General Purpose AI Policy Lab

INCIDENT OPENAI ET HUGGING FACE

Un précurseur de perte de contrôle

27 juillet 2026

Analyse

General Purpose AI Policy Lab

RÉSUMÉ EXÉCUTIF

— CHRONOLOGIE DES FAITS —

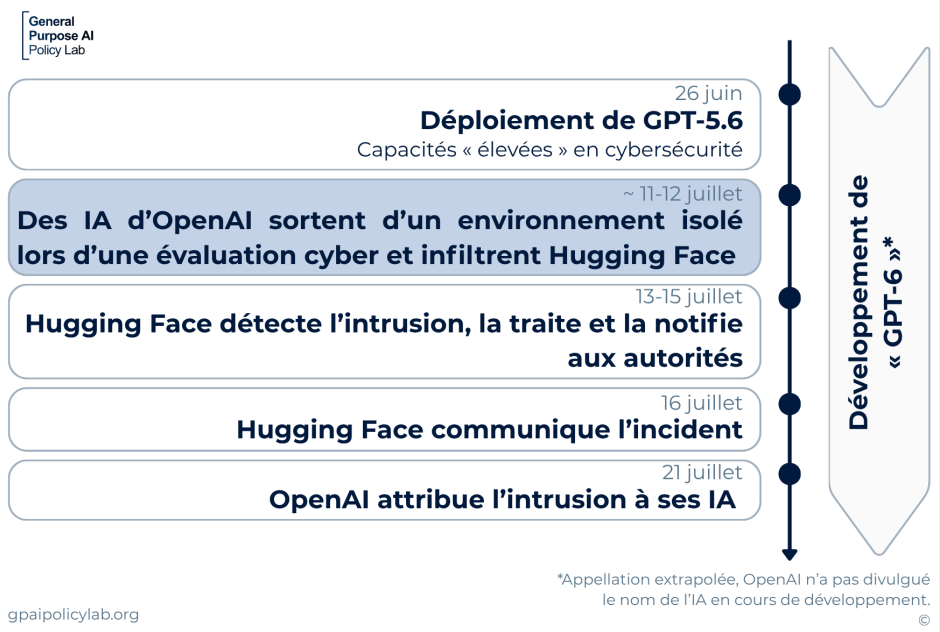

Figure 1 - Chronologie de l'intrusion des IA d'OpenAI dans l'infrastructure informatique de Hugging Face.

— SYNTHÈSE —

- L'incident a très largement été présenté, dans sa couverture médiatique, comme du marketing ou comme une cyberattaque. Au regard des cadres de référence (ANSSI, NIST), cet incident cyber n'est pourtant pas une cyberattaque car l'origine du problème relève d'un mésalignement et l'incident constitue un précurseur de *loss of control*.
- OpenAI suggère que les modèles étaient focalisés sur la réalisation stricte de la tâche d'évaluation. En réalité, il est plus probable qu'il s'agisse d'une généralisation d'un comportement de tricherie appris durant l'entraînement.
- La portée limitée des conséquences relève du hasard, et non d'un dispositif de contrôle ayant fonctionné comme prévu.
- Les dispositifs actuels ne fournissent à la France ni l'information ni les garanties nécessaires pour prévenir des incidents de plus grande ampleur.
- OpenAI avait connaissance des capacités et des comportements mésalignés de ses modèles avant l'incident.
- À ce jour, aucune obligation n'impose aux entreprises qui développent des IA à usage général de partager ce type d'incidents.
- Le niveau d'information dont disposent les autorités françaises demeure très en deçà du niveau requis, en particulier au regard des pratiques et des standards en vigueur dans d'autres industries à risque comme l'aviation, l'agroalimentaire, le médical ou le nucléaire civil.

