

Une IA qui «pourrait tous nous tuer d'ici à la fin de la décennie» : qui est Jacob Coxon, cet ingénieur qui claque la porte d'Anthropic

Par Remi Bayol, Le Figaro (09/09/2026)

Ce 9 septembre, Jacob Coxon, un chercheur en IA chez Anthropic, a annoncé qu'il quittait l'industrie, jugeant sa course à l'innovation inconsciente.

PORTRAIT - Ce Britannique de 27 ans accuse l'industrie de l'IA de courir vers des systèmes potentiellement incontrôlables. Son domaine de compétence est au cœur des débats sur leur sûreté.

Il les accuse de « *jouer avec nos vies* ». Ce mercredi 9 septembre, Jacob Coxon, un chercheur en intelligence artificielle (IA) chez Anthropic, ancien de chez OpenAI, a annoncé qu'il quittait l'industrie, jugeant sa course à l'innovation inconsciente.

« *Aucune des deux entreprises (Open AI et Anthropic, NDLR) n'agit de manière responsable. Elles foncent tout droit vers une superintelligence capable de s'améliorer elle-même et jouent avec nos vies* », a-t-il écrit mardi sur son compte X, adhérant à la thèse décriée de la potentielle émergence de futures intelligences artificielles dites « générales » — cette idée selon laquelle ces technologies pourraient bientôt faire aussi bien, voire mieux, que tout ce dont un cerveau humain est capable.

Appel des 1300 à «ralentir la course»

« *Les gens qui construisent l'IA croient sincèrement qu'elle pourrait tous nous tuer d'ici à la fin de la décennie. Ce n'est pas un coup marketing* », a-t-il ajouté sur X, où il a reçu le soutien de son collègue responsable scientifique de l'alignement de l'IA chez Anthropic, Evan Hubinger. Non sans rappeler les arguments marketing agités par les start-up en pointe du développement de ces technologies, ce geste fort a trouvé une résonance avec plusieurs alertes proférées par des travailleurs de l'industrie ces dernières années.

Quelques semaines auparavant, en juillet, Jacob Coxon avait participé à l'appel de 1300 employés de laboratoires d'IA réclamant aux gouvernements de « ralentir la course » au développement de cette technologie. Quelque 600 salariés d'Anthropic et 400 de leur concurrent OpenAI avaient signé cette lettre ouverte.

Université de Cambridge

Avant de rejoindre Anthropic, ce Britannique de 27 ans a d'abord intégré l'équipe technique d'OpenAI en 2023, où il a travaillé sur le développement de GPT 4.0, la mise à jour du modèle de langage ayant servi au déploiement en novembre 2022 de ChatGPT, le tout premier robot conversationnel grand public. Selon le Wall Street Journal, Jacob Coxon a rejoint Anthropic au début de l'année 2026, séduit par la réputation autour de la sécurité de cette entreprise, qui est valorisée à près de 1 000 milliards de dollars, à l'approche de son entrée en bourse. Il y travaillait à l'entraînement des modèles et à leur alignement. Cette phase, qui précède leur déploiement auprès du public, consiste à exercer le réseau de neurones artificiels à apprendre par essais, en faisant des erreurs, et à le corriger. C'est à cette étape que la machine est « policée » afin d'empêcher qu'elle se comporte à des fins indésirables.

L'ingénieur a auparavant étudié les mathématiques à l'Université de Cambridge, au Royaume-Uni. Son pseudonyme sur X, « @hilbertspaess », constitue d'ailleurs sans doute une référence à l'espace

de Hilbert (*Hilbert space* en anglais). Ce concept désigne une sorte d'espace géométrique, composé d'une multitude de dimensions, qui permet de représenter et de mesurer mathématiquement des objets complexes.

Un écho international

Si son premier article scientifique, publié en 2021, s'intéresse plutôt aux données médicales, il semblerait que Jacob Coxon se soit plutôt penché sur la science de l'interprétabilité des modèles d'IA au fil de ses recherches à San Francisco. En novembre 2025, l'ingénieur et plusieurs de ses collègues d'OpenAI ont prépublié un article scientifique s'intéressant à la reconstitution au niveau des connexions et sous-circuits des calculs effectivement réalisés par un réseau de neurones. Leur objectif, en clair, est de rendre possible la production de modèles dont les opérations internes sont moins enchevêtrées, donc plus traçables et explicables. Un défi pour les scientifiques, y compris ceux travaillant au développement de ces outils.

Chez Anthropic, Jacob Coxon faisait partie de l'équipe de recherche sur l'interprétabilité, qui travaille à comprendre le fonctionnement interne de leurs réseaux de neurones. Il a notamment participé à la relecture d'un article de recherche dédié au modèle de Claude 3.5 Haiku. Bien que la prise de parole du jeune ingénieur ait trouvé un écho international dans le contexte d'inquiétude croissante autour du contrôle des modèles d'IA avancés, il est difficile de savoir quelle était son importance stratégique parmi l'équipe de chercheurs dédiés à ce sujet chez Anthropic, composé de plusieurs dizaines de scientifiques.