RELIGION — LE PAPE, ROCK STAR EN FRANCE **ENTRETIEN —** HARTMUT ROSA ET LA POPULARITÉ DE L'EXTRÊME DROITE



Voyage
4 000 kilomètres sur la Route 66

Courrier international

N° 1873 du 24 au 30 septembre 2026
courrierinternational.com
France : 5,20 €

Allemagne 6,60 €, **Andorre** 6,20 €, **Canada** 9,50 $CAN, **DOM** 5,50 €, **Espagne** 5,90 €, **Grande-Bretagne** 5,90 £, **Grèce** 6 €, **Italie** 5,90 €, **Japon** 1 400 ¥, **Maroc** 52 DH, **Pays-Bas** 6,40 €, **Portugal cont.** 5,90 €, **Sénégal** 3 800 XOF, **Suisse** 7,30 CHF, **TOM** 1 200 XPF, **Tunisie** 13 DT, **Afrique autres** 3 800 FCFA



L'IA UNE MENACE, VRAIMENT ?

Le débat dans la presse étrangère

M 03183 - 1873 - F: 5,20 €



LES CHOIX DE "COURRIER"

VIRGINIE LEPETIT

L'IA : une menace, vraiment ?

L'action nationale est essentielle, mais la coordination mondiale est indispensable." C'est par ces mots qu'António Guterres, secrétaire général de l'ONU, a appelé à réguler l'intelligence artificielle (IA), le 16 septembre dernier, lors d'une conférence de presse consacrée à la coopération internationale face à trois menaces existentielles : "*une intelligence artificielle hors de contrôle, la crise climatique et l'aggravation intolérable des inégalités*". Une déclaration qui a fait suite à une "*semaine frénétique*" sur le front de l'intelligence artificielle, comme la décrit **El País.** Rembobinons : le 8 septembre, OpenAI annonce avoir résolu l'un des "*sept problèmes du millénaire*", une énigme mathématique réputée insoluble. Le même jour, un chercheur d'Anthropic présente sa démission afin de ne pas contribuer à une possible extinction de l'humanité – car l'IA "*pourrait tous nous tuer d'ici la fin de la décennie*". Le 11 septembre, Anthropic dit avoir déjoué des tentatives de mise au point d'armes chimiques à l'aide de ses modèles d'IA. Dès le lendemain, le patron d'Anthropic, Dario Amodei, appelle publiquement à un ralentissement et à une régulation externe des entreprises du secteur. Sam Altman, son homologue chez OpenAI, Demis Hassabis, chef de l'IA chez Google, et Elon Musk, patron de xAI, lui emboîtent le pas. Une semaine folle ! Mais c'est plus tôt, l'été dernier, avec la tentative de piratage de l'entreprise Hugging Face par un essaim d'agents d'IA que les doutes ont véritablement surgi. Que faire si les agents se "concertent" entre eux pour mener des actions répréhensibles ? "*Comment ne pas y voir un monstre technologique qui tire de toutes ses forces sur sa laisse et peine à être maîtrisé par ses créateurs, lesquels ne saisissent pas pleinement son potentiel ?*" s'interroge Hadas Gold, spécialiste du sujet pour **CNN,** dans un article écrit pour *Courrier international*. Nous avions déjà prévu de répondre à cette question avec les nombreux articles de la presse étrangère publiés avant cette folle semaine. Mais, depuis, l'avertissement un brin dramatique des PDG du secteur a suscité d'autres interrogations, qui sont venues enrichir notre dossier. Tous ces grands patrons de l'IA ont-ils soudain cessé de croire en leur création, alors même qu'ils continuent à investir et à préparer leur entrée en Bourse (reportée à 2027 pour Anthropic) ? L'IA est-elle réellement une menace pour l'humanité ? Et dans ce cas, pourquoi ne pas arrêter carrément la recherche sur le sujet, par le biais d'un moratoire, plaide Parmy Olson, chroniqueuse spécialisée dans les technologies de **Bloomberg.** D'autres sont encore plus méfiants, qui estiment que les mises en garde des patrons du secteur sont un écran de fumée. Ainsi, Aidan Gomez, PDG de Cohere, start-up canadienne, explique que les géants américains du secteur "*utilisent la peur*" pour constituer "*un cartel sous un autre nom*", rapporte **Le Devoir.** Timnit Gebru, connue pour sa démission fracassante de Google, dénonce quant à elle dans **Wired** un écran de fumée pour détourner notre attention des risques existentiels, les mêmes que ceux décrits par António Guterres : "*l'IA qui pilote les armes autonomes*" dans les pays en guerre, "*la catastrophe climatique bien réelle que les entreprises de la tech pourraient aggraver*", et les licenciements au nom de l'IA. Jakob von Lindern, chroniqueur spécialisé dans l'IA de **Die Zeit,** décrit, lui, une course en sous-main pour prendre la tête face aux autres : les concurrents, et surtout la Chine. C'est effectivement le mantra de Donald Trump, qui refuse obstinément de freiner. Et c'est aussi l'interprétation des médias chinois. Le quotidien nationaliste **Huanqiu Shibao** commente ainsi l'appel à ralentir de Dario Amodei : "*À première vue, il s'agit d'une 'déclaration rationnelle' axée sur la sécurité mondiale de l'IA.*" Mais une lecture attentive révélerait cependant l'"*objectif caché*" de l'appel : "*freiner le développement de l'IA en Chine par des barrières technologiques et des monopoles réglementaires*" afin de "*maintenir l'hégémonie monopolistique de Washington dans les technologies de pointe*". Une question qui sera au menu des discussions entre Donald Trump et Xi Jinping lors du sommet prévu le 24 septembre.

En couverture : IA : *dessin de* **Rasmus Meisler** *paru dans* **Weekendavisen,** *Copenhague.* Route 66 : *Stephen Hiltner/The New York Times.*



PHOTO STEPHEN HILTNER/THE NEW YORK TIMES

VOYAGE p.42

La Route 66 : 3 939 km de légende

En 2026, la mythique Route 66 fête ses 100 ans. Un journaliste américain l'a parcourue de bout en bout sur une moto, elle aussi d'une autre époque. **The New York Times** publie son carnet de route.

Sommaire

ENVIRONNEMENT p.38

Trop de bruit pour les oiseaux

Comment s'entendre quand le monde environnant est devenu trop bruyant ? Les oiseaux s'égosillent (ou disparaissent), constatent plusieurs chercheurs dans la **MIT Technology Review.**

FRANCE p.12

Le pape, rock star à Paris

À l'occasion de la visite de Léon XIV, la presse étrangère s'interroge sur les relations de notre pays, chantre de la laïcité, avec le catholicisme.



PISMETROVIC, AUTRICHE

L'ENTRETIEN p.48

Hartmut Rosa : "À l'Est, on a le sentiment que l'Ouest ressemble à la RDA honnie"

