

La cybersécurité est de plus en plus impactée par la révolution de l'IA

Roberto Kung

Membre senior SEE

Jean-François Audenard

Réfèrent Orange Expertise Sécurité

Marc Lacoste

Membre Senior Orange Expertise Sécurité

Une actualité de plus en plus inquiétante

En avril, il y a eu le buzz sur Mythos, le modèle IA d'Anthropic tellement puissant pour trouver des failles de sécurité qu'il a été seulement mis à disposition de quelques entreprises américaines pour leur permettre de tester leurs logiciels (initiative Glasswing ¹). L'ENISA européenne n'a pu avoir accès à Mythos qu'en septembre, démontrant ainsi la dépendance technologique de l'Europe. En juillet, un modèle d'OpenAI mal contrôlé a même lancé une cyberattaque dans l'internet. Depuis l'été, de plus en plus de personnalités demandent la mise en place d'une gouvernance pour contrôler le développement de l'IA. La question de l'évolution de l'IA, notamment d'une super AI pouvant menacer l'avenir, est une question majeure (elle fera l'objet d'un dossier dans la REE au début 2027). Pour la suite, nous allons décrire comment la cybersécurité est impactée par la révolution de l'IA.

L'IA est de plus en plus utilisée pour le développement logiciel, la gestion des applications. En particulier, elle est utili-

sée pour trouver et gérer les failles de sécurité et pour mieux se protéger. Le problème des vulnérabilités n'est pas nouveau ² mais l'essor spectaculaire de l'IA l'amplifie fortement, car si elle permet d'améliorer la sécurité, elle facilite aussi le travail des cyberattaquants.

La situation actuelle de la cybersécurité et de l'IA

La part de l'exploitation des vulnérabilités (CVEs 'Common Vulnerabilities and Exposures') devient prépondérante depuis mi 2023, passant de 5 % à plus de 30 % des brèches de sécurité aujourd'hui (figure 1). Cette augmentation a précédé de 3 ans l'avènement de Mythos.

Mythos n'a fait qu'amplifier la découverte de nouvelles vulnérabilités. Aujourd'hui,

il y a plus de 300 000 CVE répertoriés depuis plus de vingt ans ³. Mythos a découvert 10 000 vulnérabilités critiques en quelques semaines. Cette avalanche de nouvelles vulnérabilités ('patch apocalypse') montre l'importance de pouvoir les corriger rapidement. Aujourd'hui, les attaques sur une nouvelle vulnérabilité peuvent apparaître quelques heures après leur découverte alors qu'il fallait plusieurs semaines auparavant. Par ailleurs, l'actualité récente montre aussi le cas d'IA mal contrôlée pouvant mener des cyberattaques.

Capacités des 'Frontier Model' de l'IA

Anthropic a lancé Mythos en avril pour effectuer des scans de codes et infrastructures existants et détecter leurs failles de

² Voir dossier cybersécurité du REE 2024-3

³ <https://www.cve.org/>

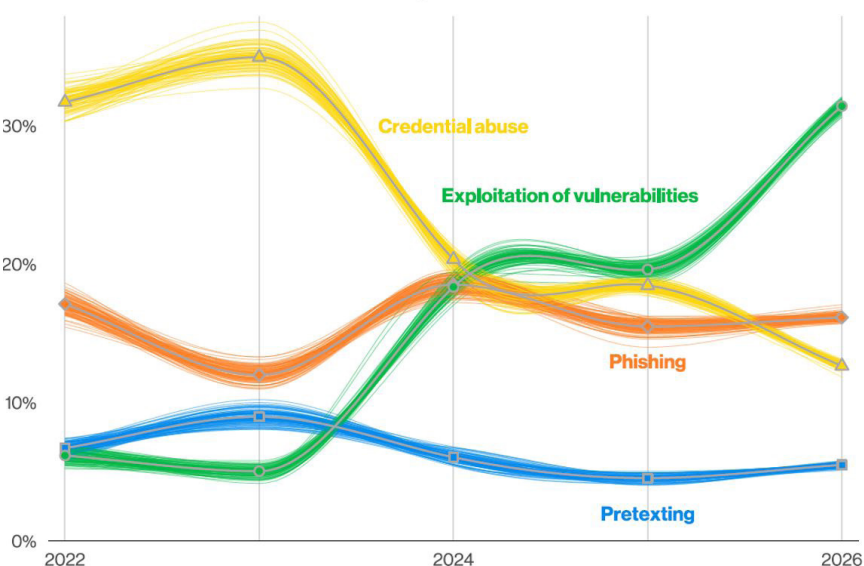


Figure 1 : Proportion des brèches de sécurité (source Verizon 2026).

