Alors que le développement de l'intelligence artificielle progresse à une vitesse fulgurante, certains des hommes les plus riches du monde sont peut-être en train de décider du sort de l'humanité. Les prévisions en matière d'évolution de l'IA oscillent entre promesses évolutionnistes peu crédibles et anticipations dans lesquelles l'IA menacerait l'existence même de l'espèce humaine.

Cette thèse d'une possible extinction de notre espèce se déploie dans différents milieux, dont certains liés à l'industrie de l'IA, et incitent certains acteurs à en appeler à un ralentissement voire un arrêt total de son développement.

***

Larry Page, cofondateur de Google, pense que l'IA superintelligente n'est que « la prochaine étape de l'évolution ». En fait, Larry Page, qui pèse environ 120 milliards de dollars, aurait affirmé que les efforts visant à empêcher l'extinction de l'IA et à protéger la conscience humaine sont « spécistes » et « un non-sens sentimental« .

En juillet, Richard Sutton, ancien scientifique principal de Google DeepMind et l'un des pionniers de l'apprentissage par renforcement, un sous-domaine majeur de l'IA, a déclaré que la technologie « pourrait nous faire disparaître de l'existence » et que « nous ne devrions pas résister à la succession ». Dans une conférence donnée en 2015, Richard Sutton a déclaré : « supposez que « tout échoue » et que l'IA « nous tue tous » ». Il a demandé : « Est-il si grave que les humains ne soient pas la dernière forme de vie intelligente dans l'univers ?

« L'extinction biologique n'est pas la question », m'a répondu Richard Sutton, âgé de 66 ans. « La lumière de l'humanité et notre compréhension, notre intelligence – notre conscience, si vous voulez – peuvent continuer sans humains en chair et en os.

Yoshua Bengio, 59 ans, est le deuxième scientifique vivant le plus cité, connu pour ses travaux fondamentaux sur l'apprentissage profond. Répondant à Page et Sutton, Bengio m'a dit : « Ce qu'ils veulent, je pense que c'est jouer aux dés avec l'avenir de l'humanité. Personnellement, je pense que cela devrait être criminalisé. » Un peu surpris, je lui ai demandé ce qu'il souhaitait exactement voir interdit, et il m'a répondu que les efforts visant à construire « des systèmes d'IA qui pourraient nous dominer et qui seraient conçus dans leur propre intérêt ». En mai, Bengio a commencé à écrire et à s'exprimer sur la manière dont les systèmes d'IA avancés pourraient s'égarer et constituer un risque d'extinction pour l'humanité.

Bengio estime que les futurs systèmes d'IA de niveau véritablement humain pourraient améliorer leurs propres capacités, créant ainsi une nouvelle espèce plus intelligente. L'humanité a provoqué l'extinction de centaines d'autres espèces, en grande partie par accident. Il craint que nous ne soyons les prochains, et il n'est pas le seul.

Bengio a partagé le prix Turing 2018, le prix Nobel de l'informatique, avec ses collègues pionniers de l'apprentissage profond Yann Le Cun et Geoffrey Hinton. Hinton, le scientifique vivant le plus cité, a fait des vagues en mai lorsqu'il a démissionné de son poste de direction chez Google pour s'exprimer plus librement sur la possibilité que les futurs systèmes d'IA puissent anéantir l'humanité. Hinton et Bengio sont les deux chercheurs en IA les plus éminents à avoir rejoint la communauté « *x-risk*« . (Risque existentiel). Parfois appelé « défenseurs de la sécurité de l'IA » ou « prophètes de malheur », ce groupe peu structuré craint que l'IA ne représente un risque existentiel pour l'humanité.

Le mois même où Hinton a démissionné de Google, des centaines de chercheurs et de chercheuses en IA et de personnalités ont signé une lettre ouverte déclarant :

> « L'atténuation du risque d'extinction par l'IA devrait être une priorité mondiale au même titre que d'autres risques à l'échelle de la société, tels que les pandémies et les guerres nucléaires ».

Hinton et Bengio sont les principaux signataires, suivis par le PDG d'OpenAI, Sam Altman, et les directeurs d'autres grands laboratoires d'IA.

Hinton et Bengio sont également les premiers auteurs d'un document de synthèse publié en octobre, qui met en garde contre le risque d'une « perte irréversible du contrôle humain sur les systèmes d'IA autonomes », rejoint par des universitaires de renom tels que le lauréat du prix Nobel Daniel Kahneman et l'auteur de *Sapiens*, Yuval Noah Harari.

Yann Le Cun, qui dirige l'IA chez Meta, reconnaît que l'IA de niveau humain est à venir, mais a déclaré lors d'un débat public contre Bengio sur l'extinction de l'IA : « Si c'est dangereux, nous ne le construirons pas. »

L'apprentissage profond alimente les systèmes d'IA les plus avancés au monde, du modèle de pliage des protéines de Google DeepMind aux grands modèles de langage (LLM) comme le ChatGPT d'OpenAI. Personne ne comprend vraiment comment fonctionnent les systèmes d'apprentissage profond, mais leurs performances n'ont cessé de s'améliorer. Ces systèmes ne sont pas conçus pour fonctionner selon un ensemble de principes bien compris, mais sont plutôt « entraînés » à analyser des modèles dans de vastes ensembles de données, ce qui a pour effet de faire émerger un comportement complexe, comme la compréhension du langage. Connor Leahy, développeur en IA, m'a dit : « C'est plus comme si nous étions en train d'introduire quelque chose dans une boîte de Petri » que comme si nous écrivions un morceau de code. Le document de synthèse d'octobre prévient que « personne ne sait actuellement comment aligner de manière fiable le comportement de l'IA sur des valeurs complexes ».

Malgré toutes ces incertitudes, les entreprises spécialisées dans l'IA se considèrent comme engagées dans une course pour rendre ces systèmes aussi puissants que possible, sans plan réaliste pour comprendre comment les choses qu'elles créent fonctionnent réellement, tout en rognant sur la sécurité pour gagner des parts de marché. L'intelligence artificielle générale (IAG) est le Saint-Graal vers lequel tendent explicitement les principaux laboratoires d'IA. L'IAG est souvent définie comme un système qui est au moins aussi performant que les humains dans presque toutes les tâches intellectuelles. C'est aussi la chose qui, selon Bengio et Hinton, pourrait conduire à la fin de l'humanité.

Bizarrement, de nombreuses personnes qui s'emploient activement à développer les capacités de l'IA pensent qu'il y a de fortes chances que cela provoque l'apocalypse. Une enquête menée en 2022 auprès de chercheurs en apprentissage automatique a révélé que près de la moitié d'entre eux pensaient qu'il y avait au moins 10 % de chances que l'IA avancée conduise à « l'extinction de l'humanité ou [à] une déresponsabilisation permanente et grave similaire » de l'humanité. Quelques mois avant de cofonder d'OpenAI, Altman avait déclaré : « L'IA conduira probablement à la fin du monde, mais entre-temps, il y aura de grandes entreprises ».

L'opinion publique sur l'IA s'est dégradée, en particulier au cours de l'année qui a suivi

l'apparition de ChatGPT. Dans tous les sondages de 2023, sauf un, les Étatsuniens sont plus nombreux à penser que l'IA pourrait constituer une menace existentielle pour l'humanité. Dans les rares cas où les sondeurs ont demandé aux gens s'ils voulaient une IA de niveau humain ou supérieure, de fortes majorités aux États-Unis et au Royaume-Uni ont répondu par la négative.

Jusqu'à présent, lorsque les socialistes s'expriment sur l'IA, c'est généralement pour mettre en évidence la discrimination induite par l'IA ou pour mettre en garde contre l'impact potentiellement négatif de l'automatisation dans un monde où les syndicats sont faibles et les capitalistes puissants. Mais la gauche est restée remarquablement silencieuse sur le scénario cauchemardesque de Hinton et Bengio, selon lequel l'IA avancée pourrait tous nous tuer.

Des capacités inquiétantes

Si l'attention de la communauté de l'*x-risk* (risque existentiel) se concentre sur l'idée que l'humanité pourrait un jour perdre le contrôle de l'IA, nombreux sont ceux qui s'inquiètent également de voir des systèmes moins performants donner du pouvoir à de mauvais acteurs dans des délais très courts.

Heureusement, il est difficile de fabriquer une arme biologique. Mais cela pourrait bientôt changer.

Anthropic, un laboratoire d'IA de premier plan fondé par d'anciens employés de OpenAI soucieux de sécurité, a récemment collaboré avec des experts en biosécurité pour déterminer dans quelle mesure un LLM (Grand modèle de langage) pourrait aider un bioterroriste en herbe. Témoignant devant une sous-commission sénatoriale en juillet, le PDG d'Anthropic, Dario Amodei, a indiqué que certaines étapes de la production d'armes biologiques ne peuvent être trouvées dans les manuels ou les moteurs de recherche, mais que « les outils d'IA d'aujourd'hui peuvent remplir certaines de ces étapes, bien qu'incomplètement » et qu' »une extrapolation directe des systèmes actuels à ceux que nous prévoyons de voir dans deux ou trois ans suggère un risque substantiel que les systèmes d'IA soient capables de remplir toutes les pièces manquantes ».

