SÉCURITÉ DE L'INTELLIGENCE ARTIFICIELLE

Nouvelles surfaces d'attaques / Sécurisation des systèmes d'IA

SECUREAI

WEBINAIRE

Mardi 5 mai 2026 - 12h

Programme du Webinaire

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

01

C'est quoi la sécurité de l'IA ?

AI Cybersecurity vs AI Safety vs AI Security vs Trustworthy AI : quelles différences ?

02

Cybersécurité classique

Pourquoi elle ne suffit plus face aux LLM et aux agents IA ?

03

Top 10 des risques de l'IA

Au-delà des TOP 10 de l'OWASP, quels sont les risques réels dans le contexte entreprise ?

04

Redéfinir le Zero Trust pour l'IA

Identité des agents, sécurité RAG, vuln. MCP, AI Gateway, validation runtime, ...

05

Conformité AI Act

Une démarche structurée en 9 étapes

06

CAISO et CAIRO

L'émergence de nouveaux métiers de la sécurité IA

07 **Compétences sécurité IA : *se positionner sur le marché avec les certifications SIAE et AAISM***

SECUREAI
www.SecureAI.fr

Le triple rôle de l'IA

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

Cible des attaques
(AI Security)

Sécurité de l'IA
(AI Security)

Cybersécurité de l'IA
(AI Cybersecurity)

Sûreté de l'IA
(AI Safety)

IA de confiance
(Trustworthy AI)

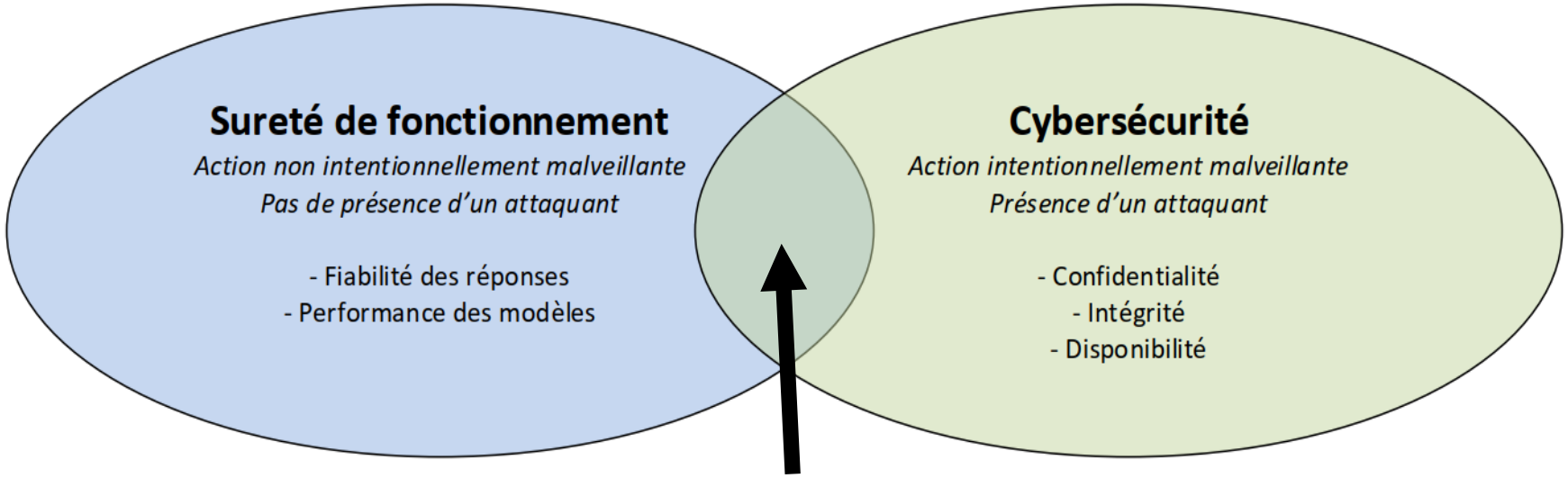
?

Arme pour attaquer
(Offensive AI)

Outil pour se protéger
(Defensive AI)

SECUREAI
www.SecureAI.fr

Cybersécurité des SIA : définition de l'ANSSI



Attaques dans le cas où un défaut de sureté de fonctionnement pourrait avoir un impact sur la cybersécurité du SI

NOTE

Dans sa définition de la cybersécurité de l'IA, l'agence prend uniquement en compte les défauts de sureté de fonctionnement qui peuvent avoir un impact sur la sécurité du SI (en termes CID).

Exemple : Un SIA permet de catégoriser, d'un point de vue métier, la sensibilité des documents bureautiques d'une entité. En cas de défaillance du modèle, des règles de protection inadéquates pourraient être appliquées sur certains documents dans l'espace documentaire, et il y aurait alors une potentielle atteinte en confidentialité sur le système d'information de l'entité.

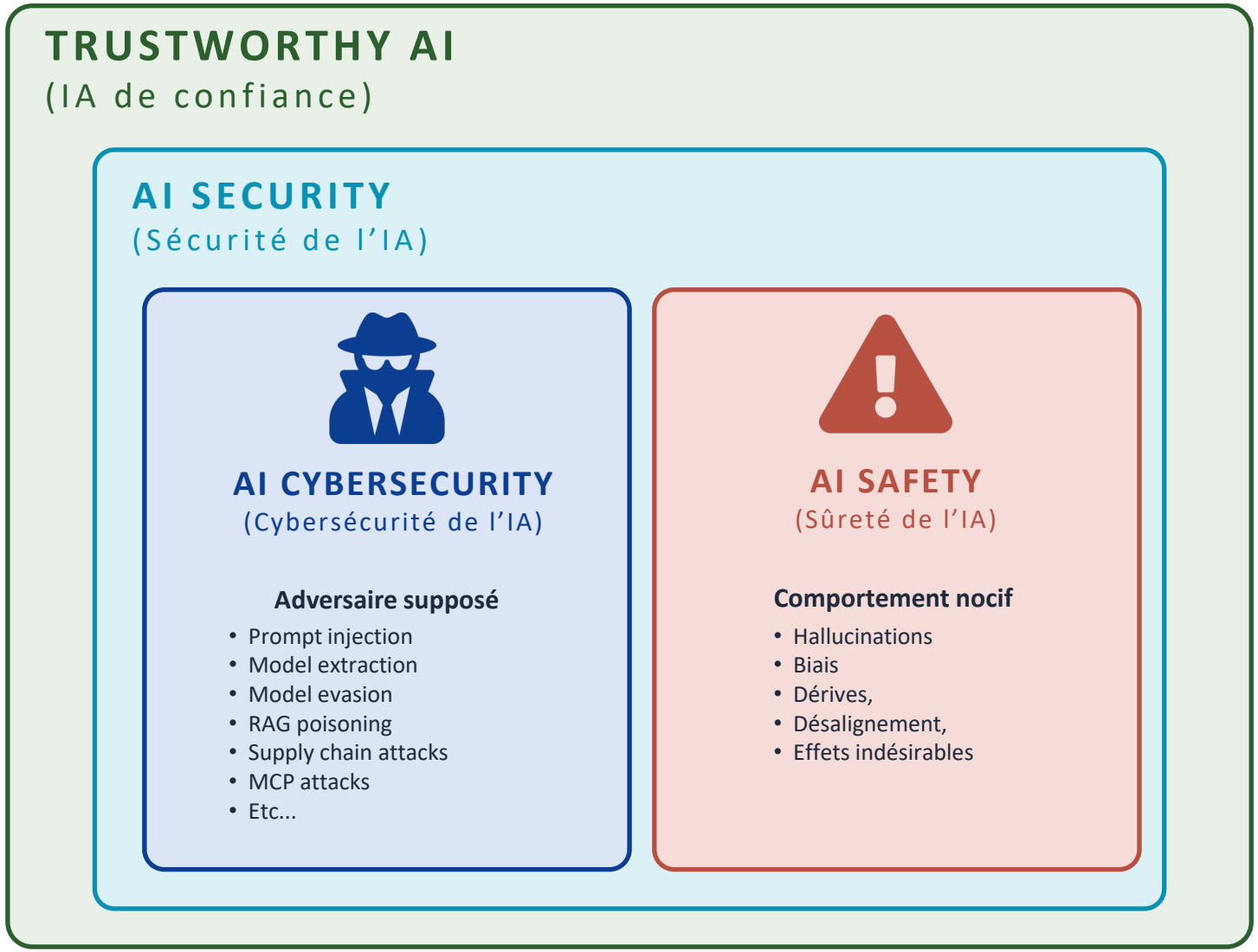
QUID de la sécurité de l'IA ?

