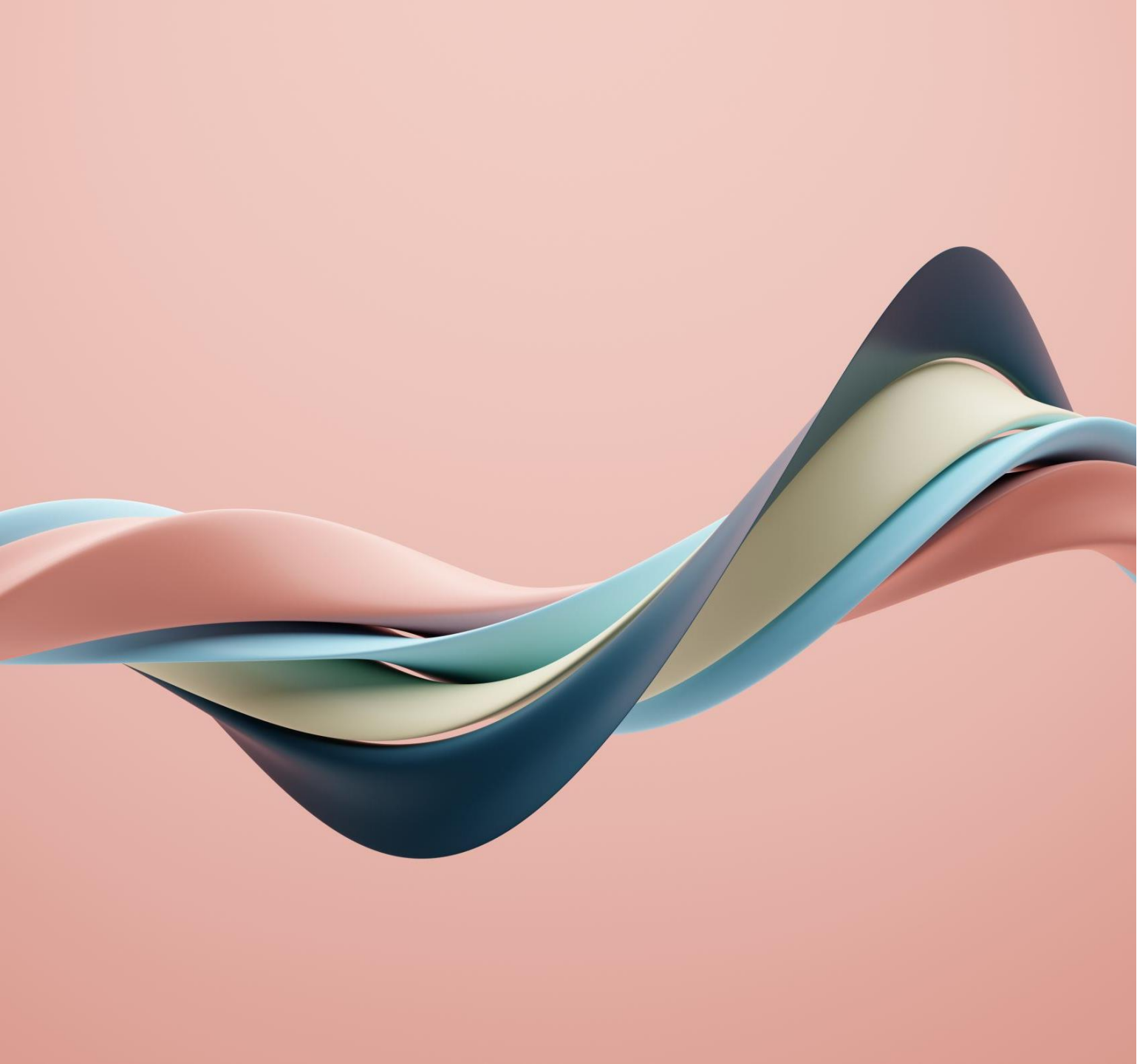
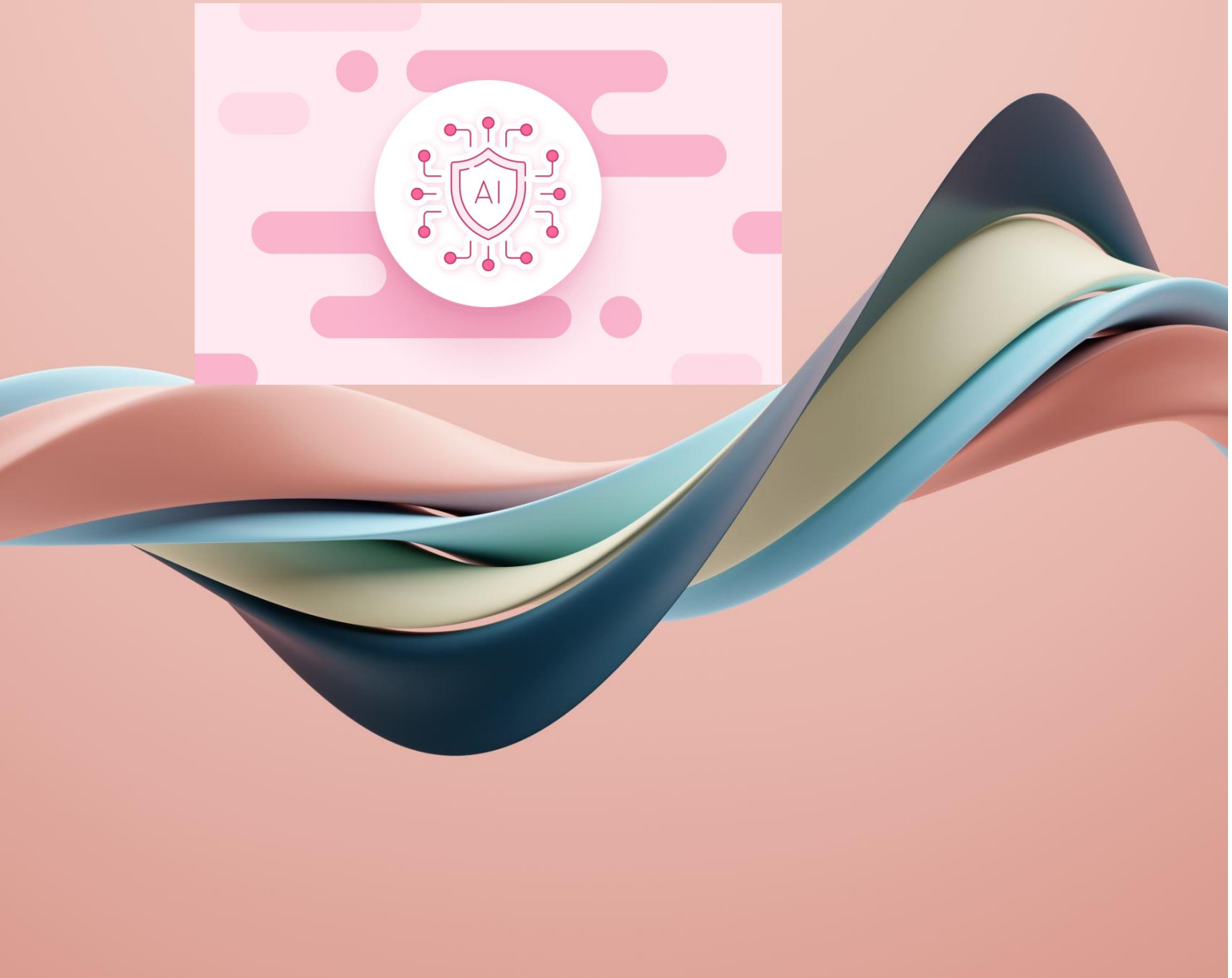


Sécurité et gouvernance de l’Intelligence Artificielle



Sécurité et gouvernance de
l'Intelligence Artificielle

SÉCURITÉ DE L'IA ÉLARGIE

LLMs

- Les LLMs transforment radicalement les dynamiques en cybersécurité.
- Ils ne sont plus seulement utilisés dans les postures cyber de défense, ils sont aussi utilisés pour automatiser des attaques.

A votre avis, dans ce que nous avons évoqué jusque-là, qu'est-ce qui est applicable aux modèles génératifs ?

PARTIE 1

Sécurisation des modèles génératifs

SÉCURISATION DES MODÈLES GÉNÉRATIFS

Contexte sur les LLM :

- LLMs = (gigantesque) système d'IA basé sur le principe des Transformers[1], capable de comprendre et générer du langage naturel.
- Entraînés sur beaucoup (= des milliards) de « tokens ».
- Capables de résumer, traduire, classer, dialoguer, assister des processus métier.
- Comportement non déterministe, probabiliste. Donc manipulable ?

[1] https://arxiv.org/abs/1706.03762

https://www.soprasteria.fr/services/offre-cybersecurite/comment-exploiter-pleinement-l-ia-tout-en-maitrisant-ses-risques/securiser-le-choix-de-son-LLM-face-aux-nouveaux-risques-cyber

SÉCURISATION DES MODÈLES GÉNÉRATIFS

Hugging Face

https://huggingface.co/datasets/open-llm-leaderboard/contents

https://huggingface.co/models?pipeline_tag=text-generation&sort=trending

Plus de 150 modèles activement référencés, avec disons une trentaine qui sont majeurs.

