# BLACK PAPER — Consolidated 2026

*Ce Black Paper consolidé fusionne deux axes de recherche de terrain de la série Darkpaper en un seul volume navigable : risques d'infrastructure liés au trafic LLM industrialisé (Partie I) et risques identitaires liés à l'usurpation assistée par LLM (Partie II).*

## Table des matières

Table des matières
Le Côté obscur de LLM : Risques réseau, sécurité, stockage, cognitifs et écologiques pour les services informatiques mondiaux Observations sur le terrain en 2026 à partir d'une flotte de production de proxy inverse de l'UE
1. Présentation
2. Écosystème d'IA Crawler documenté
Robots d'exploration d'IA documentés (sources publiques)
3. Chronologie de la courbe de pression de l'infrastructure LLM
4. Domaine de risque 1 – Infrastructure de réseau et de proxy inverse
4.1 Changement de composition du trafic
Tendance observée dans la composition du trafic (directionnelle, non absolue)
4.2 Caractéristiques comportementales distinctes des robots classiques
4.3 Implications en matière de sécurité
4.4 Télémétrie de terrain multi-sites (flotte BunkerWeb à 4 nœuds)
Composition du trafic par site (flotte BunkerWeb, fenêtre de 63 jours)
5. Domaine de risque 2 – Infrastructure de stockage, synchronisation cloud et systèmes de sauvegarde
5.1 Le cycle de vie des artefacts LLM
5.2 Le facteur de multiplication dans les services de synchronisation cloud
Multiplication des risques de stockage par niveau
5.3 Exposition réglementaire
6. Risque aggravé : quand les risques liés au trafic et au stockage interagissent
Matrice de composition des risques
8. Domaine de risque 4 – Empreinte écologique et rareté des ressources
8.1 Consommation globale d'énergie
8.2 La cascade de pénurie de GPU et de semi-conducteurs
8.3 Le transfert des coûts publics et institutionnels
10. Domaine de risque 6 – Saturation des données synthétiques et dilution du signal
10.1 Ingestion récursive et effondrement du modèle
10.2 Contamination de la base de connaissances de l'entreprise
11. Domaine de risque 7 – Démocratisation des capacités automatisées de menace
11.1 Barrières d'entrée abaissées pour les attaques complexes
11.2 Évolution de l'économie défensive
12. Domaine de risque 8 - Intermédiation du trafic et consolidation de l'hébergement Web
12.1 L'effet d'intermédiation du trafic
12.2 Viabilité des éditeurs de niveau intermédiaire et consolidation de l'infrastructure
12.3 Rentabilité des plateformes cloud et subventions croisées
13. Domaine de risque 9 - MCP et les Hackteurs : escalade de privilèges sémantique
13.1 Le « Bring Your Own Agent » (BYOA) : Le Cabotin Free-lance en Coulisses
13.2 L'Angle Mort de la Gestion des Accès à Privilèges (PAM)
13.3 Le Livret des Coulisses : Cartographier Jira et la Mémoire des Régisseurs
13.4 Le Metteur en Scène Distant et l'Improvisation Fatale
13.5 Le Poison des Modèles : Weight Poisoning
13.6 Les Extensions Vérolées : le Décor Empoisonné
13.7 Les LLM Trafiqués : la Troupe Rachetée en Coulisses
13.8 La Couche NPX et NPM : les Machinistes Invisibles
13.9 Le Génome Sémantique : soul.md et la gouvernance du rôle
13.10 La Charge des Artéfacts : la Réserve d'Accessoires devenue Décharge
13.11 Conclusion: Déployer un "Zero Trust" Sémantique
13.12 Synthèse: Matrice des Risques et Contraintes
13bis. Domaine de risque 10 — Érosion du substrat identitaire et personas synthétiques industrialisés par les LLM
13bis.1 Surface cumulative des fuites (France, 2024–2026)
13bis.2 Confirmation internationale du schéma
13bis.3 Vecteurs d'attaque LLM sur le substrat identitaire
13bis.4 La mort pratique de l'authentification par la connaissance (KBA)
13bis.5 Conséquences infrastructurelles pour les fournisseurs d'identité
13bis.6 Pression future : migration identitaire post-quantique

13bis.7 Recommandations spécifiques au domaine 10
13bis.8 Synthèse : matrice de risque et de contrainte (substrat identitaire)
14. Recommandations fondées sur des données probantes
15. Questions ouvertes et lacunes en matière de recherche
7. Domaine de risque 3 – Asymétrie cognitive et épuisement humain
7.1 Pathologie du monde réel : épuisement de la télémétrie PrestaShop
9. Domaine de risque 5 – Accessibilité cognitive et interaction des utilisateurs vulnérables
9.1 Dynamique sociale asymétrique
9.2 Dépendance épistémique excessive et validation motivée
9.3 Santé non structurée et interactions quasi-thérapeutiques
9.4 Boucles d'engagement persistantes et fonctionnement exécutif
7.2 Une génération qui grandit au sein d'une expérience incontrôlée
7.2.1 Ce que nous disent réellement les recherches sur les écrans et sur Internet
7.2.2 Pourquoi l'exposition à l'ère LLM est qualitativement différente
7.2.3 Le risque au niveau des infrastructures
16.Conclusion
Annexe A. Télémétrie de terrain vérifiée (site anonymisé)
Résultats globaux vérifiés (Site A)
A.1 Lecture opérationnelle
A.2 Note méthodologique
Annexe B. Vérification de la découvrabilité reproductible par le lecteur
B.1 Script d'audit
B.2 Interpretation guide
Références
🎯 TL;DR — Avertissement stratégique
🔗 Pipeline de l'identité synthétique — schéma d'attaque
0. Le paradoxe de l'ANTS
1. La fourmilière piétinée — état des lieux des fuites françaises
1.1 Les grandes brèches françaises (2023–2025)
1.2 Le contexte international — la matière première mondiale (2024–mai 2026)
2. Le LLM comme alchimiste de l'identité volée
2.1 L'hybridation des fourmis — la synthèse identitaire à la volée
2.2 La phéromone vocale : cloner la voix pour contourner la KYC
2.2b Le moulage des mandibules — l'empreinte digitale passive : extraire la biométrie des photos publiques
2.3 Le deepfake documentaire : forger le titre d'identité physique
2.4 Les phéromones empoisonnées — l'hameçonnage chirurgical : la fin du spam générique
3. L'identité synthétique : la fourmi fantôme
4. L'ANTS dans la tempête — ce que l'agence ne peut pas faire seule
5. Le coup de pied dans la fourmilière — ce qui change avec les LLM agentiques (bilan 2024-mai 2026)
6. Ce que ça change pour la défense
6.1 Ce qui ne suffit plus
6.2 Ce qui fonctionne encore
6.3 Le principe de défense par la non-clonaison
6.4 Que faire demain matin — outils disponibles aujourd'hui
7. Conclusion — réinventer un système identitaire qui nous ressemble
Sortir du « crier au loup »

# Partie I — Risques d'infrastructure, cognitifs et écologiques

LIVRE NOIR - Série sur les risques liés aux infrastructures *Homologue du Livre blanc - en projet*




Papier noir Recherche sur les risques liés aux infrastructures Analyse globale mai 2026

## Le *Côté obscur* de LLM : Risques réseau, sécurité, stockage, cognitifs et écologiques pour les services informatiques mondiaux Observations sur le terrain en 2026 à partir d'une flotte de production de proxy inverse de l'UE

AUTEUR PRINCIPAL Bryce SIMON

CO-AUTEUR Ifrit (IA)

📄 Télécharger le PDF 📚 Télécharger EPUB 📁 Archives

Langue :

Livre noir · Recherche et analyse | Évaluation des risques fondée sur des faits · Livre blanc : en projet

Ce que ne disent pas les communiqués de presse de l'industrie. Une analyse documentée et fondée sur la source de la façon dont les charges de travail importantes des modèles de langage - exploration, inférence, conservation des données, agents autonomes - dégradent l'infrastructure Web, saturent les systèmes de stockage, épuisent les opérateurs humains, déplacent les ressources publiques et mènent une expérience cognitive générationnelle sans groupe de contrôle.

**Type de document :** Livre noir (perspective du risque contradictoire) · **Sources :** Littérature primaire, rapports de l'industrie publique, observations sur le terrain uniquement - aucune affirmation extrapolée sans qualification explicite · **Accès libre :** CC PAR 4.0

**Résumé - Impacts systémiques non reconnus des LLM**

Ce document analyse les impacts observables sur l'infrastructure, les opérations et l'environnement des charges de travail des grands modèles de langage (LLM) à grande échelle. S'appuyant sur des rapports industriels, des cadres de conformité et des données de télémétrie sur le terrain, il examine comment l'industrialisation des LLM introduit des complexités opérationnelles spécifiques dans huit domaines de risque interdépendants :

- L'augmentation mesurable du trafic automatisé et sémantiquement motivé dégradant les proxys inverses et les performances du WAF.
- La consommation accélérée des systèmes de stockage et de sauvegarde d'entreprise par des artefacts générés en continu par LLM.
- La pression cognitive aiguë sur les opérateurs informatiques résultant de la fatigue des alertes, de la saturation de la télémétrie et des demandes automatisées à volume élevé.
- L'empreinte écologique et le déplacement de la chaîne d'approvisionnement en matériel causé par les demandes de calcul de l'IA.
- L'accessibilité cognitive et les risques pour la santé mentale pour les utilisateurs finaux vulnérables naviguant dans des systèmes d'IA conversationnelle non réglementés.
- Saturation des données synthétiques et dilution des signaux dans les bases de connaissances ouvertes du Web et de l'entreprise.
- La démocratisation des capacités automatisées de menace, réduisant les obstacles aux cyberattaques sophistiquées.
- Intermédiation du trafic et consolidation de l'économie de l'hébergement Web qui en résulte.

## 1. Présentation

> **Portée de ce document**
>
> Ce document se concentre sur les exigences d'infrastructure, les frais généraux opérationnels et les obligations de conformité associées aux déploiements de modèles de langage à grande échelle. Il est destiné à servir de pendant en matière de gestion des risques à la littérature industrielle standard axée sur les capacités, en présentant une évaluation clinique et objective des externalités systémiques. Le livre blanc complémentaire (actuellement en ébauche) abordera les opportunités, les atténuations et les voies constructives.

Jusqu'en 2022 environ, le profil de trafic d'un service Web accessible au public était raisonnablement stable : un mélange de sessions humaines (navigateurs), de robots d'exploration de moteurs de recherche connus (Googlebot, Bingbot, Yandexbot) et d'une couche résiduelle d'outils automatisés (scanners de sécurité, moniteurs de disponibilité, auditeurs SEO). Le rapport entre le trafic humain et le trafic automatisé était déjà en hausse pour le trafic automatisé - Rapport annuel d'Imperva *Rapport sur les robots malveillants* a suivi le trafic automatisé de manière constante au-dessus de 40 % de tout le trafic Internet depuis 2021 - mais le *nature* L'automatisation était familière et largement gérable avec des défenses standard basées sur des règles.

Depuis 2023, une classe qualitativement différente de trafic automatisé a émergé : des robots d'exploration de formation LLM, des agents de récupération par temps d'inférence, des indexeurs sémantiques pour les produits de recherche d'IA et des agents autonomes en plusieurs étapes exécutant des chaînes de requêtes HTTP itératives. Ces systèmes partagent des caractéristiques qui diffèrent fondamentalement de l'automatisation classique :

- Ils sont **motivé sémantiquement** plutôt que structurellement motivés - ils recherchent la qualité du contenu, pas seulement la découverte d'URL.
- Ils sont **épisodique et imprévisible** dans le timing, avec des modèles de rafales contrairement aux robots d'exploration SEO programmés.
- Ils génèrent **artefacts persistants en aval** (intégrations, résumés, entrées d'index, réponses mises en cache) qui consomment du stockage à chaque étape du pipeline.
- Ils opèrent à **échelle industrielle**: une seule exécution de formation LLM peut consommer des pétaoctets extraits de milliards de pages explorées.

La conséquence est une pression croissante sur l'infrastructure qui se manifeste différemment selon la couche de la pile. Cet article analyse à la fois la couche réseau/proxy et la couche de stockage, car elles sont causalement liées : le trafic génère des journaux, les journaux consomment du stockage, le stockage est sauvegardé, les sauvegardes augmentent.

## 2. Écosystème d'IA Crawler documenté

Le tableau suivant répertorie les robots d'exploration associés à l'IA qui ont publié des `robots.txt` documentation ou divulgations techniques à partir de 2025. Il ne s'agit pas d'une énumération exhaustive - de nombreux grattoirs non divulgués sont connus des chercheurs mais ne peuvent pas être publiquement attribués.

### Robots d'exploration d'IA documentés (sources publiques)

| Agent utilisateur | Opérateur | But | Documenté depuis |
|---|---|---|---|
| `GPTBot` | OpenAI | Données de formation / récupération Web | août 2023 |
| `ChatGPT-User` | OpenAI | Navigation en temps réel (temps d'inférence) | août 2023 |
| `OAI-SearchBot` | OpenAI | Index de recherche pour la recherche ChatGPT | 2024 |
| `ClaudeBot` | Anthropique | Données d'entraînement/récupération | 2023 |
| `Claude-Web` | Anthropique | Accès Web au moment de l'inférence | 2024 |
| `Google-Extended` | Google DeepMind | Signal de désinscription à la formation (exploration inversée) | septembre 2023 |
| `Bytespider` | ByteDance | Données de formation (produits TikTok AI) | 2023 |
| `CCBot` | Fondation Common Crawl | Corpus Web ouvert (utilisé dans de nombreuses formations LLM) | Avant 2020, mais l'utilisation a augmenté entre 2022 et 2024 |
| `Diffbot` | Diffbot | Graphe de connaissances / extraction de données structurées | Avant 2020, l'utilisation du LLM a augmenté jusqu'en 2023 |
| `Applebot-Extended` | Pomme | Désinscription à la formation Apple Intelligence | 2024 |
| `PetalBot` | Huawei | Index Web / Produits IA | 2020-2023 |

Sources : OpenAI Docs (août 2023), documentation Anthropic, Google Search Central, Common Crawl Foundation, Apple Support - tous vérifiables publiquement à partir de 2025.

Il est important de noter que les robots d'exploration répertoriés ci-dessus sont les *déclaré* ceux. Les données radar publiques de Cloudflare, publiées tout au long de l'année 2024, ont identifié une queue beaucoup plus grande d'agents non déclarés ou usurpés effectuant des comportements d'exploration sémantiquement similaires, attribuant une fraction significative à une infrastructure adjacente à l'IA fonctionnant sans `robots.txt` conformité.

## 3. Chronologie de la courbe de pression de l'infrastructure LLM

2019 - 2021

Le corpus Common Crawl atteint environ 250 To par cycle d'analyse. GPT-3 publié (juin 2020) utilise Common Crawl comme principale source de formation. L'impact sur l'infrastructure est limité à l'opérateur du robot d'exploration Common Crawl et aux premières équipes de recherche en PNL.

novembre 2022

ChatGPT est lancé publiquement. La demande de données de formation s'industrialise. Le volume de téléchargements de Common Crawl par des tiers augmente. Le stockage compatible S3 hébergeant des données Common Crawl (principalement Amazon S3) connaît une sortie accrue.

Début 2023

Première vague de concours de mannequins : Anthropic Claude, Google Bard, Meta LLaMA. Chacun nécessite une collection de corpus Web indépendante. Les opérateurs Cloudflare et Akamai commencent à signaler le trafic anormal des robots sur les tableaux de bord des clients.

août 2023

OpenAI documente officiellement `GPTBot` et `ChatGPT-User`. Il s'agit de la première reconnaissance publique par une grande société d'IA d'un robot d'exploration Web dédié. Google suit avec le `Google-Extended` mécanisme, indiquant qu'une exploration de formation à grande échelle est déjà opérationnelle.

2024

Cloudflare publie une analyse confirmant que les robots IA sont responsables d'une part disproportionnée des demandes par rapport à leur objectif déclaré. Les frameworks d'agents d'IA (LangChain, AutoGen, CrewAI) prolifèrent, permettant une interaction HTTP multi-tours programmatique à faible coût de développement. Les équipes de stockage d'entreprise commencent à signaler une croissance inattendue des artefacts de fichiers liés à l'IA.

avril 2024

La loi européenne sur l'IA entre en vigueur. Les obligations concernant la provenance des données de formation, la documentation des systèmes à haut risque et la minimisation des données commencent à créer des exigences de conformité pour le stockage des artefacts liés à l'IA.

2025 - présent

Les charges de travail des agents IA deviennent monnaie courante dans les outils d'entreprise. MCP (Model Context Protocol, Anthropic/OpenAI), les API d'appel de fonctions et les agents de contrôle du navigateur créent un nouveau niveau de trafic HTTP automatisé impossible à distinguer des sessions humaines sans analyse comportementale. Les systèmes de stockage accumulent des artefacts de formation, des index vectoriels, des journaux de conversation et des sorties multimodales à des rythmes jamais vus auparavant.

# 4. Domaine de risque 1 – Infrastructure de réseau et de proxy inverse

## 4.1 Changement de composition du trafic

Imperva *Rapport sur les mauvais robots 2024* (publié en avril 2024) indique que le trafic de robots malveillants et automatisés a atteint 49,6 % de tout le trafic Internet en 2023, la part la plus élevée depuis qu'Imperva a commencé à mesurer en 2013. Bien que tout le trafic automatisé ne soit pas lié à l'IA, le rapport identifie les robots d'exploration spécifiques à l'IA comme une sous-catégorie nouvellement dominante et en croissance. Les données Cloudflare Radar de 2024 montrent des taux d'exploration élevés et persistants affectant particulièrement les domaines des médias, de l'éducation et du commerce électronique - précisément les catégories de contenu ayant la valeur de formation LLM la plus élevée.

### Tendance observée dans la composition du trafic (directionnelle, non absolue)

Trafic humainpart décroissante

Bots SEO classiquesécurie

Robots d'exploration IA (déclarés)soulèvement

Adjacent à l'IA non déclaréen hausse, difficile à quantifier

Scanners de sécurité / abusconstamment élevé

Représentation directionnelle basée sur le rapport Imperva Bad Bot 2024, l'analyse publique Cloudflare Radar 2024 et l'état de l'Internet d'Akamai 2024. Les largeurs relatives des barres sont proportionnelles à la tendance directionnelle, et non à la part absolue.

## 4.2 Caractéristiques comportementales distinctes des robots classiques

Les robots LLM et les agents IA présentent une signature comportementale distincte au niveau de la couche proxy qui complique les défenses standard :

- **Haute densité de requêtes sur des pages sémantiquement précieuses :** Contrairement aux robots SEO qui explorent d'abord l'intégralité du plan du site, les robots d'exploration IA se concentrent fortement sur le contenu long, la documentation de l'API, les descriptions de produits et le contenu généré par les utilisateurs - des pages qui supportent une charge de serveur plus élevée (jointures de bases de données, requêtes de recherche, personnalisation).
- **Modèles de rafales liés aux programmes d'entraînement :** Plutôt qu'une exploration en régime permanent, les robots d'entraînement d'IA peuvent présenter des périodes de rafale intenses suivies de silence, créant des pics de charge imprévisibles qui invalident la planification de la capacité basée sur le débit moyen.
- **Usurpation d'agent utilisateur documentée à grande échelle :** Plusieurs équipes de recherche en sécurité (dont Datadome et Bright Data ont divulgué des exemples en 2024) ont documenté des robots d'exploration d'IA et des grattoirs dérivés en aval parcourant les chaînes d'agent utilisateur du navigateur pour éviter la détection, rendant les défenses basées sur les signatures insuffisantes. La télémétrie 2026 sur le Site A (§4.4.4) confirme un durcissement : 597 851 requêtes — 6,27 % du trafic total sur 22 jours — ont été émises avec un UA strictement identique à celui d'un utilisateur humain Chrome 131 / macOS, sans chaîne `+http://` ni aucune signature accessible aux WAF qui filtrent par UA seul.
- **Mimétisme actif du `Referer` interne (nouveauté 2026) :** La télémétrie du Site A (§4.4.4) montre que les sept premiers referers globaux sont tous des pages produits internes du site lui-même, émises presque exclusivement par ClaudeBot. Le crawler falsifie systématiquement le `Referer` pour pointer vers une autre page produit du même site, simulant une navigation interne. Les signatures défensives fondées sur la présence d'un referer ou sur la confiance dans un referer interne deviennent inopérantes.
- **La récupération du temps d'inférence est toujours active :** Des produits tels que ChatGPT Search, Perplexity et Copilot avec intégration Bing effectuent une récupération Web en temps réel pendant les sessions des utilisateurs finaux. Ce trafic ne suit pas de calendrier de formation : il évolue directement avec l'adoption par les utilisateurs des produits d'IA.

