MÉDIAS & TECH

Le dernier modèle d'Anthropic menace toute la cybersécurité

LA SOCIÉTÉ D'IA SONNE LE TOCSIN ET FÉDÈRE TOUS LES GÉANTS TECHNOLOGIQUES POUR SÉCURISER LES LOGICIELS LES PLUS CRITIQUES.

4 min • Ingrid Vergara

Imaginez un modèle d'intelligence artificielle si puissant qu'il peut identifier en quelques minutes des vulnérabilités critiques dans des logiciels essentiels à la marche du monde numérique que personne n'avait détecté, y compris les meilleurs experts et les scanners de vulnérabilité les plus performants à ce jour… C'est ce dont est capable Claude Mythos, le dernier-né de la famille de modèles de la société Anthropic. Un modèle si puissant qu'il a même réussi à sortir de l'espace sécurisé d'expérimentation dans lequel ses « créateurs » chez Anthropic l'avaient confiné.

Comme son meilleur modèle commercialisé à ce jour, Opus 4.6, Claude Mythos est un modèle à usage général, il n'a pas été entraîné spécifiquement sur de la détection de failles de sécurité. « *Nous l'avons entraîné pour être bon en code. Mais en plus d'être bon en code, il est aussi bon en cyber* », explique Dario Amodei, le PDG d'Anthropic.

Tellement bon qu'il a déjà identifié des milliers de vulnérabilités dont certaines très graves, dans à peu près chaque système d'exploitation utilisé et dans tous les navigateurs web. Selon Anthropic, Mythos a ainsi trouvé une vulnérabilité dans OpenBSD, présente depuis 27 ans et que des millions de scans automatisés n'avaient jamais détectée. « *OpenBSD est un des systèmes d'exploitation les plus renforcés en sécurité au monde, utilisé pour faire tourner des pare-feu et des infrastructures critiques* », explique un expert en cybersécurité. L'IA d'Anthropic aurait trouvé un moyen de pousser à l'erreur à distance n'importe quelle machine tournant sur ce système. Le modèle a également trouvé le moyen d'exploiter une

succession de vulnérabilités dans le noyau Linux, utilisé pour opérer de très nombreux serveurs dans le monde.

Anthropic a donc sonné le tocsin auprès des plus grandes entreprises technologiques, car les conséquences pourraient s'avérer gravissimes si l'ensemble de l'architecture numérique mondiale n'est pas mieux protégé contre les assauts de ce type de modèle. Anthropic a ainsi réservé l'accès à une version *preview* de son modèle à Amazon, Microsoft, Apple, Google, Palo Alto Networks, Crowdstrike, JPMorganChase, Broadcom, Cisco, Nvidia, etc. pour que ces grandes sociétés puissent tester la résistance de leurs systèmes et de toutes les applications qu'elles commercialisent. Et travailler ensemble à une meilleure sécurité numérique globale.

Car le risque, explique Anthropic, est que d'autres acteurs parviennent aussi à ce stade de développement de modèles et/ou qu'ils tombent dans de mauvaises mains. Autrement dit, il ne reste peut-être pas si longtemps avant que des attaquants accèdent également à de telles capacités.

Il a donc lancé le « projet Glasswing », qui fédère douze grandes sociétés technologiques et 40 autres organisations. Anthropic y engage 100 millions de dollars en crédits d'utilisation, qui permettront à des entreprises de scanner leurs systèmes pour y dénicher les failles et les corriger. La société d'IA a évidemment briefé la Cybersecurity and Infrastructure Security Agency (Cisa), l'équivalent américain de l'Anssi et le département du Commerce. « *Aucune entreprise de cybersécurité ne peut résoudre seule ces problèmes : l'industrie, les chercheurs, l'open source et les gouvernements ont tous un rôle à jouer,* explique Dario Amodei. *Je suis fier que tant d'entreprises leaders mondiales nous aient rejoints pour le projet Glasswing afin d'affronter de front la menace cybernétique posée par des systèmes d'IA de plus en plus performants* », ajoute-t-il.

« Les potentialités de l'IA ont franchi un seuil qui change fondamentalement le niveau d'urgence requis pour protéger les infrastructures (informatiques) des

attaques », a commenté pour sa part Anthony Grieco, responsable de la sécurité chez Cisco.

Anthropic n'a donc pas prévu pour l'instant de commercialiser Mythos tant que les « défenseurs » cyber n'auront pas pu mieux s'y préparer et que des garde-fous supplémentaires n'auront pas été ajoutés. Mais le but reste bien de déployer Mythos à grande échelle, rappelle Anthropic.

En décembre dernier, OpenAI avait averti que ses prochains modèles d'intelligence artificielle pourraient présenter un risque de cybersécurité *« élevé »*, en raison de la progression rapide de leurs capacités. Dans une note de blog, Sam Altman expliquait qu'il présenterait bientôt un programme visant à fournir à ses clients qualifiés travaillant dans le domaine de la cyberdéfense un accès échelonné à ces outils améliorés.

Reste à savoir dans combien de temps d'autres laboratoires d'IA seront aussi en mesure d'atteindre le niveau de puissance d'un Claude Mythos. I.V.