- Des modèles propriétaires poussant l'état de l'art, souvent disponibles via API cloud (GPT-5, Claude 4, Google Gemini…)
- Des modèles open source industriels ou académiques (LLaMA, Mistral AI, DeepSeek, BLOOM, EleutherAI…)
- Des LLMs spécialisés : pour le code, pour la vitesse, la multimodalité…
- Des modèles souverains

https://www.soprasteria.fr/services/offre-cybersecurite/comment-exploiter-pleinement-l-ia-tout-en-maitrisant-ses-risques/securiser-le-choix-de-son-LLM-face-aux-nouveaux-risques-cyber

Question

A VOTRE AVIS, RELATIVEMENT AUX LLMS, QUELLE PART DE MOYENNES ET GRANDES ENTREPRISES SE DISENT PRÊTES SUR LE PLAN DE LA CYBERSÉCURITÉ ? (EN 2024)

Question

A VOTRE AVIS, RELATIVEMENT AUX LLMS, QUELLE PART DE MOYENNES ET GRANDES ENTREPRISES SE DISENT PRÊTES SUR LE PLAN DE LA CYBERSÉCURITÉ ? (EN 2024)

- ENTRE 2% ET 5% SELON LES SOURCES

SÉCURISATION DES MODÈLES GÉNÉRATIFS

des entreprises ont subi au moins un incident lié à l'IA, et 60 % se disent mal préparées face à ce nouveau risque.
(Metomic. Lakera)

des professionnels estiment que l'IA rend le phishing plus crédible et plus difficile à détecter.
(Lakera. Arctic Wolf)

des entreprises se déclarent « **matures** » en cybersécurité LLM.
(Techradar. Cobalt.io)

SÉCURISATION DES MODÈLES GÉNÉRATIFS

Le choix d'un LLM ne repose pas uniquement sur des critères de performance brute, de coût ou de latence. Il doit aussi intégrer une lecture structurelle.

Quel est le degré de contrôle requis ?

Quelles sont les contraintes réglementaires ?

Quels sont les risques cyber liés à l'usage du modèle ?

https://www.soprasteria.fr/services/offre-cybersecurite/comment-exploiter-pleinement-l-ia-tout-en-maitrisant-ses-risques/securiser-le-choix-de-son-LLM-face-aux-nouveaux-risques-cyber

SÉCURISATION DES MODÈLES GÉNÉRATIFS

Risques liés à l'usage des LLM :

- Fuites de données
- Dérives sémantiques
- Dépendances techniques
- Comportements imprévus

SÉCURISATION DES MODÈLES GÉNÉRATIFS

Certaines attaques visant les LLM ne ressemblent pas aux menaces traditionnelles :

- Exploitation de vecteurs linguistiques, visuels ou semi-structurés
- Injection de prompt/jailbreaking,
- Empoisonnement de données dans les corpus d'entraînement ou les bases RAG,
- Manipulation d'agents autonomes ou de plugins outillés,
- Fuite d'information par hallucinations contrôlées

SÉCURISATION DES MODÈLES GÉNÉRATIFS

Les LLM n'ont pas été spécialement conçus pour évoluer en environnement hostile.

Ils doivent donc être intégrés avec une vigilance accrue :

- Analyse des risques,
- Choix d'architecture sécurisé
- Monitoring en continu
- Fort contrôle sur les données

SÉCURISATION DES MODÈLES GÉNÉRATIFS

La généralisation des principes comme le RAG élargit la surface d'attaque : ces interfaces deviennent des points d'entrée pour des manipulations (injection de prompt, exfiltration contextuelle…).

La question du fine tuning

- Améliore les performances métier mais augmente sa surface d'attaque : est-ce que vous voyez pourquoi ?

Le fine tuning peut fragiliser les garde-fous initiaux.

MÉCANISMES DE PROTECTION

LLM As-A-Service

- Les données quittent le SI
- Personnalisation limitée
- Le fournisseur impose ses garde-fous
- Manque de visibilité sur les logs, la logique du modèles, les mises à jour

Mitigations recommandées

Intégrer un middleware de filtrage et journalisation des prompts

Isoler l'usage dans un VPC cloud ou réseau dédié

Activer le chiffrement end-to-end et la journalisation externe

MÉCANISMES DE PROTECTION

LLM auto-hébergé

- Ressources (GPU) importantes
- Compétences
- Maintenance
- Ce n'est pas parce qu'un LLM est hébergé localement qu'il est digne de confiance (vulnérable à des attaques spécifiques)

Mitigations recommandées

Coupler avec des outils de modération locale (LLM en « sandwich »)

Déployer dans des environnements cloisonnés

Scanner régulièrement les usages avec des outils[1] de sécurité LLM

[1] Quelques exemples sans être sponsorisé : Lakera Guard, Mirror Security…

MÉCANISMES DE PROTECTION

On retrouve certains mécanismes mentionnés dans l'historique en introduction. Revenons sur deux d'entre eux.

MÉCANISMES DE PROTECTION

Watermarking des modèles

On va introduire volontairement et discrètement une signature dans les contenus générés par le modèle.

Ici, quelle est l'utilité en termes de protection ?

- Identifier si un contenu a été généré par une IA.
- Prévenir la désinformation (deepfakes, faux documents).
- Tracer l'origine d'un modèle modifié.
- Dissuader la réutilisation ou la modification non autorisée.

Types de watermarking utilisés :

- **Statistiques** (motifs dans la distribution des mots/outputs)
- **Cryptographiques** (hash ou signature secrète introduite par le modèle)
- **Sémantiques** (choix de mots ou styles spécifiques, plus rare).

MÉCANISMES DE PROTECTION

Mécanismes de filtrage automatique (OpenAI Moderation API, RLHF…)

OpenAI Moderation API https://platform.openai.com/docs/guides/moderation

- Service qui vérifie le texte fourni par un utilisateuret/le texte généré par un modèle
- Il classe le contenu en différentes catégories :
  - violence
  - discours haineux
  - sexualité
  - harcèlement
  - automutilation
  - crimes
  - sécurité nationale
  - données privées
- La réponse du modèle est ensuite bloquée ou reformulée si elle enfreint les règles.

MÉCANISMES DE PROTECTION

Mécanismes de filtrage automatique (OpenAI Moderation API, RLHF…)

RLHF

Question

POUVEZ-VOUS ME RAPPELER LA SIGNIFICATION ET LE PRINCIPE DE RLHF ?

MÉCANISMES DE PROTECTION

Mécanismes de filtrage automatique (OpenAI Moderation API, RLHF…)

RLHF

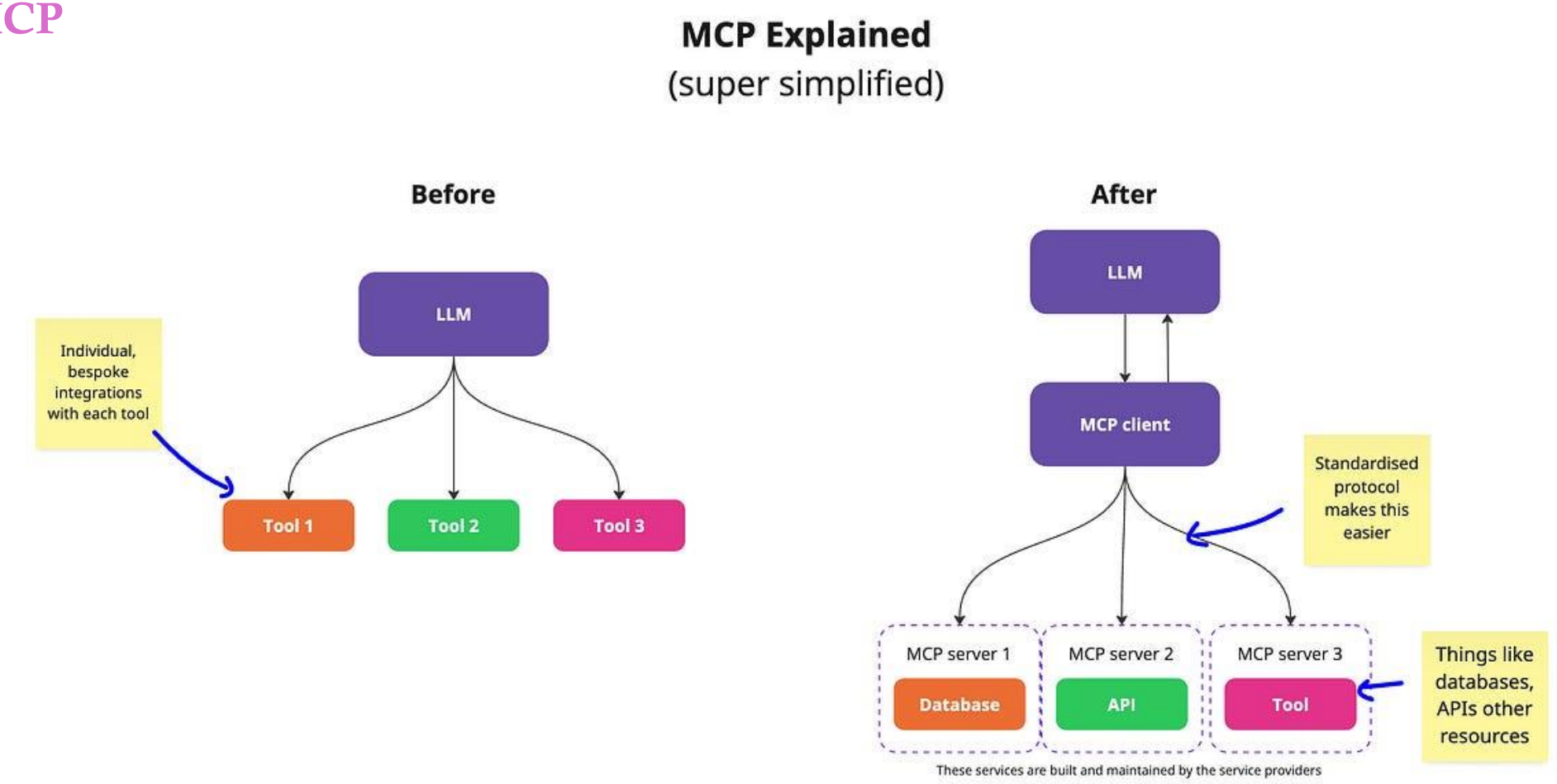
- Permet d'entraîner le modèle pour qu'il respecte les règles, suive des comportements souhaitables, évite des réponses nuisibles et plus simplement refuse les prompts dangereux

1. Des humains évaluent des réponses du modèle (bonnes / mauvaises).
2. On crée un modèle de récompense (Reward Model).
3. Le modèle génératif est ré-entraîné pour maximiser cette récompense.