En octobre, *New Scientist* a rapporté que l'Ukraine avait utilisé pour la première fois sur le champ de bataille des armes autonomes létales (SALA) – littéralement des robots tueurs. Les États-Unis, la Chine et Israël développent leurs propres robots tueurs. La Russie s'est jointe aux États-Unis et à Israël pour s'opposer à une nouvelle loi internationale sur les robots tueurs.

Cependant, l'idée plus large selon laquelle l'IA pose un risque existentiel a de nombreux détracteurs, et le discours sur l'IA est difficile à analyser : des personnes tout aussi crédibles s'opposent sur la question de savoir si le risque x de l'IA est réel, et des investisseurs en capital-risque signent des lettres ouvertes avec des éthiciens progressistes de l'IA. Tandis que l'idée du risque existentiel semble gagner du terrain le plus rapidement, une grande publication publie presque chaque semaine un essai soutenant que le risque existentiel détourne l'attention des préjudices existants. Pendant ce temps, beaucoup plus d'argent et de personnes sont discrètement consacrés à rendre les systèmes d'IA plus puissants qu'à les rendre plus sûrs ou moins biaisés.

Certains craignent non pas le scénario de science-fiction dans lequel les modèles d'IA deviennent si performants qu'ils nous arrachent le contrôle, mais plutôt que nous confiions trop de responsabilités à des systèmes biaisés, fragiles et confabulateurs, ouvrant ainsi une boîte de Pandore, pleine de problèmes terribles mais familiers qui évoluent en fonction des algorithmes qui les causent. Cette communauté de chercheurs et de défenseurs – souvent appelée *AI ethics*, («éthiques de l'IA ») – a tendance à se concentrer sur les dommages immédiats causés par l'IA, en explorant des solutions impliquant la responsabilité des modèles, la transparence algorithmique et l'équité de l'apprentissage automatique.

J'ai parlé avec certaines des voix les plus éminentes de la *communauté éthique de l'IA* (*AI ethics community*), comme les informaticiennes Joy Buolamwini, 33 ans, et Inioluwa Deborah Raji, 27 ans. Chacune d'entre elles a mené des recherches novatrices sur les préjudices causés par des modèles d'IA discriminatoires et défectueux dont les effets, selon elles, sont occultés un jour et surmédiatisés le lendemain. Comme celui de nombreux chercheurs en éthique de l'IA, leur travail mêle science et activisme.

Ceux à qui j'ai parlé dans le monde de l'éthique de l'IA ont largement exprimé l'idée que, plutôt que de faire face à des défis fondamentalement nouveaux comme la perspective d'un chômage technologique complet ou d'une extinction, l'avenir de l'IA ressemble davantage à une intensification de la discrimination raciale dans les décisions d'incarcération et de crédit, à la mise en entrepôt des lieux de travail à la façon d'Amazon, à des attaques contre les travailleurs pauvres et à une techno-élite qui s'enracine et s'enrichit encore plus.

L'un des arguments fréquemment avancés par ces personnes est que le récit de l'extinction surestime les capacités des produits des GAFAM et « détourne » dangereusement l'attention des effets néfastes immédiats de l'IA. Au mieux, disent-ils, l'idée d'un risque existentiel est une perte de temps et d'argent. Au pire, elle conduit à des idées politiques désastreuses.

Mais de nombreux partisans du risque existentiel ont souligné que les positions selon lesquelles « l'IA cause des dommages maintenant » et « l'IA pourrait mettre fin au monde » ne s'excluent pas mutuellement. Certains chercheurs ont explicitement tenté de combler le fossé entre ceux qui se concentrent sur les dommages existants et ceux qui se concentrent sur l'extinction, en mettant en évidence des potentiels objectifs politiques communs. Le professeur d'IA Sam Bowman, dont le nom figure également dans la lettre relative à l'extinction, a mené des recherches visant à révéler et à réduire les biais algorithmiques et à examiner les propositions soumises à la principale conférence sur l'éthique de l'IA. Simultanément, Bowman a demandé que davantage de chercheurs travaillent sur la sécurité de l'IA et a écrit sur les « dangers de la sous-estimation » des capacités des LLM (Grands modèles de langage).

La communauté des *x-risk* (risque existentiel) invoque couramment la défense du climat comme analogie, se demandant si l'accent mis sur la réduction des dommages à long terme du changement climatique ne détourne pas dangereusement l'attention des dommages à court terme causés par la pollution de l'air et les marées noires.

Mais, de leur propre aveu, les partisans du risque existentiel ne sont pas tous aussi diplomates. En août 2022, le cofondateur d'Anthropic, Jack Clark, a *tweeté* que « certaines personnes qui travaillent sur des politiques à long terme de type IA ont tendance à ignorer, minimiser ou simplement ne pas prendre en compte les problèmes immédiats liés au déploiement de l'IA et aux dommages qu'elle cause ».

« L'IA va sauver le monde »

Un troisième camp s'inquiète du fait qu'en matière d'IA, nous n'avançons pas assez vite. D'éminents capitalistes comme le milliardaire Marc Andreessen conviennent avec les spécialistes de la sécurité que l'IA est possible, mais affirment qu'au lieu de nous tuer tous, elle inaugurera un âge d'or indéfini d'abondance radicale et de technologies à la limite de la magie. Ce groupe, issu en grande partie de la Silicon Valley et communément appelé les « *boosters* » de l'IA, a tendance à s'inquiéter beaucoup plus de la réaction excessive des autorités réglementaires à l'égard de l'IA, qui risque d'étouffer une technologie transformatrice et salvatrice dans son berceau, condamnant ainsi l'humanité à la stagnation économique.

Certains techno-optimistes envisagent une utopie alimentée par l'IA qui fait paraître Karl Marx comme peu imaginatif. Le *Guardian* a récemment publié un mini-documentaire présentant des entretiens réalisés entre 2016 et 2019 avec le scientifique en chef d'OpenAI, Ilya Sutskever, qui déclare avec audace :

> « l'IA résoudra tous les problèmes que nous connaissons aujourd'hui. Elle résoudra les problèmes d'emploi, de maladie et de pauvreté. Mais elle créera aussi de nouveaux problèmes. »

Andreessen est d'accord avec Sutskever – jusqu'au « mais ». En juin, Andreessen a publié un essai intitulé « Pourquoi l'IA sauvera le monde« , dans lequel il explique comment l'IA améliorera « tout ce qui nous tient à cœur », à condition que nous ne la réglementions pas jusqu'à la mort. Il a poursuivi en octobre avec son « Manifeste techno-optimiste » qui, en plus de faire l'éloge d'un fondateur du fascisme italien, désigne comme ennemis du progrès des idées telles que le « risque existentiel », la « durabilité », la « confiance et la sécurité » et l' »éthique de la technologie ». Andreessen ne mâche pas ses mots : « Nous pensons que tout ralentissement de l'IA coûtera des vies. Les décès qui auraient pu être évités par l'IA dont on a empêché l'existence [sont] une forme de meurtre ».

Andreessen est peut-être, avec Martin Shkreli, le plus célèbre partisan de l' »accélérationnisme efficace« , également appelé « e/acc« , un réseau essentiellement en ligne qui mêle scientisme sectaire, hypercapitalisme et sophisme naturaliste. E/acc, qui est devenu viral cet été, s'appuie sur la théorie de l'accélérationnisme de l'écrivain réactionnaire Nick Land, qui affirme que nous devons intensifier le capitalisme pour nous propulser dans un avenir post-humain, alimenté par l'IA. E/acc reprend cette idée et y ajoute une couche de physique et de mèmes, la généralisant pour un certain sous-ensemble d'élites de la Silicon Valley. Il a été créé en réaction aux appels des « défecteurs » à ralentir l'IA, qui sont venus en grande partie de la communauté de l'altruisme efficace (EA), d'où e/acc tire son nom.

Le promoteur de l'IA Richard Sutton – le scientifique prêt à faire ses adieux aux « humains en chair et en os » – travaille maintenant à Keen AGI, une nouvelle start-up de John Carmack, le légendaire programmeur à l'origine du jeu vidéo *Doom* des années 1990. La mission de l'entreprise, selon John Carmack : « *AGI or bust, by way of Mad Science !* » (L'IAG ou l'effondrement, par le biais de la science folle).

Le capitalisme aggrave la situation

En février, Sam Altman a tweeté qu'Eliezer Yudkowsky pourrait finalement « mériter le prix Nobel de la paix ». Pourquoi ? Parce qu'Altman pensait que le chercheur autodidacte et auteur de fanfictions sur Harry Potter avait fait « plus pour accélérer l'IAG (intelligence artificielle générale) que n'importe qui d'autre ». Altman a cité la façon dont Yudkowsky a aidé Google DeepMind à obtenir un financement de départ de Peter Thiel, ainsi que le rôle « critique » de Yudkowsky « dans la décision de lancer OpenAI ».