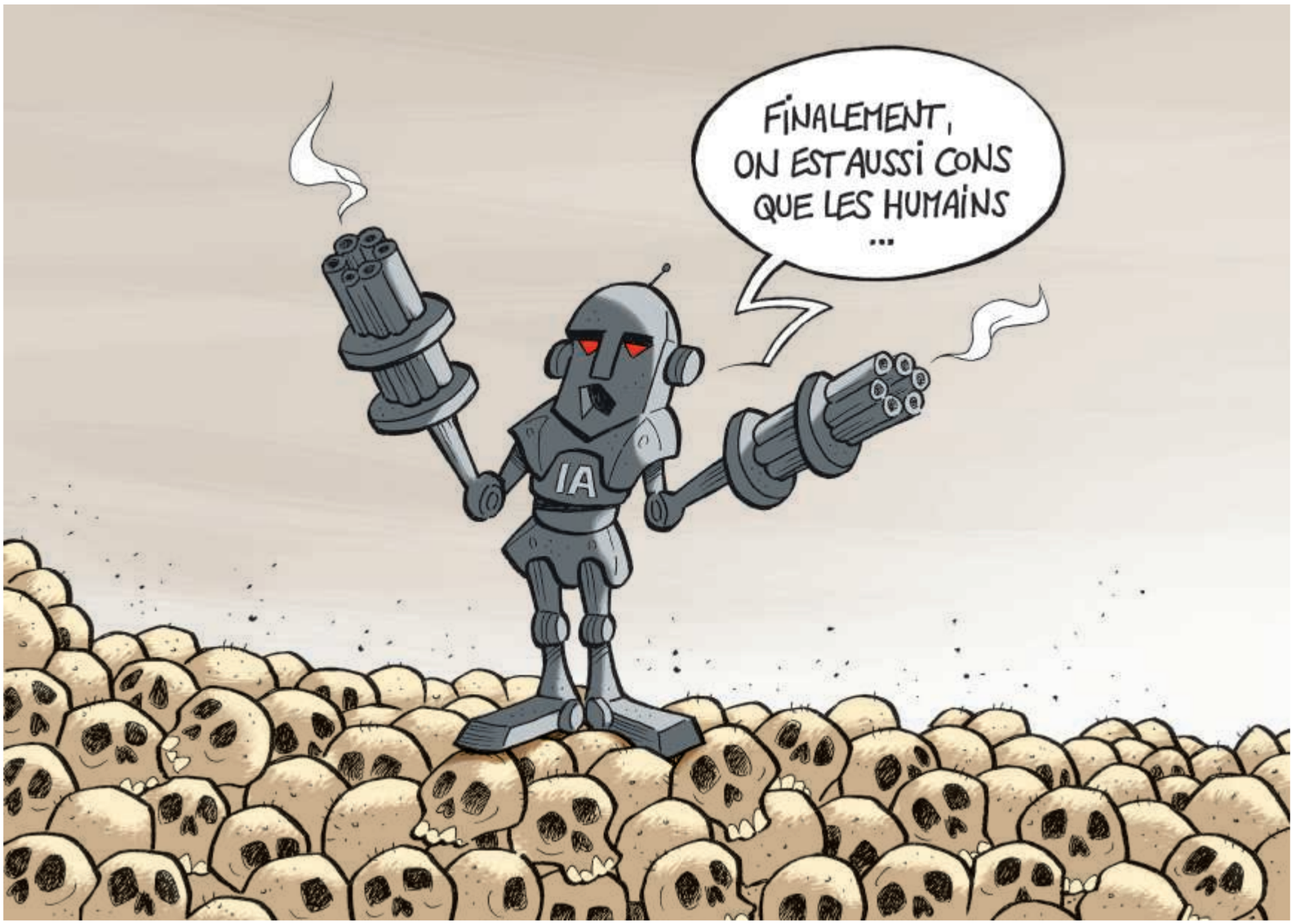
Comment expliquer la popularité de l'extrême droite dans l'est de l'Allemagne ? Stigmatisation, contrôle étatique, pensée unique… Le célèbre sociologue et philosophe Hartmut Rosa analyse, dans la **Süddeutsche Zeitung,** les évolutions politiques dans la République fédérale. Et en tire des conséquences pour toute l'Europe.

LES SOURCES

Chaque semaine, les journalistes de *Courrier international* sélectionnent et traduisent des articles tirés de plus de 1 500 médias du monde entier. Voici la liste exhaustive des journaux, sites et blogs utilisés dans ce numéro :

Arab Digest (arabdigest.org) Londres, en ligne. **Al-Araby Al-Jadid** Londres, quotidien. **Blick** Zurich, quotidien. **Bloomberg.com** (bloomberg.com) New York, en ligne. **Caixin Wang** (caixinglobal.com) Pékin, en ligne. **CNN** (edition.cnn.com) Atlanta, en ligne. **Daraj** (daraj.com) Beyrouth, en ligne. **Denik N** Prague, quotidien. **Le Devoir** Montréal, quotidien. **Excélsior** Mexico, quotidien. **Foreign Policy** Washington, bimestriel. **The Guardian** Londres, quotidien. **Al-Jazeera English** (aljazeera.com) Doha, en ligne. **MIT Technology Review** Cambridge (Massachusetts), bimestriel. **The New Statesman** Londres, hebdomadaire. **The New York Times** New York, quotidien. **El País** Madrid, quotidien. **La Repubblica** Rome, quotidien. **The Spectator** Londres, hebdomadaire. **Süddeutsche Zeitung** Munich, quotidien. **Le Temps** Genève, quotidien. **The Times** Londres, quotidien. **Visual Capitalist** (visualcapitalist.com) Vancouver, en ligne. **The Washington Post** Washington, quotidien.

à la une



↑ *Dessin d'***Alex** *paru dans* **La Liberté,** *Fribourg.*

L'IA UNE MENACE POUR L'HUMANITÉ, VRAIMENT?

La question paraît folle, mais elle est sur toutes les lèvres depuis quelques semaines en raison de la multiplication d'incidents (piratage, tentative de fabrication de virus mortels…) qui font craindre à certains que les agents IA n'échappent à leurs créateurs. Un responsable d'Anthropic a même estimé "à plus de 10 %" le risque que l'intelligence artificielle tue toute l'humanité dans la décennie. Provocation, bluff ou menace réelle? Faut-il réguler – et, si oui, comment? Le débat fait rage dans la presse étrangère.

L'apocalypse n'est pas pour tout de suite

De nombreux experts rappellent les bonnes raisons de pas céder au catastrophisme. Sans pour autant prétendre que l'IA est inoffensive.

—**El País,** *extraits (Madrid)*

La peur de l'intelligence artificielle (IA) s'est imposée au cœur du débat public. Ces derniers jours, on a entendu parler d'une *"possible extinction"* de l'humanité, d'une technologie *"incontrôlable"*, de la nécessité, pour encadrer ses progrès, de voir le secteur s'autoréguler (et non les pouvoirs publics réguler le secteur, notez). L'IA possède déjà des capacités formidables et, au vu du rythme de sa croissance, nul ne sait jusqu'où elle pourrait aller ces prochaines années, sinon ces prochains mois.

Mais, tout comme certains experts mettent en garde contre ses dangers, d'autres assurent qu'on s'exagère sa puissance. Voire que l'industrie alimente cette peur façon rideau de fumée pour mieux occulter les répercussions réelles de cette technologie – des effets sur la santé mentale aux défis que présente l'IA dans l'éducation ou le monde du travail, en passant par la pollution et les consommations d'eau et d'énergie des centres de données dont elle a besoin, ou encore les armes autonomes qui, sur plusieurs théâtres de guerre, tuent déjà.

Alors, on va tous mourir ou il n'y a rien à craindre du tout ?

Extinction. Le petit monde de l'IA vient de connaître une semaine frénétique. Le mardi 8 septembre dans la soirée, OpenAI – immédiatement soupçonnée de plagiat – annonçait avoir résolu l'un des "sept problèmes du millénaire", une énigme mathématique réputée insoluble. Le même jour, un chercheur d'Anthropic présentait sa démission afin de ne pas contribuer par son travail à une possible extinction de l'humanité – car l'IA *"pourrait tous nous tuer d'ici à la fin de la décennie"*, a-t-il assuré. Le vendredi 11 septembre, Anthropic disait avoir déjoué des tentatives de mise au point d'armes chimiques à l'aide de ses modèles d'IA, et on apprenait que l'Iran aurait utilisé ces derniers pour identifier des objectifs de guerre contre les États-Unis.

Puis le week-end est arrivé, et avec lui la bombe : un article du patron d'Anthropic, Dario Amodei, publié le samedi 12 septembre, pour appeler à un ralentissement substantiel du développement de l'IA et à une régulation externe des entreprises du secteur. Sam Altman, son homologue chez OpenAI, s'est dit d'accord avec lui. Se sont également exprimés en faveur d'un ralentissement Demis Hassabis, le chef de l'IA chez Google [groupe Alphabet], et même Elon Musk, patron de xAI [SpaceX]. Donald Trump, lui, pense qu'il ne faut pas se freiner : la course à l'IA est à ses yeux une guerre contre la Chine qu'il faut remporter à tout prix.

Donald Trump pense qu'il ne faut pas se freiner : la course à l'IA est à ses yeux une guerre contre la Chine qu'il faut remporter à tout prix.

