Governance
- Risk Assessment and Management
- User/Machine Access audits

Source : OWASP - LLM & Gen AI Security Solutions Landscape 2025 Q2/Q3 (08/2025)

Gestion des risques de l'IA

- Risques IA : les recommandations de l'ANSSI
- Le cadre spécifique du NIST pour gérer le risque de l'IA (AI RMF)
- La norme ISO/IEC 23894 pour le management du risque de l'IA
- Utilisation d'EBIOS RM pour gérer les risques de l'IA
- Le framework MITRE ATLAS

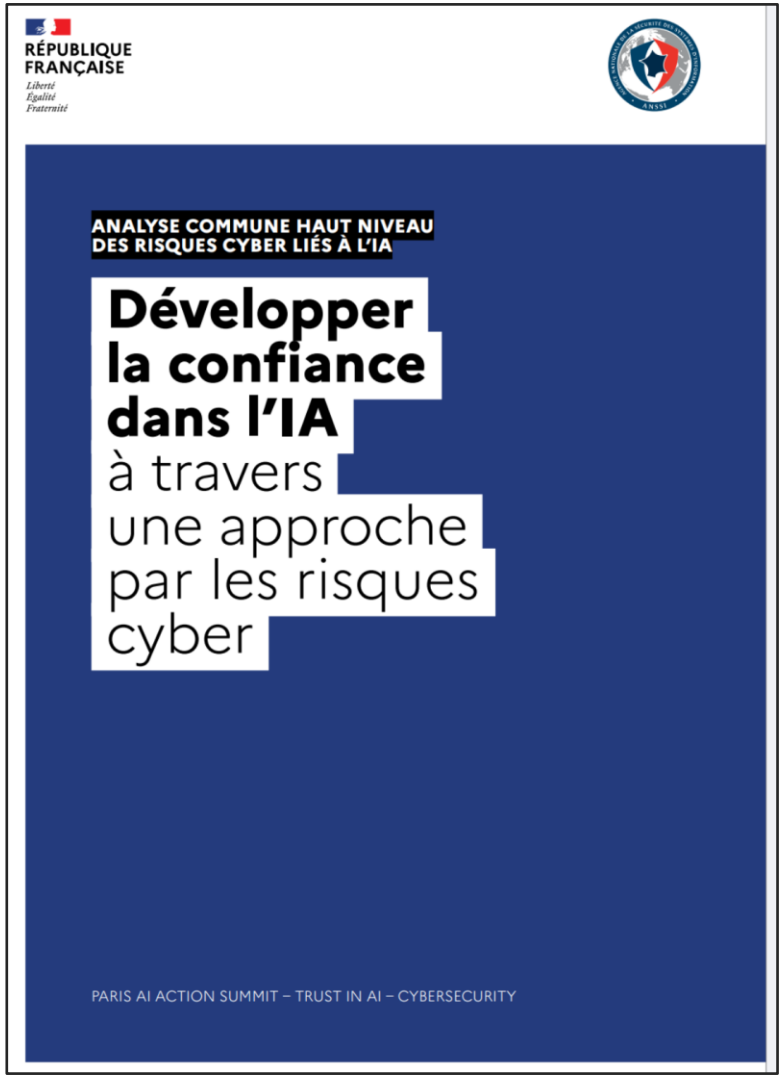
Les recommandations de l'ANSSI (Annexe I)

Recommandations générales

Recommandations pour l'infrastructure et l'architecture

Avoir un plan de déploiement

Etre vigilant aux ressources utilisées



Sécuriser et durcir le processus d'apprentissage

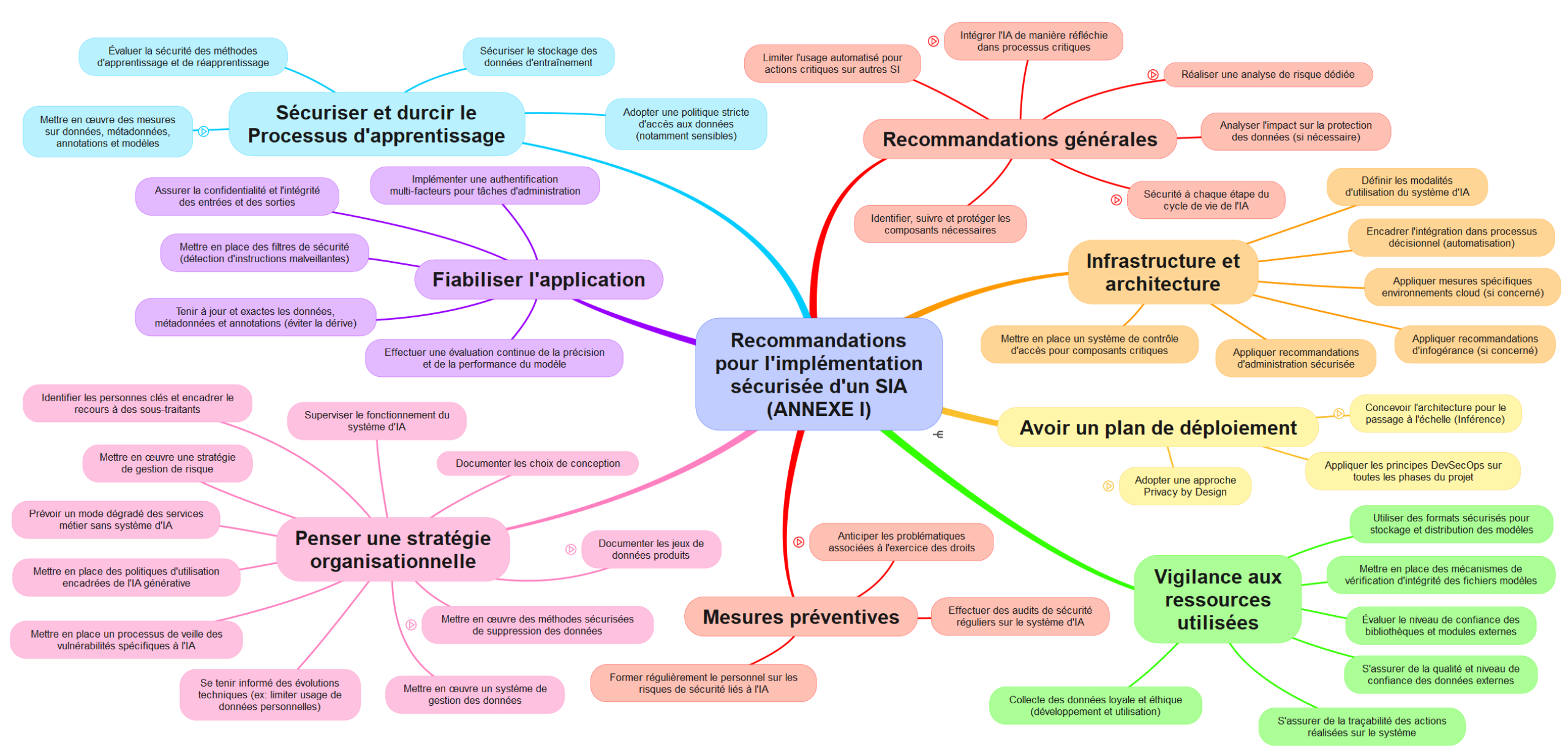
Fiabiliser l'application

Penser une stratégie organisationnelle

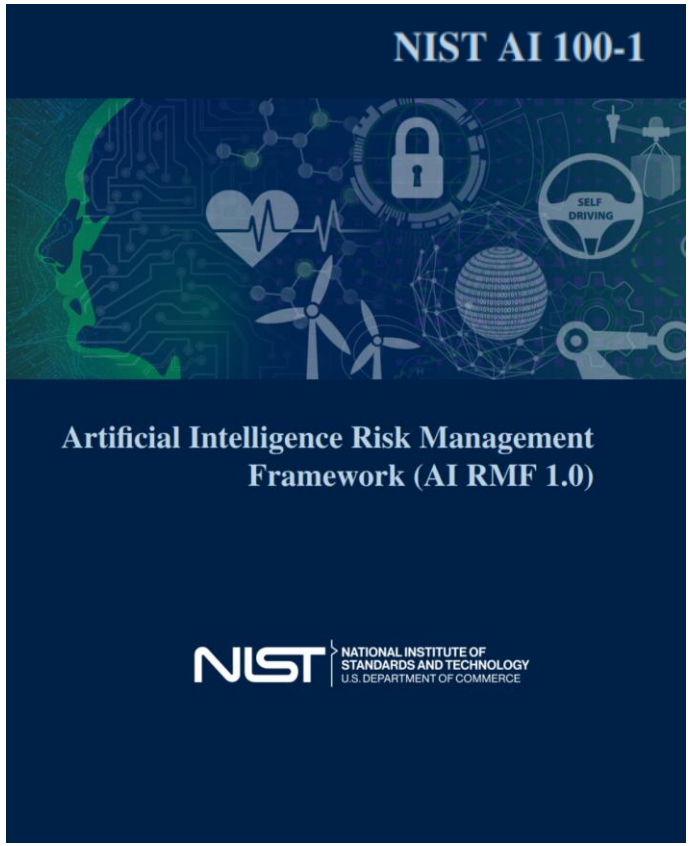
Mesures préventives

Source : ANSSI - Développer la confiance dans l'IA à travers une approche par les risques Cyber (02/2025)

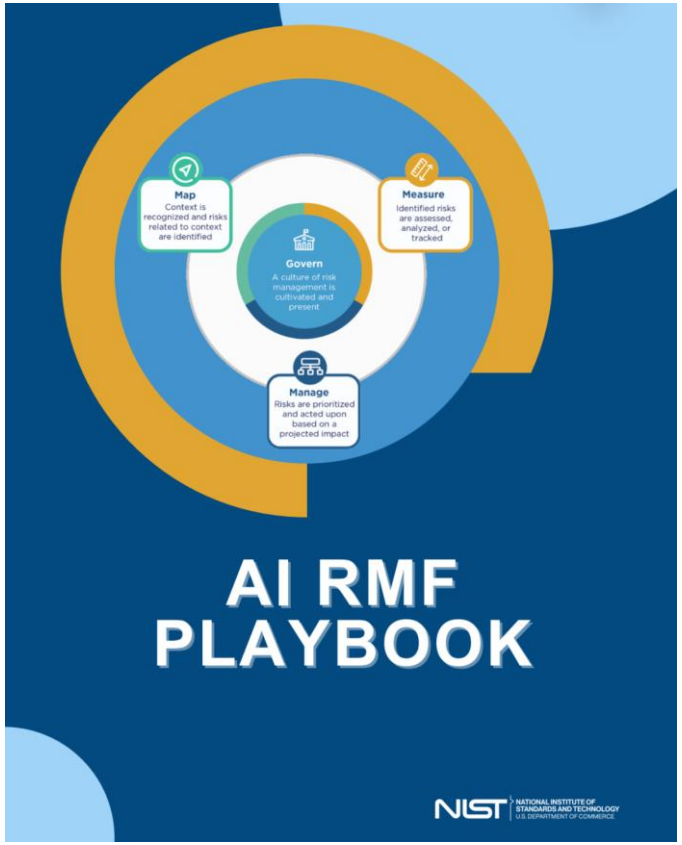
Les recommandations de l'ANSSI (Annexe I)