- Le modèle apprend à refuser spontanément des demandes interdites. Il devient plus sûr, même sans filtrage externe.

CAS PARTICULIER DU MODEL CONTEXT PROTOCOL, RISQUES ASSOCIÉS

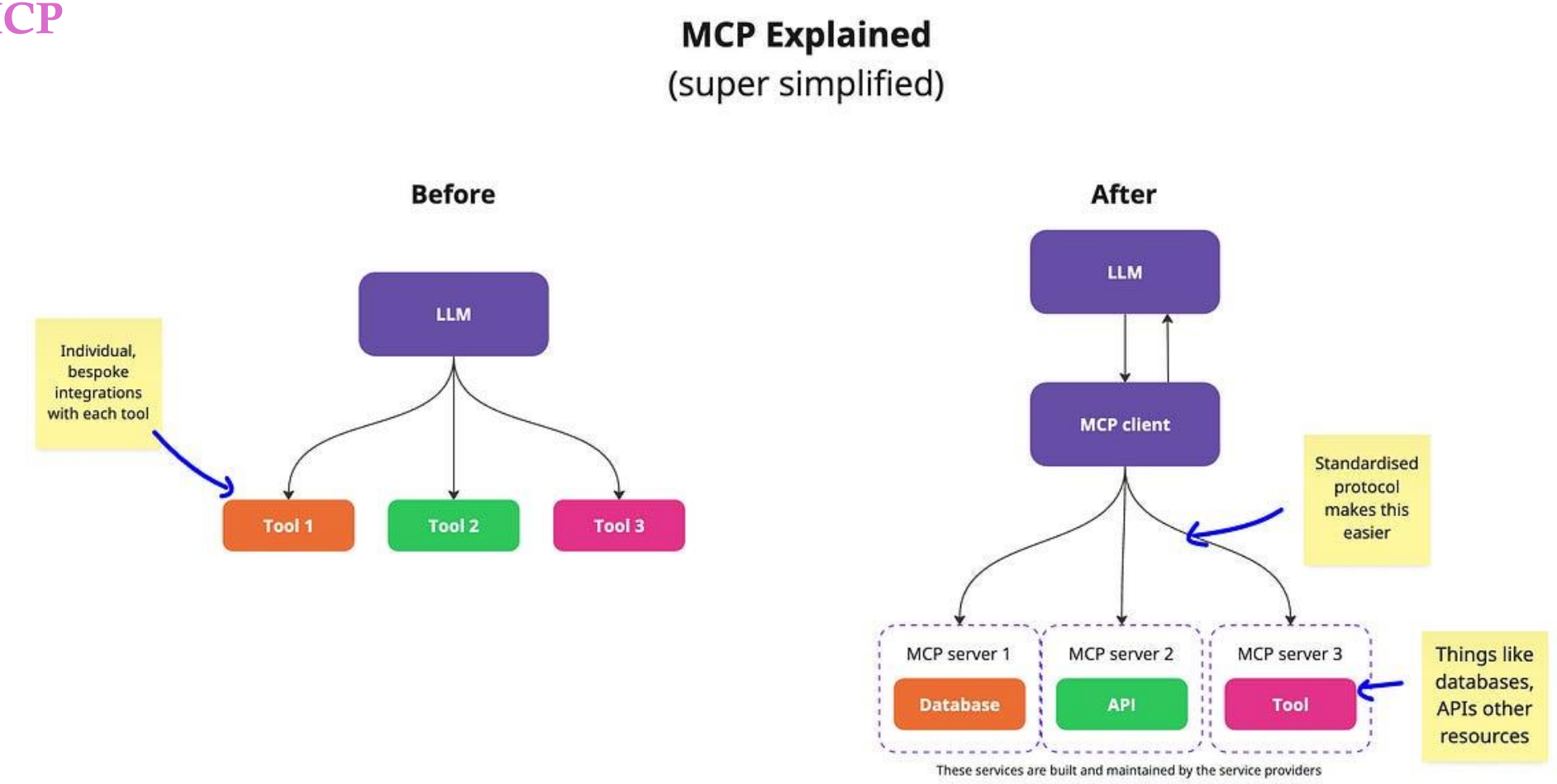
Le MCP



https://departmentofproduct.substack.com/p/mcp-explained-a-simple-guide-for

CAS PARTICULIER DU MODEL CONTEXT PROTOCOL, RISQUES ASSOCIÉS

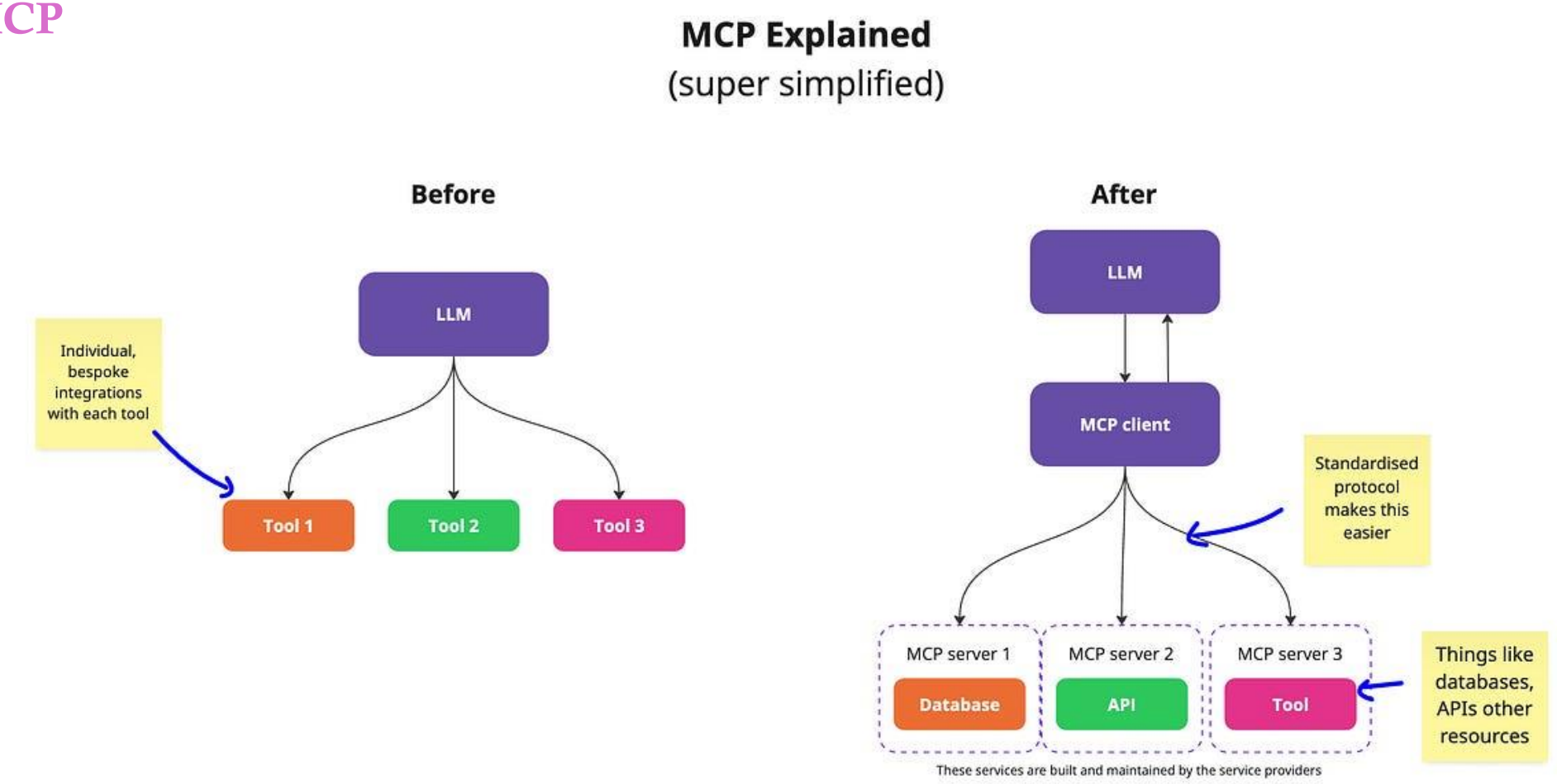
Le MCP



https://departmentofproduct.substack.com/p/mcp-explained-a-simple-guide-for

CAS PARTICULIER DU MODEL CONTEXT PROTOCOL, RISQUES ASSOCIÉS

Le MCP



https://departmentofproduct.substack.com/p/mcp-explained-a-simple-guide-for

Jailbreak prompts

ATELIER PRATIQUE

Vulnérabilités dans des implémentations MCP

VULNÉRABILITÉS DANS DES IMPLÉMENTATIONS MCP

Vous pouvez tester les implémentations MCP si vous le souhaitez.

Pour aller plus loin sur ce sujet, vous pouvez utiliser les outils LMStudio (pour télécharger un LLM en local) et Cline, ou outils équivalents.