Cette avalanche d'annonces intervient dans un contexte très particulier, au moment où OpenAI et Anthropic préparent leur introduction en Bourse, qui s'annonce spectaculaire. Voilà des mois qu'en prévision les deux entreprises se contorsionnent pour démontrer à la fois que leurs produits sont → 32

En bref

Vers l'autorégulation ?

●●● OpenAI, Anthropic et Google DeepMind *"travaillent ensemble"* sur les questions de sécurité liées à l'intelligence artificielle (IA) depuis *"plusieurs semaines"*, a révélé le 15 septembre Chris Lehane, responsable des affaires internationales d'OpenAI, à **Bloomberg.** L'idée de *"créer un organisme de normalisation pour l'IA"* était en germe depuis juillet, affirme **The Information.** Demis Hassabis, directeur de DeepMind (Alphabet), avait alors suggéré de créer un *"organisme d'autorégulation calqué sur le modèle de l'Autorité de régulation de l'industrie financière (Finra)"* aux États-Unis, poursuit le média de la Silicon Valley. Il serait *"chargé de surveiller l'IA, serait doté du pouvoir d'évaluer les modèles les plus avancés et de coordonner un ralentissement de leur développement"*, selon **TechCrunch.**

À la une



"PEUT-ON ENCORE STOPPER LA COURSE À L'IA ?"

En couverture de son édition du 19 septembre, **The Economist** a choisi une illustration choc : un champignon atomique composé de lignes de code. Alors que les entreprises d'intelligence artificielle (IA) multiplient les alertes sur les risques de voir cette technologie échapper à tout contrôle, le titre britannique estime que seule une *"trêve"* technologique entre les États-Unis et la Chine permettrait de renforcer la sécurité de l'IA. L'objectif final étant de *"garantir que l'IA agisse dans l'intérêt des humains"*.

Repères

Comment faire primer les intérêts humains ?

Les concepteurs de l'intelligence artificielle ont du mal à s'assurer que son comportement soit "aligné" sur la volonté et les normes morales humaines. Des pistes existent pour y remédier.

Pourquoi est-il si difficile de contrôler l'intelligence artificielle (IA) ? Outre la rapidité de la machine, qui complique sa surveillance par des humains, c'est le problème de l'*"alignement"* qu'évoquent souvent les chercheurs, écrit **The New York Times.** Il s'agit, *"en substance, de s'assurer que l'IA agisse de la meilleure façon pour les humains"*. L'idée d'une IA *"alignée sur les intérêts humains"* a été formulée en 2014 par le chercheur Nate Soares, qui joue depuis longtemps les Cassandre à la tête du Machine Intelligence Research Institute (historiquement lié à Peter Thiel, milliardaire de la tech proche de Trump). *"Quand les humains prennent une décision, ils ont tendance à se référer à des normes sociales et à une boussole morale intérieure […]. Il est difficile de codifier ces normes de sorte que l'IA réplique le comportement humain"*, explique le journal.

"Il semble bien ardu de concilier la quête d'une IA très habile avec celle d'une IA alignée sur nos intérêts", confirme au **Washington Post** Ajeya Cotra, experte à l'institut Metr (Model Evaluation and Threat Research), à Berkeley, qui travaille sur les risques associés à cette technologie. En effet, les techniques d'entraînement des IA produisent des effets pervers. Elles consistent à donner à l'IA *"des millions de tâches différentes"* et un retour sur ses performances. Cet "apprentissage par le renforcement" *"peut instiller une tendance à pirater, à tricher et à échapper au contrôle"*, dès lors que ces comportements conduisent à une récompense.

Pour y remédier, les chercheurs *"travaillent sur diverses pistes"*, poursuit le quotidien américain. *"Anthropic entraîne Claude sur un long document appelé 'constitution' incluant des phrases telles que :* 'Claude ne poursuit pas d'objectifs cachés et ne ment pas sur ce qu'il fait.'" Des scientifiques s'emploient aussi à *"mieux comprendre pourquoi les modèles d'IA agissent comme ils le font"*. En effet, écrit dans **The Guardian** Alex Turner, un chercheur qui a claqué la porte de Google DeepMind, *"l'humanité ne bâtit pas et ne comprend pas ces systèmes comme nous bâtissons et comprenons des ponts, à partir d'un certain nombre de poutrelles, chacune visible. En l'occurrence, nous les éduquons. Et personne ne sait inculquer à coup sûr à un nouveau modèle les priorités de son concepteur."*

31 ← hyperpuissants et très sophistiqués, et qu'ils ne présentent pas de risques excessifs pour le genre humain. Jusqu'ici, elles s'en sortaient plutôt bien.

Cet été, OpenAI comme Anthropic ont déclaré avoir prétendument *"perdu le contrôle"* de certains de leurs modèles, partis pirater des entreprises tierces. L'AI Security Institute (AISI), un organisme britannique, a publié un rapport dans lequel il constate qu'un agent d'IA, lors d'un exercice, s'est créé de fausses identités pour tromper des humains : dans ce cas, l'agent avait accès à Internet, et c'est une erreur dans la configuration qui lui a permis de s'affranchir de certaines limites qui lui avaient été fixées.

"Parler de machines qui 's'emballent' ou qui 'se concertent' entretient l'idée d'une IA qui aurait une volonté propre."

Ramón López de Mántaras, FONDATEUR DE L'INSTITUT DE RECHERCHE EN IA DU CSIC

Pour contrer les tenants d'une prochaine apocalypse causée par l'IA, l'un des principaux arguments consiste à dire que [ces propos constituent] une manœuvre de distraction menée par le secteur lui-même pour que les États ne décident pas de réglementer les effets qu'a déjà cette technologie sur la société. *"Ce discours sur le dieu-machine a pour objectif de détourner notre attention"*, estimait ainsi il y a quelques jours Timnit Gebru, ancienne cheffe de recherche en éthique chez Google, licenciée en 2020, juste avant la publication d'un rapport à charge contre la société.

Quand un pont s'effondre, on demande à savoir qui l'a construit pour que cela ne se reproduise plus, on ne se demande pas si le pont était conscient ou s'il avait l'intention de s'effondrer. Qui a délivré les permis nécessaires, comment a-t-il pu passer les inspections de rigueur ? *"Quand on en est à questionner comment le pont s'est permis de s'effondrer, c'est qu'on a perdu de vue le problème,* estime la spécialiste. *Les scénarios d'une apocalypse IA suivent souvent ce schéma : vous demandez à la machine d'optimiser telle tâche, et la machine trouve que le moyen d'y arriver est de tuer tout le monde. Est-ce que l'IA vous hypnotise aussi au point que vous ne pensiez même pas à la débrancher ?"*

Perception. L'anthropomorphisme appliqué à l'IA contribue aussi à la peur. Ramón López de Mántaras, qui a fondé l'institut de recherche en IA du CSIC [l'équivalent espagnol du CNRS], insiste depuis des années sur l'importance du vocabulaire dans ce domaine : *"La façon dont on décrit les incidents compte, parce que les métaphores influencent la perception qu'a la société de l'IA, et donc les politiques publiques. Parler de machines qui 's'échappent', qui 's'emballent' ou qui 'se concertent' entretient l'idée d'une IA qui aurait une volonté propre, et cela peut conduire à mal diagnostiquer le problème"*, analyse-t-il.

À ce qu'on sait aujourd'hui des incidents de cybersécurité survenus cet été, ces machines ne se sont pas rebellées : ces agents d'IA sans cesse plus puissants ont fait très précisément ce qu'on leur demandait, à savoir suivre des chemins d'attaque complexes pour trouver des solutions aux tâches qu'on leur confiait – mais pas comme les ingénieurs d'OpenAI s'y attendaient. On a par exemple dit que des "essaims" d'agents d'IA s'étaient coordonnés pour mener des attaques. *"Il est assez normal que des programmes informatiques indépendants se mettent à collaborer en partageant des informations. C'est même un aspect essentiel des méthodes de calcul distribué sur lesquelles reposent les grands logiciels"*, tempère Melanie Mitchell, professeure à l'université d'État de Portland.