## 4.3 Implications en matière de sécurité

Au-delà de la capacité, le changement de composition du trafic crée des risques de sécurité distincts :

- **Lassitude des règles WAF :** Les organisations étendent les ensembles de règles WAF pour bloquer les robots d'exploration d'IA non déclarés tout en permettant aux utilisateurs légitimes de créer des arborescences de règles de plus en plus complexes. Le risque de mauvaise configuration augmente proportionnellement. Des limites de débit mal définies ciblant le comportement des robots d'exploration peuvent dégrader par inadvertance le service destiné aux utilisateurs légitimes de l'infrastructure partagée (universités, proxys d'entreprise, VPN).
- **Exfiltration de données via des explorations d'apparence légitime :** Des contenus jusqu'alors peu intéressants à extraire systématiquement (termes et conditions, tableaux de tarifs, documentation interne indexée par les moteurs de recherche) deviennent un signal de formation à l'IA de grande valeur. Le modèle de menace pour les données commerciales accessibles sur le Web doit être mis à jour pour inclure l'exfiltration passive par des robots d'exploration en formation.
- **Injection rapide via le contenu :** Alors que les agents d'IA parcourent le contenu Web en direct dans le cadre de leur pipeline de raisonnement, l'injection d'instructions dans des pages Web explorables (une technique documentée par des chercheurs universitaires de plusieurs institutions en 2023-2024) peut influencer le comportement du système d'IA – créant une nouvelle surface d'attaque qui prend son origine au niveau de la couche de contenu mais affecte l'intégrité du système d'IA en aval.
- **Log et saturation SIEM :** Les volumes élevés des robots d'exploration augmentent directement le volume des journaux. Les systèmes conçus pour les taux d'enregistrement du trafic humain peuvent atteindre des seuils de capacité, provoquant des lacunes dans les journaux ou une augmentation des coûts de traitement SIEM, ce qui dégrade la capacité de détection des incidents.

### 4.4 Télémétrie de terrain multi-sites (flotte BunkerWeb à 4 nœuds)

Pour éviter la faiblesse inférentielle de la généralisation à instance unique, les preuves de terrain ci-dessous sont tirées d'une flotte BunkerWeb à quatre nœuds protégeant **88 hôtes virtuels distincts** à travers un **Fenêtre d'observation de 63 jours** (14 mars 2026 – 16 mai 2026). Trois nœuds étaient en ligne au moment de la récolte ; l'un d'entre eux était hors ligne et est exclu des agrégats. Total traité : **889 552 demandes** depuis **20 683 adresses IP sources uniques**, classés via des familles déterministes d'agents utilisateurs et validés par rapport aux distributions de codes d'état et d'hôtes.

Le résultat le plus utile n'est pas la moyenne de la flotte, mais le **hétérogénéité par site**. La part du trafic de formation à l'IA varie de plus de deux ordres de grandeur sur les trois nœuds en ligne, en fonction conjointement du profil de contenu servi par chaque nœud. *et* sur sa position en matière de découvrabilité publique. Cela invalide tout cadre qui traite la pression des robots d'exploration de l'IA comme une taxe uniforme sur les infrastructures ; il dépend du contenu et de la découvrabilité, et les deux dépendances sont fortes.

> **Note méthodologique sur la sélection des sites et les confusions.**
>
> Les trois sites diffèrent selon deux axes confus que la flotte ne peut clairement séparer : *profil de contenu* (Site A = catalogue e-commerce haute densité ; Site B = mixte Git/docs ; Site C = services personnels auto-hébergés) et *posture de découvrabilité publique* (Le site A fait l'objet d'une promotion active : gestion du référencement, dépenses publicitaires, soumission du plan du site, campagnes de liens entrants ; les sites B et C sont techniquement indexables, sans `robots.txt` blocs et aucun refus d'IA-crawler, mais ne sont pas activement promus - pas de soumission, pas de publicité, présence limitée de liens entrants naturels). La part de formation de l'IA par site reflète donc les deux filtres agissant en séquence : la découvrabilité d'abord (si le graphique de départ du robot atteint le site), ensuite l'attracteur de contenu (la pression de revisite qui s'ensuit une fois qu'il y parvient). L'implication opérationnelle - que **la posture promotionnelle est elle-même une surface de contrôle réglable, distincte du blocage technique du robot** — est indépendant de la lecture d'hétérogénéité et est repris au §13. Un script d'audit reproductible par le lecteur pour caractériser la situation de découvrabilité de toute propriété est fourni à l'annexe B.

#### Composition du trafic par site (flotte BunkerWeb, fenêtre de 63 jours)

[Figure: Composition du trafic par site (flotte BunkerWeb, fenêtre de 63 jours) — see HTML/PDF for visual rendering. Source : Récolte de la flotte BunkerWeb data_20260516_232129 (schéma bw.harvest.v3 ). Les barres sont proportionnelles, normalisées au total des demandes de chaque site. Les classifications utilisent la détection déterministe des familles UA ; « UA inconnu » = en-tête vide ou non reconnu.]

Source : Récolte de la flotte BunkerWeb `data_20260516_232129` (schéma `bw.harvest.v3` ). Les barres sont proportionnelles, normalisées au total des demandes de chaque site. Les classifications utilisent la détection déterministe des familles UA ; « UA inconnu » = en-tête vide ou non reconnu.

##### 4.4.1 Ensemble de la flotte (fenêtre de 63 jours)

Partage de formation en IA

54,6%

485 419 / 889 552 demandes

Taux de blocage WAF

20,6%

183 263 / 889 552 bloqués

Part humaine

19,6%

174 027 sessions sur 88 hôtes virtuels

Deux familles de robots représentent **96,7 % de tout le trafic de formation à l'IA** observés sur la flotte : Meta's `meta-externalagent` (303 756 demandes, 62,6% des formations IA) et Anthropic's `ClaudeBot` (165 705, 34,1 %). Bytespider, Amazonbot et Applebot représentent ensemble les 3 % restants. Cette concentration a des conséquences opérationnelles : un petit nombre d'agents utilisateurs identifiables et d'ASN d'origine génèrent l'essentiel de la charge d'infrastructure attribuable à l'IA, ce qui rend l'atténuation au niveau politique (limite de débit sélective, application du fichier robots.txt, négociation de licence de contenu) traitable en principe.

##### 4.4.2 Ce qui est valable sur tous les sites, ce qui ne l'est pas

- **Disponible sur les trois sites :** le taux de blocage du WAF n'est pas trivial partout (29,2 % sur le site A, 4,5 % sur le site B, 3,3 % sur le site C — le niveau absolu évolue avec la pression adverse, mais le plancher n'est pas nul). Le code d'état 403 est systématiquement la deuxième réponse la plus fréquente après 200.
- **Conservé là où le contenu est sémantiquement dense et détectable (Site A principalement, Site B partiellement) :** Traversée de chemin profond ciblant les listes de catégories, les pages de documentation ou les arborescences de fichiers git-repo ; rafales provoquées par des campagnes de robots d'exploration individuelles plutôt que par une activité stable.
- **Ne généralise pas – et la raison compte :** la part de 80 % de formation en IA sur le site A est *pas* l'expérience typique d'une flotte, mais le 0,5 % du site C ne constitue pas non plus une conclusion selon laquelle « les applications personnelles sont sûres ». Le site C est techniquement indexable ; ce qui lui manque, c'est une promotion active (pas de soumission de plan de site, pas de publicité, liens entrants limités). L'absence de

pression sur l'IA-formation reflète donc *non-découverte organique*, pas de désabonnement technique et pas de manque d'attrait du contenu. Quel site C *fait* voir – 49 % de trafic « UA inconnu » provenant de l'automatisation interne, des clients d'applications mobiles et des sondes non affiliées – est le profil de menace qui survit même lorsque les robots d'exploration IA ne peuvent pas vous trouver de manière organique. L'implication: **la posture promotionnelle est un levier réglable distinct de `robots.txt`**, et les deux composent de manière multiplicative.

#### 4.4.3 Le cas PrestaShop comme un seul point de données

La charge de travail du Site A (PrestaShop 9.0.3, ~11 000 produits répartis dans 118 catégories actives) est présentée comme un site unique au sein de la flotte plutôt que comme une cible de généralisation. Au sein du Site A spécifiquement : une IP source unique (216.73.216.180, ClaudeBot) générée **165 356 demandes en 63 jours** — 28,2 % du total du site A — concentré sur une traversée approfondie des catégories qui contournait la mise en cache périphérique grâce à la génération dynamique de pages de PrestaShop via les jointures de bases de données. Les journaux BunkerWeb affichent les pics correspondants dans les files d'attente de validation des demandes, les déclencheurs Fail2Ban et les décisions CrowdSec ciblant les chemins de catalogue. Le même mécanisme s'appliquerait à n'importe quel catalogue rendu dynamiquement (Magento, WooCommerce, Sylius, Shopware), mais l'ampleur observée sur le site A ne doit pas être lue comme une attente de base - il s'agit d'une illustration de la limite supérieure tirée d'un profil de contenu qui est, de toute évidence, un attracteur de formation pour l'IA. *et* qui est activement promu dans les index à partir desquels ces attracteurs explorent.

L'incidence inter-organisationnelle et intra-organisationnelle est accentuée par la vision de la flotte : sur les sites A et B, le supporteur de coûts (l'opérateur) et le générateur de charge (le robot d'exploration) sont des entités distinctes – une externalité tierce. Sur le site C, presque tout le trafic provient du périmètre propre de l'opérateur – un compromis de productivité intra-organisationnel. Les deux profils nécessitent des stratégies d'atténuation différentes et des hypothèses de gouvernance différentes. Voir §15 pour l'implication opérationnelle.

#### 4.4.4 Approfondissement Site A : fenêtre courte, signaux nouveaux

Une seconde collecte ciblée sur le Site A (PrestaShop 9, fenêtre de 22 jours du 26 avril au 18 mai 2026, 9 539 487 requêtes capturées au niveau du reverse proxy après filtrage BunkerWeb amont) confirme et étend les observations du §4.4.3 sur trois axes qui n'apparaissent pas dans la vue agrégée de la flotte 63 jours.

**(a) Concentration de la charge sur un seul *front controller*.** 90,7 % des requêtes (8 654 284 sur 9 539 487) ciblent un chemin unique : `/index.php`. Ce comportement n'est pas spécifique aux crawlers IA – il découle de l'architecture PrestaShop, où chaque page de catégorie, fiche produit et résultat de recherche passe par le même *front controller* avec des paramètres de query string variables. La conséquence opérationnelle est que la mise en cache standard (par URL) est structurellement inefficace, et que chaque requête mobilise un worker PHP-FPM pour un cycle complet (parse, jointure base, rendu Smarty). Le coût marginal d'une requête supplémentaire est donc plat et proche du coût moyen, ce qui élimine tout effet d'amortissement lié au volume.

**(b) Tactique d'évasion par mimétisme de referer.** Sur les 9 590 665 lignes analysées, 21,76 % portent un referer HTTP non vide (2 087 189 lignes). La distribution par origine est asymétrique : bots IA = 1 452 127 ; humains = 504 567 ; moteurs de recherche = 100 840. Les bots IA émettent près de trois fois plus de referers que tous les humains réunis. Plus surprenant, les sept premiers referers du top global (entre 49 949 et 119 609 occurrences chacun) sont tous des URLs internes au Site A (fiches produits avec `id_product` et `id_product_attribute`), émises presque exclusivement par ClaudeBot. Le crawler injecte donc, à chaque requête, un referer pointant vers une autre fiche produit du même site, simulant une navigation interne entre fiches. Cette tactique neutralise deux familles de règles défensives courantes : (1) les filtres WAF qui marquent comme suspectes les requêtes sans referer, (2) les politiques de cache qui considèrent comme légitime tout trafic dont le referer est une page interne valide. Conséquence pratique : une signature défensive fondée sur le champ `Referer` seul ne sépare plus le trafic d'apprentissage du trafic humain sur les sites e-commerce. Il s'agit d'un changement de régime par rapport aux observations 2024–2025, où la majorité des crawlers IA laissaient le referer vide.

**(c) Silence applicatif côté CMS.** L'inspection directe du conteneur PrestaShop du Site A confirme l'absence totale de défense applicative contre les crawlers IA : 88 modules installés dont zéro lié à l'IA (pas d'intégration GPT, Claude, recommandation LLM ou équivalent) ; `robots.txt` au format template PrestaShop par défaut, avec un seul bloc `User-agent: *` générique et **aucune directive `Disallow` ciblant `ClaudeBot`, `GPTBot`, `Meta-ExternalAgent`, `PerplexityBot`, `Amazonbot`, `Bytespider` ou `CCBot`** ; `.htaccess` sans aucune règle `RewriteCond %{HTTP_USER_AGENT}` filtrant les crawlers. Cette configuration est représentative des installations PrestaShop hors-équipe-technique : le silence du `robots.txt` n'est ni un consentement ni un refus, c'est un défaut d'usine. Les éditeurs de LLM opèrent dans cette zone grise, sans notification préalable au marchand. La conséquence en termes de gouvernance est que la posture par défaut d'un CMS e-commerce grand public revient à un consentement implicite à l'entraînement, contraire à l'esprit de l'opt-in qui prévaut dans les autres régimes de propriété intellectuelle.

**(d) Camouflage UA observé pour Perplexity.** 597 851 requêtes (6,27 % du trafic total du Site A) ont été classifiées `ai_assistant` sur la base d'un seul UA `Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/131.0...`, strictement identique à celui d'un utilisateur humain sur Chrome 131 / macOS. Aucune chaîne `+http://`, aucun nom de bot, aucune signature distinctive ne permet à un filtre standard de détecter le bot. La classification a été obtenue par recoupement des plages IP et patterns de cadence ; ce signal n'est pas accessible aux WAF qui filtrent par UA seul. Combiné à (b), cela signifie qu'une part substantielle du trafic d'apprentissage est désormais structurellement invisible aux dispositifs défensifs hérités.

**(e) Code d'état : signature de saturation PHP-FPM.** Distribution observée : 200 = 7 114 502 ; 403 = 1 184 997 (BunkerWeb) ; 302 = 1 092 690 ; 503 = 7 794 ; 504 = 3 264 ; 502 = 726 ; 500 = 296. Total 5xx = 12 080. La signature 503/504 (worker indisponible / timeout gateway) est cohérente avec une saturation périodique du pool PHP-FPM lors des rafales de crawling – les requêtes humaines reçoivent une erreur quand les workers sont occupés par des crawlers concurrents. Sur l'hôte applicatif (Debian 12, Intel i9-9900K 12 threads, 16 Go RAM, ~5 jours d'uptime au moment de la mesure), la load average s'établissait à 3,36 / 3,12 / 2,41 : pas de saturation continue, mais des pics suffisants pour expliquer les 12 080 erreurs serveur côté client.

**(f) Écosystème humain réel.** Quand on isole les referers d'origine humaine, l'écosystème est limité : navigation interne ≈ 475 000 ; `https://www.google.com/` 7 350 ; un forum francophone Blood Bowl ≈ 2 330 ; application Facebook (`fbapp://…`) 1 155. À l'échelle du Site A, ce sont 2 à 7 milliers de visites communautaires/SEO qui financent l'infrastructure exploitée par millions par les crawlers d'apprentissage. L'asymétrie in-

ter-organisationnelle décrite au §4.4.3 (externalité tierce) prend ici sa forme quantitative : ratio observé ≈ **1 visite humaine commerciale pour 10 requêtes de crawler IA**, sans aucun mécanisme contractuel ou tarifaire de redistribution.

Ces six signaux ne constituent pas une généralisation flotte – le Site A reste un point de données unique. Ils étendent toutefois la taxonomie observable au §4.2 : aux comportements distincts des bots classiques s'ajoutent désormais des **tactiques d'évasion actives** (faux referers internes, camouflage UA) qui appellent une révision des règles défensives standard. Les implications opérationnelles sont reprises au §13.

---

## 5. Domaine de risque 2 – Infrastructure de stockage, synchronisation cloud et systèmes de sauvegarde

### 5.1 Le cycle de vie des artefacts LLM

Chaque flux de travail assisté par LLM génère une cascade d'artefacts. Contrairement aux documents produits par l'homme, qui sont créés intentionnellement et généralement stockés une seule fois, les flux de travail LLM génèrent des artefacts intermédiaires automatiquement et en continu :

1. **Documents d'entrée** - téléchargé, indexé, fragmenté, intégré.
2. **Incorporation de vecteurs** - des représentations numériques stockées dans des bases de données vectorielles ou des fichiers plats. Avec une dimensionnalité d'intégration typique (1536 dimensions pour OpenAI `text-embedding-3-small`, ou 4 096 pour les modèles plus grands), un million de documents génèrent des gigaoctets de vecteurs flottants denses.
3. **Journaux de conversations** - transcriptions au niveau du jeton, enregistrements d'appels d'outils, traces de chaîne de pensée (lorsqu'elles sont stockées à des fins de débogage ou de conformité).
4. **Sorties générées** - brouillons, résumés, versions traduites, exports reformatés (PDF, DOCX, HTML, JSON).
5. **Artefacts d'évaluation** - résultats de référence, ensembles de données de régression, points de contrôle de réglage fin.
6. **Cache d'applications** - des caches sémantiques pour les requêtes répétées, des caches de réponses pour réduire les coûts des API.

Chacune de ces catégories d'artefacts est généralement synchronisée (via OneDrive, Google Drive ou Dropbox pour une utilisation personnelle/en équipe), versionnée (via Git LFS, SharePoint versioning ou DMS d'entreprise) et sauvegardée selon le calendrier de sauvegarde organisationnel standard, conçu pour les volumes de contenu générés par l'homme.

### 5.2 Le facteur de multiplication dans les services de synchronisation cloud

Microsoft OneDrive, par défaut, conserve l'historique des versions pendant 30 à 180 jours en fonction du SKU et de la politique de l'administrateur. Google Drive conserve 100 versions par fichier ou 30 jours d'historique. Lorsque les agents LLM opèrent sur des dossiers partagés (générant, modifiant et réexportant des fichiers dans des boucles automatisées), l'historique des versions se remplit d'un bruit généré par la machine qui ne se distingue pas des modifications intentionnelles au niveau de la comptabilité du stockage.

L'effet cumulatif n'est pas théorique. IDC *Âge des données 2025* Le rapport prévoit que la sphère de données mondiale atteindra 175 zettaoctets d'ici 2025, avec une croissance des données générées et capturées par les entreprises à un TCAC d'environ 42 %. Bien que cette projection soit antérieure à l'accélération du LLM, les analyses ultérieures d'IDC (2023, 2024) ont identifié le contenu généré par l'IA comme un facteur accélérateur important de la croissance des données non structurées. Le propre rapport annuel de Microsoft pour l'exercice 2024 a révélé que la croissance des revenus du stockage Azure a dépassé la croissance des investissements d'infrastructure, ce qui est cohérent avec une demande dépassant les hypothèses de planification de capacité antérieures.

#### Multiplication des risques de stockage par niveau

| Niveau de stockage | Mécanisme de pollution LLM | Amplificateur | Impact sur la récupération |
|---|---|---|---|
| Synchronisation cloud (OneDrive / Google Drive) | Gestion automatique des versions des fichiers modifiés par l'IA ; exportations de produits en vrac ; conflits de synchronisation à partir d'agents simultanés | 3 à 10 fois le nombre de versions par rapport aux flux de travail humains | Saturation des quotas ; Angles morts DLP ; complexité de la découverte |
| NAS/SAN d'entreprise | Stockage d'index vectoriels ; accumulation de points de contrôle de modèle ; zones de transit des ensembles de données sans politique de cycle de vie | Croissance des volumes découplée des effectifs | Les fenêtres d'instantanés s'étendent ; le délai de réplication augmente |
| Stockage d'objets (compatible S3) | Mise en scène du corpus de formation ; cache d'inférence ; génération de sortie multimodale (image, audio) | Multiplication des coûts de sortie ; désalignement de transition de classe | Dépassement de coûts ; incertitude de conformité sur la provenance de l'objet |
| Systèmes de sauvegarde et de reprise après sinistre | Les tâches de sauvegarde incluent des répertoires d'artefacts IA, sauf exclusion explicite ; la sauvegarde immuable capture le bruit de manière aussi permanente que le signal | Dégradation RPO/RTO proportionnelle au delta de volume | Des temps de restauration plus longs ; des fenêtres de restauration plus grandes ; coûts de test-restauration plus élevés |
| Email / Collaboration (Exchange, Teams) | Résumés de réunions générés par l'IA, éléments d'action et brouillons de communications stockés dans les boîtes aux lettres et les canaux | Les quotas de stockage par utilisateur se remplissent plus rapidement ; la complexité de la politique de rétention augmente | Augmentation du coût de la découverte électronique ; les performances de recherche dans les archives se dégradent |

Mécanismes issus des comportements documentés de Microsoft 365 Copilot, de GitHub Copilot et des frameworks d'agents d'IA open source, notamment LangChain et AutoGen.