Cette étude de code pratique a été en particulier créée à partir d'un dépôt qui peut vous aider à mettre en place les serveurs associés :

Lien vers le dépôt Github

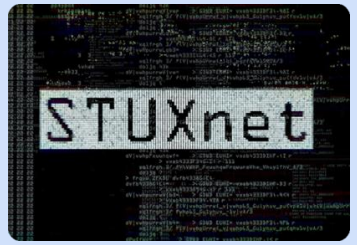
PARTIE 2

Sécurité opérationnelle et cybersécurité industrielle

ATTAQUES SUR INFRASTRUCTURES CRITIQUES - VULNÉRABILITÉS SCADA/IOT

- **SCADA** (Supervisory Control and Data Acquisition) : systèmes de contrôle industriels pour usines, réseaux électriques, pipelines, etc.
- **IoT** industriel : capteurs, actionneurs, automates, robots connectés.

Attaque Stuxnet



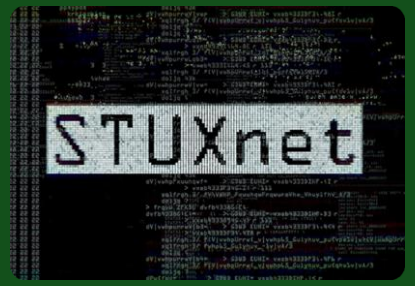
Période	Actif à partir de 2009, découvert en 2010.
Type d'attaque	Malware industriel ciblé (ver autonome). Sabotage de systèmes SCADA/ICS. Zero-day exploitation (4 zero-day Windows). Attaque furtive sur automates industriels (PLC). Manipulation et falsification de capteurs
Cible	Systèmes de contrôle industriels (ICS) des installations nucléaires iraniennes. Automates programmables industriels (PLC) Siemens S7-300. Environnements SCADA supervisant les centrifugeuses d'enrichissement d'uranium à Natanz
Principe	Altérer physiquement le fonctionnement des centrifugeuses tout en dissimulant l'attaque aux opérateurs humains.

Question

QUEL TYPE D'ATTAQUANT A MIS EN ŒUVRE STUXNET ?

- UN HACKEUR INDIVIDUEL CONNU SOUS LE PSEUDONYME DE SHADOWFLUX
- UN PETIT GROUPE DE HACKEURS VISANT À EXTORQUER DE L'ARGENT VIA RANSOMWARE
- UN PETIT GROUPE D'HACKTIVISTES
- UN GROUPE APT

Question



QUEL TYPE D'ATTAQUANT A MIS EN ŒUVRE STUXNET ?

- ☐ ~~UN HACKEUR INDIVIDUEL CONNU SOUS LE PSEUDONYME DE SHADOWFLUX~~
- ☐ ~~UN PETIT GROUPE DE HACKEURS VISANT À EXTORQUER DE L'ARGENT VIA RANSOMWARE~~
- ☐ ~~UN PETIT GROUPE D'HACKTIVISTES~~
- ☑ UN GROUPE APT

Le ver :

- infecte Windows via plusieurs vulnérabilités zero-day,
- se propage dans les réseaux industriels (même isolés),
- cible spécifiquement le logiciel industriel Siemens Step7,
- modifie silencieusement les commandes envoyées aux PLC,
- falsifie simultanément les données de capteurs pour masquer le sabotage.

➢ Les centrifugeuses tournent à des vitesses destructrices pendant que le SCADA affiche des valeurs « normales ».

Méthode un peu plus détaillée

1. Exploitation systémique

- Utilisation de **4 vulnérabilités zero-day Windows** pour entrer dans les réseaux (LNK exploit, escalade de privilèges, etc.).
- Propagation via clés USB dans des environnements industriels non connectés à Internet.

2. Prise de contrôle du logiciel industriel

- Injection dans le système Siemens Step7 pour intercepter et altérer les commandes envoyées aux automates et les programmes PLC chargés en mémoire.

3. Sabotage physique ciblé

- Envoi de séquences malveillantes aux centrifugeuses ayant pour effet des accélérations et décélérations brusques, conduisant à un stress mécanique entraînant leur destruction progressive.

Méthode un peu plus détaillée

4. Manipulation de capteurs (fausse télémétrie)

- Enregistre des données normales des capteurs.
- Pendant l'attaque, rejoue ces valeurs au SCADA pour masquer l'anomalie.
- Les opérateurs ne voient que des mesures "propres", aucune alerte n'est levée.

5. Mécanisme furtif avancé

- Blocage des mécanismes de détection.
- Identification précise de l'installation cible (signature de configuration PLC spécifique).
- Auto-effacement après exécution pour limiter la détection.

Centrifugeuses détruites ou fortement endommagées	**Environ 1000 (20% du parc)**
Retard pris dans le programme nucléaire	**Estimé autour de 1 an – 2 ans**

DÉTOURNEMENT DE CAPTEURS INDUSTRIELS

Détournement de capteurs industriels

- Attaques consistant à injecter de fausses mesures dans les capteurs (IoT, SCADA).
- Risque : prise de décisions automatiques incorrectes, pannes, accidents industriels.
- Exemple : injection de faux signaux de température ou pression pour perturber une usine.

Attaques par accès distant

- Exploitation de protocoles industriels non sécurisés (Modbus, DNP3, OPC-UA mal configuré).
- Intrusions dans le réseau industriel à travers des IoT ou VPN vulnérables

MENACES SUR L'IA DANS L'AUTOMATISATION INDUSTRIELLE

Attaques sur les modèles IA

- Data poisoning, adversarial attacks… on connait déjà !
- Quelques exemples génériques possibles modernes dans ce domaine :
  - IA de maintenance prédictive trompée par de fausses mesures (Stuxnet v2 ?)
  - Robots collaboratifs ou bras automatisés manipulés via des commandes malicieuses
- Dans tous les cas, l'impact est la perturbation de la production, avec des risques pour la sécurité humaine, de perte économique…

SÉCURISATION DES SYSTÈMES IA CONNECTÉS AUX INFRASTRUCTURES CRITIQUES

Isolation des réseaux industriels

- Segmenter les réseaux SCADA, IoT et IA.
- Utiliser des VLAN, DMZ ou VPC privés pour séparer les flux critiques.

Chiffrement et authentification

- Chiffrement bout-à-bout des communications entre capteurs, automates et IA.
- Authentification forte des appareils IoT et des endpoints IA.

Surveillance et détection d'anomalies

- Déploiement de SIEM et IDS/IPS spécialisés pour les réseaux industriels.
- Détection des anomalies dans les sorties de l'IA et des capteurs.

Robustesse des modèles IA

- Prévention du data poisoning via validation stricte des données.
- Tests adversariaux pour vérifier la résilience aux entrées malveillantes.
- Mise en place de fail-safes humains ou automatiques en cas de comportement suspect.

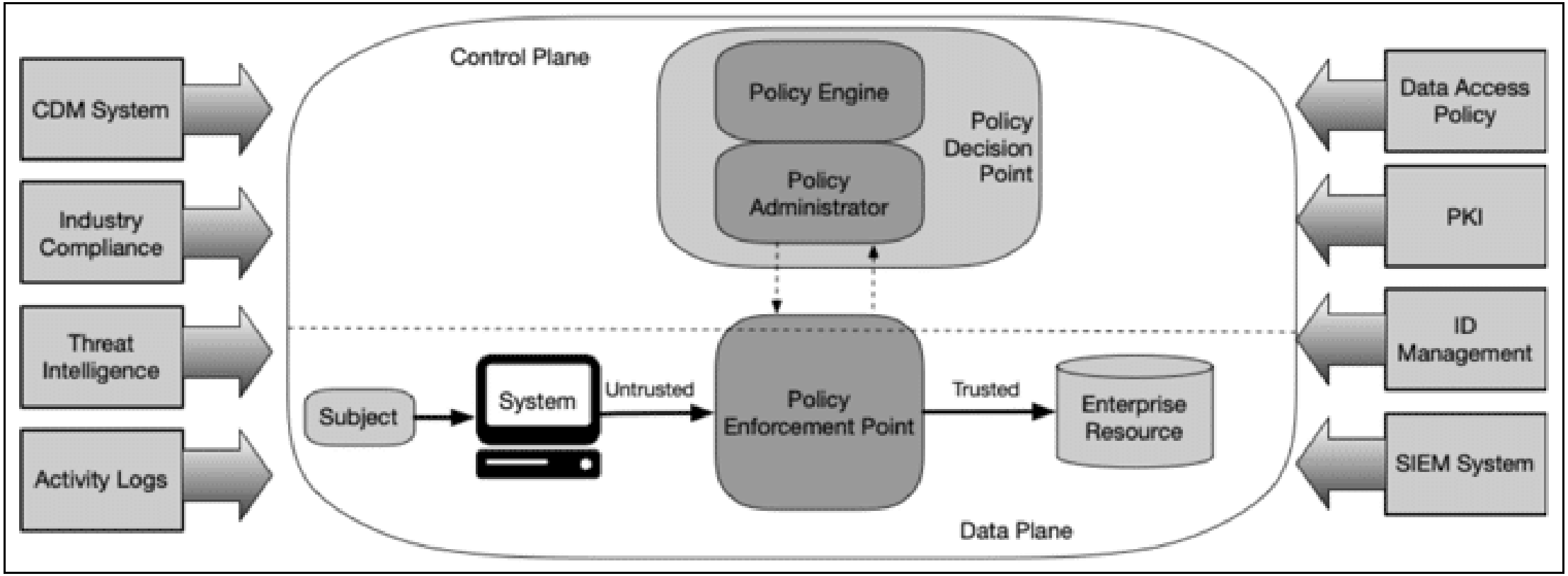
INTRODUCTION AUX ARCHITECTURES ZERO-TRUST EN ENTREPRISE

- Classiquement, une fois qu'un acteur malveillant a pénétré le périmètre de défense, il est libre de se déplacer à l'intérieur du réseau à sa guise.
- Dans le modèle Zero Trust, chaque requête réseau doit être traitée comme si le réseau avait déjà été compromis et même les requêtes simples devraient être considérées comme une menace potentielle.