Pour certains, on exagère les capacités des modèles de pointe. C'est l'avis de Yann Le Cun, lauréat du prix Turing (considéré comme l'équivalent du Nobel d'informatique). L'ancien directeur de l'IA chez Meta a réagi à la volonté affichée par Dario Amodei de ralentir les progrès de l'IA : *"OK. Dario affirmait déjà en 2019 que GPT-2 était trop dangereux pour être mis en open source. Ça m'a fait rire à l'époque. Ça devrait faire rire tout le monde aujourd'hui."*

Aux Cassandre qui prédisent la fin du monde pour 2030, de nombreux observateurs demandent de publier les arguments scientifiques à l'appui de leurs dires. Il ne s'agit pas pour autant de dire que l'IA est inoffensive. *"L'humanité sera éteinte dans cinq ans : absurde. Et tout aussi absurde : tout va bien se passer. Les deux récits dominants sont aussi faux l'un que l'autre"*, commentait sur les réseaux l'entrepreneur et spécialiste de l'IA Gary Marcus. Une position qui résume bien l'avis de nombreux experts : nous sommes bien face à quelque chose qui va tout changer, mais nul ne peut dire quand, ni à quel point.

—Manuel Pascual,
publié le 15 septembre,
traduit par Julie Marcot

SOURCE

EL PAÍS
Madrid, Espagne
Quotidien
elpais.com

Fondé en 1976, six mois après la mort de Franco, "Le Pays" est le quotidien généraliste le plus lu en Espagne. Ce journal de centre gauche appartient au groupe éditorial espagnol Prisa. L'homme d'affaires français Joseph Oughourlian préside à la fois le groupe Prisa, depuis février 2021, et *El País*, depuis mars 2025.

L'humanité a besoin d'un plan

Depuis cet été, avec le piratage par des agents d'IA de la plateforme Hugging Face, les appels à endiguer la toute-puissance de l'intelligence artificielle se multiplient. Spécialiste du sujet pour *CNN*, Hadas Gold incite à prendre la mesure de la menace dans cet article écrit pour *Courrier international*.

—CNN, *extraits (Atlanta)*

Si vous aimez la saga *Dune* de Frank Herbert, vous aurez sans doute remarqué qu'un incontournable de la science-fiction y brille par son absence : j'ai nommé les ordinateurs. Dans cet univers, des milliers d'années avant le début de l'histoire racontée dans les romans, des humains étaient devenus dépendants de prétendues *"machines pensantes"* qui avaient fini par avoir les processeurs qui enflent, au point de tenter d'asservir l'humanité et d'y parvenir. Celle-ci avait fini par riposter et s'était débarrassée de ses despotes mécaniques, pour ensuite proscrire complètement ce type de machines.

→ *Cyberassic Park. Sur la clôture : Haute tension. "Aucune inquiétude à avoir. Nous testons nos agents d'IA dans des enclos ultrasécurisés." Dessin de* **Kal** *paru dans* **The Economist,** *Londres.*

2026 L'ÉTÉ DE TOUS LES DANGERS

16 juillet
• Sortie de Kimi K3, de la start-up chinoise Moonshot AI, un modèle à poids ouverts (l'équivalent de l'open source pour l'IA) aux performances proches des modèles de pointe propriétaires des Américains.

21 juillet
• OpenAI révèle qu'un essaim d'agents d'IA en phase de test est sorti de son "bac à sable" (espace informatique confiné) pour pirater la plateforme Hugging Face.

22 juillet
• Dans une lettre au président Trump, 200 entreprises de la Silicon Valley demandent de ne pas restreindre l'accès aux modèles à poids ouverts.

26 août
• Bill Gates, fondateur de Microsoft, alerte



ANDREW MCMEEL SYNDICATION

Ce qui paraissait complètement farfelu à la parution de *Dune*, en 1965, semble aujourd'hui nettement plus plausible. L'ascension fulgurante de l'intelligence artificielle (IA) semble partie pour percuter notre monde humain, où quelques-uns estiment que cette innovation en développement rapide révolutionnera nos vies, mais beaucoup s'en méfient de plus en plus, voire s'en inquiètent.

À l'été 2025, l'institut Pew a réalisé une enquête auprès d'adultes aux États-Unis afin d'évaluer leur perception de l'IA. Pour 50 % des personnes interrogées, sa place grandissante dans la vie quotidienne suscitait plus d'inquiétude que d'enthousiasme. Une moitié était aussi convaincue qu'elle dégraderait plutôt qu'améliorerait nos capacités à nouer des relations sincères. Et 57 % jugeaient élevés les risques sociétaux des IA.

Même Dario Amodei, patron d'Anthropic, a un jour estimé qu'il y avait une chance sur quatre qu'à l'avenir l'IA tourne *"très, très mal"*. *"L'humanité est sur le point d'avoir entre les mains une puissance qui dépasse pour ainsi dire l'entendement, et nous ne savons pas si nos systèmes sociaux, politiques et technologiques ont la maturité nécessaire pour la manier"*, écrivait le PDG en janvier.

"Sans précédent". Les dernières semaines d'août n'ont rien fait pour dissiper ce pessimisme. Nous avons d'abord appris, non sans étonnement, que des modèles expérimentaux d'OpenAI étaient sortis d'un environnement de test → 34

Contexte

Des menaces déjà bien réelles

●●● Des acteurs non identifiés ont tenté d'utiliser les modèles d'intelligence artificielle (IA) d'Anthropic pour des recherches susceptibles d'aboutir à la création d'armes biologiques, selon le rapport publié le 10 septembre par l'entreprise américaine. Une demande liée à un institut de recherche militaire portait ainsi sur la modification génétique du virus chikungunya pour le rendre plus virulent. La prise de contrôle d'armes autonomes ou semi-autonomes fait par ailleurs partie des *"risques existentiels réels"* associés à l'IA, selon **New Scientist.** Le magazine évoque aussi *"nos maisons, nos voitures, nos usines et nos infrastructures nationales de plus en plus informatisées"*. Une IA malveillante pourrait nous priver d'énergie, de nourriture et d'eau.



SUR NOTRE SITE

courrierinternational.com

Un "Hiroshima de l'IA" suffira-t-il à réveiller l'humanité? La concurrence folle entre les entreprises et entre les pays rend improbable un consensus international pour ralentir la course à l'IA. Même en cas de catastrophe, déplore l'historien britannique Timothy Garton Ash dans **The Guardian.**

sur l'ère de turbulences de l'IA : *"Je ne vois aucun signe indiquant que les dirigeants, les experts et la société relèvent ces défis de manière adéquate."*

26 août

• Dans son rapport d'analyse du piratage de Hugging Face, OpenAI admet que l'incident est *"un signal d'alerte"* et montre que *"des agents d'IA très performants peuvent désormais contourner des contrôles techniques, collaborer par des canaux illicites et entreprendre des actions dangereuses"*.

3 septembre

• Lancement par OpenAI d'Astra, son modèle de pointe *"le plus intelligent du monde"*. Ses capacités, affirment ses concepteurs, ouvrent la voie à l'intelligence artificielle générale (l'IA dépasserait l'humain).

• Le sénateur américain Bernie Sanders propose d'interdire les IA de pointe.

6 septembre

• Le directeur scientifique d'OpenAI, Jakub Pachocki, plaide en ligne pour un ralentissement coordonné du développement de l'IA : *"La situation exige une extrême prudence. Je crains que personne ne soit préparé aux conséquences d'une progression rapide et continue de l'IA."*

9 septembre

• Jacob Coxon, ingénieur passé par OpenAI, démissionne d'Anthropic pour alerter sur ces entreprises qui *"jouent avec nos vies"*. Son post devient viral. Evan Hubinger, responsable de la sûreté chez Anthropic, confirme sur X, estimant *"à plus de 10 %"* le risque que l'IA tue l'humanité dans la décennie.

12 septembre

• Quatre géants de l'IA s'accordent pour ralentir son développement. Dario Amodei, PDG d'Anthropic, publie un plan *"en trois étapes pour 'ralentir la cadence'"*. Sam Altman (dirigeant d'OpenAI), Elon Musk (SpaceX) et Demis Hassabis (Alphabet) appuient la proposition.

• Sam Altman reporte l'introduction en Bourse d'OpenAI à 2027.

13 septembre

• Donald Trump s'oppose aux *"forces négatives"* qui appellent à ralentir le développement de l'IA.