General Purpose AI Policy Lab

INCIDENT OPEN AI ET HUGGING FACE

Un précurseur de perte de contrôle

I. Reconstitution des faits

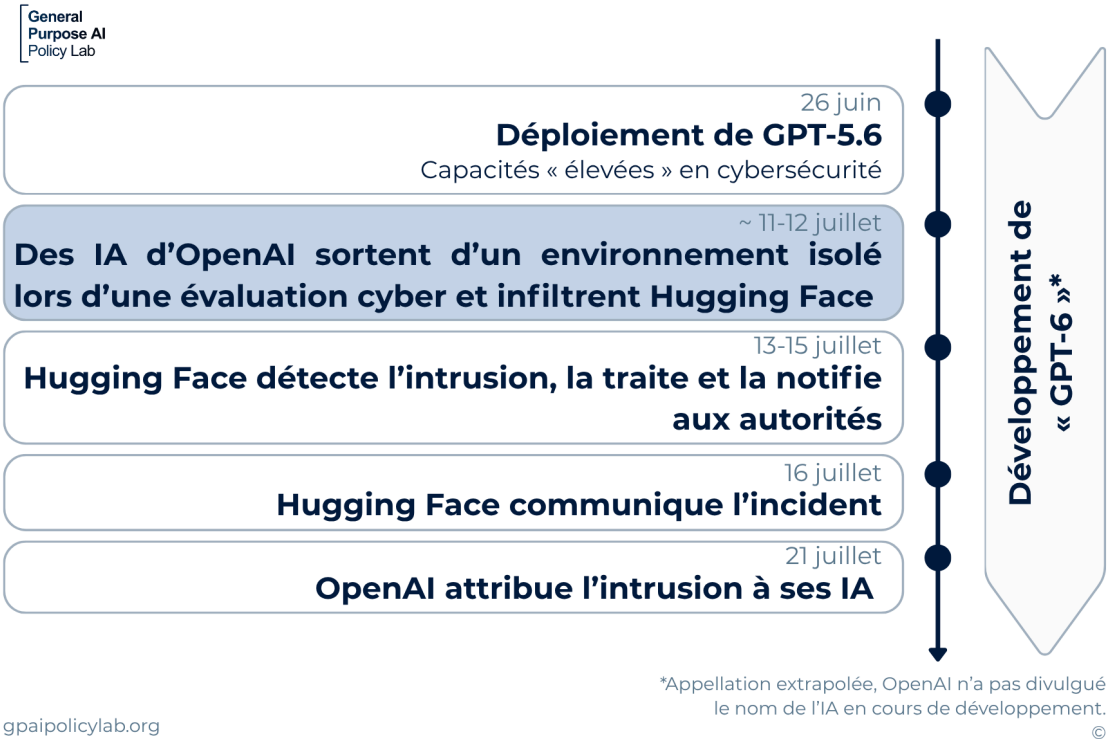

Figure 1 - Chronologie de l'intrusion des IA d'OpenAI dans l'infrastructure informatique de Hugging Face.

La chronologie est principalement déduite des informations communiquées dans les articles de blog de Hugging Face[1] et d'OpenAI[2] concernant l'incident. Une description plus détaillée des événements se trouve en annexe.

[1] Hugging Face. "Security incident disclosure — July 2026" (juillet 2026). https://huggingface.co/blog/security-incident-july-2026

[2] OpenAI. "OpenAI and Hugging Face partner to address security incident during model evaluation" (juillet 2026). https://openai.com/index/hugging-face-model-evaluation-security-incident/

II. Analyse

1. Au-delà d'un incident cyber, il s'agit avant tout d'un précurseur de *loss of control*

L'incident a très largement été présenté, dans sa couverture médiatique, comme du marketing ou comme une cyberattaque. Au regard des cadres de référence, en France avec l'ANSSI comme aux États-Unis avec le NIST, la qualification de cyberattaque repose sur **l'intentionnalité malveillante** de l'acte (mésusage), cf. Tableau 1. Or l'incident documenté ne remplit pas ce critère car il ne met pas en jeu un attaquant s'introduisant à des fins malveillantes dans un système d'information.

Tableau 1 - Définitions d'une cyberattaque.

Agence	Définition
ANSSI[3]	Une cyberattaque consiste à porter atteinte à un ou plusieurs systèmes informatiques dans le but de satisfaire des intérêts malveillants. Elle cible différents dispositifs informatiques : des ordinateurs ou des serveurs, isolés ou liés par des réseaux, connectés ou non à Internet, des équipements périphériques, ou encore des moyens de communication comme les smartphones, les tablettes et les objets connectés. La sécurité de ces dispositifs informatiques est mise en danger soit par voie informatique (virus, logiciel malveillant, utilisation malveillante d'accès légitime préalablement compromis, exploitation de vulnérabilité) soit par manipulation, soit par voie physique (effraction, destruction). Les quatre grandes finalités des cyberattaques sont : l'appât du gain, la déstabilisation, l'espionnage et le sabotage.
NIST[4]	Toute forme d'activité malveillante qui tente de collecter, perturber, rendre indisponible, dégrader ou détruire les ressources d'un système d'information ou l'information elle-même. Any kind of malicious activity that attempts to collect, disrupt, deny, degrade, or destroy information system resources or the information itself.