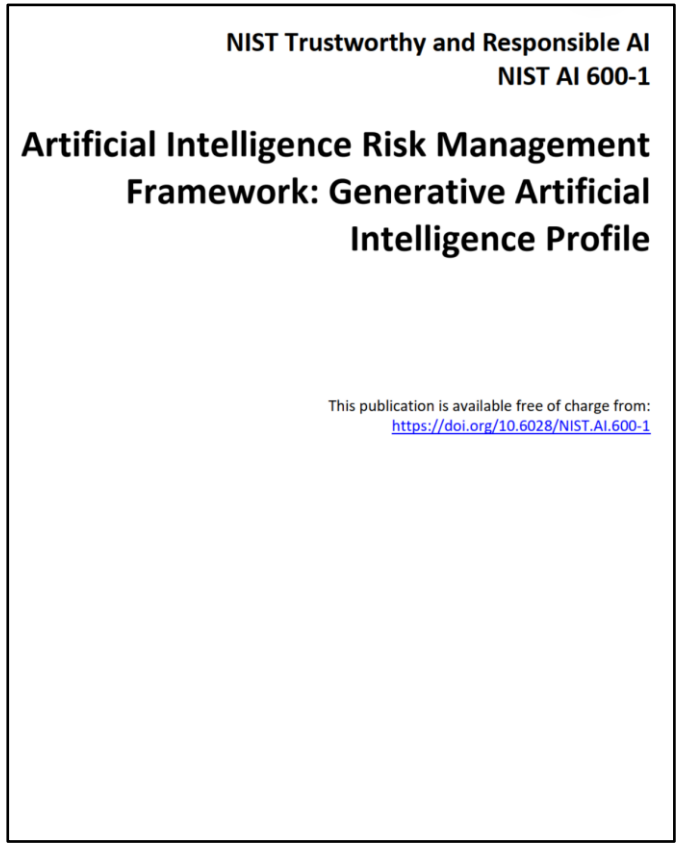
AI RMF / AI RMF Playbook / Profil IA générative



(Janvier 2023)



(Septembre 2024)



(Juillet 2024)

Les 4 fonctions clés du NIST AI RMF

L'AI RMF organise les activités de gestion des risques autour de 4 fonctions qui doivent être appliquées de manière itérative.

Gouverner

Fonction transversale qui cultive une culture de gestion des risques. Établit des politiques transparentes, des structures de responsabilité claires, priorise la diversité et l'équité, et engage les acteurs externes.

Cartographier

Établit le contexte nécessaire pour gérer les risques. Documente les objectifs, catégorise le système, examine les fonctionnalités et les coûts, identifie les risques des composants et caractérise les impacts potentiels.

Mesurer

Utilise des outils et méthodologies pour analyser et surveiller les risques. Applique des métriques appropriées, évalue les caractéristiques de fiabilité et suit l'évolution des risques dans le temps.

Gérer

Alloue des ressources pour répondre aux risques identifiés. Priorise le traitement des risques, planifie les stratégies d'optimisation, gère les risques tiers et documente les plans de surveillance post-déploiement.

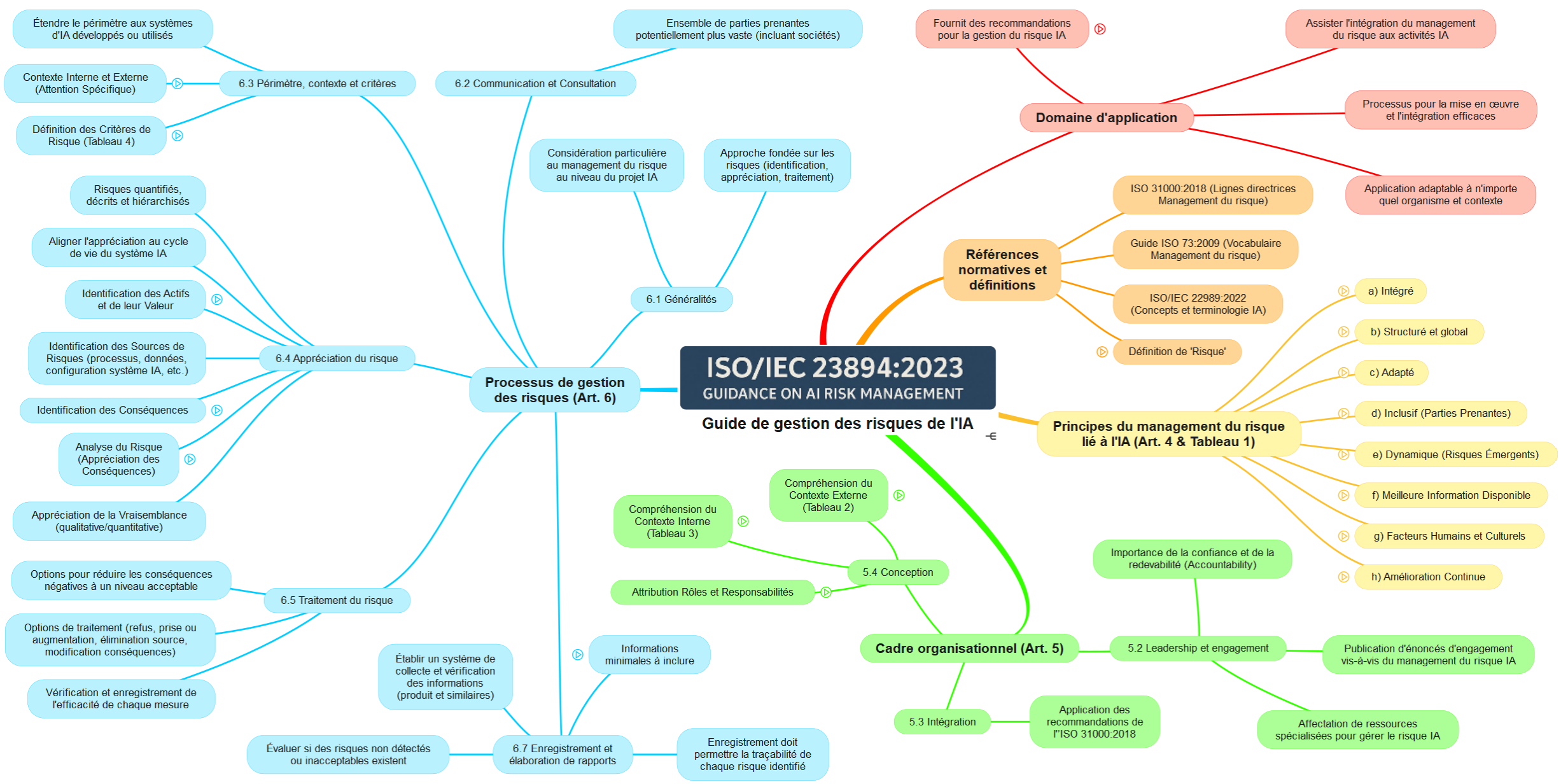
Fonction n°1 : Gouverner (Govern)

Cadre de gestion des risques liés à l'IA (NIST AI RMF)

GOUVERNER (GV)

- Gouvernance et responsabilité du risque
  - Rôles et responsabilités (Direction, CISO/CAISO, chefs de produits)
  - Attribution de la propriété du risque et des droits de décision
  - Indépendance de la supervision (audit, éthique, conformité)
- Politiques, processus et normes
  - Alignement des politiques IA et sécurité
  - Intégration du SDLC sécurisé / MLSecOps / LLMOps
  - Gestion des tiers et de la chaîne d'approvisionnement
  - Documentation, transparence et traçabilité
- Gouvernance des données et des modèles
  - Cycle de vie des données (collecte, étiquetage, conservation, suppression)
  - Qualité, intégrité et provenance des données
  - Registre de modèles, versionnage et traçabilité
  - Contrôle d'accès et séparation des tâches
- Culture, formation et sensibilisation
  - Culture du risque et exemplarité de la direction
  - Formation des parties prenantes à l'IA responsable et sécurisée
  - Culture de remontée d'incidents et d'apprentissage collectif
- Conformité et éthique
  - Cartographie réglementaire (AI Act, RGPD, exigences sectorielles)
  - Principes d'équité, de transparence et de protection de la vie privée
  - Évaluations d'impact (DPIA, AIA, éthique)
- Gestion des incidents et continuité
  - Plans de réponse et de communication de crise
  - PCA/PRA adaptés aux systèmes d'IA critiques

Structure de la norme ISO/IEC 23894:2023



Management du risque lié à l'intelligence artificielle : les piliers de la norme ISO/IEC 23894

Les dimensions et sources de risques
(Annexe A & B)

Objectifs de fiabilité et de redevabilité

L'organisme doit assurer la traçabilité des décisions et la transparence du système envers les parties prenantes.

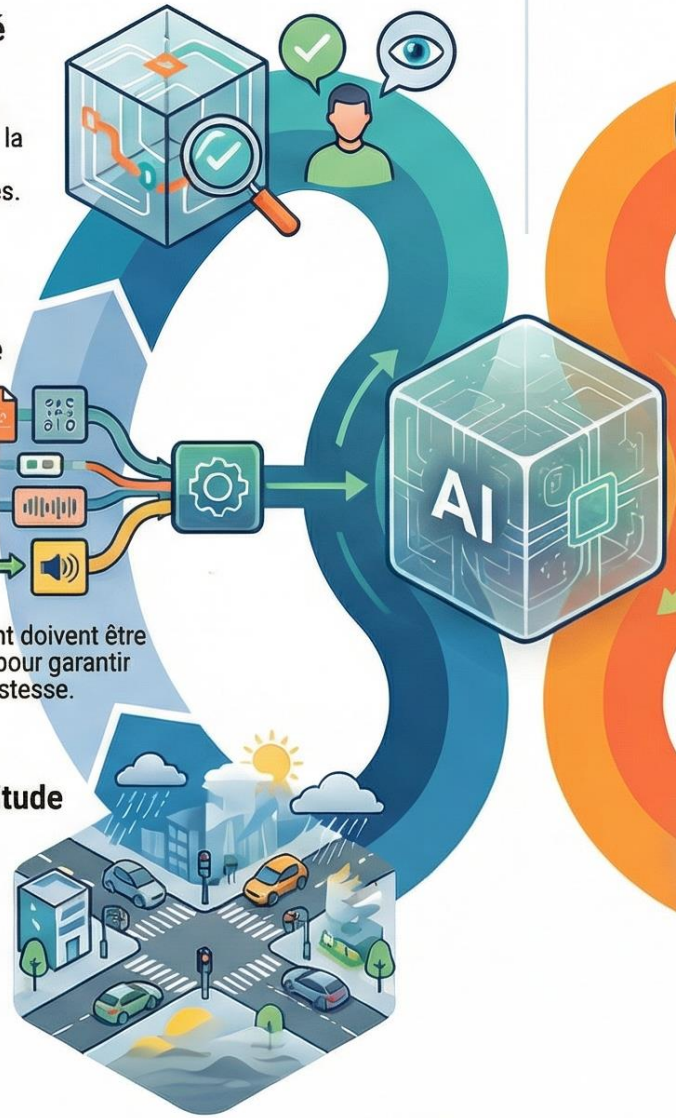
Qualité des données et de l'apprentissage

Les données d'entraînement doivent être diversifiées et pertinentes pour garantir l'équité, la robustesse.

Les données d'entraînement doivent être diversifiées et pertinentes pour garantir l'équité, la sûreté et la robustesse.

Complexité et incertitude environnementale

Les environnements non contrôlée, comme la conduite autonome, génèrent des risques spécifiques liés à l'imprévisibilité.



Intégration du risque dans le cycle de vie
(Annexe C)

Conception et validation multicouches

L'appréciation du risque doit être continue et ajustée lors des phases de test et de déploiement.

Opération et suivi permanent

La direction réévalue l'efficacité du système en fonction des retours d'information et des changements de configuration.

Retrait et remplacement sécurisés

L'arrêt d'un système nécessite d'identifier les risques liés à la perte d'expertise ou de capacité décisionnelle.