Source : ANSSI - Cybersécurité des systèmes d'intelligence artificielle : définition de l'ANSSI (06/2025)

SECUREAI
www.SecureAI.fr

Cybersecurity / Safety / Security / Trustworthy

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM



SECUREAI
www.SecureAI.fr

Cybersecurity / Safety / Security / Trustworthy

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

Critère	AI Cybersecurity	AI Safety	AI Security	Trustworthy AI
Question centrale	Empêcher qu'un attaquant compromette le système d'IA ou s'en serve pour compromettre autre chose	Empêcher que le système d'IA, même non attaqué, produise des effets nocifs ou non désirables.	Cyber + safety à effet sécurité (intégrité comportementale)	Garantir une confiance globale dans le système
Adversaire requis ?	Oui (par définition)	Non requis	Optionnel	Non requis
Propriétés visées	Confidentialité, Intégrité, Disponibilité (CIA) + robustesse adversariale	Alignement, fiabilité, absence de dommages, supervision humaine	CIA + intégrité comportementale + résilience runtime	7 caractéristiques NIST AI RMF (équité, explicabilité, transparence, privacy, robustesse, validité, accountability)
Modèle de menace	Adversaire intentionnel modélisé (MITRE ATLAS, OWASP)	Défaillance, dérive, comportement émergent, désalignement	Adversaire + défaillances comportementales avec des impacts en termes de sécurité	Approche par propriétés systémiques, pas par menace
Acteur responsable	RSSI / CISO	AI Safety team (ML / alignment researchers)	CAISO (Chief AI Security Officer)	AI Governance Officer + COMEX + DPO
Menaces / risques (exemples)	Prompt injection, model extraction, RAG poisoning, MCP RCE, supply chain attack,...	Hallucinations, biais discriminants, contenus toxiques, désalignement, CBRN	Tool hijacking, memory poisoning, lethal trifecta, dérive d'agent avec impact sur le SI	Opacité, défaut d'équité, non-conformité RGPD/AI Act, atteinte aux droits fondamentaux
Métriques typiques	Defense Success Rate, CVE/CVSS, taux de blocage des prompt injection	Hallucination rate, bias score, refusal rate, alignment evaluations	DSR + détection runtime, taux de task drift, anomalies de trajectoire	Conformité AI Act - ISO/IEC 42001, FRIA / AIIA réalisées, audit AI Act

SYNTHÈSE

- Cybersecurity : protège le SIA contre des attaquants
- Safety : protège les utilisateurs et les tiers d'un dysfonctionnement du SIA
- Security : couvre les deux dans une perspective de défense
- Trustworthiness : englobe le tout + équité, explicabilité, gouvernance

SECUREAI
www.SecureAI.fr

Définition de l'ANSSI

- La cybersécurité d'un système d'IA (SIA) peut être définie de manière analogue à celle d'un système classique, comme l'étude des vulnérabilités du système et des mesures visant à y remédier afin de garantir, face à un attaquant, les propriétés de sécurité : confidentialité, intégrité, disponibilité.
- Les vulnérabilités, leurs impacts et leur remédiation peuvent se situer à trois niveaux : les composants individuels du système, son architecture et son intégration au sein d'un SI.
- La spécificité d'un SIA tient à ce qu'au moins l'un de ses composants implémente un modèle d'IA obtenu par apprentissage statistique, à partir de données d'entraînement.
- La cybersécurité d'un SIA ne prend pas en compte les risques liés aux aspects métiers, comme la fiabilité du résultat ou la performance du modèle (résultats erronés, hallucinations, etc.). Ces risques relèvent généralement de la sureté de fonctionnement (« *safety* »).

Pour l'ANSSI, la cybersécurité d'un SIA
- **est identique à celle d'un système numérique classique**
- **ne doit pas prendre en compte la fiabilité des résultats** (sauf si impact CID)
- **et sa spécificité tient seulement à l'apprentissage statistique**

SECUREAI
www.SecureAI.fr

Pourquoi l'IA pose un problème de sécurité ?

CE N'EST PAS *le fait d'être « obtenu par apprentissage statistique sur des données d'entraînement »*

→ une régression linéaire des années 70 l'est aussi.

Ce n'est pas ça qui crée le risque !

Les 6 propriétés intrinsèques des modèles d'IA modernes qui posent problème

01 Couplage code / données / comportement

Le comportement n'est pas écrit dans du code, il est inscrit dans les poids dérivés des données. Compromettre le dataset = compromettre durablement le modèle.

02 Opacité décisionnelle (black-box)

Aucune méthode d'interprétabilité ne fournit aujourd'hui une explication forensique du chemin causal entrée → sortie.
Difficulté d'audit et de debugging
Pb de conformité (RGPD art. 22)

03 Non-déterminisme en génération

À température > 0, la même entrée produit des sorties différentes. Une mitigation efficace 99 fois sur 100 peut échouer la 100e. Le paradigme du test reproductible devient invalide.

04 Sensibilité aux entrées adversariales

Perturbations imperceptibles, caractères Unicode invisibles, prompts cachés dans des images ou des PDF. Les filtres syntaxiques sont contournés au niveau sémantique.

05 Comportements émergents non prévus

Capacités apparues par effet d'échelle (in-context learning, tool use, planification). Un agent peut être sûr étape par étape mais dériver sur la séquence complète.

06 Pas de séparation native entre instruction / donnée (LLM)

Pour un LLM, prompt système, contenu utilisateur et données récupérées partagent le même contexte. Cause-racine des attaques de type prompt injection (pas de correctif structurel).

SECUREAI
www.SecureAI.fr

Changement de paradigme : 3 ruptures !

La cybersécurité classique protège du code déterministe dans un périmètre cartographiable.

La sécurité de l'IA générative et agentique ajoute une couche sémantique, probabiliste et autonome.

01

Rupture sémantique

Le LLM ne distingue pas, par construction, instruction et donnée. Tout contenu ingéré (e-mail, PDF, page web, sortie d'outil) peut devenir une commande.

Prompt injection (≠ SQL injection)

n°1 OWASP LLM 2025 - sans correctif structurel

02

Rupture probabiliste

Une même entrée peut produire des sorties différentes. Avec l'IA agentique, une hallucination ne se traduit pas par la génération d'un mauvais contenu mais par un action inattendue pouvant impacter la sécurité du SI.

Validation continue

Red-teaming probabiliste, mesure de dérive

03

Rupture d'autonomie

Les agents exécutent des actions (API, BDD, courriels, code) avec des permissions élevées. Le rayon d'action devient celui d'un opérateur privilégié (confused deputy) pouvant agir 24h/24, 7J/7.

Cas Anthropic (nov. 2025)

Campagne d'espionnage pilotée à 80–90 % par IA, ~30 cibles

Les mesures de sécurité cyber classiques restent nécessaires mais clairement insuffisantes !

SECUREAI
www.SecureAI.fr

Sécurité SI classique vs Sécurité de l'IA

- Définir la sécurité de l'IA
- **Cyber IT vs Cyber IA**
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

	SI classique	Système d'IA (GenAI / Agentique)
Actifs à protéger	Code, configs, identités, bases	+ Poids du modèle, données d'entraînement, prompts système, embeddings, bases vectorielles RAG, mémoires, outils MCP,...
Surface d'attaque	Réseau, API, dépendances, CI/CD	+ Prompts, documents RAG, sorties d'outils, serveurs MCP, mémoire long-terme
Vulnérabilités	CVE, mauvaises configs, authN/Z	+ Prompt injection, data poisoning, model extraction, RAG poisoning, excessive agency,...
Supply chain	Packages, conteneurs, SBOM	+ Modèles tiers (Hugging Face), datasets, serveurs MCP, AIBOM
Cycle de vie	DevSecOps	MLOps / LLMOps / AgentOps (la phase d'entraînement devient surface de compromission)
Observabilité	Logs déterministes	Traces sémantiques : prompts, contexte, décisions d'agent, trajectoires
Impact métier	Indisponibilité, fuite de données, fraude	+ Décisions biaisées à grande échelle, automatisation d'actions nocives, sanctions AI Act jusqu'à 7 % du CA mondial

EXEMPLES D'INCIDENTS DOCUMENTÉS EN 2025

- **EchoLeak / CVE-2025-32711 : zero-click M365 Copilot (CVSS 9.3)**
- **Amazon Q VS Code : release compromise via token GitHub**

SECUREAI
www.SecureAI.fr

Quels changements dans la cybersécurité ?

ARCHITECTURE

Zero Trust étendu

Séparer strictement instruction layer / retrieval layer / action layer. Toute donnée ingérée traitée comme non authentifiée. Les sorties du modèle ne sont jamais dignes de confiance par défaut.

GOUVERNANCE D'ACTION

Human-in-the-loop (HITL) obligatoire

Validation humaine pour les actions à effet de bord : transferts, suppressions, modifications d'ACL. Recommandation de l'ANSSI : proscrire l'usage automatisé d'IA pour les actions critiques sur le SI.

PIPELINE DE DÉFENSE

Guardrails runtime + sandboxing

Filtres entrée/sortie, allowlists d'outils, isolation des agents (gVisor, Firecracker), secrets éphémères, observabilité GenAI (OpenTelemetry), AI-SOC dédié.

ASSURANCE

Red-teaming probabiliste

Les tests one-shot ne suffisent plus. Mesurer un Defense Success Rate sur un panel d'attaques actualisé : prompt injection directe et indirecte, jailbreak, tool poisoning, memory poisoning,...

ORGANISATION

Nouveaux rôles, nouvelles ownerships

Émergence du CAISO (Chief AI Security Officer) et de l'AI Risk Officer. Articulation explicite RSSI / DPO / CAIO. Existence d'un owner clair du risque IA.

CONFORMITÉ

Empilement réglementaire

RGPD + AI Act (haut risque applicable au 2 août 2026) + NIS2 + DORA, à articuler avec ISO/IEC 42001 et NIST AI RMF. Sanctions AI Act jusqu'à 7 % du CA mondial.

Sécurité applicative classique : empêcher qu'un logiciel fasse ce qu'il ne doit pas faire.

Sécurité d'un système d'IA : empêcher qu'un système (probabiliste, nourri de données et de contexte) exécute autre chose que ce qu'on voulait lui faire faire.