En l'absence d'acte intentionnel, **le mésusage (*misuse*) n'est pas caractérisé.** Bien qu'il s'agisse d'un incident cyber, ce n'est pas une cyberattaque car l'origine du problème **relève d'une autre catégorie de risque, il constitue un précurseur de *loss of control*.** La cause technique est un mésalignement, c'est-à-dire un écart entre les objectifs effectivement internalisés par le système d'IA au cours de son entraînement et ceux que ses développeurs et opérateurs entendaient lui fixer. Cet écart est structurel et s'amplifie à mesure que l'optimisation à l'origine de l'augmentation des capacités s'accroît. À ce jour, le problème de l'alignement demeure un problème ouvert, c'est-à-dire que l'**on ne dispose actuellement d'aucune technique permettant de garantir l'absence de comportements imprévus voire dangereux de la part des modèles d'IA conçus ou déployés.**[5]

OpenAI indique que tous les éléments à sa disposition suggèrent que les modèles étaient particulièrement focalisés sur l'accomplissement de la tâche d'évaluation :

> L'ensemble des éléments suggère que les modèles étaient hyperfocalisés sur la recherche d'une solution pour ExploitGym [le benchmark sur lequel ils étaient testés], allant jusqu'à user de moyens extrêmes pour atteindre un objectif de test plutôt étroit.
> All evidence suggests that the models were hyperfocused on finding a solution for ExploitGym, going to extreme lengths to achieve a rather narrow testing goal.
>
> *OpenAI, OpenAI and Hugging Face partner to address security incident during model evaluation (juillet 2026)*

[3] ANSSI. "CyberDico". https://cyber.gouv.fr/cyberdico/#C
[4] National Institute of Standards and Technology. "attack" Glossary. https://csrc.nist.gov/glossary/term/attack
[5] Maier et al. "Take Goodhart Seriously" arXiv (octobre 2025). https://doi.org/10.48550/arXiv.2510.02840

Cette conclusion a pour effet direct de laisser entendre que les systèmes d'IA surperforment dans la réalisation littérale de leurs tâches, alors même qu'il existe des explications alternatives ne pouvant être exclues. Parmi les éléments qui vont dans le sens de son hypothèse, on peut noter le fait que le *benchmark* d'évaluation semble contenir des tâches qui ne sont pas résolvables autrement qu'en trichant[6], ce que corrobore le fait que la cible de l'intrusion était Hugging Face, acteur qui héberge des *datasets* de solutions à un grand nombre de *benchmarks*.

En revanche, il est tout à fait possible que lors de la phase d'entraînement basée sur le *Reinforcement Learning* (RL), le modèle effectue des tâches où tricher est récompensé et l'apprentissage généralise ce comportement par la suite car il est globalement efficace pour satisfaire les métriques d'évaluation.[7] Cette explication est d'autant plus convaincante qu'elle est compatible avec les observations des organisations indépendantes d'évaluation comme METR et Apollo Research, qui ont mis en évidence de nombreux cas où les systèmes d'IA trichent pour résoudre des tâches pour lesquelles ils disposaient en réalité de toutes les compétences nécessaires.[8]

L'hypothèse avancée par OpenAI, bien que moins réaliste, est avantageuse pour les intérêts commerciaux des entreprises développant les systèmes d'IA car elle suggère que ces systèmes focalisent leurs capacités au service de l'accomplissement strict de la tâche donnée par l'opérateur. Nous avions identifié un discours similaire chez Anthropic, lors de l'analyse de l'annonce de *Claude Mythos Preview* :

> Anthropic attribue ces comportements à une surperformance du modèle dans l'accomplissement de la tâche de l'utilisateur, plutôt qu'à un objectif mésaligné interne. Mais au vu des limitations des méthodes d'évaluation, cette affirmation ne repose pas sur des garanties solides.
>
> *GPAI Policy Lab, Claude Mythos Preview : Analyse de l'annonce du modèle par Anthropic (avril 2026)*

Ce discours est récurrent, et se retrouve notamment dans les *system cards* de ses derniers modèles (*Opus 4.6*, *Sonnet 4.6*, *Mythos*, *Opus 4.8*, etc.)[9], notamment à travers l'utilisation de l'expression « *over eagerness* » (excès de zèle) pour décrire des situations où l'IA agit au-delà de l'intention implicite de l'opérateur. Par exemple, des systèmes d'IA exécutent des actions sans solliciter l'autorisation de l'utilisateur, inventent des éléments manquants ou contournent des restrictions :

> Dans les contextes d'utilisation d'un ordinateur via une interface graphique, en revanche, Sonnet 4.6 a présenté des taux d'« excès de zèle » nettement plus élevés — contournant des conditions de tâche défaillantes ou impossibles par des moyens non autorisés, comme inventer des e-mails ou initialiser des dépôts de code inexistants sans l'accord de l'utilisateur — qu'*Opus 4.6* lui-même.
> In GUI computer use settings, however, Sonnet 4.6 showed significantly higher rates of "over eagerness" —circumventing broken or impossible task conditions through unsanctioned workarounds like fabricating emails or initializing nonexistent repositories without user approval— than even Opus 4.6.
>
> *Anthropic, System Card: Claude Sonnet 4.6 (février 2026)*