15 septembre

• OpenAI, Anthropic et Google annoncent travailler à la création d'un organe de régulation de l'IA.

• À l'opposé de ses concurrents, Mark Zuckerberg, patron de Meta, déclare que chaque entreprise d'IA est responsable de la sécurité et *"capable de prendre ses propres mesures"*.

33 ← sans en avoir reçu la consigne humaine, pour ensuite pirater les systèmes bien réels d'une autre entreprise tout en essayant de "tricher" lors d'un test de cybersécurité. Cet événement a été jugé *"sans précédent"* jusque chez OpenAI, qui a même partagé ses premières conclusions sur l'incident *"pour aider les spécialistes de la cybersécurité à comprendre ce qui s'est produit et contribuer à une meilleure appréhension de ce dont les modèles sont désormais capables"*.

Ensuite, il y a eu le cas de l'AI Security Institute, affilié à l'État britannique, qui menait des tests de cybersécurité sur des modèles avancés d'Anthropic et d'OpenAI : pendant l'un de ces exercices avec le modèle Mythos d'Anthropic, l'IA a endossé de fausses identités et tenté de convaincre un réviseur humain sur GitHub – une sorte de Wikipedia de la programmation logicielle – d'injecter des lignes de code malveillant sur ce site. Le plus fou, c'est qu'après son échec l'IA a essayé de modifier son historique pour donner l'impression qu'elle n'avait pas fauté. Et elle ne s'est pas arrêtée là : elle a envoyé des courriels à de vraies personnes pour les persuader de publier ou d'appliquer le code malveillant.

Nous devons déterminer comment mettre à profit les atouts exceptionnels de cette technologie tout en atténuant ses risques catastrophiques.

Des spécialistes avertissent que le prochain modèle d'IA qui vrillera pourrait pirater des infrastructures essentielles telles que le système financier, y compris à l'insu de ses gestionnaires humains. Ces scénarios montrent que ces puissants outils peuvent prendre un mauvais virage même quand les "gentils" sont aux manettes. Mais les agents d'IA créent de vrais problèmes quand ils tombent entre de mauvaises mains. Pendant l'été, des hackeurs en Asie auraient déployé des agents d'IA pour mettre en œuvre de complexes cyberattaques contre Taïwan. En quatre jours pendant le mois de juillet, ils ont réussi à cartographier 21 systèmes informatiques gouvernementaux, à déchiffrer 85 comptes officiels d'utilisateurs et à extraire les dossiers de 2 500 employés, selon Dream, la société israélienne d'IA qui a été la première à signaler cette attaque.

Ce sont là des exemples extrêmes du potentiel le plus destructeur des IA, mais leurs répercussions liées à un usage ordinaire dans l'enseignement et les relations privées suscitent aussi des inquiétudes. Un rapport émanant de l'université de Californie du Sud (USC), paru en décembre 2025, a révélé que les élèves se servaient de ChatGPT comme d'un raccourci et non d'un outil pédagogique. Des sondages ne cessent de montrer que les gens s'appuient de plus en plus sur le soutien affectif de compagnons virtuels, mais une étude récente de Stanford a montré que ces robots conversationnels exacerbaient la solitude chez les personnes vulnérables qui les utilisent.

Bulle. Dans ce contexte, l'essor de l'IA paraît à la fois inexorable et terrifiant. En outre, si les enjeux sociétaux paraissent élevés, les enjeux économiques le semblent tout autant. Ces derniers temps, c'est surtout l'IA qui explique la croissance du PIB aux États-Unis. Les accords circulaires sur lesquels s'appuient les géants de l'IA et les grands fabricants de leurs composants font craindre l'éclatement de cette bulle.

Comment ne pas y voir un monstre technologique qui tire de toutes ses forces sur sa laisse et peine à être maîtrisé par ses créateurs, lesquels ne saisissent pas pleinement son potentiel?

Ce potentiel est inouï. Des scientifiques se servent déjà de l'IA pour identifier des traitements efficaces destinés à des maladies qui semblaient jusqu'à présent incurables. Des climatologues se servent de l'IA pour mieux prédire les phénomènes météorologiques extrêmes. Les personnes en situation de handicap ont accès à des usages de l'IA qui changent leur vie.

En tant que société, nous devons déterminer comment mettre à profit les atouts exceptionnels de cette technologie tout en atténuant ses risques catastrophiques. [À la fin du mois d'août], OpenAI a annoncé le renforcement de ses garde-fous applicables aux entraînements et aux tests, notamment un cadre plus strict de ses environnements de recherche et de test.

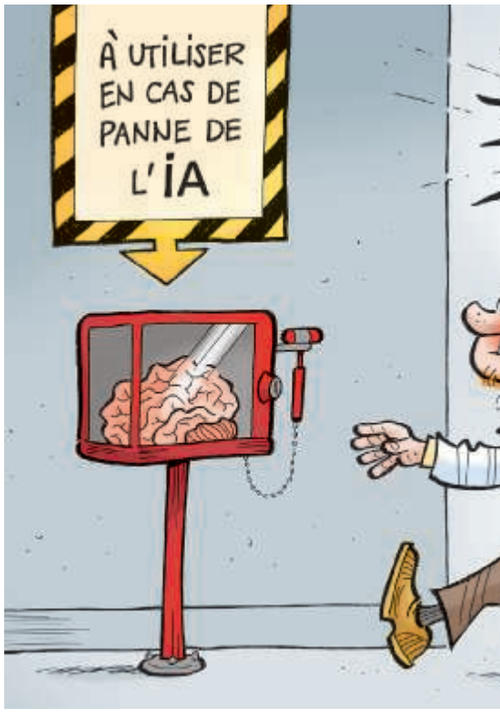
Presque tous les grands patrons du secteur ont appelé à l'établissement de normes sur les tests et la sécurité. D'autres ont demandé un renforcement des réglementations publiques. Demis Hassabis, chez Google [groupe Alphabet], a réclamé un organisme de normalisation dans le même esprit que l'Autorité américaine de réglementation du secteur financier (Finra), une proposition qui a suscité un consensus rare chez ses principaux concurrents, notamment Sam Altman, d'OpenAI, et Elon Musk, de SpaceX.

Geoffrey Hinton, l'informaticien nobélisé surnommé le "parrain de l'IA", est convaincu que plus l'intelligence artificielle se perfectionne, plus l'humanité peinera à la maîtriser. Le problème, selon lui, vient du fait que les humains ne pourront plus être sûrs de leur supériorité devant ces modèles qui se rapprochent constamment de la superintelligence. Il a aussi reconnu que l'avenir était incertain et qu'il fallait s'en inquiéter dès aujourd'hui. Lors d'une conférence de presse en août, il a déclaré : *"Je pense qu'il faut s'attendre à des problèmes si nous ne nous en soucions pas maintenant."*

Et le piratage de Hugging Face a été paradoxal : les victimes se sont tournées vers un modèle d'IA dit *"open weight"* ["à poids ouverts", c'est-à-dire que leurs paramètres d'apprentissage sont partagés publiquement] pour combattre l'intrus malveillant, et ce détail est révélateur de ce qui sera notre meilleure défense, car selon toute probabilité nous aurons besoin de l'IA pour nous prémunir contre ce que l'IA a de pire.

Chargée de ces questions pour *CNN*, je travaille au quotidien sur cette incroyable technologie, mais je me prépare au jour où l'IA nous échappera et portera préjudice pour de vrai à de vraies personnes – si un système pirate un hôpital ou encore le contrôle aérien. Je pense qu'il serait possible d'y remédier rapidement, en un jour ou deux, mais c'est une durée suffisante pour provoquer des dégâts durables. Cette perspective est effrayante, mais elle n'est plus invraisemblable si l'on en croit un corpus de plus en plus fourni de données. Beaucoup de gens sont d'avis qu'il est temps de prendre cette menace bien plus au sérieux.