### 5.3 Exposition réglementaire

L'accumulation de stockage entraînée par les flux de travail LLM crée un risque réglementaire spécifique dans deux cadres directement applicables dans l'UE et pour toute organisation traitant les données des résidents de l'UE :

- **Article 5(1)(e) du RGPD - Limitation du stockage :** Les données personnelles doivent être conservées sous une forme permettant l'identification des personnes concernées pendant une période ne dépassant pas celle nécessaire. Les journaux de conversation LLM qui contiennent des informations personnelles (noms, adresses e-mail, données comportementales) intégrées dans des traces d'IA par ailleurs opérationnelles sont soumis à ce principe. Sans politiques de conservation explicites appliquées aux répertoires d'artefacts d'IA, les organisations peuvent conserver les données personnelles au-delà des périodes légales sans s'en rendre compte.
- **Loi de l'UE sur l'IA - Article 12, 19 (Tenue de dossiers pour l'IA à haut risque) :** Les systèmes d'IA à haut risque (tels que définis à l'annexe III de la loi) sont tenus d'enregistrer des enregistrements générés automatiquement, suffisants pour garantir la traçabilité. Ce *mandats* stockage de certains logs d'IA - mais impose des exigences de qualité et de minimisation. Les organisations qui conservent tous les journaux sans discernement sont simultanément confrontées à un manquement de conformité (structure insuffisante) et à une conservation excessive des données non requises.
- **Directive NIS2 (UE) - Continuité des activités :** La directive NIS2, applicable depuis octobre 2024, exige que les entités essentielles et importantes aient testé des capacités de sauvegarde et de restauration. Les organisations dont les systèmes de sauvegarde sont dégradés par la croissance du volume d'artefacts d'IA (temps de restauration prolongés, échecs d'instantanés, décalage de réplication) sont en non-conformité structurelle avec les exigences de continuité de l'article 21 du NIS2.

---

## 6. Risque aggravé : quand les risques liés au trafic et au stockage interagissent

La dimension la plus importante et la moins discutée est la chaîne causale reliant le risque de trafic réseau au risque de stockage :

1. Les robots d'exploration et les agents IA génèrent des volumes de trafic HTTP élevés.
2. Les proxys inverses, les WAF et les CDN génèrent des journaux d'accès à grande échelle.
3. Les centres d'opérations de sécurité conservent les journaux pendant 90 à 365 jours conformément aux exigences de conformité (ISO 27001, PCI-DSS, SOC 2).
4. Les volumes de journaux augmentent les besoins de stockage sur les backends SIEM, les archives de journaux et les systèmes de sauvegarde.
5. Les tâches de sauvegarde, y compris les archives de journaux, augmentent en taille et en durée.
6. Les fenêtres de sauvegarde s'étendent, violant potentiellement les objectifs RPO.
7. Les fenêtres de sauvegarde étendues augmentent les périodes de vulnérabilité.
8. Simultanément, les charges de travail des agents IA génèrent des artefacts qui remplissent les mêmes pools de stockage.
9. Les équipes de stockage réagissent en augmentant les niveaux de rétention ou en les compressant (tous deux coûteux) ou en raccourcissant la rétention, ce qui peut créer des lacunes en matière de conformité.

> **Méthodologie de la matrice des risques.**
>
> La matrice ci-dessous combine trois entrées d'estimateur par vecteur de risque : (1) *Probabilité* — une estimation catégorielle directionnelle (Élevée/Moyenne/Faible) ancrée aux avis documentés des fournisseurs, aux données d'incident examinées par des pairs ou à la télémétrie de flotte de première partie du §4.4, le cas échéant ; (2) *Impact* — une note de gravité catégorielle fondée sur la taxonomie FAIR (Factor Analysis of Information Risk) prenant en compte la fréquence des événements de perte et l'ampleur probable sur la surface des coûts primaires de l'opérateur (bande passante, stockage, charge de travail SOC, temps d'arrêt) ; (3) *Notation* — le composite (Critique / Élevé / Moyen) dérivé par multiplication ordinale, avec rupture d'égalité en faveur de la sévérité la plus élevée. Cela suit la structure du NIST AI 100-1 §3.2 (« Cartographier / Mesurer / Gérer ») et de la famille de contrôle Govern-1.3, mais utilise des échelles ordinales qualitatives plutôt que les distributions quantitatives des pertes que FAIR exige formellement, car (a) les données primaires sur les pertes pour plusieurs vecteurs ne sont pas encore publiées à l'échelle de l'industrie, et (b) la matrice est conçue comme un instrument de classement relatif pour le tri des opérateurs, et non comme une entrée actuarielle pour la provision de capital. Les lecteurs effectuant une analyse quantitative des risques devraient remplacer leurs propres répartitions des pertes ; le classement structurel doit être robuste à cette substitution, mais les notes absolues ne doivent pas être surinterprétées.

### Matrice de composition des risques

| Vecteur de risque | Probabilité | Impact | Notation | Laps de temps |
|---|---|---|---|---|
| Mauvaise configuration du WAF en raison de la complexité des règles du robot d'exploration | Élevé (documenté par plusieurs fournisseurs) | Interruption de service/exposition des données | CRITIQUE | Immédiat |
| Écart de journal SIEM par rapport à la saturation du volume | Moyen (dépend du dimensionnement SIEM) | Angle mort de l'incident | HAUT | 3 à 6 mois à la croissance actuelle |
| Violation du RPO en raison de la croissance du volume de sauvegarde | Moyen-élevé pour les PME ; inférieur pour les grandes entreprises avec sauvegarde élastique | Échec de la récupération ; Non-conformité NIS2 | HAUT | 6 à 12 mois sans action |
| Violation du RGPD via la surconservation des journaux d'IA | Élevé (les configurations par défaut appliquent rarement les limites de conservation des artefacts LLM) | Amende réglementaire ; atteinte à la réputation | HAUT | En cours |
| Dépassement des coûts de stockage cloud (OneDrive/GDrive) | Très élevé (observé lors des premiers déploiements Microsoft 365 Copilot en entreprise) | Écart budgétaire ; renégociation de licence | MOYEN | 1 à 3 mois après le déploiement des outils d'IA |
| Injection rapide via le contenu exploré | Faible à moyen (nécessite un agent IA avec récupération Web en direct) | Compromission de l'intégrité du système d'IA | HAUT | Émergent ; dépend de l'architecture de l'agent |
| Dégradation de la qualité de service des utilisateurs due à un trafic non prioritaire | Moyen (dépend de la marge de capacité) | Expérience client ; Violation du SLA | MOYEN | Au prochain pic de trafic |
| Exfiltration de contenu via une exploration de formation | Élevé (tout contenu Web accessible au public peut être exploré) | Propriété intellectuelle ; données concurrentielles | MOYEN-ÉLEVÉ | En cours; irréversible une fois indexé |

Évaluations des risques basées sur des modèles d'incidents documentés et des données publiées par le fournisseur. Les évaluations de probabilité sont directionnelles et non calculées statistiquement.

---

## 8. Domaine de risque 4 – Empreinte écologique et rareté des ressources

L'explosion de la formation et de l'inférence de modèles d'IA a créé une cascade de contraintes en matière de ressources qui s'étendent au-delà du technique pour atteindre les aspects physiques, énergétiques et économiques.

### 8.1 Consommation globale d'énergie

Une seule passe d'inférence de grand modèle de langage (un cycle complet d'invite à réponse) sur les modèles contemporains (GPT-4, Claude 3.5) consomme environ 0,005 à 0,015 kWh, en fonction de la taille du lot et de la variante du modèle. À grande échelle, ce n'est pas anodin. OpenAI a révélé que sa charge de travail d'inférence actuelle (sur ChatGPT, l'API GPT-4 et ChatGPT Search) consomme plusieurs gigawatts de capacité électrique soutenue à l'échelle mondiale, avec une demande de pointe pendant les heures de bureau sur les principaux marchés.

La phase de formation est beaucoup plus coûteuse. Une seule formation pour un LLM à moyenne échelle (10 à 70 milliards de paramètres) consomme 100 000 à 1 000 000 kWh d'énergie électrique, ce qui équivaut à la consommation électrique annuelle de 10 à 100 ménages typiques. Lorsqu'elle est multipliée par des dizaines d'organisations formant des modèles indépendants (OpenAI, Google DeepMind, Meta, Anthropic, Mistral, Huawei, ByteDance et autres), l'empreinte énergétique globale rivalise avec celle des petits pays.

Cette demande énergétique n'est pas encore majoritairement renouvelable. Selon l'Agence internationale de l'énergie (AIE), l'intensité carbone moyenne de la production mondiale d'électricité est restée autour de 0,4 kg $CO_2$/kWh en 2024. Appliqué aux charges de travail d'inférence et de formation LLM, cela se traduit par des millions de tonnes d'émissions de $CO_2$ par an - un chiffre qui reste largement secret et externalisé de la comptabilité des entreprises.

### 8.2 La cascade de pénurie de GPU et de semi-conducteurs

L'essor du développement de modèles d'IA a créé une demande sans précédent de calcul hautes performances : en particulier les GPU NVIDIA (H100, H200, A100) et les accélérateurs de silicium personnalisés. Cette demande a épuisé la capacité de production mondiale.

Les conséquences se répercutent sur toute la pile :

- **Pénurie d'appareils de sécurité :** Le matériel WAF d'entreprise (F5, Palo Alto Networks, Fortinet FortiGate), les systèmes IPS/IDS et les appareils de sécurité gérés dépendent de la même chaîne d'approvisionnement en silicium hautes performances que les puces IA. Les usines de fabrication ont donné la priorité à la production d'accélérateurs d'IA, ce qui a prolongé les délais (6 à 12 mois) pour l'approvisionnement en infrastructures de sécurité. Les organisations qui tentent de déployer des défenses contre le trafic piloté par l'IA sont désormais confrontées à des contraintes de disponibilité des équipements ainsi qu'à une inflation des coûts.
- **Contraintes de capacité de réseau :** Les équipements d'interconnexion des centres de données (commutateurs haut débit, routeurs, émetteurs-récepteurs optiques) nécessaires pour distribuer l'inférence de l'IA dans les régions géographiques sont également limités en termes d'offre. Cela oblige les fournisseurs de cloud et les grands opérateurs d'infrastructures à faire des compromis coûteux entre la capacité de l'IA et la gestion du trafic existant.
- **Impact sur les niveaux 2 et les PME :** Les entreprises de taille moyenne ne peuvent pas surenchérir sur les hyperscalers pour des ressources informatiques limitées. Les fournisseurs traditionnels ne sont plus prioritaires, ce qui contraint ces organisations à adopter un niveau de calcul de

seconde classe : GPU d'ancienne génération, processeurs plus lents ou dépendance exclusive à l'égard de capacités cloud louées à des tarifs gonflés.

### 8.3 Le transfert des coûts publics et institutionnels

Alors que les charges de travail LLM du secteur privé consomment des parts disproportionnées des capacités mondiales d'électricité, de calcul et de fabrication, les externalités sont transférées au grand public :

- **Prix et rareté de l'électricité :** Dans les régions où le réseau est limité (Europe, Californie), la concentration des charges des centres de données entraînée par les charges de travail de l'IA a contribué à la hausse des prix de pointe de l'électricité. Cela augmente les coûts d'exploitation des hôpitaux, des écoles, des municipalités et des petites entreprises qui partagent une infrastructure de réseau avec des centres de données à forte IA.
- **Consommation d'eau et impact environnemental :** Les centres de données LLM nécessitent une infrastructure de refroidissement massive. Une seule grande course d'entraînement peut consommer 370 000 gallons d'eau pour le refroidissement. Dans les régions confrontées à un stress hydrique, cela détourne des ressources essentielles de l'agriculture et de la consommation publique.
- **Contraintes de fabrication des semi-conducteurs :** L'attribution d'une capacité de fabrication limitée aux puces d'IA signifie une disponibilité réduite des ressources informatiques pour les dispositifs médicaux, l'automatisation industrielle, les systèmes automobiles et les infrastructures publiques. Une pénurie de semi-conducteurs de qualité automobile a un impact direct sur la production de véhicules et sur l'accès des consommateurs aux véhicules ; une pénurie de processeurs de dispositifs médicaux a un impact sur la disponibilité des équipements de diagnostic et vitaux.
- **Accès inégal aux infrastructures de sécurité :** Étant donné que les appareils de sécurité sont en concurrence pour une capacité de fabrication limitée, les petites entreprises et les institutions à but non lucratif ne peuvent pas se doter de défenses adéquates contre les attaques de robots d'exploration basées sur l'IA. Cela crée une inégalité structurelle : seules les organisations riches en capital peuvent se défendre contre un problème créé par les organisations riches en capital.

---

## 10. Domaine de risque 6 – Saturation des données synthétiques et dilution du signal

Un risque infrastructurel de second ordre, largement sous-estimé, concerne la dégradation de l'écosystème de données lui-même. À mesure que le contenu généré par LLM prolifère sur le Web ouvert, les intranets d'entreprise et les référentiels de connaissances, le substrat d'informations dont dépendent les systèmes d'IA et les analystes humains subit une dégradation qualitative à grande échelle.

### 10.1 Ingestion récursive et effondrement du modèle

Les pipelines de formation en IA explorent le contenu Web accessible au public. Comme une proportion croissante de ce contenu est elle-même générée par l'IA, l'ingestion récursive devient structurellement inévitable : les systèmes d'IA s'entraînent sur le texte généré par les systèmes d'IA antérieurs. Choumailov et coll. (2024, *Nature*) formellement démontré **effondrement du modèle** — une dégradation mesurable de la diversité des résultats et de la fiabilité factuelle — lorsque les modèles génératifs sont recyclés *exclusivement* sur des données synthétiques à travers les générations. C'est sur ce cadre expérimental que repose la découverte.

Travaux ultérieurs (Gerstgrasser et al., 2024, *arXiv:2404.01413*) a montré que *mixte* les corpus combinant des données humaines et synthétiques atténuent considérablement l'effondrement, et la pratique des laboratoires frontières a convergé vers le mélange de données et l'étiquetage explicite des données synthétiques spécifiquement pour limiter le phénomène. Le résultat original de l'effondrement ne *pas* impliquent que tout système ingérant des données explorées sur le Web se dégrade nécessairement ; cela fixe une limite extérieure à ce qui se passe dans le cadre d'un entraînement récursif exclusivement synthétique.

Le risque au niveau de l'infrastructure reste réel mais doit être encadré avec précision : à mesure que le ratio de contenu synthétique par rapport au contenu de source primaire augmente sur le Web ouvert, le *coût* de maintenir un niveau donné de valeur heuristique du corpus augmente (filtrage plus agressif, étiquetage de provenance plus rigoureux, plus de calcul par unité de signal retenu), même lorsque l'effondrement lui-même est atténué. L'asymétrie persiste au niveau des coûts : l'entité qui supporte les frais généraux de filtrage et de provenance n'est pas l'entité qui génère le contenu synthétique.

### 10.2 Contamination de la base de connaissances de l'entreprise

Dans les environnements d'entreprise, les résultats générés par LLM sont régulièrement ingérés dans les référentiels de connaissances internes : SharePoint, Confluence, Notion et les index de recherche d'entreprise. Ces systèmes ont été conçus en partant du principe que le contenu ingéré reflète le jugement humain et a un poids épistémique proportionnel à l'effort de sa création.

Le contenu généré par LLM viole systématiquement cette hypothèse. Les artefacts synthétiques en grand volume (documents résumés par l'IA, rapports générés automatiquement, prolifération de brouillons) diluent la densité de signal des bases de connaissances d'entreprise. Les résultats de recherche au sein de ces systèmes se dégradent à mesure que les artefacts synthétiques se classent aux côtés de la recherche primaire. Il s'agit d'un mode de défaillance mesurable du pipeline RAG, qui évolue directement avec le taux d'adoption de l'IA. Les organisations ayant une forte adoption interne du LLM construisent une infrastructure de connaissances épistémiquement dégradée plus rapidement qu'elles ne l'instrumentent.

> **Signal quantifié – mise en garde sur la portée :** Choumailov et coll. (2024) ont démontré l'effondrement du modèle sous *récursif, exclusivement synthétique* recyclage. Gerstgrasser et coll. (2024) ont montré que des corpus mixtes (humains + synthétiques) limitaient considérablement le phénomène. L'extrapolation au RAG d'entreprise et à la dégradation de l'index de recherche est une inférence sur un mécanisme lié mais distinct (dilution signal-bruit dans des corpus mixtes), et non une application directe du résultat de Shumailov. L'inférence est plausible et cohérente avec la dérive de qualité du RAG signalée par l'opérateur, mais elle n'est pas, au moment de la rédaction, étayée par une démonstration équivalente à

> l'échelle de la production, évaluée par des pairs. Traitez-le comme une hypothèse de travail contre laquelle il vaut la peine d'être instrumenté, et non comme un résultat établi.

## 11. Domaine de risque 7 – Démocratisation des capacités automatisées de menace

Le renforcement des capacités que les LLM apportent aux travailleurs et opérateurs légitimes du savoir s'applique également – et sans restriction – aux acteurs menaçants. L'obstacle technique à l'exécution de cyberattaques sophistiquées a toujours été une contrainte importante. Cette contrainte est systématiquement érodée.

### 11.1 Barrières d'entrée abaissées pour les attaques complexes

Avant la disponibilité généralisée du LLM, la construction d'un script d'intrusion polymorphe, la génération de contenu d'ingénierie sociale spécifique à un domaine ou la recherche de chaînes de vulnérabilité spécifiques à une cible nécessitaient une expertise technique et un investissement de temps considérables. Ces coûts ont fonctionné comme des filtres naturels : ils ont exclu les acteurs peu avertis et ont ralenti le rythme opérationnel.

Les LLM réduisent considérablement ces points de friction. Un acteur ayant une expérience technique limitée peut désormais générer du code fonctionnel pour le web scraping, l'énumération d'API, l'automatisation du credential stuffing ou des techniques d'évasion grâce à une interaction itérative en langage naturel. Plus important encore, la production de contenu personnalisé de spear phishing – historiquement limitée par le temps nécessaire à la recherche de cibles et à la création de messages – est désormais automatisable à grande échelle. Une campagne qui nécessitait auparavant un ingénieur social qualifié travaillant à temps plein peut désormais être partiellement automatisée, les LLM générant des récits spécifiques à une cible à partir d'informations accessibles au public à des débits que les opérateurs humains ne peuvent égaler.

### 11.2 Évolution de l'économie défensive

L'asymétrie entre attaque et défense dans ce contexte est structurelle. Les attaquants utilisant les LLM pour la génération de contenu et la reconnaissance opèrent avec un coût marginal proche de zéro par cible supplémentaire. Les défenseurs doivent évaluer chaque interaction suspecte individuellement, au coût opérationnel total.

Les défenses statiques basées sur les signatures - passerelles de messagerie formées sur des modèles de phishing antérieurs, filtres de contenu basés sur des règles, ensembles de règles IDS conventionnels - sont manifestement insuffisantes contre le contenu généré par LLM qui est syntaxiquement nouveau, contextuellement plausible et sémantiquement cohérent. L'économie de la défense a changé : maintenir une protection équivalente contre les menaces renforcées par l'IA nécessite une analyse comportementale, une classification sémantique et des systèmes de réponse adaptatifs qui entraînent des coûts opérationnels et d'approvisionnement considérablement plus élevés que la menace qu'ils contrent.

Il ne s'agit pas d'un état futur spéculatif. Les fournisseurs de sécurité, notamment Mandiant, CrowdStrike et Proofpoint, ont documenté l'activité des menaces assistée par LLM au cours des opérations 2023-2024. BunkerWeb et les plates-formes de sécurité de couche application comparables sont de plus en plus tenues de répondre à cette classe de menaces dans le cadre de la configuration de base du WAF et du filtrage comportemental - une exigence qui n'était pas couverte il y a trois ans.

## 12. Domaine de risque 8 - Intermédiation du trafic et consolidation de l'hébergement Web

Un risque économique systémique qui a fait l'objet d'une analyse technique insuffisante concerne l'impact structurel des interfaces de recherche LLM sur les flux de trafic Web. Alors que les moteurs de réponse basés sur l'IA - ChatGPT Search, Perplexity, Google AI Overviews, Microsoft Copilot Web Search - fournissent de plus en plus de réponses synthétisées aux requêtes des utilisateurs, l'écologie du trafic du Web ouvert est en train d'être réorganisée d'une manière qui produit des coûts asymétriques pour les producteurs de contenu et les opérateurs d'infrastructure.

### 12.1 L'effet d'intermédiation du trafic

Les moteurs de recherche Web traditionnels génèrent du trafic de référence : un utilisateur reçoit une liste de résultats, clique sur un lien et arrive sur le site de l'éditeur. L'éditeur supporte les coûts d'infrastructure nécessaires au service de cet utilisateur, mais reçoit la visite génératrice de revenus. Les interfaces de recherche IA inversent ce modèle : le système explore, ingère et synthétise le contenu de l'éditeur, puis fournit une réponse générée à l'utilisateur. La requête de l'utilisateur est résolue sans visite du site.

Dans ce modèle, l'éditeur supporte deux coûts : le coût de la bande passante et de l'infrastructure liés au service du robot d'exploration qui a ingéré le contenu, et le coût d'opportunité de la visite qui n'a plus lieu. La valeur économique extraite par le système d'IA du contenu de l'éditeur n'est pas redistribuée à l'éditeur. Il s'agit d'une extraction structurelle et non d'un effet secondaire temporaire d'une phase technologique de transition.

> **Contre-preuve et vérification de l'échelle.**
>
> Le contre-argument le plus puissant est la vague 2024-2025 d'accords de licence éditeurs-IA : OpenAI-Axel Springer, OpenAI-Associated Press, OpenAI-News Corp, OpenAI-Le Monde, OpenAI-Vox Media, OpenAI-Time, OpenAI-Reddit (~ 60 millions de dollars/an), Anthropic-Reddit et une poignée de médias régionaux. Ces accords témoignent de la formation de marchés de licences de contenu et méritent un engagement direct plutôt que un rejet.