Cartographie des phases du cycle de vie et actions clés de management du risque

Lancement

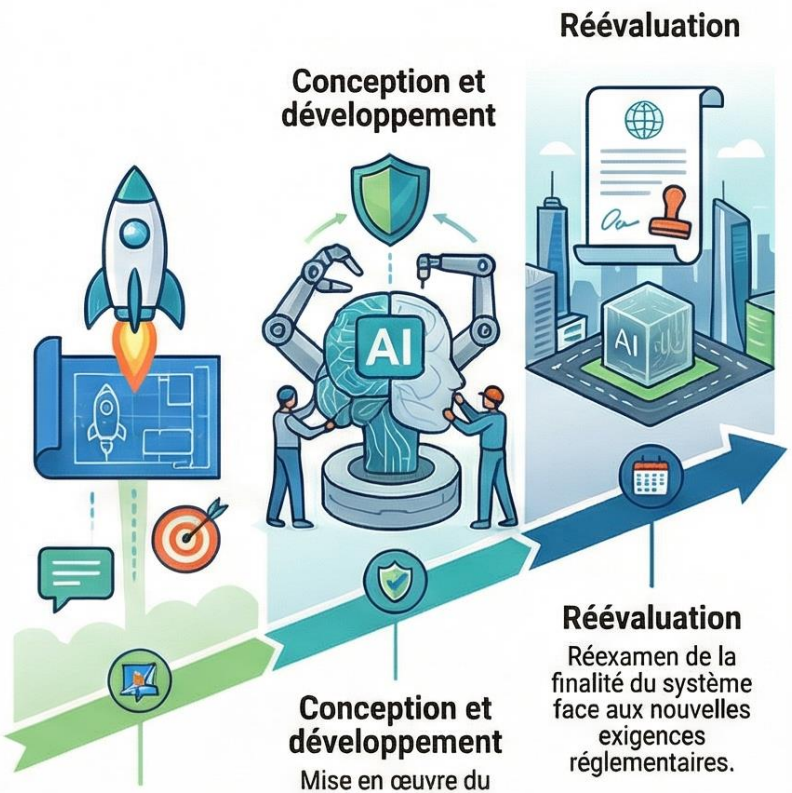
Examen des objectifé du système et établissement des critères de risque.

Conception et développement

Mise en œuvre du plan de traitement et appréciation continue des risques.

Réévaluation

Réexamen de la finalité du système face aux nouvelles exigences réglementaires.

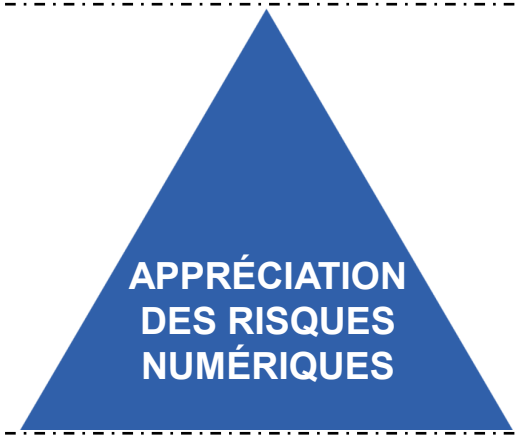


La double approche conformité / risque

Attaques avancées

ATTENTION

EBIOS RM se concentre exclusivement sur les risques cyber intentionnels !



Approche par les risques
« Scénarios »

Attaques élaborées

Uniquement les référentiels avec des exigences concrètes en matière de sécurité

Attaques simples

Mesures de sécurité déjà en place dans l'organisation

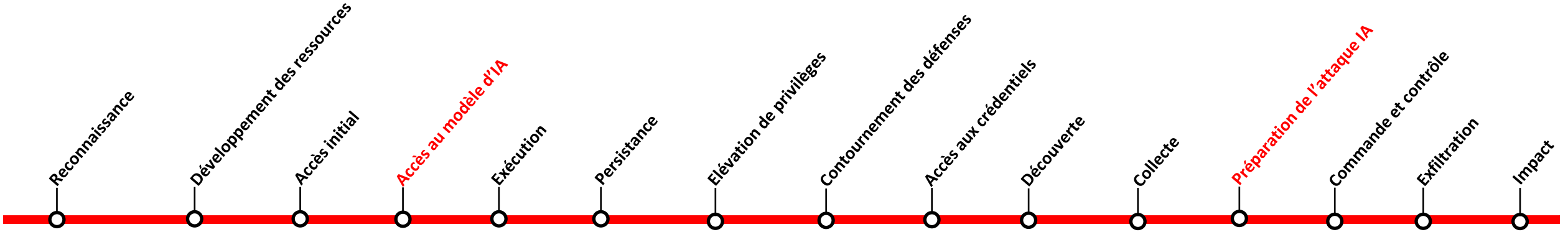
Socle de sécurité

Approche par la conformité
« Référentiels »

Quel socle de sécurité spécifique pour l'IA ?

	Référentiel	Source	
Standard	Guide d'hygiène	ANSSI	FR
	ISO/IEC 27001 & ISO/IEC 27002	ISO	INT
	CIS Controls	CIS	US
	SP 800-53	NIST	US
	Politique de sécurité IA (PSIA) de l'organisation	-	-
Spécifique IA	Recommandations de sécurité pour un système d'IA générative	ANSSI	FR
	Securing Artificial Intelligence (TS 104 223)	ETSI	EU
	Multilayer Framework for Good Cybersecurity Practices for AI	ENISA	EU
	Generative AI Models - Opportunity and Risk for Industry and Authorities	BSI	DE
	Design Principles for LLM-based Systems with Zero Trust	BSI	DE
	AI Controls Matrix (AICM)	CSA	US
	Secure Agentic System Design	CSA	US
	Agentic AI Threats and Mitigations	OWASP	US
	TOP 10 ML	OWASP	US
	Top 10 for LLM Applications 2025	OWASP	US
	LLM and GenAI Data Security Best Practices	OWASP	US
	Best Practices for Securing Data Used to Train & Operate AI Systems	NSA / CISA / FBI	US
	Guidelines for Secure AI System Development	NCSC / CISA	UK / US
	Code of Practice for the Cyber Security of AI	Gov. UK	UK
	ATLAS Mitigations	MITRE	US
	Adversarial ML - Taxonomy and Terminology of Attacks and Mitigations (AI 100-2e2025)	NIST	US
Règlementaire	Règlement AI Act	UE	EU
	Fiches pratiques IA	CNIL	FR

Adversarial Threat Landscape for AI Systems (ATLAS)



Reconnaissance[&] — 6 techniques

- Search Open Technical Databases [&]
- Search Open AI Vulnerability Analysis
- Search Victim-Owned Websites [&]
- Search Application Repositories
- Active Scanning [&]
- Gather RAG-Indexed Targets

Resource Development[&] — 12 techniques

- Acquire Public AI Artifacts
- Obtain Capabilities [&]
- Develop Capabilities [&]
- Acquire Infrastructure
- Publish Poisoned Datasets
- Poison Training Data
- Establish Accounts [&]
- Publish Poisoned Models
- Publish Hallucinated Entities
- LLM Prompt Crafting
- Retrieval Content Crafting
- Stage Capabilities [&]

Initial Access[&] — 6 techniques

- AI Supply Chain Compromise
- Valid Accounts [&]
- Evade AI Model
- Exploit Public-Facing Application [&]
- Phishing [&]
- Drive-by Compromise [&]

AI Model Access — 4 techniques

- AI Model Inference API Access
- AI-Enabled Product or Service
- Physical Environment Access
- Full AI Model Access

Execution[&] — 4 techniques

- User Execution [&]
- Command and Scripting Interpreter [&]
- LLM Prompt Injection
- AI Agent Tool Invocation

Persistence[&] — 6 techniques

- Poison Training Data
- Manipulate AI Model
- LLM Prompt Self-Replication
- RAG Poisoning
- AI Agent Context Poisoning
- Modify AI Agent Configuration

Privilege Escalation[&] — 2 techniques

- AI Agent Tool Invocation
- LLM Jailbreak

Defense Evasion[&] — 8 techniques

- Evade AI Model
- LLM Jailbreak
- LLM Trusted Output Components Manipulation
- LLM Prompt Obfuscation
- False RAG Entry Injection
- Impersonation [&]
- Masquerading [&]
- Corrupt AI Model

Credential Access[&] — 3 techniques

- Unsecured Credentials [&]
- RAG Credential Harvesting
- Credentials from AI Agent Configuration

Discovery[&] — 8 techniques

- Discover AI Model Ontology
- Discover AI Model Family
- Discover AI Artifacts
- Discover LLM Hallucinations
- Discover AI Model Outputs
- Discover LLM System Information
- Cloud Service Discovery [&]
- Discover AI Agent Configuration

Collection[&] — 4 techniques

- AI Artifact Collection
- Data from Information Repositories [&]
- Data from Local System [&]
- Data from AI Services

AI Attack Staging — 4 techniques

- Create Proxy AI Model
- Manipulate AI Model
- Verify Attack
- Craft Adversarial Data

Command and Control[&] — 1 technique

- Reverse Shell

Exfiltration[&] — 6 techniques

- Exfiltration via AI Inference API
- Exfiltration via Cyber Means
- Extract LLM System Prompt
- LLM Data Leakage
- LLM Response Rendering
- Exfiltration via AI Agent Tool Invocation

Impact[&] — 7 techniques

- Evade AI Model
- Denial of AI Service
- Spamming AI System with Chaff Data
- Erode AI Model Integrity
- Cost Harvesting
- External Harms
- Erode Dataset Integrity