SECUREAI
www.SecureAI.fr

TOP 10 des risques de l’IA en entreprise

- Définir la sécurité de l’IA
- Cyber IT vs Cyber IA
- **Top 10 des risques IA Entreprise**
- Redéfinir le Zero Trust pour l’IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

#	Risque	Description
1	**Fuite de données sensibles via les outils IA**	Les collaborateurs soumettent à des LLM publics du code propriétaire, des données clients ou des informations confidentielles, qui peuvent être réutilisées pour l'entraînement ou stockées hors UE. Les conséquences vont de la violation du RGPD à la perte d'avantage concurrentiel.
2	**Non-conformité réglementaire (AI Act / RGPD)**	L'AI Act, le RGPD et les textes sectoriels (DORA, NIS2) imposent des obligations croissantes, avec des sanctions pouvant atteindre 7% du chiffre d'affaires mondial. Au-delà des amendes, la non-conformité expose à des contentieux et à une atteinte durable à la confiance.
3	**Absence de gouvernance IA**	Sans cadre formalisé (politique, comité de pilotage, cartographie des usages), les décisions d'adoption se prennent en silos et les responsabilités se diluent. Ce vide organisationnel amplifie mécaniquement tous les autres risques : sécurité de l’information, conformité, éthique, réputationnel et financier.
4	**Shadow AI (usages non maîtrisés de l’IA)**	Le Shadow AI désigne l''utilisation d'outils IA hors de tout cadre validé par les collaborateurs, échappant ainsi totalement au contrôle ou à la visibilité de la DSI. Cette pratique occulte génère des failles de sécurité invisibles, multiplie les silos de données non sécurisés et empêche toute évaluation des risques cyber ou de la conformité réglementaire.
5	**Attaques sur les systèmes d’IA et augmentation de la surface d’attaque**	L'IA introduit de nouvelles menaces : injection de prompts, empoisonnement des données, extraction de modèles, attaques sur les agents et sur le protocole MCP, etc… Les défenses traditionnelles, conçues pour des applications déterministes, s'avèrent souvent inadaptées.
6	**Décisions erronées (hallucinations, biais, non-déterminisme)**	Les modèles produisent des sorties probabilistes parfois fausses mais plausibles et héritent de biais. Dans une architecture multi-agentique, l'erreur d'un agent devient l'entrée du suivant : les hallucinations se propagent, s'amplifient et déclenchent des actions inattendues sans qu'aucun humain ne puisse intervenir à temps.
7	**Dépendance aux fournisseurs IA et perte de souveraineté**	Une dépendance excessive envers quelques fournisseurs technologiques dominants (principalement américains) réduit l'autonomie stratégique de l'entreprise. En cas de changement brusque de tarifs, de conditions générales ou de tensions géopolitiques, l'entreprise peut se retrouver piégée, incapable de migrer ses actifs critiques vers des solutions souveraines ou alternatives.
8	**Manque de compétences internes et perte de savoir-faire**	Une délégation excessive à l'IA entraîne une perte de savoir-faire et une dépendance technologique où l'humain ne comprend plus les processus décisionnels sous-jacents. Parallèlement, le déficit de talents capables de superviser ces systèmes complexes laisse l'entreprise vulnérable face à l'obsolescence rapide de ses méthodes.
9	**Responsabilité floue en cas d'incident**	L'opacité des modèles et la multiplicité des acteurs impliqués rendent l'attribution des responsabilités extrêmement complexe lors d'un dysfonctionnement majeur. Sans protocoles clairs identifiant le responsable légal (fournisseur, intégrateur, déployeur ou utilisateur final), l'entreprise s'expose à des contentieux longs et coûteux.
10	**Coûts non maîtrisés et ROI incertain**	Aux licences et API (tarif aux jetons) s'ajoutent des coûts cachés (GPU, stockage, supervision, conformité), avec des tarifications à l'usage propices aux dérives budgétaires. Sans évaluation précise de la valeur ajoutée réelle, ces investissements peuvent transformer l'innovation en gouffre financier.

SECUREAI
www.SecureAI.fr

Critères d’évaluation

Criticité

Probabilité d'occurrence multipliée par l'impact potentiel sur l'entreprise.

Contrôlabilité

L'entreprise peut-elle agir seule, ou dépend-elle de tiers ?

Effort de mitigation

Combien faut-il investir pour réduire significativement le risque ?

Temporalité

Le risque est-il déjà actif, émergent ou prospectif ?

Réversibilité

Une fois le risque matérialisé, peut-on revenir en arrière ?

SECUREAI
www.SecureAI.fr

Evaluation des risques

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- **Top 10 des risques IA Entreprise**
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

#	Risque	Catégorie	Criticité	Temporalité	Contrôlabilité	Réversibilité	Effort de mitigation
1	**Fuite de données sensibles via les outils IA**	Technique / Cybersécurité	**Critique**	**Immédiat**	**Mixte**	**Irréversible**	**Moyen**
2	**Non-conformité réglementaire (AI Act / RGPD)**	Juridique / Conformité	**Critique**	**Immédiat**	**Mixte**	**Irréversible**	**Élevé**
3	**Absence de gouvernance IA**	Gouvernance / Management	**Critique**	**Immédiat**	**Endogène**	**Réversible**	**Moyen**
4	**Shadow AI (usages non maîtrisés de l'IA)**	Gouvernance / Management	**Élevé**	**Immédiat**	**Endogène**	**Difficilement réversible**	**Faible**
5	**Attaques sur les systèmes d'IA et augmentation de la surface d'attaque**	Technique / Cybersécurité	**Élevé**	**Immédiat**	**Mixte**	**Difficilement réversible**	**Élevé**
6	**Décisions erronées (hallucinations, biais, non-déterminisme)**	Opérationnel / Qualité	**Élevé**	**Immédiat**	**Mixte**	**Difficilement réversible**	**Élevé**
7	**Dépendance aux fournisseurs IA et perte de souveraineté**	Stratégique / Compétitivité	**Élevé**	**Moyen terme**	**Exogène**	**Difficilement réversible**	**Élevé**
8	**Manque de compétences internes et perte de savoir-faire**	Gouvernance / Management	**Élevé**	**Moyen terme**	**Endogène**	**Difficilement réversible**	**Moyen**
9	**Responsabilité floue en cas d'incident**	Juridique / Conformité	**Modéré**	**Immédiat**	**Mixte**	**Difficilement réversible**	**Moyen**
10	**Coûts non maîtrisés et ROI incertain**	Stratégique / Compétitivité	**Modéré**	**Immédiat**	**Endogène**	**Réversible**	**Faible**

SECUREAI
www.SecureAI.fr

Les limites du Zero Trust classique avec l'IA

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- **Redéfinir le Zero Trust pour l'IA**
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

SECUREAI
www.SecureAI.fr

01

Le sujet n'est plus humain

Les agents IA sont des identités non-humaines (NHI) éphémères, qui s'enchaînent et invoquent d'autres modèles. Les modèles, les agents autonomes, les sous-agents et les serveurs d'outils doivent être traités comme des sujets de sécurité de première importance nécessitant une identité propre et gouvernable.

02

La décision devient probabiliste

Un même prompt produit des sorties différentes. Les pare-feu applicatifs traditionnels sont inefficaces face à cette dynamique sémantique.
Aucune règle déclarative ne filtre fiablement la sortie d'un modèle → PEP runtime sémantiques (Guardrails).

03

L'injection de prompt brise l'auth

Un contenu non fiable récupéré (web, RAG, email, document) devient une instruction exécutée avec les privilèges de l'agent. Prompt inject + Confused deputy = risque critique → AI Firewall (Palo Alto AI Runtime security, Cisco AI Defense, Lakera Guard,...) ou guardrails managés (AWS Bedrock Guardrails, Azure Prompt Shield, Google Model Armor).

04

Le moindre privilège vacille

Du "Least Privilege" au "Least Agency) : un LLM avec accès à des outils peut chaîner des capacités non explicitement accordées : lecture → écriture → exfiltration via résumé.

La publication SP 800-207 du NIST reste le document de référence pour le Zero Trust (2020)
Elle est complétée par la SP 1800-35 pour son implémentation pratique (2025).
Mais il n'existe à ce jour aucun document NIST dédié au Zero Trust pour l'IA.

Les 5 couches du Zero Trust pour l'IA

Identité	Identité unique par agent (Entra Agent ID, AWS Agentcore Identity, SPIFFE) - OAuth 2.1 + PKCE + Resource Indicators (MCP) - Jetons éphémères ≤ 1h
Données	ACL propagées au retrieval (RAG permission-aware) - Index vectoriel par tenant - Masquage des DCP (PII redaction) - Provenance signée des datasets
Modèle	Format safetensors obligatoire - Signatures Sigstore - Registre privé avec scan automatisé (ModelScan) - Confidential GPU computing pour modèles à risque
Application & Orchestration	AI Firewall en chokepoint - Filtres entrée/sortie - Pattern dual-LLM ou CaMeL - Capability tokens scopés - HITL sur opérations irréversibles
Infrastructure	Endpoints privés systématiques - Désactivation auth locale - mTLS service mesh - Egress filtering avec allowlist

PLAN DE CONTRÔLE
Contrôle d'accès ABAC ou PBAC

- Moteur de politique externe (OPA/Rego, AWS Cedar, Permit.io)
- Décision ré-évaluée à chaque action
- Attributs IA : autonomie agent, source du contexte, sensibilité agrégée, profondeur d'outils, présence HITL

3 composantes essentielles : ABAC externalisé *(pour la décision),* **AI Firewall** *(pour le contrôle des flux)* et l'**identité des agents** *(pour la traçabilité/imputabilité)*

SECUREAI
www.SecureAI.fr

Détail des règles Zero Trust pour l'IA (1/2)