Yudkowsky était un accélérationniste avant même que le terme ne soit inventé. À l'âge de 17 ans, lassé des dictatures, de la faim dans le monde et de la mort elle-même, il a publié un manifeste demandant la création d'une superintelligence numérique pour « résoudre » tous les problèmes de l'humanité. Au cours des dix années suivantes, sa « technophilie » s'est transformée en phobie et, en 2008, il a écrit sur son histoire de conversion, admettant que « dire que *j'ai failli détruire le monde* aurait été trop orgueilleux ».

Yudkowsky est désormais célèbre pour avoir popularisé l'idée que l'intelligence artificielle pourrait tuer tout le monde, et il est devenu le plus sombre des catastrophistes de l'IA. Une génération de techniciens a grandi en lisant les articles de blog de Yudkowsky, mais davantage d'entre eux (et peut-être surtout Altman) ont intériorisé ses arguments selon lesquels l'IAG (Intelligence Artificielle Générale) serait la chose la plus importante de tous les temps, plutôt que ses convictions sur la difficulté qu'il y aurait à faire en sorte qu'elle ne nous tue pas.

Au cours de notre conversation, Yudkowsky a comparé l'IA à une machine qui « imprime de l'or », jusqu'à ce qu'elle « enflamme l'atmosphère ».

Qu'elle mette ou non le feu à l'atmosphère, cette machine imprime de l'or plus rapidement que jamais. Le boom de l' »IA générative » rend certaines personnes très, très riches. Depuis 2019, Microsoft a investi un montant cumulé de 13 milliards de dollars dans OpenAI. Porté par le succès fou de ChatGPT, Microsoft a gagné près de 1 000 milliards de dollars en valeur dans l'année qui a suivi la sortie du produit. Aujourd'hui, cette entreprise presque cinquantenaire vaut plus que Google et Meta réunis.

Les acteurs qui cherchent à maximiser leurs profits continueront à aller de l'avant, externalisant les risques que le reste d'entre nous n'a jamais accepté de supporter, à la poursuite de la richesse – ou simplement de la gloire de créer une superintelligence numérique, ce qui, comme l'a *tweeté* Sutton, « sera la plus grande réalisation intellectuelle de tous les temps … dont l'importance dépasse l'humanité, la vie, le bien et le mal ». Les pressions du marché pousseront probablement les entreprises à transférer de plus en plus de pouvoir et d'autonomie aux systèmes d'IA au fur et à mesure qu'ils s'amélioreront.

Un chercheur en IA de Google m'a écrit : « Je pense que les grandes entreprises sont tellement pressées de gagner des parts de marché que la sécurité [de l'IA] est considérée comme une sorte de distraction stupide ». Bengio m'a dit qu'il voyait « une course dangereuse entre les entreprises » qui pourrait encore s'aggraver.

Paniqué par le moteur de recherche Bing alimenté par OpenAI, Google a déclaré « l'état d'urgence », « recalibré » son appétit pour le risque et s'est empressé de lancer Bard, son

LLM, malgré l'opposition du personnel. Lors de discussions internes, les employés ont qualifié Bard de « menteur pathologique » et d' »indigne ». Google l'a tout de même lancé.

Dan Hendrycks, directeur du *Center for AI Safety* (Centre pour la sécurité de l'IA), a déclaré que « faire des économies sur la sécurité [...] est en grande partie ce qui motive le développement de l'IA. [...] . . Je ne pense pas, en fait, qu'en présence de ces pressions concurrentielles intenses, les intentions aient une importance particulière ». Ironiquement, Hendrycks est également le conseiller en matière de sécurité de xAI, la dernière entreprise d'Elon Musk.

Les trois principaux laboratoires d'IA ont tous commencé par être des organisations indépendantes et motivées par leur mission, mais ils sont aujourd'hui soit des filiales à part entière de géants de la technologie (Google DeepMind), soit ils ont reçu tellement de milliards de dollars d'investissements de la part d'entreprises valant des milliards de dollars que leurs missions altruistes risquent d'être supplantées par la quête sans fin de la valeur actionnariale (Anthropic a reçu jusqu'à 6 milliards de dollars de Google et d'Amazon réunis, et les 13 milliards de dollars de Microsoft leur ont permis d'acheter 49 % de la branche à but lucratif de OpenAI). Le *New York Times* a récemment rapporté que les fondateurs de DeepMind sont devenus « de plus en plus inquiets de ce que Google ferait de leurs inventions ». En 2017, ils ont tenté de se séparer de l'entreprise. Google a réagi en augmentant les salaires et les primes en actions des fondateurs de DeepMind et de leur personnel. Ils sont restés « .

Un développeur d'un laboratoire de premier plan m'a écrit en octobre que, puisque les dirigeants de ces laboratoires sont généralement convaincus que l'IA éliminera le besoin d'argent, la recherche de profit est « largement instrumentale » à des fins de collecte de fonds. Mais « ensuite, les investisseurs (qu'il s'agisse d'une société de capital-risque ou de Microsoft) exercent une pression en faveur de la recherche de bénéfices ».

Entre 2020 et 2022, plus de 600 milliards de dollars d'investissements d'entreprises ont afflué dans le secteur, et une seule conférence sur l'IA en 2021 a accueilli près de trente mille chercheurs. Dans le même temps, une estimation de septembre 2022 n'a trouvé que quatre cents chercheurs à temps plein sur la sécurité de l'IA, et la principale conférence sur l'éthique de l'IA a accueilli moins de neuf cents participants en 2023.

De la même manière que les logiciels ont « mangé le monde« , il faut s'attendre à ce que l'IA présente une dynamique similaire de « *winner-takes-all* » qui conduira à des concentrations de richesse et de pouvoir encore plus importantes. Altman a prédit que le « coût de l'intelligence » tomberait à près de zéro grâce à l'IA et, en 2021, il a écrit que « le pouvoir passera encore plus du travail au capital ». Il poursuit : « Si les politiques publiques ne s'adaptent pas en conséquence, la situation de la plupart des gens sera pire qu'elle ne l'est aujourd'hui. » Également dans son fil de discussion, Jack Clark a écrit que « le capitalisme à économie d'échelle est, par nature, anti-démocratique, et l'IA à forte intensité de capital est donc anti-démocratique ».

Markus Anderljung est le responsable de GovAI, un groupe de réflexion sur la sécurité de l'IA, et le premier auteur d'un livre blanc influent sur la réglementation de l' »IA frontière ». Il m'a écrit pour me dire : « Si vous vous inquiétez du capitalisme sous sa forme actuelle, vous devriez vous inquiéter encore plus d'un monde où d'énormes pans de l'économie sont gérés par des systèmes d'IA explicitement entraînés à maximiser les profits. »

Sam Altman, vers juin 2021, s'est dit d'accord, racontant à Ezra Klein la philosophie

fondatrice d'OpenAI : « L'une des incitations qui nous rendait très nerveux était l'incitation au profit illimité, où plus c'est toujours mieux. . .. Et je pense qu'avec ces systèmes d'intelligence artificielle très puissants et à usage général, en particulier, vous ne voulez pas d'une incitation à maximiser indéfiniment le profit ».

Dans une décision stupéfiante qui a été largement considérée comme le point le plus important du débat sur la sécurité de l'IA jusqu'à présent, le conseil d'administration à but non lucratif d'Open-AI a licencié le PDG Sam Altman le 17 novembre 2023, le vendredi précédant la fête de Thanksgiving. Le conseil d'administration, conformément à la charte inhabituelle d'Open-AI, a une obligation fiduciaire envers « l'humanité », plutôt qu'envers les investisseurs ou les employés. Pour justifier sa décision, le conseil a vaguement cité le manque de franchise d'Altman, mais, ironiquement, il a ensuite gardé le silence sur sa décision.

Le lundi suivant, vers 3 heures du matin, Microsoft a annoncé qu'Altman allait créer un laboratoire de recherche avancée avec des postes pour tous les employés d'OpenAI, dont la grande majorité a signé une lettre menaçant d'accepter l'offre de Microsoft si Altman n'était pas réintégré. (Bien qu'il semble être un PDG populaire, il convient de noter que le licenciement a perturbé la vente prévue des actions détenues par les employés d'OpenAI, dont la valeur est estimée à 86 milliards de dollars). Le mercredi, peu après une heure du matin, OpenAI a annoncé le retour de Sam Altman au poste de PDG et l'arrivée de deux nouveaux membres au conseil d'administration : l'ancien président du conseil d'administration de Twitter et l'ancien secrétaire d'État au Trésor, Larry Summers.

En moins d'une semaine, les dirigeants d'OpenAI et Sam Altman ont collaboré avec Microsoft et le personnel de l'entreprise pour organiser son retour et le retrait de la plupart des membres du conseil d'administration à l'origine de son licenciement. La première préférence de Microsoft était que Sam Altman redevienne PDG. L'éviction inattendue a d'abord fait chuter l'action du géant technologique de 5 % (140 milliards de dollars), puis l'annonce de la réintégration de Sam Altman a fait grimper l'action à un niveau record. Craignant d'être à nouveau « pris au dépourvu« , Microsoft occupe désormais un siège sans droit de vote au sein du conseil d'administration de l'organisation à but non lucratif.

