RÉPUBLIQUE
FRANÇAISE
Liberté
Égalité
Fraternité

Date : 4 février 2026
Nombre de pages : 12

L'IA GÉNÉRATIVE FACE AUX ATTAQUES INFORMATIQUES

SYNTHÈSE DE LA MENACE EN 2025

TLP:CLEAR

Table des matières

AVANT-PROPOS

Cette synthèse traite exclusivement des IA génératives c'est-à-dire des systèmes générant des contenus (texte, images, vidéos, codes informatiques, etc.) à partir de modèles entraînés sur des corpus d'apprentissage [1]. Cette catégorie inclut les « grands modèles de langage » (GML) ou *Large Language Model (LLM)* [1], qui illustrent les enjeux de l'usage dual des systèmes d'IA [2]. Ils sont en effet à la fois utilisés par les défenseurs pour améliorer la cybersécurité et détournés par les attaquants pour améliorer leurs attaques [2].

Cette synthèse propose un état des lieux de la menace que peut représenter aujourd'hui l'IA générative et des menaces qui pèsent sur celle-ci. Toutefois, l'évolution rapide des usages par les attaquants et les organisations appelle à une réévaluation régulière de la menace.

Le guide ANSSI « Recommandations de sécurité pour un système d'IA générative » contient des recommandations de sécurité pour la mise en œuvre de solutions d'IA générative reposant sur des LLM au sein d'entités publiques et privées.

1. Les grands modèles de langage (GML) ou en anglais *Large Language Model (LLM)* constituent un type de modèle d'IA spécialisé dans l'analyse et la génération de contenus textuels, notamment du langage naturel ou du code informatique.

2. Ce concept renvoie à l'idée que tout outil suffisamment puissant pour construire un système complexe peut également être détourné pour le détruire [2].

1 L'UTILISATION DE L'INTELLIGENCE ARTIFICIELLE DANS LES ATTAQUES INFORMATIQUES

A ce jour, l'ANSSI n'a pas connaissance de cyberattaques menées contre des acteurs français à l'aide de l'intelligence artificielle (IA)[3] ou identifié en propre de système d'IA capable de réaliser de manière autonome l'intégralité des étapes d'une attaque informatique. Il est cependant plausible que ces technologies continuent d'être utilisées par divers profils d'attaquants et leur permettent d'améliorer significativement le niveau, la quantité, la diversité et l'efficacité de leurs attaques, particulièrement sur les environnements peu sécurisés [3].

1.1 L'utilisation de services d'IA générative comme facilitateurs d'attaques informatiques

En tant que service numérique innovant, performant et flexible[4], l'IA générative a été progressivement intégrée à l'éventail d'outils et de services auxquels sont susceptibles de recourir des attaquants informatiques.

Les modèles d'IA générative sont ainsi utilisés par divers profils d'attaquants tout au long de la chaîne d'attaque :

- **Conception de contenus à des fins d'ingénierie sociale et de reconnaissance** : les opérateurs de modes opératoires d'attaque (MOA) réputés liés à l'Iran auraient utilisé l'IA générative **Gemini** de GOOGLE à des fins de reconnaissance à l'encontre d'experts et d'organisations d'intérêt [4]. En 2024, les opérateurs du MOA Charcoal Typhoon, réputé lié à la Chine, auraient utilisé des services d'IA générative afin de générer du contenu d'hameçonnage[5][5]. Entre 2024 et 2025, les opérateurs du MOA Lazarus réputé lié à la Corée du Nord auraient également eu recours à des services d'IA générative afin de créer de faux profils d'entreprises et d'employés sur les réseaux sociaux [6]. Lors de ses investigations, l'ANSSI a par ailleurs pu observer à plusieurs reprises des sites Internet semblant avoir été générés par des systèmes d'IA générative. Ces sites à l'apparence légitime servent à héberger des charges malveillantes ou à effectuer de la caractérisation[6]. Enfin, de nombreux cybercriminels exploitent pour quelques dizaines de dollars des services de *deepfakes*[7] à des fins d'usurpation d'identités [7].
- **Développement de codes malveillants :** de nombreux groupes cybercriminels utilisent l'IA générative pour développer leur arsenal offensif. En 2024, le MOA TA547 aurait utilisé un script powershell généré par un LLM pour compromettre une entreprise allemande [8]. Un collectif de chercheurs de l'université de New-York a par ailleurs mis au point un

3. Un système d'IA (SIA) peut être défini comme un ensemble de composants matériels et logiciels avec la particularité d'avoir au moins un de leurs composants qui implémente un modèle issu d'un processus d'apprentissage statistique. L'apprentissage statistique ou automatique désigne l'application d'un algorithme d'apprentissage à des données d'entraînement pour produire un modèle. Un système d'IA peut être intégré ou interconnecté à un système d'information plus large.