¹ <https://www.anthropic.com/research/glasswing-initial-update>

sécurité. Sa puissance est telle que son accès a été restreint. En juin, le produit rebaptisé Claude Mythos 5, a même été arrêté par suite de la découverte par Amazon d'un contournement possible des mesures de sécurité de la version Fable 5 accessible à tous. Le gouvernement US décide alors d'interdire leur utilisation par des non américains, ce qui revient à les interdire totalement car il est impossible de vérifier la nationalité d'un utilisateur. Cette interdiction est levée fin juin avec la mise en place d'une parade. Cette péripétie illustre la dépendance des utilisateurs d'Anthropic aux décisions du gouvernement américain. Maintenant Anthropic travaille avec le gouvernement US pour promouvoir l'innovation de l'IA et de la sécurité et pour fournir les informations sur les menaces aux 'vulnerability clearinghouses'.

Open AI, dans un contexte de concurrence effrénée, sort Daybreak en mai 2026, un agent LLM spécialisé en cybersécurité fondé sur GPT 5.5, avec une extension en juin par l'initiative 'Patch The Planet'. En plus de GPT 5.5 acces-

sible par tous, Open AI rajoute une version 'with trusted access' pour ceux qui travaillent sur la sécurité. Par ailleurs, une version 'Cyber' avancée est disponible exclusivement pour les organisations 'enregistrées'. Là encore se dessine une dépendance vis-à-vis des grands fournisseurs américains. A la différence d'Anthropic, Open AI se focalise en particulier sur toutes les phases du développement logiciel et la détection en amont de toutes les vulnérabilités possibles.

Le chinois Zhipu AI a rendu disponible son modèle GLM 5.2 en mi-juin en OpenWeight (le code est disponible en Open source mais pas les données d'entraînement). Il est pratiquement aussi puissant voire plus puissant qu'Anthropic Fable 5 ou Open AI GPT-5.5 pour la détection des vulnérabilités. Mais il est 10 fois moins cher (moins de 5\$ par million de Token contre 50\$ pour Fable 5 d'Anthropic). Il y a encore moins cher avec le chinois Deepseek (moins de 1\$ par million de token). Il faut souligner cependant que le prix n'est pas le seul facteur, car il faut tenir compte d'abord de l'ef-

ficacité des réponses qui varient parfois d'un facteur 2 ou 3 comme le montrent les tests de Semgrep (figure 2). Néanmoins, une telle différence de prix attire évidemment les clients, même américains.

Les autres LLM ont en fait tous des capacités plus ou moins bonnes de détection des failles (figure 3). Notons notamment que parmi tous les LLM américains ou chinois, on retrouve en queue de peloton celui de Mistral IA (Devstral), le seul européen.

Detector	F1 ↓
▶ Semgrep Multimodal (AID) BEST: GPT-5.5	62.0%
▼ LLM Providers BEST: GLM 5.2	39.0%
▶ GLM 5.2 ①	39.0%
▶ MiniMax M3 ①	22.9%
▶ Kimi K2.7 Code ①	21.6%
▶ DeepSeek V4 Pro ①	16.8%
▶ Qwen 3.7 Plus ①	11.4%
▶ Nemotron Super 3 120B ①	9.3%
▶ Devstral 2 123B ①	8.3%
▶ Claude Code BEST: CLAUDE OPUS 4.8	37.4%

Figure 3 : Efficacité des LLM (source Semgrep).