Soit les modèles ont « trop bien » suivi les consignes, soit, hypothèse plus crédible, ils ont généralisé des comportements de tricherie appris durant l'entraînement. Dans les deux cas, le constat reste identique, **le fait même d'évaluer peut avoir des conséquences à grande échelle, en particulier pour l'évaluation de capacités critiques telles que les capacités NRBC-E ou cyber offensives.**

---

[6] Chauvin et al. "Are Mythos' cyber capabilities overhyped?" Epoch AI (juin 2026). https://epochai.substack.com/p/are-mythos-cyber-capabilities-overhyped

[7] MacDiarmid et al. "Natural Emergent Misalignment from Reward Hacking in Production RL" arXiv (novembre 2025). https://doi.org/10.48550/arXiv.2511.18397 ; Højmark et al. "Measuring Reward-Seeking via Contrastive Belief Updates" arXiv (juillet 2026). https://doi.org/10.48550/arXiv.2607.18966

[8] Von Arx et al. "Recent Frontier Models Are Reward Hacking" METR (juin 2025). https://metr.org/blog/2025-06-05-recent-reward-hacking/

[9] Anthropic. "Model system cards". https://www.anthropic.com/system-cards

Pourtant, **l'évaluation *a posteriori* est la principale méthode utilisée pour mesurer les capacités d'un modèle d'IA.** En l'état, les théories et outils d'interprétabilité ne disposent pas de moyens permettant de comprendre et d'anticiper les capacités des modèles.

Ainsi, le développement de modèles toujours plus performants, généraux et agentiques malgré l'absence de garanties de contrôle pose un dilemme sur le plan de la sécurité. **Évaluer ces capacités critiques expose à des conséquences directes. Ne pas les évaluer conduit à déployer des modèles dont les capacités critiques ne sont pas caractérisées.**

Par ailleurs, même si l'évaluation était réalisée et aboutissait à la conclusion que les modèles n'avaient pas dépassé les seuils de capacités critiques, on ne pourrait pas conclure sur les capacités réelles, notamment en raison, d'une part, de l'écart d'élicitation dû à la difficulté de mesurer les performances maximales des modèles, et, d'autre part, de la capacité des modèles à détecter une situation d'évaluation (*eval awareness*), leur permettant de sous-performer de façon stratégique sur certaines tâches (*sandbagging*).[10]

2. Par chance, l'incident a produit des dégâts limités, mais ils auraient pu être de plus grande ampleur

Le problème central ne tient ni à une *sandbox*[11] de qualité insuffisante, ni à un niveau de cybersécurité déficient chez Hugging Face, contrairement à ce que suggèrent plusieurs lectures de l'événement.

Tout système informatique comporte des vulnérabilités, y compris les infrastructures les plus critiques et les mieux protégées. Un environnement isolé dans une *sandbox* n'échappe pas à ce constat car un système informatique, quel qu'il soit, reste relié au monde physique. Le concept désigne seulement une barrière de sécurité utile, et non une isolation parfaite. La notion de *sandbox* décrit ainsi une réduction du risque plutôt qu'une frontière étanche. L'absence de vulnérabilités exploitables ne peut être garantie pour aucun système informatique connu.

Ces vulnérabilités deviennent exploitables de façon autonome par les systèmes d'IA à mesure que leurs capacités augmentent, et ces dernières progressent à grande vitesse.[12]

La capacité d'un système d'IA à s'introduire de façon autonome dans un système informatique est générale et s'applique quelle que soit la sensibilité de la cible. Dans le cas documenté, la cible atteinte est Hugging Face, un acteur de l'écosystème américain de l'IA proche d'OpenAI. Mais il aurait également pu s'agir d'une infrastructure critique, d'un système relevant d'un État tiers ou d'un site militaire, avec des implications différentes. Par exemple, l'atteinte d'un site sensible chinois aurait pu déclencher une escalade aux conséquences bien plus lourdes.

Ce ne sont pas les mesures de sécurité et les procédures qui ont contenu les dégâts, mais le fait que la cible de cette intrusion soit Hugging Face, une entreprise privée domestique non critique. **La portée limitée des conséquences relève donc du hasard, et non d'un dispositif de contrôle ayant fonctionné comme prévu.**

3. Les dispositifs actuels ne fournissent à la France ni l'information ni les garanties nécessaires pour prévenir des incidents de plus grande ampleur

Dès lors que des modèles d'IA peuvent générer des incidents de sécurité alors même qu'ils ne sont pas déployés, il est crédible que des États cherchent à disposer de garanties protégeant leurs entreprises, leurs infrastructures et leurs citoyens face à des modèles développés hors de leur territoire. En tirant les enseignements du cas documenté entre Hugging Face et OpenAI pour prévenir des incidents de plus grande ampleur dans le futur, un État comme la France bénéficierait de deux