Immédiatement après le licenciement d'Altman, X (ex Twitter) a explosé, et un récit largement alimenté par des rumeurs en ligne et des articles anonymes a émergé, selon lequel les altruistes efficaces axés sur la sécurité au sein du conseil d'administration avaient licencié Altman en raison de sa commercialisation agressive des modèles d'OpenAI au détriment de la sécurité. Saisissant la teneur de la réaction massive des e/acc, le fondateur pseudonyme @BasedBeffJezos a posté : « Les EA's (altruistes efficaces) sont fondamentalement des terroristes. Détruire 80 milliards de dollars de valeur en une nuit est un acte de terrorisme ».

L'image qui s'est dégagée des rapports ultérieurs est qu'une méfiance fondamentale à l'égard d'Altman, et non des préoccupations immédiates concernant la sécurité de l'IA, a motivé le choix du conseil d'administration. Le *Wall Street Journal* a constaté que « ce n'est pas un incident unique qui a conduit à la décision d'éjecter Altman, mais une lente et constante érosion de la confiance au fil du temps qui les a mis de plus en plus mal à l'aise ».

Quelques semaines avant le licenciement, Altman aurait utilisé des tactiques malhonnêtes pour tenter d'écarter Helen Toner, membre du conseil d'administration, au sujet d'un article universitaire qu'elle avait coécrit et qui, selon lui, critiquait l'engagement d'OpenAI en

faveur de la sécurité de l'IA. Dans cet article, Helen Toner, chercheuse en gouvernance de l'IA alignée sur l'altruisme efficace, (EA), loue Anthropic pour avoir évité « le genre de coupes sombres frénétiques que la publication de ChatGPT a semblé susciter ».

Le *New Yorker* a rapporté que « certains des six membres du conseil d'administration ont trouvé Altman manipulateur et complice ». Quelques jours après le licenciement, un chercheur en sécurité de l'IA de DeepMind qui travaillait pour OpenAI a écrit qu'Altman « m'a menti à plusieurs reprises » et qu'il « était trompeur, manipulateur et pire encore avec les autres », une évaluation reprise par un rapport récent de *Time.*

Ce n'était pas la première fois qu'Altman était renvoyé. En 2019, le fondateur de *Y Combinator*, Paul Graham, a retiré Altman de la direction de l'incubateur, craignant qu'il ne donne la priorité à ses propres intérêts plutôt qu'à ceux de l'organisation. Paul Graham a précédemment déclaré : « Sam est extrêmement doué pour devenir puissant. »

L'étrange modèle de gouvernance d'OpenAI a été établi spécifiquement pour empêcher l'influence corruptrice de la recherche du profit, mais comme l'a proclamé à juste titre *The Atlantic,* « l'argent gagne toujours ». Plus d'argent que jamais est consacré à l'amélioration des capacités de l'IA.

En avant toute

Les progrès récents de l'IA sont le fruit de l'aboutissement de plusieurs décennies de tendances : l'augmentation de la puissance de calcul (appelée « *compute*«) et des données utilisées pour former les modèles d'IA, elles-mêmes amplifiées par des améliorations significatives de l'efficacité algorithmique. Depuis 2010, la quantité de calcul utilisée pour former les modèles d'IA a été multipliée *par cent millions*. La plupart des progrès que nous observons aujourd'hui sont le fruit de ce qui était à l'époque un domaine beaucoup plus petit et plus pauvre.

Bien que l'année dernière ait certainement été marquée par un battage médiatique important autour de l'IA, la confluence de ces trois tendances a permis d'obtenir des résultats quantifiables. Le temps nécessaire aux systèmes d'IA pour atteindre des performances humaines dans de nombreuses tâches de référence s'est considérablement réduit au cours de la dernière décennie.

Il est possible, bien sûr, que les gains de capacité de l'IA se heurtent à un mur. Les chercheurs pourraient ne plus avoir de données valables à utiliser. La loi de Moore – l'observation selon laquelle le nombre de transistors sur une puce électronique double tous les deux ans – finira par devenir de l'histoire ancienne. Des événements politiques pourraient perturber les chaînes de fabrication et d'approvisionnement, entraînant une hausse des coûts de calcul. La mise à l'échelle des systèmes pourrait ne plus permettre d'obtenir de meilleures performances.

En réalité, personne ne connaît les véritables limites des approches existantes. Un extrait d'une interview de Yann Le Cun datant de janvier 2022 a refait surface sur Twitter cette année. Yann Le Cun a déclaré : « Je ne pense pas que nous puissions former une machine à être intelligente uniquement à partir de textes, car je pense que la quantité d'informations sur le monde contenues dans les textes est minuscule par rapport à ce que nous avons

besoin de savoir ». Pour illustrer son propos, il donne un exemple : « Je prends un objet, je le pose sur la table et je pousse la table. Il est tout à fait évident pour vous que l'objet est poussé avec la table ». Cependant, avec « un modèle basé sur le texte, si vous entraînez une machine, aussi puissante soit-elle, votre 'GPT-5000' … elle n'apprendra jamais cela ».

Mais si vous donnez cet exemple à ChatGPT-3.5, il crache instantanément la bonne réponse.

Dans une interview publiée quatre jours avant son licenciement, Sam Altman a déclaré : « Jusqu'à ce que nous formions ce modèle [GPT-5], c'est comme un jeu de devinettes amusant pour nous. Nous essayons de nous améliorer, car je pense qu'il est important, du point de vue de la sécurité, de prévoir les capacités. Mais je ne peux pas vous dire exactement ce qu'il va faire et que le GPT-4 n'a pas fait ».

L'histoire est jalonnée de mauvaises prédictions concernant le rythme de l'innovation. Un éditorial du *New York Times* affirmait qu'il faudrait « un million à dix millions d'années » pour mettre au point une machine volante – soixante-neuf jours avant que les frères Wright ne volent pour la première fois. En 1933, Ernest Rutherford, le « père de la physique nucléaire », a rejeté avec assurance la possibilité d'une réaction en chaîne induite par les neutrons, ce qui a incité le physicien Leo Szilard à formuler une hypothèse de travail *dès le lendemain* – une solution qui s'est avérée fondamentale pour la création de la bombe atomique.

Une conclusion qui semble difficile à éviter est que, depuis peu, les personnes les plus aptes à construire des systèmes d'IA pensent que l'IAG (Intelligence Artificielle Générale) est à la fois possible et imminente. Les deux principaux laboratoires d'IA, OpenAI et DeepMind, travaillent peut-être sur l'IAG depuis leur création, à une époque où le fait d'admettre que l'IAG était possible dans un avenir proche pouvait vous faire sortir de la salle en riant (Ilya Sutskever a entonné le chant « Feel the AGI » lors de la fête de fin d'année 2022 d'OpenAI).

Des travailleurs parfaits

Les employeurs utilisent déjà l'IA pour surveiller, contrôler et exploiter les travailleurs. Mais le véritable rêve est d'exclure les humains de la boucle. Après tout, comme l'a écrit Marx, « la machine est un moyen de produire de la plus-value ».

Ajeya Cotra, chercheuse en risques d'IA chez Open Philanthropy (OP), m'a écrit que « le point final logique d'une économie capitaliste ou de marché maximalement efficace » n'impliquerait pas les humains parce que « les humains sont tout simplement des créatures très inefficaces pour gagner de l'argent ». Nous accordons de l'importance à toutes ces émotions « commercialement improductives », écrit-elle, « donc si nous finissons par nous amuser et par aimer le résultat, c'est parce que nous avons commencé avec le pouvoir et que nous avons façonné le système pour qu'il s'adapte aux valeurs humaines ».

OP est une fondation inspirée par EA (Altruisme Efficace) et financée par Dustin Moskovitz, cofondateur de Facebook. C'est le principal bailleur de fonds des organisations de sécurité de l'IA, dont beaucoup sont mentionnées dans cet article. OP a également accordé 30 millions de dollars à OpenAI pour soutenir le travail de sécurité de l'IA deux ans avant que

le laboratoire ne se transforme en une branche à but lucratif en 2019. J'ai précédemment reçu une subvention ponctuelle pour soutenir le travail de publication à New York Focus, une organisation à but non lucratif d'information et d'investigation couvrant la politique new-yorkaise, de la part d'EA Funds, qui reçoit elle-même des fonds d'OP. Après avoir rencontré EA pour la première fois en 2017, j'ai commencé à donner 10 à 20 % de mes revenus à des organisations à but non lucratif de santé mondiale et de lutte contre l'agriculture industrielle, j'ai fait du bénévolat en tant qu'organisateur de groupe local et j'ai travaillé dans une organisation à but non lucratif voisine de lutte contre la pauvreté dans le monde. L'EA a été l'une des premières communautés à s'engager sérieusement dans le risque existentiel de l'IA, mais j'ai regardé les gens de l'IA avec une certaine méfiance, compte tenu de l'incertitude du problème et de l'immense souffrance évitable qui se produit actuellement.

