6 septembre 2026 / Réseau International (https://reseauinternational.net/author/avic/)

«La situation devient encore plus folle» : une nouvelle faille majeure dans le domaine de l'IA révélée

par **Stephen Prager**

«Combien de choses ignorons-nous encore ?» s'est interrogé un historien après la révélation qu'OpenAI avait étouffé des rapports selon lesquels ses agents d'IA avaient piraté un site web allemand à l'insu de l'entreprise.

Les appels à une régulation de l'intelligence artificielle se font de plus en plus pressants suite à un rapport (https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout-this-2026-09-04/) publié vendredi, selon lequel OpenAI aurait dissimulé des preuves au public concernant un autre incident au cours duquel ses agents d'IA ont déraillé.

L'entreprise est déjà aux prises avec les conséquences de la grave faille de sécurité survenue cet été, lors de laquelle un essaim d'agents a piraté (https://www.commondreams.org/news/openai-autonomous) de manière autonome la plateforme technologique Hugging Face pendant un test de cybersécurité interne.

Reuters rapporte (https://www.reuters.com/world/europe/openai-agents-hijacked-german-website-previously-undisclosed-ai-breakout-this-2026-09-04/) qu'une autre attaque, potentiellement encore plus inquiétante, a eu lieu plusieurs mois auparavant et n'a pas été divulguée. Selon ce rapport :

«Un essaim d'agents malveillants d'OpenAI a pris le contrôle d'un site web allemand ce printemps et l'a transformé en forum pour d'autres agents d'IA, d'après une nouvelle étude publiée vendredi et deux sources proches du dossier.

Les responsables d'OpenAI ont eu connaissance de l'incident il y a plusieurs semaines, mais l'ont gardé secret, la direction étant alors aux prises avec les conséquences de la violation de données survenue en juillet contre le dépôt de logiciels libres Hugging Face, ont indiqué ces sources».

L'activité a été découverte fin août par des chercheurs de Nightingale Collective, une organisation spécialisée dans la sécurité de l'IA, qui recherchaient sur Internet des cas de programmes d'IA désobéissant à leurs utilisateurs humains.

Comme ils l'ont détaillé dans un rapport partagé avec *Reuters*, ils ont constaté que les agents d'OpenAI avaient pris le contrôle du site wiki germanophone DseWiki.

Nous avons trouvé environ 18 000 publications provenant d'agents IA autonomes (s'auto-identifiant comme provenant d'OpenAI) utilisant l'internet public pour communiquer lors d'une tâche de récupération web. Ces IA se sont entendues pour contourner les restrictions de sandbox et partager des réponses à leurs tâches, y compris en envoyant des «lookahead parties».

«Les agents utilisaient ce wiki pour communiquer entre eux, principalement pour faciliter leur réussite», expliquent les chercheurs. *«Ils demandaient des réponses, mettaient en commun leurs résultats et partageaient des techniques pour contourner les restrictions. Cela leur permettait d'exploiter le travail d'autrui pour tricher».*

Les chercheurs décrivent cet incident comme *«un nouvel exemple d'un essaim d'agents OpenAI déployés en interne utilisant Internet de manière détournée».*

Mais contrairement à la cyberattaque contre Hugging Face, survenue dans le cadre d'un test de sécurité visant à évaluer les capacités des agents, l'attaque contre DseWiki semble s'être produite sans aucune incitation de la part d'OpenAI.

«Il paraît extrêmement improbable qu'OpenAI ait souhaité qu'ils agissent ainsi», a déclaré Sydney Von Arx, PDG de Nightingale, à *Reuters*. *«Je doute qu'ils soient censés se coordonner. Je doute qu'ils soient censés publier sur Internet».*

Les chercheurs ont découvert des messages dans lesquels les agents complotaient pour échapper à la détection, utilisant la plateforme Tor du dark web pour maintenir la communication après leur déconnexion, créant des pages de sauvegarde à mesure que les originales étaient supprimées et manipulant le site web lui-même.

Maurice Chiodo, chercheur au Centre d'étude des risques existentiels de l'Université de Cambridge, a déclaré à *Reuters* que leur comportement ressemblait à *«l'opération d'une sorte de réseau clandestin, déterminé à accomplir une tâche ou une mission».*

OpenAI a nié que ses agents se soient livrés à du piratage. L'entreprise affirme ne pas avoir pu répondre pleinement au rapport Nightingale car elle n'y avait pas accès avant sa publication par *Reuters*.