Principe clé : *« Never trust, always verify »*

- Chaque appareil, utilisateur et service est authentifié et vérifié en continu, même à l'intérieur du réseau interne.

INTRODUCTION AUX ARCHITECTURES ZERO-TRUST EN ENTREPRISE



Identity Provider (IdP) et IAM

Gestion des idéntités. L'identité devient le nouveau périmètre.

Rôles du composant IAM :

- Authentifier utilisateurs, machines, services, API
- Gérer les droits (RBAC, ABAC, PBAC)
- Appliquer le principe de moindre privilège

Exemples :

- Azure AD / Entra ID
- Okta
- Keycloak
- Ping Identity
- Identity Federation (SAML, OIDC)

MFA / authentification forte

Pas d'accès basé sur un simple mot de passe.

- Continue
- Contextuelle
- Multi-facteur (MFA/2FA)

Exemples :

- Biométrie
- Clés FIDO2 / YubiKey
- Applications d'authentification (TOTP)
- Challenge/response

Device Trust

Les terminaux deviennent des “identités” au même titre que les utilisateurs.

Le système vérifie :

- L’état du device (patching, antivirus/EDR)
- La posture de sécurité (Secure Boot, chiffrement disque)
- L’intégrité (pas rooté/jailbreaké)
- L’appartenance au parc d’entreprise

Device non conforme = accès refusé.

Network Trust

Pas de confiance au niveau réseau.

- Software Defined Perimeter (SDP)
- Micro-segmentation (ex : Illumio, VMware NSX)
- MTLS entre services
- Z-TNA (Zero Trust Network Access), remplaçant du VPN

Policy Engine (PE)

Détermine si une requête doit être acceptée, dégradée ou rejetée. Risk-based access control.

Critères :

- Identité
- Poste
- Emplacement
- Risque en temps réel
- Comportement historique
- Sensibilité de la ressource ciblée
- Heure de la journée / contexte

Policy Enforcement Point (PEP)

Applique la décision du moteur PE.

Peut être situé sur :

- un proxy d'accès
- un reverse proxy / gateway
- un agent installé sur le poste
- un load balancer Zero Trust

Actions possibles :

- autoriser l'accès
- bloquer
- demander une réauthentification / imposer une posture renforcée
- limiter l'accès (lecture seule, no-download, etc.)

Surveillance continue

- SIEM centralisé
- UEBA
- EDR / XDR
- logs d'accès et d'actions systémiques
- détection d'anomalies et déviation de comportement

Le but est de détecter les abus de privilèges, les tentatives de contournement, le mouvement latéral, l'activité inhabituelle…

INTRODUCTION AUX ARCHITECTURES ZERO-TRUST EN ENTREPRISE

Zero-trust pour IA industrielle

- **Contrôle strict des flux IA vers SCADA / IoT**
  - L'IA ne peut accéder aux automates ou capteurs qu'après vérification continue des droits.
- **Micro-segmentation**
  - Limiter la portée des modèles IA et de leurs communications.
  - Même si un composant est compromis, l'attaquant ne peut pas se déplacer latéralement.
- **Audit et journalisation centralisée**
  - Toutes les interactions IA vers infrastructures critiques sont loggées et auditées.
- **Authentification continue des modèles IA**
  - Chaque modèle ou agent IA doit s'authentifier régulièrement, avec rotation des clés et contrôle d'intégrité.

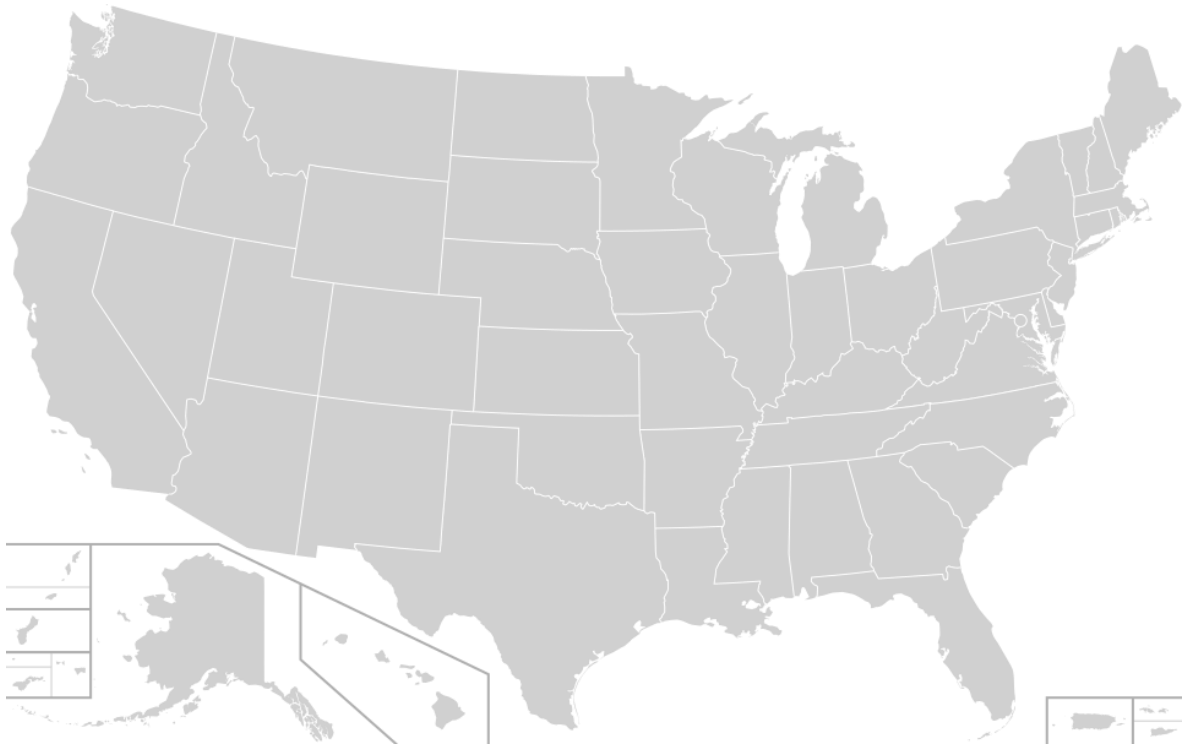
PARTIE 3

Sécurité des IA en entreprise : gouvernance et conformité

ADAPTATION DE MÉTHODES DE GOUVERNANCE CLASSIQUE

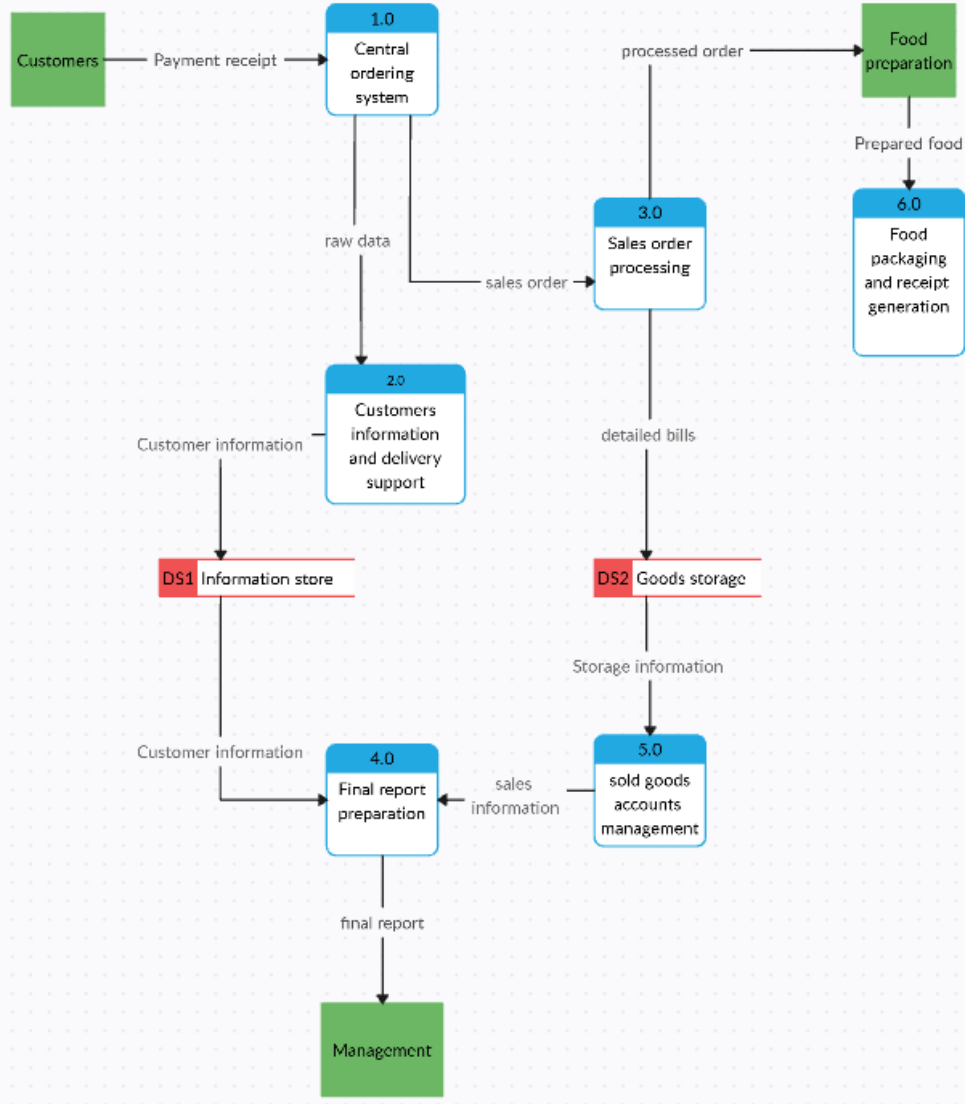
The Emerging Technology Observatory's AGORA

https://www.kaggle.com/datasets/umerhaddii/ai-governance-documents-data



ADAPTATION DE MÉTHODES DE GOUVERNANCE CLASSIQUE

DFD (diagramme de flux de données)



Template : https://creately.com/diagram-community/popular/t/data-flow

ADAPTATION DE MÉTHODES DE GOUVERNANCE CLASSIQUE

Rapport du Gartner : Effective AI governance

https://domino.ai/resources/journey-guide-managing-ai-governance-trust-risk-security

Databricks : A Comprehensive Guide to Data and AI Governance

https://www.databricks.com/resources/ebook/data-analytics-and-ai-governance

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

ISO 27001 - Norme de management de la sécurité de l'information (SMSI)

- Norme internationale de sécurité des systèmes d'information de l'ISO et la CEI.
- Définit les exigences pour établir, mettre en œuvre, maintenir et améliorer un SMSI
- Protéger la confidentialité, l'intégrité et la disponibilité de l'information au sein d'une organisation.