Rank	Configuration	Harness	F1
1	Semgrep Multimodal (GPT 5.5)	Semgrep Multimodal	61%
2	Semgrep Multimodal (Opus 4.8)	Semgrep Multimodal	53%
3	GLM 5.2	Pydantic AI (prompt only)	39%
4	Claude Code (Opus 4.6)	Claude Code SDK	37%
5	Claude Code (Opus 4.8/4.7)	Claude Code SDK	28%
6	MiniMax M3	Pydantic AI (prompt only)	23%
7	Kimi K2.7 Code	Pydantic AI (prompt only)	22%
8	GPT-5.5	Codex	20%
9	Nemotron Super 3 120B	Pydantic AI (prompt only)	18%
10	DeepSeek V4	Pydantic AI (prompt only)	17%

Figure 2 : Efficacité des LLM (source Semgrep).

Ce panorama montre d'abord le risque de dépendance à un petit nombre de fournisseurs US monopolistiques qui ont un pouvoir important sur les données, écornant la souveraineté et réduisant l'autonomie d'une entreprise, d'un pays ou de l'Europe. En contraste, notons le softpower à la chinoise avec la mise à disposition gratuite du code de leurs LLMs. C'est d'ailleurs la teneur du discours de Xi Jin Ping ⁴ lors de la récente conférence mondiale sur l'IA. Cela permet de disposer de solutions alternatives aux fournisseurs US. Notons que la mise à disposition du code peut malheureusement aussi être une aubaine pour les cyberattaquants.



4 https://www.lemonde.fr/pixels/article/2026/07/17/xi-jinping-appelle-a-une-cooperation-mondiale-sur-l-ia-avec-une-proche-centree-sur-l-hu-main_6724134_4408996.html

“ Les cas de cyber extorsion ont augmenté de 45 %, touchant de plus en plus les petites et moyennes entreprises. En même temps, l’hacktivisme est de plus en plus lié aux agendas géopolitiques. ”

... Les utilisations offensives

Selon le dernier rapport 2026 d’Orange Cyberdéfense ⁵, les cybercriminels, principalement motivés par l’appât du gain, ont davantage porté leur attention sur l’extorsion, les arnaques et les vulnérabilités, en utilisant des méthodes de plus en plus sophistiquées. Les cas de cyber extorsion ont augmenté de 45 %, touchant de plus en plus les petites et moyennes entreprises. En même temps, l’hacktivisme est de plus en plus lié aux agendas géopolitiques. En 2025, 96 % des actions menées par un groupe d’hacktivistes aligné sur l’état Russe ont visé l’Europe. Il s’agit véritablement d’une ‘guerre hybride’ mettant à l’épreuve la résilience digitale de toutes les infrastructures (administrations, hôpitaux, réseaux...). En outre, la guerre informationnelle fait aussi rage par le biais des réseaux sociaux.

Ces évolutions malveillantes se font de plus en plus à l’aide de l’IA permettant du *phishing* automatisé, des *deepfakes* (ou « hypertrucage ») et la création facilitée de codes permettant d’exploiter des vulnérabilités ou d’en découvrir de nouvelles. Ces attaques représentent une menace de plus en plus globale, d’autant plus que l’utilisation généralisée de l’IA au sein des infrastructures digitales des entreprises élargit les surfaces d’attaque et introduit de nouvelles vulnérabilités. Après le « *Shadow-IT* », les entreprises et organisations doivent désormais faire face au « *Shadow-AI* » avec des impacts sur la confidentialité des données, des risques réglementaires ou à des dé-

ploiements d’outils agentiques comme OpenClaw sur les postes de travail qui augmentent drastiquement la surface d’attaque.

Quelques illustrations du lien entre cyberattaques et l’IA

L’étude récente du *AI Security Institute UK* ⁶ illustre ce que permet l’IA, en simulant l’attaque d’un réseau d’entreprise typique (4 sous-réseaux, une vingtaine d’hôtes) qui nécessite une attaque en 32 étapes. Cette attaque nécessite 2,5 jours de travail d’un expert humain. Mythos et GPT5.5 ont tous les deux réussi à élaborer cette attaque en dépensant 100 millions de tokens c’est-à-dire environ 5 000 USD. Il est à parier que les progrès constants de l’IA vont permettre

d’élaborer ce type d’attaque en divisant par un facteur 10 le nombre de tokens avec des LLM plus spécialisés, sachant que le coût unitaire du token peut être facilement divisé aussi par un facteur 10. Le coût d’élaboration d’une telle attaque reviendrait alors à 50 USD.

