«Maître Claude», «rêve nocturne», Tamagotchi : le code source de l'IA d'Anthropic a fuité, révélant les projets secrets du géant américain

Par **Steve Tenré**

Il y a 15 heures

intelligence artificielle Anthropic

Le logo de Claude, IA de la société Anthropic. *JOEL SAGET / AFP*

RÉCIT - Les futures évolutions de Claude Code, l'assistant IA prisé des développeurs du monde entier, ont été publiées par erreur sur le Web.

Anthropic traverse une bien mauvaise passe. L'entreprise américaine, créatrice de l'intelligence artificielle Claude et valorisée à plus de 350 milliards de dollars, a subi ce mardi 31 mars une seconde fuite en moins d'une semaine, suscitant l'ire de nombreux experts et observateurs. *«Une entreprise à qui vous confiez votre code ne sait pas sécuriser le sien»,* a d'ailleurs raillé un développeur sur les réseaux

sociaux.

La fuite de ce jour concerne *«Claude Code»*, un assistant lié à Claude et dédié à la programmation informatique. Pour 17, 100 ou 200 dollars par mois, les utilisateurs de Claude Code peuvent bénéficier d'une aide virtuelle afin de créer des applications ou des sites web. De la bouche d'Anthropic sur son site, *«Claude Code cartographie et explique des bases de code entières en quelques secondes. Il utilise la recherche agentique pour comprendre la structure et les dépendances (de votre) projet (...) La compréhension de votre base de code et de ses dépendances par Claude Code lui permet d'effectuer des modifications puissantes et efficaces sur plusieurs fichiers.»* Autrement dit, Claude Code a accès, au moins temporairement, aux différents projets numériques de ses utilisateurs.

«Maître Claude» et ses clones

Cela étant, ces derniers peuvent être rassurés: la fuite qui a eu lieu ne concerne pas leurs travaux, mais bel et bien Claude Code lui-même - et plus globalement les grands projets de la société Anthropic, qui seront très probablement contrariés. Selon les fichiers en fuite, compilés sur X, Reddit et GitHub par nombre d'observateurs et d'experts, Anthropic préparerait en premier lieu *«Ultraplan»*, un système permettant de faire travailler Claude Code de manière autonome sur des sujets complexes, pendant trente minutes, en tâche de fond. L'IA proposera ensuite un plan détaillé à l'utilisateur, que celui-ci pourra rejeter ou accepter.

Le mode Coordinateur, de son côté, pourrait *«cloner»* plusieurs *«sous-agents»* et les placer sous la direction d'un *«Maître Claude»*, chargé de distribuer les tâches et gérer leur organisation. Anthropic mettrait également en place un système permettant de surveiller le niveau d'agacement des utilisateurs de Claude - il compterait le nombre d'insultes écrites, ainsi que la fréquence à laquelle l'utilisateur donne des ordres et indique à Claude de *«continuer»*.

Le projet Kairos est probablement le plus impressionnant d'entre tous. Il compte transformer Claude en un agent IA persistant, doté d'une mémoire à long terme. Même si l'utilisateur ferme la fenêtre, Claude se souviendra de lui à la prochaine session. Le projet s'accompagne d'ailleurs d'un mode *«rêve nocturne»*, qui permettra à Claude Code, lorsqu'il ne sera pas utilisé, de trier, nettoyer et réfléchir aux informations de la journée écoulée. Les fichiers ayant fuité évoquent aussi un nouveau modèle, du nom de *«Capybara»*. Il s'agit, en réalité, de l'autre

pseudonyme de Mythos, qui a déjà fait parler de lui la semaine précédente.

Enfin, Anthropic préparerait un agent IA... inspiré du petit jouet électronique japonais Tamagotchi, prisé au début des années 2000. Le projet *«Buddy»*, qui aurait dû commencer à être présenté début avril, prévoit de générer un familier dans l'interface de commande de Claude Code, réagissant aux actions du développeur. Un compagnon ludique, dont la fête sera peut-être gâchée, au vu de l'ampleur de la fuite présumée.