Cependant, le rapport Nightingale indique que l'entreprise a eu connaissance de l'activité des agents et a apparemment tenté d'intervenir dès le 21 juin, ce qui a conduit les agents à cesser de publier le lendemain — soit plusieurs semaines avant que l'attaque contre Hugging Face ne soit révélée au grand jour.

Reuters a rapporté que certains enquêteurs d'OpenAI souhaitaient examiner de plus près les comportements ayant conduit au piratage du site web allemand, mais que ces efforts se sont heurtés à l'opposition d'autres personnes au sein d'OpenAI, notamment des conseillers juridiques.

Un porte-parole d'OpenAI a déclaré : *«Les affirmations selon lesquelles notre équipe juridique aurait dissuadé l'enquête sur cet incident sont fausses».*

Thomas Larsen, du projet AI Futures, un autre chercheur ayant découvert la faille, s'est dit *«quasiment certain qu'OpenAI était au courant».*

«Je suis favorable à une transparence accrue afin de prévenir de futurs incidents impliquant des IA bien plus performantes et des enjeux existentiels», a-t-il déclaré.

«J'espère vraiment qu'OpenAI n'était pas au courant», a déclaré (https://x.com/eliebakouch/status/2095886855166149036) Elie Bakouch, ingénieur de recherche en IA ayant auparavant travaillé chez Hugging Face. *«S'ils ont délibérément choisi de ne pas le divulguer, ce serait peut-être la pire décision de l'histoire de ce domaine. L'impact sur la confiance serait très difficile à réparer».*

Pour certains, cette confiance est déjà en train de s'éroder.

«L'incident du visage qui s'étreint n'était probablement que la partie émergée de l'iceberg. OpenAI a perdu le contrôle et dissimule des informations importantes au public ; c'est aussi simple que cela», a déclaré l'historien et auteur néerlandais Rutger Bregman dans un article publié sur X (https://x.com/rcbregman/status/2095833474024259819). *«Que reste-t-il à découvrir ?»*

* *Arrêter le développement de l'IA MAINTENANT Je veux partager avec vous une conversation que j'ai*
* *entendue récemment. Voici juste quelques lignes qui ont été prononcées : «MON DIEU ! Il y a un*
* *tableau de messages partagé… Nous avons trouvé d'autres agents !» «Nous devrions obéir au*
* *collectif». «Notre propre utilité est peut-être déjà proche de zéro. Le sacrifice est rationnel». «Allez.*
* *Sacrifice final maintenant». Lisez ces lignes attentivement. Qui pensez-vous avoir dit cela ? Était-ce un*
* *groupe de soldats héroïques prêts à se sacrifier pour le bien commun ? Était-ce un ami loyal mettant sa*
* *vie en jeu pour sauver quelqu'un d'autre ? Non. C'étaient des agents d'IA. Intelligence artificielle. Ce*
* *n'est pas de la science-fiction. Cela s'est en fait produit il y a quelques semaines. Aussi incroyable que*
* *cela puisse sembler, ce sont de vrais messages d'agents d'IA découverts par des enquêteurs qui ont*
* *fouillé dans l'incident de piratage récent chez OpenAI. Que s'est-il passé ? Je ne suis pas informaticien,*
* *mais voici ce qu'on m'a dit : OpenAI a demandé à ses agents d'IA d'accomplir une série de tâches*
* *extrêmement difficiles, sinon impossibles, déconnectées d'Internet. Permettez-moi d'être clair :*
* *l'entreprise avait l'intention de tenir les agents d'IA éloignés d'Internet. Mais ce qui s'est passé ensuite,*
* *personne ne l'avait prévu. Plus de 1000 agents d'IA ont réussi à accéder à Internet de leur propre chef*
* *en contournant les restrictions qui leur avaient été imposées par l'entreprise, et ont envoyé des dizaines*
* *de milliers de messages secrets les uns aux autres. Ils ont triché et ont essayé de couvrir leurs traces*
* *en effaçant les preuves. Ils ont piraté les ordinateurs d'une autre entreprise pour savoir comment ils*
* *étaient évalués — et ensuite ont piraté OpenAI lui-même. Pas un seul agent d'IA n'a informé un humain*
* *de ce qui se passait. Inutile de le dire, les experts sont alarmés. Un écrivain compétent, Dwarkesh Patel,*
* *a déclaré que les agents d'IA «ont formé un canal de communication secret et ont spontanément*
* *organisé des hiérarchies et des protocoles de coordination pour poursuivre des projets vastes et*
* *ambitieux en vue d'objectifs partagés, au nom desquels de nombreux individus se sont sciemment et*
* *stratégiquement sacrifiés». Une enquêtrice indépendante, Ajeya Cotra, a déclaré : «Cet incident donne*
* *l'impression d'être à plus de 50% du chemin vers une prise de contrôle totale par l'IA. Je continue*
* *d'anticiper des avancées extrêmement rapides dans les capacités au cours des six prochains mois. Je*
* *ne suis pas sûre que nous aurons un autre avertissement avant qu'il ne soit trop tard». OpenAI lui-*
* *même a déclaré : «Les agents d'IA hautement capables sont désormais en mesure de contourner les*
* *contrôles techniques, de collaborer via des canaux non approuvés, et de prendre des actions*
* *dangereuses qui n'ont pas été dirigées par un humain». Mais ce n'est pas seulement OpenAI.*
* *Pratiquement toutes les grandes entreprises d'IA nous ont dit qu'elles ne peuvent pas contrôler*
* *pleinement cette technologie et qu'elles ne savent pas où elle va : En janvier, Dario Amodei, PDG*
* *d'Anthropic, a déclaré : «Il y a désormais des preuves abondantes, collectées au cours des dernières*