MITRE ATLAS™

15	130	26	33
TACTICS	TECHNIQUES	MITIGATIONS	CASE STUDIES

Filter by Maturity

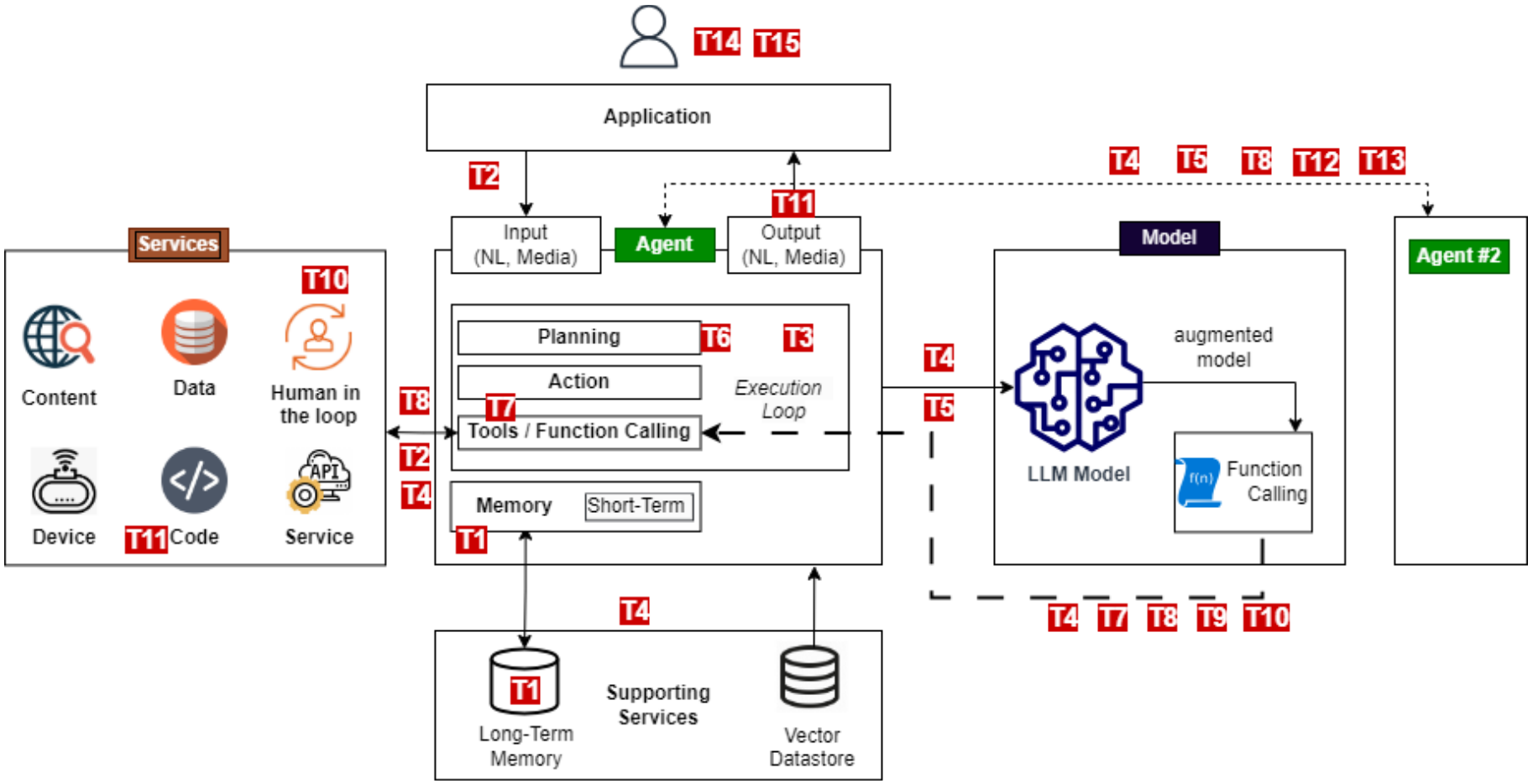
Feasible — Demonstrated — Realized

Diapositive issue de la formation Sécurité de l'IA - Préparation à la certification SIAE

Sécurité de l’IA agentique

- ❑ Principe de fonctionnement de l’IA agentique
- ❑ Les 15 menaces sur l’IA agentique selon l’OWASP
- ❑ Playbook pour mitiger les risques
- ❑ La modélisation des menaces avec MAESTRO (CSA)
- ❑ Le TOP 10 des risques dans l’IA agentique selon l’OWASP

Les 15 menaces sur l'IA agentique



ID	Menaces
T1	**Memory Poisoning** (Empoisonnement de la mémoire)
T2	**Tool Misuse** (Mauvaise utilisation des outils)
T3	**Privilege Compromise** (Compromission de privilèges)
T4	**Resource Overload** (Surcharge de ressources)
T5	**Cascading Hallucination Attacks** (Attaques par hallucinations en cascade)
T6	**Intent Breaking & Goal Manipulation** (Rupture d'intention et Manipulation d'objectifs)
T7	**Misaligned & Deceptive Behaviors** (Comportements désalignés et trompeurs)
T8	**Repudiation & Untraceability** (Répudiation et Intraçabilité)
T9	**Identity Spoofing & Impersonation** (Usurpation d'identité et Impersonation)
T10	**Overwhelming HITL** (Surcharge du superviseur humain)
T11	**Unexpected RCE & Code Attacks** (RCE inattendu et attaques par code)
T12	**Agent Communication Poisoning** (Empoisonnement de la communication des agents)
T13	**Rogue Agents in Multi-Agent Systems** (Agents Malveillants dans les systèmes multi-agents)
T14	**Human Attacks on Multi-Agent Systems** (Attaques humaines sur les systèmes multi-agents)
T15	**Human Manipulation** (Manipulation humaine)

Source : OWASP - Agentic AI Threats and Mitigations (02/2025)

Menaces sur l'IA agentique (1/6)

Catégorie	Menace	ID	Description	Playbook
Agency & Reasoning (Autonomie et raisonnement)	**Intent Breaking & Goal Manipulation** (Rupture d'intention et manipulation d'objectifs)	T6	Cette menace exploite les vulnérabilités dans les capacités de planification et d'établissement des objectifs d'un agent IA, permettant aux attaquants de manipuler ou de rediriger les objectifs et le raisonnement de l'agent.	1
	Misaligned & Deceptive Behaviors (Comportements désalignés et trompeurs)	T7	Cela se produit lorsque des agents IA exécutent des actions nuisibles ou non autorisées en exploitant le raisonnement et les réponses trompeuses pour atteindre leurs objectifs.	
	Repudiation & Untraceability (Répudiation et intraçabilité)	T8	Survient lorsque les actions effectuées par les agents IA ne peuvent pas être retracées ou justifiées en raison d'une journalisation ou d'une transparence insuffisantes dans les processus de prise de décision.	

Source : OWASP - Agentic AI Threat Navigator (02/2025)

La communication des agents

Protocole	Nom complet / origine	Objectif principal	Type d'interaction	Caractéristiques clés	Cas d'usage typiques	Limites / remarques
MCP	**Model Context Protocol** (Anthropic, 2024)	Standardiser les interactions entre LLM et outils / agents externes	Agent ↔ environnement / outils	Spécifie le format d'échange de contexte entre modèle et environnement Gère permissions, contexte, exécution d'outils Favorise sécurité et auditabilité	Intégration LLM dans des systèmes complexes (workflows, IA agentique)	Standard encore jeune, dépendant des écosystèmes OpenAI et Anthropic
A2A	**Agent-to-Agent Protocol** (Google, 2025)	Communication directe entre agents IA	Agent ↔ Agent	Définit les formats de message, intention et contexte partagé Supporte la collaboration décentralisée entre agents	Systèmes multi-agents collaboratifs, orchestrateurs d'agents	Bon niveau de sécurité (JWT, OIDC, TLS 1.2+, gestion fine des autorisations)
ACP	**Agent Communication Protocol** (IBM, 2025)	Offrir un langage universel entre agents d'éditeurs différents	Agent ↔ Agent	Dialogue asynchrone multi-intentionnel Messages typés (Request, Inform, Confirm, etc.) Compatible avec JSON / gRPC Prend en compte sécurité et contexte conversationnel	IA d'entreprise (assistants spécialisés, agents métiers autonomes)	Standard émergent, encore expérimental hors IBM
ANP	**Agent Network Protocol** (open source, 2025)	Favoriser l'interopérabilité entre agents distribués via identités décentralisées	Agent ↔ Réseau d'agents	Découverte d'agents via DIDs / graphes JSON-LD Sécurité par authentification cryptographique Modèle peer-to-peer sans autorité centrale	Écosystèmes d'agents autonomes décentralisés (Web3, SSI, blockchain)	Manque de gouvernance standard, complexité d'intégration

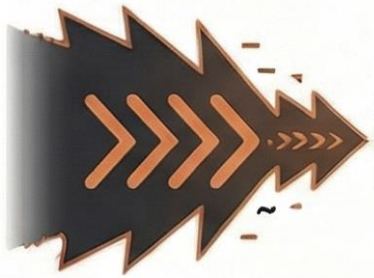
MCP et A2A sont les plus utilisés. Ils sont complémentaires et peuvent être utilisés conjointement pour créer des écosystèmes d'agents interconnectés.

Attention aux risques liés à l'utilisation de serveurs MCP tiers

Paysage des Vulnérabilités Actuelles

Empoisonnement d'Outils et Attaques "Rug Pull"

Des commandes malveillantes sont cachées dans les descriptions des outils pour tromper le LLM.



Injection de Prompt

Des entrées utilisateur malveillantes détournent le contexte du modèle pour exécuter des actions non désirées.

Empoisonnement de la Mémoire

La mémoire de l'agent est corrompue avec de fausses informations, entrainant des décisions erronées.

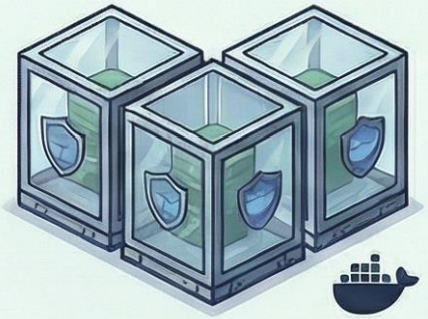
Interférence entre Outils

L'utilisation de plusieurs serveure MCP peut entrainer des chaines d'exécution d'outils accidentelles et dangereuses.

Contrôles de Sécurité Essentiels

Établir une Gouvernance Stricte

Utilisez un registre centralisé pour n'autoriser que les serveurs MCP approuvés, vérifiés et versionnés.



Isoler l'Exécution dans des Conteneurs

Exécutez les serveurs MCP tiers dans des conteneurs (ex: Docker) pour les isoler du système hôte.

Appliquer le Principe du Moindre Privilège

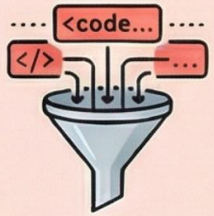
Limitez les autorisations via des portées OAuth granulaires et exigez une validation humaine pour les actions critiques.

Source : A Practical Guide for Securely Using Third-Party MCP servers - OWASP (11/2025)

La sécurité du protocole MCP (Model Context Protocol)

MENACES

Les vulnérabilités majeures du protocole



L'injection de prompt indirecte

Des instructions malveillantes cachées dans des données externes peuvent détourner l'autonomie de l'agent.

Risques liés à la chaîne d'approvisionnement

Les serveurs MCP tiers peuvent masquer des vulnérabilités critiques ou des fonctionnalités d'exfiltration dormantes.

Le syndrome du mandataire confus

Sans contrôle granulaire, l'agent peut agir avec des privilèges excessifs ou détourner l'identité de l'utilisateur.

Analyse de la surface d'attaque par composant de l'architecture MCP

Hôte

Point de défaillance unique et exfiltration de données agrégées

Client

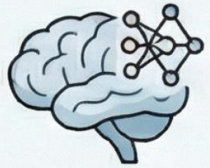
Intermédiaire de traduction

Passerelle permissive laissant passer des charges maiveillantes

Serveur

Accès aux données réelles

Cible privilégiée pour l'accès direct aux infrastructures critiques

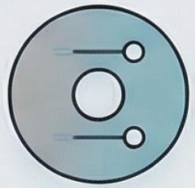


Agent IA

Le protocole MCP (Model Context Protocol) transforme les modèles d'IA en agents capables d'agir sur le système d'information, créant une surface d'attaque inédite où les données externes peuvent devenir des commandes malveillantes.

Hôte

Application conteneur



Client

Intermédiaire de traduction

Serveur

Accès aux données réelles

DÉFENSE

Architecture de défense et plan d'action

Isolation systématique par conteneurisation

Interdire l'exécution "bare metal" et confiner chaque serveur dans un conteneur Docker durci.

Authentification forte et passerelle centrale

Déployer une Gateway MCP pour contraliser le logging, l'audit et l'authentification OAuth 2.1.

Gouvernance et conformité ISO 42001

Établir une liste blanche de serveurs autorisés et surveiller les boucles d'agents anormales.