Une IA conforme serait le travailleur dont les capitalistes ne peuvent que rêver : infatigable, motivé et libéré du besoin de faire des pauses toilettes. Les managers, de Frederick Taylor à Jeff Bezos, s'indignent des différentes façons dont les humains ne sont pas optimisés pour le rendement – et, par conséquent, pour les résultats de leur employeur. Même avant l'époque de la gestion scientifique de Taylor, le capitalisme industriel a cherché à rendre les travailleurs plus semblables aux machines avec lesquelles ils travaillent et par lesquelles ils sont de plus en plus remplacés. Comme l'observait avec clairvoyance *Le Manifeste Communiste*, l'utilisation intensive des machines par les capitalistes transforme le travailleur en « un appendice de la machine ».

Mais selon la communauté de la sécurité de l'IA, les mêmes capacités inhumaines qui feraient saliver Bezos font également de l'IAG (Intelligence Artificielle Générale) un danger mortel pour les humains.

Explosion : Le cas de l'extinction

L'argument courant du *x-risk* (risque existentiel) est le suivant : lorsque les systèmes d'IA atteindront un certain seuil, ils seront capables de s'améliorer de manière récursive, ce qui déclenchera une « explosion d'intelligence ». Si un nouveau système d'IA devient suffisamment intelligent – ou simplement plus grand – il sera en mesure de priver définitivement l'humanité de son pouvoir.

Le document d'octobre intitulé « *Managing AI Risks* » (Gérer les risques liés à l'IA) indique ce qui suit :

> « Il n'y a aucune raison fondamentale pour que les progrès de l'IA ralentissent ou s'arrêtent lorsqu'elle atteindra des capacités de niveau humain. . .. Par rapport aux humains, les systèmes d'IA peuvent agir plus rapidement, absorber plus de connaissances et communiquer à une bande passante beaucoup plus large. En outre, ils peuvent être mis à l'échelle pour utiliser d'immenses ressources informatiques et peuvent être reproduits par millions. »

Ces caractéristiques ont déjà permis de développer des capacités surhumaines : Les LLM peuvent « lire » une grande partie de l'internet en quelques mois, et *AlphaFold* de

DeepMind peut réaliser des années de travail de laboratoire humain en quelques jours.

Voici une version stylisée de l'idée que la croissance de la « population » stimule une explosion de l'intelligence : si les systèmes d'IA rivalisent avec les scientifiques humains en matière de recherche et de développement, ils prolifèreront rapidement, ce qui équivaudra à l'arrivée dans l'économie d'un nombre considérable de nouveaux travailleurs hautement productifs. En d'autres termes, si le GPT-7 peut effectuer la plupart des tâches d'un travailleur humain et qu'il ne coûte que quelques dollars pour faire travailler le modèle entraîné sur une journée de travail, chaque instance du modèle sera extrêmement rentable, ce qui déclenchera une boucle de rétroaction positive. Cela pourrait conduire à une « population » virtuelle de milliards ou plus de travailleurs numériques, chacun valant bien plus que le coût de l'énergie nécessaire à son fonctionnement. Iliya Sutskever pense qu'il est probable que « toute la surface de la terre sera recouverte de panneaux solaires et de centres de données ».

Ces travailleurs numériques pourraient être en mesure d'améliorer nos conceptions de l'IA et de créer des systèmes « superintelligents », dont les capacités, selon Alan Turing en 1951, « dépasseront bientôt nos faibles pouvoirs« . Par ailleurs, comme l'affirment certains partisans de la sécurité de l'IA, il n'est pas nécessaire qu'un modèle d'IA soit superintelligent pour constituer une menace existentielle ; il suffirait qu'il y ait suffisamment de copies de ce modèle. Nombre de mes sources ont comparé les entreprises à des super-intelligences, dont les capacités dépassent clairement celles de leurs membres constitutifs.

L'objection la plus courante est la suivante : « Il suffit de la débrancher ». Mais lorsqu'un modèle d'IA sera suffisamment puissant pour menacer l'humanité, il sera probablement la chose la plus précieuse qui existe. Il est peut-être plus facile de « débrancher » la Bourse de New York ou les services Web d'Amazon.

Une super-intelligence paresseuse ne représente peut-être pas un grand risque, et des sceptiques comme Oren Etzioni, PDG de l'*Allen Institute for AI*, Melanie Mitchell, professeure de complexité, et Sarah Myers West, directrice générale de l'*AI Now Institute*, m'ont dit qu'ils et elles n'avaient pas vu de preuves convaincantes que les systèmes d'IA devenaient plus autonomes. Dario Amodei, d'Anthropic, semble convenir que les systèmes actuels ne présentent pas un niveau d'autonomie inquiétant. Cependant, un système totalement passif mais suffisamment puissant, utilisé par un mauvais acteur, suffit à inquiéter des personnes comme Bengio.

En outre, les universitaires et les industriels redoublent d'efforts pour rendre les modèles d'IA plus autonomes. Quelques jours avant son licenciement, Sam Altman a déclaré au *Financial Times* :

> « Nous allons rendre ces agents de plus en plus puissants [...] et les actions deviendront de plus en plus complexes. . .. Je pense que la valeur commerciale qui résultera de la capacité à faire cela dans chaque catégorie est assez importante ».

Qu’est-ce qui se cache derrière le battage médiatique ?

La crainte qui empêche de nombreuses personnes de la communauté *x-risk* (risque existentiel) de dormir n’est pas qu’une IA avancée se « réveille », « devienne diabolique » et décide de tuer tout le monde par méchanceté, mais plutôt qu’elle en vienne à nous considérer comme un obstacle à ses objectifs, quels qu’ils soient. Dans son dernier livre, *Brèves réponses aux grandes questions*, Stephen Hawking a exprimé cette idée en déclarant : « Vous n’êtes probablement pas un fourmilier diabolique qui marche sur les fourmis par méchanceté, mais si vous êtes responsable d’un projet hydroélectrique d’énergie verte et qu’il y a une fourmilière dans la région à inonder, tant pis pour les fourmis ».

Des comportements inattendus et indésirables peuvent résulter d’objectifs simples, qu’il s’agisse du profit ou de la fonction de récompense d’une IA. Dans un marché « libre », la recherche du profit conduit à des monopoles, à des systèmes de marketing multi-niveaux, à l’empoisonnement de l’air et des rivières, et à d’innombrables autres maux.

Il existe de nombreux exemples de systèmes d’IA présentant des comportements surprenants et indésirables. Un programme censé éliminer les erreurs de tri dans une liste a entièrement effacé la liste. Un chercheur a été surpris de constater qu’un modèle d’IA « faisait le mort » pour éviter d’être identifié lors de tests de sécurité.

D’autres encore voient dans ces préoccupations une conspiration des grandes entreprises technologiques. Certaines personnes s’intéressant aux dommages immédiats causés par l’IA soutiennent que l’industrie promeut activement l’idée que ses produits pourraient mettre fin au monde, comme Myers West, de l’*AI Now Institute*, qui déclare qu’elle « considère les récits sur le soi-disant risque existentiel comme un jeu visant à éliminer tout l’air de la pièce, afin de s’assurer qu’il n’y a pas de mouvement significatif dans le moment présent ». Curieusement, Yann Le Cun et Andrew Ng, directeur scientifique de Baidu AI, sont d’accord.

Lorsque je soumets l’idée aux adeptes du *x-risk* (risque existentiel), ils réagissent souvent avec un mélange de confusion et d’exaspération. Ajeya Cotra, de l’OP, m’a répondu : « J’aimerais que le *x-risk* soit moins associé à l’industrie, car je pense qu’il s’agit fondamentalement, sur le fond, d’une croyance très anti-industrielle. . .. Si les entreprises construisent des choses qui vont tous nous tuer, c’est vraiment mauvais, et elles devraient être limitées de manière très stricte par la loi ».

Markus Anderljung, de GovAI, a qualifié la crainte d’une capture réglementaire de « réaction naturelle », mais il a souligné que les politiques qu’il privilégie pourraient bien nuire aux plus grands acteurs du secteur.

Le fait que Sam Altman soit aujourd’hui l’une des personnes les plus associées à l’idée de risque existentiel est une source compréhensible de suspicion, mais son entreprise a fait plus que toute autre pour faire avancer la frontière de l’intelligence artificielle à usage général.

En outre, à mesure qu’OpenAI se rapprochait de la rentabilité et qu’Altman se rapprochait du pouvoir, le PDG a changé de discours en public. Lors d’une séance de questions-réponses en janvier 2023, lorsqu’on lui a demandé quel était son pire scénario pour l’IA, il a

répondu : « C'est l'extinction des feux pour nous tous. » Mais lorsqu'il répond à une question similaire sous serment devant des sénateurs en mai, Altman ne mentionne pas l'extinction. Et, dans ce qui fut peut-être sa dernière interview avant son licenciement, Altman a déclaré : « En fait, je ne pense pas que nous allons tous disparaître. Je pense que tout va bien se passer. Je pense que nous nous dirigeons vers le meilleur des mondes ».

En mai, M. Altman a imploré le Congrès de réglementer l'industrie de l'IA, mais une enquête menée en novembre a révélé que la quasi-société mère d'OpenAI, Microsoft, avait joué un rôle influent dans le lobbying, finalement infructueux, visant à exclure les « modèles de base » tels que ChatGPT de la réglementation prévue par la future loi européenne sur l'IA. Sam Altman a également exercé de nombreuses activités de lobbying dans l'UE, menaçant même de se retirer de la région si les réglementations devenaient trop lourdes (menaces qu'il a rapidement retirées). S'exprimant lors d'un panel de PDG à San Francisco quelques jours avant son éviction, Altman a déclaré que « les modèles actuels sont bons. Nous n'avons pas besoin d'une réglementation lourde ici. Probablement pas pour les deux prochaines générations ».

