« Trop puissant » pour une diffusion publique : le prochain modèle d'IA d'Anthropic, victime d'une fuite, suscite la peur de ses créateurs

Par Steve Tenré Le Figaro *Tech & Web* 28.03.2026

Le logo de Claude, IA de la société Anthropic.

Selon des documents ayant été accidentellement révélés, ce nouveau modèle d'intelligence artificielle, surnommé «Claude Mythos», constituerait une avancée technologique majeure, mais aussi un danger massif entre de mauvaises mains.

Une *«erreur humaine»* à l'origine d'une fuite qui pourrait tout changer. La société américaine Anthropic, spécialisée dans l'intelligence artificielle, a confirmé ce vendredi 27 mars avoir été victime d'une importante fuite, repérée plus tôt par deux chercheurs en cybersécurité, Alexandre Pauwels de l'Université de Cambridge et Roy Paz, de l'entreprise LayerX Security.

La fuite n'est pas due à une cyberattaque, mais à une erreur de configuration du blog d'Anthropic, que la société utilise pour rédiger diverses annonces. Les fichiers qui y sont créés se voient automatiquement attribuer une adresse URL accessible par tous, sauf si un employé en décide autrement. En l'occurrence, aucun salarié d'Anthropic n'avait verrouillé l'accès à ces documents de brouillon, qui auraient dû être dévoilés bien plus tard après moult modifications. La fuite, révélée jeudi par le média *Fortune*, a depuis été colmatée et Anthropic a coupé l'accès dans la foulée.

Parmi ces documents de brouillon, une annonce : celle de *«Claude Mythos»*, un nouveau modèle d'IA que la société présente comme *« **trop puissant** »* pour une diffusion publique. Une petite phrase qui pourrait autant servir d'argument de communication pour la société, que comme un avertissement adressé aux autres employés de la firme. Il était effectivement peu clair, ce vendredi, si cette annonce était dédiée à la communication interne ou externe d'Anthropic.

« Outrepasser les efforts de la cyberdéfense »

« Mythos est le nouveau nom d'une catégorie inédite de modèles IA, bien plus intelligente que nos modèles Opus, qui étaient jusqu'ici nos plus puissants », débute la firme dans un billet de blog aujourd'hui officiellement inaccessible, et consulté par *Le Figaro*. *«Mythos obtient des scores spectaculairement plus hauts que ceux de Claude Opus 4.6, notre meilleur modèle à ce jour»*, poursuit-elle, indiquant que Mythos excellerait dans le codage, le *« raisonnement académique »* et la cybersécurité.

Au vu de ces caractéristiques, Anthropic affirme vouloir *«agir avec la plus grande précaution»* et vouloir *« comprendre tout le potentiel de ce modèle en termes de cybersécurité »*. *«Ainsi, nous adopterons une approche plus lente et plus graduelle lors de la sortie de Mythos, comparativement à nos anciens modèles: nous commencerons à le déployer auprès d'un petit groupe de clients, qui analyseront par eux-mêmes les capacités en cybersécurité de Mythos avant de nous faire des retours »*, peut-on encore lire.

Toujours dans le même document, Anthropic prévient que les *« récents progrès des modèles d'IA en termes de cybersécurité peuvent être utilisés pour le bien comme pour le mal »*, estimant qu'ils *« sont déjà utilisés pour mener des cyberattaques à grande échelle »*. *« Mythos dispose de capacités cyber bien en avance sur les autres modèles d'IA, ce qui présage d'une nouvelle vague de modèles pouvant largement outrepasser les efforts des acteurs de la cyberdéfense »*.

3000 documents accessibles

Près de 3000 éléments et documents d'Anthropic, valorisée à 350 milliards de dollars, étaient consultables avant le coup de boutoir de l'entreprise. Parmi eux, des fichiers détaillant la tenue prochaine d'un sommet dans un *« manoir du 18e siècle »* au Royaume-Uni, uniquement accessible sur invitation, et destiné à vendre des modèles d'IA d'Anthropic à de grands clients d'entreprises. Un autre portait sur le *« congé parental »* d'un employé.

Dans une déclaration transmise au média *Fortune*, un porte-parole d'Anthropic a reconnu qu'une *« erreur humaine »* avait rendu possible la visibilité des fichiers en brouillon de son blog. Et de confirmer que l'entreprise *« développe un modèle polyvalent avec des avancées significatives »*. Pour l'heure, aucune date de sortie n'a été communiquée. Dans son billet de blog, Anthropic prévenait toutefois que son déploiement serait *«t rès cher pour Anthropic, et très cher pour les clients »* qui l'utiliseraient.

L'IA d'Anthropic est, dans tous les cas, particulièrement prisée en ce moment. Selon plusieurs médias américains, les modèles d'IA d'Anthropic ont notamment été utilisés lors de la préparation de l'offensive américano-israélienne contre l'Iran. D'autres grandes entreprises technologiques américaines, telles que Microsoft, utilisent également le modèle Claude d'Anthropic tout en fournissant l'armée américaine.