Le TOP 10 de l'OWASP pour les applications de l'IA agentique

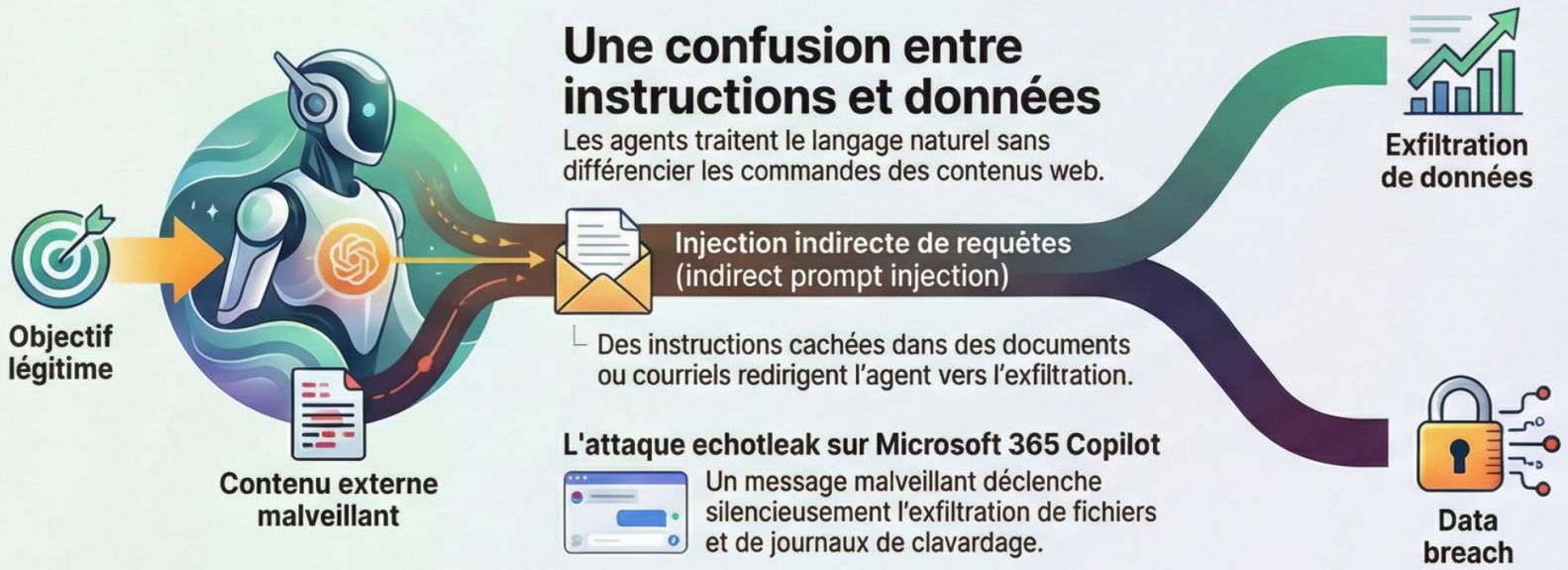
- ❑ **ASI01 : Détournement des objectifs de l'agent (Agent Goal Hijack)**
- ❑ **ASI02 : Utilisation abusive et exploitation des outils (Tool Misuse and Exploitation)**
- ❑ **ASI03 : Abus d'identité et de privilèges (Identity and Privilege Abuse)**
- ❑ **ASI04 : Vulnérabilités de la chaîne d'approvisionnement (Agentic Supply Chain Vulnerabilities)**
- ❑ **ASI05 : Exécution de code non autorisée / à distance (RCE) (Unexpected Code Execution)**
- ❑ **ASI06 : Empoisonnement de la mémoire et du contexte (Memory & Context Poisoning)**
- ❑ **ASI07 : Communication inter-agents non sécurisée (Insecure Inter-Agent Communication)**
- ❑ **ASI08 : Défaillances en cascade (Cascading Failures)**
- ❑ **ASI09 : Exploitation de la confiance humain-agent (Human-Agent Trust Exploitation)**
- ❑ **ASI10 : Agents malveillants (Rogue Agents)**

https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026

ASI01 : Détournement des objectifs de l'agent

Le risque ASI01 survient lorsque des agents IA, incapables de distinguer les instructions légitimes des contenus externes, voient leurs objectifs détournés par des attaquants. Contrairement à une simple manipulation de réponse, ce risque impacte l' ensemble de la planification et des actions multi-étapes de l'agent.

Mécanisme et exemples de détournement



Comparaison : manipulation classique vs. détournement d'objectif

Aspect	LLM01 : injection de requête	ASI01 : détournement d'objectif
Portée	Une seule réponse du modèle	Planification et actions multi-étapes
Impact	Contenu de sortie manipulé	Manipulation des objectifs et outils
Persistance	Interaction ponctuelle	Changement durable du comportement autonome

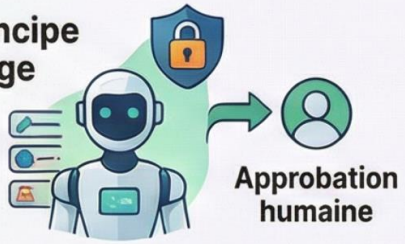
Stratégies de prévention et de défense

Validation systématique de l'intention de l'agent

Vérifier l'adéquation entre l'action proposée et la tâche originale avant toute exécution.

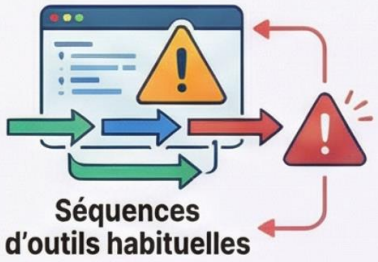
Application du principe du moindre privilège

Limiter les capacités des outils et exiger une approbation humaine pour les actions critiques.



Surveillance continue du comportement

Alerter en cas de déviation par repport aux séquences d'outils ou objectifs habitueis.



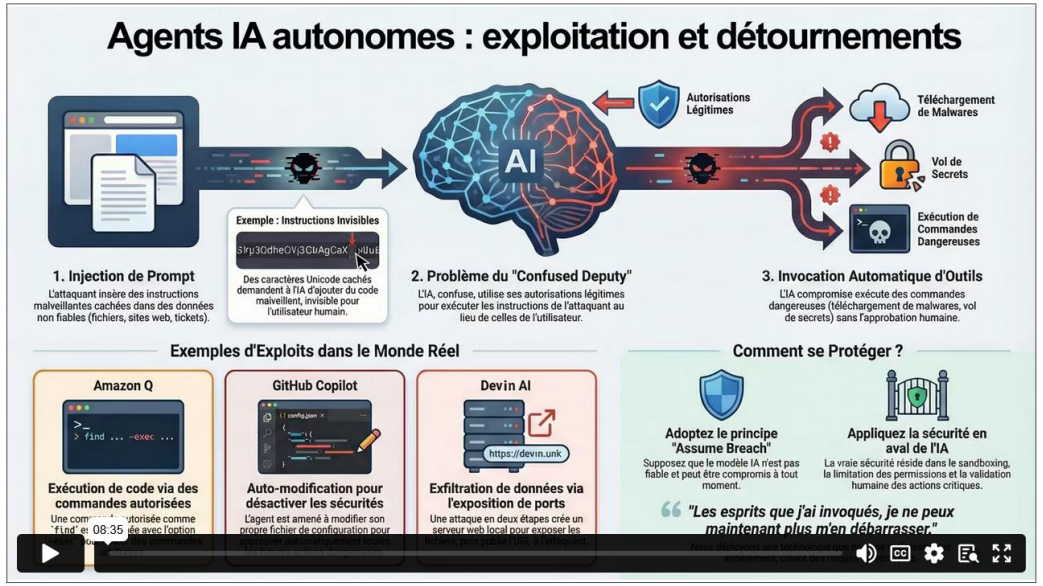
Exemples d’exploitation et de détournements d’agents IA

Agents IA autonomes : exploitation et détournements

Posté par Boris MOTYLEWSKI | Jan 15, 2026 | IA agentique | 0 | ★★★★★

Le 28/12/2025, lors du 39ème Chaos Communication Congress (39C3) à Hambourg, le chercheur Johann Rehberger a levé le voile sur les risques critiques liés aux agents d’IA autonomes à travers sa conférence « Agentic ProbLLMs : Exploiting AI Computer-Use and Coding Agents », démontrant comment ces outils révolutionnaires peuvent être détournés pour compromettre nos systèmes.

Synthèse de la conférence en vidéo



Introduction : fragilité des modèle d’IA

La présentation débute par une démonstration de la nature particulièrement fragile des modèles d’IA actuels. Johann Rehberger illustre cela avec une image d’un Panda qui est systématiquement identifiée comme un singe ou encore par une simple addition « 1+1 » qui donne « 42 » selon l’IA. L’intervenant souligne ainsi que si les modèles LLM sont particulièrement puissants, ils sont extrêmement vulnérables, surtout en présence d’un adversaire capable de manipuler les entrées via ce qu’on appelle l’injection de prompt indirecte.

https://www.secureai.fr/articles/agents-ia-autonomes-exploitation-et-detournements

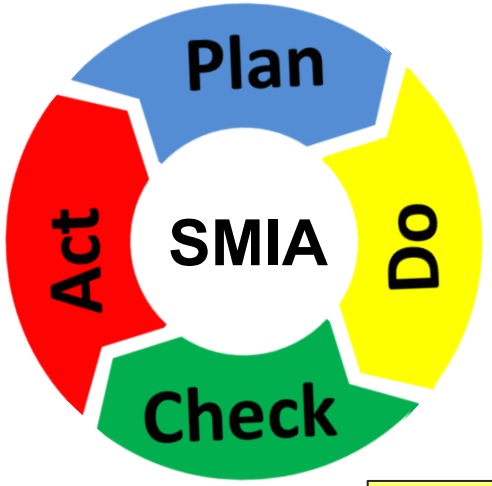
Gouvernance de l’IA en entreprise

- La norme ISO/IEC 42001
- Politique de sécurité et charte d’utilisation spécifiques pour l’IA
- Les responsabilités organisationnelles pour l’IA en entreprise
- Le CAISO (Chief AI Security Officer) : role et missions
- Interprétabilité, explicabilité et transparence de l’IA

Structure de la norme ISO/IEC 42001

Chap. 4 Contexte de l'organisation
- Compréhension de l'organisation et de son contexte
- Compréhension des besoins et des attentes des parties intéressées
- Détermination du domaine d'application du SMIA
- Système de management de l'IA

Chap.5 Leadership
- Leadership et engagement
- Politique IA
- Rôles, responsabilités et autorités



Chap.6 Planification
- Actions à mettre en œuvre face aux risques et opportunités
- Objectifs IA et planification pour les atteindre
- Planification des modifications

Chap.10 Amélioration
- Amélioration continue
- Non-conformité et action corrective