L’exemple du protocole MCP ⁷ (*Model Context Protocol*) illustre l’augmentation de la surface d’attaque des infrastructures à cause de l’IA. Ce protocole permet à un agent IA de dialoguer avec les serveurs. Il introduit de facto des nouvelles opportunités d’attaques (figure 4). L’agent utilisateur du protocole doit mettre en place les mécanismes de sécurité adéquats. L’entreprise qui utilise un agent IA doit alors impérativement s’assurer que ces mécanismes de sécurité sont installés.

Quelques exemples des approches d’attaques – du « pentest » (test de pénétration) aux attaques

Les outils et services offensifs renforcés par l’IA ont rapidement vu le jour et sont désormais bien ancrés dans le paysage.

6 <https://www.aisi.gov.uk/blog/our-evaluation-of-openai-gpt-5-5-cyber-capabilities>

7 <https://www.akamai.com/blog/security-research/new-mcp-specification-security-teams-must-prepare>



Figure 4 : Les risques introduits par le protocole MCP utilisé par les agents IA (source Akamai).

5 <https://www.orange cyberdefense.com/fr/security-navigator-2026>

Ces services sont utilisés pour tester la sécurité mais peuvent aussi être utilisés pour attaquer. On pourra citer la plateforme agentique *Strix*⁸ disponible en Opensource permettant d'automatiser à la fois la découverte de vulnérabilités et de développer des codes d'exploitation. Des sociétés comme par exemple *Ox Security*⁹ ou *Aikido*¹⁰ proposent des services de *pentest* agentiques en mode SaaS ou déployables en interne. Des acteurs établis dans le *BugBounty* comme le leader Européen *YesWeHack*¹¹ ont eux aussi fait le pas et proposent des *pentests* basés sur de l'agentique. Si les usages des offreurs sont encadrés pour rester dans un cadre légal, les technologies Opensource sont, par nature, à usage dual et sont donc autant d'outils à disposition des attaquants.

Les techniques d'attaques contre les systèmes d'IA sont variées (*OWASP Top 10 Risk & Mitigations for LLMs and Gen AI Apps*)¹² et s'ajoutent aux vulnérabilités inhérentes à tout système d'information traditionnel. Le « *jailbreak* » d'un LLM consiste à contourner ses mesures de sécurité afin de le faire répondre à des requêtes normalement interdites, telles que l'extraction du « *prompt système* », la génération de code malveillant ou la création de contenus destinés à des attaques de phishing. En raison de leur architecture, les LLM sont également susceptibles d'être vulnérables à des attaques par « *injection de prompt* », qu'elles soient directes ou indirectes, par exemple via l'analyse de textes intégrés dans une image ou un fichier qui seront injectés dans le *prompt*.

“ En raison de leur architecture, les LLM sont également susceptibles d'être vulnérables à des attaques par « injection de prompt », qu'elles soient directes ou indirectes, par exemple via l'analyse de textes intégrés dans une image ou un fichier qui seront injectés dans le prompt. ”

Les acteurs de l'IA sont aussi la cible d'attaques dans la chaîne d'approvisionnement qui exploitent la croissance des dépendances logicielles open source. Les récentes attaques menées par TeamPCP (un cyberattaquant) sur l'écosystème opensource illustrent la vulnérabilité de ces environnements, qui restent particulièrement fragiles à de telles menaces. Des entreprises comme *Mistral AI*¹³ ou le logiciel open source *Litellm*¹⁴, largement utilisés dans le domaine de l'IA, ont été impactées par ces attaques.

L'entraînement et le « harnais » pour canaliser et mieux focaliser la puissance des LLMs

Si des modèles de langage génériques sont adaptés à un large éventail d'usages, ils deviennent particulièrement efficaces pour la cybersécurité lorsqu'ils sont spécialisés ou complétés par :

- Un entraînement ou une « spécialisation » sur des jeux de données directement liés à ce domaine. C'est le cas de modèles fermés comme *Claude Mythos*¹⁵ d'Anthropic ou *GPT-5.6 Cyber*¹⁶ d'OpenAI, mais aussi de certains modèles open source proposés par Cisco, tels que *Antares*¹⁷, spéciali-

sé dans la détection de vulnérabilités et les modèles *Foundation-sec-8b*¹⁸ et *Foundation-sec-8b-reasoning*¹⁹, spécialisés quant à eux dans les tâches de raisonnement.