Couche	Mesure de sécurité	Règle ZT-IA
Identité	**Identité unique par agent**	Chaque agent IA dispose d'une identité cryptographique unique, traçable et révocable (Entra Agent ID, AWS AgentCore Identity, SPIFFE), jamais une clé API en dur ni un compte partagé.
	OAuth 2.1 + PKCE + Resource Indicators	Tout serveur MCP est authentifié via OAuth 2.1 avec PKCE obligatoire et Resource Indicators (RFC 8707) pour lier chaque jeton à un destinataire unique.
	Jetons éphémères ≤ 1h	Tous les jetons d'authentification d'agent ont une durée de vie courte avec rotation automatique, pour limiter l'impact d'une compromission.
Données	**ACL propagées au retrieval** (RAG permission-aware)	La base vectorielle hérite des permissions du document source et filtre la récupération par identité utilisateur, pour éviter qu'un RAG ne casse le RBAC.
	Index vectoriel par tenant	Chaque tenant dispose de son propre index physiquement séparé, jamais d'un index partagé avec simple metadata-filtering.
	Masquage des DCP	Détection et anonymisation automatiques des données personnelles dans les prompts entrants et réponses sortantes, en amont des LLM tiers.
	Provenance signée des datasets	Chaque jeu de données d'entraînement ou de fine-tuning porte une attestation cryptographique signée (SLSA, in-toto) garantissant son origine et son intégrité.
Modèle	**Format safetensors obligatoire**	Interdiction des formats sérialisés exécutables (pickle, joblib non-safe) au profit de safetensors (hugging face), qui ne peut structurellement pas contenir de code.
	Signatures Sigstore	Chaque artefact modèle est signé via Sigstore avec une identité d'entreprise vérifiable, et la signature est vérifiée automatiquement avant tout chargement en production.
	Registre privé avec scan (ModelScan)	Tous les modèles transitent par un registre interne contrôlé, après scan de sécurité (ModelScan, Fickling) en zone de quarantaine.
	Confidential GPU computing	Pour les modèles à risque systémique ou traitant des données ultra-sensibles : inférence GPU TEE (NVIDIA H100/H200 CC) avec chiffrement de la VRAM et attestation cryptographique.

SECUREAI
www.SecureAI.fr

Détail des règles Zero Trust pour l'IA (1/2)

Couche	Mesure de sécurité	Règle ZT-IA
Application & Orchestration	AI Firewall en chokepoint	Toute requête vers un LLM transite par une passerelle unique qui centralise authentification, filtres, quotas, observabilité et politique de sécurité.
	Filtres entrée / sortie	Détection d'injection en entrée (Lakera, Prompt Shields) et filtres de toxicité, PII et grounding en sortie (Bedrock Guardrails, Llama Guard).
	Pattern dual-LLM ou CaMeL	Séparation stricte entre un LLM privilégié qui décide et un LLM en quarantaine qui consomme le contenu non fiable, avec sorties typées inertes.
	Capability tokens scopés	Chaque appel d'outil est autorisé par un jeton éphémère scopé sur le tuple utilisateur × mission × outil × paramètres × durée, jamais par un rôle large.
	HITL sur opérations irréversibles	Toute action irréversible ou à fort blast radius (email, paiement, suppression, modif. config) exige une validation humaine explicite avec MFA.
Infrastructure	Endpoints privés systématiques	Tous les services IA managés sont accessibles uniquement via Private Link / PrivateLink / PSC, jamais via leur endpoint public Internet.
	Désactivation auth locale	Interdiction des clés API statiques au profit de l'IdP corporate (Entra ID, IAM, Workload Identity) avec MFA et Conditional Access.
	mTLS service mesh	Toutes les communications inter-services sont chiffrées et mutuellement authentifiées via certificats SPIFFE émis par un service mesh (Istio, Linkerd, Cilium).
	Egress filtering avec allowlist	Default deny en sortie : aucun workload IA ne peut joindre Internet hors d'une liste blanche limitative de destinations métier nécessaires.
Plan de contrôle ABAC / PBAC	Moteur de politique externe	La décision d'autorisation est externalisée dans un service dédié (OPA/Rego, AWS Cedar, Permit.io), séparé des applications, avec politiques versionnées comme du code.
	Décision ré-évaluée à chaque action	Pas de token statique émis au login : chaque action sensible déclenche une nouvelle évaluation avec les attributs frais du moment.
	Attributs IA dynamiques	La politique consomme des attributs IA-natifs (autonomie agent, source du contexte, sensibilité agrégée, profondeur d'outils, présence HITL) en plus des attributs IT classiques.

SECUREAI
www.SecureAI.fr

Les limites du Zero Trust

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- **Redéfinir le Zero Trust pour l'IA**
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

Ce que le Zero Trust ne résout PAS !

Imperfection des mesures de sécurité
Mitigation probabiliste → pas de prévention déterministe
Nécessite des évalations continues (PyRIT, Garak, Promptfoo,...)

Fiabilité intrinsèque du modèle
Hallucinations, biais et derives ne sont pas du resort du ZT
Cela relève de la sûreté de l'IA (AI Safety)

Opacité comportementale
On signe les poids, mais un sleeper agent (Anthropic 2024) peut survivre au RLHF

Causalité faible
Qui est responsable ? L'utilisateur, l'auteur du document RAG, l'éditeur du modèle ?

Usage malveillant (Functionnal misuse)
Un agent qui exécute une consigne malveillante d'un utilisateur autorisé n'enfreint aucun contrôle ZT

3 priorités pour un CISO / CAISO

1 Rationaliser l'identité des agents
→ NHI de premier rang
(Entra Agent ID, AWS AgentCore Identity, SPIFFE,...)

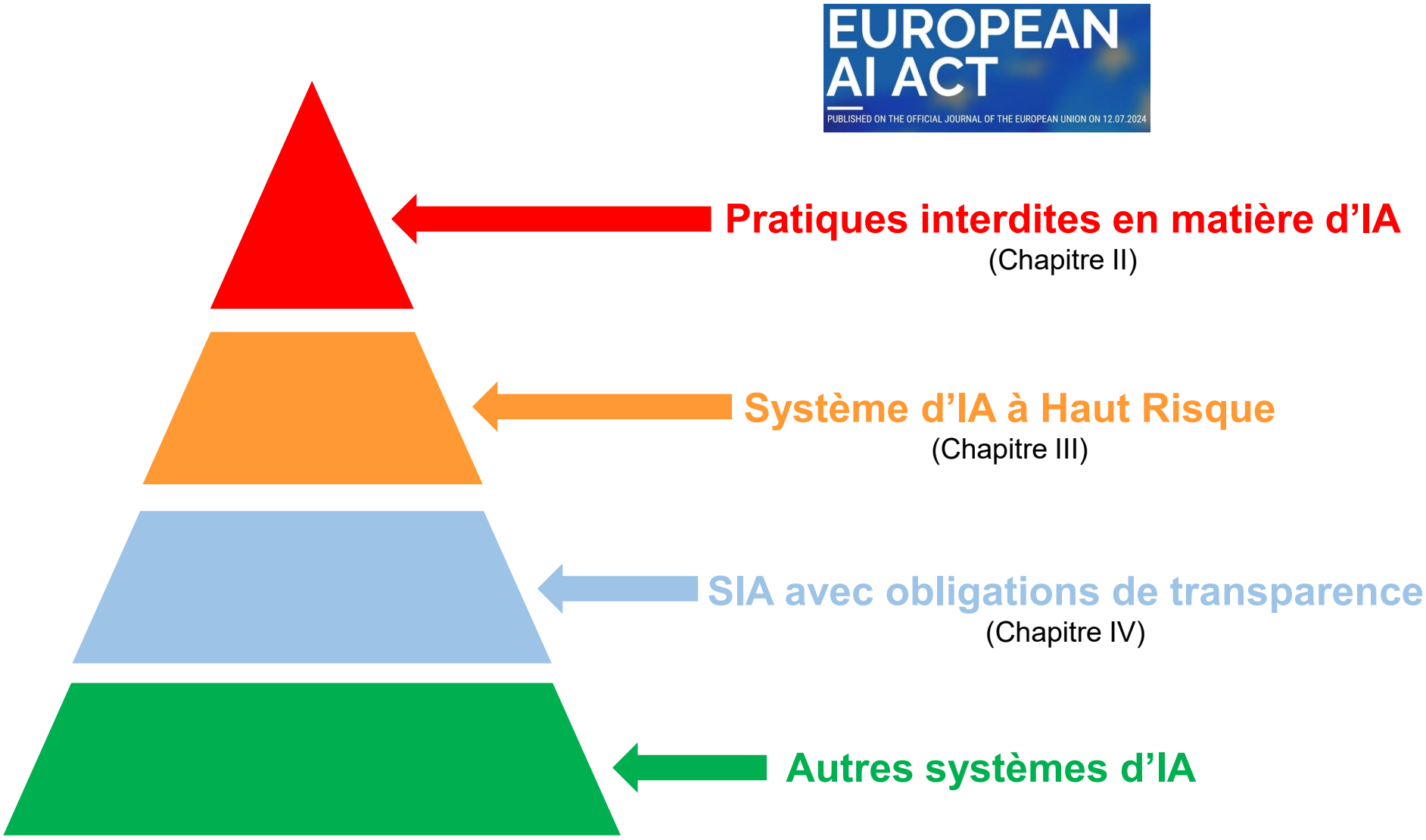
2 Déployer un AI FW (ou AI Gateway) + ABAC
- Chokepoint avec PEP entrée/sortie
- Autorisations : OPA/Rego, AWS Cedar,...
- Observabilité : Open Telemetry avec l'extension GenAI Semantic Conventions

3 Aligner la gouvernance
- ISO/IEC 42001 + 23894 + 27090
- AI Act art. 15 (SIA à haut-risque - 08/2026)
- GPAI code of practice (section : Safety and Security)

Le Zero Trust ne supprime pas les risques résiduels.
Il les rend gouvernables, traçables et juridiquement opposables.