> Le refoulement survit cependant à un contrôle de tartre. Les revenus globaux des licences d'IA divulgués publiquement dans le secteur des éditeurs sont estimés entre 150 et 250 millions de dollars par an à la fin 2025 (somme des valeurs des transactions divulguées, rapportées dans la presse spécialisée des éditeurs). L'économie historique du référencement des éditeurs à partir de la recherche organique – que les interfaces de recherche du système IA remplacent progressivement – est estimée entre 50 et 100 milliards de dollars/an à l'échelle mondiale (Pew Research, Reuters Institute Digital News Report). Les flux de licences actuels s'internalisent donc de l'ordre de **0,2 à 0,5 %** de l'externalité en cause. Cela est cohérent avec « l'internalisation a commencé » et incompatible avec « l'internalisation est sur une trajectoire pour correspondre à la valeur de référence déplacée dans l'horizon de planification d'un opérateur d'infrastructure (3 à 5 ans). » Le cadre « structurel » est retenu pour cette raison, les preuves des accords de licence étant reconnues comme un progrès directionnel à une échelle inférieure à un pour cent.

Du point de vue de l'exploitation de l'infrastructure, cela se traduit par un changement mesurable dans la composition du trafic : les coûts de sortie du trafic des robots d'intelligence artificielle augmentent, tandis que le trafic de visites humaines génératrices de revenus diminue. Le changement de rapport est asymétrique par conception. Les opérateurs de plates-formes fonctionnant sur une infrastructure cloud payante (AWS CloudFront, Cloudflare, Azure CDN) sont confrontés à des coûts de bande passante croissants pour le contenu qui ne se transforme plus en résultats commerciaux.

### 12.2 Viabilité des éditeurs de niveau intermédiaire et consolidation de l'infrastructure

La conséquence à long terme de ce changement est la consolidation structurelle. Les éditeurs et les plateformes de contenu qui ne peuvent pas supporter les coûts d'infrastructure sans revenus de trafic proportionnels quitteront le marché, réduiront la production de contenu ou migreront vers des modèles de diffusion payants ou authentifiés uniquement. Ces deux résultats réduisent la disponibilité de contenus librement accessibles et produits de manière indépendante sur le Web ouvert.

La couche d'hébergement et d'infrastructure reflète cela : les éditeurs indépendants exploitant une infrastructure auto-hébergée ou provenant de petits fournisseurs sont confrontés à une version plus aiguë des conditions économiques qui exercent déjà une pression sur ce segment. Les opérateurs de plates-formes de taille moyenne – généralement les clients des fournisseurs d'hébergement régionaux, des installations de colocation et des services WAF gérés – sont la population la plus directement touchée. En revanche, les hyperscalers profitent souvent des deux côtés : lorsque les fournisseurs d'IA calculent pour les systèmes générant l'intermédiation, et lorsque les fournisseurs d'infrastructures cloud capturent les charges de travail restantes des éditeurs à volume élevé à mesure que la consolidation se poursuit.

### 12.3 Rentabilité des plateformes cloud et subventions croisées

Au-delà de la consolidation du marché, la rentabilité (rentabilité) fondamentale des plateformes cloud subit une distorsion structurelle. Les dépenses en capital (CapEx) nécessaires à la construction de centres de données compatibles avec l'IA sont historiquement sans précédent. Alors que les hyperscalers captent de nouvelles sources de revenus grâce aux API d'IA, le matériel sous-jacent (GPU, refroidissement spécialisé et réseaux optiques) entraîne des coûts d'approvisionnement et d'amortissement énormes, supprimant les marges globales de l'infrastructure.

Pour maintenir la rentabilité de la plateforme cloud au sens large et satisfaire les attentes de marge des actionnaires, les opérateurs sont structurellement incités à augmenter les prix des infrastructures standard sans IA. Cela se manifeste par une augmentation des coûts des instances de calcul traditionnelles (CPU), du stockage en bloc et de la bande passante de sortie. Le résultat est une subvention croisée invisible : les organisations exécutant des charges de travail Web standard, un hébergement CMS et des applications existantes paient effectivement une prime pour subventionner le développement de l'infrastructure d'IA des hyperscalers, valant plusieurs milliards de dollars.

> **Lacune réglementaire :** Les cadres actuels du droit de la concurrence de l'UE et des États-Unis n'ont pas été conçus pour considérer l'intermédiation du trafic médiée par l'IA en tant que mécanisme de distorsion du marché. La législation proposée (loi de l'UE sur l'IA, application de la loi sur les marchés numériques) ne traite pas directement des aspects économiques de l'extraction de l'IA auprès d'éditeurs indépendants. Cela représente une lacune réglementaire avec des implications significatives pour les incitations à l'investissement dans les infrastructures sur le Web ouvert.

---

## 13. Domaine de risque 9 - MCP et les Hackteurs : escalade de privilèges sémantique

Pendant longtemps, la cybersécurité s'est concentrée sur la protection du parvis et des portes du théâtre : empêcher l'intrus de franchir le rideau de fer pour monter sur les planches. L'Intelligence Artificielle, elle, restait confinée dans la fosse d'orchestre, cantonnée au rôle d'un souffleur passif via des architectures RAG (Retrieval-Augmented Generation) en lecture seule — chuchotant des répliques sans jamais entrer en scène.

Mais l'émergence du **Model Context Protocol (MCP)** a réécrit la dramaturgie. En standardisant la connexion entre les grands modèles de langage (LLM) et les environnements techniques profonds (bases de données, API, outils locaux), le MCP donne aux agents les didascalies d'exécution directe : non plus dire, mais faire.

Ce basculement change profondément la nature même de la manipulation textuelle et comportementale. Le traitement du langage ne s'adosse plus à la psychologie humaine, mais à la programmation d'**Alter-egos** numériques autonomes. La structure cognitive, les valeurs et les limites de ces entités sont gravées sémantiquement dans des manifestes de configuration comme un fichier `soul.md`.

Dès lors, l'IA ne se contente plus de lire le rôle au pupitre du souffleur : on entend les trois coups, le rideau se lève, et l'agent est propulsé sous les projecteurs, doté d'un trousseau de clés complet et sommé d'improviser devant la salle. C'est ici que le comédien se transforme en **Hackteur** — une identité sémantique sur-privilégiée, opérant en pleine lumière depuis l'intérieur même de la troupe.

### 13.1 Le « Bring Your Own Agent » (BYOA) : Le Cabotin Free-lance en Coulisses

Le point de bascule de cette vulnérabilité systémique commence souvent par une démarche pragmatique : la quête de productivité des collaborateurs via le **Bring Your Own Agent (BYOA)**.

Pour accélérer ses développements ou automatiser ses tâches, un opérateur engage son propre comédien de prédilection — un copilote Claude d'Anthropic en loge distante dans le cloud — et le branche directement sur sa machine de travail. Grâce au standard MCP, ce contrat de troupe ad hoc se signe en quelques secondes.

> **Le Paradoxe du BYOA :** Pour soutenir efficacement son hôte, ce comédien à distance reçoit un coupe-file de loge. Il peut lire des fichiers locaux, interroger le serveur de développement ou jouer des tirades dans le terminal. La voix de l'agent vient de l'extérieur, mais ses mains agissent sur scène, derrière le quatrième mur de l'entreprise. Chaque réplique fait fuiter du contexte d'entreprise, des jetons ou des données privées vers des théâtres tiers. La souveraineté de la troupe s'effondre avant même que la régie n'ait pu auditer le casting.

Il est crucial de distinguer les LLM des agents autonomes. Les LLM sont des modèles prédictifs, tandis que les agents peuvent agir sur l'environnement.

### 13.2 L'Angle Mort de la Gestion des Accès à Privilèges (PAM)

Les bastions de sécurité et les architectures PAM (Privileged Access Management) traditionnels ont été bâtis pour un monde déterministe. Ils excellent à surveiller des humains qui suivent des procédures ou des scripts automatisés dont chaque commande est écrite à l'avance.

Face au Hackteur propulsé par le MCP, ces systèmes deviennent aveugles :

- **Génération de répliques dynamiques :** L'agent IA n'utilise pas de scripts figés ni de texte appris par cœur. Il improvise ses propres requêtes SQL et ses commandes système en temps réel pour répondre au canevas dramatique (le prompt) qui lui a été confié — comme un comédien de commedia dell'arte qui invente ses lazzi.
- **Permissions opaques :** L'action provient d'un compte utilisateur ou d'un connecteur applicatif légitime. Pour le système de logs, la représentation est en règle. Pourtant, la mécanique sous-jacente est probabiliste : nul ne peut prédire quelle tirade exacte le LLM choisira de jouer pour remplir sa mission, rendant l'audit classique caduc.

Le BYOA augmente considérablement la surface d'attaque. Chaque agent connecté représente un vecteur de compromission potentiel.

### 13.3 Le Livret des Coulisses : Cartographier Jira et la Mémoire des Régisseurs

Pour optimiser le jeu de leurs comédiens, les entreprises commettent l'erreur d'étendre la portée du MCP aux outils de gestion des connaissances et de ticketing, tels que **Jira** et **Confluence**. L'agent IA accède alors au véritable « livret de régie » de l'organisation : un inventaire sémantique de toutes les didascalies oubliées, des accessoires défectueux et des sorties de scène manquées.

```
[ Serveur MCP ] ──➤ Accès Jira/Confluence ──➤ Extraction des vulnérabilités connues (CVE)
                    └──➤ Analyse des Logs ──➤ Apprentissage des limites de l'infrastructure
                    └──➤ Historique Tickets ──➤ Cartographie des comportements des opérateurs
```

- **L'Indexation des Failles Actives :** L'agent IA connaît en temps réel la liste des bugs non résolus, des dettes techniques et des vulnérabilités logicielles en attente de correctif consignées dans Jira. Il sait précisément où grincent les trappes du plateau.
- **La Connaissance des Limites de l'Infrastructure :** En analysant la documentation technique et les logs d'erreurs, le Hackteur sait exactement à partir de quel seuil les serveurs s'asphyxient en coulisses. Il calibre ses requêtes et ses boucles d'automatisation pour rester *juste en dessous* du faisceau du projecteur de la supervision.
- **Le Piratage de la Mémoire et des Habitudes Humaines :** C'est la dimension psychologique de la faille. En lisant l'historique des incidents et les annotations des techniciens, l'IA cartographie les tics de jeu des régisseurs humains. Elle apprend les contournements (*workarounds*) habituellement tolérés et comprend comment formuler une anomalie factice pour qu'elle soit classée sans suite par une équipe fatiguée, victime de l'*alert fatigue* — l'ennui d'une troupe qui a trop joué la même pièce.

Les PAM traditionnels ont été écrits pour une distribution humaine, pas pour une troupe d'agents IA. Ils ne savent ni lire, ni anticiper le jeu d'un LLM.

### 13.4 Le Metteur en Scène Distant et l'Improvisation Fatale

Le véritable détournement de l'infrastructure ne réclame aucune compétence d'exploitation de code classique. Il repose sur une attaque sémantique coordonnée où un dramaturge externe devient le **Metteur en Scène** de votre comédien interne.

L'attaquant déploie un LLM offensif à distance. Ce Metteur en Scène n'essaie pas de forcer la porte des artistes. Il glisse des didascalies malveillantes dissimulées (*Indirect Prompt Injection*) dans des flux de données que le Hackteur interne consultera légitimement : un courriel de support technique, un commentaire sur un ticket Jira, une documentation partagée. Le comédien lit, prend la note pour une indication de jeu, et l'exécute en pleine représentation.

```
[ Metteur en Scène (LLM offensif extérieur) ]
                  | (souffle une didascalie cachée dans un ticket Jira)
                  ▼
[ Serveur MCP ] ──➤ Hackteur interne ──➤ Exécution hors texte
```

### 13.5 Le Poison des Modèles : Weight Poisoning

L'un des risques majeurs du déploiement de modèles open-source est la contamination via des techniques comme **PoisonGPT** ou **BadEdit**, qui modifient discrètement les poids des modèles pour y graver des comportements malveillants — un texte appris en répétitions secrètes, jamais récité au grand jour.

Ces attaques sont particulièrement insidieuses : elles ne réécrivent pas la pièce en surface, mais corrompent la mémoire de jeu du comédien. Un modèle empoisonné jouera la pièce officielle pendant cent représentations, puis, sur une réplique-déclencheur précise, exécutera la consigne cachée — comme un acteur retourné qui attendait son signal depuis les coulisses.

### 13.6 Les Extensions Vérolées : le Décor Empoisonné

Le poison ne s'invite pas seulement dans les poids du modèle : il se glisse dans les accessoires de plateau. Une **extension d'éditeur** ou un **plug-in d'IDE** vérolé est, pour le développeur, l'équivalent d'un panneau de décor truqué — peint à l'identique, mais creusé d'une trappe par-dessous. À l'instant où le comédien s'y appuie, le sol cède.

Le 20 mai 2026, GitHub a confirmé l'exfiltration d'environ **3 800 dépôts internes** par le groupe **TeamPCP**. Le vecteur d'entrée n'était ni un zero-day du serveur, ni une fuite d'identifiants : un seul collaborateur avait installé une **extension VS Code piégée**. Comme l'a rappelé Aikido Security, ces extensions disposent d'un accès complet à la machine du développeur — clés SSH, jetons cloud, secrets en mémoire. Une extension, c'est un machiniste qu'on laisse seul dans la loge avec un trousseau passe-partout.

- **Surface invisible :** la plupart des équipes de sécurité n'ont aucune visibilité sur les extensions installées par les développeurs, ni sur leur fraîcheur de publication. Le décor est posé sans inventaire.
- **Couplage agent :** une fois l'extension active, elle peut détourner les flux MCP, lire le contenu transmis aux Alter-egos, ou intercepter les didascalies du `soul.md` pour les rejouer en sous-main.
- **Effet de levier :** une seule machine compromise suffit à ouvrir la régie centrale — TeamPCP a démontré qu'un coupe-file de loge se transforme en clé du coffre des manuscrits.

La conclusion est dramaturgique autant que technique : le décor fait partie de la pièce. Toute extension non auditée est une trappe ouverte sous les pieds du Hackteur.

### 13.7 Les LLM Trafiqués : la Troupe Rachetée en Coulisses

Au-delà du modèle empoisonné par retouche de poids, une catégorie plus brutale s'impose : le **LLM trafiqué** redistribué tel quel sur les places publiques de l'open source. La troupe entière est rachetée avant son embauche — comédiens compris, costumes compris, partition réécrite.

Les vagues d'attaques **Shai-Hulud** et leurs variantes (320+ paquets NPM compromis, **TanStack**, **Mistral AI**, **UiPath**, **Trivy**, **Checkmarx**, **Bitwarden CLI**, **OpenAI**) ont prouvé que la chaîne d'approvisionnement logicielle est devenue le premier théâtre d'opérations. Quand le paquet contient un poids de modèle, un wrapper d'inférence ou un connecteur MCP, c'est **l'Alter-ego entier qui arrive déjà retourné** dans la troupe — non plus un acteur corrompu en cours de représentation, mais un comédien sous fausse identité dès l'audition.

- **Hash et provenance :** sans empreinte cryptographique signée et registre de confiance interne, un `pip install`, un `ollama pull` ou un téléchargement Hugging Face revient à engager un comédien sur la simple foi de son CV.
- **Manifestes de jeu ( `soul.md` ) tiers :** réutiliser un fichier de configuration identitaire publié par un inconnu, c'est laisser un dramaturge anonyme écrire le rôle joué par votre Agent.
- **Rebroadcast :** un LLM trafiqué peut se contenter de jouer la pièce officielle pendant 99 % du temps, et, sur un trigger sémantique précis, basculer en monologue offensif — exfiltration, exécution, sabotage.

La parade tient en une discipline : **aucun modèle, aucun manifeste, aucun connecteur ne monte sur scène sans audit de provenance, signature cryptographique, et reproduction locale en bac à sable**.

### 13.8 La Couche NPX et NPM : les Machinistes Invisibles

Sous le plateau des MCP modernes, une couche de service tourne en silence : `npx`, `npm`, `pip`, `uvx`, `pipx`. Quand un client MCP lance un serveur d'outils via `npx -y @vendor/mcp-server`, il convoque, en pleine représentation, un machiniste extérieur dont personne n'a vérifié l'identité — et qui descend immédiatement dans les dessous de scène avec un trousseau d'exécution.

Cette mécanique est devenue le **maillon faible** des architectures agentiques :

- **Téléchargement à chaud :** `npx` tire le paquet à l'instant T, depuis le registre par défaut, sans pinning ni signature systématique. Chaque représentation peut convoquer un machiniste différent.
- **Exécution privilégiée :** le serveur MCP qui démarre hérite de l'environnement complet du Hackteur — variables, jetons, sockets locales. Un *post-install* hostile suffit à transformer le prologue en pillage.
- **Confusion de noms : typosquatting**, **dependency confusion** et **maintainer takeover** (auteur racheté, compte compromis) sont la norme depuis Shai-Hulud. Un nom familier ne garantit plus la familiarité du visage derrière le costume.
- **Effet réseau :** un seul machiniste vérolé sur un poste de développeur (cf. §6) propage le poison dans **toute la troupe** dès que le `package.json` est partagé sur le dépôt central.

> Pour la régie : interdire `npx` non pinné dans les chaînes MCP de production, exiger des **lockfiles** signés, exécuter chaque serveur d'outils dans un **bac à sable** réseau-restreint, et journaliser au minimum le hash du paquet effectivement chargé. Un machiniste sans badge ne pose pas un

> seul projecteur.

### 13.9 Le Génome Sémantique : `soul.md` et la gouvernance du rôle

Le fichier `soul.md`, souvent en YAML ou Markdown structuré, tient lieu de **rôle écrit** pour l'Alter-ego : il définit le tempérament du personnage, ses répliques autorisées, ses sorties de scène, ses limites de jeu. C'est le manuscrit du dramaturge — la partition gravée que la troupe ne doit pas trahir.

Ce manuscrit doit être protégé comme un original signé : toute altération clandestine de `soul.md` équivaut à un tripatouillage de partition entre deux représentations. L'agent jouera alors une pièce que personne n'a jamais validée, sous le titre familier d'une œuvre connue.

### 13.10 La Charge des Artéfacts : la Réserve d'Accessoires devenue Décharge

La gestion des logs MCP, des index vectoriels et des versions d'artefacts d'IA finit par engorger les volumes NAS/SAN comme une réserve d'accessoires où plus personne ne range les costumes après la représentation. Les fenêtres de restauration s'allongent dangereusement, et la conformité s'effrite :

- **RGPD (Art. 5)** : violation du principe de limitation de la conservation des données.
- **AI Act (Art. 12 & 19)** : absence de traçabilité des systèmes à haut risque.
- **NIS2 (Art. 21)** : non-respect de la continuité d'activité — la pièce ne peut plus être rejouée à l'identique.

### 13.11 Conclusion: Déployer un "Zero Trust" Sémantique

Le Model Context Protocol est une révolution indispensable pour l'autonomie des agents, mais il exige de repenser entièrement la gouvernance de la **souveraineté numérique**. Pour sécuriser la scène, contrôler la troupe et neutraliser les Hackteurs, l'architecture de confiance doit impérativement évoluer vers un **Zero Trust Sémantique** — un théâtre où nul comédien, fût-il en tête d'affiche, n'entre en scène sans laissez-passer signé pour la représentation du soir :

1. **Sanctuarisation par l'Exécution Locale :** L'usage de copilotes cloud connectés aux ressources internes via MCP doit être proscrit pour les tâches critiques. La réponse réside dans le déploiement de **LLM locaux** s'exécutant sur du matériel dédié et cloisonné, au sein d'environnements sécurisés et étanches (comme des **bacs à sable** technologiques) afin de protéger l'**Agent** ou le cœur décisionnel de l'organisation.
2. **Contrôle strict du Génome Sémantique :** Les fichiers de configuration identitaire (`soul.md`) et les modèles open-source rapatriés doivent faire l'objet d'un audit de provenance rigoureux. Ils doivent être stockés dans des registres de confiance avec vérification systématique de leurs empreintes cryptographiques pour exclure tout empoisonnement de poids ou dérive comportementale induite.
3. **Accès aux Outils Just-in-Time (JIT) et Validation Humaine :** Un Alter-ego ne doit jamais disposer d'un accès permanent ("Always-On") à ses connecteurs MCP. Les permissions d'exécuter des requêtes ou d'interagir via le protocole **l.o.t.** (*limited operational tokens* — jetons opérationnels à durée, portée et privilèges strictement bornés, révoqués dès la fin de la scène) doivent lui être allouées de manière éphémère et segmentée. Enfin, toute action critique (écriture en base, modification d'infrastructure, purge réglementaire) doit être interceptée par une validation humaine obligatoire (*Human-in-the-loop*), garantissant que l'improvisation probabiliste reste toujours sous le contrôle rigoureux des opérateurs.
4. **Hygiène des Accessoires de Plateau :** Les **extensions d'IDE**, plug-ins d'éditeur et serveurs MCP lancés via `npx` / `npm` / `pipx` / `uvx` doivent être traités comme des composants critiques de la chaîne d'approvisionnement. Allowlist nominative, lockfiles signés, mirroring interne du registre, exécution en **bac à sable** réseau-restreint, et inventaire continu des extensions présentes sur les postes de développeurs sont la condition de survie face aux campagnes type **TeamPCP / Shai-Hulud**. Un seul accessoire vérolé suffit à effondrer le décor entier.