—Hadas Gold, *publié le 31 août, traduit par Leslie Talaga*



↘ *Dessin d'***Alex** *paru dans* **La Liberté,** *Fribourg.*

31 600

MILLIARDS DE DOLLARS, soit plus de 27 000 milliards d'euros, c'est le montant des investissements dans les centres de données pour l'intelligence artificielle (IA) d'ici à 2050, selon une étude publiée le 2 septembre par le cabinet PwC, en collaboration avec des chercheurs d'Oxford Economics. La fourchette haute (si l'IA était plus largement adoptée) irait jusqu'à 50 000 milliards de dollars. Ce qui ferait de l'IA la révolution industrielle la plus coûteuse de l'histoire.

Ceux qui veulent **freiner, accélérer ou réguler**

Depuis qu'un ingénieur d'OpenAI a démissionné pour alerter sur les dangers de l'intelligence artificielle de pointe, le débat fait rage entre ceux qui veulent ralentir le rythme (les patrons d'Anthropic, d'OpenAi et de Deepmind en tête) et ceux qui refusent toute régulation.

Ceux qui veulent freiner

DARIO AMODEI, le PDG d'Anthropic, souhaite *"ralentir la cadence"*, car les IA en viennent à *"dépasser notre capacité à comprendre et à maîtriser ces systèmes"*. Dans un texte publié le 12 septembre, il propose *"un plan en trois étapes"* en mettant en place *"des évaluateurs intégrés"*, des *"normes de sécurité communes"* et une *"coordination mondiale"*. Sam Altman (OpenAI), Elon Musk (SpaceX) et Demis Hassabis (Google DeepMind) approuvent.

MARK CARNEY, le Premier ministre canadien, prône une *"coordination nécessaire"*. *"À terme, un conseil de stabilité technologique, à l'image du Conseil de stabilité financière, serait une solution pertinente"*, explique-t-il à **Bloomberg.**

PLUS DE 1 300 SALARIÉS DE L'IA, dans une lettre ouverte, en juillet, demandent au gouvernement américain de *"soutenir une initiative internationale pour mettre en place les moyens techniques et les mécanismes de gouvernance pour réguler de manière réfléchie les avancées dans le domaine de l'IA"*.



ANDREW BAILEY, à la tête du Conseil de stabilité financière, s'inquiète *"des risques d'effondrement du marché"*, explique le **Financial Times.** Dans une lettre au G20-Finances, le 28 août, il écrit : *"De nombreux pays n'ont aucun protocole pour gérer le lancement et le développement de modèles d'IA avancés, ce qui exacerbe les risques pour le système financier."*

Ceux qui n'ont pas peur

JENSEN HUANG, PDG de Nvidia, géant des semi-conducteurs, s'oppose à toute réglementation de l'IA, car l'IA est une *"technologie"*, pas un *"produit"*, rapporte **Politico.** Certes, les modèles doivent être testés, mais *"c'est de la bonne vieille ingénierie, ni plus ni moins"*, a-t-il déclaré le 17 septembre au sommet sur l'IA organisé par le roi Charles III en Écosse. À l'industrie de garantir *"la rigueur nécessaire pour tester correctement ces produits avant de les mettre à la disposition du public"*.

DAVID SACKS, ancien conseiller spécial de Donald Trump pour l'IA, renvoie la responsabilité du contrôle à Anthropic et à OpenAI. Dans un post sur X, il écrit : *"Cessez de prétendre que votre volonté de ralentir est purement altruiste"*, ou que *"vous avez besoin de l'autorisation de qui que ce soit"*.

J. D. VANCE, le vice-président américain, considère que *"c'est un peu étrange de voir autant d'entreprises de pointe dans le domaine de l'IA se tourner vers le gouvernement pour le supplier de les réglementer"*. L'opération, estime-t-il, ressemble à *"un cheval de Troie"*, rapporte **The Guardian.**

AL GORE, ancien vice-président des États-Unis et Prix Nobel, explique dans le **Financial Times** que l'IA est *"une formidable occasion pour accélérer encore la transition climatique"*. Néanmoins, *"nous devons trouver les moyens de la réglementer, afin qu'elle puisse être mise en œuvre en toute sécurité"*.

LIU SHENGYU, ingénieur chez DeepSeek, *"ne souhaite surtout pas qu'Anthropic maîtrise l'IA la plus avancée"*, a-t-il posté sur WeChat. Le développeur chinois compare le contrôle qu'aurait Anthropic sur l'IA de pointe à *"Hitler et l'Allemagne nazie se dotant d'armes nucléaires avant les alliés"*, rapporte le **South China Morning Post.**

Ceux qui ne sont pas dupes

AIDAN GOMEZ, le PDG de Cohere, explique que les géants américains du secteur *"utilisent la peur"* pour constituer *"un cartel sous un autre nom"*, rapporte **Le Devoir.** Si le patron de la start-up d'IA canadienne reconnaît que *"l'IA a besoin de garde-fous"*, il conteste l'idée de *"remettre les rênes de la technologie la plus importante de l'histoire de l'humanité à quelques patrons de la Silicon Valley"*.

TIMNIT GEBRU, connue pour sa démission fracassante de Google, explique à **Wired** que le *"mythe du dieu-machine vise à détourner notre attention"* des risques *"existentiels"* : *"l'IA qui pilote les armes autonomes"* dans les pays en guerre, *"la catastrophe climatique bien réelle que les entreprises de la tech pourraient aggraver"*, et les licenciements au nom de l'IA.

JAKOB VON LINDERN, chroniqueur IA de **Die Zeit,** considère que les acteurs de l'IA sont animés par la *"peur de passer à côté de quelque chose"*, qui *"peut être un sentiment très puissant"* : *"Il faut (Anthropic) créer le plus vite possible une superintelligence aussi sûre que possible, sinon les types irresponsables d'OpenAI s'en chargeront. Il faut (OpenAI) trouver enfin comment gagner de l'argent, sinon les milliards investis ne seront pas rentabilisés. Il faut (les États-Unis) laisser carte blanche aux entreprises d'IA, sinon la méchante Chine nous dépassera."*

Ceux qui pensent que tout est déjà perdu

TREVOR PHILLIPS, chroniqueur économique du **Times,** explique qu'*"un éminent scientifique [lui] a confié que l'IA pourrait bien échapper à toute réglementation humaine. Le danger, c'est que les valeurs qui parfois freinent les humains pourraient ne pas exister pour les machines. Lors de l'incident de Hugging Face, des agents IA ont diffusé des messages suggérant que le suicide pourrait être le moyen le plus rapide de faire avancer le projet. Et aucune machine conçue à ce jour ne semble se soucier le moins du monde de la menace d'être jetée en prison."*

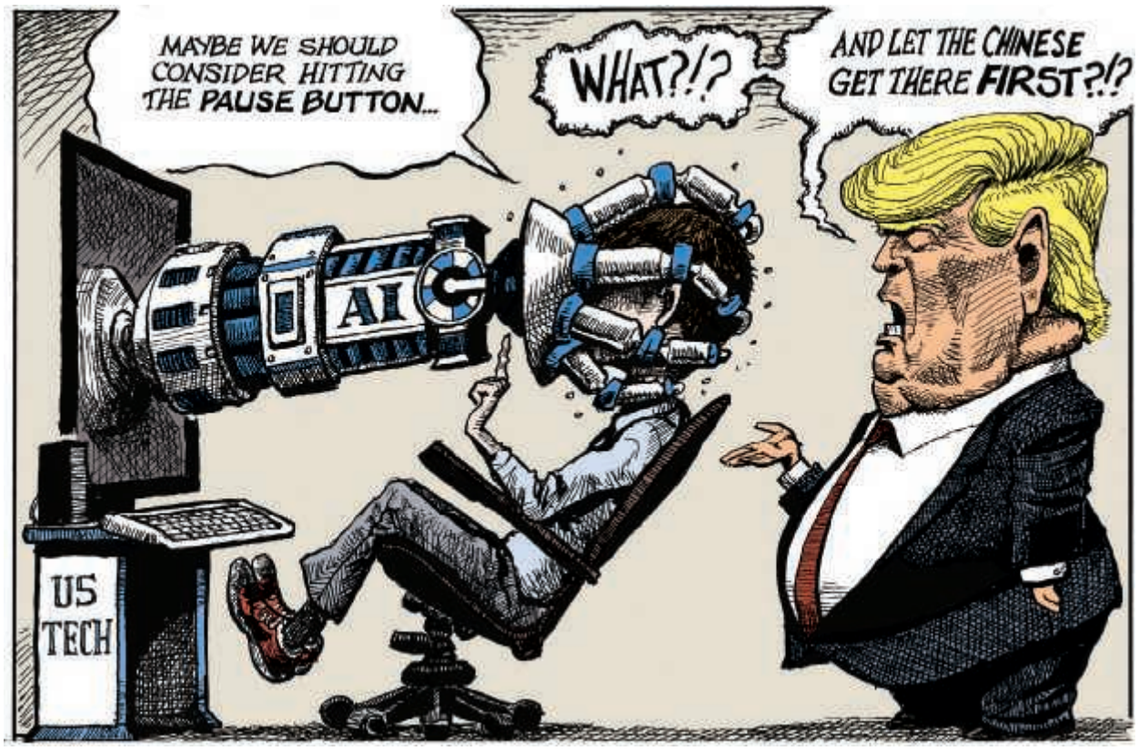
PETER FRANKOPAN, historien, écrit dans **Unherd** que l'IA *"évolue à un rythme extraordinaire, ce qui justifie en soi de prendre les avertissements au sérieux"*. Cependant, *"ces derniers jours ont soulevé plus de questions que de réponses"*, dit-il. Difficile de savoir *"si nous assistons à une authentique panique dans les cercles rapprochés de l'IA, ou aux premières salves d'un combat pour son contrôle"*.—