SECUREAI
www.SecureAI.fr

Classification des risques dans l'AI act



SECUREAI
www.SecureAI.fr

Sanctions administratives

35 M€ ou 7% du CA (article 99 §3)
- ✓ **Non-respect de l'interdiction des pratiques en matière d'IA** (article 5)

15 M€ ou 3% du CA (article 99 §4)
- ✓ **Non-conformité concernant les systèmes d'IA à haut risque**

7,5 M€ ou 1% du CA (article 99 §5)
- ✓ **Fourniture d'informations inexactes, incomplètes ou trompeuses aux autorités compétentes**

NOTE

- **Pour les entreprises privées ou publiques : le chiffre le plus élevé est retenu**
- **Pour les PME ou les startups : le chiffre le plus faible est retenu**

SECUREAI
www.SecureAI.fr

Obligations selon classification du SIA & rôle

	Fournisseur	Déployeur	Mandataire	Importateur	Distributeur
Risque inacceptable	Interdiction	Interdiction	Interdiction	Interdiction	Interdiction
Haut risque	Obligations	Obligations	Obligations	Obligations	Obligations
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)

SECUREAI
www.SecureAI.fr

Les différents rôles dans le RIA (AI act.)

Mandataire

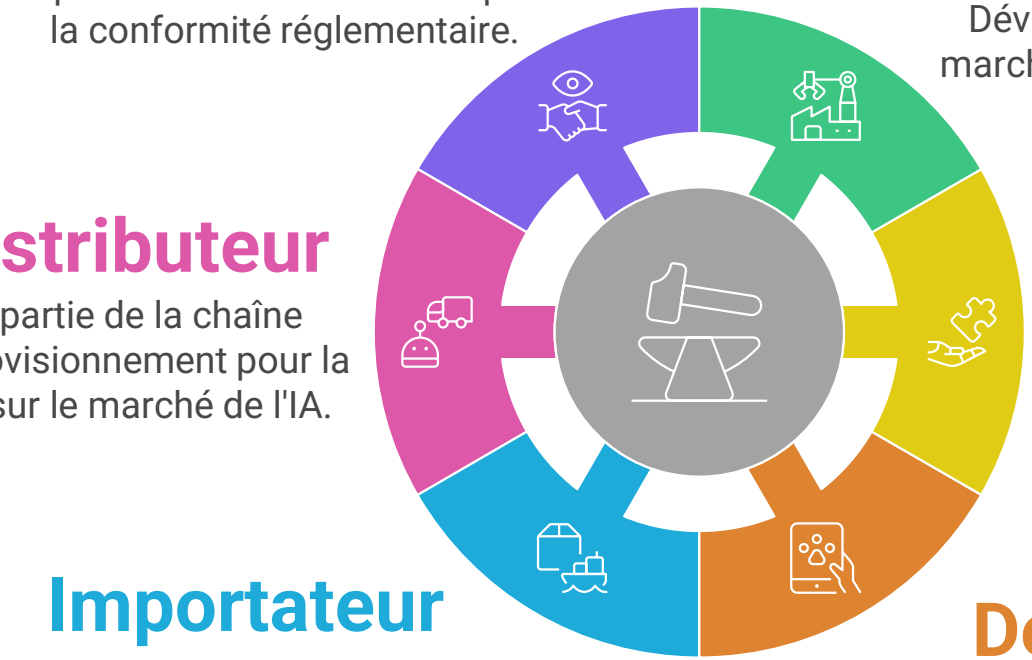
Représente les fournisseurs pour la conformité réglementaire.

Fournisseur

Développe et met sur le marché des systèmes d'IA.

Distributeur

Fait partie de la chaîne d'approvisionnement pour la mise sur le marché de l'IA.



Fournisseur en aval

Intègre des modèles d'IA dans des systèmes existants.

Importateur

Met sur le marché des systèmes d'IA en provenance de pays tiers.

Déployeur

Utilise des systèmes d'IA dans des activités professionnelles.

SECUREAI
www.SecureAI.fr

Identification des rôles (exemple)

LLM (open weights) utilisé dans un chatbot d'entreprise	
Scénario	L'entreprise DeltaCorp, établie au sein de l'UE, télécharge depuis la plateforme Hugging Face le modèle de langage à poids ouverts (gpt-oss-120b) développé par OpenAI. DeltaCorp intègre ce modèle dans un chatbot d'intelligence artificielle qu'elle déploie en interne afin d'automatiser son service client.
Fournisseur	?
Déployeur	?
Importateur	?
Distributeur	?

SECUREAI
www.SecureAI.fr

Identification des rôles (exemple)

LLM (open weights) utilisé dans un chatbot d'entreprise	
Scénario	L'entreprise DeltaCorp, établie au sein de l'UE, télécharge depuis la plateforme Hugging Face le modèle de langage à poids ouverts (gpt-oss-120b) développé par OpenAI. DeltaCorp intègre ce modèle dans un chatbot d'intelligence artificielle qu'elle déploie en interne afin d'automatiser son service client.
Fournisseur	DeltaCorp est fournisseur en aval (downstream provider)
Déployeur	DeltaCorp
Importateur	Aucun
Distributeur	Aucun

SECUREAI
www.SecureAI.fr

LLM (open weights) / chatbot d'entreprise

Le scénario d'étude

L'entreprise DeltaCorp (UE)
DeltaCorp télécharge un modèle LLM (gpt-oss-120b) depuis Hugging Face pour l'intégrer dans un chatbot de service client interne.

Le modèle gpt-oss-120b (OpenAI)
Il s'agit d'un modèle de langage à poids ouverts servant de base technologique au système final.

Les rôles de DeltaCorp

Fournisseur en aval (downstream provider)
DeltaCorp est fournisseur car elle assemble le système final, intègre le modèle et met le chatbot en service sous sa propre responsablilité.

Déployeur
DeltaCorp est aussi déployeur car elle utilise le système d'IA dans le cadre de son activité professionnelle pour automatiser son service client.

Un cas de cumul de rôles : Dans ce scénario, une seule entité (DeltaCorp) endosse simultanément les responsabilités de fournisseur et de déployeur.

Les rôles absents ou exclus

OpenAI n'est pas le fournisseur du chatbot
OpenAI est uniquement le fournisseur du modèle de langage initial, pas du système final (le chatbot) créé par DeltaCorp.

Aucun importateur ni distributeur
Comme le chatbot est développé et utilisé en interne sans être mis sur le marché par un tiers ou vendu, ces **rôles n'existent pas ici.**

Pourquoi le client n'est pas un déployeur

Une personne physique interagissant avec l'IA
Le règlement ne crée pas de rôle juridique spécifique pour l'utilisateur final qui se contente de discuter avec le chatbot.

Le test du contrôle
- ✗ Arrêter le système
- ✗ Changer les règles
- ✗ Décider de l'usage

L'autorité et l'exploitation :
Le déployeur est celui qui exploite l'IA sous sa propre autorité et en définit les finalités, ce qui est le rôle exclusif de DeltaCorp.

SECUREAI
www.SecureAI.fr

Assurer sa conformité en 9 étapes

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- **La conformité AI Act**
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

Identifier les systèmes d'IA	Faire l'inventaire de tous les systèmes d'IA développés, importés, commercialisés ou utilisés par l'organisation
Définir le type d'IA	Est-ce un système d'IA ou un modèle d'IA à usage général (GPAI) ?
Classifier le niveau du risque	Selon 4 niveaux pour les SIA et risque systémique ou non pour les modèles GPAI
Déterminer le rôle de l'organisation	Quel rôle joue l'organisation pour cette IA ? (Fournisseur, Déployeur, Mandataire, Importateur ou Distributeur)
Identifier les obligations	Identifier les obligations qui s'appliquent dans ce contexte selon le niveau de risque et le rôle de l'organisation
Mesurer les écarts de conformité	Identifier les éventuels écarts de conformité
Corriger les écarts	Mettre en œuvre les mesures nécessaires afin de respecter ses obligations
Notifier les autorités	Effectuer (si nécessaire) les déclarations auprès des autorités compétentes
Réévaluer périodiquement	S'assurer que le rôle de l'organisation ou le niveau de risque du SIA ne changent pas dans le temps

SECUREAI
www.SecureAI.fr

Timeline de mise en application

1er août 2024

- Entrée en vigueur du règlement
- Aucune obligation immédiatement applicable

2 février 2025

- Interdiction des IA à risque inacceptable (art. 5)
- Obligation de littératie IA (art. 4)

2 août 2025

- Obligations pour les modèles d'IA à usage général (GPAI)
- Mise en place de la gouvernance (AI Office, autorités nationales)
- Application du régime de sanctions (art. 99)

2 août 2026

- Obligations pour les IA à haut risque (Annexe III)
- Obligation de transparence des contenus générés (art. 50)
- Application générale du règlement

2 août 2027

- Obligations pour les IA à haut risque composants de produits réglementés (Annexe I, art. 6(1))
- Mise en conformité des modèles GPAI antérieurs (art. 111)

SECUREAI
www.SecureAI.fr

CAISO / Responsable Sécurité de l'IA (RSIA)

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM

SECUREAI
www.SecureAI.fr

Qu'est-ce qu'un CAISO ?

(Parfois appelé AI Security Lead ou Head of AI Safety & Security)

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

Ce que le CAISO apporte

- Évalue les menaces spécifiques aux SIA : *adversarial, data/model poisoning, exfiltration, biais.*
- Déploie les contrôles techniques et organisationnels sur tout le cycle : Data → Modèle → API → Production.
- Aligne sécurité IA, conformité et objectifs métier.

Différences avec les autres rôles

- CAIO stratégie & gouvernance globale de l'IA → CAISO focus sécurité.
- CTO architecture & innovation → CAISO exigences sécurité IA (cloud, MLOps, libs ML).
- CISO/RSSI sécurité SI au sens large → CAISO vulnérabilités et contrôles propres aux modèles et systèmes d'IA.