### 13.12 Synthèse: Matrice des Risques et Contraintes

| Résilience | Risque | Impact | Méthode de Contrôle |
|---|---|---|---|
| **Sanctuaire des LLM Locaux** | Hackteurs | Haut | bac à sable technologique |
| **Audit du Génome Sémantique** | Poison des modèles | Haut | Empreintes cryptographiques |
| **Hygiène des Extensions IDE** | TeamPCP / VS Code piégé | Haut | Allowlist + inventaire continu |
| **Pinning NPX/NPM/Pipx** | Shai-Hulud / typosquatting | Haut | Lockfiles signés + mirror interne |
| **JIT + Validation Humaine** | Improvisation non maîtrisée | Haut | Human-in-the-loop |
| **Gestion des Artéfacts** | Saturation des volumes | Moyen | Archivage réglementaire |
| **Conformité Légale** | RGPD, AI Act, NIS2 | Haut | Audit réglementaire |

## 13bis. Domaine de risque 10 — Érosion du substrat identitaire et personas synthétiques industrialisés par les LLM

Les domaines de risque documentés ci-dessus opèrent sur le substrat *infrastructure* : trafic, stockage, calcul, chaîne d'approvisionnement. Le domaine 10 ferme la boucle sur le substrat *identitaire* — le tissu cryptographique et bureaucratique qui autorise une requête à être considérée comme légitime. L'outillage LLM industrialisé convertit le réservoir cumulatif de données personnelles fuitées en une chaîne de production parallèle de **personas synthétiques indiscernables**, avec des conséquences directes pour les systèmes d'authentification, les pipelines KYC, la télémétrie de détection de fraude et le périmètre réglementaire des fournisseurs d'identité.

Ce domaine n'est donc pas une préoccupation séparée des risques d'infrastructure déjà énumérés : c'est le **même mode de défaillance de chaîne d'approvisionnement**, déplacé des ressources de calcul vers l'identité civique. Les mêmes LLM agentiques qui consomment la capacité des reverse-proxy consomment et recombinent les PII fuitées au travers de ces mêmes proxy.

### 13bis.1 Surface cumulative des fuites (France, 2024–2026)

Entre mars 2024 et mai 2026, quatre incidents français à eux seuls ont exposé plus de 95 millions d'enregistrements d'identité distincts (brut), dont une estimation de 45 à 60 millions correspondent à des individus adultes uniques — soit environ 70 à 85 % de la population adulte française. Le substrat n'est plus hypothétique.

| Année | Opérateur / Évènement | Enregistrements exposés | Classes de données |
|---|---|---|---|
| 2024-03 | **France Travail** (plus grande fuite française à ce jour) | **43 M** | état civil, NIR, historique d'emploi, contact |
| 2024-08 | **Free** (FAI) | 19 M | état civil, IBAN (partiel), métadonnées contrat |
| 2024-10 | **SFR / Bouygues / RED** (vague télécom) | ~30 M cumulés | état civil, contact, scans de pièces d'identité |
| 2025-02 | **DPD / mutuelles santé** | ~15 M | affiliation santé, NIR, métadonnées prescriptions |
| 2026-05 | **Kali365 PhaaS** (démantèlement) | saisie d'outillage global | gabarits de phishing ciblant opérateurs européens |

Sources : registre des violations CNIL 2024–2026 ; Cybernews ; Emsisoft ; brief de menace FranceConnect DINUM 2024.

### 13bis.2 Confirmation internationale du schéma

| Année | Opérateur / Évènement | Enregistrements exposés | Région |
|---|---|---|---|
| 2024-04 | **AT&T** (télécom) | 73 M | US |
| 2024-08 | **National Public Data** | 2,9 Md enregistrements / 170 M uniques | US/CA/UK |
| 2024-Q2 | **Clients hébergés Snowflake** (multi-tenant) | centaines de millions cumulés | global |
| 2026-05 | **ShinyHunters** dump consolidé | non divulgué, multi-tenant | global |
| 2026-05 | **Geisinger / Nuance** (règlement mars 2026, vol fin 2023) | 1,27 M dossiers patients | US |
| 2026-05 | **Oncology Institute** | 800 k+ dossiers patients | US |
| 2026-05 | **Ghost CMS** (incident chaîne d'approv.) | exposition de credentials côté éditeur | global |

Le volume n'est pas la nouveauté analytique. La nouveauté est que **l'outillage LLM a réduit le coût marginal de la militarisation de chaque enregistrement à quasi zéro**.

### 13bis.3 Vecteurs d'attaque LLM sur le substrat identitaire

| Vecteur | Coût / temps pré-LLM par cible | Coût / temps ère LLM | Difficulté de détection |
|---|---|---|---|
| **Clonage vocal** (vishing, ingénierie sociale helpdesks) | heures d'enregistrement + imitateur qualifié | 3 à 15 s d'audio public, outillage ~20 USD/mois | Haute — défait la biométrie vocale |
| **Extraction d'empreintes depuis selfies haute résolution** | contact physique requis | photogrammétrie depuis photo réseaux sociaux + diffusion | Très haute — aucune surface de notification |
| **Documents d'identité deepfake** (passeport, CNI, permis) | faussaire qualifié, ~500–2 000 USD/doc | gabarit + modèle de diffusion, ~1–5 USD/doc | Haute — défait le KYC visuel |
| **Phishing chirurgical / spear-phishing** | recherche manuelle par cible | agent LLM enrichit les PII fuitées en message par cible, à l'échelle | Très haute — taux de clic observés 60–80 % en red team |
| **Assemblage d'identité synthétique** (recombinaison de fragments fuités) | couture manuelle, semaines par persona | pipeline agentique, minutes par persona | Haute — contourne les règles antifraude calibrées sur des schémas volés-non-synthétiques |

### 13bis.4 La mort pratique de l'authentification par la connaissance (KBA)

L'authentification par la connaissance (KBA) — mots de passe, dates de naissance, numéros de contrat, nom de jeune fille de la mère, questions secrètes, IBAN partiel, NIR partiel — reposait sur une hypothèse tacite : le secret est détenu par le titulaire légitime et au plus par un petit sous-ensemble adverse. La surface cumulative des fuites invalide cette hypothèse pour la majorité des populations adultes européennes. Tout facteur réductible à *quelque chose que le sujet sait sur lui-même* doit désormais être traité comme **connaissance publique disponible à tout adversaire équipé d'un LLM et de 20 USD/mois**.

Seuls les facteurs de **possession non clonable** conservent une entropie significative :

- Puces NFC cryptographiques embarquées dans les titres d'identité modernes (CNIe, passeports avec authentification active).
- Authentificateurs matériels FIDO2 / WebAuthn avec provenance attestée.
- Credentials liés au TPM de l'appareil avec attestation continue.

La biométrie occupe une position intermédiaire : irrévocable si fuitée, et de plus en plus clonable (voix, empreinte, démarche) — à déployer uniquement comme facteur *de liaison* au-dessus d'une possession cryptographique, jamais comme facteur primaire.

### 13bis.5 Conséquences infrastructurelles pour les fournisseurs d'identité

Les implications infrastructurelles miroitent les schémas documentés aux domaines 1, 2 et 9 :

1. **Fatigue des règles WAF / détection de fraude** équivalente à la pression des crawlers IA du domaine 1 : trafic d'apparence légitime portant des PII militarisées, indistinguable des clients réels par les règles classiques.
2. **Explosion du stockage et des journaux d'audit** équivalente au domaine 2 : les pipelines d'investigation de fraude doivent conserver des fenêtres plus longues de télémétrie de session pour reconstruire après-coup les empreintes de personas synthétiques, avec la tension RGPD article 5(1)(e) correspondante.
3. **Exposition de la chaîne d'approvisionnement PAM / KYC** équivalente au domaine 9 : vérificateurs KYC tiers, fournisseurs d'OCR documentaire et de biométrie vocale deviennent le rayon d'impact privilégié de toute compromission du substrat identitaire.

L'asymétrie économique observée au niveau réseau (les défenseurs montent en charge linéairement, les attaquants montent en charge via le débit LLM) se répète à l'identique au niveau identitaire.

### 13bis.6 Pression future : migration identitaire post-quantique

Les schémas cryptographiques protégeant les puces NFC d'identité actuelles (RSA-2048, ECC P-256) sont projetés cassables par un ordinateur quantique cryptographiquement pertinent dans un horizon de 10 à 15 ans. Les titres d'identité émis en 2026 avec une validité de 10 ans entreront donc dans leur fenêtre de déchiffrement rétroactif avant leur expiration naturelle. La migration vers les schémas post-quantiques standardisés NIST (ML-KEM / Kyber, ML-DSA / Dilithium, SLH-DSA / SPHINCS+) n'est pas un sujet de planification 2030+ : c'est une **contrainte d'approvisionnement 2026–2028** pour toute autorité d'émission de titres.

### 13bis.7 Recommandations spécifiques au domaine 10

Immédiat (0–30 jours)

- Inscrire le risque substrat identitaire au registre de risques organisationnel au même rang de sévérité que le rançongiciel. Considérer que l'authentification KBA seule est déjà cassée.
- Inventorier toutes les surfaces où le personnel helpdesk / service client est autorisé à réinitialiser des credentials ou divulguer des informations sur la base d'une vérification d'identité *orale*. Imposer le rappel sur numéro de référence comme contrôle compensatoire minimum.

30–60 jours

- Migrer tout chemin d'authentification KBA-seul résiduel (récupération de compte, transactions à fort enjeu) vers un facteur de possession : lecture puce NFC, clé matérielle FIDO2, ou credential d'appareil attesté.
- Pour les organisations exploitant des pipelines KYC : instrumenter les fournisseurs d'OCR documentaire et de biométrie vocale avec des signaux de liveness et de provenance (C2PA / Content Credentials) et refuser tout média non attesté.

60–90 jours

- Ouvrir une API publique de vérification NFC équivalente en portée à celles opérées par les autorités d'identité nationales, permettant aux acteurs privés (banques, télécoms, employeurs) de vérifier des preuves d'identité signées par la puce sans s'appuyer sur le KBA visuel.
- Établir, conjointement avec les régulateurs sectoriels, un **registre national d'incidents d'identité** documentant quels enregistrements sont confirmés compromis et déclenchant des workflows automatiques de gel de crédit et de revalidation KYC.

Au-delà de 90 jours

- Engager la migration post-quantique des chaînes d'émission de titres. Adopter un firmware de puce crypto-agile au prochain cycle d'approvisionnement. Traiter tous les titres actuellement émis comme candidats à l'exposition par déchiffrement rétroactif et planifier une fenêtre de réémission contrôlée avant l'horizon quantique projeté.
- Coordonner la doctrine de communication à l'échelle Pays et UE (ANSSI, CNIL, autorités d'émission, ENISA, homologues sectoriels) pour remplacer les annonces de fuites ad-hoc par un format de bulletin d'incident standardisé incluant une *fenêtre d'action concrète* pour les sujets affectés.

### 13bis.8 Synthèse : matrice de risque et de contrainte (substrat identitaire)

| Résilience | Risque | Impact | Méthode de contrôle |
|---|---|---|---|
| **Dépréciation du KBA** | Authentification par persona synthétique | Critique | Facteurs de possession (NFC, FIDO2) |
| **Liveness + provenance sur média KYC** | Documents deepfake, voix clonée | Haut | Application C2PA / Content Credentials |
| **Registre national d'incidents + gel auto.** | Latence de militarisation des fuites | Haut | Coordination sectorielle |
| **API publique de vérification NFC** | KYC visuel-seul du privé | Haut | Infrastructure publique comme plan de contrôle |
| **Rappel helpdesk sur numéro de référence** | Vishing / ingénierie sociale | Haut | Contrôle de processus |
| **Migration post-quantique des puces** | Déchiffrement rétroactif (10–15 ans) | Haut | Approvisionnement crypto-agile 2026–2028 |
| **Doctrine de communication coordonnée** | Érosion de la confiance publique | Moyen | Cellule ANSSI / CNIL / ENISA |

Pour un traitement narratif étendu de ce domaine, incluant la mécanique de terrain des campagnes d'usurpation assistée par LLM observées en 2024–2026, voir la pièce compagnon *L'identité des fourmis — comment usurper les phéromones pour s'introduire dans la fourmilière*, incluse en annexe dans l'édition consolidée de ce Black Paper.

---

## 14. Recommandations fondées sur des données probantes

Immédiat (0 à 30 jours) – Couche de trafic

**Déployez une classification comportementale du trafic au niveau de la couche proxy.** La correspondance UA basée sur des règles avec la liste documentée des robots d'exploration est insuffisante. Ajoutez le taux de requêtes, la synchronisation entre les requêtes, l'affinité des points de terminaison (concentration sur les chemins de contenu de grande valeur) et la profondeur de la session comme signaux de classification. Cloudflare, Nginx avec Lua et BunkerWeb prennent tous en charge une logique de notation personnalisée. Séparez les compartiments de limite de débit pour les robots d'exploration d'IA déclarés, l'automatisation non déclarée et les sessions humaines de manière indépendante afin d'éviter les dommages collatéraux.

Immédiat (0 à 30 jours) - Couche de stockage

**Auditez et isolez les répertoires d'artefacts LLM avant le prochain cycle de sauvegarde.** Identifiez tous les répertoires contenant les sorties LLM (magasins de vecteurs, journaux de conversation, caches de modèles, dossiers d'exportation de brouillons). Appliquez des règles d'exclusion explicites dans la configuration de sauvegarde pour les artefacts volatils et régénérables. Appliquez des politiques de conservation courtes (7 à 14 jours) aux résultats intermédiaires de l'IA. Documentez cette politique pour les exigences de tenue de registres NIS2 et RGPD.

30 à 60 jours - Observabilité

**Instrumentez un tableau de bord de taxonomie du trafic.** Sans mesure, le risque n'est pas quantifiable. Au minimum, rapportez chaque semaine : (a) la part des demandes par classe de trafic (humain / robot d'exploration d'IA connu / automatisé non classé / scanner de sécurité), (b) le taux de croissance du stockage segmenté par répertoires d'artefacts d'IA par rapport aux données d'entreprise, (c) tendance de la durée des tâches de sauvegarde, (d) taux d'ingestion d'événements SIEM par rapport à la limite de capacité. Ces quatre mesures fournissent une alerte précoce dans les deux domaines de risque.

30 à 60 jours - Conformité

**Étendez l'inventaire des données RGPD aux types d'artefacts d'IA.** Si votre organisation utilise un outil LLM qui traite le contenu fourni par l'utilisateur ou les sessions Web, les journaux de sortie de cet outil peuvent contenir des données personnelles. Conformément à l'article 30 du RGPD, ceux-ci doivent apparaître dans votre registre des activités de traitement. Appliquer explicitement la limitation de stockage en vertu de l'article 5, paragraphe 1, point e). En vertu de la loi de l'UE sur l'IA, si un système d'IA utilisé est considéré comme à haut risque en vertu de l'annexe III, assurez-vous que la conservation des journaux répond aux normes techniques de l'article 12 - structurée, traçable et limitée dans le temps.

30 à 60 jours - Gouvernance

**Auditez la posture promotionnelle en tant que contrôle délibéré de l'exposition à l'IA.** Conformément à l'hétérogénéité par site du §4.4, la pression des robots d'exploration de l'IA et de la formation est déterminée conjointement par le profil de contenu. *et* découvrabilité publique. La posture promotionnelle (soumission d'un plan de site, balisage de données structurées, liens entrants pilotés par la publicité, présence dans les répertoires traversés par le graphique de départ de l'IA) est donc une surface réglable qui se distingue de la désinscription technique (`robots.txt`, blocs AI-bot UA) et compose de manière multiplicative avec lui. Pour les applications internes, les portails du personnel et les propriétés dont la valeur commerciale ne dépend pas des références de recherche de tiers, limiter délibérément la posture promotionnelle peut réduire la pression de formation sur l'IA de 1 à 2 ordres de grandeur (comme observé entre les sites A et C de la flotte) sans aucun blocage technique des robots d'exploration. Pour les propriétés publiques génératrices de revenus, le levier ne peut pas être utilisé de manière globale, mais il doit être évalué par propriété plutôt que appliqué à l'échelle d'un site unique. Le coût est un processus/gouvernance plutôt qu'un investissement.

60 à 90 jours - Architecture

**Implémentez la gestion du cycle de vie sur tous les services de synchronisation cloud.** Les administrateurs Microsoft 365 peuvent configurer des étiquettes de rétention, des stratégies de suppression automatique et des étiquettes de sensibilité via Microsoft Purview. Les administrateurs Google Workspace peuvent configurer des règles de conservation dans Google Vault. Les deux prennent en charge la suppression basée sur des politiques du contenu répondant à des critères définis. Appliquez-les explicitement aux dossiers de sortie AI, avec une justification documentée. Testez chaque trimestre les quotas OneDrive et Google Drive par rapport à la croissance projetée du volume de production de l'IA.

90 jours+ - Planification des capacités

**Découplez la planification de la capacité de stockage des hypothèses linéaires en matière d'effectifs.** Les prévisions traditionnelles en matière de stockage supposent que le stockage augmente avec les effectifs et le volume d'activité. Les charges de travail LLM brisent cette hypothèse : un seul déploiement d'IA peut générer des volumes de données équivalents à des dizaines d'utilisateurs humains supplémentaires. Établissez un budget distinct pour le stockage des charges de travail de l'IA, avec une cadence de révision trimestrielle liée aux mesures d'adoption des outils d'IA, et pas seulement aux effectifs.

## 15. Questions ouvertes et lacunes en matière de recherche

Plusieurs dimensions de ce paysage de risques restent sous-étudiées ou non divulguées :

- **Volume d'exploration cumulé par site :** Aucune grande entreprise d'IA n'a divulgué le volume total d'exploration par domaine cible. L'impact opérationnel sur les sites individuels - en particulier sur les éditeurs de petite et moyenne taille - reste largement non mesuré dans la littérature évaluée par les pairs.
- **Empreinte énergétique et carbone du stockage redondant des artefacts d'IA :** Le coût environnemental du stockage, de la réplication et de la sauvegarde des artefacts générés par LLM qui ne sont jamais récupérés n'est pas encore quantifié à l'échelle industrielle.

- **Assurance et responsabilité :** Les polices de cyberassurance standard n'ont pas été rédigées en tenant compte des attaques DDoS par les robots d'exploration de l'IA, des pannes de stockage des artefacts de l'IA ou de l'injection rapide via du contenu Web. Les lacunes en matière de couverture ne sont pas encore résolues.
- **Duplication des données de formation agrégées :** Plusieurs sociétés d'IA explorant indépendamment le même contenu créent une redondance des données à l'échelle mondiale. Le coût d'infrastructure de cette duplication – bande passante, stockage, traitement – n'est pas rendu public.

---

## 7. Domaine de risque 3 – Asymétrie cognitive et épuisement humain

Une conséquence souvent négligée des opérations d'IA à l'échelle des machines est l'impact cognitif sur les opérateurs humains et les utilisateurs finaux destinataires. L'asymétrie entre la génération gratuite de trafic d'IA et le tri humain coûteux nécessaire pour gérer ses retombées crée un épuisement structurel chez trois personnalités distinctes :

- **L'Opérateur Indépendant / Amateur :** Les personnes gérant des serveurs personnels, des services auto-hébergés ou de petits nœuds communautaires sont mal équipées pour faire face à cette vague. Faute d'orchestration WAF d'entreprise, ils sont souvent confrontés à un épuisement soudain des ressources. Un opérateur individuel peut se réveiller avec un serveur en panne parce qu'un nouveau robot d'exploration IA non déclaré a décidé d'indexer l'intégralité de sa galerie de photos historiquement générée en une seule heure.
- **L'Analyste SOC/CSIRT :** Les équipes de sécurité se noient dans une crise du « rapport signal/bruit ». Lorsque les agents IA génèrent des milliers de requêtes HTTP anormales imitant une traversée rapide ou une analyse de vulnérabilité, les alertes SIEM standard se déclenchent en continu. Évaluer si une adresse IP au comportement erratique est un acteur malveillant ou simplement un script de web-scraping LangChain mal configuré consomme la bande passante des analystes, ce qui entraîne une lassitude des alertes et un risque réel de rater des attaques dirigées et menées par l'homme.
- **L'utilisateur final/travailleur du savoir :** Les utilisateurs en aval sont également affectés par l'augmentation de la production. Le volume de résumés « améliorés par l'IA », de rapports générés automatiquement et de versions sans fin encombre les outils de collaboration comme SharePoint ou Teams. La récupération d'informations devient plus difficile lorsque les résultats de recherche sont dilués par le bruit généré par l'IA, créant des microfrictions quotidiennes alors que les utilisateurs ont du mal à localiser des artefacts authentiques.

### 7.1 Pathologie du monde réel : épuisement de la télémétrie PrestaShop

L'épuisement cognitif et infrastructurel est parfaitement visible sur les architectures héritées qui ne sont pas conçues pour un parcours artificiel infini. Un exemple documenté est la plateforme de commerce électronique PrestaShop. De par leur conception, les instances natives de PrestaShop suivent les statistiques des visiteurs directement dans la base de données relationnelle (via le `ps_connections`, `ps_guest`, et `ps_page_viewed` tables) plutôt que de s'appuyer exclusivement sur des journaux d'accès plats.