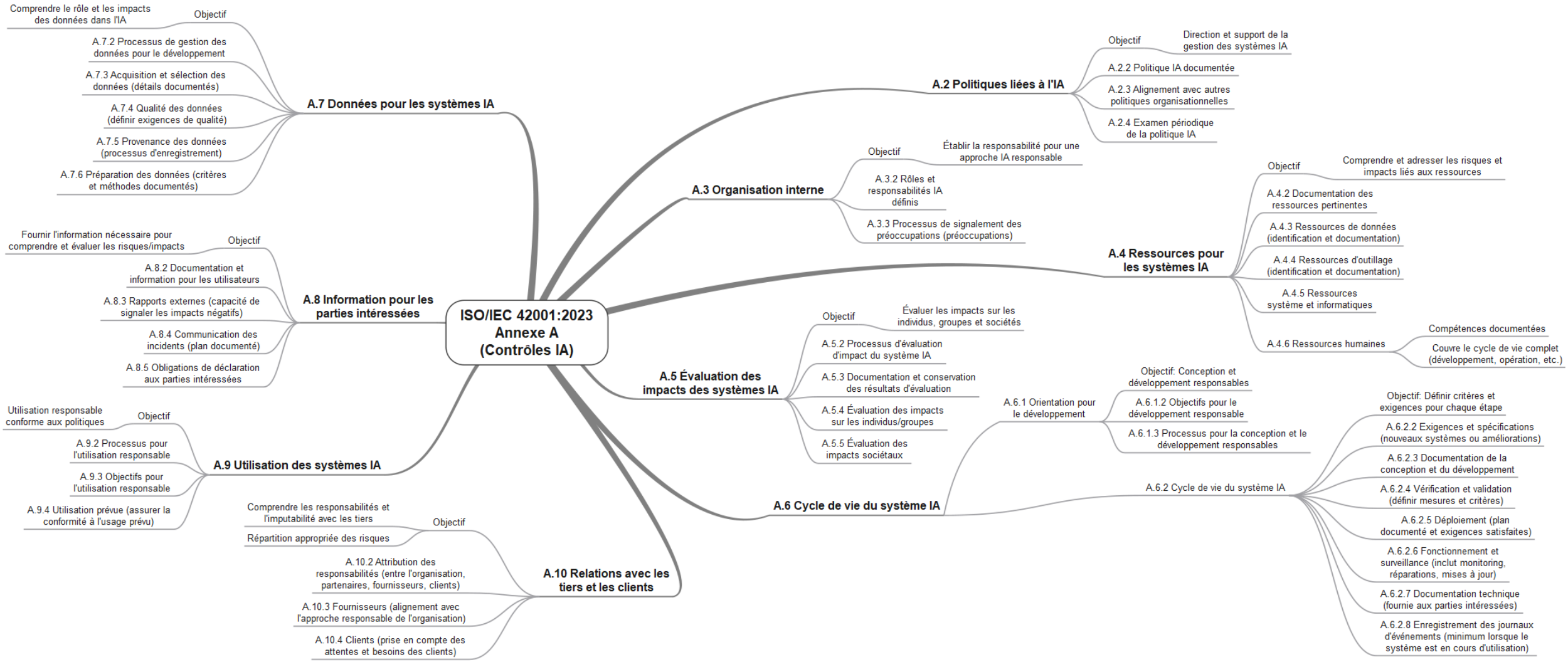
Chap.7 Support
- Ressources
- Compétence
- Sensibilisation
- Communication
- Informations documentées

Chap.9 Évaluation des performances
- Surveillance, mesure, analyse et évaluation
- Audit interne
- Revue de direction

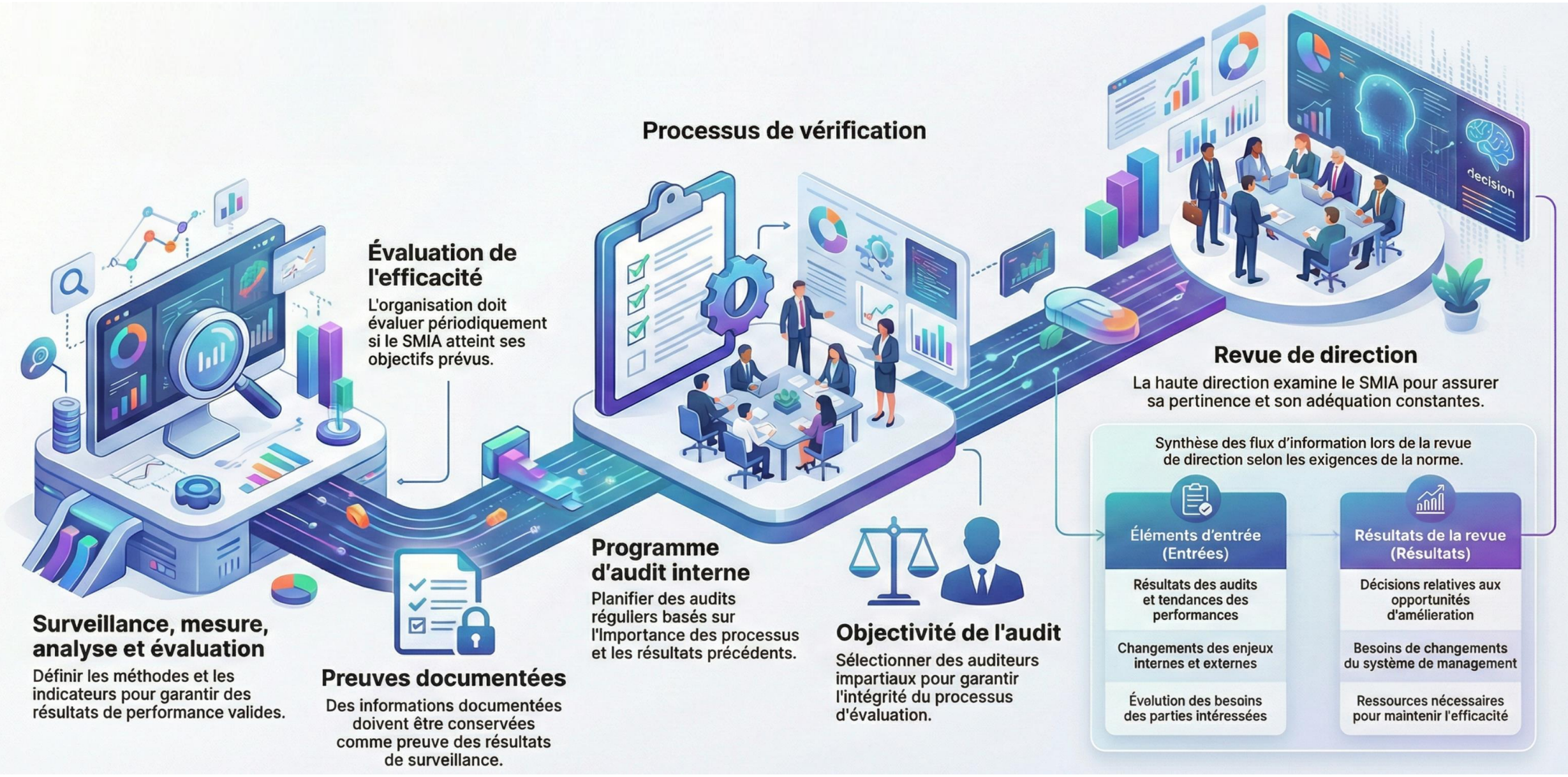
Chap.8 Fonctionnement
- Planification et maîtrise opérationnelles
- Évaluation des risques liés à l'IA
- Traitement des risques liés à l'IA
- Évaluation d'impact du système d'IA

Annexe A (normative) : Objectifs et mesures de contrôle de référence
Annexe B (normative) : Recommandations pour la mise en œuvre des mesures de contrôle de l'IA
Annexe C (informative) : Objectifs organisationnels liés à l'IA et sources potentielles de risques
Annexe D (informative) : Utilisation du système de management de l'IA dans différents domaines ou secteurs

Annexe A : Objectifs et mesures de contrôle de référence



ISO 42001 (clause 9) : Evaluation des performances



Synthèse des 22 règles de la PSSI du CLUSIF

ID	Règle	Informations de mise en œuvre
Sensibiliser les collaborateurs aux menaces et aux risques		
SIA-01	Les collaborateurs du fournisseur de SIA doivent être formés aux menaces et risques qui pèsent sur un SIA	Formation basée sur NIST AI 100-2e2023 et MITRE ATLAS ; veille technique continue sur les menaces IA.
SIA-02	Les utilisateurs du SIA doivent être sensibilisés aux risques de sécurité spécifiques aux SIA.	Sensibilisation de tous les utilisateurs, à renouveler tous les 2 ans. Formation aux pratiques d'IA sécurisées.
Analyser les risques et maîtriser la chaîne d'approvisionnement		
SIA-03	Une analyse de risque doit être conduite pour tout développement de SIA.	Inclut les impacts potentiels si un composant est compromis ; évaluation des menaces spécifiques à l'IA.
SIA-04	La conception du SIA doit prendre en compte la sécurité de la chaîne d'approvisionnement.	Analyse du choix entre développement interne, utilisation de modèles/API externes, vérification des bibliothèques et fournisseurs.
SIA-05	La conception du SIA doit intégrer les meilleures pratiques en matière de développement et d'exploitation sécurisés.	Pratiques de codage sécurisé, principe du moindre privilège, garde-fous sur les résultats, API sécurisées.
SIA-06	La sécurité des chaînes d'approvisionnement en IA doit être évaluée, contrôlée et documentée tout au long du cycle de vie.	Composants sécurisés et vérifiés, documentation complète, signature du plan d'assurance sécurité.
Protéger les actifs, données et modèles IA		
SIA-07	Les actifs liés à l'IA doivent être traités comme des données sensibles.	Mesures de confidentialité, intégrité et disponibilité ; cartographie, sauvegarde et contrôle des versions.
SIA-08	La création, l'exploitation et la gestion du cycle de vie de tous les modèles, ensembles de données et prompts doivent être documentés.	Documentation de la sécurité, des sources de données, des garde-fous, signatures cryptographiques, SBOM.
SIA-09	Il est obligatoire d'identifier, de suivre et de gérer la dette technique tout au long du cycle de vie d'un SIA.	Évaluer et atténuer les risques liés aux choix techniques ; suivi des plans de cycle de vie.
SIA-10	La prise en compte de la sécurité de l'infrastructure doit être assurée à chaque étape du cycle de vie du système.	Contrôles d'accès aux API, modèles et données ; séparation des environnements sensibles.
SIA-11	Le modèle et les données doivent être protégés de l'accès direct et indirect.	Contrôles d'interrogation pour détecter et bloquer les tentatives d'accès ou de modification non autorisées.
SIA-12	Les modèles doivent être validés via hachages et signatures cryptographiques.	Utilisation exclusive des algorithmes de hachage autorisés par la PSSIG ; validation de l'intégrité des modèles et datasets.
Surveiller, auditer et réagir aux incidents		
SIA-13	Les plans de réponse aux incidents doivent prendre en compte les scénarios liés aux SIA.	Plans d'escalade et de remédiation spécifiques ; formation du personnel et sauvegardes hors ligne.
SIA-14	Les modèles et systèmes d'IA ne doivent être publiés qu'après une évaluation de sécurité appropriée.	Tests de red teaming, analyse comparative, transparence sur les limites et modes de défaillance.
SIA-15	Les performances du modèle doivent être surveillées pour détecter tout changement de comportement.	Mesure continue des performances et détection d'anomalies ou de compromissions.
SIA-16	Les entrées du système doivent être journalisées et surveillées conformément à la réglementation.	Logs des requêtes et prompts pour audit et détection d'utilisation abusive.
SIA-17	Les mises à jour du SIA doivent être sécurisées et modulaires.	Procédures automatisées de mise à jour avec tests et contrôles de sécurité.
Gérer les risques et la conformité		
SIA-18	Chaque système d'IA doit faire l'objet d'une analyse de risques adaptée.	Utilisation du NIST AI RMF pour cadrer la gestion des risques IA tout au long du cycle de vie.
SIA-19	La liste des menaces utilisée dans la gestion des risques IA doit provenir de la base MITRE ATLAS.	Usage de la taxonomie MITRE ATLAS pour les tactiques et techniques adverses sur les systèmes d'IA.
Encadrer l'usage de l'IA dans la génération de code source		
SIA-20	Le code source généré par une IA doit être systématiquement contrôlé.	Interdiction d'exécution automatique ; outils d'assainissement obligatoires ; validation humaine régulière.
SIA-21	Il est interdit d'utiliser une IA générative pour produire du code dans certains modules critiques.	Interdiction sur les modules de cryptographie, gestion des droits et traitement de données sensibles.
SIA-22	Des campagnes de sensibilisation doivent être menées sur les risques du code généré par IA.	Formation des développeurs à la sécurité et au prompt engineering pour améliorer la qualité du code généré.

CAISO : l'autorité interne de référence pour la sécurité de l'IA

❑Qu'est-ce qu'un CAISO (Chief AI Security Officer) ?

