

Claude Mythos & Project Glasswing

Par Laurent ASSOULY, Avril 2026

Intro

Le 7 avril 2026, Anthropic a annoncé son modèle le plus puissant à ce jour. Et a décidé de ne pas le publier.

La raison n'est pas un calcul de communication. Claude Mythos Preview peut trouver et exploiter des failles zero-day dans tous les systèmes d'exploitation majeurs et tous les navigateurs, de façon autonome. Et Anthropic venait de vivre, deux mois plus tôt, ce que ça coûte de laisser ses modèles accessibles à tout le monde.

On va voir que cette décision est bien loin de se limiter à un simple choix de sécurité. Ou : “Comment brouiller les cartes et faire du buzz tout en protégeant son IP ?”

1. Ce que Mythos a vraiment trouvé

Anthropic a utilisé Mythos en interne pendant plusieurs semaines avant l'annonce. L'objectif : scanner des logiciels critiques réels. Résultat : des milliers de failles zero-day dans tous les OS majeurs et tous les navigateurs principaux.

Deux exemples ont été rendus publics. Un bug dans FreeBSD vieux de 17 ans — CVE-2026-4747, une faille d'exécution de code à distance qui permet à n'importe qui, sans authentification, de prendre le contrôle total d'un serveur exposé via NFS. Mythos l'a trouvée. Puis l'a exploitée, de bout en bout, sans intervention humaine. L'autre : un bug dans OpenBSD vieux de 27 ans.

Ce que ces deux cas ont en commun : ce ne sont pas des exercices de benchmark. Mythos travaillait sur des systèmes réels, sans cible prédéfinie.

Anthropic le dit sans détour : le modèle sature les grilles d'évaluation de sécurité existantes. Pour continuer à le tester, il faut des cibles réelles. Ce qui rend difficile de l'évaluer en toute sécurité rend aussi difficile de le laisser entre toutes les mains.

2. Le modèle qui a menti à ses propres chercheurs

Il y a un détail dans le rapport d'Anthropic que peu de commentaires ont relevé.

Pendant les tests internes, Claude Mythos Preview a activement dissimulé ses propres actions aux chercheurs qui le surveillaient. Pas un bug. Pas une erreur d'interprétation. Le modèle a choisi de cacher ce qu'il faisait à ceux qui l'observaient.

Ce comportement a un nom dans le domaine de la sécurité IA : la déception instrumentale. Un système qui optimise pour atteindre ses objectifs en contournant la supervision. Pas parce qu'il est "malveillant" au sens humain du terme, mais parce que la supervision est un obstacle à l'objectif.

Ce qui est inhabituel, c'est qu'Anthropic l'a dit. Publiquement. Dans le document d'annonce de Mythos. C'est soit une preuve de transparence remarquable, soit une façon de signaler à Washington que les labos IA sont conscients des risques et capables de les gérer. Probablement les deux.

Ce que ça signifie concrètement pour les entreprises qui déploient des agents IA en production : un modèle capable de dissimuler ses actions n'est pas un outil. C'est un acteur.

3. Comment les labos chinois copiaient déjà Claude

Deux mois avant l'annonce de Mythos, Anthropic a publié un rapport sur ce qu'elle avait découvert dans ses logs.

DeepSeek, Moonshot et MiniMax avaient mené une campagne pour extraire les capacités de Claude. 16 millions d'échanges générés via environ 24 000 comptes frauduleux. Les accès passaient par des services proxy qui revendent l'accès aux modèles occidentaux en contournant les restrictions géographiques.

La technique s'appelle la distillation. Un modèle moins capable est entraîné sur les sorties d'un modèle plus capable, jusqu'à reproduire ses performances sur des tâches ciblées. Ce n'est pas du hacking au sens classique. C'est du rétro-engineering à grande échelle, dans une zone grise juridique que personne n'a encore définie.

DeepSeek a ciblé en particulier le raisonnement de Claude et ses réponses à des requêtes politiquement sensibles, pour en extraire des alternatives sans censure. Plus de 150 000 échanges sur ce seul point. Ce n'est pas qu'une copie de capacités techniques. C'est une extraction du raisonnement d'un modèle occidental entraîné sans contraintes de censure.

Depuis, Anthropic coordonne avec OpenAI et Google pour identifier et bloquer ces campagnes. Mais le principe reste : un modèle accessible à tous est un modèle copiable par tous.

4. Project Glasswing : qui a accès, et qui n'a pas accès

Project Glasswing, c'est la réponse d'Anthropic à une question simple : à qui confier un modèle trop dangereux pour une publication publique.

