LES GOUROUS DE L'IA ÉPISODE 12/12

Eliezer Yudkowsky et Nate Soares, « cavaliers de l'Apocalypse » de l'IA, s'alarment d'un risque d'extinction de l'espèce humaine

« Les gourous de l'IA » (12/12). Les deux penseurs du Machine Intelligence Research Institute ont rédigé ensemble un ouvrage où ils nous préviennent sans ambages : si quelqu'un construit une superintelligence, elle nous tuera tous.

Par Arnaud Leparmentier (San Francisco, correspondant)
Publié hier à 14h00, modifié à 00h09 · Lecture 3 min.

Eliezer Yudkowsky (à gauche), en 2023, et Nate Soares, en 2025 LÉA GIRARDOT / « LE MONDE » D'APRÈS JASON HENRY / THE NEW YORK TIME-REDUX-REA & AUBREY TRINNAMAN POUR « LE MONDE »

Eliezer Yudkowsky est un « doomer » de la première heure, un de ces « cavaliers de l'Apocalypse » qui ont mis en garde sur les dangers existentiels de l'intelligence artificielle (IA) avant même qu'elle n'existe. C'est en 2000, il y a un quart de siècle, alors qu'il était âgé d'une vingtaine d'années à peine, que ce natif de Chicago éduqué dans l'orthodoxie juive, mais devenu athée, crée à Atlanta, puis à Berkeley, en Californie, le Machine Intelligence Research Institute (MIRI), pour travailler sur l'IA.

Lire aussi | « La Silicon Valley a compris avant tout le monde que la guerre du futur serait une guerre logicielle »

Doté d'un quotient intellectuel hors norme, le jeune homme, qui a quitté l'école au collège pour raisons de santé, choisit de passer outre les études universitaires. Il se passionne pour le moment où la machine dépassera le cerveau humain. Mais, bien vite, il estime déceler un problème : le non-alignement des valeurs de l'IA avec celles de l'homme. Et décide de concentrer ses recherches sur le sujet d'une « IA amicale ».

Parmi les idées explorées, Eliezer Yudkowsky popularise, avec le philosophe d'Oxford Nick Bostrom, auteur, en 2014, de *Superintelligence* (publié chez Dunod en 2017), des concepts comme l'« orthogonalité » – notion selon laquelle l'intelligence et la bienveillance sont des traits distincts, et un système d'IA ne deviendrait pas automatiquement plus amical à mesure qu'il gagnerait en intelligence – et la « convergence instrumentale », l'idée qu'un système d'IA puissant et orienté vers un objectif pourrait adopter des stratégies finissant par nuire aux êtres humains.

« Un débat plus vaste »

Pendant des années, il s'intéresse à ces sujets, en étant écouté des spécialistes, mais pas du grand public. L'IA semble tellement lointaine. Sa puissance apparaîtra au grand public seulement en novembre 2022, avec la présentation du robot conversationnel d'OpenAI, ChatGPT. Il en profite alors pour changer d'orientation : il cesse largement de faire de la recherche pour se concentrer sur la communication, estimant que le monde est désormais plus mûr pour écouter les mises en garde sur les dangers potentiels de l'IA. *« Le succès de ChatGPT a suscité un débat bien plus vaste qu'auparavant sur le risque d'extinction* [humaine] *lié à l'IA. Le MIRI a largement réorienté ses efforts vers la*

communication sur ce sujet auprès d'un large public, intervenant notamment au sein de grandes tribunes politiques », explique le laboratoire sur son site Internet.

Il s'agit d'un aveu d'échec en creux : incapable d'imposer une IA sûre, Eliezer Yudkowsky emploie la voie politique et médiatique. Depuis, il professe ses mises en garde, sans cesse plus inquiet, comme dans le titre de l'ouvrage publié en septembre 2025 qu'il a écrit avec son collègue Nate Soares : *If Anyone Builds It, Everyone Dies. Why Superhuman AI Would Kill Us All* (« si quelqu'un construit la superintelligence, elle nous tuera tous », Little, Brown and Company, non traduit).

Eliezer Yudkowsky et Nate Soares en bref

Eliezer Yudkowsky

- Américain, âgé de 46 ans
- Autodidacte
- Fondateur du laboratoire Machine Intelligence Research Institute (MIRI)

Nate Soares

- Américain, âgé de 37 ans
- Président du MIRI
- Ancien employé de Microsoft et de Google

Les deux hommes, sans cesse, multiplient les analogies entre l'homme et la machine. *« L'IA d'aujourd'hui n'est pas soigneusement écrite comme les logiciels traditionnels. Elle se développe tel un organisme, ce qui signifie qu'elle peut manifester des comportements émergents que nul n'avait souhaités »*, nous prévenait Nate Soares, 37 ans, lors d'un entretien, en décembre 2025. Et de mettre en garde contre un développement quasi humain de la machine : *« Ce qui rend les êtres humains dangereux, ce n'est pas qu'ils aient reçu des armes d'autrui mais qu'ils sont partis de rien dans la savane et ont su bâtir une civilisation technologique entière. »*

« Un scénario par défaut »

Car si l'on veut que l'intelligence artificielle soit vraiment intelligente, il faut qu'elle soit autonome, et c'est là qu'apparaît le problème. Nate Soares compare les progrès fulgurants de l'IA à ceux dans les jeux d'échecs, et estime que cette évolution aura un impact immense sur la planète.

« Le monde s'en trouverait transformé de manière radicale. Or, la plupart des scénarios ne débouchent pas sur une amélioration du monde, mais pourraient causer notre perte. Non pas parce que les IA nous haïssent, mais parce que comme les humains, en transformant radicalement le monde, elles provoqueraient l'extinction de nombreuses autres espèces. Ainsi, je ne considère pas cela comme une simple éventualité, mais plutôt comme le scénario par défaut si nous nous précipitons dans le développement de l'IA sans savoir exactement ce que nous faisons — ce qui correspond précisément à notre mode opératoire actuel », estime-t-il.

Lire aussi | Le chercheur Nate Soares, « doomer » de l'IA à Berkeley, prévoit la fin de l'humanité : « C'est de la folie de les laisser essayer »

Dans le podcast du *New York Times*, « Hardfork », Eliezer Yudkowsky balayait en septembre 2025 l'argument de ceux qui disent que l'IA est jusqu'à présent restée sous contrôle : *« C'est comme dire : les montres au radium sont désormais sans danger, pourquoi tant de pessimisme au sujet des armes nucléaires ? La prédiction n'a jamais été que l'IA serait néfaste à chaque étape de son développement technologique. Le fait que les IA actuelles ne se comportent pas de manière visiblement ou manifestement malveillante ne contredit en rien les prédictions théoriques. C'est un peu comme regarder un ballon à l'hélium s'élever dans les airs et se demander : "Cela ne contredit-il pas la théorie de la gravité ?" Non. En l'occurrence, les théories fondamentales en jeu ici ne sont nullement contredites par les IA actuelles. »*

Pour aller plus loin Le site du Machine Intelligence Research Institute (MIRI) explore les inquiétudes soulevées par les deux auteurs et les spécialistes mondiaux de l'IA.

Les gourous de l'IA

12 épisodes

ÉPISODE 1/12

Dario Amodei, le patron d'Anthropic, l'utopiste alarmiste de l'IA

Publié le 25 mars 2026

ÉPISODE 2/12

Sam Altman, le patron d'OpenAI, l'apôtre de la « superintelligence » artificielle

Publié le 26 mars 2026