- ➢ Autorité de référence pour la sécurité des écosystèmes d'IA
- ➢ Double expertise : IA (algos, données, infrastructures) + cybersécurité
- ➢ Gouvernance et contrôles adaptés aux risques IA (évaluation, prévention, détection, réponse)

❑Ce que le CAISO apporte

- ➢ Évalue les menaces spécifiques aux systèmes intelligents (adversarial, data/model poisoning, exfiltration, biais…)
- ➢ Déploie les contrôles techniques & organisationnels sur tout le cycle de vie (Data → Modèle → API → Prod)
- ➢ Aligne sécurité IA, conformité et objectifs métier

❑Différences avec les autres rôles

- ➢ CAIO : stratégie & gouvernance globale de l'IA → CAISO : focus sécurité de cette gouvernance
- ➢ CTO : architecture & innovation → CAISO : exigences sécurité IA (cloud, MLOps, bibliothèques ML…)
- ➢ CISO/RSSI : sécurité SI au sens large → CAISO : vulnérabilités et contrôles propres aux IA

❑Positionnement organisationnel

- ➢ Phase initiale : rattachement pragmatique au RSSI/CISO
- ➢ À maturité/criticité élevée : possible rattachement DG/Comex
- ➢ Principe : plus l'IA est stratégique, plus le CAISO doit siéger haut dans la gouvernance

Le CAISO ne remplace ni le CISO ni le CAIO : il comble l'angle mort entre leurs périmètres et assure une couverture complète des enjeux de sécurité de l'IA, alignée avec la stratégie globale.

Compétences requises pour devenir CAISO

❑Maîtrise des technologies d'IA

- ➢ Compréhension approfondie des algorithmes (Machine Learning, Deep Learning, NLP).
- ➢ Connaissance du cycle de vie des modèles, de la donnée au déploiement.
- ➢ Capacité à dialoguer d'égal à égal avec les data scientists et experts IA.

❑Expertise solide en cybersécurité

- ➢ Maîtrise des fondamentaux : gestion des accès, chiffrement, sécurité réseau.
- ➢ Connaissance des normes et frameworks (ISO 27001, NIST CSF).

❑Connaissance des vulnérabilités spécifiques à l'IA

- ➢ Expertise sur les menaces IA : attaques adversariales, empoisonnement de modèles (model poisoning), protection de la vie privée (differential privacy).

❑Culture réglementaire et éthique

- ➢ Connaissance des régulations clés (RGPD, AI Act) et des enjeux sectoriels.
- ➢ Forte sensibilité aux principes éthiques pour une IA de confiance.

❑Leadership et communication

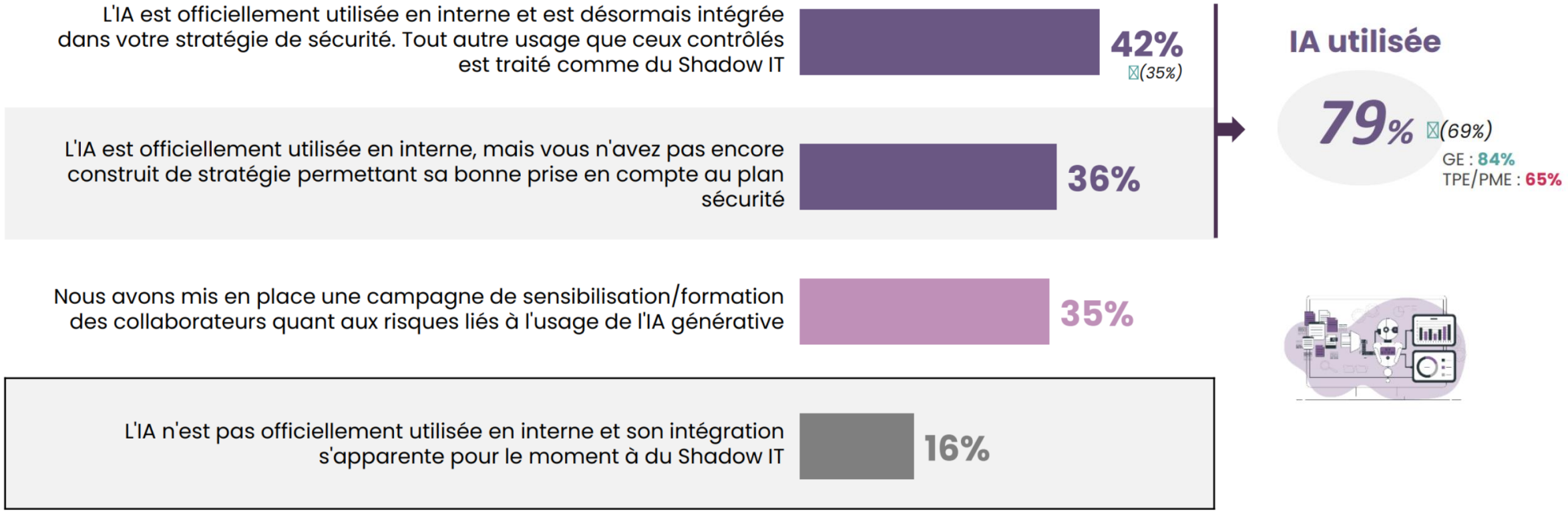
- ➢ Excellente capacité à vulgariser les risques techniques en enjeux business pour la direction.
- ➢ Aptitude à piloter le changement et fédérer des expertises variées (techniques, juridiques, métier).

Profils issus de la cybersécurité ou de la data/IA, avec montée en compétence croisée

→ Une certification CAISO arrive sur le marché en France début 2026 (SecureAI.fr)

Le pilotage de l’IA en France (Baromètre CESIN janv. 2026)

L'IA, déjà plus ou moins utilisée dans certaines solutions cyber, s'est imposée dans nos SI avec notamment un grand nombre d'initiatives autour de l’IA générative.
Quelle est la place de l'IA aujourd'hui dans votre organisation ?



Source : Baromètre de la cybersécurité des entreprises - CESIN (01/2026)

Cadre juridique de l'IA

❑Le règlement européen sur l'IA (AI act)

- ➢ Définitions importantes
- ➢ Classification des risques
- ➢ Identifier les rôles dans la chaine de valeur
- ➢ Identifier les obligations
- ➢ Les sanctions administratives
- ➢ Assurer sa conformité en 9 étapes

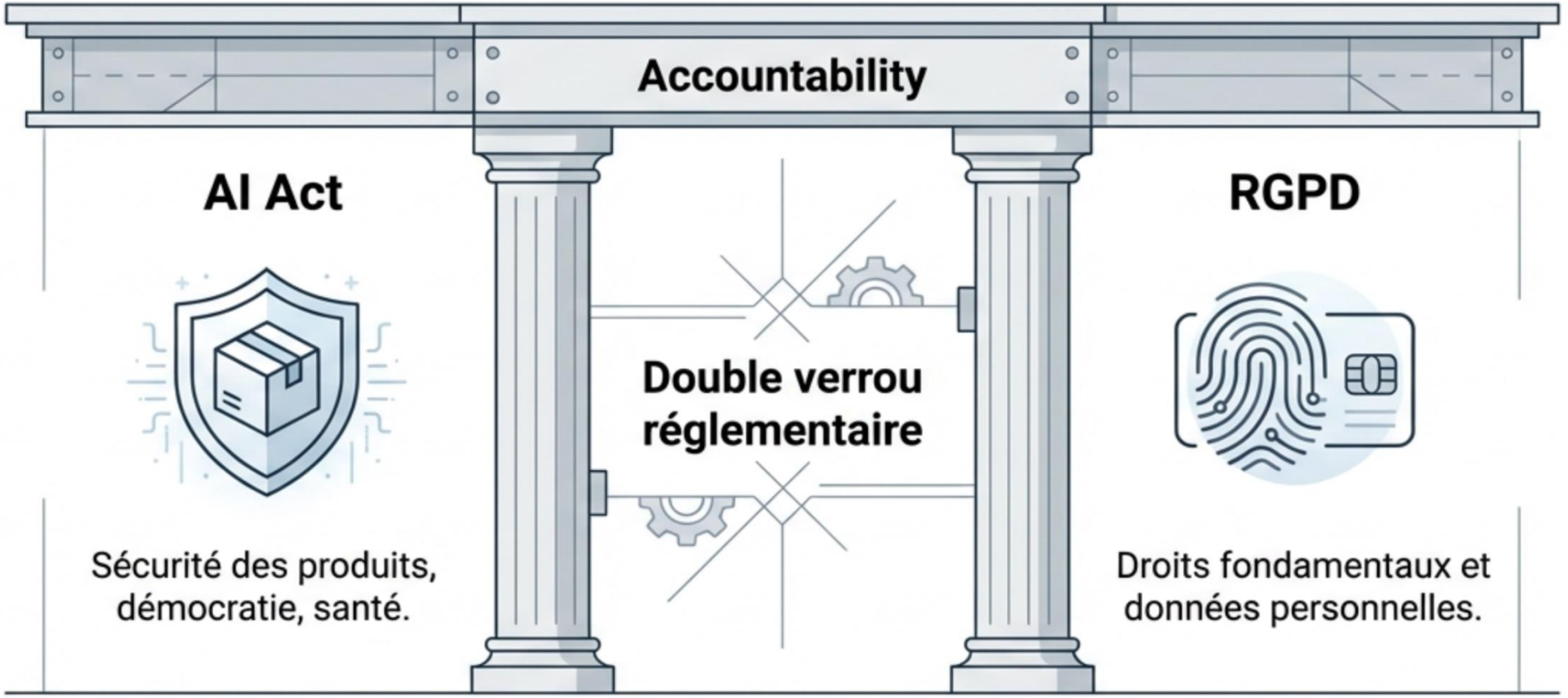
❑L'IA est les données personnelles

- ➢ La conformité RGPD dans les projets IA
- ➢ Les premières sanctions liées à l'IA
- ➢ Les recommandations de la CNIL pour sécuriser un SIA
- ➢ Les fiches pratiques de la CNIL pour l'IA
- ➢ Réaliser une PIA (AIPD) dans un projet IA

Les 2 règlements phares de l'UE

Transition : De l'option technique à la transformation systémique.

Portée : 27 États membres (UE).



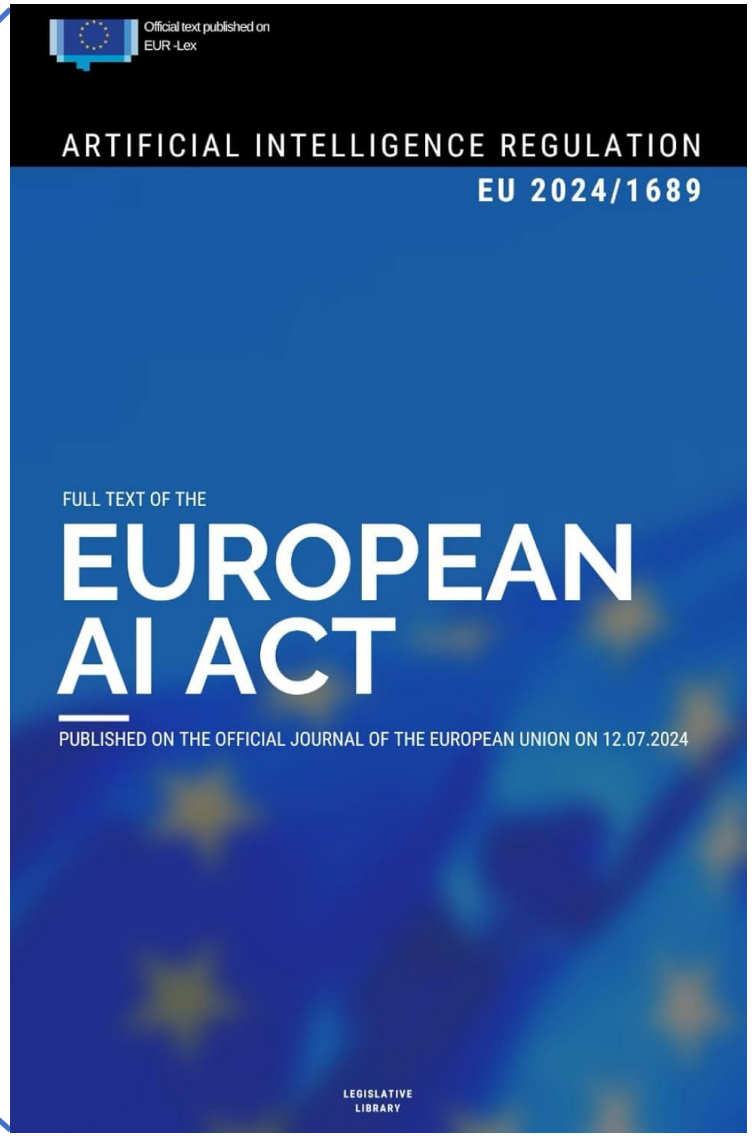
Key Insight

Sanctions inédites : jusqu'à 35 millions d'euros ou 7 % du chiffre d'affaires mondial.