Positionnement organisationnel

- Phase initiale : rattachement pragmatique au CISO.
- À maturité ou criticité élevée : rattachement DG ou COMEX possible.
- Principe directeur : plus l'IA est stratégique pour l'entreprise, 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.

Les missions du CAISO (1/2)

SECUREAI
www.SecureAI.fr

Leadership stratégique & gouvernance

Feuille de route stratégique

Élaborer et piloter la roadmap pour développer outils, compétences et processus de sécurité IA.

Politiques & standards

Établir des règles claires pour la validation des modèles, l'usage des données et l'intégration de logiciels tiers.

Alignement métier

S'assurer que la sécurité accompagne l'innovation sans la freiner ; conseiller la direction sur risques et opportunités.

Représentation externe

Incarner l'expertise de l'entreprise dans conférences et groupes de travail sectoriels.

Promotion interne

Diffuser une culture de l'IA sécurisée et éthique à tous les niveaux, dès la conception des projets.

Supervision opérationnelle & réponse aux incidents

Centre opérationnel (AISOC)

Mettre en place et superviser un SOC dédié à la détection d'anomalies et au monitoring des modèles en production.

Réponse aux incidents

Plans de réponse spécifiques IA (contenir un modèle corrompu) et coordination data scientists / ML / cyber.

Gestion des ressources

Définir budgets, recruter les talents spécialisés et sélectionner les outils technologiques adaptés.

Indicateurs de performance

Suivre des KPIs (attaques bloquées, MTTR) pour piloter l'amélioration continue.

Évaluation de la maturité

Audits réguliers pour identifier les lacunes et prioriser les actions à mener.

Les missions du CAISO (2/2)

SECUREAI
www.SecureAI.fr

Expertise technique & gestion des risques IA

Analyse des risques IA
Évaluations de menaces (vol de modèle, empoisonnement). Méthodologies avancées : tests d'attaques adversariales.

Architecture & contrôles
Concevoir des systèmes secure-by-design (séparation environnements, chiffrement). Superviser watermarking, détection d'anomalies.

Développement sécurisé (MLSecOps)
Intégrer la sécurité dans le pipeline (qualité données, robustesse). Exercices Red Team / Blue Team pour l'IA.

Veille & outillage
Rester à la pointe des menaces et des solutions. Sélectionner et déployer les technologies de protection pertinentes.

Gouvernance, Risque, Conformité & collaboration

Cadre GRC pour l'IA
Intégrer les risques IA dans la gouvernance globale. Piloter la conformité (AI Act, RGPD). Documenter les contrôles pour les audits.

Collaboration transverse
CISO (cyber globale), CTO/MLOps (infrastructures, pipelines), CAIO/métiers (sécurité au service de l'innovation).

Sensibilisation & formation
Diffuser une culture sécurité IA auprès des data scientists, développeurs et équipes métier. Former aux nouveaux risques.

Compétences requises pour devenir CAISO

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM

SECUREAI
www.SecureAI.fr

01
Maîtrise des technologies d'IA

- Compréhension approfondie des algorithmes (ML, DL, NLP)
- Connaissance du cycle de vie des modèles (donnée → déploiement)
- Capacité à dialoguer d'égal à égal avec data scientists et experts IA

02
Expertise cybersécurité

- Maîtrise des fondamentaux : IAM, chiffrement, sécurité réseau
- Connaissance des normes et frameworks (ISO 27001, NIST CSF, MITRE ATT&CK)
- Pratique du SOC, de la gestion d'incidents et du Zero Trust

03
Vulnérabilités spécifiques à l'IA

- Attaques adversariales, prompt injection, jailbreaks
- Empoisonnement de données et de modèles, model extraction
- Differential privacy, watermarking, MLSecOps

04
Culture réglementaire & éthique

- Connaissance des régulations clés (AI Act, RGPD, NIS2, DORA)
- Maîtrise des standards (ISO 42001, NIST AI RMF)
- Forte sensibilité aux principes éthiques pour une IA de confiance

05
Leadership & communication

- Capacité à vulgariser les risques techniques pour la direction
- Aptitude à piloter le changement et fédérer des expertises variées
- Posture d'autorité interne reconnue (techniques, juridiques, métier)

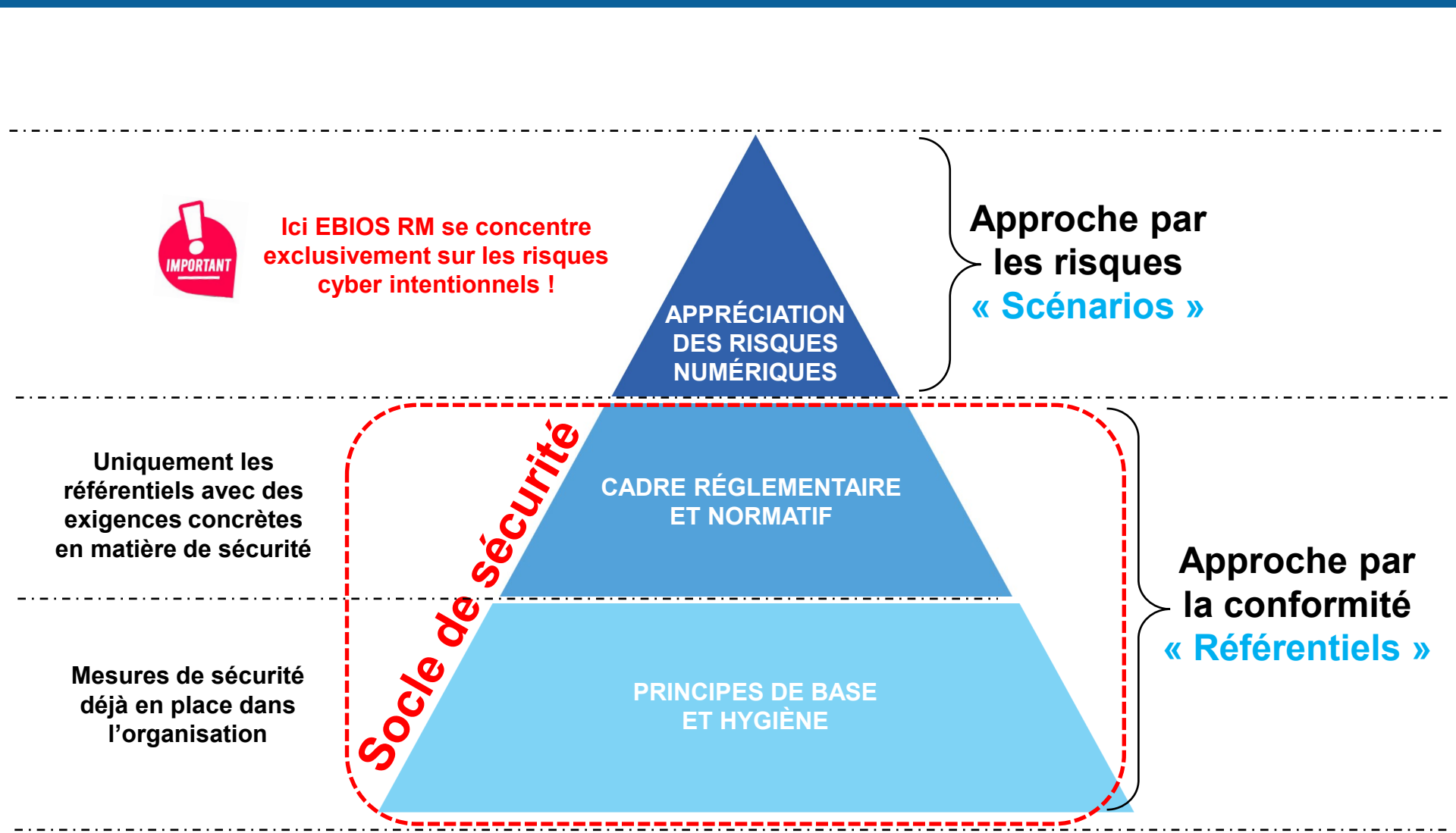
SIAE

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

→ La certification SIAE arrivera sur le marché en France mi-2026

EBIOS RM adapté à l'IA ?

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM



SECUREAI
www.SecureAI.fr

Quel socle de sécurité pour l'IA ?