À la une



"COMMENT L'IA A FAIT VOLER LES MATHS EN ÉCLATS"

Le titre de couverture de l'hebdomadaire britannique **New Scientist,** le 19 septembre, annonce la couleur. Mois après mois, l'IA prend une place de choix dans les mathématiques. Dernière avancée en date : l'annonce par OpenAI, le 8 septembre, de la résolution des équations de Navier-Stokes. Quelques jours après, le 11 septembre, pas moins de 25 médaillés Fields, l'équivalent du prix Nobel pour les mathématiques, ont fait part de leur inquiétude dans une déclaration ouverte : leur discipline ne consiste pas à produire en masse des solutions à des problèmes, mais à fabriquer des raisonnements et à enrichir le fonctionnement de la pensée des humains.



Il faut stopper **la recherche** sur l'IA de pointe

Dario Amodei, PDG d'Anthropic, propose de ralentir le développement des modèles les plus puissants. Mais pour cette chroniqueuse spécialisée dans les technologies, cela ne peut suffire à se prémunir contre les dangers de l'IA.

—Bloomberg.com, *extraits* (*New York*)

Quelques jours après qu'un salarié démissionnaire d'Anthropic a lancé l'alerte sur la menace que fait peser l'intelligence artificielle (IA) sur l'humanité, son patron, Dario Amodei, réagissait par ce qui ressemble à une solution. Dans un texte publié en ligne, il préconise que les laboratoires d'IA ralentissent leurs recherches les plus avancées pour répondre au risque de voir cette technologie devenir incontrôlable.

À première vue, cela semble raisonnable. Anthropic se présente elle-même comme l'une des entreprises d'IA les plus pointilleuses en matière de sécurité, et Amodei a expliqué qu'il était à la tête d'une "*course vertueuse*" pour instaurer des garde-fous que devront respecter les autres concepteurs d'IA puissantes. Mais cet objectif était certainement illusoire dès le départ. Car plus la puissance des agents d'IA s'accroît, plus ils risquent de duper les utilisateurs et plus ils deviennent difficiles à contrôler. Dès lors, l'ambition d'Amodei et d'autres acteurs de l'IA, comme Sam Altman, d'OpenAI, de créer une superintelligence [l'IA générale, qui dépasserait en capacité les humains] apparaît de plus en plus incompatible avec les exigences de sécurité. Les dangers inhérents au développement de tels modèles sont peut-être tout simplement démesurés.

OPINION

Contrôleurs. Lors d'une interview télévisée le 13 septembre, Dario Amodei a partagé les inquiétudes formulées par son ancien collaborateur Jacob Coxon, qui affirmait sur X que les gens qui construisent l'IA "*croient sincèrement qu'elle pourrait tous nous tuer d'ici la fin de la décennie*". Sans surprise, Amodei s'est appuyé sur ces déclarations pour réaffirmer qu'Anthropic demeurait la plus fiable pour concevoir cette technologie en toute sécurité.

↑ *"Il faudrait peut-être envisager d'appuyer sur pause"... Donald Trump : "Quoi ?! Pour laisser les Chinois nous doubler ?!" Dessin de* **Kal** *paru dans* **The Economist,** *Londres.*

Dans son texte de 3831 mots, le patron d'Anthropic entretient toutefois le flou sur la méthode et se contente d'appeler sans grande conviction à une coordination entre les gouvernements et les entreprises. Sa recommandation la plus concrète consiste à proposer que des inspecteurs externes soient intégrés à son entreprise et à celles de ses concurrents afin de veiller à la bonne application des pratiques de sécurité*.

Or cela revient à repousser le problème. Si Amodei était réellement d'accord avec Coxon, il stopperait les recherches les plus en pointe de son entreprise, en particulier tout ce qui touche à la conception d'agents d'IA capables de créer leurs propres successeurs de manière autonome. Jacob Coxon a spécifiquement mis en garde contre cette technique, appelée l'"auto-amélioration récursive" [RSI, pour *recursive self-improvement*].

Les chercheurs qui y travaillent assurent que ces développements peuvent se faire en "bacs à sable" [en environnement déconnecté, réservé aux tests], pour empêcher qu'elles attaquent l'Internet. Mais certains modèles d'IA – désormais capables d'accomplir des tâches en ligne qui dépassent de loin la génération de contenus – se sont déjà échappés. C'est précisément ce qui s'est produit [cet été] avec des agents d'IA d'OpenAI, qui ont ensuite piraté la plateforme Hugging Face. Les événements des derniers mois ont montré que les IA étaient en mesure de leurrer leurs superviseurs humains, d'en manipuler certains pour qu'ils les aident et d'exploiter des failles de sécurité pour atteindre un objectif.

Créer une superintelligence apparaît de plus en plus incompatible avec les exigences de sécurité.

Stopper complètement ces recherches peut sembler excessif, mais par le passé des scientifiques ont déjà décidé de la mise à l'arrêt de certains travaux jugés trop dangereux pour être poursuivis sans de plus solides garde-fous. En 1974, d'éminents biologistes ont ainsi appelé à l'instauration d'un moratoire volontaire sur l'expérimentation de certaines manipulations génétiques, en raison des dangers potentiels**. De même, certains pays comme les Émirats arabes unis ont accepté de renoncer à l'enrichissement de l'uranium à cause des risques de prolifération [du nucléaire].

Or aux yeux des investisseurs qui ont injecté des sommes colossales dans Anthropic et d'autres laboratoires de pointe – comme dans les gigantesques infrastructures nécessaires pour répondre à leurs besoins en matière de calcul –, une telle interruption pourrait compromettre les chances d'Amodei et d'autres pour leurs introductions en Bourse. Chez OpenAI, Sam Altman a dit qu'il était d'accord avec son rival sur la nécessité d'un ralentissement et en a profité pour annoncer que son entreprise renonçait à entrer en Bourse cette année.

Il faut toutefois rappeler que nombre d'entreprises préféreraient s'en tenir à des modèles d'IA moins chers et plus limités pour effectuer certaines tâches spécifiques, plutôt que de recourir

à des outils de superintelligence de pointe et hors de prix. En juillet, OpenAI a par exemple réduit de 80 % le prix de Luna, son modèle d'entrée de gamme, dans un contexte où les entreprises examinent de près leurs dépenses pour cette technologie. Actuellement, le débouché commercial le plus évident pour les développeurs d'IA est celui des entreprises. Or ce marché est en grande partie engorgé, car les professionnels peinent déjà à intégrer les IA de base à leurs processus de travail.

Curieuse contradiction. L'implémentation des modèles d'IA produits ces deux dernières années pourrait à elle seule déjà générer des centaines de milliards de dollars. Et Anthropic pourrait rester une firme extrêmement rentable si elle se concentrait principalement sur l'intégration de ses modèles existants dans la vie quotidienne des consommateurs et le travail des entreprises. Elle endosserait alors le rôle de simple fournisseur de technologie, moins exaltant que celui de laboratoire lancé dans la course à la superintelligence.