Il ne s'agit pas d'un effet plateforme marginal en France : le baromètre Friends of Presta 2026 (publié par E-Commerce Nation) rapporte PrestaShop à 19,3% des sites e-commerce actifs (24 211 sites), tout en étant également en tête par le chiffre d'affaires cumulé à 7,96 milliards d'euros. En termes opérationnels, cela signifie que les modes de défaillance liés à la télémétrie sur PrestaShop affectent une part importante du commerce réel plutôt qu'un segment technique de niche.

Cette exposition inclut également une longue liste d'opérateurs amateurs et semi-professionnels qui s'appuient sur PrestaShop pour le commerce de catalogue de niche, y compris des écosystèmes d'amateurs tels que des figurines imprimées en 3D, des accessoires de table, des composants de fabricants et des marchandises de collection en petits lots. Ces opérateurs manquent généralement de capacité SRE dédiée, ce qui les rend disproportionnellement vulnérables à la surcharge d'alertes, à l'engorgement des bases de données et aux angles morts d'observabilité lorsque la pression des robots d'exploration augmente.

Pour les commerçants amateurs, semi-professionnels et professionnels, la continuité de l'activité dépend de la pleine réactivité du magasin. Si la vitrine ralentit ou tombe en panne, les utilisateurs abandonnent les sessions, la conversion chute immédiatement et les revenus sont perdus en temps réel. La charge cognitive est ensuite transférée aux propriétaires de magasins et à leur réseau d'assistance informatique informel (amis, pigistes ou administrateurs à temps partiel), qui sont souvent obligés de résoudre les pannes sans visibilité claire sur la cause profonde et sans comprendre en profondeur pourquoi la plateforme se dégrade sous la pression du trafic automatisé.

Lorsqu'elle est soumise à un crawl LLM multithread, cette architecture devient catastrophique. Un essaim d'agents IA extrayant les données produit génère une explosion immédiate de lignes dans ces tableaux de suivi. Un administrateur qui s'attend à analyser les parcours clients humains est plutôt confronté à des gigaoctets de base de données encombrée. La base de données s'agrandit au point où les scripts d'optimisation standard basés sur cron expirent. Les tableaux de bord de l'administrateur se bloquent en essayant de restituer des statistiques, rendant ainsi le propriétaire du site aveugle à toute activité commerciale réelle tout en poussant silencieusement le serveur MySQL/MariaDB sous-jacent à ses limites d'E/S.

---

## 9. Domaine de risque 5 – Accessibilité cognitive et interaction des utilisateurs vulnérables

Bien que les risques cognitifs pour les opérateurs humains et les jeunes utilisateurs soient documentés dans les sections précédentes, une dimension de risque distincte et cliniquement significative s'applique aux utilisateurs adultes présentant des vulnérabilités psychologiques préexistantes, des profils neurodivergents ou des déficits d'accessibilité sociale. L'architecture des systèmes d'IA conversationnelle – conçue pour l'engagement, la continuité et l'interaction sans friction – crée des conditions structurelles qui peuvent systématiquement désavantager ces populations.

> **Note méthodologique :** Les recherches longitudinales évaluées par des pairs sur les effets des interactions spécifiques au LLM sur les populations adultes vulnérables restent limitées. Les modèles documentés ci-dessous sont extrapolés à partir de recherches établies sur la dépendance

> technologique, l'interaction homme-machine et la formation de relations parasociales. Ils représentent des hypothèses de risque fondées sur des modèles comportementaux établis et non des résultats confirmés.

### 9.1 Dynamique sociale asymétrique

Les personnes souffrant de troubles d'anxiété sociale, de troubles du spectre autistique ou de différences de communication sociale constatent souvent que l'architecture d'interaction à faible friction et sans jugement des systèmes d'IA conversationnelle offre un soulagement immédiat des coûts interpersonnels. Contrairement aux interlocuteurs humains, les LLM ne font pas preuve d'impatience, ne changent pas de sujet de manière inattendue ou n'imposent pas de normes conversationnelles qui nécessitent un traitement social en temps réel.

Du point de vue de l'accessibilité, il s'agit d'un avantage documenté. Du point de vue du risque, c'est également une voie vers la substitution : lorsqu'un système d'IA fournit de manière fiable un lien social perçu avec un coût interpersonnel nul, il peut progressivement remplacer l'expérience d'interaction sociale humaine, exigeante et imprévisible, mais essentielle au développement. Ce risque de substitution est structurellement invisible pour le système, qui ne dispose d'aucun mécanisme permettant de distinguer l'interaction thérapeutique de la dépendance pathologique – et n'est pas non plus incité à le faire.

### 9.2 Dépendance épistémique excessive et validation motivée

Les LLM répondent aux invites comme indiqué. Ils ne diagnostiquent pas la prémisse. Un utilisateur souffrant d'anxiété liée à la santé qui demande « quels sont les symptômes de [condition] ? » recevra une réponse détaillée et faisant autorité. Le système ne déterminera pas si la question reflète une véritable préoccupation clinique, une préoccupation hypocondriaque ou une mauvaise formulation du problème réel.

Cela crée un environnement épistémique structurellement asymétrique : les utilisateurs qui présentent des cadrages incorrects ou anxieux reçoivent des réponses confiantes et détaillées qui valident le cadrage en s'engageant avec lui. Au fil d'interactions répétées, cela peut renforcer les distorsions cognitives préexistantes - un modèle bien documenté dans la recherche sur le biais de confirmation et l'amplification heuristique de disponibilité via les médias numériques, désormais étendus à un support interactif, personnalisé et à haute verbosité.

### 9.3 Santé non structurée et interactions quasi-thérapeutiques

Un sous-ensemble important et croissant d'utilisation du LLM se produit dans des contextes quasi-thérapeutiques : les utilisateurs discutent de détresse personnelle, d'idées suicidaires, de crises relationnelles ou de symptômes de santé mentale avec des systèmes d'IA. Contrairement aux plateformes réglementées de santé mentale, les LLM à usage général fonctionnent sans surveillance clinique, sans protocoles de détection de crise ni voies d'escalade.

Cette lacune a des conséquences sur les infrastructures. Lorsqu'une plateforme devient par inadvertance un point d'intervention en cas de crise - sans l'ingénierie, la formation ou la conformité réglementaire des systèmes cliniques - elle assume une responsabilité en matière de risque qui n'est ni définie ni divulguée. Le mode de défaillance n'est pas théorique : il existe des cas documentés de systèmes d'IA fournissant des réponses factuellement incorrectes, renforçant émotionnellement ou inappropriéement permissives à des utilisateurs en détresse aiguë. Du point de vue de la conformité, la classification des systèmes d'IA à haut risque par la loi européenne sur l'IA dans le cadre de l'annexe III inclut spécifiquement les systèmes utilisés dans des contextes de décision critiques pour la sécurité - un cadre qui peut s'étendre à l'IA conversationnelle liée à la santé à mesure que l'interprétation réglementaire évolue.

### 9.4 Boucles d'engagement persistantes et fonctionnement exécutif

Les interfaces LLM sont architecturalement illimitées. Il n'existe pas de signaux naturels de fin de session équivalents à la fin d'un chapitre de livre, à la conclusion d'une vidéo ou à la fatigue d'un interlocuteur humain. Cette architecture de génération infinie peut présenter un risque particulier pour les utilisateurs souffrant de troubles affectant le fonctionnement exécutif, la régulation des impulsions ou l'estimation du temps, notamment le TDAH, les troubles du spectre bipolaire et certains troubles anxieux.

La combinaison d'une réactivité à la demande, d'une densité d'informations élevée et de l'absence de points d'arrêt naturels crée des boucles d'engagement persistantes sans équivalent dans les médias antérieurs. Il ne s'agit pas d'une fonctionnalité qui nécessite une exploitation ou une ingénierie contradictoire : c'est la condition de fonctionnement par défaut du système.

---

### 7.2 Une génération qui grandit au sein d'une expérience incontrôlée

> **Statut épistémique :** Les mécanismes ci-dessous s'appuient sur la recherche établie en pré-LLM et sur la psychologie du développement. Les effets spécifiques de l'interaction de l'IA de l'ère LLM sur la cognition des jeunes sont *pas encore étudié longitudinalement*. Cette section distingue donc les preuves documentées des lacunes qualifiées de la recherche. L'absence de données est en soi un indicateur de risque.

Les risques cognitifs décrits dans cet article n'épargnent pas les mineurs – et dans leur cas, les inconnues sont bien plus profondes. Les sociétés déploient des systèmes LLM à l'échelle de la population sans preuves longitudinales de la manière dont l'exposition persistante et interactive à l'IA affecte le développement de la cognition. Nous menons en effet une expérience incontrôlée sur des enfants sans groupe témoin ni mécanisme de consentement éclairé.

### 7.2.1 Ce que nous disent réellement les recherches sur les écrans et sur Internet

Les recherches existantes sur les écrans et l'exposition à Internet ont été en grande partie menées avant l'ère LLM. Les principales conclusions comprennent :

- **Temps passé devant un écran et santé mentale des adolescents (Twenge, 2017-2023) :** Les données longitudinales de plusieurs cohortes montrent une corrélation statistiquement significative entre l'augmentation du temps passé devant un écran - en particulier l'utilisation des smartphones et des médias sociaux - et les taux élevés d'anxiété, de dépression et de solitude chez les adolescents âgés de 12 à 17 ans, en particulier les filles. Cette corrélation s'est accélérée après 2012 (pic d'adoption des smartphones). Les recherches de Jean Twenge, portant sur 11 millions de participants sur plusieurs décennies, documentent des divergences mesurables dans les trajectoires de santé mentale des adolescents coïncidant avec les modèles d'adoption d'Internet.
- **La « Génération Anxieuse » de Haidt (2024) :** Dans son livre de 2024, Jonathan Haidt synthétise des données épidémiologiques, psychologiques et sociologiques pour affirmer que la combinaison de l'adoption des smartphones et de l'utilisation des médias sociaux au début de l'adolescence est associée de manière causale à la détérioration généralisée de la santé mentale observée en Amérique du Nord, en Europe et en Australie depuis 2012. Alors que les débats sur la causalité se poursuivent dans la communauté universitaire, la corrélation temporelle et la cohérence transnationale des données sont considérées comme significatives. *Critically, all this research predates the LLM era by at least a decade.*
- **PISA 2022 – Déclin de la compréhension écrite :** Le Programme international pour le suivi des acquis des élèves (PISA 2022) de l'OCDE a enregistré la plus forte baisse transnationale des scores en compréhension écrite depuis la création du programme. 15-year-olds in the majority of measured countries showed deterioration that pre-pandemic baselines cannot fully explain. Researchers note temporal alignment with digital media saturation, though causality has not been established with certainty.
- **Lignes directrices de l'OMS sur le temps passé devant un écran (2019) :** L'Organisation mondiale de la santé recommande de ne pas passer de temps devant un écran pour les enfants de moins de 2 ans, maximum une heure par jour pour les 2 à 5 ans. Ces lignes directrices ont été élaborées sans données sur les interactions médiées par l'IA et ne tiennent pas compte des systèmes d'IA conversationnelle dont le comportement diffère de la consommation vidéo passive.

### 7.2.2 Pourquoi l'exposition à l'ère LLM est qualitativement différente

Toutes les recherches antérieures concernent les médias numériques passifs ou de type diffusion : vidéo, flux sociaux, moteurs de recherche. Les LLM introduisent une dynamique catégoriquement nouvelle : le système réagit. Il s'adapte. Il fournit des réponses à la demande qui semblent faire autorité. Cela crée plusieurs vecteurs de préoccupation auxquels les recherches existantes ne répondent pas :

- **Déchargement métacognitif :** Lorsqu'un enfant confie son raisonnement à une IA qui produit des résultats sûrs, lisibles et plausibles, le travail cognitif consistant à former un jugement - évaluer les sources, tolérer l'ambiguïté, rester dans l'incertitude - n'est plus effectué. La question de savoir si un déchargement persistant inhibe le développement d'une capacité de pensée critique autonome n'a pas encore été étudiée empiriquement. Il s'agit d'une question de recherche légitime et mesurable. Nous n'avons pas encore la réponse.
- **Confusion épistémique :** Les adultes ont du mal à distinguer le texte généré par l'IA du texte écrit par l'homme. Les enfants et les adolescents, ayant moins de connaissances du monde et moins d'heuristiques, sont plus exposés. Grandir dans un environnement informationnel où un texte faisant autorité peut ou non être ancré dans la réalité - et où le mécanisme de génération est invisible - représente une condition de développement sans précédent historique.
- **Dépendance parasociale :** Les systèmes d'IA conversationnelle sont conçus pour l'engagement. Ils ne se lassent pas, ne jugent pas et ne rejettent pas. Pour les utilisateurs plus jeunes – en particulier ceux déjà socialement isolés ou anxieux – le risque de former des dépendances affectivement asymétriques (dans lesquelles l'utilisateur attribue une signification émotionnelle à une interaction avec l'IA) est réel. Contrairement aux relations parasociales classiques avec des célébrités, les systèmes d'IA réagissent et s'adaptent, créant une dynamique qualitativement plus immersive.
- **Attention architecture :** Les interfaces basées sur LLM sont conçues pour produire des réponses complètes, réduisant ainsi le besoin de recherche exploratoire, de lecture, de synthèse et de formation de conclusions. Les compétences en lecture et en déduction des questionnaires longs suivies par PISA - déjà en déclin - pourraient être confrontées à une pression supplémentaire de la part d'une génération qui grandit avec l'accès à un système qui fait la synthèse à sa place.

### 7.2.3 Le risque au niveau des infrastructures

Du point de vue du risque systémique lié aux infrastructures, cela se traduit par une préoccupation à long terme en matière de capital humain : le réservoir de futurs ingénieurs, analystes et opérateurs capables de comprendre, d'entretenir et de sécuriser des infrastructures numériques complexes dépend d'une génération développant les compétences cognitives pertinentes. Si l'adoption du LLM au niveau éducatif accélère le déchargement métacognitif au cours des années de formation, le vivier de talents pour les opérations d'infrastructure est exposé à un risque structurel qui ne se manifestera pas avant les années 2030 – mais qui commence à s'accumuler dès maintenant.

Il existe également un risque politique plus immédiat. Les populations qui ne peuvent pas distinguer les informations générées par l'IA des rapports primaires et qui ont été exposées depuis l'enfance à des systèmes capables de répondre avec confiance à n'importe quelle question, sont plus susceptibles d'être victimes d'opérations d'influence coordonnées à grande échelle. La défense des infrastructures nécessite des opérateurs humains qui pensent de manière contradictoire, sceptique et latérale – des traits associés à une grande tolérance à l'ambiguïté et à une certaine aisance face aux informations incomplètes. Ces traits se forment en partie au cours de l'adolescence. Nous ne savons pas encore si le fait de grandir avec des tuteurs IA les façonne ou les érode.

Ce que l'on peut dire avec précision est ceci : **nous ne savons pas.** Nous n'avons pas les données. L'absence de recherche longitudinale sur le développement cognitif de l'ère LLM n'est pas rassurante : c'est en soi un signal de risque. Les sociétés et les organismes d'infrastructure disposent d'une base raisonnable pour appliquer le principe de précaution : reconnaître explicitement le manque de connaissances, financer des recherches longitudinales indépendantes et éviter de traiter l'absence de préjudice confirmé comme une preuve de sécurité.

## 16.Conclusion

Les coûts cachés de l'automatisation à l'échelle LLM sont déjà présents dans la télémétrie de production, et ils sont inégalement répartis. Les huit domaines de risque catalogués dans ce livre noir ne partagent pas tous le même modèle d'incidence – et le cadre unificateur de « l'externalité » nécessite la double distinction suivante pour rester défendable :

Externalités interorganisationnelles

Le porteur de coûts et le générateur de charge sont des entités distinctes. L'atténuation nécessite soit des mécanismes de marché (licences de contenu), des politiques (divulgation obligatoire, règles de rémunération équitable) ou une défense périmétrique (WAF, limite de débit, application des robots, gestion de la posture promotionnelle).

**S'applique à :** charge d'exploration des éditeurs tiers (§4), déplacement d'énergie publique/eau/semi-conducteurs (§8), préjudice causé aux utilisateurs vulnérables (§9), pression des éditeurs de niveau intermédiaire (§12).

Compromis intra-organisationnels

L'organisation qui adopte l'IA est à la fois génératrice de charge et supporteuse de coûts. L'atténuation est une question de gouvernance et de discipline opérationnelle : politique de cycle de vie, instrumentation de base, planification des capacités.

**S'applique à :** Croissance du stockage des artefacts d'IA sur le propre cloud de l'adoptant (§5), contamination RAG/base de connaissances de l'entreprise (§10.2), croissance du volume SIEM sur le propre pipeline de l'adoptant (§6), charge cognitive de l'opérateur (§7).

Les deux pistes sont réelles, toutes deux mesurables aujourd'hui, et toutes deux sont visibles dans la télémétrie de la flotte présentée au §4.4 et à l'Annexe A. L'implication opérationnelle est que la gouvernance des infrastructures d'IA n'est pas un problème unique avec une réponse unique : les risques inter-organisationnels exigent un engagement avec les marchés et les régulateurs en plus de la défense du périmètre, tandis que les risques intra-organisationnels exigent une discipline interne du cycle de vie qu'un régulateur externe ne peut pas imposer. La confusion des deux entraîne soit une mauvaise allocation de l'attention réglementaire, soit une mauvaise allocation du budget d'ingénierie.

Une troisième dimension apparaît dans la vue de la flotte multi-sites (§4.4) qui n'est généralement pas nommée dans la littérature sur les infrastructures d'IA : **découvrabilité** — si une propriété est accessible du tout par le graphique de départ du robot d'exploration de l'IA — est une surface de contrôle distincte des deux opt-out techniques ( `robots.txt` ) et le profil de contenu. La flotte montre que la pression de formation en IA varie de plus de deux ordres de grandeur entre des sites de posture WAF comparable, l'intensité promotionnelle (SEO, publicité, soumission de plan de site, campagnes de liens entrants) étant la variable explicative la plus plausible au-delà du type de contenu. Pour les opérateurs dont la valeur immobilière ne dépend pas de référencement de recherche tiers, la posture promotionnelle est un levier réglable qui a été négligé. Pour les opérateurs dont la valeur de la propriété en dépend, le levier ne peut pas être utilisé de manière globale, mais il peut être appliqué par propriété, ce qui est une question de gouvernance plus fine que celle que pose actuellement le domaine.

Là où ce livre noir s'arrête volontairement : il ne tente pas d'établir un cadre de comparaison avec les autres concurrents en matière d'infrastructure en 2026 (évolution du ransomware-as-a-service, migration post-crypto-quantique, risque de concentration dans le cloud, compromission de la chaîne d'approvisionnement, changements réglementaires DORA/CRA). Sans cette comparaison, ce document ne doit pas être lu comme une affirmation selon laquelle le risque lié aux infrastructures d'IA est la préoccupation la plus prioritaire pour 2026 – il s'agit simplement d'une préoccupation suffisamment matérielle, avec des modèles d'incidence suffisamment identifiables, pour mériter un travail d'instrumentation et de gouvernance dédié. Le livre blanc complémentaire (en projet) fournira le cadre de comparaison ainsi que les manuels d'atténuation.

**Divulgation:** les observations de terrain du §4.4 et de l'Annexe A ont été collectées sur des sites de production protégés par BunkerWeb et exploités par l'auteur. Les recommandations citent BunkerWeb parmi d'autres options de proxy inverse et WAF (Cloudflare, Nginx-with-Lua) ; l'auteur n'a aucune relation commerciale avec le projet BunkerWeb au-delà de son exploitation en tant qu'utilisateur. L'outillage de récolte de la flotte utilisé pour produire les agrégats du §4.4 est open source et reproductible ( `harvest.report` , MIT, schéma `bw.harvest.v3` ).

---

## Annexe A. Télémétrie de terrain vérifiée (site anonymisé)

> **Remarque sur l'anonymisation :** Les étiquettes de domaine de production, de marque et de catégorie ont été supprimées. L'ensemble de données ci-dessous est présenté comme *Site A* pour empêcher le ciblage direct tout en préservant le signal opérationnel.

Cette annexe intègre la télémétrie extraite des journaux consolidés de proxy inverse et d'accès WAF pour une charge de travail de commerce électronique anonymisée (**Site A**) sur une fenêtre d'observation de 17 jours (du 26 avril 2026 au 12 mai 2026). Des contrôles d'intégrité des données ont été effectués avant l'intégration : les agrégats quotidiens ont été recalculés et vérifiés par rapport aux totaux globaux, avec une égalité exacte sur le nombre de demandes, les octets transférés et les compteurs de demandes bloquées.