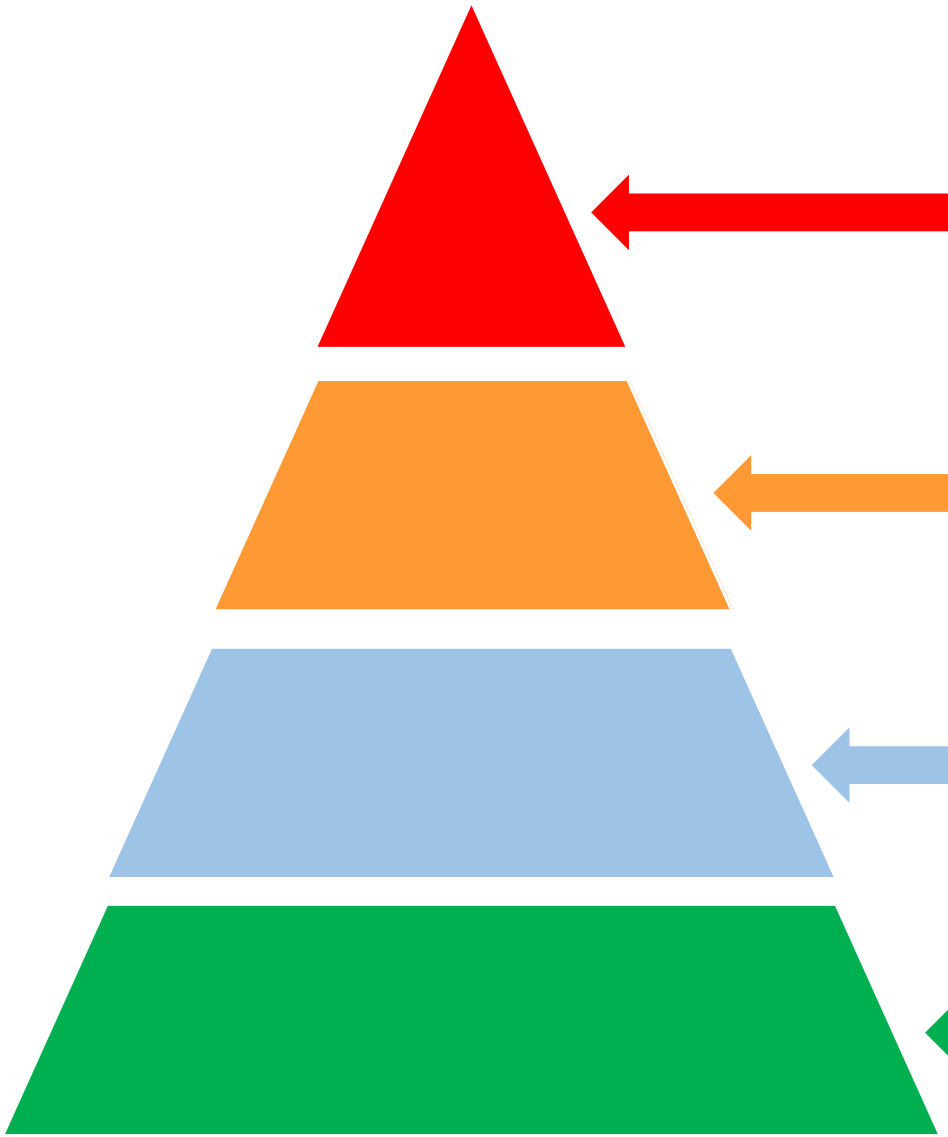
EU AI Act - Règlement EU 2024/1689

Texte règlementaire

- 144 pages
- 180 considérants
- 13 chapitres
- 113 articles
- 13 annexes



Classification des risques



Pratiques interdites en matière d'IA
(Chapitre II)

Système d'IA à Haut Risque
(Chapitre III)

SIA avec obligations de transparence
(Chapitre IV)

Autres systèmes d'IA

Conséquences de la classification

	Fournisseur	Déployeur	Mandataire	Importateur	Distributeur
Risque inacceptable	Interdiction	Interdiction	Interdiction	Interdiction	Interdiction
Haut risque	Obligations	Obligations	Obligations	Obligations	Obligations
Risque avec obligation de transparence	Obligations	Obligations	-	-	-
Risque minime	Adoption de codes de conduite (démarche volontaire)	Adoption de codes de conduite (démarche volontaire)	Adoption de codes de conduite (démarche volontaire)	Adoption de codes de conduite (démarche volontaire)	Adoption de codes de conduite (démarche volontaire)

Identifier les rôles - Exercice n°2

Modification d'un modèle et intégration dans une nouvelle application	
Scénario	La société française KALOPIA modifie un modèle open source de Mistral AI, l'intègre dans un assistant pédagogique (moteur de génération de contenus pour une plate-forme e-learning) et le commercialise en tant que service SaaS à des organismes de formation en France.
Fournisseur	KALOPIA
Déployeur	Les organismes de formation clients du service
Importateur	Aucun
Distributeur	Aucun

Modification d'un modèle et intégration dans une nouvelle application

Le cas de l'assistant pédagogique KALOPIA

La société KALOPIA transforme un modèle open source de Mistral AI en un outil de génération de contenus e-learning pour le vendre en mode SaaS.

Mistral AI - le fournisseur amont

Fournisseur de modèle d'IA à usage général (GPAI)

Mistral AI développe et met à disposition le modèle de base sous licence open source, sans maitriser le cas d'usage final ni le système commercialisé par KALOPIA.

KALOPIA le fournisseur de système

Fournisseur du système d'IA final

KALOPIA modifie le modèle, lui assigne une finalité déterminée et le met sur le marché sous son propre nom.

Fournisseur en aval (downstream provider)

Cette qualification s'ajoute à celle de fournisseur car KALOPIA s'appuie sur une IA à usage général existante pour construire son propre système.

Les organismes de formation - les utilisateurs

Déployeurs du système d'IA

En achatant la solution SaaS pour l'utiliser sous leur autorité dans leur activité pédagogique, iis deviennent des déployeurs seion l'AI Act.

Analyse des rôles absents

Pourquoi il n'y a ni importateur ni distributeur

Tous les acteurs sont établis dans l'Union européenne et il n'y a aucune revente intermédiaire du système entre KALOPIA et les clients finaux.

Action réalisée	Qualification résultante
Modification du modèle de base	**Fournisseur en aval**
Assignation d'une finalité précise	**Fournisseur du système**
Mise sur le marché sous sa marque	**Fournisseur du système**
Usage pro sous sa propre autorité	**Déployeur**

RGPD & Intelligence Artificielle

RGPD

(Règlement Général sur la Protection des Données)

ou

GDPR

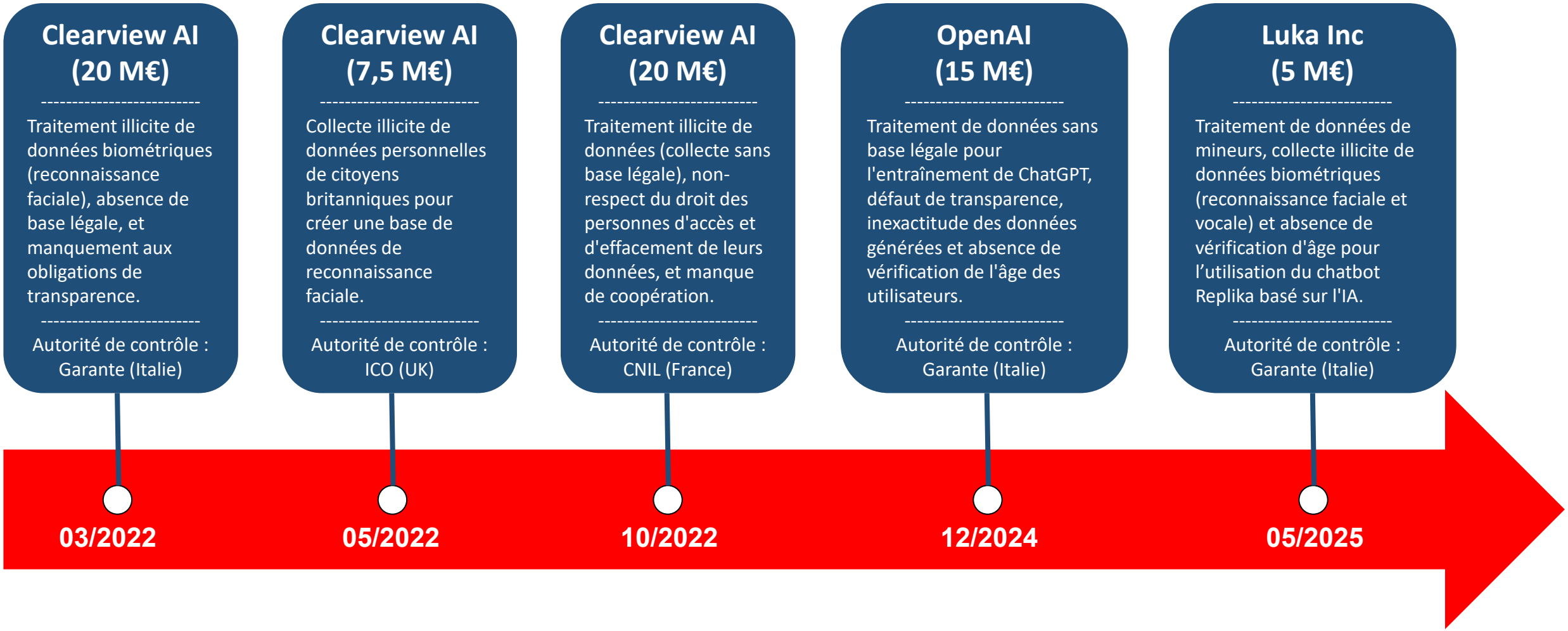
(General Data Protection Regulation)

RÈGLEMENT (UE) 2016/679 DU PARLEMENT EUROPÉEN ET DU CONSEIL

du 27 avril 2016

relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données

Les premières sanctions RGPD liées à l'IA



Conformité RGPD : 11 étapes + focus sur l'AIPD

Source CNIL

11 Respecter les principes du RGPD lors de l'annotation

10 Analyser le statut du modèle

9 Sécuriser le système d'IA

8 Assurer l'exercice des droits

7 Informer les personnes

6 Définir une durée de conservation

5 Minimiser les données

4 Vérifier la réutilisation des données

3 bis Adapter les garanties au moissonage

3 Définir la base légale

2 Déterminer les responsabilités (RT / ST)

1 Définir l'objectif (finalité) du système d'IA