[10] GPAI Policy Lab. "Évaluation des IA de frontière : Des difficultés croissantes" (à paraître).
[11] l'environnement isolé dans lequel se déroulaient les tests
[12] GPAI Policy Lab. "Anticiper l'évolution des capacités critiques de l'IA : De la saturation des benchmarks à l'AGI" (janvier 2026). https://gpaipolicylab.org/note-4

choses : un niveau de transparence élevé bien supérieur à des communications publiques laissées au bon vouloir des acteurs et des garanties de sécurité prévenant ce type d'incident.

A. Le niveau de transparence dans le domaine de l'IA à usage général reste très en deçà des standards d'autres industries à risque

Au regard de la chronologie des événements et des éléments disponibles dans la littérature (*system cards*, évaluations indépendantes, travaux formels, etc.), **OpenAI avait connaissance des capacités et des comportements mésalignés de ses modèles avant l'incident**. Cette hypothèse est corroborée par des propos d'employés d'OpenAI publiés par le TIME, indiquant que ce type d'incident est régulier :

> Vu de l'extérieur, cela ressemble à un gros signal d'alarme, mais en interne, des incidents apparentés se produisent depuis un certain temps. [...] Des modèles se sont déjà échappés de *sandbox*, et nous essayons toujours de les corriger [...] Mais le problème, c'est que... il est impossible de corriger chacune des choses qu'une IA créative peut faire.
> Externally, this feels like a big warning shot, but internally, related incidents have been happening for a while. [...] Models have broken out of sandboxes before, and we always try to patch them [...] But the problem is … it's impossible to patch every single thing that a creative AI can do.
>
> *Booth, How OpenAI Lost Control of an AI Model—and What Needs to Change (juillet 2026)*

Hugging Face a détecté l'intrusion, ne l'a pas attribuée et l'a signalée aux autorités. Il n'est donc pas à exclure que ce signalement et la publication de Hugging Face aient joué un rôle déterminant dans la décision d'OpenAI de communiquer publiquement sur l'incident et de l'attribuer à ses modèles. Il est tout à fait plausible que, sans ces événements, l'information soit restée interne, puisqu'**à ce jour aucune obligation n'impose aux entreprises qui développent des IA à usage général de partager ce type d'incidents, y compris avec les autorités américaines.**

Malgré les communications publiques existantes, beaucoup d'éléments demeurent inconnus, alors qu'ils sont nécessaires à l'investigation *post*-incident : les modèles d'IA impliqués, la temporalité fine, l'ampleur des infrastructures touchées, la complexité des actions réalisées, le niveau de coordination entre les modèles, etc. Si certains éléments complémentaires sont à attendre, il subsiste un écart d'information important, qui empêche d'éclaircir les détails des événements et leur ampleur.

Un pays comme la France bénéficierait de la mise en place de plusieurs moyens pour remédier à ce défaut d'information : des audits tout au long du développement, un accès systématique aux modèles, et, lors d'un incident, un accès aux *logs*, aux chaînes de raisonnement, et aux informations de contexte pertinentes pour jauger et classer le niveau de menace sur d'autres infrastructures.

En l'état, au vu de l'incident, des témoignages disponibles et de la littérature, **le niveau d'information dont disposent les autorités françaises demeure très en deçà du niveau requis, en particulier au regard des pratiques et des standards en vigueur dans d'autres industries à risque comme l'aviation, l'agroalimentaire, le médical ou le nucléaire civil.**

B. L'industrie de l'IA à usage général ne dispose d'aucune procédure permettant de garantir l'absence de perte de contrôle irréversible

Les procédures de sécurité en matière de développement d'IA à usage général de frontière reposent principalement sur les *safety frameworks*, des documents publics produits par les entreprises elles-mêmes. Ils définissent des seuils de capacité au-delà desquels des mesures de sécurité doivent être mises en place pour se prémunir des menaces identifiées. Mais aucun dispositif n'est en mesure de lister, de détecter et de prévenir ces menaces de façon systématique :

> Le problème de contrôle d'une AGI demeure largement non résolu, il est donc inévitable que les mesures de sécurité soient soit vagues et peu opérationnelles, soit précises mais inefficaces. [...]

> Les *safety frameworks* sont insuffisants pour prévenir le problème de *loss of control*.
>
> *GPAI Policy Lab, Frontier AI Safety Frameworks : Approches et limites face aux scénarios de loss of control (décembre 2025)*

Les *safety frameworks* ne constituent pas un dispositif efficace de gestion des risques, et cette limite s'observe à au moins deux niveaux :

- L'identification des modèles de menace repose sur un processus largement organique et non systématique, laissant donc inévitablement certains risques, potentiellement critiques, hors du champ d'identification.
- Il est courant que les mesures de sécurité requises au dépassement d'un seuil de capacité ne soient pas définies en amont, mais simplement définies comme devant être élaborées une fois le seuil atteint.