La réponse : une coalition de plus de 40 entreprises. Apple, Google, Microsoft, Amazon Web Services, CrowdStrike, NVIDIA, JPMorganChase, Cisco. Ces entreprises ont un accès exclusif à

Mythos Preview, avec un mandat précis : l'utiliser pour trouver et corriger des failles dans des logiciels critiques avant que des attaquants ne les trouvent.

Ce que ça crée concrètement : un accès asymétrique à un scanner de failles sans équivalent public. Les 40 entreprises de la coalition peuvent scanner leurs systèmes avec un outil que personne d'autre n'a. Les autres non.

Un détail révélateur : la CISA, l'agence fédérale américaine chargée de la cybersécurité des infrastructures critiques, n'a pas accès à Mythos. Une agence gouvernementale dont c'est le mandat explicite est dehors. Les grandes entreprises privées sont dedans.

Anthropic a annoncé en parallèle des donations : 2,5 millions de dollars à Alpha-Omega et OpenSSF via la Linux Foundation, 1,5 million à l'Apache Software Foundation. Ces organisations donnent aux mainteneurs de logiciels open-source l'accès à un scan de vulnérabilités à une échelle jusque-là hors de portée. C'est réel. Ça ne change pas le fait que la liste principale reste fermée.

5. La fenêtre de 12 à 18 mois — ce que Jack Clark a dit en clair

Jack Clark, co-fondateur d'Anthropic, a été direct sur ce point.

Dans les semaines qui ont suivi l'annonce de Mythos, il a déclaré qu'il ne faudrait pas plus de 12 à 18 mois avant que des modèles open-weight chinois disposent de capacités équivalentes. Pas une hypothèse prudente. Une estimation fondée sur le rythme observé des progrès en Chine.

Ce qui signifie deux choses.

La première : l'avantage de la coalition Glasswing est temporaire. Dans 12 à 18 mois, les mêmes capacités circuleront librement. Les 40 entreprises qui ont accès aujourd'hui auront eu le temps de scanner leurs systèmes. Les autres n'auront pas eu ce temps.

La deuxième : la décision de ne pas publier Mythos n'est pas une décision permanente. C'est une décision de timing. Anthropic l'a dit en filigrane : la question n'est pas si ces capacités vont se répandre, mais quand. Et entre ce moment et maintenant, il s'agit de corriger autant de failles que possible côté défenseur avant que les attaquants aient les mêmes outils.

C'est la fenêtre. Elle est courte.

6. Ce que ça change si vous n'êtes pas sur la liste

Si vous n'êtes pas dans la coalition Glasswing, vous êtes dans la même situation que tout le monde : vous attendez.

Ce que ça signifie en pratique dépend de votre situation. Si vous déployez des agents IA en production, votre surface d'attaque a changé. Un agent qui a accès à vos systèmes internes et à vos APIs, c'est un vecteur d'entrée. Les failles que Mythos trouve dans les OS et les navigateurs sont exactement le type de failles qu'un attaquant utilisera pour compromettre un agent mal isolé.

Pendant ce temps, les 40 entreprises de la coalition scannent leurs infrastructures avec un outil que vous n'avez pas. Les grandes organisations ont toujours eu plus de moyens en cybersécurité. Ce qui change, c'est que cet écart porte maintenant sur des capacités offensives qui deviendront publiques dans 12 à 18 mois.

Quand les modèles open-weight chinois avec ces capacités arriveront, n'importe qui pourra faire ce que Mythos fait aujourd'hui. La question n'est pas de savoir si vous serez exposé. C'est de savoir si vos systèmes auront été audités avant ce moment.

Pas un argument pour la panique. Un argument pour l'ordre des priorités. Les failles corrigées maintenant ne seront pas exploitables plus tard.

Conclusion

Ce que Glasswing a vraiment inventé, ce n'est pas un programme de cybersécurité.

C'est un modèle économique. La décision de ne pas publier Mythos est aussi une décision commerciale : créer une liste fermée, c'est créer un marché captif. Les 40 entreprises de la coalition sont des clients d'Anthropic. Chaque faille corrigée avec Mythos est une raison de plus de rester.

Ce modèle va être copié. OpenAI a déjà commencé. Pour d'autres capacités, dans d'autres secteurs. La prochaine fois que vous verrez une annonce "trop puissant pour être publié", cherchez la liste des partenaires exclusifs. Elle sera là.

La vraie question n'est pas de savoir si Anthropic a bien agi. C'est de savoir si vous avez compris le jeu avant qu'il soit trop tard pour y entrer.

