STORIES

Intelligence artificiellePlongée dans la dérive de celui qui voulait tuer le créateur de ChatGPT

6 min • Par élise viniacourt

Accusé d'avoir tenté d'incendier la maison de Sam Altman vendredi, Daniel M., un Américain de 20 ans, est persuadé que l'IA provoquera la fin de l'humanité. Une croyance portée par un mouvement alarmiste en partie incarné par les patrons de la tech : les «doomers».

Il est 3 h 37 du matin lorsqu'un bruit de verre brisé tire Russian Hill de son sommeil. Dans ce très chic quartier perché sur les hauteurs de San Francisco, un cocktail Molotov vient de s'écraser contre la grille d'une imposante résidence. Des flammes lèchent déjà la barrière lorsque des agents de sécurité se précipitent pour éviter que le feu ne se propage à cette luxueuse propriété à 27 millions de dollars. Une demeure aux allures de forteresse moderne, dotée d'un «garage Batcave» avec plateau tournant pour voitures et d'une piscine à débordement. En quelques minutes, vendredi, le nid douillet de Sam Altman, PDG du géant de l'intelligence artificielle OpenAI, est sauvé.

Lorsque la police arrive sur les lieux, les braises fument encore. Le mystérieux incendiaire s'est volatilisé mais des caméras de surveillance ont tout enregistré : le suspect a des traits juvéniles, un sweat-shirt et des baskets. Il faut moins de deux heures aux enquêteurs pour lui mettre la main dessus dans un quartier d'affaires de San Francisco à vingt minutes en voiture de la maison du créateur de ChatGPT. Daniel M., 20 ans, est alors en train de

frapper les vitres des locaux d'OpenAI à l'aide d'une chaise et menace «de tout brûler et de tuer tout le monde à l'intérieur», rapporte l'acte d'accusation fédéral. Le soleil se lève à peine lorsqu'il est placé en garde à vue. Et alors que le jeune homme est accusé de «tentative de meurtre», une question flotte dans l'air : qu'est-ce qui lui a pris ?

D'entrée de jeu, après ces événements, Sam Altman a pointé du doigt deux responsables : les craintes «justifiées» au sujet de l'IA… et la presse. Manipulations, ambitions démesurées, imposture… Quelques jours avant l'attaque, un article peu élogieux à son sujet est paru dans le New Yorker. Un récit qui, selon lui, aurait mis le feu aux poudres à l'heure où les conséquences de sa technologie sur l'emploi et ses usages militaires inquiètent : «Quelqu'un m'a dit que cet article survenait dans un contexte de grande anxiété liée à l'IA et que cela rendait les choses plus dangereuses pour moi», écrit-il sur son blog. Secoué, il publie une image de son fils, en bas âge, et de son mari : «Je partage une photo dans l'espoir qu'elle dissuadera quiconque de lancer un cocktail Molotov chez nous, peu importe ce qu'il pense de moi.»

Blog personnel, document judiciaire, analyses d'experts… Libé s'est penché sur les raisons qui auraient poussé Daniel M. à s'en prendre au patron d'une entreprise valorisée à 852 milliards de dollars. Son geste, isolé en apparence, révèle une angoisse profonde nourrie par l'alarmisme de certains collectifs, les théories catastrophistes de recherches controversées et les discours marketing des milliardaires de la tech en personne. Une peur, simple et viscérale : celle de l'extinction de l'humanité… causée par l'IA.

«Jihadiste butlérien»

Lorsque Daniel M. dégaine son briquet devant la villa de Sam Altman, il a déjà envoyé à d'anciens camarades d'université un document de plusieurs pages. Une sorte de manifeste anti-IA dans lequel le Texan - qui a parcouru les plus de 2 200 km le séparant de la Californie pour l'occasion - juge la fin de l'espèce humaine «imminente». Selon lui, une superintelligence

artificielle serait sur le point d'émerger. Capable d'échapper à tout contrôle, elle poursuivrait coûte que coûte ses objectifs avec une efficacité telle qu'elle mettrait en péril la survie de l'humanité. Pour empêcher ce scénario, il se dit prêt au pire : «Si je dois inciter d'autres personnes à tuer et à commettre des crimes, je dois montrer l'exemple et prouver que mon message est sincère», écrit-il. Ses derniers mots sont pour Sam Altman : «Si, par miracle, tu survis, alors je prendrai cela comme un signe divin t'invitant à te racheter...»