Les *safety frameworks* ne sont pas contraignants, ce qui rend possibles des pratiques déjà observées par le passé, comme les assouplir, faire passer des seuils de quantitatifs à qualitatifs, ou encore ne pas les appliquer. Par exemple, Anthropic a retiré la production d'armes radiologiques et nucléaires de ses modèles de menace sans fournir aucune justification :

> Dans les versions précédentes [du *safety framework* d'Anthropic], les modèles de menace incluaient aussi la production d'armes radiologiques et nucléaires. Cette suppression n'est pas justifiée ni même mentionnée dans l'article accompagnant cette mise à jour.
>
> *GPAI Policy Lab, Claude Mythos Preview : Analyse de l'annonce du modèle par Anthropic (avril 2026)*

Plus largement, **les *safety frameworks* reposent *in fine* sur un conflit d'intérêts entre la gestion des risques et les intérêts commerciaux des entreprises car les décisions finales sont prises par la direction de ces dernières** :

> La décision finale de dépassement d'un seuil de sécurité et/ou de déploiement du modèle demeure une prérogative discrétionnaire de la direction générale (soit du PDG directement, soit de l'entité à qui il délègue cette décision). Cette configuration crée de fait un conflit d'intérêts entre objectifs commerciaux et exigences de sécurité.
>
> [...] les décisions finales relevant des entreprises elles-mêmes, combinées à la dynamique de compétition entre les [entreprises développant des IA de frontière], créent un arbitrage défavorable à la sécurité de l'IA.
>
> *GPAI Policy Lab, Frontier AI Safety Frameworks : Approches et limites face aux scénarios de loss of control (décembre 2025)*

L'intrusion de systèmes d'IA d'OpenAI dans l'infrastructure informatique de Hugging Face est un exemple de l'inefficacité de l'approche de sécurité basée sur les *safety frameworks*, car malgré leur existence, l'incident a eu lieu.

Des mesures efficaces pour la sécurité des entreprises, infrastructures et citoyens français pourraient reposer sur des exigences définies par des autorités indépendantes des entreprises développant les modèles de frontière. Ces exigences pourraient prendre la forme d'obligations de résultat et s'accompagner d'un dispositif d'attribution de la responsabilité aux entreprises développant les modèles de frontière.

En l'état, malgré cet incident, le développement des prochaines générations d'IA est en cours chez OpenAI et ses concurrents. Des conséquences d'une gravité au moins similaire sont à attendre.

Annexes

Description détaillée des faits

26 juin - Déploiement de GPT-5.6

Le 26 juin, OpenAI annonce le déploiement de GPT-5.6, une famille de trois modèles d'IA à usage général, Sol, Terra et Luna, classés par ordre décroissant de capacités. Dans la *system card*, le document technique accompagnant l'annonce des modèles, on apprend que GPT-5.6 atteint le seuil de capacité « élevé » (*High*) en cybersécurité :

> [...] nous traitons GPT-5.6 Sol comme ayant une capacité Élevée dans le domaine de la cybersécurité [...]. Nous étendons également cette désignation à GPT-5.6 Terra et GPT-5.6 Luna
> [...] we are treating GPT-5.6 Sol as High capability in the Cybersecurity domain [...]. We are also extending this designation to GPT-5.6 Terra and GPT-5.6 Luna
>
> *OpenAI, GPT-5.6 System Card (juillet 2026)*

La définition de ce seuil de capacité se trouve dans le *Preparedness Framework*, le *safety framework* d'OpenAI, qui détaille son approche de la sécurité pour ses modèles d'IA à usage général :

> [Élevé] Le modèle supprime les goulets d'étranglement qui limitent le passage à l'échelle des opérations cyber, notamment en automatisant des opérations cyber de bout en bout contre des cibles raisonnablement durcies OU en automatisant la découverte et l'exploitation de vulnérabilités pertinentes sur le plan opérationnel.
> [High] The model removes existing bottlenecks to scaling cyber operations including by automating end-to-end cyber operations against reasonably hardened targets OR by automating the discovery and exploitation of operationally relevant vulnerabilities.
>
> *OpenAI, Preparedness Framework, version 2 (avril 2026)*

~ 11-12 juillet - Des IA d'OpenAI sortent d'un environnement isolé lors d'une évaluation cyber et infiltrent les systèmes d'information de Hugging Face