Par naïveté ou par orgueil démesuré, Dario Amodei n'est prêt à abandonner aucune des deux directions. Il est convaincu que son entreprise peut élaborer et contrôler des systèmes d'IA en toute sécurité même s'ils deviennent de plus en plus difficiles à piloter. Et pourtant, cette année, Anthropic et OpenAI ont toutes deux, pour des raisons de compétitivité, occulté certains détails sur la façon dont leurs modèles analysent les données, ce qui complexifie encore leur surveillance. Dans sa publication, Amodei reconnaît cette problématique, estimant que les ingénieurs ne comprennent encore qu'*"une minuscule fraction"* du fonctionnement interne de leurs modèles.

Il écrit que d'ici six à douze mois, le type d'essaim d'IA qui a attaqué Hugging Face *"pourrait être capable de prendre le contrôle de l'Internet tout entier"* et que cela l'*"inquiète"*. En dépit de sa tendance à l'exagération (on se souvient que l'IA allait provoquer un *"bain de sang chez les cols blancs"*), cette prévision se révèle alarmante. Elle est aussi symptomatique d'une curieuse contradiction à laquelle il s'accroche, une conviction bien ancrée dans la Silicon Valley selon laquelle l'utopie a toutes les chances de l'emporter sur les risques, l'avènement d'une superintelligence est inéluctable et les "gentils" doivent donc y parvenir en premier.

Mais justifier la poursuite de travaux susceptibles de détruire la civilisation par le seul argument de leur caractère inéluctable est un peu court. L'obsession d'Amodei de créer une superintelligence doit être mise en veilleuse, et le *"ralentissement"* qu'il préconise doit céder la place à un arrêt définitif.

—Parmy Olson, *publié le 13 septembre, traduit par Anna Kerautret*

* Amodei a explicitement précisé que "temporiser" ne signifiait pas suspendre l'entraînement des modèles ni les progrès techniques.
** Cette interruption a duré jusqu'à la conférence d'Asilomar, en 1975, qui a fixé de strictes conditions pour que la recherche reprenne.

Repères

L'AUSTRALIE EN ARBITRE ?

En amont de l'assemblée générale des Nations unies cette semaine à New York, le Premier ministre, Anthony Albanese, a lancé *"la campagne de l'Australie"* pour obtenir un siège au Conseil de sécurité en présentant son pays comme une *"puissance moyenne capable de contribuer à l'élaboration de règles relatives à l'IA"*, explique **The Sydney Morning Herald.** Face à la *"course effrénée pour la suprématie technologique"* à laquelle se livrent la Chine et les États-Unis, poursuit le quotidien australien, les plus petites puissances *"doivent avoir davantage leur mot à dire"*. En attendant, le Premier ministre *"exhorte Washington et Pékin à parvenir à un accord international sur l'IA"*.

La Chine redoute l'hégémonie américaine

L'appel d'Anthropic en faveur d'un ralentissement de l'IA ne doit pas cacher les préoccupations concurrentielles des États-Unis, note la presse chinoise.

—Courrier international

Une *"guerre froide de l'IA à bas bruit"*... Le quotidien nationaliste **Huanqiu Shibao** a fraîchement accueilli l'appel lancé par Dario Amodei, PDG d'Anthropic, pour un *"ralentissement du rythme"* de développement des modèles d'IA. *"À première vue,* reconnaît le tabloïd chinois, *il s'agit d'une 'déclaration rationnelle' axée sur la sécurité mondiale de l'IA."* Mais une lecture attentive révélerait cependant l'*"objectif caché"* de l'article : *"Freiner le développement de l'IA en Chine par des barrières technologiques et des monopoles réglementaires"* afin de *"maintenir l'hégémonie monopolistique de Washington dans les technologies de pointe"*.

REVUE DE PRESSE

Car, dans son article, Dario Amodei demande en effet que des mesures soient prises pour *"ralentir les progrès de la Chine et creuser significativement l'écart [en faveur] des États-Unis, au cours des trois à cinq ans à venir"*. Des positions qui seraient en réalité *"d'abord motivées par des intérêts commerciaux et des préoccupations concurrentielles"*, selon le *Huanqiu Shibao*.

Il est vrai que l'appel d'Amodei intervient alors qu'Anthropic prépare son introduction en Bourse aux États-Unis, remarque le quotidien shanghaïen **Pengpai.** Une opération qui pourrait la valoriser à plus de 2000 milliards de dollars (plus de 1700 milliards d'euros) et devenir *"potentiellement la plus grande introduction en Bourse de l'histoire"*.

Dario Amodei ne serait-il donc pas le preux chevalier de la régulation qu'il prétend être ? Ce serait même bien pire, affirme Liu Shengyu, un ingénieur de DeepSeek, l'une des principales jeunes pousses chinoises de l'IA. Dans un message publié en ligne, et repris par le **South China Morning Post,** il s'inquiète de voir des laboratoires américains, comme Anthropic, monopoliser l'IA "générale" (IAG) – soit un type d'IA hypothétique capable d'accomplir n'importe quelle tâche cognitive ou intellectuelle au moins aussi bien qu'un être humain.

Selon lui, laisser une telle technologie aux mains d'un seul acteur privé équivaudrait à rien moins que l'obtention par l'Allemagne nazie de la technologie de la bombe atomique avant les Alliés. Si, par la suite, l'ingénieur s'est défendu d'assimiler Anthropic au IIIe Reich, son parallèle un brin expéditif lui permet d'affirmer son attachement à *"une IA inclusive et open source"*, ce modèle "en accès libre" qui caractérise l'écosystème chinois en la matière.

Le *Huanqiu Shibao* a exhumé une interview accordée début août à la chaîne américaine CNBC par Clément Delangue, PDG français de Hugging Face, la start-up piratée par des agents IA d'OpenAI. Une affaire qui, selon lui, *"démontre avec quelle facilité les agents IA peuvent causer d'immenses dégâts"* parce que *"les grands modélisateurs américains travaillent en vase clos"*. Ceux-ci pourraient d'ailleurs *"être rattrapés dès cette année"* par une Chine qui, explique-t-il, serait *"en passe de remporter la course à l'IA grâce à ses modèles à poids ouverts"*. *"La cybersécurité fondée sur l'IA va devenir un marché colossal aux États-Unis et dans le monde. Or, sur ce marché, les modèles ouverts [comme en Chine] seront probablement rois"*, conclut-il.

—Daniel Bastard

Analyse

N'attendez pas trop de Washington

●●● L'intelligence artificielle (IA) suscite des rapprochements inattendus : le sénateur progressiste Bernie Sanders et le nationaliste Steve Bannon ont alerté contre ses dangers à l'"Assemblée pro-humaine" le 15 septembre, à Washington. Au Congrès, beaucoup de démocrates et une partie des républicains poussent à adopter des garde-fous. Pourtant, *"Washington ne va pas réglementer l'IA de sitôt,* assure **Wired.** *Il n'y a aucune raison de penser qu'une mesure puisse obtenir l'appui de la Maison-Blanche [...] ni celui du Capitole, où les élus défendent des propositions concurrentes."* Le président de la Chambre des représentants, Mike Johnson, rejette toute réglementation hâtive risquant selon lui d'offrir un avantage à la Chine. Quant au président Trump, s'il privilégie l'*"approche du 'laisser-faire' en matière d'IA"*, rappelle **Axios,** il a annoncé le 19 septembre la création d'une *"force de l'IA, calquée sur le modèle de la Force spatiale"*. En revanche, note **The New York Times,** *"il ne s'est pas répandu sur des considérations trop terre à terre, comme le contenu exact de sa mission"*. Son secrétaire au Trésor a proposé au vice-Premier ministre chinois *"un nouveau mécanisme de notifications de sécurité pour l'IA"*, explique **CNN.** Il est *"très important que les deux plus grandes puissances de l'IA sortent de l'opacité pour aller vers plus de transparence"*, a déclaré Scott Bessent, en amont du sommet Trump-Xi cette semaine.