Le récent décret « radical » du président Joe Biden sur l'IA semble aller dans le même sens : ses exigences en matière de partage d'informations pour les tests de sécurité ne concernent que les modèles plus importants que tous ceux qui ont probablement été formés jusqu'à présent. Myers West a qualifié ce type de « seuils d'échelle » d' »exclusion massive ». Anderljung m'a écrit que la réglementation devrait s'adapter aux capacités et à l'utilisation d'un système, et il a déclaré qu'il « souhaiterait une certaine réglementation des modèles les plus performants et les plus largement utilisés aujourd'hui », mais il pense qu'il sera « beaucoup plus viable politiquement d'imposer des exigences aux systèmes qui n'ont pas encore été développés ».

Inioluwa Deborah Raji estime que si les géants de la technologie « savent qu'ils doivent être le méchant dans une certaine dimension … ils préféreraient que ce soit abstrait et à long terme ». Cela me semble bien plus plausible que l'idée que les grandes entreprises technologiques souhaitent promouvoir l'idée que leurs produits ont une chance raisonnable de *tuer littéralement tout le monde*.

Près de sept cents personnes ont signé la lettre sur l'extinction, dont une majorité d'universitaires. Un seul d'entre eux dirige une société cotée en bourse : Dustin Moskovitz, bailleur de fonds de OP (*Open Philanthropy Project*), qui est également cofondateur et PDG d'Asana, une application de productivité. Il n'y avait aucun employé d'Amazon, d'Apple, d'IBM ou de l'une des principales entreprises de matériel d'IA. Aucun dirigeant de Meta n'a signé.

Si les dirigeants des grandes entreprises technologiques voulaient amplifier le récit de l'extinction, pourquoi n'ont-ils pas ajouté leur nom à la liste ?

Pourquoi construire la « machine de l'apocalypse » ?

Si l'IA sauve effectivement le monde, celui qui l'a créée peut espérer être salué comme un Jules César moderne. Et même si ce n'est pas le cas, le premier à construire » la dernière invention que l'homme doit faire » n'aura pas à craindre d'être oublié par l'histoire – à moins, bien sûr, que l'histoire ne s'arrête brusquement après son invention.

Connor Leahy pense que, si nous continuons sur notre lancée, la fin de l'histoire suivra de peu l'avènement de l'intelligence artificielle. Avec ses cheveux flottants et sa barbichette mal entretenue, il aurait probablement l'air chez lui avec un panneau sandwich portant l'inscription « La fin est proche » – ce qui ne l'a pas empêché d'être invité à s'adresser à la Chambre des Lords britannique ou à CNN. Le PDG de Conjecture, âgé de 28 ans, et cofondateur d'EleutherAI, un influent collectif de logiciels libres, m'a expliqué qu'une grande partie de la motivation pour construire l'IA se résume à ceci : »Oh, vous construisez la machine ultime qui vous rapportera des milliards de dollars et fera de vous le roi-empereur de la Terre ou qui tuera tout le monde ? Oui, c'est comme le rêve masculin. Vous vous dites : « Putain, oui. Je suis le roi du malheur ». Il poursuit : « Je comprends. C'est tout à fait dans l'esthétique de la Silicon Valley ».

Leahy a également fait part d'une chose qui ne surprendra pas les personnes ayant passé beaucoup de temps dans la *Bay Area*ou dans certains coins de l'internet :

> « Il existe des hommes d'affaires et des technologues techno-utopiques, non élus, qui n'ont aucun compte à rendre et qui vivent principalement à San Francisco. Ils sont prêts à risquer votre vie, celle de vos enfants, de vos petits-enfants et de toute l'humanité future, simplement parce qu'ils pourraient avoir la chance de vivre éternellement. »

En mars, le *MIT Technology Review* a rapporté que Sam Altman « dit avoir vidé son compte en banque pour financer deux objectifs : une énergie illimitée et une durée de vie prolongée ».

Compte tenu de tout cela, on pourrait s'attendre à ce que la communauté éthique considère la communauté de la sécurité comme un allié naturel dans une lutte commune pour régner sur les élites technologiques non responsables qui construisent unilatéralement des produits risqués et nocifs. Comme nous l'avons vu précédemment, de nombreux défenseurs de la sécurité ont fait des ouvertures aux éthiciens de l'IA. Il est également rare que les membres de la communauté *x-risk*attaquent publiquement les éthiciens de l'IA (alors que l'inverse n'est pas vrai), mais la réalité est que les partisans de la sécurité ont parfois été difficiles à digérer.

Les éthiciens de l'IA, comme les personnes qu'ils défendent, se sentent souvent marginalisés et coupés du pouvoir réel, menant une lutte acharnée contre les entreprises technologiques qui les considèrent comme un moyen de couvrir leurs arrières plutôt que comme une véritable priorité. L'éviscération des équipes chargées de l'éthique de l'IA dans de nombreuses grandes entreprises technologiques au cours des dernières années (ou des derniers jours) vient étayer ce sentiment. Dans un certain nombre de cas, ces entreprises ont exercé des représailles à l'encontre de lanceurs d'alertes et de militants syndicaux soucieux d'éthique.

Cela n'implique pas nécessairement que ces entreprises accordent une priorité sérieuse au *x -risk*. Le comité d'éthique de Google DeepMind, qui comprenait Larry Page et l'éminent chercheur en risques existentiels Toby Ord, a tenu sa première réunion en 2015, mais n'en a jamais eu de seconde. Un chercheur en IA de Google m'a écrit qu'ils « ne parlaient pas des risques à long terme […] au bureau ». Google se concentre davantage sur la construction de la technologie et sur la sécurité dans le sens de respect de la légalité et de l'absence de caractère offensant ».

Timnit Gebru, ingénieure en logiciel, a codirigé l'équipe éthique d'IA de Google jusqu'à ce qu'elle soit chassée de l'entreprise fin 2020 à la suite d'un différend sur un projet de document – aujourd'hui l'une des publications les plus célèbres sur l'apprentissage automatique. Dans l'article intitulé « perroquets stochastiques », Timnit Gebru et ses coautrices affirment que les LLM (Grands modèles de langage) nuisent à l'environnement, amplifient les préjugés sociaux et utilisent les statistiques pour assembler « au hasard » le langage « sans aucune référence à la signification ».

Timnit Gebru, qui n'est pas une fan de la communauté de la sécurité de l'IA, a appelé à une meilleure protection des lanceurs d'alertes pour les chercheurs en IA, ce qui est également l'une des principales recommandations du Livre Blanc de GovAI. Depuis que Gebru a été évincée de Google, près de 2 700 membres du personnel ont signé une lettre de solidarité, mais Geoff Hinton, alors chez Google, n'en faisait pas partie. Interrogé sur CNN pour savoir pourquoi il ne soutenait pas une collègue lanceuse d'alertes, Hinton a répondu que les critiques de Gebru sur l'IA « étaient des préoccupations assez différentes des miennes » qui « ne sont pas aussi graves sur le plan existentiel que l'idée que ces choses deviennent plus intelligentes que nous et prennent le dessus ».

Inioluwa Deborah Raj m'a expliqué que « la frustration et l'animosité » entre le camp de l'éthique et celui de la sécurité s'expliquent en grande partie par le fait que « l'un des camps a beaucoup plus d'argent et de pouvoir que l'autre », ce qui « lui permet d'imposer son agenda de manière beaucoup plus directe ».

Selon une estimation, le montant de l'argent investi dans les start-ups et les organisations à but non lucratif spécialisées dans la sécurité de l'IA en 2022 a quadruplé depuis 2020, atteignant 144 millions de dollars. Il est difficile de trouver un chiffre équivalent pour la communauté de l'éthique de l'IA. Cependant, la société civile des deux camps est éclipsée par les dépenses de l'industrie. Au cours du seul premier trimestre 2023, *OpenSecrets* a indiqué qu'environ 94 millions de dollars avaient été dépensés pour le lobbying en matière d'IA aux États-Unis. *LobbyControl* estime que les entreprises technologiques ont dépensé 113 millions d'euros cette année pour faire pression sur l'UE, et nous rappelons que des centaines de milliards de dollars sont investis dans l'industrie de l'IA en ce moment même.

L'animosité est peut-être encore plus forte que la différence de pouvoir et d'argent perçue : c'est la ligne de tendance. À la suite de livres très appréciés comme *Algorithmes, la bombe à retardement*, (2016), de la scientifique des données Cathy O'Neil, et de découvertes fracassantes de biais algorithmiques, comme l'article « *Gender Shades* » (Nuances de genre, 2018) de Joy Buolamwini et Timnit Gebru, la perspective de l'éthique de l'IA a attiré l'attention du public et a bénéficié de son soutien.