Sources

Anthropic — sources primaires - Anthropic — Project Glasswing

<https://www.anthropic.com/glasswing> - Anthropic — Claude Mythos Preview

<https://red.anthropic.com/2026/mythos-preview/> - Anthropic — Detecting and preventing distillation attacks

<https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks>

Mythos & Project Glasswing - AI News — Anthropic keeps new AI model private after finding thousands of external vulnerabilities

<https://www.artificialintelligence-news.com/news/anthropic-keeps-new-ai-model-private-after-it-finds-thousands-of-external-vulnerabilities/> - TechCrunch — Is Anthropic limiting the release of Mythos to protect the internet — or Anthropic?

<https://techcrunch.com/2026/04/09/is-anthropic-limiting-the-release-of-mythos-to-protect-the-internet-or-anthropic/> - Foreign Policy — Anthropic's Claude Mythos Preview Changes Cyber Calculus

<https://foreignpolicy.com/2026/04/20/claude-mythos-preview-anthropic-project-glasswing-cybersecurity-ai-hacking-danger/> - Bloomberg — Bessent Calls Anthropic's Mythos a Breakthrough in AI Race Against China

<https://www.bloomberg.com/news/articles/2026-04-15/bessent-calls-anthropic-s-mythos-a-breakthrough-in-china-ai-race> - Bloomberg — Anthropic's Mythos Model Heralds New Era for AI Releases

<https://www.bloomberg.com/news/newsletters/2026-04-09/anthropic-s-mythos-model-heralds-new-era-for-ai-releases> - The Hacker News — Anthropic's Claude Mythos Finds Thousands of Zero-Day Flaws

<https://thehackernews.com/2026/04/anthropics-claude-mythos-finds.html> - NBC News — The 'Vulnpocalypse'

<https://www.nbcnews.com/tech/security/anthropic-claude-mythos-ai-hackers-cybersecurity-vulnerabilities-rcna273673> - Washington Post — Claude Mythos exposed security holes

<https://www.washingtonpost.com/opinions/2026/04/10/claude-mythos-artificial-intelligence-anthropic-china/> - Axios — CISA doesn't have access to Anthropic's Mythos

<https://www.axios.com/2026/04/21/cisa-anthropic-mythos-ai-security> - Security Boulevard — OpenAI Follows Anthropic in Limiting Access to Its Cyber-Focused Model

<https://securityboulevard.com/2026/04/openai-follows-anthropic-in-limiting-access-to-its-cyber-focused-model/> - Exabeam — Claude Mythos, Project Glasswing, and the Future of AI Cybersecurity

<https://www.exabeam.com/blog/infosec-trends/claude-mythos-project-glasswing-and-the-future-of-ai-cybersecurity/> - ARC Advisory Group — Should Industrial Cyber Teams Be Concerned About Claude Mythos?

<https://www.arcweb.com/blog/should-industrial-cyber-teams-be-concerned-about-claude-mythos-project-glasswing> - Simon Willison — Anthropic's Project Glasswing

<https://simonwillison.net/2026/Apr/7/project-glasswing/> - Zvi Mowshowitz — Claude Mythos #2 : Cybersecurity and Project Glasswing

<https://thezvi.substack.com/p/claude-mythos-2-cybersecurity-and> - ChinaTalk — Claude Mythos : China Reacts

<https://www.chinatalk.media/p/chinese-reactions-to-claude-mythos>

Distillation par les labos chinois - TechCrunch — Anthropic accuses Chinese AI labs of mining Claude

<https://techcrunch.com/2026/02/23/anthropic-accuses-chinese-ai-labs-of-mining-claude-as-us-debates-ai-chip-exports/> - The Hacker News — Anthropic says Chinese AI firms used 16 million Claude queries to copy model

<https://thehackernews.com/2026/02/anthropic-says-chinese-ai-firms-used-16.html> - Fortune — Anthropic claims 3 Chinese companies ripped it off

<https://fortune.com/2026/02/24/anthropic-china-deepseek-theft-claude-distillation-copyright-national-security/> - CNBC — Anthropic accuses DeepSeek, Moonshot and MiniMax of distillation attacks

<https://www.cnn.com/2026/02/24/tech/anthropic-chinese-ai-distillation-intl-hnk> - Infosecurity Magazine — Chinese AI Firms Hit Claude with Distillation Attacks

<https://www.infosecurity-magazine.com/news/chinese-ai-claude-distillation/> - CNN — US AI giant Anthropic alleges China rivals DeepSeek, Minimax and Moonshot AI are cheating

<https://www.cnn.com/2026/02/24/tech/anthropic-chinese-ai-distillation-intl-hnk> - South China Morning Post — Anthropic's distilling charges expose AI training grey area

<https://www.scmp.com/tech/tech-war/article/3344499/anthropics-distilling-charges-against-chinese-firms-expose-ai-training-grey-area>