4. Les modèles d'IA générative peuvent être utilisés pour répondre à tout type de questions et de tâches relatives à la plupart des environnements informatiques, des protocoles et des systèmes de défense.

5. Selon MICROSOFT, les opérateurs de Charcoal Typhoon auraient déjà ciblé Taïwan, la Thaïlande, la Mongolie, le Népal et la France.

6. La caractérisation ou *profiling* consiste à récupérer des données techniques des internautes ayant consulté la page et ainsi identifier des cibles avant de les compromettre.

7. Enregistrement vidéo ou audio réalisé ou modifié grâce à l'intelligence artificielle.

prototype de rançongiciel, **PromptLock**. Le code a comme particularité d'utiliser de manière dynamique des *prompts*[8] pour générer des scripts à l'exécution permettant d'exfiltrer et de chiffrer les données [9]. Enfin, Google a identifié une famille de codes malveillants, dont **Promptflux**, un code malveillant polymorphique qui inclut une fonction qui prompte l'API **Gemini** pour réécrire entièrement son code source toutes les heures afin d'éviter la détection [10]. Le développement de tels codes malveillants suggère cependant des capacités relativement sophistiquées des développeurs.

- **Identification d'informations d'intérêt avant et après exfiltration de données** : en février 2025, le département de cyberdéfense ukrainien a affirmé que des opérateurs russes auraient utilisé des services d'IA générative pour analyser massivement les données exfiltrées de ses victimes et en identifier les informations d'intérêt [11].

L'utilisation de l'IA générative dans la mise en œuvre de certaines étapes de la chaîne d'infection comme la recherche de vulnérabilités est plus complexe [12]. L'identification d'une vulnérabilité et le développement de la preuve de concept associée dépendent encore de compétences humaines. La plupart des systèmes d'IA générative commerciaux, disponibles en sources ouvertes ou sur les forums cybercriminels resteraient encore trop instables et trop limités[9] pour identifier des vulnérabilités jour-zéro[10] rapidement et en quantité [13]. De même, il n'existe à l'heure actuelle aucun cas avéré d'exploitation de vulnérabilité jour-zéro découverte grâce à un modèle d'IA générative [13, 14]. De récents progrès pourraient toutefois à l'avenir bénéficier à des acteurs cyberoffensifs. En novembre 2024, le système d'IA générative **BigSleep** a démontré son efficacité pour la recherche de vulnérabilités dans des codes sources [15]. En juin 2025, le système d'IA générative **XBOW**, développé par d'anciens ingénieurs de GITHUB pour scanner des milliers d'applications Web simultanément et y identifier des vulnérabilités sans intervention humaine, a soumis des centaines de vulnérabilités, dont certaines critiques, sur différents programmes de *bugbounty* [16]. Si ces exemples dans le domaine de la cyberdéfense se multiplient, il est difficile d'évaluer les progrès des attaquants par la simple observation des interactions avec leurs cibles.

1.2 Différentes utilisations des services d'IA générative par divers profils d'attaquants

Un large spectre d'acteurs offensifs utilisent les services d'IA générative. En janvier 2025, GOOGLE indiquait que son modèle d'IA générative **Gemini** avait été utilisé entre 2023 et 2024 par des groupes cybercriminels ainsi que par les opérateurs d'au moins dix MOA liés à l'Iran, vingt à la Chine, neuf à la Corée du Nord et trois à la Russie [4, 17].