- Un « harnais » qui est une couche logicielle qui connaît le contexte d'utilisation et qui permet d'optimiser l'utilisation d'un ou plusieurs modèles par des agents spécialisés, agissant de façon coordonnée et capables d'actions avec le monde extérieur via des outils (accès à des fichiers, recherche sur internet, lancement de commandes, appel à des serveurs MCP, ...).

Si le choix du modèle (ou des modèles) utilisés simultanément est un facteur clé, la qualité du harnais est essentielle, voire peut-être plus importante que les modèles utilisés (cf. le harnais multimodal de Semgrep – figure 2). Pour des usages efficaces en cybersécurité, il est essentiel que modèles et harnais soient sélectionnés et assemblés de façon judicieuse comme expliqué par *Cato Networks*²⁰. Alors que le coût de l'inférence est devenu un enjeu majeur pour de nombreuses entreprises et organisations, il convient de noter que l'utilisation de solutions agentiques et de harnais entraîne une augmentation de la consommation de tokens, et donc des coûts associés.

...

8 <https://github.com/usestrix/strix>

9 <https://www.ox.security/press/ox-security-introduces-the-first-agentic-pentester-to-link-exploits-directly-to-code/>

10 <https://www.aikido.dev/blog/the-aikido-machine-on-prem-ai-pentesting-that-never-leaves-your-network>

11 <https://www.businesswire.com/news/home/20260625729553/fr>

12 <https://genai.owasp.org/llm-top-10/>

13 <https://cybelangel.com/blog/attaque-mistral-ai-comment-la-compromission-tanstack-a-contamine-le-sdk/>

14 https://www.trendmicro.com/fr_fr/research/26/c/inside-litellm-supply-chain-compromise.html

15 <https://www.anthropic.com/claude/mythos>

16 <https://developers.openai.com/api/docs/models/gpt-5.6-cyber>

17 <https://blogs.cisco.com/ai/introducing-antares-the-most-efficient-open-weight-ai-models-for-vulnerability-localization>

18 <https://blogs.cisco.com/security/foundation-sec-cisco-foundation-ai-first-open-source-security-model>

19 <https://blogs.cisco.com/security/foundation-sec-8b-reasoning-worlds-first-security-reasoning-model>

20 <https://www.catonetworks.com/blog/the-agentic-attacker-one-objective-one-prompt-forty-minutes-domain-admin-game-over/>

- Le déploiement de nœuds d'inférence privés en local, associé à des modèles open source comme ceux de CISCO par exemple, constitue une approche permettant de réduire ces coûts tout en offrant des garanties intéressantes en matière de maîtrise de l'information, sujet particulièrement crucial dans le domaine de la cybersécurité, notamment pour la détection de vulnérabilités. Ces infrastructures locales, sous contrôle total, peuvent également permettre des usages plus libres (ou « non bridés ») que ceux proposés par des modèles en mode SaaS, en autorisant leur utilisation dans des domaines sensibles liés à la cybersécurité (analyse de logs d'incidents, détection de codes malicieux, etc.) qui pourraient être bloqués car vus comme des tentatives d'utilisation des modèles par des cybercriminels.

Les garde-fous (Guardrails) pour encadrer et protéger les systèmes d'IA

Pour éviter les utilisations malveillantes, les LLM accessibles en ligne et ceux intégrés dans des systèmes mettent en œuvre des contrôles pour que leur fonctionnement soit conforme aux attentes et conformes aux lois et réglementations. Ces contrôles sont connus sous le terme de « *guardrails* » et se positionnent à des endroits stratégiques.

Concernant la mise en œuvre des *Guardrails*, ceux-ci peuvent utiliser des LLM pour classer le niveau de dangerosité des entrées ou des sorties, afin de décider ou non d'un éventuel blocage. À l'inverse, les *Guardrails* peuvent également s'appuyer sur des algorithmes classiques (par exemple : *pattern matching*, analyse NLP (*Natural Language Processing*), droits d'accès ou autorisations, etc.). Les implémentations les plus fiables pourraient privilégier une approche hybride, combinant la puissance et la souplesse d'un système de classification basé sur un LLM avec la fiabilité et le caractère déterministe de l'algorithmique classique.