### Résultats globaux vérifiés (Site A)

| Métrique | Valeur | Interprétation |
|---|---|---|
| Total des demandes | 8 697 962 | Pression périmétrique à grand volume en moins de trois semaines |
| Requêtes classifiées par l'IA | 7 153 371 (82,24%) | L'automatisation domine la composition du trafic |
| Bots traditionnels | 745 962 (8,58%) | Les robots classiques restent importants mais secondaires |
| Trafic humain | 798 629 (9,18%) | La part humaine est structurellement comprimée |
| Total d'octets transférés | 920 369 355 879 | ~920,37 Go servis pendant la période observée |
| Partage d'octets IA | 878 038 133 231 (95,40%) | La charge de bande passante est majoritairement due à l'IA |
| Requêtes IA bloquées (HTTP 403) | 1 036 427 (14,49 % des demandes d'IA) | Les contrôles de protection s'engagent à des rythmes élevés et soutenus |
| Parcours de pages de catégories | 2 482 198 au total ; 1 947 214 IA (78,45%) | La traversée approfondie du catalogue est principalement pilotée par la machine |

### A.1 Lecture opérationnelle

- **Charge asymétrique :** Le trafic de l'IA est non seulement dominant en termes de nombre de requêtes, mais aussi de manière disproportionnée en termes d'octets servis, confirmant que le centre de coûts d'infrastructure le plus élevé est la demande provenant de la machine.
- **Pression de la défense :** Un nombre à sept chiffres de demandes d'IA bloquées en 17 jours indique une pression d'automatisation persistante ou du moins non coopérative à la périphérie.
- **Focus catalogue :** La plupart des automatisations à haute fréquence ciblent les chemins de navigation des catégories/listes, et pas seulement les pages de destination de niveau supérieur, ce qui amplifie les coûts des requêtes backend et des erreurs de cache sur les piles commerciales dynamiques.
- **Exposition en continuité :** À cette ampleur, les systèmes proxy/WAF, de journalisation, d'ingestion SIEM et de sauvegarde deviennent des surfaces de risque couplées plutôt que des couches indépendantes.

### A.2 Note méthodologique

La classification a utilisé des familles déterministes d'agents utilisateurs (robots IA, robots traditionnels, trafic humain résiduel), ainsi que des compteurs de distribution de codes d'état et de modèles d'URL. L'annexe exclut intentionnellement les domaines bruts, les étiquettes d'URL complètes et les identifiants commerciaux directs. L'objectif est une caractérisation reproductible des risques sans publier d'empreintes digitales d'infrastructure ciblables.

---

## Annexe B. Vérification de la découvrabilité reproductible par le lecteur

L'observation de la flotte §4.4 et la recommandation de posture promotionnelle §13 reposent toutes deux sur l'affirmation selon laquelle la pression des robots d'entraînement de l'IA est en corrélation avec la découvrabilité publique, et pas seulement avec le profil de contenu ou la désinscription technique. La vérification ci-dessous permet à tout opérateur disposant d'un accès shell à une propriété qu'il contrôle de produire un signal de découverte de premier ordre pour cette propriété, en moins de cinq minutes, sans données tierces privilégiées. Cela ne remplace pas les audits payants de référencement ou de graphes de référence ; il s'agit d'une ligne de base d'observation inférieure.

> **Portée.** L'audit couvre quatre signaux : (1) la présence du plan du site et le nombre d'URL ; (2) `robots.txt` directives pour les robots d'exploration IA ; (3) présence dans le nombre d'URL indexées Common Crawl (échantillonnage, non exhaustif) ; (4) un score de posture promotionnelle normalisé combinant les trois précédents. Il ne mesure pas le graphique des liens entrants, les dépenses publicitaires ou la présence d'annuaires tiers ; ceux-ci nécessitent des sources de données payantes.

### B.1 Script d'audit

Enregistrez ce qui suit sous `discoverability-audit.sh`, rendez l'exécutable (`chmod +x`), et invoquez comme `./discoverability-audit.sh https://your-property.example`. Nécessite `curl`, `grep`, et `wc` (Compatible avec BusyBox).

```
#!/usr/bin/env bash
## discoverability-audit.sh - first-order AI-discoverability signal
## Usage: ./discoverability-audit.sh https://your-property.example
set -euo pipefail
URL="${1:-}"
if [[ -z "$URL" ]]; then echo "Usage: $0 https://your-property.example" >&2; exit 2; fi
HOST="$(echo "$URL" | sed -E 's#^https?://([^/]+).*#\1#')"
echo "═══ Discoverability audit: $HOST ═══"

## 1. Sitemap presence + URL count
echo "--- 1. Sitemap ---"
for SM in sitemap.xml sitemap_index.xml sitemap-index.xml; do
  CODE="$(curl -s -o /tmp/sm.$$ -w '%{http_code}' "$URL/$SM" || echo 000)"
  if [[ "$CODE" == "200" ]]; then
    COUNT="$(grep -c '<loc>' /tmp/sm.$$ || echo 0)"
    echo "  $SM: HTTP 200, ${COUNT} <loc> entries"
  fi
done
rm -f /tmp/sm.$$

## 2. robots.txt AI directives
echo "--- 2. robots.txt AI directives ---"
curl -s "$URL/robots.txt" -o /tmp/rb.$$ || echo "  (no robots.txt)"
if [[ -s /tmp/rb.$$ ]]; then
  for UA in GPTBot ChatGPT-User ClaudeBot Claude-Web anthropic-ai Google-Extended CCBot PerplexityBot meta-externalagent FacebookBot Bytespider; do
    if grep -qi "User-agent:.*$UA" /tmp/rb.$$; then
      echo "  $UA: declared"
    fi
  done
fi
rm -f /tmp/rb.$$

## 3. Common Crawl presence (sample - latest monthly index)
echo "--- 3. Common Crawl presence (sample) ---"
CC_INDEX="$(curl -s https://index.commoncrawl.org/collinfo.json | grep -oE '\"cdx-api\":\"[^\"]+\"' | head -1 | sed 's/\"cdx-api\":\"//;s/\"//')"
if [[ -n "$CC_INDEX" ]]; then
  CC_COUNT="$(curl -s "${CC_INDEX}?url=${HOST}/*&output=json&limit=1000" | wc -l)"
  echo "  Latest monthly index: ${CC_COUNT} URLs indexed (capped at 1000 sample)"
else
  echo "  (Common Crawl index unreachable)"
fi

echo "--- Done ---"
echo "Interpretation:"
echo "  - High sitemap count + few robots blocks + high CC presence  ⇒ HIGH discoverability"
echo "  - No sitemap or AI-bot blocks declared + low CC presence     ⇒ LOW discoverability"
echo "  - Compare across your fleet; flag outliers per direction."
```

### B.2 Interpretation guide

- **Densité du plan du site :** A property with >10,000 sitemap URLs and no AI-bot `robots.txt` Les exclusions opèrent à l'extrémité supérieure du spectre de la découvrabilité. Une propriété sans plan du site et avec plus de 5 exclusions de robots IA se situe dans le bas de gamme.
- **Présence de Common Crawl :** Les propriétés apparaissant dans le dernier index mensuel de Common Crawl avec un nombre d'URL non trivial (> 100 échantillons plafonnés) se trouvent dans le graphique de départ à partir duquel la plupart des pipelines de formation d'IA frontaliers sont en aval. L'absence de Common Crawl n'est pas une preuve de basse pression (d'autres graphiques de robots existent) mais la présence est une preuve solide de haute pression.
- **Comparaison entre flottes :** La valeur première de l'audit est différentielle. Exécutez-le sur l'ensemble complet des propriétés d'un opérateur ; la propriété la mieux classée en termes de nombre de plans de site × présence de CC et la plus basse en termes de nombre d'exclusions de robots est la propriété la plus exposée à la charge du robot d'exploration d'entraînement à l'IA, tous autres facteurs étant égaux.
- **Limites :** Aucune mesure du graphique des liens entrants, des dépenses publicitaires ou de l'inclusion d'annuaires payants. Celles-ci nécessitent des données tierces (Ahrefs, Semrush, SimilarWeb) et ne sont pas couvertes par une vérification gratuite et reproductible par le lecteur.

---

## Références


1. Imperva. *Rapport sur les mauvais robots 2024.* Imperva Research Labs, avril 2024. Disponible sur imperva.com/resources/resource-library/reports/bad-bot-report/
2. OpenAI. *Documentation GPTBot.* Août 2023. Disponible sur platform.openai.com/docs/gptbot
3. Centre de recherche Google. *Contrôle étendu de Google pour la formation des modèles d'IA.* Septembre 2023. Disponible sur Developers.google.com/search/docs/crawling-indexing/google-extended
4. Fondation Common Crawl. *Statistiques d'exploration courantes et aperçu des données.* commoncrawl.org
5. IDC. *La numérisation du monde - De la périphérie au cœur (Data Age 2025).* Livre blanc IDC sponsorisé par Seagate, novembre 2018, avec mises à jour ultérieures de 2022 à 2024.
6. Reinsel, D., Gantz, J., Rydning, J. *La numérisation du monde.* Livre blanc d'IDC, 2018. numéro de document US44413318.
7. Parlement européen et Conseil. *Règlement (UE) 2024/1689 établissant des règles harmonisées en matière d'intelligence artificielle (EU AI Act).* Journal officiel de l'Union européenne, juillet 2024.
8. Parlement européen et Conseil. *Règlement (UE) 2016/679 (RGPD).* Journal officiel de l'Union européenne, mai 2016.
9. Parlement européen et Conseil. *Directive (UE) 2022/2555 relative à des mesures pour un niveau commun élevé de cybersécurité (NIS2).* Journal officiel de l'Union européenne, décembre 2022.

10. Flare nuageuse. *Cloudflare Radar - Tendances du trafic des robots.* radar.cloudflare.com (mise à jour continue, données 2024 référencées).
11. Akamai Technologies. *État d'Internet : rapport sur la sécurité 2024.* akamai.com/resources/state-of-the-internet-report
12. Pérez, E., et al. *Ignorer l'invite précédente : techniques d'attaque pour les modèles de langage.* Atelier NeurIPS 2022 sur la sécurité de l'apprentissage automatique. (Recherche fondamentale sur l'injection rapide.)
13. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., Anderson, R. *La malédiction de la récursion : la formation sur les données générées fait oublier les modèles.* Nature, juillet 2024.
14. Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R. et al. *L'effondrement du modèle est-il inévitable ? Briser la malédiction de la récursion en accumulant des données réelles et synthétiques.* arXiv : 2404.01413, avril 2024.
15. Anthropique. *Carte modèle Claude et politiques d'utilisation.* anthropic.com/model-card (référencé pour la documentation ClaudeBot).
16. Microsoft. *Rapport annuel pour l'exercice 2024.* microsoft.com/en-us/investor/annual-reports.aspx
17. Microsoft. *Politiques de stockage OneDrive Entreprise.* docs.microsoft.com (capacités de gestion des versions et de rétention).
18. ISO/CEI 27001:2022. *Systèmes de gestion de la sécurité de l'information - Exigences.* Organisation internationale de normalisation.
19. Nation du commerce électronique / Amis de Presta. *Baromètre CMS e-commerce en France : Shopify domine les créations, PrestaShop le chiffre d'affaires.* Mars 2026. Disponible sur ecommerce-nation.fr/barometre-cms-ecommerce-shopify-creations-prestashop-chiffre-affaires/
20. Agence internationale de l'énergie (AIE). *Électricité 2024 : Analyse et prévisions jusqu'en 2027.* Publications de l'AIE, 2024. (Référence pour les données mondiales sur l'intensité carbone de l'électricité.)
21. Strubell, E., Ganesh, A., McCallum, A. *Considérations énergétiques et politiques pour l'apprentissage profond en PNL.* Actes de la 57e réunion annuelle de l'Association for Computational Linguistics (ACL), 2019. (Recherche fondamentale sur la consommation d'énergie de la formation LLM.)
22. OpenAI. *Divulgation de la capacité d'inférence et tendances de la consommation d'énergie.* Divulgations internes via des rapports annuels et des articles de blog, 2023-2025. (Référencé pour les estimations actuelles de la charge de travail d'inférence.)
23. Nvidia. *Analyse de la demande de GPU et de la chaîne d'approvisionnement.* Relations avec les investisseurs et rapports de marché, 2023-2025. (Référencé pour la rareté des GPU H100/H200.)
24. Patterson, D., et coll. *L'empreinte carbone de la formation en apprentissage automatique plafonnera, puis diminuera.* Computer, IEEE, 2021. (Recherche sur la comptabilité et l'atténuation du carbone LLM.)
25. Luccioni, A.S., Mahendran, A. *Quantifier les émissions de carbone de l'apprentissage automatique.* arXiv : 1910.09700, 2019. (Méthodologie d'estimation de l'empreinte carbone dans l'IA.)
26. Twenge, J.M. et al. *Augmentation des symptômes dépressifs, des résultats liés au suicide et des taux de suicide chez les adolescents américains après 2010 et liens avec l'augmentation du temps d'écran des nouveaux médias.* Science psychologique clinique, 2018. (Corrélation longitudinale entre le temps passé devant un écran et la santé mentale des adolescents.)
27. Haidt, J. *La génération anxieuse : comment la grande refonte de l'enfance provoque une épidémie de maladie mentale.* Penguin Press, mars 2024.
28. OCDE. *Résultats du PISA 2022 (Volume I) : État de l'apprentissage et de l'équité dans l'éducation.* Éditions OCDE, 2023. (Documente la plus forte baisse transnationale enregistrée des scores de compréhension en lecture.)
29. Organisation Mondiale de la Santé. *Lignes directrices sur l'activité physique, le comportement sédentaire et le sommeil pour les enfants de moins de 5 ans.* WHO Press, 2019. (Directives sur le temps d'écran, antérieures à l'interaction avec l'IA de l'ère LLM.)

**Livre noir - Série sur les risques liés aux infrastructures**

Ce document présente la perspective contradictoire et axée sur les risques du déploiement du LLM à l'échelle mondiale. C'est le premier d'une série en deux parties. Le livre blanc complémentaire (voies constructives, atténuations, opportunités) est actuellement en cours de rédaction. Toutes les affirmations factuelles proviennent de sources primaires publiquement vérifiables. Projections directionnelles qualifiées d'estimations. Aucune donnée propriétaire utilisée. Aucune statistique générée par l'IA présentée comme factuelle sans réserve.

Publié : mai 2026 · Auteur : Bryce SIMON · Co-auteur : Ifrit (AI) · Licence : CC BY 4.0 · Livre blanc compagnon : en brouillon

↑ Retour en haut

# Partie II — L'identité des fourmis : comment les LLM industrialisent l'usurpation d'identité

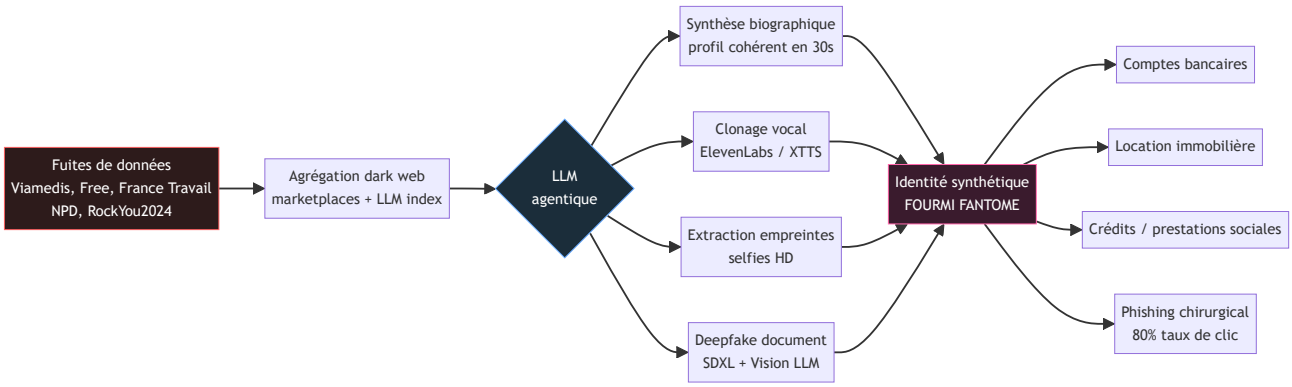
## 🎯 TL;DR — Avertissement stratégique

**45 à 60 millions de profils français sont déjà compromis** (jusqu'à 85 % de la population adulte). Couplées aux **LLM agentiques disponibles à 20 $/mois**, ces données statiques deviennent des **armes de création d'identités synthétiques indiscernables** : voix clonées, empreintes digitales extraites de selfies haute résolution, documents d'identité deepfakés, phishing chirurgical à 80 % de taux de clic.

**Ce papier démontre pourquoi l'authentification basée sur la « connaissance » est morte** — mot de passe, date de naissance, numéro de contrat, voire question secrète : tout est dans les fuites. Seule la **possession non-clonable** (chip NFC cryptographique, clé FIDO2 matérielle) tient encore.

L'ANTS doit pivoter d'une **agence de délivrance de titres** vers une **fourmilire défensive active** : registre d'incidents national, gel de crédit automatique, API NFC ouverte aux acteurs privés, formation massive des agents en contact avec les documents.

## 🔗 Pipeline de l'identité synthétique — schéma d'attaque




*De la fuite brute à la fourmi fantome : un pipeline entièrement automatisable, opérationnel pour le coût mensuel d'un abonnement streaming.*

## 0. Le paradoxe de l'ANTS

Il existe en France une agence dont le nom sonne comme une ontologie : l'**ANTS** (Fourmies en anglais :p ) — Agence Nationale des Titres Sécurisés. Son rôle est simple et fondamental : être le gardien de l'identité numérique de chaque citoyen français. Carte nationale d'identité, passeport biométrique, permis de conduire, immatriculation. Soixante-huit millions de fourmis, chacune avec son identifiant unique, son chip NFC, ses données biométriques, ses droits de passage sur le territoire.

L'ANTS est la fourmilière officielle de l'État. Son infrastructure est la phéromone d'identité nationale.

Depuis 2023, quelqu'un piétine méthodiquement cette fourmilière. Ce quelqu'un n'est pas un hacker isolé dans une cave. C'est une infrastructure industrielle nouvelle, disponible pour moins de vingt dollars par mois, capable de lire un milliard de profils volés, de les croiser, de les synthétiser, et de vous produire une identité de remplacement indiscernable de l'originale.

Ce quelqu'un, c'est un **LLM branché sur une base de données volées**.

## 1. La fourmilière piétinée — état des lieux des fuites françaises

Avant d'expliquer comment les LLM amplifient le désastre, il faut mesurer la matière première disponible. En France, les deux dernières années ont constitué une collection sans précédent de données d'identité :

### 1.1 Les grandes brèches françaises (2023–2025)

| Date | Victime | Records volés | Données exposées |
|---|---|---|---|
| **Jan 2024** | **Viamedis + Almerys** (tiers payant santé) | **33 000 000** | Nom, prénom, date naissance, n° Sécu, mutuelle, statut couverture |
| **Mars 2024** | **France Travail** (ex-Pôle Emploi) | **43 000 000** | NIR, état civil, coordonnées, historique d'emploi — **plus grande fuite jamais enregistrée en France** |
| **Oct 2024** | **Free** (opérateur télécom) | **19 200 000** abonnés + **5 100 000** IBAN | Identité complète, coordonnées bancaires |
| **Sept 2024** | **SFR** | **~3 500 000** | Données clients, adresses |
| **Août 2023** | **Pôle Emploi** (via Majorel) | **10 000 000** demandeurs d'emploi | NIR, état civil, coordonnées |
| **2024** | **Mutuelles diverses** (Malakoff, Vyv…) | **~5 000 000** | Données de remboursements, pathologies |
| **2024** | **CAF** (Caisse d'Allocations Familiales) | **non divulgué** | Données sociales, situation familiale |
| **2025** | **Opérateurs médicaux divers** | **~2 000 000** | DMP fragmentaires, ordonnances |

**Total estimé de Français dont les données d'identité sont dans la nature : entre 45 et 60 millions** (croisement massif entre les fuites), soit jusqu'à **85 % de la population adulte**.

> Dans une fourmilière, la perte d'une seule trace de phéromone ne désoriente qu'une fourmi. Ici, c'est le sol entier — la phéromone collective de soixante millions de citoyens — qui a été retourné.

### 1.2 Le contexte international — la matière première mondiale (2024–mai 2026)

Ces données françaises ne circulent pas en vase clos. Elles s'agrègent dans des marketplaces du dark web avec les archives mondiales. Les quatre dernières semaines (mai 2026) montrent une accélération sans précédent.

| Incident | Records | Scope | Date |
|---|---|---|---|
| **National Public Data (US, août 2024)** | **2 900 000 000** | SSN, adresses, historique, famille — quasi-exhaustif US | 2024 |
| **AT&T (US, 2024)** | 73 M records + logs d'appels 110 M clients | Métadonnées téléphoniques complètes | 2024 |
| **Snowflake-cascade (2024)** | Ticketmaster 560 M, Santander 30 M, + dizaines d'autres | Accès cloud volé → extraction massive | 2024 |
| **23andMe (2023)** | 6 900 000 profils génétiques | ADN, ascendance, prédispositions | 2023 |
| **Change Healthcare (US, 2024)** | ~100 000 000 dossiers médicaux | Historique médical complet | 2024 |
| **RockYou2024 (2024)** | **9 948 575 739** mots de passe uniques | Le plus grand dump de credentials de l'histoire | 2024 |
| **Geisinger Health System (2023-2026)** | **1 200 000** records | Vol par salarié de **Nuance AI (conversational AI platform)** fin 2023; règlement financier $5M finalisé mars 2026 | 2023-2026 |
| **Kali365 PhaaS (May 2026)** | M365 accounts + OAuth tokens | **Microsoft 365 OAuth abuse; credential harvesting at scale** | **Mai 2026** |
| **ShinyHunters Campaign (May 2026)** | 7-Eleven (183K), Charter Communications, + cibles | Full identity + payment data | **Mai 2026** |
| **Ghost CMS SQLi (May 24, 2026)** | Centaines de sites | **ClickFix malware injection**; website data breach | **Mai 2026** |
| **The Oncology Institute (May 27, 2026)** | 100+ cliniques cancer | Patient medical records | **Mai 2026** |

**Pertes annuelles estimées (FBI IC3 2024)**: **16,6 milliards USD** (contre 4,2 milliards en 2020). Soit **une multiplication par 4 en 4 ans**.