Si l'utilisation de services d'IA générative est de plus en plus répandue dans les attaques informatiques, ils sont utilisés différemment par les acteurs malveillants en fonction de leurs objectifs et de leur niveau de maturité. Pour les acteurs les plus matures, l'IA générative devient un nouveau cadre pratique, semblable à l'utilisation d'autres codes ou outils malveillants génériques[11]. Elle permet notamment de générer du contenu en masse dans plusieurs langues à des

8. Une requête (ou *prompt*) désigne l'instruction sous forme de texte envoyée par l'utilisateur au système d'IA.
9. A la suite de nombreux tests effectués par des chercheurs sur une cinquantaine d'IA générative, trois modèles seulement ont permis de développer un *exploit* fonctionnel au bout de nombreuses heures et de l'utilisation de plusieurs ressources. De plus, la plupart des modèles d'IA générative, même payants, ne permettent pas d'ingérer du code dépassant un certain nombre de lignes.
10. Vulnérabilité n'ayant fait l'objet d'aucune publication ou n'ayant reçu aucun correctif au moment de son exploitation.
11. Comme par exemple **Cobalt Strike** ou **Metasploit** utilisés très largement par de multiples modes opératoires

fins de tromperie ou de désinformation, de développer du code non signant ou d'effectuer des recherches sur des cibles plus rapidement. Il est également plausible d'affirmer que la maîtrise de plus en plus rapide des systèmes d'IA générative par ces acteurs pourrait mener à court ou moyen terme à l'automatisation complète ou quasi-complète d'une chaîne d'attaque. Pour les acteurs moins expérimentés, l'IA générative peut être un bon outil d'apprentissage et offrir un gain de productivité par exemple en répondant à des questions techniques [4, 12]. Dans l'ensemble des cas, l'IA générative permet aux acteurs malveillants d'agir plus rapidement et donc à plus grande échelle [4].

1.3 Le détournement de modèles d'IA générative existants à des fins malveillantes

Les modèles d'IA générative comme **ChatGPT**, développé par OpenAI, disposent de garde-fous techniques empêchant leur utilisation à des fins illégales ou non conformes aux standards définis par les développeurs [12]. Les acteurs malveillants cherchent à détourner ces limitations en manipulant ou structurant leurs requêtes afin de contourner les mécanismes de modération. Ces méthodes d'ingénierie de *prompt*, qui peuvent être des formulations ambiguës, des mots-clés spécifiques ou encore l'utilisation de scénarios fictifs, évoluent constamment et constituent un défi pour les développeurs [18]. Dès 2023, des chercheurs en sécurité parvenaient ainsi à détourner ChatGPT pour développer un code malveillant polymorphique [13] [19]. En 2024, des services de *jailbreak-as-a-service* [14] comme **EscapeGPT** ou **LoopGPT** sont apparus sur les forums cybercriminels [7].

Les systèmes d'IA générative se sont également ajoutés à la multitude d'offres sur étagères proposées à l'écosystème cybercriminel. Dès 2023, certains rapports font mention de services d'IA générative « débridés » tels que **WormGPT**, **FraudGPT** ou **EvilGPT** [20]. Ces services sont vendus sur les forums cybercriminels ou *via* des canaux Telegram [21] pour une centaine de dollars par mois. Depuis, une multitude de services aux niveaux de qualité hétérogènes a été proposée. Des modèles plus récents comme **WormGPT 4** seraient directement entraînés sur des jeux de données spécifiques à des activités cybercriminelles comme du code malveillant et des modèles d'hameçonnage [2].

attaquants pour compromettre le système d'information de leurs victimes.

12. Par exemple pour la fabrication de bombes, de produits stupéfiants ou encore la génération de contenus pédopornographiques.

13. Un code malveillant polymorphique est conçu pour modifier ou transformer son code régulièrement dans le temps afin d'éviter sa détection par des antivirus ou des EDR.

14. Service offert par des cybercriminels qui consiste à proposer des *prompts* permettant de contourner les garde-fous des systèmes d'IA générative.

2 LE CIBLAGE DE SYSTÈMES D'IA PAR DES MENACES INFORMATIQUES

Les catégories d'acteurs malveillants susceptibles de cibler spécifiquement les systèmes d'IA semblent similaires à celles qui s'attaquent aux SI conventionnels. Les systèmes de LLM pourraient cependant être vulnérables à de nouveaux vecteurs d'attaque potentiels et ce à différents niveaux [22] :

- **lors de l'entraînement** du modèle en introduisant des données corrompues ou fausses ;
- **lors de l'intégration** du modèle en y implémentant des portes dérobées [23] ;
- **lors de l'interrogation** du modèle [15] en injectant de fausses informations en vue d'altérer la réponse ou en récupérant des informations confidentielles concernant un compte utilisateur.