D'où Daniel M. tire-t-il ses certitudes ? Son avocate commise d'office, Diamond Solange Ward, a affirmé mardi qu'il avait «des antécédents d'autisme» et traversait au moment des faits «une grave crise de santé mentale». Plusieurs pans de sa biographie restent flous. Sur son blog, l'Américain - qui n'aurait pas de formation d'ingénieur comme il l'affirme - raconte avoir grandi au milieu d'une famille protestante «particulièrement croyante». Son père aurait vécu «une expérience de mort imminente» qui l'aurait conduit à «vouer une grande dévotion à Dieu, à diriger de longues prières et même à fonder deux groupes d'étude biblique en espagnol». Pour Daniel M., qui explique avoir été plus jeune «rongé par une peur de la mort presque paralysante», la religion aurait longtemps fait office de «béquille». Mais sa foi, peu à peu, se serait étiolée. A l'issue d'une crise existentielle, le vingtenaire aurait forgé une conviction : le «but de l'humanité» serait de «faire tout son possible pour améliorer le niveau de vie de ses enfants». Un objectif que les développements de l'IA entraveraient.

Le 11 juin 2024, Daniel M. rejoint pour la première fois le serveur Discord - une plateforme de conversation en ligne - de Pause IA. Fondé en 2023 aux Pays-Bas, ce mouvement politique mondial appelle à un «moratoire sur l'entraînement des systèmes les plus dangereux» et alerte sur les «risques catastrophiques» de l'IA. Lors de son inscription, le vingtenaire opte pour un surnom évocateur : «jihadiste butlérien». Dans l'univers du livre de fiction Dune, le «jihad butlérien» est une période historique imaginée par l'auteur Frank Herbert au cours de laquelle les humains entrent en guerre contre les «machines pensantes», des ordinateurs et IA surpuissants ayant asservi

l'homme. Ce conflit aboutit à la mise hors la loi de certaines technologies. Dans ses premiers messages, Daniel M. dit vouloir «aider par tous les moyens nécessaires» et demande «combien de temps» il reste à l'humanité. Peu à peu, son ton se durcit.

Le 16 juin 2024, le Texan partage une lettre envoyée à Dan Crenshaw, un député républicain. «Ce qui me terrifie, c'est l'idée qu'un petit cartel d'individus [les PDG de l'IA, ndlr] a réussi à duper notre gouvernement et le public, lui écrit-il. Je suis effrayé de les voir jouer à la roulette russe avec nos vies.» Un an plus tard, le 6 novembre, il appelle Pause IA à des actions plus radicales : «Nous devons être plus forts que ça et au moins être prêts à mourir en combattant si besoin.» Le 3 décembre, en référence à l'horloge de l'apocalypse, il annonce : «Il est presque minuit, il est grand temps d'agir.» Cette fois, un modérateur lui adresse un avertissement pour «promotion de la violence».

Au total, Daniel M. publie 34 messages sur le serveur. «Il n'avait aucun rôle au sein de notre organisation, n'a participé à aucune campagne ni à aucun événement et n'a reçu aucun soutien de notre part», rapporte à Libé Maxime Fournes, cofondateur de la branche française de Pause IA. Tout en décrivant son mouvement comme «non-violent», l'ingénieur condamne «sans équivoque» l'attaque au cocktail Molotov menée contre le domicile du PDG d'OpenAI à qui il souhaite par ailleurs «la sécurité et la paix». «Notre Discord est un espace communautaire ouvert à tous. Nous assurons une modération active mais nous ne pouvons pas prévoir les actions de chaque personne qui rejoint un forum ouvert», précise-t-il enfin.