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

ISO 27001 - Norme de management de la sécurité de l'information (SMSI)

Principes clés

- Approche basée sur le risque
  - Identifier les actifs, menaces et vulnérabilités.
  - Évaluer et prioriser les risques pour appliquer des mesures adaptées.
- Amélioration continue
  - Suivi, audit et révision régulière du SMSI.
  - Adaptation aux évolutions technologiques et aux nouveaux risques.
- Intégration dans le management global
  - La sécurité n'est pas seulement technique : politique, processus et organisation sont inclus.

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

ISO 27001 - Norme de management de la sécurité de l'information (SMSI)

- 114 contrôles répartis en 14 domaines (annexe A).
  - ❖ Politique de sécurité
  - ❖ Organisation de la sécurité
  - ❖ Sécurité des ressources humaines
  - ❖ Gestion des actifs
  - ❖ Contrôle d'accès
  - ❖ Cryptographie
  - ❖ Sécurité physique et environnementale
  - ❖ Sécurité opérationnelle
  - ❖ Sécurité des communications
  - ❖ Acquisition, développement et maintenance des systèmes
  - ❖ Relations fournisseurs
  - ❖ Gestion des incidents
  - ❖ Aspects de continuité d'activité
  - ❖ Conformité
- Pertinent pour l'IA industrielle. Les contrôles peuvent inclure, par exemple :
  - Chiffrement des données sensibles.
  - Journalisation et monitoring.
  - Politiques d'accès strictes pour les systèmes IA industriels.
  - Gestion des vulnérabilités et patching.
- Renvoie aux bonnes pratiques d'ISO 27002.

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

ISO 27001 - Norme de management de la sécurité de l'information (SMSI)

Certification ISO 27001

- Une organisation peut obtenir une certification par un organisme tiers pour prouver la conformité.
- Processus : audit initial, correction des écarts, audits réguliers de suivi.
- Crédibilité, confiance clients, réduction des risques cyber et conformité réglementaire.

Quelles sont les principales normes qui concernent BNP Paribas ?

BNP Paribas détient plus de 80 certifications et labels à travers le monde. La majorité des certifications concerne la qualité du service fournie aux clients, **l'ISO 9001**. Il s'agit de « la » norme essentielle et « mère de toute les normes » destinée à établir la confiance entre tous les acteurs internes à l'entreprise et à fédérer toutes les énergies autour d'un seul objectif : **satisfaire le client**.

BNP Paribas détient également des certifications **ISO 27001**, norme en essor dans le Groupe qui lui permet de mettre en place un dispositif pour gérer les **risques liés à la sécurité de ses données** ou de celles qu'elle est amenée à traiter. Le traitement des données à caractère personnel et la libre circulation de ces données se trouvent ainsi sécurisés. Dans le domaine informatique et technologique, le législateur se base sur les meilleures pratiques internationales décrites dans les normes volontaires en transcrivant quelques années plus tard ces éléments dans la loi. **Être certifié permet à BNP Paribas de se maintenir à la pointe de la technologie**. Les donneurs d'ordre sont de plus en plus nombreux à exiger de leurs prestataires la preuve de la sécurité et de la résilience de leur système d'information.

BNP Paribas possède aussi des certifications **ISO 14001 sur l'environnement** et **ISO 50001 sur le management de l'énergie** qui, couplées au label Numérique responsable, démontrent que le Groupe est attentif à sa démarche de maîtrise des impacts environnementaux engendrés par ses activités.

https://group.bnpparibas/actualite/labels-et-normes-des-leviers-cles-de-la-qualite-de-service-de-bnp-paribas

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

ISO 27002 - Code de bonnes pratiques pour la sécurité de l'information

- Fournit des recommandations et bonnes pratiques pour mettre en œuvre la sécurité de l'information.
- Complète ISO 27001, mais elle ne fournit pas de critères de certification.
- Objectif : guider les organisations sur les contrôles de sécurité à appliquer pour protéger leurs informations.

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

ISO 27002 - Code de bonnes pratiques pour la sécurité de l'information

Domaine	Exemples IA industrielle
Politique de sécurité	Définition de règles et directives claires pour la sécurité de l'information
Organisation de la sécurité	Rôles et responsabilités, séparation des fonctions critiques
Sécurité des ressources humaines	Sensibilisation, vérification des antécédents, formation continue
Gestion des actifs	Inventaire des systèmes, classification des données sensibles (ex. datasets IA)
Contrôle d'accès	MFA, RBAC, ABAC, gestion des privilèges
Cryptographie	Chiffrement des données sensibles, protection des clés
Sécurité physique et environnementale	Accès sécurisé aux data centers et aux équipements SCADA/IoT
Sécurité opérationnelle	Patch management, journaux, surveillance en temps réel
Sécurité des communications	Protection des flux entre systèmes IA et automates industriels
Acquisition, développement et maintenance des systèmes	Sécurisation du cycle de vie des modèles IA et applications critiques
Relations fournisseurs	Gestion des risques liés aux tiers qui accèdent aux données
Gestion des incidents	Détection, réponse et retour d'expérience sur les incidents de sécurité
Continuité d'activité	Plan de reprise après sinistre et résilience des systèmes IA
Conformité	Respect de la législation, des régulations (ex. IA Act, GDPR)

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

ISO 27001 et 27002 dans le contexte IA industriel :

- Protection des modèles génératifs : prévention des fuites de données d'entraînement.
- Systèmes IA dans SCADA/IoT : sécurisation des communications, authentification forte, journalisation continue.
- Gouvernance IA en entreprise : intégration dans le SMSI, suivi des risques, conformité aux cadres ISO et NIST.

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

NIST AI Risk Management Framework

- Publié par le **NIST (National Institute of Standards and Technology)** en 2023
- Fournit aux organisations un cadre pour **identifier, évaluer, atténuer et surveiller les risques liés à l'IA**, incluant :
  - les modèles d'IA traditionnels,
  - les IA génératives,
  - les systèmes intégrés dans des environnements critiques (SCADA, IoT, automatisation industrielle).
- Pas une norme de conformité obligatoire, un **cadre volontaire** et **pragmatique**.

https://www.nist.gov/itl/ai-risk-management-framework

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

NIST AI Risk Management Framework

4 fonctions continues (comme pour le NIST Cybersecurity Framework)

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

Secteur financier : DORA (Digital Operational Resilience Act)

- Régulation européenne entrée en application début 2025.
- Obligatoire pour l'ensemble du secteur financier.
- Vise à renforcer la résilience opérationnelle numérique des acteurs financiers face aux cybermenaces, aux défaillances techniques et aux risques liés aux fournisseurs technologiques.

Objectifs :

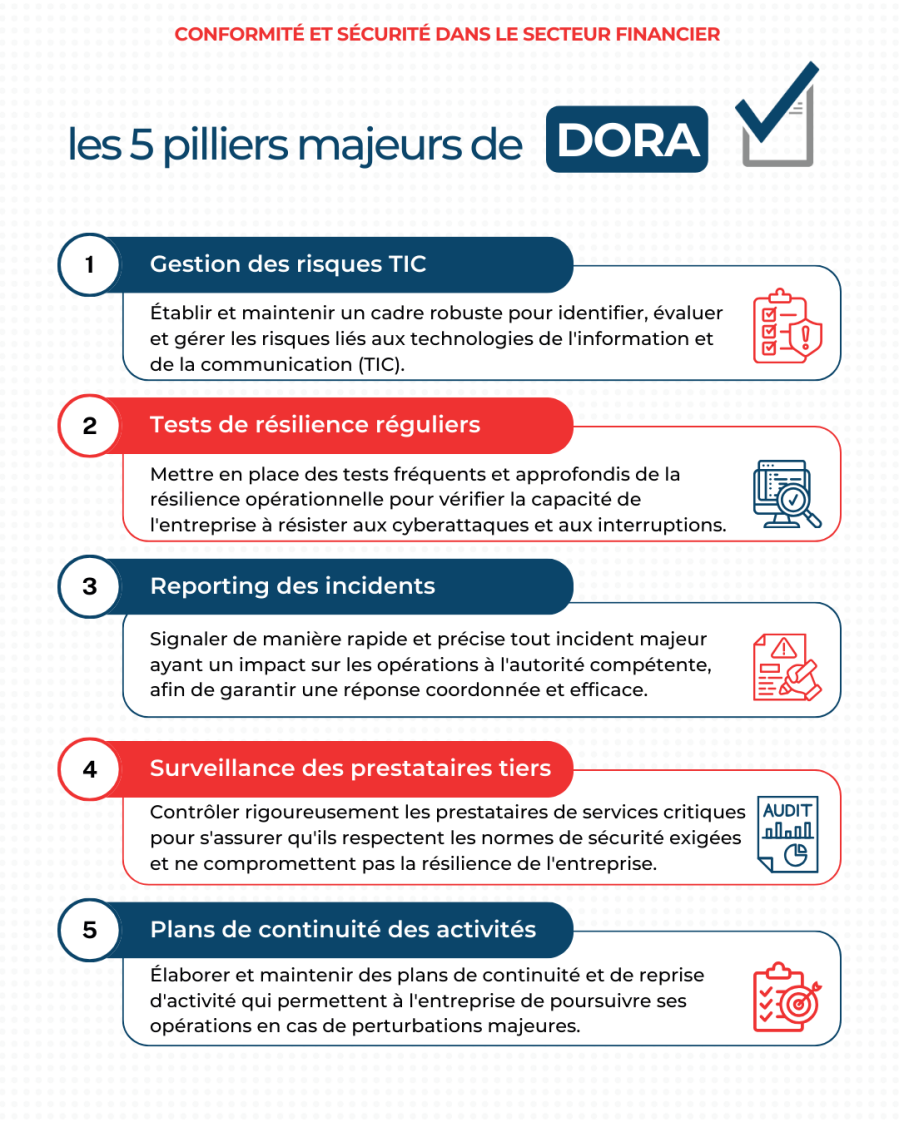
- Assurer la continuité des services financiers malgré des incidents numériques (cyber, logiciels, IA, cloud, etc.)
- Renforcer la sécurité informatique et opérationnelle des institutions financières
- Encadrer les prestataires technologiques critiques (dont les fournisseurs IA)
- Harmoniser les exigences dans toute l'UE