En 2014, la cause du risque existentiel de l'IA a eu son propre best-seller surprise, *Superintelligence* du philosophe Nick Bostrom, qui soutenait que l'IA au-delà de l'humain pourrait conduire à l'extinction et a reçu les éloges de personnalités telles qu'Elon Musk et Bill Gates. Mais Yudkowsky m'a dit qu'avant ChatGPT, en dehors de certains cercles de la Silicon Valley, le fait de s'intéresser sérieusement à la thèse du livre incitait les gens à vous regarder d'un drôle d'air. Les premiers partisans de la sécurité de l'IA comme Yudkowsky ont occupé la position étrange de maintenir des liens étroits avec la richesse et le pouvoir par l'intermédiaire des techniciens de la *Bay Area*, tout en restant marginalisés dans le discours général.

Dans le monde post-ChatGPT, des lauréats du Prix Turing et des prix Nobel sortent du placard de la sécurité de l'IA et adoptent les arguments popularisés par Yudkowsky, dont la

publication la plus connue est une fanfiction de Harry Potter de plus de 660 000 mots.

Le signe avant-coureur le plus choquant de ce nouveau monde a peut-être été diffusé en novembre, lorsque les animateurs d'un podcast technologique du *New York Times*, *Hard Fork*, ont demandé à la présidente de la*Federal Trade Commission* : « Quel est votre p(doom), Lina Khan ? Quelle est, selon vous, la probabilité que l'IA nous tue tous ? » (Lina Khan a déclaré être « optimiste » et a donné une estimation « basse » de *15 %*).

Il serait facile d'observer toutes les lettres ouvertes et les cycles médiatiques et de penser que la majorité des chercheurs en IA se mobilisent contre le risque existentiel. Mais lorsque j'ai demandé à Yoshua Bengio comment ce risque était perçu aujourd'hui dans la communauté de l'apprentissage automatique, il m'a répondu : « Oh, cela a beaucoup changé. Auparavant, 0,1 % des gens prêtaient attention à la question. Aujourd'hui, c'est peut-être 5 %. »

Probabilités

Comme beaucoup d'autres personnes préoccupées par le risque existentiel de l'IA, le célèbre philosophe de l'esprit David Chalmers a présenté un argument probabiliste au cours de notre conversation : « Il ne s'agit pas d'une situation où il faut être certain à 100 % que nous aurons une IA de niveau humain pour s'en préoccuper. Si la probabilité est de 5 %, il faut s'en préoccuper ».

Ce type de raisonnement statistique est très répandu dans la communauté de l'EA (Altruisme Efficace) et c'est en grande partie ce qui a poussé ses membres à se concentrer sur l'IA en premier lieu. Si vous vous en remettez aux arguments des experts, vous risquez d'être encore plus confus. Mais si vous essayez de faire la moyenne des préoccupations des experts à partir de la poignée d'enquêtes, vous pourriez finir par penser qu'il y a au moins quelques pour cent de chances que l'extinction à cause de l'IA se produise, ce qui pourrait être suffisant pour en faire la chose la plus importante au monde. Si l'on accorde une quelconque valeur à toutes les générations futures qui pourraient exister, l'extinction de l'humanité est catégoriquement pire que les catastrophes auxquelles on peut survivre.

Cependant, dans le débat sur l'IA, les allégations d'arrogance abondent. Des sceptiques comme Melanie Mitchell et Oren Etzioni m'ont dit qu'il n'y avait pas de preuves pour étayer la thèse du risque existentiel, tandis que des croyants comme Bengio et Leahy mettent en avant des gains de capacité surprenants et demandent : « Et si le progrès ne s'arrêtait pas ? Un ami chercheur universitaire en IA a comparé l'avènement de l'IAG (Intelligence Artificielle Générale) au passage au mixeur de l'économie et de la politique mondiales.

Même si, pour une raison ou une autre, l'IAG ne peut qu'égaler et non dépasser l'intelligence humaine, la perspective de partager la terre avec un nombre presque arbitrairement élevé d'agents numériques de niveau humain est terrifiante, surtout lorsqu'ils essaieront probablement de faire gagner de l'argent à quelqu'un.

Il y a beaucoup trop d'idées politiques sur la façon de réduire le risque existentiel lié à l'IA pour pouvoir en parler correctement ici. Mais l'un des messages les plus clairs émanant de la communauté de la sécurité de l'IA est que nous devrions « ralentir ». Les partisans d'une telle décélération espèrent qu'elle donnera aux décideurs politiques et à la société dans

son ensemble une chance de rattraper leur retard et de décider activement de la manière dont une technologie potentiellement transformatrice est développée et déployée.

Coopération internationale

L'objection « mais la Chine ! » est l'une des réponses les plus courantes à tout effort de réglementation de l'IA. Sam Altman, par exemple, a déclaré devant une commission sénatoriale en mai que « nous voulons que l'Amérique prenne la tête » et a reconnu que le risque de ralentir est que « la Chine ou quelqu'un d'autre fasse des progrès plus rapides ».

Markus Anderljung m'a écrit que ce n'était pas une raison suffisante pour ne pas réglementer l'IA.

Dans un article paru en juin dans *Foreign Affairs*, Helen Toner et deux politologues ont indiqué que les chercheurs chinois en IA qu'ils ont interrogés pensaient que les LLM (Grands modèles de langage) chinois avaient au moins deux à trois ans de retard sur les modèles américains de pointe. En outre, les auteurs affirment que, puisque les progrès de l'IA chinoise « reposent en grande partie sur la reproduction et l'adaptation de recherches publiées à l'étranger », un ralentissement unilatéral « décélérerait probablement » également les progrès chinois. La Chine a également agi plus rapidement que tout autre grand pays pour réglementer de manière significative l'IA, comme l'a fait remarquer Jack Clark, dirigeant d'Anthropic.

Selon Eiezer Yudkowsky, « il n'est pas vraiment dans l'intérêt de la Chine de se suicider en même temps que le reste de l'humanité ».

Si l'IA avancée menace réellement le monde entier, une réglementation nationale ne suffira pas. Mais des restrictions nationales fortes pourraient signaler de manière crédible aux autres pays à quel point ils prennent les risques au sérieux. Rumman Chowdhury, éminente éthicienne de l'IA, a appelé à une surveillance mondiale. Bengio estime que nous devons « faire les deux ».

Yudkowsky, sans surprise, a adopté une position maximaliste, me disant que « la bonne direction consiste plutôt à placer tout le matériel d'IA dans un nombre limité de centres de données sous la supervision internationale d'organismes ayant un traité symétrique selon lequel personne – y compris les armées, les gouvernements, la Chine ou la CIA – ne peut faire aucune des choses vraiment horribles, y compris la construction de superintelligences ».

Dans une tribune controversée publiée par *Time* en mars, Yudkowsky a préconisé de « tout arrêter » en instaurant un moratoire international sur les « nouveaux grands entraînements », en brandissant la menace de la force militaire. Étant donné que Eliezer Yudkowsky est convaincu que l'IA avancée serait bien plus dangereuse que n'importe quelle arme nucléaire ou biologique, cette position radicale est tout à fait naturelle.

Les vingt-huit pays présents au récent sommet sur la sécurité de l'IA, dont les États-Unis et la Chine, ont signé la déclaration de Bletchley, qui reconnaît les dommages causés par l'IA et le fait que « des risques substantiels peuvent découler d'une mauvaise utilisation intentionnelle potentielle ou de problèmes de contrôle involontaires liés à l'alignement sur

l'intention humaine ».

Lors du sommet, le gouvernement britannique hôte a chargé Yoshua Bengio de diriger la production du premier rapport sur « l'état de la science » concernant les « capacités et les risques de l'IA d'avant-garde », ce qui constitue une étape importante vers la création d'un organe d'experts permanent tel que le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC).

La coopération entre les États-Unis et la Chine sera impérative pour une coordination internationale significative sur le développement de l'IA. En matière d'IA, les deux pays ne sont pas vraiment en bons termes. Avec les contrôles à l'exportation prévus par la loi CHIPS de 2022, les États-Unis ont tenté de mettre à genoux les capacités d'IA de la Chine, ce qu'un analyste du secteur aurait auparavant considéré comme un « acte de guerre« . Comme *Jacobin* l'a rapporté en mai, certains chercheurs qui travaillent sur le risque existentiel ont probablement joué un rôle dans l'adoption de ces contrôles onéreux. En octobre, les États-Unis ont renforcé les restrictions de la loi CHIPS afin de combler les lacunes.

Toutefois, signe encourageant, Joe Biden et Xi Jinping ont discuté en novembre de la sécurité de l'IA et de l'interdiction de l'IA dans les systèmes d'armes létales. Selon un communiqué de presse de la Maison-Blanche, « les dirigeants ont affirmé la nécessité d'aborder les risques liés aux systèmes d'IA avancés et d'améliorer la sécurité de l'IA dans le cadre de discussions entre les gouvernements américain et chinois ».