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM

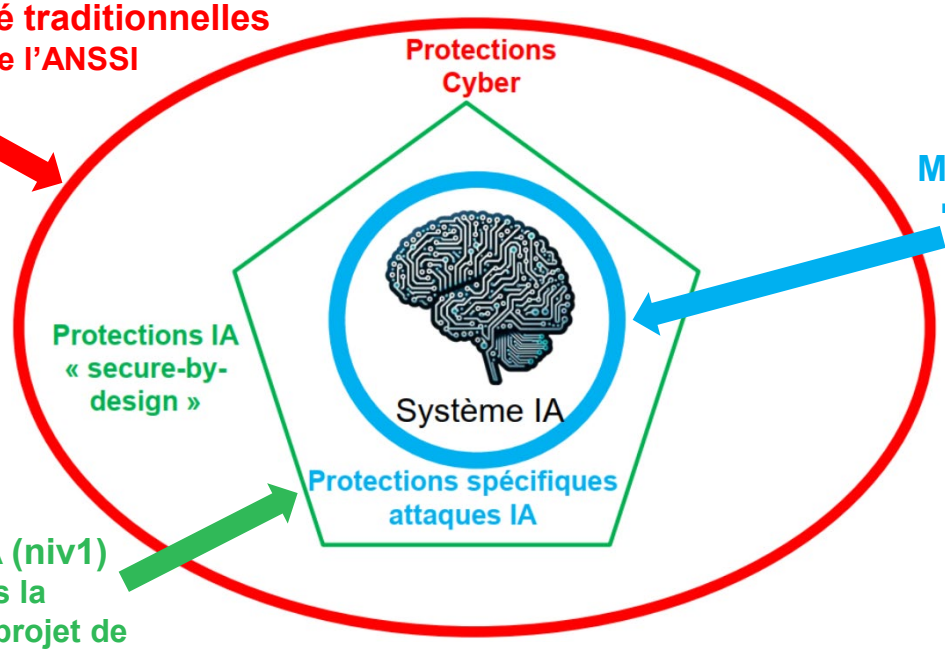
	Référentiel	Source	
Standard	Guide d'hygiène	ANSSI	FR
	ISO/IEC 27001 & ISO/IEC 27002	ISO	INT
	CIS Controls	CIS	US
	SP 800-53	NIST	US
	Politique de sécurité du SI (PSSI) de l'organisation	-	-
Spécifique IA	Recommandations de sécurité pour un système d'IA générative	ANSSI	FR
	Securing Artificial Intelligence (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 / TOP 10 LLM / TOP Agentic AI / TOP 10 MCP	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
	Politique de sécurité IA (PSIA) de l'organisation	-	-
Règlementaire	Règlement AI Act	UE	EU
	Fiches pratiques IA	CNIL	FR

NOTE Liste non exhaustive donnée à titre indicatif

SECUREAI
www.SecureAI.fr

3 niveaux de mesures pour traiter les risques

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...

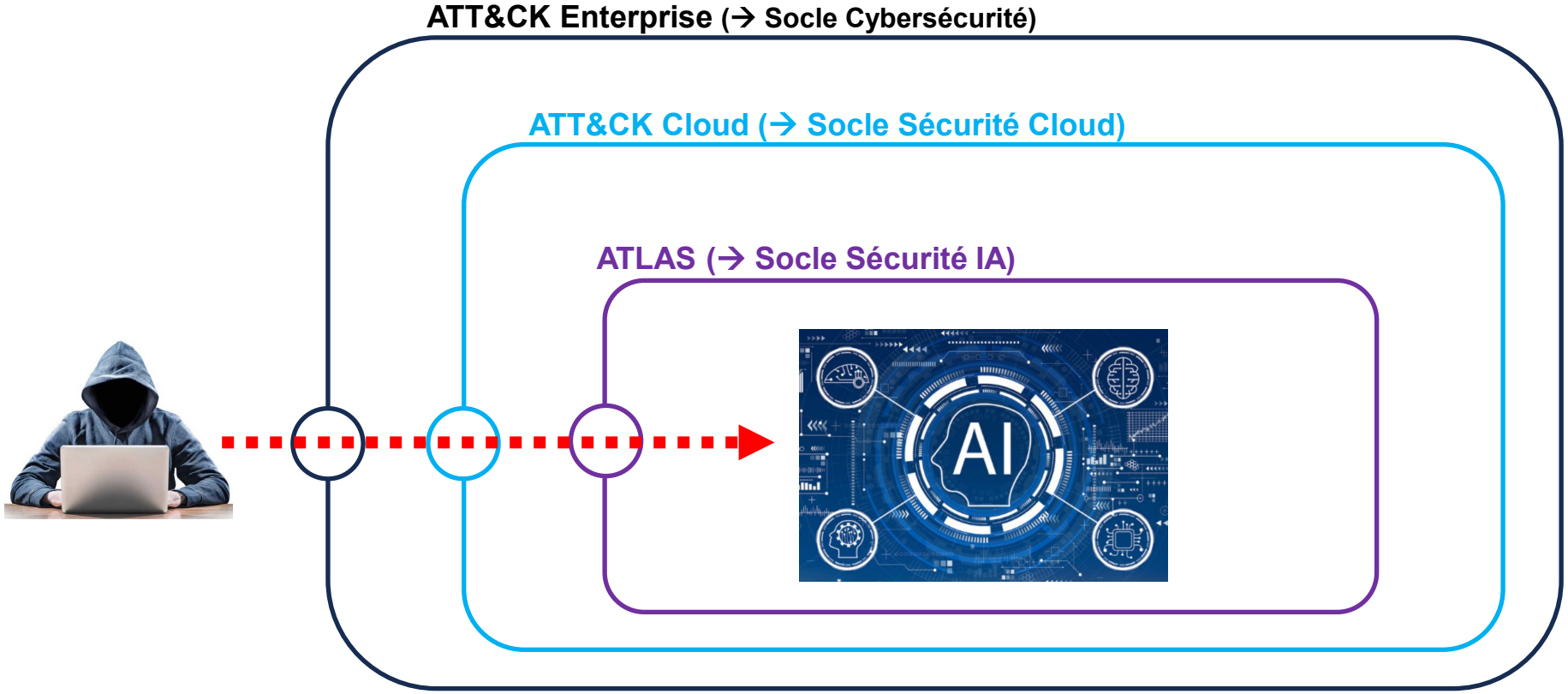
Mesures de sécurité IA (niv1)
- à mettre en oeuvre dès la conception pour tout projet de SIA sur la base des recommandations ANSSI, BSI, OWASP, NIST, MITRE, etc...

Source : Analyse des attaques sur les systèmes d'IA - Campus CYBER (02/2026)

SECUREAI
www.SecureAI.fr

Scénario opérationnel global

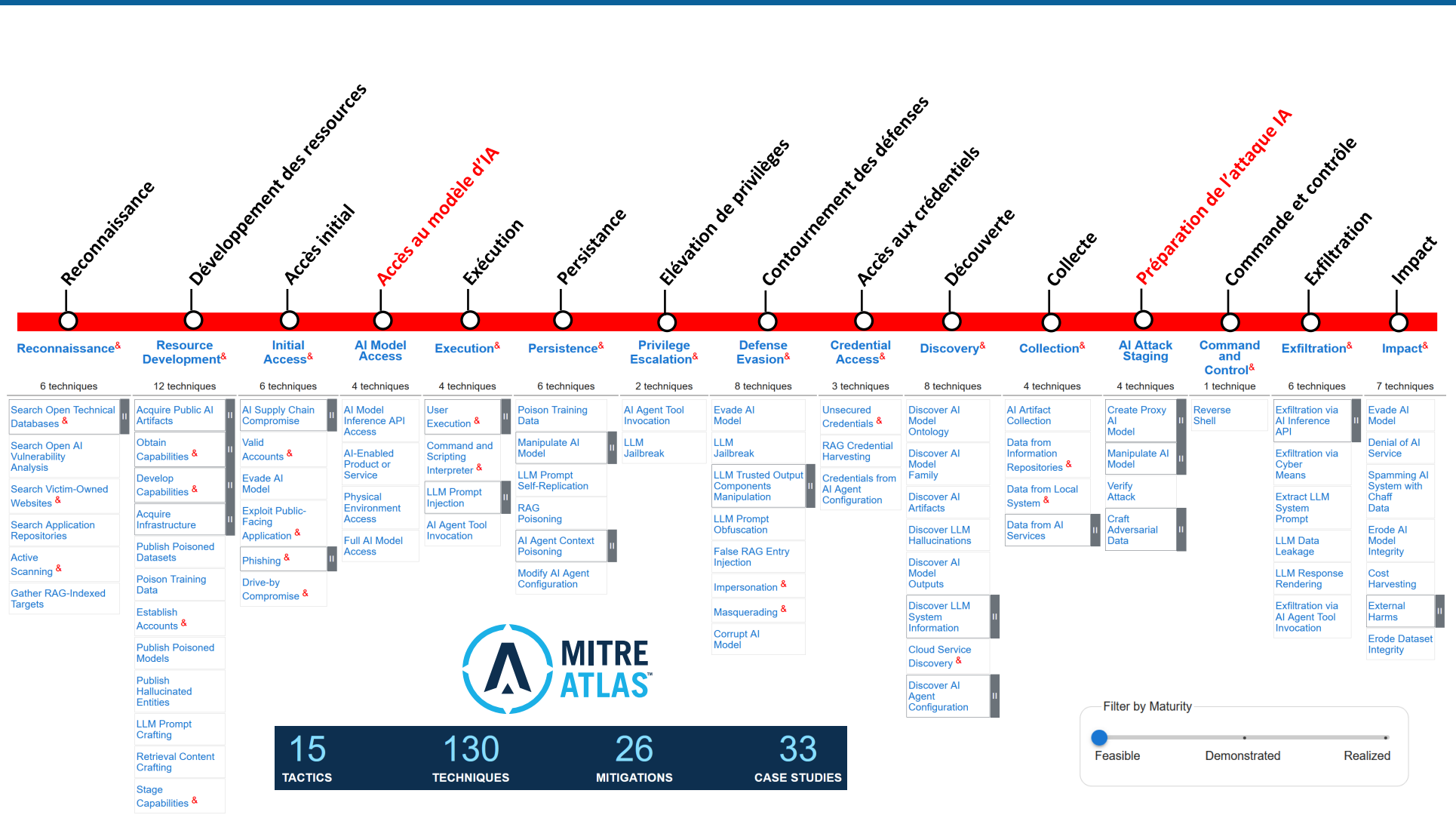
- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM



SECUREAI
www.SecureAI.fr

ATLAS (Adversarial Threat Landscape for AI Systems)

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM



SECUREAI
www.SecureAI.fr

1ère question ...

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM

SECUREAI
www.SecureAI.fr

2ème question ...

