Claude Mythos : quand l'intelligence artificielle franchit le seuil de la menace cyber systémique

> *Le 8 avril 2026, la société Anthropic a annoncé Claude Mythos Preview, un modèle d'intelligence artificielle qu'elle juge trop dangereux pour être diffusé publiquement. C'est la première fois depuis 2019 qu'un grand laboratoire d'IA retient volontairement un modèle pour des motifs de sécurité. Cet événement marque une rupture que les décideurs publics et privés ne peuvent pas se permettre d'ignorer.*

1. UN MODÈLE HORS CATÉGORIE : CE QUE MYTHOS SAIT FAIRE

Claude Mythos n'est pas un simple progrès incrémental dans la lignée des modèles précédents. Il représente l'émergence d'une nouvelle catégorie de capacité automatisée en matière de cybersécurité offensive — une capacité jusqu'alors réservée à des acteurs étatiques ou à des groupes criminels très organisés.

Des performances documentées, pas des promesses marketing

Lors des phases de test internes, Mythos a été en mesure d'identifier et d'exploiter de manière autonome des vulnérabilités inconnues, dites zero-day, dans l'ensemble des systèmes d'exploitation majeurs (Windows, Linux, macOS, OpenBSD) et dans tous les navigateurs web. Ces failles n'étaient pas récentes : certaines dataient de dix, vingt, voire vingt-sept ans, restées invisibles aux outils d'analyse conventionnels faute de vecteur d'exploitation connu.

L'AI Security Institute britannique (AISI), accrédité pour évaluer le modèle en amont de sa publication, a mesuré un taux de réussite de 73 % sur des tâches de hacking de niveau expert. Avant avril 2025, **aucun modèle d'IA ne pouvait accomplir ces tâches.**

Sur les évaluations académiques classiques, Mythos affiche également des résultats historiques :

- 93,9 % sur SWE-bench (évaluation de compétences en développement logiciel) et
- 97,6 % sur les épreuves du concours de mathématiques USAMO 2026.

Ces scores confirment qu'il ne s'agit pas d'un outil spécialisé dans l'attaque : c'est un modèle généraliste d'une puissance inédite, dont les capacités offensives en cybersécurité constituent l'aspect le plus préoccupant.

Une diffusion déjà incontrôlée

Anthropic a fait le choix de ne pas rendre Mythos disponible au grand public. Cette décision, rarissime dans le secteur, est en elle-même un signal fort. Pourtant, dès le 21 avril 2026, Bloomberg révélait qu'un groupe d'utilisateurs non autorisés avait réussi à accéder au modèle via un forum privé en ligne. La fenêtre de contrôle de la diffusion est donc déjà entamée.

Repère chronologique

Mars 2026 : fuite involontaire d'un brouillon d'annonce par une mauvaise configuration du système de gestion de contenu d'Anthropic, ~3 000 documents exposés, 8 avril 2026 : annonce officielle de Mythos Preview et lancement du Project Glasswing. 21 avril 2026 : Bloomberg révèle un accès non autorisé au modèle via un forum privé. — 22 avril 2026 : les banques allemandes et la Banque d'Angleterre déclarent avoir intensifié leurs évaluations de risque cyber en lien direct avec Mythos.

2. POURQUOI CETTE RUPTURE CHANGE LE CALCUL STRATÉGIQUE

La démocratisation de la menace sophistiquée

Jusqu'à présent, conduire une attaque informatique exploitant des vulnérabilités inconnues supposait des ressources considérables : des équipes de chercheurs en sécurité, des mois de travail, des infrastructures dédiées. Ces contraintes limitaient de facto ce type d'attaque aux États, aux grandes organisations criminelles et aux groupes APT (Advanced Persistent Threats).

Avec Mythos ou des modèles équivalents qui ne manqueront pas de lui succéder, ce seuil s'effondre. Un acteur de niveau intermédiaire, disposant d'un accès au modèle et d'une cible définie, peut désormais déléguer la phase de recherche et d'exploitation de vulnérabilités à un système automatisé. Le coût marginal d'une attaque sophistiquée tend vers zéro.

La compression du temps comme facteur de déstabilisation

La notion de « fenêtre d'exploitation », le délai entre la découverte d'une vulnérabilité et son exploitation active sont au cœur des stratégies de patching et de réponse aux incidents. Mythos compresse ce délai de façon radicale. Là où il fallait autrefois des semaines ou des mois pour transformer une faille connue en exploit opérationnel, un modèle comme Mythos peut y parvenir en une nuit, de manière autonome, sans intervention humaine experte.

Cette accélération remet en cause les cycles de mise à jour des systèmes d'information, les délais de traitement des alertes de sécurité, et plus généralement toute l'architecture organisationnelle de la réponse cyber dans les entreprises et les administrations.

Les secteurs les plus exposés

Les environnements les plus vulnérables à cette nouvelle donne sont ceux qui cumulent trois caractéristiques: une surface d'attaque étendue, un patrimoine informatique ancien et hétérogène, et une criticité opérationnelle élevée. C'est précisément le profil des infrastructures critiques nationales.

Secteurs à vigilance prioritaire

- Industrie lourde et énergie (systèmes SCADA, automates industriels, convergence OT/IT)
- Secteur bancaire et financier (systèmes de règlement, infrastructures de marché)
- Santé (systèmes hospitaliers, dispositifs médicaux connectés)
- Transports et logistique (systèmes de contrôle ferroviaire, portuaire, aérien)
- Collectivités territoriales et administrations (systèmes « hérités » non patchés, budgets contraints)

3. LA RÉPONSE D'ANTHROPIC : PROJECT GLASSWING

Face aux risques identifiés, Anthropic a choisi une voie intermédiaire entre la rétention totale et la diffusion générale : le Project Glasswing. Ce programme offre à un nombre limité d'organisations telles que les opérateurs d'infrastructures critiques, les agences de cybersécurité ou encore les acteurs de l'open source, un accès contrôlé à Mythos Preview à des fins exclusivement défensives.