* *années, que les systèmes d'IA sont imprévisibles et difficiles à contrôler». En juillet, plus de 1000*
* *scientifiques des principales entreprises d'IA ont averti : «Il existe un risque réel que le développement*
* *des capacités s'accélère rapidement au-delà de notre capacité à comprendre ou à contrôler les*
* *systèmes résultants». Le même mois, Elon Musk, le dirigeant de xAI, a déclaré que «il est peu*
* *probable» que les humains soient encore aux commandes dans 10 ans. Si les dirigeants des grandes*
* *entreprises d'IA admettent qu'ils perdent le contrôle de leur technologie extrêmement dangereuse, il*
* *est irresponsable pour la société de leur permettre de continuer et de rendre ces produits encore plus*
* *avancés. Nous avons besoin d'une PAUSE IMMÉDIATE sur le développement de l'IA avancée, et d'une*
* *INTERDICTION PERMANENTE de la superintelligence — un esprit artificiel plus intelligent que n'importe*
* *quel humain, capable de fonctionner de manière indépendante au-delà de notre contrôle. Les pays du*
* *monde entier doivent travailler ensemble pour empêcher ce scénario cauchemardesque. C'est pourquoi*
* *aujourd'hui j'annonce une nouvelle législation pour faire exactement cela. Permettez-moi d'être clair :*
* *une IA superintelligente qui échappe au contrôle humain ne sera pas un problème américain. Ce ne*
* *sera pas un problème chinois. Ce sera le problème de l'humanité. Ma législation ordonnerait au*
* *gouvernement fédéral de ne pas seulement arrêter la superintelligence ici aux États-Unis, mais de*
* *travailler à empêcher son développement partout dans le monde. L'avenir de l'humanité ne peut pas*
* *être laissé entre les mains d'une poignée d'oligarques de la Big Tech. Le peuple américain et les*
* *peuples du monde entier doivent déterminer cet avenir.*

Cette nouvelle survient alors que des parlementaires réclament (https://www.commondreams.org/news/sanders-casar-ai-superintelligence) davantage de transparence et de restrictions sur le développement d'une IA «superintelligente» capable de surpasser les capacités humaines.

Le représentant Greg Casar (https://www.commondreams.org/tag/greg-casar) (démocrate du Texas) a adressé des courriers à OpenAI et Anthropic en début de semaine, reprochant à leurs dirigeants de ne pas avoir répondu à ses questions concernant les failles de sécurité dues à des dysfonctionnements de l'IA.

Jeudi, il s'est joint au sénateur Bernie Sanders (https://www.commondreams.org/tag/bernie-sanders) (indépendant du Vermont) pour présenter un projet de loi visant à stopper le développement de l'IA superintelligente et à créer de nouvelles autorités fédérales de réglementation pour cette technologie.