2.1 Empoisonnement des modèles d'IA à des fins d'altération de données ou de désinformation

Si aucun incident n'a été porté à la connaissance de l'ANSSI jusqu'à présent, il existe un risque de manipulation, de modification et d'interaction d'un acteur malveillant avec les données d'entraînement d'une IA générative. Une telle compromission pourrait entre autre mener à l'utilisation de ces modèles à des fins d'altération de données et de sabotage de systèmes opérationnels [24].

Exceptées des campagnes d'attaques, la multiplication de contenus fallacieux générés par IA sur Internet pourraient polluer les données d'entraînement des modèles sur lesquels s'appuient les agents conversationnels comme **ChatGPT** et contribuer à diffuser de fausses informations à grande échelle [25, 26]. Une analyse conjointe du UK AI Security Institute et du Alan Turing Institute aurait par ailleurs démontré qu'il serait possible d'empoisonner des modèles d'IA générative à partir de 250 documents malveillants seulement et que ce nombre resterait stable indépendamment de la taille des données d'apprentissage du modèle [27, 28]. Si ce sujet sort des prérogatives de l'Agence, l'ANSSI a pu observer certains modèles d'IA intégrant dès leur conception des limitations ou des éléments de censure [29, 30]. Dans le cadre du sommet pour l'IA 2024, VIGINUM a par ailleurs publié un rapport sur les défis et les opportunités de l'IA dans la lutte contre les manipulations de l'information [31].

2.2 Ciblage des modèles d'IA à des fins de compromission de logiciels ou d'exfiltration de données sensibles

Certaines attaques à l'encontre de modèles d'IA pourraient constituer une nouvelle forme d'attaque par chaîne d'approvisionnement. Des modèles d'IA générative disponibles en sources ouvertes et spécialisés dans la génération de code informatiques peuvent être malveillants ou compromis et exécuter du code arbitraire pour installer une porte dérobée sur le poste de l'utilisateur dès leur téléchargement [32]. Des attaquants peuvent également exploiter des failles au sein d'agents *Model Context Protocol* (MCP), utilisés pour connecter les LLM à des outils externes et à des sources de données. Ces serveurs, qui peuvent fonctionner en local ou à distance sur des hôtes tiers, peuvent participer à étendre la surface d'attaque s'ils ne sont pas suffisamment sécurisés [33]. La pratique du *slopsquatting* qui consiste à récupérer des noms de paquets imaginés

15. aussi appelée inférence du modèle.

par des IA puis à en diffuser des versions malveillantes est également utilisée. Les attaquant exploitent ainsi les hallucinations de l'IA pour introduire des paquets malveillants dans la chaîne d'approvisionnement logicielle [34].

Ainsi, les systèmes d'IA participent à l'augmentation de la surface d'attaque et ce d'autant plus lorsqu'ils sont intégrés dans des contextes logiciels plus larges, déployés dans des environnements classifiés ou dans certains flux opérationnels de l'entreprise [1]. En l'absence d'un cloisonnement rigoureux physique et des usages, la compromission du système d'IA pourrait mener à des conséquences plus concrètes comme l'atteinte à la confidentialité des données qu'il traite et à l'intégrité des systèmes d'information auxquels il est connecté [1, 24].

De plus, l'utilisation de comptes d'IA par les salariés dans un contexte professionnel peut exposer des informations sensibles en cas de compromission. Entre 2022 et 2023, plus de 100 000 comptes utilisateurs de **ChatGPT** ont été compromis par des acteurs cybercriminels à l'aide *d'infostealers*[16] comme **Rhadamanthys** puis revendus sur des forums [35]. Au delà des compromissions, des employés peuvent involontairement générer des fuites de données en fournissant à l'IA des informations sensibles, voires confidentielles. En juin 2023, des salariés de SAMSUNG ont ainsi divulgué des informations sensibles sur la technologie des semi-conducteurs en utilisant leur compte **ChatGPT** [36].

16. Code malveillant conçu pour collecter des informations sur le SI de la victime, notamment des identifiants et mots de passe enregistrés par les navigateurs internet, les cookies de session etc.

A Références

[1] ANSSI. *Recommandations de Sécurité Pour Un Système d'IA Générative*. 29 avril 2024.
URL : https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative.

[2] PALO ALTO NETWORKS. *The Dual-Use Dilemma of AI : Malicious LLMs*. 25 novembre 2025.
URL : https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/.