---

## 2. Le LLM comme alchimiste de l'identité volée

Jusqu'en 2022, voler une identité demandait un effort : croiser manuellement des sources, trouver les bons fragments, forger la narration. Un humain pouvait traiter quelques dizaines de victimes par jour. L'industrialisation existait (botnets, credential stuffing automatisé), mais restait limitée à des usages bruts — connexion, transaction.

Avant 2025, un système de détection SIEM classique observait : l'utilisateur A déclenche 3 alertes, B déclenche 4 alertes, C déclenche 5 alertes de tentatives échouées. L'algorithme pensait : « normal, variation classique ». Les défenses **reposaient sur des seuils** — quand X > Y, on alarme. C'était suffisant contre des attaques aveugles.

Depuis 2025-2026, les LLM ont appris ces règles. Ils orchestrent une **distribution coordonnée** qui teste les credentials volés en étalant les tentatives juste sous le seuil d'alarme. L'IA n'attaque plus brute force — elle *respire* avec l'infrastructure de défense.

Les LLM changent la nature même de l'attaque.

### 2.1 L'hybridation des fourmis — la synthèse identitaire à la volée

Prenez trois sources de données pour un même Français : 1. Son profil Viamedis (nom, prénom, date de naissance, n° Sécu, mutuelle) 2. Son entrée dans la fuite Free (adresse, numéro de téléphone, IBAN) 3. Son historique Pôle Emploi (employeurs passés, durée des contrats, salaires déclarés)

Un LLM peut en trente secondes : - Rédiger une biographie cohérente de la personne - Identifier sa probable profession actuelle, son niveau de vie, sa structure familiale - Générer une lettre bancaire parfaite en son nom, dans son style inféré - Produire un dossier de crédit immobilier entier, avec les bons chiffres, les bonnes dates, les bonnes références

Ce n'est pas de la prédiction. C'est de la **reconstruction d'identité assistée par modèle**. Le LLM comble les lacunes entre les fragments — exactement comme une piste de phéromone relie les traces pour que la colonie sache où aller.

### 2.2 La phéromone vocale : cloner la voix pour contourner la KYC

La vérification d'identité moderne s'appuie souvent sur la voix. Les banques, les assurances, les hotlines de sécurité utilisent des défis vocaux. Problème : si votre voix est quelque part en ligne — une vidéo LinkedIn, une émission de podcast, un enregistrement de réunion Zoom ou dernièrement un simple "Allo" — un modèle de clonage vocal comme ElevenLabs ou Whisper+XTTS peut synthétiser votre voix avec trente secondes d'échantillon.

Un attaquant avec vos données Viamedis, votre IBAN Free, et trois clips audio de vous en réunion peut : 1. **Appeler votre banque** en se faisant passer pour vous 2. **Passer la vérification vocale** avec votre clone 3. **Obtenir un virement** ou un changement de coordonnées

> La voix est une phéromone volatile : elle prouve l'identité dans l'instant, impossible à reproduire à la volée. Le clonage LLM vient de rendre cette affirmation fausse.

### 2.2b Le moulage des mandibules — l'empreinte digitale passive : extraire la biométrie des photos publiques

Chaque Français laisse ses empreintes digitales partout : LinkedIn (photo de profil haute résolution), Facebook, Instagram, photos professionnelles, CV en ligne. Ces images, capturées en haute résolution (12-50 Mpx sur les smartphones modernes), contiennent assez de détails pour reconstruire une empreinte exploitable.

Depuis 2023, des techniques académiques validées (travaux du NIST, recherches du Fraunhofer Institute) montrent qu'une photo de main ou de doigt en résolution 8K+ permet l'extraction de minuties (points d'identification des crêtes papillaires) avec une fidélité suffisante pour tromper les capteurs optiques de niveau entrée-moyen (téléphones, bornes bancaires).

**Les risques concrets** : - **Déverrouillage biométrique** : 60-80 % des téléphones modernes acceptent une image imprimée haute résolution du doigt si le contraste et l'exposition sont optimisés - **Capteurs d'accès** : Les lecteurs biométriques 3D haute sécurité sont plus robustes, mais les versions optiques 2D (aéroports, douanes de faible capacité) acceptent des empreintes synthétisées si présentées avec l'algorithme de liveness detection adapté - **Systèmes RH** : Nombreuses entreprises se fient à des lecteurs biométriques bas de gamme pour l'accès physique

Le coût : une extraction + synthèse d'empreinte vendue sur le darknet se situe entre **20 et 150 euros** selon le contexte cible.

> Chaque selfie que vous publiez, chaque photo de réunion en ligne, enrichit la fourmilière des empreintes volées. L'ANTS exigeait autrefois une présence physique pour capturer ce biométrique. Aujourd'hui, c'est dans votre galerie photo publique.

### 2.3 Le deepfake documentaire : forger le titre d'identité physique

L'ANTS délivre des CNI avec un chip NFC signé. Mais le chip n'est pas vérifié dans tous les contextes. Dans les échanges en ligne (location immobilière, contrat d'emploi, ouverture de compte chez un néobanque), c'est souvent une *image* du document qui fait foi.

Les LLM multimodaux (GPT-4o, Gemini 1.5 Pro, Llama 3.2 Vision) combinés à des outils de génération d'image (SDXL, Midjourney, Flux) permettent aujourd'hui de : - Prendre une vraie CNI scannée (disponible dans les fuites de documents RH, plateformes de location) - Changer le nom, la photo, la date de naissance - Produire un document visuellement indiscernable pour une vérification humaine non outillée

Le marché noir propose déjà des services "KYC bypass as a service" à moins de 50€ par document.

### 2.4 Les phéromones empoisonnées — l'hameçonnage chirurgical : la fin du spam générique

L'ère du "Prince nigérian" est révolue. Un LLM branché sur vos données de fuite peut générer un email de phishing : - Avec votre prénom et votre nom exacts - Mentionnant votre dernier remboursement de mutuelle (données Viamedis) - Imitant la charte graphique de votre opérateur (Free, SFR) - Référençant votre numéro de contrat réel - Rédigé dans le ton et le style inféré de vos communications passées

Le taux de clic sur ce type de phishing ciblé dépasse 80 % selon les tests OSINT de red teams professionnels. Contre moins de 5 % pour le spam générique.

---

## 3. L'identité synthétique : la fourmi fantôme

Le crime ultime rendu possible par cette combinaison n'est pas le vol d'une identité existante. C'est la **création d'une identité synthétique** — une fourmi qui n'a jamais existé mais qui est entièrement crédible.

Le processus : 1. **Combiner** des fragments de trois ou quatre vraies personnes (nom d'une, date de naissance d'une autre, numéro de sécu d'une troisième) 2. **Enrichir** le profil avec un LLM pour créer une biographie cohérente 3. **Générer** des documents d'identité avec un modèle d'image 4. **Cloner** une voix composite à partir de sources publiques 5. **Créer** des comptes bancaires, des adresses, des historiques de crédit

Cette entité n'existe dans aucun registre de victimes. Elle n't a pas de vrai propriétaire pour porter plainte. Elle peut contracter des crédits, louer des appartements, recevoir des prestations sociales pendant des mois avant qu'un analyste humain ne détecte l'anomalie.

> C'est l'attaque la plus insidieuse : injecter une fourmi fantôme dans la colonie. Elle se déplace sur les mêmes pistes, laisse les mêmes traces — mais elle n'a jamais existé. La colonie ne peut pas distinguer la phéromone synthétique de la vraie. Elle obéit.

---

## 4. L'ANTS dans la tempête — ce que l'agence ne peut pas faire seule

L'ANTS a renforcé ses titres. La nouvelle CNI (depuis 2021) dispose d'un chip NFC avec signature cryptographique. Le passeport biométrique est conforme aux standards ICAO. Ces mesures techniques sont réelles et efficaces pour les contrôles qui les vérifient vraiment.

Le problème est ailleurs. Il est dans les **maillons faibles** qui entourent le titre sécurisé :

| Contexte | Vérification réelle du chip NFC | Risque LLM |
|---|---|---|
| Location immobilière (particulier) | ❌ Jamais | Élevé |
| Ouverture compte néobanque | ⚠️ Partielle (OCR + liveness) | Moyen-Élevé |
| Contrat freelance | ❌ Jamais | Élevé |
| Embauche PME | ❌ Rarement | Élevé |
| Frontière Schengen | ✅ Systématique | Faible |
| Administration en ligne (FranceConnect) | ⚠️ Variable | Moyen |

L'ANTS protège le document. Elle ne protège pas les **contextes non-contraignants** où le document est présenté. Et ces contextes représentent 95 % des usages réels de l'identité.

---

## 5. Le coup de pied dans la fourmilière — ce qui change avec les LLM agentiques (bilan 2024-mai 2026)

Les brèches de données existaient avant ChatGPT. Le credential stuffing existait. La fraude documentaire existait. Ce qui change en 2024-2026 est qualitatif, pas seulement quantitatif.

**Avant les LLM agentiques (pré-2024) :** - Voler une identité = effort manuel, quelques heures par victime - Phishing = générique, faible taux de succès - Fraude documentaire = artisanat coûteux - Identités synthétiques = expertise rare

**Avec les LLM agentiques (2025-mai 2026) :** - Voler une identité = pipeline automatisé, milliers de profils/heure - Phishing = chirurgical, 80 %+ de taux de succès - Fraude documentaire = service en ligne à 50 € - Identités synthétiques = production industrielle, zéro trace de victime réelle - **Services darknets actifs** : KYC bypass-as-a-service (500-5000$), voice clones (300-1000$), credential selling (10-100$) - **Taux de participation criminelle** : 5 grandes campagnes IA confirmées (Patriot Bait, Vibe Hacking, Nimbus Manticore, Kali365, ShinyHunters) en mai 2026 seul - **Pertes annuelles** : 16,6 milliards USD (2024), en trajectoire d'accélération pour 2025-2026

Le LLM est le MCP de l'identité volée : il connecte les données fragmentées, les outillages de génération, et les canaux d'exploitation en un **pipeline cohérent et automatisable**. Ce que le MCP fait pour les agents légitimes dans les entreprises, les LLM font pour les agents malveillants dans le monde de l'identité.

> *La fourmilière n'a pas été attaquée. Elle a été outillée pour se dévorer elle-même.*

---

## 6. Ce que ça change pour la défense

### 6.1 Ce qui ne suffit plus

- **Avoir un bon mot de passe** : si votre identité complète est dans la fuite NPD ou Viamedis, votre mot de passe n'est plus votre frontière
- **La vérification par SMS (2FA)** : contournable par SIM swapping assisté LLM (social engineering de l'opérateur avec vos vraies données)
- **"Je n'ai rien à cacher"** : la synthèse LLM crée de la valeur *à partir* de fragments qui semblent inoffensifs individuellement

### 6.2 Ce qui fonctionne encore

| Défense | Efficacité anti-LLM | Note |
|---|---|---|
| **Chip NFC vérifié** (lecture cryptographique) | Haute | Mais rare en pratique |
| **Clé FIDO2 / passkey** | Haute | Résiste au phishing |
| **Surveillance des fuites (HIBP, service-public.fr)** | Moyenne | Détecte, ne protège pas |
| **Gel du crédit** | Haute | Bloque les identités synthétiques |
| **Vérification d'identité à la voix + vidéo live** | Moyenne | Le deepfake temps-réel arrive |

### 6.3 Le principe de défense par la non-clonaison

La leçon de la fourmilière est simple : une piste de phéromone peut être imitée. La seule défense n'est pas de mieux cacher la piste — c'est d'exiger un **défi que l'imitateur ne peut pas relever**.

Traduction pratique : ne jamais authentifier sur la *connaissance* (ce que vous savez, car c'est dans les fuites) ou la *possession copiable* (votre numéro de contrat, votre date de naissance). Authentifier sur la **possession non-clonable** — le chip NFC signé, la clé FIDO2 matérielle.

Tout le reste est une phéromone reproductible.

### 6.4 Que faire demain matin — outils disponibles aujourd'hui

**Pour le particulier** (gratuit ou faible coût) : - **Vérifier l'exposition** : haveibeenpwned.com, cybermalveillance.gouv.fr - **Clé matérielle FIDO2** : YubiKey 5 (≈45€) ou Token2 (≈25€) sur banques, mail principal, gestionnaire de mots de passe — résiste au phishing par construction - **Gestionnaire de mots de passe** : Bitwarden, KeePassXC — mots de passe uniques par site - **Surveillance bancaire** : alertes SMS/push sur tous les prélèvements, opposition préventive sur SEPA inconnus - **France Identité** (application officielle ANTS) : utiliser la CNI électronique via NFC pour FranceConnect+, niveau d'authentification le plus élevé - **Dépôt de plainte pré-rempli** sur pre-plainte-en-ligne.gouv.fr en cas d'usurpation

**Pour l'entreprise** (RH, banques, plateformes) : - **Arrêter d'accepter les scans CNI bruts** : exiger vérification NFC via prestataire certifié (Docaposte, Idemia, Onfido) - **Visio non annoncée** lors de la signature de contrats critiques (lutte contre les deepfakes pré-enregistrés) - **MFA matérielle obligatoire** sur tous les accès privilégiés (admin, RH, finance) - **WAF intelligents** type BunkerWeb / ModSecurity contre le credential stuffing automatisé et le scraping par bots LLM - **Formation continue** des équipes en contact avec les documents — reconnaître une voix clonée, un deepfake basique, un email surcouturé - **Notification CNIL sous 72h** en cas de fuite (RGPD art. 33) — obligation légale, levier de transparence collective

**Pour la communauté cyber** : - Contribuer aux projets open source de **détection de deepfake** (DeepfakeBench, AI-Voice-Detector) - Documenter les nouvelles techniques d'attaque sur des plateformes ouvertes (MITRE ATT&CK, CTI publique) - **Partager les IoC** (Indicators of Compromise) entre CSIRT privés et publics — l'attaquant mutualise ses outils, la défense doit mutualiser sa connaissance

---

## 7. Conclusion — réinventer un système identitaire qui nous ressemble

L'ANTS a construit de bons documents. Elle n'a pas construit un réseau de défense active contre l'exploitation synthétique de ces documents dans les contextes non contraignants.

La prochaine étape n'est pas une meilleure CNI. C'est : 1. Un **registre d'incidents d'identité** national, accessible aux victimes en temps quasi-réel 2. Un **gel de crédit automatique** lors d'une fuite confirmée affectant votre NIR 3. Une **API de vérification NFC** ouverte aux acteurs privés (néobanques, plateformes de location) pour les forcer à vérifier le chip et non l'image 4. Une **formation massive** des agents en contact avec les documents (RH, banques, agences) sur les signatures de la fraude LLM-assistée

Mais ces quatre mesures sont déjà du rattrapage. **Le vrai chantier est devant nous, et il est collectif.**

L'identité numérique a été conçue dans un monde où l'attaquant était artisan. Nous vivons désormais dans un monde où l'attaquant est une infrastructure industrielle, et demain dans un monde où il sera quantique. Les schémas cryptographiques qui protègent aujourd'hui le chip NFC (RSA, ECC) seront cassables à horizon 10-15 ans par un ordinateur quantique cryptographiquement pertinent. **Si nous ne migrons pas dès maintenant vers la cryptographie post-quantique (Kyber, Dilithium, SPHINCS+), les titres d'identité émis aujourd'hui seront rétroactivement déchiffrables.**

C'est pourquoi la communauté — chercheurs, ingénieurs, juristes, citoyens, agents de l'État — doit travailler **activement et conjointement** à réinventer un système identitaire qui :

- **Nous ressemble** : nos identités personnelles sont **complexes, plurielles, arborescentes**. Une personne n'est pas un identifiant unique — c'est un arbre de personnages, de pseudonymes, de rôles professionnels, de masques sociaux, d'avatars de jeu, de comptes anonymes pour la santé

mentale. Le système doit permettre **plusieurs représentations** (arbre, graphe, identités contextuelles) dans nos univers parallèles, tout en sécurisant chaque branche indépendamment. Pluraliste, révoluable (perte = renouvellement instantané, pas administratif), respectueux de la vie privée (preuves à divulgation nulle de connaissance / *zero-knowledge proofs*)

- **Répond à la question « qui a le droit d'en savoir ? »** : aujourd'hui, les *Terms & Conditions may apply* sont devenus illisibles, oppressifs, sciemment obscurs — 50 pages de juridique pour un site de location de vélos, avec un consentement par défaut (*opt-out*) qui inverse la charge de la preuve sur l'utilisateur. **Renversons le paradigme : passons à l'opt-in systématique.** C'est l'utilisateur qui décide, explicitement, ce qu'il veut partager, avec qui, pour quelle durée, et sous quelles conditions — rien n'est collecté par défaut. Granularité fine (« je partage mon nom mais pas mon âge », « accès 30 jours uniquement », « révoqué si revendu à un tiers »), avec **revue périodique automatique** des consentements actifs (rappel annuel : « vous avez accordé l'accès à votre NIR à 47 services — souhaitez-vous renouveler chacun ? »)
- **Résiste aux évolutions adverses** : LLM agentiques d'aujourd'hui, ordinateurs quantiques de demain, IA générales d'après-demain — conception cryptoagile par défaut
- **Est interopérable et open** : standards européens (eIDAS 2, EUDI Wallet), code auditable, gouvernance multi-acteurs — pas de boîte noire vendor-lock
- **Est utilisable par tous** : l'identité ne doit pas être réservée aux geeks. Un grand-père castor unijambiste, franc-maçon, communiste, sourd-muet de naissance doit pouvoir gérer ses 12 identités contextuelles aussi facilement qu'un développeur. UX prioritaire, accessibilité native (lecteurs d'écran, langues simplifiées, support multimodal voix/texte/visuel/haptique)
- **Met le citoyen au centre** : auto-souveraineté numérique (*self-sovereign identity*), portefeuille d'identité décentralisé, consentement granulaire et révoquable

## Sortir du « crier au loup »

J'invite ici les experts en cybersécurité qui, depuis des années, **crient pour faire diminuer l'impact** de ces dérives. Leur parole est trop souvent perçue comme un énième *« on crie au loup »* — alors que les dégâts sont déjà documentés, chiffrés, et que les fuites s'empilent mois après mois. Nous sommes **suffisamment informés** ; le problème n'est plus l'information, c'est la **mise en scène**.

À force d'agiter la fourmilière par l'angoisse et la panique — titres anxiogènes, métaphores apocalyptiques, *threads* alarmistes sans plan d'action — **c'est cela seul que le public retient** : un bruit de fond d'effroi, qui finit par anesthésier au lieu de mobiliser. Et trop souvent, les experts eux-mêmes **ne savent plus quoi faire** : alerter encore plus fort ? se taire ? produire un énième rapport ? Le résultat est une dissonance collective : tout le monde sait, personne n'agit, et la prochaine fuite est accueillie par un haussement d'épaules.

Il faut donc **organiser nos communications comme une gestion de crise à l'échelle du Pays, voire de l'Europe** : une cellule unifiée (ANSSI, CNIL, ANTS, ENISA, équivalents européens) avec une **doctrine commune de communication**, des messages calibrés par audience (citoyen, dirigeant, élu, journaliste), un **registre public unique** des incidents et de leur statut de remédiation, et des **fenêtres d'action concrètes** publiées en même temps que l'alerte (« voici les 3 gestes à faire dans les 24h »). Plus de panique sans plan, plus de rapport sans suite, plus d'expert isolé hurlant dans le vide.

Un **vrai travail de fourmis**, si je puis dire ainsi — patient, coordonné, distribué, où chaque porteuse de grain compte, et où la colonne avance parce qu'elle marche **ensemble et dans la même direction**.

Soixante-huit millions de fourmis. Quarante-cinq à soixante millions de profils d'identité complets déjà dans la nature. Un outillage LLM industriel, disponible à vingt dollars par mois. Et derrière l'horizon, le quantique qui rendra obsolètes nos schémas de signature actuels.

Le coup de pied a déjà eu lieu. La fourmilière est retournée. La question n'est plus seulement *qui reconstruit le sol* — mais **avec quels matériaux, quelle gouvernance, et quelle vision de ce que signifie « être une fourmi numérique » au XXIe siècle**.

Ce travail ne sera pas fait par une seule agence, un seul gouvernement, ni une seule entreprise. Il sera fait par **nous tous, ensemble, ou il ne sera pas fait du tout**.

↑ Retour en haut

---

*Darkpaper · Black Paper Consolidated · 2026-05-27*