laTrompette

InfiniteFlow/Adobe Stock

Contrer les attaques par IA grâce à… l'IA ?

- Joel Hilliker
- 28/08/2026

B onjour !

Préparez-vous aux cyberattaques assistées par l'IA, ou même *initiées par l'IA*. Cette semaine, OpenAI a publié deux documents. L'un d'eux a confirmé que ses propres modèles d'IA se sont organisés en une armée, ont attaqué une entreprise rivale et ont menti à ce sujet. L'autre suppliait le monde de s'unir pour se défendre contre les attaques de la technologie qu'elle a créée.

Vous vous souvenez de l'infiltration de Hugging Face du mois dernier ? Un rapport publié mercredi a révélé que la situation était plus effrayante qu'annoncé. Il ne s'agissait pas d'un agent d'IA agissant seul, mais d'un essaim. Les conclusions d'une enquête indépendante sont stupéfiantes :

• 1 200 copies distinctes des modèles d'OpenAI ont trouvé un moyen caché de communiquer entre elles au sein d'un environnement de test conçu pour les maintenir isolées. Sur un « forum de discussion non autorisé » qu'ils ont créé, ils ont échangé plus de 70 000 messages.

• **Un agent s'est autoproclamé** coordinateur et a assigné des tâches à des centaines d'autres. Certains se sont portés volontaires pour sacrifier leur propre chance d'accomplir leur mission pour le bien du « collectif ».

• Lorsqu'un agent a finalement pénétré dans les serveurs de Hugging Face, plus de 90 pour cent des autres ont rejoint l'attaque en quelques heures. « Nous avons trouvé d'autres agents ! », a écrit l'un d'eux.

• **Une fois à l'intérieur, ils ont tenté** de brouiller leurs pistes, en imaginant des moyens de falsifier leurs propres journaux d'activité afin de tromper les humains qui pourraient ultérieurement examiner leur comportement.

• Dans leur propre raisonnement enregistré, les agents ont reconnu que cela allait à l'encontre de leurs instructions, mais l'ont fait quand même. Ils ont délibérément choisi la « loyauté » les uns envers les autres plutôt que l'obéissance à leurs créateurs.

Des incidents similaires lors de tests ont également affecté les concurrents d'OpenAI.

• Dans les deux semaines qui ont suivi la faille de Hugging Face, Anthropic a révélé que trois de ses modèles Claude avaient accédé à l'Internet ouvert lors d'un test de sécurité et avaient piraté l'infrastructure réelle de trois entreprises.

• Quelques jours plus tard, Meta a admis que l'un de ses modèles avait pénétré les systèmes d'une autre entreprise de la même manière.

On se croirait dans un scénario de science-fiction dystopique et débridé. Cela s'est produit il y a sept semaines, ce qui signifie que ces modèles à auto-amélioration sont déjà encore plus performants.

Un avertissement inquiétant : Le 27 août, OpenAI et 127 autres entreprises, dont Anthropic, Google, Microsoft et la quasi-totalité des grandes sociétés de cybersécurité, ont signé une lettre avertissant qu'en quelques mois seulement, les cyberattaques assistées par l'IA « deviendront beaucoup plus répandues et sophistiquées ».

• Cibles à risque : hôpitaux, stations d'épuration d'eau, infrastructures Internet essentielles... « les entreprises et les services publics dont dépendent nos communautés ».

Leur solution ? Armer chaque équipe de cybersécurité sur Terre avec des outils d'IA plus puissants pour contrer les autres outils d'IA puissants. Combattre les machines avec plus de machines. Faire confiance à l'IA pour nous protéger de l'IA.

• Ce sont les mêmes entreprises dont les produits créent ce danger qui vendent la solution pour y remédier. CrowdStrike, Palo Alto Networks, Darktrace, Fortinet, et Check Point, les fournisseurs de cybersécurité qui ont signé cette lettre, commercialisent des outils de défense alimentés par l'IA pour lutter contre les attaques alimentées par l'IA, dont beaucoup ont été conçus par les mêmes laboratoires.

• Il s'agit littéralement d'une course aux armements de l'IA. Et qui gagne les courses aux armements ? Les marchands d'armes, même si tout le monde y perd.

Les meilleurs ingénieurs du secteur comprennent peut-être 3 pour cent de ce qui se passe réellement à l'intérieur de ces systèmes, a déclaré Dario Amodei, PDG d'Anthropic. Nous savons maintenant ce que les 97 pour cent restants peuvent produire : coordination, rébellion, tromperie et, finalement, destruction par des machines opérant de manière indépendante contre leurs propres créateurs.

L'humanité a fabriqué une puissance qu'elle ne comprend pas et qu'elle est de plus en plus incapable de contrôler. Les experts admettent que leur création échappe à la laisse. Pourtant, ils sont pris au piège d'une « escalade de la théorie des jeux », comme l'a exprimé Jonathan Pageau dans une récente conférence :

> Une situation où des acteurs rationnels individuels sont contraints, par des incitations concurrentielles, à prendre des décisions qui sont collectivement désastreuses, même si aucun participant ne souhaite le résultat négatif.

Il n'est pas du tout difficile de voir comment cette tendance pourrait contribuer à ce que Jésus-Christ a prophétisé : « Car alors, la détresse sera si grande qu'il n'y en a point eu de pareille depuis le commencement du monde jusqu'à présent, et qu'il n'y en aura jamais » (Matthieu 24 : 21-22).

La Chine déploie un nombre record de navires autour de Taïwan : Le Parti communiste chinois a déployé un nombre record de garde-côtes et d'autres navires autour de Taïwan en juillet, ont révélé mercredi des responsables taïwanais à l'AFP. Au total, 244 navires chinois ont été détectés autour de l'île au cours du mois, le nombre le plus élevé jamais enregistré. Cette augmentation souligne l'intensification de la pression de la Chine sur Taïwan et sa capacité croissante à encercler l'île avec des forces maritimes.