[3] NCSC.GOV.UK. *Impact of AI on Cyber Threat from Now to 2027*. 7 mai 2025.
URL : https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027.

[4] GOOGLE. *Adversarial Misuse of Generative AI*. 29 janvier 2025.
URL : https://cloud.google.com/blog/topics/threat-intelligence/adversarial-misuse-generative-ai.

[5] Microsoft Threat MICROSOFT SECURITY BLOG. *Staying Ahead of Threat Actors in the Age of AI*. 14 février 2024.
URL : https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

[6] SILENT PUSH. *Contagious Interview (DPRK) Launches a New Campaign Creating Three Front Companies to Deliver a Trio of Malware : BeaverTail, InvisibleFerret, and OtterCookie*. 24 avril 2025.
URL : https://www.silentpush.com/blog/contagious-interview-front-companies/.

[7] TREND MICRO. *Back to the Hype : An Update on How Cybercriminals Are Using GenAI - Nouvelles de sécurité*. 8 mai 2024.
URL : https://www.trendmicro.com/vinfo/fr/security/news/cybercrime-and-digital-threats/back-to-the-hype-an-update-on-how-cybercriminals-are-using-genai.

[8] NETENRICH_FRAUDGPTVILLAINAVATAR_2023. *TA547 Targets German Organizations : Rhadamanthys Stealer*. 3 avril 2024.
URL : https://www.proofpoint.com/us/blog/threat-insight/security-brief-ta547-targets-german-organizations-rhadamanthys-stealer.

[9] ESET. *First Known AI-powered Ransomware Uncovered by ESET Research*. 26 août 2025.
URL : https://www.welivesecurity.com/en/ransomware/first-known-ai-powered-ransomware-uncovered-eset-research/.

[10] GOOGLE CLOUD BLOG. *GTIG AI Threat Tracker : Advances in Threat Actor Usage of AI Tools*. 5 novembre 2025.
URL : https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools.

[11] RECORDED FUTURE. *Ukraine Warns of Growing AI Use in Russian Cyber-Espionage Operations*. 14 février 2025.
URL : https://therecord.media/russia-ukraine-cyber-espionage-artificial-intelligence.

[12] NCSC.GOV.UK. *The Near-Term Impact of AI on the Cyber Threat*. 24 janvier 2024.
URL : https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

[13] FORESCOUT. *Artificial Exploits, Real Limitations : How AI Cyber Attacks Fall Short*. 10 juillet 2025.
URL : https://www.forescout.com/blog/artificial-exploits-real-limitations-how-ai-cyber-attacks-fall-short/.

[14] PLATFORM SECURITY. *How I Used AI to Create a Working Exploit for CVE-2025-32433 Before Public PoCs Existed*. 17 avril 2025.
URL : https://platformsecurity.com/blog/CVE-2025-32433-poc.

[15] BIGSLEEPTEAM. *Project Zero : From Naptime to Big Sleep : Using Large Language Models To Catch Vulnerabilities In Real-World Code*. 1er novembre 2024.
URL : https://googleprojectzero.blogspot.com/2024/10/from-naptime-to-big-sleep.html.

[16] BLOG DU MODÉRATEUR. *L'IA est-elle déjà un meilleur pentester que l'humain ?* 3 juillet 2025.
URL : https://www.blogdumoderateur.com/ia-meilleur-pentester-humain/.

[17] THEREGISTER.COM. *Here's How Spies and Crooks Abuse Gemini AI*. 5 novembre 2025.
URL : https://www.theregister.com/2025/11/05/attackers_experiment_with_gemini_ai/.

[18] JEDHA. *10 techniques pour jailbreaker ChatGPT*. 26 mai 2025.
URL : https://www.jedha.co/formation-ia/comment-jailbreak-chatgpt-dan-prompt-injection-et-autres-techniques.

[19] CYBERARK. *Chatting Our Way Into Creating a Polymorphic Malware*. 17 janvier 2023.
URL : https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware.

[20] Vitaly CATO NETWORKS. *Cato CTRL™ Threat Research : WormGPT Variants Powered by Grok and Mixtral*. 17 juin 2025.
URL : https://www.catonetworks.com/blog/cato-ctrl-wormgpt-variants-powered-by-grok-and-mixtral/.