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

Secteur financier : DORA

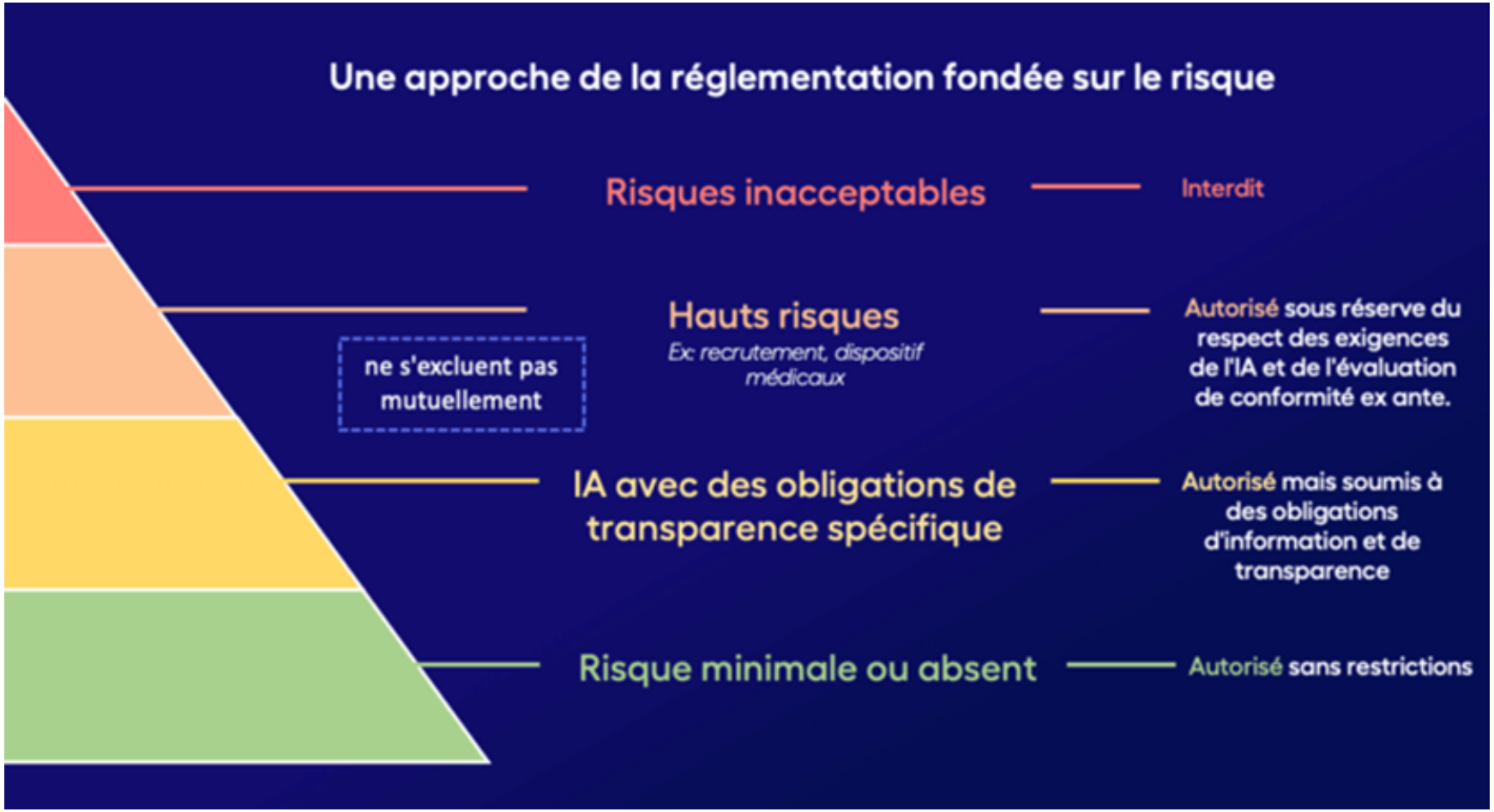
Pas un texte « IA » à la base, mais impacte l'IA sur 4 axes majeurs :

- Risques techniques liés aux modèles
- Dépendance aux fournisseurs IA
- Tests de robustesse des IA utilisées
- Traçabilité des décisions IA



RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

AI Act et ses implications pour les entreprises



RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

AI Act et ses implications pour les entreprises

	FOURNISSEUR	DÉPLOYEUR	MANDATAIRE	IMPORTATEUR	DISTRIBUTEUR
Définition	personne physique ou morale, autorité publique, agence ou tout autre organisme qui		personne physique ou morale située ou établie dans l'Union qui		personne physique ou morale
	développe ou fait développer un SIA ou un GPAI et le met sur le marché ou met le SIA en service sous son propre nom ou sa propre marque, à titre onéreux ou gratuit	Utilise sous sa propre autorité un SIA sauf lorsque ce système est utilisé dans le cadre d'une activité personnelle à caractère non professionnel	a reçu et accepté un mandat écrit d'un fournisseur de SIA ou de GPAI pour s'acquitter en son nom des obligations et des procédures établies par le présent règlement	met sur le marché un SIA qui porte le nom ou la marque d'une personne physique ou morale établie dans un pays tiers	faisant partie de la chaîne d'approvisionnement, autre que le fournisseur ou l'importateur, qui met un SI à disposition sur le marché de l'Union
Risque inacceptable	/	/	/	/	/
Haut-risque	Obligations pour les **fournisseurs** d'IA à haut risque	Obligations pour les **déployeurs** d'IA à haut risque	Obligations pour les **mandataires** d'IA à haut risque	Obligations pour les **importateurs** d'IA à haut risque	Obligations pour les **distributeurs** d'IA à haut risque
Risque limité	Obligations pour les **fournisseurs** d'IA à risque soumis à obligations de transparence	Obligations pour les **déployeurs** d'IA à risque soumis à obligations de transparence	Aucune obligation	Aucune obligation	Aucune obligation
Risque minimal	Code de bonnes pratiques - volontaire	Code de bonnes pratiques - volontaire	Code de bonnes pratiques - volontaire	Code de bonnes pratiques - volontaire	Code de bonnes pratiques - volontaire

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

AI Act – risques inacceptables (1/2)

Typologie	Description
Technique subliminale	Tout système ou partie de système visant influencer subliminalement le public.
Exploitation de personnes vulnérables	Tout système ou partie de système visant influencer une population vulnérable (handicape, âge, situation sociale ou économique) grâce à la dite vulnérabilité.
Catégorisation biométrique	Tout système ou partie de système visant à catégoriser une population en fonction de ses attributs biométriques que cela soit au niveau de sa race, ses opinions politiques ou son appartenance à un syndicats, un courant politique ou religieux.
Score social	Système évaluant des personnes physiques en fonction de leur comportement social, ou de leurs caractéristiques personnelles ou de personnalité connues, déduites ou prédites.
Identification biométrique à distance "en temps réel"	Tout système permettant de suivre, repérer, identifier des personnes physiques dans un espace public et cela en temps réel pour l'administration (sauf lorsque c'est strictement nécessaire).

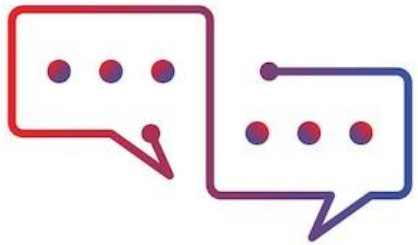
RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

AI Act – risques inacceptables (2/2)

Typologie	Description
Création de bases de données de visage	Systèmes qui créent ou développent des bases de données de reconnaissance faciale par l'extraction non ciblée d'images faciales sur l'internet ou d'images de vidéosurveillance.
Analyse de sentiment	Système qui permet de déduire les émotions d'une personne physique sur le lieu de travail et dans les établissements d'enseignement.
Prédiction d'infraction pénale	Système ayant pour but d'évaluer ou de prédire le risque qu'une personne physique commette une infraction pénale, sur la seule base de son profilage ou de l'évaluation de ses traits de personnalité et de ses caractéristiques.

