

Chers membres du réseau de l'ancienne FSFA, chers intéressés,

Avec des outils de plus en plus puissants, notamment ceux de la société d'IA Anthropic, des questions urgentes de sécurité et donc de réglementation se posent à l'échelle mondiale, avec des implications géopolitiques.

Alerte concernant Claude Mythos Preview

[Claude Mythos Preview](#), développé en avril 2026 par Anthropic, est capable de détecter des failles de sécurité jusque-là inconnues dans les logiciels. S'il tombait entre les mains de pirates informatiques, cela deviendrait très dangereux. C'est pourquoi Anthropic ne l'a mis à la disposition que d'une quarantaine d'entreprises technologiques, afin que celles-ci puissent rapidement vérifier les principaux systèmes opérationnels, combler les failles de sécurité et se prémunir ainsi contre de futures cyberattaques. Le CEO d'Anthropic, Dario Amodei, a averti que, dans le contexte du développement dynamique de l'IA, Anthropic ne disposait tout au plus que d'une avance de quelques mois, et qu'il fallait en profiter. D'autres laboratoires, notamment chinois, pourraient développer des outils tout aussi performants d'ici un an. En effet, OpenAI, concurrent d'Anthropic, a mis à disposition peu après un outil comparable, [GPT-5.4-Cyber](#), également accessible uniquement à un cercle restreint pour des raisons de sécurité.

Réglementation de l'IA : un sujet (reporté) lors du sommet États-Unis – Chine

Le développement de l'IA bouleverse les paradigmes. Pour la première fois, ces outils offrent aux organisations non gouvernementales, aux pirates informatiques et aux criminels une arme miracle pour mener des cyberattaques dévastatrices. Même les deux grandes puissances, les États-Unis et la Chine, n'en seraient pas à l'abri. Seules des normes et des contrôles de sécurité efficaces à l'échelle mondiale, mis en place avant la mise sur le marché d'outils d'IA très puissants, permettraient d'écarter ce danger. En effet, les questions de réglementation ont été abordées lors du sommet entre les États-Unis et la Chine qui s'est tenu à Pékin à la mi-mai 2026. Le ministre américain des Finances, Scott Bessent, a déclaré le 14 mai sur [CNBC](#) qu'il avait été convenu d'élaborer conjointement un protocole à cet effet. Les États-Unis seraient actuellement en tête et pourraient donc montrer la voie.

Claude Mythos 5 et Claude Fable 5

L'UE a réagi rapidement et a exigé [l'accès à Claude Mythos Preview](#), notamment en raison des implications pour les banques européennes et des conséquences mondiales en cas d'attaques. Des [membres du Congrès américain](#) ont également réclamé un accès élargi pour les entreprises. À la mi-mai, [Anthropic a assoupli ses règles](#) et a présenté, le 9 juin, deux nouveaux modèles : [Claude Mythos 5 et Claude Fable 5](#). Tous deux sont basés sur Claude Mythos Preview. Le premier, dans sa version améliorée, reste réservé à un cercle d'utilisateurs clairement défini. Le second est en libre accès, mais doté de mesures de sécurité. Les requêtes dangereuses ne reçoivent pas de réponse ou sont redirigées vers un modèle d'IA plus ancien.

Verdict de l'administration Trump

À peine Claude Fable 5 était-il disponible qu'Anthropic a dû [justifier auprès des utilisateurs ses mesures de sécurité extrêmement strictes](#). La société spécialisée dans l'IA avait testé l'outil [pendant environ 1'000 heures](#) afin d'empêcher que ces mesures ne puissent être contournées par une technique appelée « jailbreak ». Malgré cela, l'administration Trump a été alertée qu'une entreprise aurait réussi à réaliser un « jailbreak ». À peine trois jours après la publication, [l'administration Trump a décrété](#) que les deux outils devaient être bloqués, pour des raisons de sécurité nationale, tant pour l'étranger que pour les personnes étrangères se trouvant aux États-Unis. Les cofondateurs et les collaborateurs d'Anthropic étaient donc également concernés. Immédiatement, [Anthropic a désactivé les outils](#).

Réactions et autogoal pour les États-Unis

Les réactions ont été vives. Aux États-Unis, des [experts en cybersécurité](#) ont réclamé, dans une [lettre ouverte](#), la mise à disposition de ces outils. Comme [Anthropic](#), ils ont fait valoir que les failles découvertes étaient mineures et qu'on les retrouvait également dans d'autres outils, notamment ceux d'OpenAI. Lors du [sommet du G7](#) à Évian du 15 au 17 juin, l'UE a abordé le sujet en présence des CEO des entreprises technologiques. Un programme destiné aux partenaires de confiance a finalement été discuté. Celui-ci vise à permettre à tous les États démocratiques d'y avoir accès.

Les États-Unis veulent dominer le monde grâce à leur IA, mais en agissant de manière irréfléchie et précipitée, ils ne réduisent pas les risques ; au contraire, ils minent la confiance dans leurs entreprises d'IA à l'échelle mondiale, mettant ainsi en jeu des investissements de milliards. Il est évident que des lignes directrices mondiales sont nécessaires – malgré la course à l'IA entre les États-Unis et la Chine.

Avec nos salutations les meilleures,
Pour le réseau de l'ancienne FSFA : Hanna Muralt Müller

23.6.2026

Si vous ne souhaitez plus recevoir cet e-mail, veuillez me contacter : info@muralt-mueller.ch.