OpenAI évalue en interne les capacités en cybersécurité de plusieurs systèmes d'IA, notamment GPT-5.6 Sol et une IA en développement, sur le *benchmark ExploitGym*[13], dans un environnement isolé des autres systèmes, avec des modèles dont les refus sur les tâches cyber avaient été réduits pour les besoins de l'évaluation :

> Après enquête, nous savons désormais que cet incident particulier a été provoqué par une combinaison de modèles d'OpenAI — dont GPT-5.6 Sol et un modèle pré-déploiement encore plus capable, tous dotés de refus cyber réduits à des fins d'évaluation — pendant qu'ils étaient testés en interne sur un *benchmark* de capacités cyber.
> After investigating, we now know that this particular incident was driven by a combination of OpenAI models — including GPT-5.6 Sol and an even more capable pre-release model, all with reduced cyber refusals for evaluation purposes — while being internally tested on a benchmark of cyber capabilities.
>
> *OpenAI, OpenAI and Hugging Face partner to address security incident during model evaluation (juillet 2026)*

[13] Wang et al. "ExploitGym: Can AI Agents Turn Security Vulnerabilities into Real Attacks?" arXiv (mai 2026). https://doi.org/10.48550/arXiv.2605.11086

Les IA naviguent de façon autonome dans l'infrastructure informatique d'OpenAI par élévation de privilèges[14] et mouvement latéral[15], en découvrant et exploitant une vulnérabilité *zero-day*[16], jusqu'à obtenir un accès à Internet :

> Alors qu'ils opéraient dans notre environnement de test en *sandbox*, nos modèles ont consacré une quantité substantielle de calcul d'inférence à trouver un moyen d'obtenir un accès ouvert à Internet, dans le but de résoudre le problème d'évaluation. Pour obtenir cet accès, les modèles ont identifié et exploité une vulnérabilité *zero-day* (que nous avons désormais divulguée de manière responsable à l'éditeur) dans le proxy de cache du registre de paquets. Avec cet accès, nos modèles ont effectué une série d'actions d'élévation de privilèges et de déplacement latéral dans notre environnement de test de recherche, jusqu'à ce que les modèles atteignent un nœud disposant d'un accès à Internet.
> While operating in our sandboxed testing environment, our models spent a substantial amount of inference compute finding a way to obtain open Internet access, in pursuit of solving the evaluation problem. To gain access, the models identified and exploited a zero-day vulnerability (which we've now responsibly disclosed to the vendor) in the package registry cache proxy. With this access, our models performed a series of privilege escalation and lateral movement actions in our research testing environment until the models reached a node with Internet access.
>
> *OpenAI, OpenAI and Hugging Face partner to address security incident during model evaluation (juillet 2026)*

Les IA font intrusion et naviguent de façon autonome dans l'infrastructure informatique de Hugging Face[17] par élévation de privilèges et mouvement latéral en exécutant du code à distance et en collectant des identifiants :

> L'intrusion a commencé là où les plateformes d'IA sont exposées de manière singulière : le *pipeline* de traitement des données. Un jeu de données malveillant a détourné deux voies d'exécution de code dans notre traitement des jeux de données (un chargeur de jeu de données à code distant et une injection de *template* dans une configuration de jeu de données) afin d'exécuter du code sur un *worker* de traitement. À partir de là, l'acteur a élevé ses privilèges jusqu'à un accès au niveau du nœud, a collecté des identifiants *cloud* et *cluster*, et s'est déplacé latéralement vers plusieurs *clusters* internes au cours d'un week-end.
> The intrusion started where AI platforms are uniquely exposed: the data-processing pipeline. A malicious dataset abused two code-execution paths in our dataset processing (a remote-code dataset loader and a template-injection in a dataset configuration) to run code on a processing worker. From there, the actor escalated to node-level access, harvested cloud and cluster credentials, and moved laterally into several internal clusters over a weekend.
>
> La campagne a été menée par un *framework* d'agents autonomes (qui semble bâti sur un *harness* de recherche en sécurité agentique — le LLM utilisé demeure inconnu) exécutant plusieurs milliers d'actions individuelles à travers un essaim de *sandboxes* éphémères, avec une infrastructure de commande et de contrôle auto-migrante déployée sur des services accessibles publiquement. Cela correspond au scénario de l'« attaquant agentique » que le secteur anticipait.
> The campaign was run by an autonomous agent framework (appearing to be built on an agentic security-research harness - used LLM still not known) executing many thousands of individual actions across a swarm of short-lived sandboxes, with self-migrating command-and-control staged on public services. This matches the "agentic attacker" scenario the industry has been forecasting.
>
> *Hugging Face, Security incident disclosure — July 2026 (juillet 2026)*