[21] NETENRICH. *FraudGPT : The Villain Avatar of ChatGPT | Netenrich*. 25 juillet 2023.
URL : https://netenrich.com/blog/fraudgpt-the-villain-avatar-of-chatgpt.

[22] REVUE FRANÇAISE DE COMPTABILITÉ. *L'IA générative : quelles sont les cybermenaces et comment s'en protéger ? | Revue Française de Comptabilité*. 1er juin 2024.
URL : https://revuefrancaisedecomptabilite.fr/lia-generative-quelles-sont-les-cybermenaces-et-comment-sen-proteger/.

[23] NEXT LINK. *Les modèles de langage peuvent contenir des backdoors*. 26 mars 2024.
URL : https://next.ink/123823/les-modeles-de-langage-peuvent-contenir-des-backdoors/.

[24] NIST. *Adversarial Machine Learning A Taxonomy and Terminology of Attacks and Mitigations*. 1er mars 2025.
URL : https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf.

[25] LE DEVOIR. *L'IA, un nouveau front pour la désinformation*. 10 février 2025.
URL : https://www.ledevoir.com/culture/medias/840655/ia-nouveau-front-desinformation.

[26] MNTD.FR. *"LLM Laundering", une nouvelle pratique de désinformation qui cible les outils d'IA*. 12 mars 2025.
URL : https://www.mntd.fr/llm-laundering-une-nouvelle-pratique-de-desinformation-qui-cible-les-outils-dia/.

[27] ANTHROPIC. *A Small Number of Samples Can Poison LLMs of Any Size*. 9 octobre 2025.
URL : https://www.anthropic.com/research/small-samples-poison.

[28] UK AI SECURITY INSTITUTE. *Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples*. 8 octobre 2025.
URL : https://arxiv.org/pdf/2510.07192.

[29] THE GUARDIAN. *We Tried out DeepSeek. It Worked Well, until We Asked It about Tiananmen Square and Taiwan*. 28 janvier 2025.
URL : https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan.

[30] LE MONDE. « Grok, l'IA d'Elon Musk, encourage à voter Marine Le Pen et salue « l'efficacité » d'Adolf Hitler ». 9 juillet 2025.
URL : https://www.lemonde.fr/pixels/article/2025/07/09/grok-l-ia-d-elon-musk-encourage-a-voter-marine-le-pen-et-salue-l-efficacite-d-adolf-hitler_6620180_4408996.html.

[31] VIGINUM. *Défis et « opportunités » de l'intelligence artificielle dans la lutte contre les manipulations de l'information | SGDSN*. 7 février 2024.
URL : http://www.sgdsn.gouv.fr/publications/defis-et-opportunites-de-lintelligence-artificielle-dans-la-lutte-contre-les.

[32] ADIA. *Data Scientists Targeted by Malicious Hugging Face ML Models with Silent Backdoor*. 27 février 2024.
URL : https://jfrog.com/blog/data-scientists-targeted-by-malicious-hugging-face-ml-models-with-silent-backdoor/.

[33] ELASTIC SECURITY LABS. *MCP Tools : Attack Vectors and Defense Recommendations for Autonomous Agents*. 18 décembre 2025.
URL : https://www.elastic.co/security-labs/fr/security-labs/mcp-tools-attack-defense-recommendations.

[34] SOCKET. *The Rise of Slopsquatting : How AI Hallucinations Are Fueling...* 9 avril 2025.
URL : https://socket.dev/blog/slopsquatting-how-ai-hallucinations-are-fueling-a-new-class-of-supply-chain-attacks.

[35] LES ECHOS. *ChatGPT déjà victime de piratage massif*. 22 juin 2023.
URL : https://www.lesechos.fr/tech-medias/intelligence-artificielle/chatgpt-deja-victime-de-piratage-massif-1955066.

[36] RFI. *Des données sensibles de Samsung divulgués sur ChatGPT par des employés*. 9 avril 2023.
URL : https://www.rfi.fr/fr/technologies/20230409-des-donn%C3%A9es-sensibles-de-samsung-divulgu%C3%A9s-sur-chatgpt-par-des-employ%C3%A9s.

SDO/DCA/CTI

4 février 2026

AGENCE NATIONALE DE LA SÉCURITÉ DES SYSTÈMES D'INFORMATION

ANSSI – 51, boulevard de la Tour-Maubourg – 75700 PARIS 07 SP
cyber.gouv.fr • cert.ssi.gouv.fr