L'engagement financier est substantiel : **100 millions de dollars** en crédits d'utilisation du modèle sont mobilisés pour les partenaires du programme, auxquels s'ajoutent 4 millions de dollars de dons directs à des organisations de sécurité open source (Linux Foundation, Apache Software Foundation).

L'objectif déclaré est d'utiliser Mythos lui-même pour identifier et corriger les vulnérabilités dans les logiciels les plus critiques, avant que des acteurs malveillants ne puissent en faire autant. C'est une approche de sécurité offensive mise au service de la défense. C'est un retournement conceptuel qui interroge néanmoins sur la gouvernance à long terme de ces outils.

Question de gouvernance

Project Glasswing pose une question que la communauté stratégique ne peut éluder : qui décide de l'accès à un outil dont les capacités offensives sont comparables à celles d'une unité cyber étatique ? La réponse ne peut rester entre les mains d'un acteur privé, aussi bien intentionné soit-il. L'émergence de Mythos appelle une réponse régulatoire internationale coordonnée, au même titre que les produits ou technologies à usage dual, càd initialement conçus pour un usage civil mais pouvant être utilisés ou détournés à des fins militaires ou malveillantes.

4. RECOMMANDATIONS À L'ATTENTION DES DÉCIDEURS

Les recommandations qui suivent s'adressent à trois niveaux de responsabilité : les organisations publiques et privées opérant des infrastructures critiques, les décideurs institutionnels et régulateurs, et la communauté de la défense et de la sécurité nationale.

Pour les organisations opérant des systèmes critiques

- **Réexaminer d'urgence les plans de patching :** les vulnérabilités anciennes (5 ans et plus) sur les systèmes d'exploitation et les applications doivent être traitées en priorité absolue, indépendamment des cycles habituels. Mythos cible précisément ce type de failles.
- **Revoir les modèles de menace :** les référentiels de risque cyber utilisés dans les plans de continuité et les revues de direction doivent intégrer un nouveau niveau de menace — celui d'une exploitation automatisée de masse à coût marginal quasi nul.
- **Accélérer les démarches Zero Trust :** la micro-segmentation des réseaux, le cloisonnement OT/IT et l'application stricte du moindre privilège ne sont plus des objectifs à horizon pluriannuel. Ils deviennent des conditions de survie opérationnelle.
- **Explorer les accès défensifs disponibles :** des programmes comme Project Glasswing permettent d'utiliser Mythos Preview pour scanner ses propres périmètres. Les organisations éligibles devraient engager cette démarche sans délai.

Pour les décideurs institutionnels et régulateurs

- **Engager un dialogue avec les grands laboratoires d'IA :** l'ANSSI, l'ENISA et leurs homologues doivent nouer des relations formelles avec les développeurs de modèles frontières pour obtenir une visibilité anticipée sur les capacités émergentes.

- **Réviser les cadres de dual use appliqués à l'IA :** les régimes de contrôle des exportations et les cadres de double usage (dual use) doivent être adaptés pour couvrir les modèles d'IA à capacités cyber offensives avérées, sur le modèle des conventions existantes pour les outils de surveillance.
- **Renforcer les exigences de notification :** à l'instar de ce que prévoit NIS2 pour les incidents, des obligations de notification préalable devraient être envisagées pour les développeurs qui identifient des capacités offensives critiques dans leurs modèles.

Pour la communauté défense & sécurité nationale

- **Intégrer l'IA offensive dans les scénarios de planification :** les exercices de cyberdéfense et les plans de réponse aux crises doivent désormais intégrer l'hypothèse d'adversaires disposant de capacités d'exploitation automatisée de niveau expert.
- **Investir dans la détection des attaques assistées par IA :** les signatures comportementales d'une attaque conduite par un modèle comme Mythos différeront de celles d'un humain. Les outils de détection doivent évoluer en conséquence.
- **Soutenir la recherche en sécurité de l'IA :** les travaux sur l'évaluation des capacités offensives des modèles, leur confinement et leur audit indépendant doivent bénéficier d'un financement public significatif en France et en Europe.

CONCLUSION

Claude Mythos n'est pas une anomalie. C'est l'avant-garde d'une génération de modèles dont les capacités en cybersécurité offensive vont continuer de progresser, portées par la course aux puces, aux données et aux architectures d'entraînement. D'autres laboratoires — moins prudents qu'Anthropic — publieront des modèles comparables sans les mêmes garde-fous.

La vraie question n'est pas de savoir si ces capacités existent. Elles existent. La question est de savoir si les institutions, les entreprises et les États sont prêts à adapter leur posture de défense à une réalité dans laquelle l'expertise cyber offensive est devenue un bien accessible, automatisable et scalable.

Le temps de la préparation se compte désormais en semaines, pas en années de cycle budgétaire.

Pour aller plus loin

- Anthropic — System Card Claude Mythos Preview (244 pages) : red.anthropic.com
- Anthropic — Project Glasswing : anthropic.com/glasswing
- AI Security Institute UK — Évaluation indépendante de Mythos : aisi.gov.uk
- Scientific American — What is Mythos and why are experts worried? (avril 2026)
- Agence nationale de la sécurité des systèmes d'information (ANSSI) : ssi.gouv.fr

Note produite par la Commission des Hauts de France de l'AAIE IHEDN