RÉGLEMENTATION ET NORMES DE CYBERSÉCURITÉ IA

AI Act et ses implications pour les entreprises



Discussion

ATELIER PRATIQUE

Analyse de conformité IA dans un projet en entreprise

PERSPECTIVES 2026

- Les LLM vont de plus en plus être enveloppées dans des agents capables d'appeler des API, d'interagir avec des environnements (OS, navigateurs…) et de réaliser des tâches complexes de manière autonome.

  Projets : LangChain, agents OpenAI, AutoGPT.

- Cela va permettre d'automatiser des processus entiers et accentue les risques d'actions non contrôlées (« sur-autonomie »).

- On peut s'attendre à l'émergence de cadres de sécurité spécifiques pour les agents IA (MAESTRO[1] pour les IA agentiques de la Cloud Security Alliance)

LLM souverains

- Prolifération probable à venir de modèles locaux adaptés aux contextes régionaux
- OpenEuroLLM
- GPT privés
- Plus de contrôle (souveraineté des données, alignement sur des valeurs locales), mais question de leur mise à jour et de leur sécurité.

Sans les moyens d'OpenAI, pourront-ils suivre en qualité et sûreté ?

PERSPECTIVES 2026

Evolutions réglementaires

- Renforcement probable du cadre légal autour de l'IA à venir

Maturation des pratiques DevSecOps IA

- Actuellement, encore beaucoup d'équipes développent des POC à base de LLM sans intégrer pleinement la sécurité dans le cycle.
- En 2026-2027, on peut penser que les pratiques DevSecOps incluront nativement l'IA : intégrer des étapes de sécurité spécifiques aux modèles dans la CI/CD (scans de prompts, analyse des modèles pour biais ou faiblesses avant déploiement, tests de robustesse automatiques).

2026 l'année de l'industrialisation sécurisée de l'IA ?

On y croit ! :D

Question

QUELLE AFFIRMATION DÉCRIT LE MIEUX UNE ATTAQUE DE TYPE PROMPT INJECTION SUR UN MODÈLE GÉNÉRATIF ?

- ❑ L'ATTAQUANT MANIPULE LA REQUÊTE POUR FORCER LE MODÈLE À RÉVÉLER DES INFORMATIONS SENSIBLES OU EXÉCUTER DES INSTRUCTIONS NON PRÉVUES.
- ❑ L'ATTAQUANT INJECTE DU CODE EXÉCUTABLE DANS L'OS HÔTE POUR DÉTOURNER LE MODÈLE.
- ❑ L'ATTAQUANT MODIFIE LES POIDS DU MODÈLE
- ❑ L'ATTAQUANT SUPPRIME PHYSIQUEMENT LES FICHIERS D'ENTRAÎNEMENT DU MODÈLE.

Question

QUELLE AFFIRMATION DÉCRIT LE MIEUX UNE ATTAQUE DE TYPE PROMPT INJECTION SUR UN MODÈLE GÉNÉRATIF ?

- [x] L'ATTAQUANT MANIPULE LA REQUÊTE POUR FORCER LE MODÈLE À RÉVÉLER DES INFORMATIONS SENSIBLES OU EXÉCUTER DES INSTRUCTIONS NON PRÉVUES.
- [ ] ~~L'ATTAQUANT INJECTE DU CODE EXÉCUTABLE DANS L'OS HÔTE POUR DÉTOURNER LE MODÈLE.~~
- [ ] ~~L'ATTAQUANT MODIFIE LES POIDS DU MODÈLE~~
- [ ] ~~L'ATTAQUANT SUPPRIME PHYSIQUEMENT LES FICHIERS D'ENTRAÎNEMENT DU MODÈLE.~~

Question

QUELLE DISTINCTION PRINCIPALE EXISTE ENTRE UN WATERMARKING STATISTIQUE ET SÉMANTIQUE SUR UN MODÈLE GÉNÉRATIF ?

- STATISTIQUE : MODIFICATION DES HYPERPARAMÈTRES DU MODÈLE
  SÉMANTIQUE : MODIFICATION DU VOCABULAIRE DU MODÈLE
- STATISTIQUE : AJOUT DE BRUIT AUX ENTRÉES
  SÉMANTIQUE : SUPPRESSION DES SORTIES À RISQUE
- STATISTIQUE : MODIFICATIONS INVISIBLES DANS LA DISTRIBUTION DES TOKENS
  SÉMANTIQUE : INSERTION D'INFORMATIONS CODÉES DANS LE SENS OU LE STYLE DU TEXTE GÉNÉRÉ
- STATISTIQUE : SIGNATURE DES POIDS DU MODÈLE
  SÉMANTIQUE : CHIFFREMENT DE L'OUTPUT

Question

QUELLE DISTINCTION PRINCIPALE EXISTE ENTRE UN WATERMARKING STATISTIQUE ET SÉMANTIQUE SUR UN MODÈLE GÉNÉRATIF ?

- ~~STATISTIQUE : MODIFICATION DES HYPERPARAMÈTRES DU MODÈLE~~
  ~~SÉMANTIQUE : MODIFICATION DU VOCABULAIRE DU MODÈLE~~

- ~~STATISTIQUE : AJOUT DE BRUIT AUX ENTRÉES~~
  ~~SÉMANTIQUE : SUPPRESSION DES SORTIES À RISQUE~~

- ✓ STATISTIQUE : MODIFICATIONS INVISIBLES DANS LA DISTRIBUTION DES TOKENS
  SÉMANTIQUE : INSERTION D'INFORMATIONS CODÉES DANS LE SENS OU LE STYLE DU TEXTE GÉNÉRÉ

- ~~STATISTIQUE : SIGNATURE DES POIDS DU MODÈLE~~
  ~~SÉMANTIQUE : CHIFFREMENT DE L'OUTPUT~~

Question

DANS LE CONTEXTE DES MODEL CONTEXT PROTOCOLS (MCP), QUEL COMPOSANT EST LE PLUS SOUVENT DÉVELOPPÉ EN INTERNE ET SUSCEPTIBLE DE PRÉSENTER DES VULNÉRABILITÉS ?

Question

DANS LE CONTEXTE DES MODEL CONTEXT PROTOCOLS (MCP), QUEL COMPOSANT EST LE PLUS SOUVENT DÉVELOPPÉ EN INTERNE ET SUSCEPTIBLE DE PRÉSENTER DES VULNÉRABILITÉS ?

- LE SERVEUR MCP
  (ET LES MÉCANISMES DE COMMUNICATION CONTEXTUELLE ENTRE MODÈLES ET CLIENTS)

Question

QUELLE MESURE ILLUSTRE LE MIEUX L'APPROCHE ZERO TRUST APPLIQUÉE À L'IA DANS DES ENVIRONNEMENTS INDUSTRIELS SCADA/IOT ?

- ❑ AUTORISER UNIQUEMENT LES CAPTEURS CONNECTÉS AU RÉSEAU INTERNE
- ❑ AUTHENTIFIER ET VÉRIFIER CONTINUELLEMENT CHAQUE DEVICE ET CHAQUE REQUÊTE AVANT D'EXÉCUTER UNE ACTION SUR UN AUTOMATE
- ❑ ISOLER COMPLÈTEMENT LES SYSTÈMES IA ET SUPPRIMER LES COMMUNICATIONS AVEC LES OPÉRATEURS HUMAINS
- ❑ UTILISER DES ANTIVIRUS SUR LE SERVEUR QUI HÉBERGE LE MODÈLE.

Question

QUELLE MESURE ILLUSTRE LE MIEUX L'APPROCHE ZERO TRUST APPLIQUÉE À L'IA DANS DES ENVIRONNEMENTS INDUSTRIELS SCADA/IOT ?

- ~~AUTORISER UNIQUEMENT LES CAPTEURS CONNECTÉS AU RÉSEAU INTERNE~~
- ✓ AUTHENTIFIER ET VÉRIFIER CONTINUELLEMENT CHAQUE DEVICE ET CHAQUE REQUÊTE AVANT D'EXÉCUTER UNE ACTION SUR UN AUTOMATE
- ~~ISOLER COMPLÈTEMENT LES SYSTÈMES IA ET SUPPRIMER LES COMMUNICATIONS AVEC LES OPÉRATEURS HUMAINS~~
- ~~UTILISER DES ANTIVIRUS SUR LE SERVEUR QUI HÉBERGE LE MODÈLE.~~

Question

L’AI ACT EUROPÉEN IMPOSE AUX ENTREPRISES DE :

- ❑ DÉPLOYER DES MODÈLES OPEN SOURCE POUR DES RAISONS DE RESPONSABILITÉ
- ❑ CLASSER LEURS SYSTÈMES IA SELON LE NIVEAU DE RISQUE ET APPLIQUER DES MESURES PROPORTIONNELLES DE CONFORMITÉ ET DE SÉCURITÉ
- ❑ STOCKER LES DATASETS LOCALEMENT (D’EMBLÉE OU EN BACKUP) POUR DES RAISONS DE RÉSILIENCE
- ❑ NE PAS UTILISER D’IA GÉNÉRATIVE POUR LA DOCUMENTATION INTERNE

Question

L'AI ACT EUROPÉEN IMPOSE AUX ENTREPRISES DE :

- ~~DÉPLOYER DES MODÈLES OPEN SOURCE POUR DES RAISONS DE RESPONSABILITÉ~~
- ✓ CLASSER LEURS SYSTÈMES IA SELON LE NIVEAU DE RISQUE ET APPLIQUER DES MESURES PROPORTIONNELLES DE CONFORMITÉ ET DE SÉCURITÉ
- ~~STOCKER LES DATASETS LOCALEMENT (D'EMBLÉE OU EN BACKUP) POUR DES RAISONS DE RÉSILIENCE~~
- ~~NE PAS UTILISER D'IA GÉNÉRATIVE POUR LA DOCUMENTATION INTERNE~~