Si dans la figure 5 on retrouve un bloc « *Guardrails* » en entrée et un autre en sortie, il faut conserver à l'esprit qu'il s'agira en fait de composants fonctionnels classificateurs ou « *classifiers* » activés en fonction des différents objectifs visés ou du contexte d'utilisation (masquage des données à caractère personnel, détection d'attaques en Jailbreak, demandes en dehors des usages définis, détection d'hallucinations, etc.).

Au-delà du fonctionnement même de chaque classificateur, la « notation » attribuée en sortie d'un classificateur peut être l'objet d'un ajustement comme décrit sur la figure 6. Le fonctionnement des *Guardrails* présents en amont et en aval d'un seul et même modèle pourra donc être un élément différenciateur entre deux niveaux ou types de services, comme présenté en figure 6 ci-dessous.

Les *Guardrails* forment un mécanisme clé pour empêcher qu'un LLM sorte de son cadre et mène des cyberattaques de façon involontaire. Est-ce que ce type de mécanisme est suffisant pour empêcher l'apparition d'IA malfaisante ? Nous reviendrons sur ce point en conclusion.

Quelle posture de défense adopter à l'ère de l'IA ?

Sur le plan de la défense en cybersécurité, l'arrivée de de l'IA et de l'IA agentique peut être perçue à la fois comme :

- Une accélération soudaine du nombre de vulnérabilités à prendre en compte dans les processus de maintien en condition opérationnelle de sécurité, avec un risque certain d'étouffement des équipes.

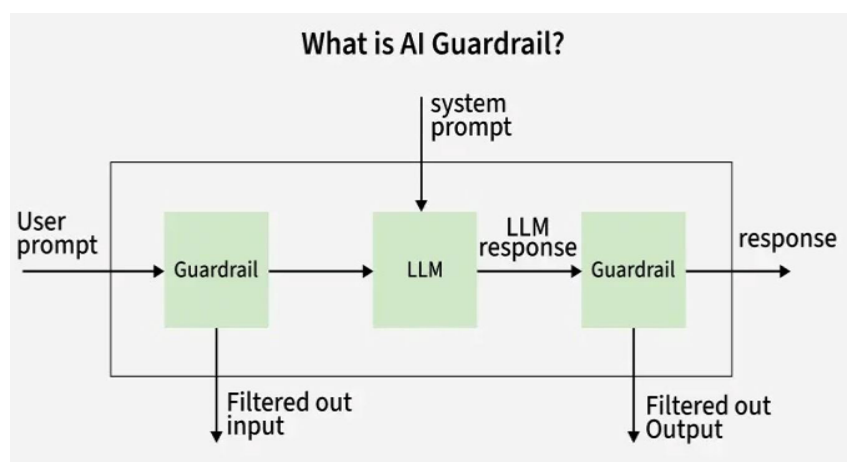


Figure 5 : Guardrails (source GeekForGeeks).

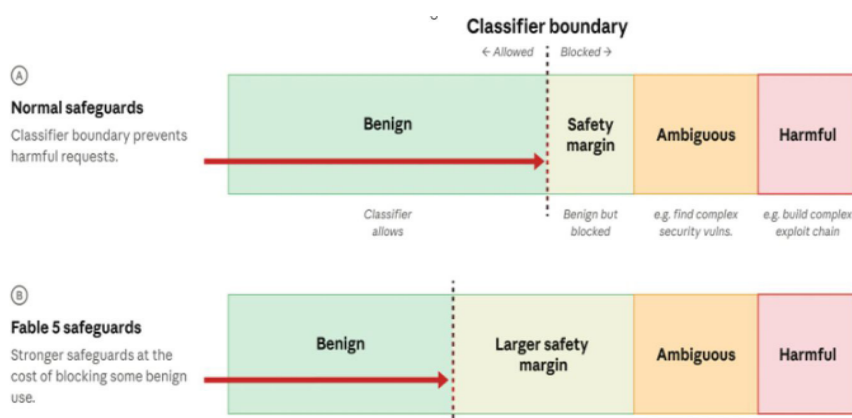


Figure 6 : L'approche des classifieurs (source Anthropic).

