La crainte d’Open AI face à l’avènement des IA “surhumaines”

L’intelligence artificielle face aux enjeux éthiques et de contrôle

L’intelligence artificielle face aux enjeux éthiques et de contrôle

L’équipe “Superalignement” d’Open AI aspire à défendre la propriété intellectuelle de l’homme face à l’IA.

Jusqu’où ira l’intelligence artificielle? Si elle permet, désormais, de réaliser une opération chirurgicale à distance, savoir si le père Noël existe vraiment ou rédiger un projet de loi, son utilisation et ses capacités posent question.

Pour OpenAI, leader mondial du raisonnement artificiel et créateur de ChatGPT, il est l’heure de se pencher sur le problème du contrôle d’une IA plus intelligente que les

humains.

Pour y parvenir, l'entreprise à but lucratif "plafonné" a décidé de lancer l'équipe "Superalignment", en juillet 2023. Elle sert à diriger, réguler et gouverner les différents systèmes d'intelligence artificielle, selon le communiqué d'Open AI lors du lancement de l'entité. Un programme qui pèse 10 millions de dollars de budget.

> *"C'est un objectif ambitieux et nous ne sommes pas assurés d'y parvenir. Seul un effort ciblé et concerté pourra résoudre le problème d'une IA trop intelligente. De nombreuses mesures se sont révélées prometteuses lors d'expériences préliminaires".*

L'équipe est dirigée par le cofondateur d'OpenAI, Ilya Sutskever, et suppléée par Jan Leike. Trois de leurs chercheurs étaient, la semaine passée, à la conférence sur les Systèmes de traitement de l'information neuronale (NeurIPS) à La Nouvelle-Orléans (Louisiane, Etats-Unis) pour présenter les derniers travaux d'OpenAI, visant à garantir le "bon comportement" de leur IA.

"Un vrai problème"

"Les progrès sur l'IA explosé cette année et je peux vous assurer qu'ils ne s'arrêteront pas là", a confié Leopold Aschenbrenner, l'un des trois chercheurs, au média spécialisé TechCrunch.

> *"Je pense que l'intelligence artificielle va bientôt atteindre un niveau humain. S'ensuivront des capacités intellectuelles surhumaines. Ce qui représente un vrai problème, peut-être le plus important de notre époque".*

La prise de position du chercheur peut faire froid dans le dos, et pourtant l'IA "superintelligente" est aux portes de

la réalité.

Sam Altman, créateur d'Open AI, utilise même comme comparaison le projet Manhattan (un projet de recherche du gouvernement américain dont l'objectif était de produire une bombe atomique au cours de la Seconde Guerre mondiale) pour évoquer les travaux d'Open AI. Ces derniers devraient permettre de "se protéger contre des risques catastrophiques" (ici, la prise de contrôle de l'IA sur l'humain).

Selon TechCrunch, l'équipe "Superalignment" tente actuellement de créer "des cadres de gouvernance et de contrôle" pour "superviser" les systèmes d'intelligence artificielle. Les chercheurs s'efforcent de faire fonctionner GPT-4 (le dernier modèle d'IA de l'entreprise) avec GPT-2 comme "chef".