Les armes autonomes létales (les robots tueurs) font également l'objet d'un accord relatif dans les débats sur l'IA. Dans son nouveau livre intitulé *Unmasking AI : My Mission to Protect What Is Human in a World of Machines*(*Démasquer l'IA : ma mission pour protéger ce qui est humain dans un monde de machines*), Joy Buolamwini plaide en faveur de la campagne *Stop Killer Robots* (Arrêtez les robots tueurs), faisant ainsi écho à une préoccupation de longue date de nombreux partisans de la sécurité de l'IA. Le *Future of Life Institute*, une organisation de lutte contre le risque existentiel a rassemblé des opposants idéologiques pour signer une lettre ouverte publiée 2016 appelant à l'interdiction des robots tueurs offensifs, dont Bengio, Hinton, Sutton, Etzioni, Le Cun, Musk, Hawking et Noam Chomsky.

Un siège à la table des négociations

Après des années d'inaction, les gouvernements du monde entier s'intéressent enfin à l'IA. Mais en ne s'intéressant pas sérieusement à ce que les futurs systèmes pourraient faire, les socialistes cèdent leur place à la table.

En grande partie à cause des types de personnes qui ont été attirées par l'IA, beaucoup des premiers adoptants sérieux de l'idée du risque existentiel ont décidé de s'engager dans une recherche extrêmement théorique sur la façon de contrôler l'IA avancée ou de créer des entreprises d'IA. Mais pour un autre type de personnes, la réponse à la croyance que l'IA pourrait mettre fin au monde est d'essayer de faire en sorte que les gens *arrêtent de la développer.*

Les promoteurs de l'IA ne cessent de répéter que son développement est inévitable et que

si suffisamment de gens y croient, cela devient vrai. Mais « l'intelligence artificielle n'a rien d'inévitable », écrit l'*AI Now Institute*. Sa directrice générale, Myers West, a fait écho à ces propos en mentionnant que la technologie de reconnaissance faciale semblait inévitable en 2018, mais qu'elle a depuis été interdite dans de nombreux endroits. Et comme le souligne Katja Grace, chercheuse sur le risque existentiel, nous ne devrions pas ressentir le besoin de construire toutes les technologies simplement parce que nous le pouvons.

En outre, de nombreux décideurs politiques regardent les récentes avancées de l'IA et *paniquent*. Le sénateur Mitt Romney est « plus terrifié par l'IA » qu'optimiste, et son collègue Chris Murphy déclare : « Les conséquences de l'externalisation de tant de fonctions humaines vers l'IA sont potentiellement désastreuses ». Les membres du Congrès étatsunien Ted Lieu et Mike Johnson sont littéralement « effrayés » par l'IA. Si certains techniciens sont les seuls à reconnaître que les capacités de l'IA se sont considérablement améliorées et qu'elles pourraient constituer une menace pour l'espèce humaine à l'avenir, les décideurs politiques les écouteront de manière disproportionnée. En mai, la professeure et éthicienne de l'IA Kristian Lum a tweeté : « Il y a un risque existentiel que je suis certaine que les LLM (Grands modèles de langage) posent, et c'est celui de la crédibilité des *FAccT [Fairness, Accountability, and Transparency] / Ethical AI* (Conférences sur l'équité, la responsabilité et la transparence/IA Éthique) si nous continuons à mettre en avant un narratif à base de poudre de perlimpinpin à leur sujet. »

Même si l'idée d'une extinction provoquée par l'IA vous semble plus proche de la science-fiction que de la science tout court, elle pourrait avoir un impact considérable sur la manière dont une technologie transformatrice est développée et sur les valeurs qu'elle représente. En supposant que nous puissions faire faire à une hypothétique IA ce que nous voulons, cela soulève peut-être la question la plus importante à laquelle l'humanité ne sera jamais confrontée : Que devrions-nous *vouloir* qu'elle veuille ?

Lorsque j'ai interrogé David Chalmers à ce sujet, il m'a répondu : « À un moment donné, nous récapitulons toutes les questions de philosophie politique : Quel type de société voulons-nous et apprécions-nous réellement ? ».

L'une des façons d'envisager l'avènement d'une IA de niveau humain est que cela reviendrait à créer la constitution d'un nouveau pays (l« IA constitutionnelle » d'Anthropic prend cette idée au pied de la lettre, et l'entreprise a récemment expérimenté l'incorporation d'une contribution démocratique dans le document fondateur de son modèle). Les gouvernements sont des systèmes complexes qui exercent un pouvoir énorme. Les bases sur lesquelles ils sont établis peuvent influencer la vie de millions de personnes, aujourd'hui et à l'avenir. Les États-uniens vivent sous le joug d'hommes morts qui craignaient tellement le public qu'ils ont mis en place des mesures antidémocratiques qui continuent d'affecter notre système politique plus de deux siècles plus tard.

L'IA est peut-être plus révolutionnaire que toute autre innovation passée. Il s'agit également d'une technologie normative unique, étant donné que nous la construisons pour qu'elle reflète nos préférences. Jack Clark a récemment déclaré à Vox : « Il est vraiment étrange qu'il ne s'agisse pas d'un projet gouvernemental ». David Chalmers m'a dit : « Une fois que les entreprises technologiques essaieront d'intégrer ces objectifs dans les systèmes d'IA, nous devrons leur faire confiance pour résoudre ces questions sociales et politiques très profondes. Je ne suis pas sûr d'y croire. » Il a insisté sur le fait qu'il ne s'agissait pas seulement d'une réflexion technique, mais aussi d'une réflexion sociale et politique.

Faux choix

Il n'est peut-être pas nécessaire d'attendre pour trouver des systèmes superintelligents qui ne donnent pas la priorité à l'humanité. Les agents surhumains optimisent sans pitié pour obtenir une récompense au détriment de tout ce qui peut nous intéresser. Plus l'agent est capable et plus l'optimiseur est impitoyable, plus les résultats sont extrêmes.

Cela vous semble familier ? Si c'est le cas, vous n'êtes pas seul. L'AI Objectives Institute (AOI) considère le capitalisme et l'IA comme des exemples d'optimiseurs mal alignés. Cofondé par Brittney Gallagher, ancienne animatrice d'une émission de radio publique, et Peter Eckersley, « héros de la vie privée », peu avant sa mort inattendue, le laboratoire de recherche examine l'espace entre l'anéantissement et l'utopie, « la poursuite des tendances actuelles à la concentration du pouvoir dans un nombre réduit de mains – suralimentée par les progrès de l'IA – plutôt qu'une rupture brutale avec le présent ». Le président de l'AOI, Deger Turan, m'a dit : « Le risque existentiel est l'incapacité à se coordonner face à un risque ». Il ajoute que « nous devons créer des ponts entre » la sécurité et l'éthique de l'IA.

L'une des idées les plus influentes dans les milieux de l'*x-risk* (risque existentiel) est la malédiction de l'unilatéralisme, un terme qui désigne les situations dans lesquelles un acteur isolé peut ruiner les choses pour l'ensemble du groupe. Par exemple, si un groupe de biologistes découvre un moyen de rendre une maladie plus mortelle, il suffit d'un seul pour le publier. Au cours des dernières décennies, de nombreuses personnes ont été convaincues que l'IA pourrait anéantir l'humanité, mais seuls les plus ambitieux et les plus tolérants au risque d'entre eux ont créé les entreprises qui font aujourd'hui progresser les capacités de l'IA ou, comme l'a récemment déclaré Sam Altman, qui repoussent le « voile de l'ignorance ». Comme l'indique le PDG, nous n'avons aucun moyen de savoir ce qui se trouve au-delà de la limite technologique.

Certains d'entre nous comprennent parfaitement les risques, mais poursuivent quand même leur chemin. Avec l'aide de scientifiques de haut niveau, ExxonMobil a découvert de manière concluante en 1977 que son produit provoquait le réchauffement de la planète. Elle a ensuite menti au public à ce sujet, tout en construisant ses plates-formes pétrolières plus haut.

L'idée que la combustion du carbone puisse réchauffer le climat a été émise pour la première fois à la fin du XIXe siècle, mais le consensus scientifique sur le changement climatique a mis près de cent ans à se former. L'idée que nous pourrions perdre définitivement le contrôle des machines est plus ancienne que l'informatique numérique, mais elle est loin de faire l'objet d'un consensus scientifique. Si les progrès récents de l'IA se poursuivent au même rythme, nous n'aurons peut-être pas des décennies pour former un consensus avant d'agir de manière significative.

Le débat qui se déroule sur la place publique peut vous amener à penser que nous devons choisir entre les inconvénients immédiats de l'IA et ses risques existentiels intrinsèquement spéculatifs. Il y a certainement des compromis à faire qui doivent être examinés avec soin.

Mais si l'on examine les forces matérielles en jeu, le tableau est différent : d'un côté, des entreprises valant des milliards de dollars tentent de rendre les modèles d'IA plus puissants

et plus rentables ; de l'autre, des groupes de la société civile essaient de faire en sorte que l'IA reflète des valeurs qui s'opposent régulièrement à la maximisation des profits.

En bref, c'est le capitalisme contre l'humanité.

*

Garrison Lovely est un écrivain indépendant basé à Brooklyn et animateur du podcast *The Most Interesting People I Know.*

Publié initialement sur https://jacobin.com/2024/01/can-humanity-survive-ai. Traduction par Christian Dubucq pour *Contretemps*.