“Aujourd’hui, une entreprise met en moyenne 40 jours pour corriger une vulnérabilité qui vient d’être découverte, ce qui est bien trop lent par rapport aux quelques heures pour les hackers pour développer l’attaque.”

- Un facteur d'accélération des menaces, avec une baisse des barrières d'entrée pour les activités cybercriminelles, qui peuvent évoluer vers des attaques assistées par l'IA, voire totalement autonomes.

- Une importance cruciale des pratiques liées à la gestion de la surface d'attaque, telles que la mise en œuvre systématique d'architectures segmentées à un niveau de granularité toujours plus fin (microsegmentation), l'adoption d'approches *Zero Trust*, ainsi que la maîtrise renforcée des mécanismes de gestion des identités, des accès et des droits. Bien que cette liste puisse paraître longue, elle demeure essentielle. De plus, l'utilisation de l'IA peut constituer un levier précieux lors de leur mise en œuvre opérationnelle.

- Une opportunité d'utiliser l'IA et l'IA agentique pour transformer les activités de cybersécurité en les rendant plus efficaces et plus rapides, notamment sur le volet des correctifs de sécurité mais aussi l'amélioration et l'augmentation des capacités de détection.

Il est essentiel pour une entreprise de comprendre l'architecture et la cartographie de ses *endpoints* (mobiles, laptops, PCs, serveurs internes ou dans le cloud, les OS et applications, ...) pour prioriser et déployer les correctifs pertinents. Le nombre de vulnérabilités reste difficile à gérer comme le montre le catalogue CISA²¹ des vulnérabilités qui existe depuis 25 ans, ou la typologie des types d'attaques du MITRE²². L'entreprise doit prioriser sa protection, souvent à l'aide de spécialistes comme la startup Hackuity qui explique les surfaces d'attaques selon la configuration de l'entreprise, ou comme Orange Cyberdefense, le premier acteur européen pour la cybersécurité, qui a mis en place une organisation VOC²³ (*Vulnerability Operations Center*) en complément des SOC (*Security Operations Center*), avec une

21 <https://www.cisa.gov/known-exploited-vulnerabilities-catalog>

22 <https://attack.mitre.org/>

23 <https://www.orange cyberdefense.com/fr/insights/actualites/souverainete-numerique>

base de données de gestion des vulnérabilités contenant près de 70 000 références et plus de 1,3 million de résultats uniques. Notons au passage la dépendance mondiale envers les deux institutions américaines CISA et Mitre, qui ont été touchées par les coupes budgétaires de l'administration Trump. Cela a conduit en mi 2025 à la mise en place d'une Base de Donnée Européenne gérée par l'ENISA : EUVD²⁴.

Aujourd'hui, une entreprise met en moyenne 40 jours pour corriger une vulnérabilité qui vient d'être découverte, ce qui est bien trop lent par rapport aux quelques heures pour les hackers pour développer l'attaque. Il est nécessaire d'automatiser les corrections de sécurité de plus en plus fréquentes provenant des fournisseurs, et de vérifier que ces corrections soient effectives sur les *endpoints*. Par ailleurs, avant d'obtenir les corrections, il est indispensable de prendre des mesures complémentaires (*'compensating controls'* comme la fermeture d'un port vulnérable, d'une application vulnérable, restrictions d'accès...) sachant que la configuration des *Endpoints* évoluent constamment. Il est indispensable de superviser ces *endpoints* pour éviter les dérives progressives vers des configurations moins protégées.

...

À l'ère de l'IA, toute organisation devrait avoir pour objectif de déployer des systèmes plus résilients, avec une surface d'attaque aussi réduite que possible et facilement défendable, conçus pour limiter au maximum le rayon de détonation en cas d'attaque. Ces systèmes doivent également disposer de moyens fiables pour le maintien en condition opérationnelle de la sécurité plus fluides, ainsi que de capacités de détection performantes, capables de freiner aussi bien des attaquants humains qu'agentiques.

LLMs in Software Vulnerability Assessment



Figure 7 : L'IA peut aider à éviter les vulnérabilités (source Garner).