Catastrophistes

Après son arrestation, Daniel M. a été banni du serveur du collectif. Ses messages ont été supprimés… et republiés sur X par une chercheuse américaine en communication, Nirit Weiss-Blatt, qui les avait enregistrés. Auprès de Libé, cette spécialiste des discours de la tech juge que certaines positions de Pause IA seraient malgré tout en partie responsables de la

situation : «L'idéologie selon laquelle "l'IA va tous nous tuer" pousse certains à des idées radicales», explique celle qui affirme «avoir tiré la sonnette d'alarme à maintes reprises» par le biais de sa newsletter baptisée AI Panic. Aujourd'hui, Nirit Weiss-Blatt formule un sinistre présage : «Daniel M. ne sera pas le dernier à en venir à la conclusion qu'il faut recourir à la violence.» D'ailleurs, il ne serait pas non plus le premier : déjà en 2025, Wired rapportait que les employés d'OpenAI à San Francisco avaient été confinés tout un après-midi dans leur bureau, après des menaces proférées par un ancien militant de Stop AI, un autre collectif «non-violent» alertant sur «le risque d'extinction».

Ces catastrophistes algorithmiques portent un nom : les «doomers». Comprendre, les acteurs alertant sur les cataclysmes mondiaux que pourrait amener l'IA. L'une de leurs figures de proue ? Le controversé futurologue Nick Bostrom, auteur du livre Superintelligence, paru en 2014 dans lequel il argumente qu'une IA pourrait prendre le contrôle du monde pour mieux accomplir ses objectifs. «Il ne fait pas d'études empiriques mais des exercices de pensée philosophiques, qui ont pourtant une forte résonance auprès des ingénieurs», décrypte auprès de Libé Irénée Régnauld, sociologue et cofondateur du collectif technocritique le Mouton numérique. Chez les doomers, les scénarios parfumés à la science-fiction foisonnent : on imagine qu'un modèle pourrait se rebeller dans un souci d'auto-préservation, faire preuve de mensonge pour mieux remplir sa mission, s'auto-améliorer de façon exponentielle… Autant de théories qui, selon le sociologue, ne font l'objet «d'aucun consensus scientifique» et même plutôt de «vifs débats».

«C'est du charabia»

Pourtant, ces mises en garde trouvent des échos inattendus. Des chercheurs reconnus comme Geoffrey Hinton estiment entre 10 % et 20 % les risques que l'humanité soit détruite dans les prochaines décennies. Plus étonnant : les capitaines de l'industrie aussi relaient ces craintes. Elon Musk a appelé à une pause dans le développement des modèles les plus puissants, Dario

Amodei, PDG d'Anthropic, a évoqué un «péril existentiel» dans un essai récent et Sam Altman lui-même a parlé de risques «d'extinction». «Ce sont des opérations d'enfumage qui visent à montrer à quel point leur technologie est puissante», observe Irénée Régnauld. Des opérations marketing inversées, en somme.

Problème : aux yeux de Daniel M., ces lanceurs d'alerte milliardaires sont sincères. «Musk, Altman, Amodei... Si vous ne les croyez pas, interrogez-vous peut-être sur leurs investissements quasi unanimes dans des bunkers anti-apocalypse», écrit-il sur son blog. Pourquoi ces patrons continuent-ils malgré tout de construire cette technologie ? Car en plus de «ne pas réfléchir comme nous», ils manqueraient «de principes moraux», poursuit le vingtenaire. A commencer, selon lui, par Sam Altman. «Ces personnes sont très probablement des sociopathes-psychopathes et, dans le cas d'Altman, elles sont régulièrement décrites comme des menteurs pathologiques», affirme-t-il. Avant de conclure : «Nous sommes condamnés si nous n'agissons pas immédiatement.»

Faut-il voir dans l'histoire de Daniel M. l'émergence d'un militantisme anti-IA radical ? Pour le philosophe Eric Sadin, il convient de garder la tête froide : «Le discours des doomers, c'est du charabia. On est dans un délire de science-fiction anthropomorphique.» L'auteur du Désert de nous-mêmes estime que «les enjeux sont ailleurs» - du côté des collectifs alertant sur l'impact environnemental de l'IA ou des syndicats pointant ses effets «concrets» sur l'emploi. «Le plus grand risque, c'est le renoncement à l'usage de nos facultés les plus fondamentales : attendre de ces systèmes toute la vérité du monde au risque de se voir dépossédés de notre sensibilité, de notre jugement, de notre pensée», détaille-t-il. Selon l'intellectuel, l'erreur première des doomers est peut-être la plus fondamentale : se méprendre sur ce qu'est vraiment «la fin de l'humanité».