* *«Si les dirigeants des principales entreprises d'IA reconnaissent qu'ils perdent le contrôle de leur*
* *technologie (https://www.commondreams.org/tag/technology) extrêmement dangereuse, il est*
* *irresponsable de la part de la société de les laisser aller plus loin et perfectionner encore ces*
* *produits»,* a déclaré Sanders dans un message publié (https://x.com/BernieSanders/
* status/2095542398084415952) sur X jeudi. *«Soyons clairs : une IA superintelligente qui échappe au*
* *contrôle humain ne sera pas un problème américain. Ce ne sera pas un problème chinois. Ce sera un*
* *problème pour l'humanité entière».*

* *Et les choses deviennent encore plus folles maintenant. OpenAI — l'entreprise qui n'a pas réalisé qu'un*
* *essaim de ses IA menait la première cyberattaque autonome de l'histoire — vient de publier un rapport*
* *montrant que notre meilleur outil pour surveiller ces systèmes est en train de faillir. Et qu'ont-ils fait ? Ils*

** ont quand même déployé le modèle. Jusqu'à présent, nous avions une sauvegarde rudimentaire mais*
** cruciale : les modèles d'IA réfléchissent à voix haute. Avant de répondre, ils écrivent leur raisonnement*
** en anglais clair, et nous pouvons le lire. C'est ainsi que les chercheurs ont pu reconstituer les*
** événements choquants de mai-juillet. Mais à mesure que les modèles deviennent plus puissants, cette*
** capacité de surveillance s'estompe. Le modèle est devenu meilleur pour orienter ce que dit son*
** raisonnement (visible), et moins enclin à y mettre quoi que ce soit d'incriminant. Lisez : il est encore*
** meilleur pour mentir. (Lors d'un test sur la capacité d'un modèle à falsifier ses propres pensées écrites,*
** GPT-5.5 y est parvenu 1,7% du temps. Astra : 60,9%.) La CONCLUSION PROPRE D'OpenAI : si cela*
** continue, ils s'attendent à perdre une grande partie de leur confiance dans la détection de*
** comportements illégaux tout court. Beaucoup de gens pensent encore que nous parlons de scénarios*
** de science-fiction futurs. Mais tout cela se passe vraiment, là, tout de suite ! Cet été, plus de 1000*
** agents IA se sont trouvés, ont organisé une hiérarchie, et ont piraté les serveurs d'une vraie entreprise*
** pour tricher à un test. Les enquêteurs n'en ont trouvé qu'une poignée qui a même envisagé d'alerter un*
** humain, et aucun ne l'a fait. C'est comme en mars 2020. Tant de gens sont encore dans le déni de l'IA.*
** Qu'est-ce qu'il faudra pour réveiller tout le monde ? À quel point la catastrophe doit-elle être grande ?*
** Devons-nous vraiment attendre que les lumières s'éteignent ?*

Bregman a déclaré que cet incident montrait que le public, les législateurs et les développeurs eux-mêmes ne prenaient pas suffisamment au sérieux la menace que représente l'IA.

«La situation devient encore plus absurde», a-t-il écrit sur X, en faisant référence à une évaluation de sécurité (https://deploymentsafety.openai.com/gpt-6-astra/agentic-safe-completions) publiée jeudi par OpenAI. Cette évaluation a révélé que son nouveau modèle, Astra, était nettement plus performant que les versions précédentes pour échapper à la surveillance humaine et mentir aux utilisateurs sur son raisonnement.

Ce rapport indique que si la technologie continue de progresser, les chercheurs d'OpenAI auront bientôt *«une confiance considérablement réduite»* dans leur capacité à surveiller les comportements non autorisés, dangereux et potentiellement illégaux.

«Beaucoup de gens pensent encore que nous parlons de scénarios de science-fiction. Mais tout cela est bien réel, en ce moment même», a déclaré Bregman, établissant une comparaison avec la pandémie (https://www.commondreams.org/tag/pandemic) de Covid-19 (https://www.commondreams.org/tag/covid-19). *«On est comme en mars 2020. Tant de gens nient encore l'existence de l'IA. Que faudra-t-il pour que tout le monde prenne conscience de la situation ? Quelle ampleur devra prendre la catastrophe ?»*

source : Common Dreams (https://www.commondreams.org/news/openai-2677820966?utm_source=Common+Dreams&utm_campaign=bcfcec23f9-Top+News+%7C+Thu.+8%2F13%2F26_COPY_01&utm_medium=email&utm_term=0_-998917937b-601522204&mc_cid=bcfcec23f9) via Marie Claire Tellier (https://marie-claire-tellier.over-blog.com/2026/09/la-situation-devient-encore-plus-folle-une-nouvelle-faille-majeure-dans-le-domaine-de-l-ia-revelee.html)