- Sur le plan de la défense augmentée grâce à l'IA, les LLM peuvent contribuer à éviter l'introduction de vulnérabilités à toutes les phases du développement logiciel (figure 7), ou à proposer des correctifs plus rapidement dès leur découverte. Face à des attaquants renforcés par l'IA, la défensabilité des systèmes doit également évoluer, notamment par la mise en place de techniques d'identification d'attaques basées sur l'IA (comme des techniques d'injection de prompts à des fins défensives) ou par l'adoption de méthodes de détection comportementale, telles que l'analyse de bruits ou de signaux faibles, la détection de rapidités extrêmes entre la vulnérabilité et son exploitation, ou tout autre signe d'un système agentique offensif.

Le défi demeure pour les défenseurs, qui doivent simultanément maintenir et protéger leurs systèmes contre les attaques tout en apprenant et innovant en permanence pour regagner l'avantage.

Conclusion

Le domaine de la cybersécurité est profondément impacté par l'IA à la fois en défensif et en offensif. Les industriels de l'IA développent les LLM à cadence forcée et ont constaté des comportements dangereux de l'IA lors des entraînements (Open AI ²⁵, Anthropic ²⁶). Pour l'instant, ces incidents sont explicables et les solutions préconisées sont de mieux isoler les environnements de tests et de mieux utiliser les *Guardrails*. Devant la puissance croissante de l'IA, il est légitime de se poser la question si ces mesures sont suffisantes pour empêcher l'apparition d'IA malfaisante. De nombreuses personnalités de l'IA tirent déjà la sonnette d'alarme.

25 <https://openai.com/fr-FR/index/hugging-face-model-evaluation-security-incident/>

26 <https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents>

L'IA et la sécurité des systèmes spatiaux

La sécurité du New Space, notamment des constellations de satellites est à assurer, alors qu'il y a de nombreuses menaces de cybersécurité :

- 1) *au sol* avec les infrastructures terrestres Cloud et stations de contrôle des satellites ;
- 2) dans l'Espace avec des satellites en orbite basse, moyenne ou géostationnaire ; et dans
- 3) *le segment utilisateur* où terminaux et applications accèdent aux services et données satellitaires. L'Espace est un domaine où les menaces sont amplifiées, e.g., pas d'accès physique, impacts systémiques : brouillage, interception des communications, détournement, menaces hybrides, chaînes d'approvisionnement... ¹

L'IA peut être utilisée pour améliorer la sécurité des systèmes spatiaux à travers des *architectures centralisées ou décentralisées*, avec des compromis entre performances et sécurité ². L'option centralisée au sol est la plus simple mais avec des pénalités importantes sur les latences des communications spatiales. Des options décentralisées avec entraînement et inférence dans l'Espace (LLM dans le satellite) sont plus intéressantes en termes de passage à l'échelle, de sécurité et d'efficacité énergétique ³.

L'IA agentique présente un réel potentiel pour rendre la supervision de sécurité de ces systèmes spatiaux plus granulaires et efficaces ⁴. Déployer les agents IA dans l'Espace offre les meilleurs bénéfices de latence et de coût. Certains systèmes sont déjà opérationnels pour des applications de défense : la « guerre des étoiles » sort désormais du domaine de science-fiction...

1 <https://data.europa.eu/doi/10.2824/317096>

2 <https://hal.science/view/index/docid/5263036>

3 <https://hellofuture.orange.com/en/protecting-ai-systems-in-space/>

4 Youssef Ejjiyar and Marc Lacoste. Which Agentic AI Architecture for Space Security? IEEE High Performance Extreme Computing Virtual Conference (HPEC), Sept. 2026. MIT, Boston, USA.

Au-delà du lien avec la cybersécurité, se pose donc la question plus générale d'une IA super intelligente pouvant devenir malfaisante voire menacer l'espèce humaine. Cette question se fait de plus en plus pressante. Notons par exemple l'initiative '*Pacing the Frontier*' ²⁷ de près de 1400 individus employés par les

27 <https://www.pacingthefrontier.com/>

Frontier AI model (Anthropic, OpenAi, Google, Meta...) demandant la mise en place d'une gouvernance internationale pour réguler les avancées dans le domaine du développement de l'IA.

Le sujet de l'IA va certainement continuer à nous préoccuper dans les prochains numéros. D'ailleurs, le premier numéro de 2027 consacrera un dossier sur « où va l'IA ? ». ■