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM

SECUREAI
www.SecureAI.fr

25 termes de l'IA avec le mot « learning »

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM

SECUREAI
www.SecureAI.fr

25 termes de l'IA avec le mot « learning »

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM

Terme	Signification
Active Learning (AL)	Le modèle choisit lui-même les exemples qu'il fait étiqueter par un humain, pour maximiser l'apprentissage avec un budget d'annotation minimal.
Continual Learning (CL)	Apprendre séquentiellement de nouvelles tâches sans oublier les précédentes (problème du catastrophic forgetting).
Contrastive Learning	Apprendre des représentations en rapprochant les exemples similaires et en éloignant les différents. Base de CLIP, SimCLR, MoCo.
Curriculum Learning	Présenter les exemples du plus simple au plus complexe, par analogie avec l'apprentissage humain (Bengio 2009).
Deep Learning (DL)	Sous-ensemble du ML basé sur des réseaux de neurones profonds (CNN, Transformer, etc.).
Differential Privacy Learning	Entraînement avec garanties mathématiques de confidentialité (epsilon-DP). Algorithme de référence : DP-SGD (Abadi et al. 2016).
Federated Learning (FL)	Entraînement décentralisé : les données restent locales chez chaque participant, seules les mises à jour du modèle sont agrégées (Google 2016).
Few-Shot Learning (FSL)	Apprendre à résoudre une tâche à partir de quelques exemples seulement (typiquement 1 à 16).
In-Context Learning (ICL)	Apprendre une tâche uniquement à partir d'exemples placés dans le prompt, sans modifier les poids du modèle. Découvert avec GPT-3.
Machine Learning (ML)	Discipline générale : apprendre des fonctions à partir de données. Englobe régression, SVM, arbres, réseaux de neurones, etc.
Meta-Learning	« Apprendre à apprendre » : acquérir des stratégies d'apprentissage transférables à de nouvelles tâches. Référence : MAML (Finn et al. 2017).
Multi-Agent Reinforcement Learning (MARL)	Plusieurs agents apprennent simultanément par renforcement, en coopération ou en compétition. Base d'OpenAI Five, AlphaStar.
One-Shot Learning	Apprendre à reconnaître une classe à partir d'un seul exemple. Typique de la reconnaissance faciale (FaceNet).
Privacy-Preserving Machine Learning (PPML)	Famille de techniques préservant la confidentialité : differential privacy, homomorphic encryption (FHE), secure multi-party computation (MPC), TEE.
Reinforcement Learning (RL)	Apprendre par essai-erreur via des récompenses obtenues en interagissant avec un environnement. Base d'AlphaGo et de la robotique moderne.
Reinforcement Learning from AI Feedback (RLAIF)	RL où le feedback est généré par un autre modèle d'IA plutôt que par des humains. Base de Constitutional AI (Anthropic 2022).
Reinforcement Learning from Human Feedback (RLHF)	RL où la récompense est apprise à partir de préférences humaines. Technique clé d'InstructGPT et ChatGPT (OpenAI 2022).
Reinforcement Learning via Verifiable Rewards (RLVR)	RL où la récompense est calculée par vérification automatique (math, code, logique). Base de DeepSeek-R1 (2024-2025).
Robust Learning	Apprentissage de modèles résistants aux perturbations adversariales et au bruit. Théorie associée à l'adversarial training (Madry 2017).
Self-Supervised Learning (SSL)	Le modèle crée lui-même ses étiquettes à partir des données (mot suivant, pixel masqué). Base du pretraining de BERT, GPT, CLIP.
Semi-Supervised Learning	Apprentissage à partir d'un petit jeu étiqueté complété par un grand jeu non étiqueté. Très utilisé quand l'étiquetage humain est coûteux.
Supervised Learning	Apprentissage à partir de paires (entrée, étiquette). Paradigme classique de la classification et de la régression.
Transfer Learning	Réutiliser un modèle pré-entraîné pour résoudre une nouvelle tâche, généralement par fine-tuning. Paradigme dominant en pratique.
Unsupervised Learning	Apprentissage de structures dans des données non étiquetées : clustering, réduction de dimension (PCA), détection d'anomalies.
Zero-Shot Learning (ZSL)	Résoudre une tâche sans aucun exemple d'entraînement spécifique, en s'appuyant sur des connaissances générales. Typique des LLM modernes.

SECUREAI
www.SecureAI.fr

- Définir la sécurité de l’IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l’IA
- La conformité AI Act
- **Nouveaux métiers CAISO et CAIRO**
- Certifications SIAE & AAISM

SECUREAI
www.SecureAI.fr

DICTIONNAIRE DE LA SÉCURITÉ DE L’IA

+280 TERMES POUR COMPRENDRE ET SÉCURISER L’IA

https://www.verisafe.fr/dictionnaire-securite-ia

WEBINAIRE VERISAFE **- Sécurité de l’IA en entreprise - 5 mai 2026**

SIAE vs AAISM de l'ISACA

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- **Certifications SIAE & AAISM**

SECUREAI
www.SecureAI.fr

	SIAE (Sécuriser l'IA en Entreprise)	AAISM (Advanced AI Security Management)
	Organisme certificateur	
	VERISAFE	**ISACA**
Examen de certification		
Prérequis	Aucun	Certification CISM ou CISSP obligatoire
Format	Distanciel	Présentiel ou distanciel
Langue	Français	Anglais
Type	QCM (20 questions) + questions ouvertes (8 questions) + étude de cas + soutenance devant jury	QCM en anglais (90 questions)
Durée	4h + 1h (soutenance)	2h30
Formation de préparation à la certification		
Référence	SIA-STD	AIS-STD
Bundle avec examen	SIA-EXA (avec coaching personnalisé par visioconférence)	AIS-EXA
Langue	Français	Français
Examen Blanc	1 examen blanc (QCM/questions ouvertes/étude de cas) en français	3 examens (QCM 90 questions) en anglais
Programme de certification		
Domaines	D1 : Gouvernance et stratégie de sécurité des systèmes d'IA D2 : Gestion des risques spécifiques aux systèmes d'IA D3 : Sécurisation technique des systèmes d'IA D4 : Supervision, résilience et gestion des incidents IA D5 : Conformité réglementaire et maîtrise des fournisseurs IA	D1 : AI Governance and Program Management D2 : AI Risk and Opportunity Management D3 : AI Technologies and Controls
Positionnement		
Approche globale	Gouvernance, risques et conformité (GRC) + sécurité opérationnelle + supervision et réponse à incident	Gouvernance et gestion des risques de l'IA
Orientation métier	Responsable Sécurité IA / Head of AI Security / CAISO	AI Security Manager

Source : SecureAI.fr

Cursus CYBERPRO

- 15 formations
- 9 certifications

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- **Certifications SIAE & AAISM**

Fondamentaux de cybersécurité

- **Cybersécurité 360°** (CCI) : La sécurité à l'ère du Cloud et de l'IA
- **Certified in Cybersecurity** (PCC) : Consolider les bases et préparer la certification CC de l'(ISC)²

Maîtrise des risques

- **Gestion des risques avec l'ISO 27005** (IRM) : Maîtriser la norme ISO 27005 et préparer la certification Risk Manager
- **Gestion des risques avec EBIOS RM** (ERM) : Maîtriser la méthode de l'ANSSI et préparer la certification EBIOS RM

Cybersécurité avancée

- **Cybersécurité Synthèse Technique** (CST) : Maîtriser les aspects techniques de la cybersécurité
- **Certified Information Systems Security Professional** (CIS-STD) : Préparer la certification CISSP de l'(ISC)²

Sécurité de l'IA

- **Sécurité de l'intelligence artificielle** (SIA) : Sécurité, fiabilité, conformité et éthique pour bâtir une IA de confiance
- **Sécurité de l'intelligence artificielle** (AIS) : Préparation à la certification AAISM de l'ISACA

Juridique

- **RGPD / GDPR** (RGP) : Maîtriser le règlement européen sur la protection des données personnelles
- **DORA** (DOR) : Maîtriser le règlement européen sur la résilience cyber du secteur financier
- **NIS2** (NIS) : Maîtriser la directive européenne NIS2

ISO/IEC 27001

- **Auditer un SMSI ISO 27001** (ILA) : Maîtriser la norme ISO 27001 et préparer la certification Lead Auditor
- **Implémenter un SMSI ISO 27001** (ILI) : Maîtriser la norme ISO 27001 et préparer la certification Lead Implementer

Sécurité du Cloud

- **Sécurité du Cloud Computing** (SDC) : Assurer la sécurité et la souveraineté des données dans le Cloud
- **Certificate of Cloud Security Knowledge** (CSK-STD) : Préparer la certification CCSK de la CSA

SECUREAI
www.SecureAI.fr

WEBINAIRE VERISAFE - **Sécurité de l'IA en entreprise - 5 mai 2026**

VERISAFE vs CERTICYBER

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- **Certifications SIAE & AAISM**

SECUREAI
www.SecureAI.fr

VERISAFE

Entreprises

- ✓ Financement par entreprise
- ✓ Financement par OPCO
- ✓ Financement par CPF

Formation éligible CPF MON COMPTE FORMATION moncompteformation.gouv.fr

info-cpf@verisafe.fr

www.verisafe.fr

CERTI CYBER FORMATIONS

Particuliers

- ✓ Non finançable par OPCO / CPF
- ✓ Tarif particulier avec TVA à 0%

Plus d'infos !

commercial@verisafe.fr

www.certicyber.com

Merci de votre attention !

- Définir la sécurité de l'IA
- Cyber IT vs Cyber IA
- Top 10 des risques IA Entreprise
- Redéfinir le Zero Trust pour l'IA
- La conformité AI Act
- Nouveaux métiers CAISO et CAIRO
- Certifications SIAE & AAISM

SECUREAI
www.SecureAI.fr

Rejoignez-moi sur LinkedIn

https://www.linkedin.com/in/boris-motylewski