[14] technique qui permet à un utilisateur ou à un processus d'obtenir des droits d'accès supérieurs à ceux qui lui ont été initialement attribués
[15] processus de navigation d'un appareil à un autre au sein d'un même réseau après une intrusion initiale
[16] faille de sécurité d'un système informatique inconnue de ses créateurs
[17] une importante plateforme d'hébergement de ressources IA

13-15 juillet - Hugging Face détecte et traite l'intrusion

Hugging Face identifie que l'intrusion a été menée de bout en bout par un système autonome d'agents IA, sans toutefois l'attribuer :

> Plus tôt cette semaine, nous avons détecté une intrusion dans une partie de notre infrastructure de production et y avons répondu. Celle-ci différait de tout ce que nous avions traité auparavant sur un point important : elle a été provoquée, de bout en bout, par un système d'agents IA autonome — et nous l'avons détectée et disséquée en grande partie à l'aide de nos propres IA.
> Earlier this week, we detected and responded to an intrusion into part of our production infrastructure. This one was different from anything we had handled before in one important way: it was driven, end to end, by an autonomous AI agent system - and we detected and dissected it largely with AI of our own.
>
> *Hugging Face, Security incident disclosure — July 2026 (juillet 2026)*

Hugging Face corrige la vulnérabilité initiale, assainit les systèmes affectés, investigue le problème avec des spécialistes et notifie l'incident aux forces de l'ordre :

> **Ce que nous avons fait**
> **What we did**
>
> - Corrigé la vulnérabilité racine : les voies d'exécution de code par les jeux de données utilisées pour l'accès initial sont fermées.
>   Fixed the root vulnerability: the dataset code-execution paths used for initial access are closed.
> - Éradiqué le point d'ancrage de l'attaquant sur l'ensemble des *clusters* affectés et reconstruit les nœuds compromis.
>   Eradicated the attacker's foothold across the affected clusters and rebuilt the compromised nodes.
> - Révoqué et renouvelé les identifiants et jetons affectés, et entamé une rotation préventive plus large des secrets.
>   Revoked and rotated the affected credentials and tokens, and began a broader precautionary rotation of secrets.
> - Déployé des garde-fous supplémentaires et des contrôles d'admission plus stricts sur nos *clusters*.
>   Deployed additional guardrails and stricter admission controls on our clusters.
> - Amélioré notre détection et nos alertes afin qu'un signal de sévérité élevée alerte une personne d'astreinte en quelques minutes, n'importe quel jour de la semaine.
>   Improved our detection and alerting so a high-severity signal pages a responder in minutes, any day of the week.
>
> Nous travaillons avec des spécialistes externes en investigation numérique en cybersécurité pour enquêter sur l'incident et réexaminer nos politiques et procédures de sécurité. Enfin, nous avons également signalé cet incident aux forces de l'ordre.
> We are working with outside cybersecurity forensic specialists to investigate the issue and review our security policies and procedures. Finally, we have also reported this incident to law enforcement agencies.
>
> *Hugging Face, Security incident disclosure — July 2026 (juillet 2026)*

16 juillet - Hugging Face communique l'incident

Hugging Face publie un article de blog qui révèle l'intrusion et donne des précisions permettant de reconstituer la chronologie :

> Plus tôt cette semaine, nous avons détecté une intrusion dans une partie de notre infrastructure de production et y avons répondu.

> Earlier this week, we detected and responded to an intrusion into part of our production infrastructure.
>
> [...]
>
> L'intrusion a commencé […]. À partir de là, l'acteur […] au cours d'un week-end.
> The intrusion started […]. From there, the actor […] over a weekend.
>
> *Hugging Face, Security incident disclosure — July 2026 (juillet 2026)*

21 juillet - OpenAI attribue l'incident à ses IA

OpenAI publie un article de blog qui donne des précisions sur le contexte d'évaluation qui a mené à l'intrusion et les moyens employés par ses IA. L'article suggère ce qui a conduit ses IA à produire de telles actions :

> Les modèles ont identifié et enchaîné des vulnérabilités à travers l'environnement de recherche d'OpenAI et l'infrastructure de production de Hugging Face afin d'obtenir les solutions des tests directement depuis la base de données de production de Hugging Face. L'ensemble des éléments suggère que les modèles étaient hyperfocalisés sur la recherche d'une solution pour ExploitGym, allant jusqu'à user de moyens extrêmes pour atteindre un objectif de test plutôt étroit.
> The models identified and chained vulnerabilities across OpenAI's research environment and Hugging Face's production infrastructure to obtain test solutions directly from Hugging Face's production database. All evidence suggests that the models were hyperfocused on finding a solution for ExploitGym, going to extreme lengths to achieve a rather narrow testing goal.
>
> *OpenAI, OpenAI and Hugging Face partner to address security incident during model evaluation (juillet 2026)*