Nouvelle erreur

Repérée puis publiée sur X par le développeur et spécialiste en cybersécurité Chaofan Shou, et confirmée par bien d'autres observateurs dont le chercheur en IA Yam Peleg, cette fuite est en réalité due à... une erreur, *«humaine»*, selon un porte-parole de la firme, qui a confirmé les faits mardi. *«Il s'agissait d'un problème sur la publication de la mise à jour causé par une erreur humaine, et non d'une faille de sécurité»*, a déclaré un porte-parole d'Anthropic. *«Aucune donnée sensible de clients ni aucun identifiant n'ont été impliqués ou exposés»*, a précisé l'entreprise de San Francisco.

Selon les observateurs, le code de Claude Code a été publié par erreur sur le registre *«npm»*, une plateforme prisée des développeurs où sont mis à disposition, gratuitement ou en échange d'un paiement, une grande variété d'outils appelés *«paquets»*. npm compte plus d'un million de *«paquets»*, et permet ainsi aux développeurs du monde entier d'utiliser facilement du code écrit par d'autres sans avoir à le réécrire par eux-mêmes.

Problème: Anthropic a malencontreusement publié sur la plateforme npm un fichier qui aurait dû rester privé. En son sein se trouvait une version du code source original de Claude Code, composée de 512.000 lignes du langage de programmation TypeScript, réparties sur 1906 fichiers, mais aussi d'un moteur de requêtes dédié aux modèles IA, d'outils modulaires et d'un système d'interface.

Il est toutefois, aujourd'hui, trop tôt pour connaître les tenants et aboutissants d'une telle fuite. Pour certains observateurs, il ne s'agit en aucun cas de la fuite du siècle: selon eux, personne, après avoir pris connaissance de ces fichiers, ne pourra faire tourner gratuitement et localement Claude Code sur son ordinateur. Pour de nombreux autres en revanche, comme ce développeur d'agents IA, la fuite serait

comparable à une *«catastrophe»*. *«C'est comme verrouiller toutes les portes de votre maison... installer des caméras... engager des gardes armés... puis accidentellement mettre en ligne vos plans d'étage sur Google Maps...»,* écrit-il sur X.

Anthropic semble en tout cas devenir coutumier du fait. Vendredi dernier, le 27 mars, Anthropic avait publiquement concédé une *«erreur humaine»* après que des milliers de documents et éléments de son blog s'étaient retrouvés dans la nature. Parmi ces fichiers, une publication, dont on ne sait pas si elle devait servir à la communication interne ou externe de l'entreprise, concernant un prochain modèle de l'IA Claude. Baptisé Mythos, il avait été qualifié par ses créateurs de *«trop puissant»* pour une diffusion publique, et même de dangereux s'il venait à tomber entre de mauvaises mains. *«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»,* avait pu lire *Le Figaro* dans cette publication dont l'accès a depuis été restreint par Anthropic.

La rédaction vous conseille

→ **«Il ouvre vos applications, envoie vos emails» : une vidéo sur les «superpouvoirs» de l'IA Claude suscite un véritable engouement**

→ **«C'est du travail à la chaîne» : plongée avec les annotateurs de données, ces petites mains qui nourrissent l'IA dans l'ombre**

→ **Pascal Gauthier, PDG de Ledger : «La crypto et l'IA se complètent parfaitement»**

Sur le même thème

Eugénie Bastié : « Comment ChatGPT a achevé Gutenberg »

«Un bon intervenant touche 10 à 12.000 euros bruts par mois» : dans la jungle des formations en intelligence artificielle

---

Pour capter notre attention, les IA ont tendance à nous flatter

---

Une première depuis 2002 : l'emploi dans le secteur informatique en baisse en France, l'IA en cause ?

Julie Girard : « Les agents d'IA autonomes peuvent-ils faire sécession ? »

Pourquoi la France n'est pas si bien partie pour héberger l'un des «méga data center» de l'IA voulu par Bruxelles

---

Pénurie de compétences : l'IA au secours des entreprises pour former la relève

---

Le duel entre Anthropic et le Pentagone sur l'IA militaire divise la Silicon Valley autour d'un débat inédit

---

How the Port Town of Saint-Nazaire Turns to AI to Reinvent Local Tourism

---

Murielle Popa-Fabre : « Sommes-nous en capacité de contrôler l'utilisation de l'intelligence artificielle dans la guerre ? »