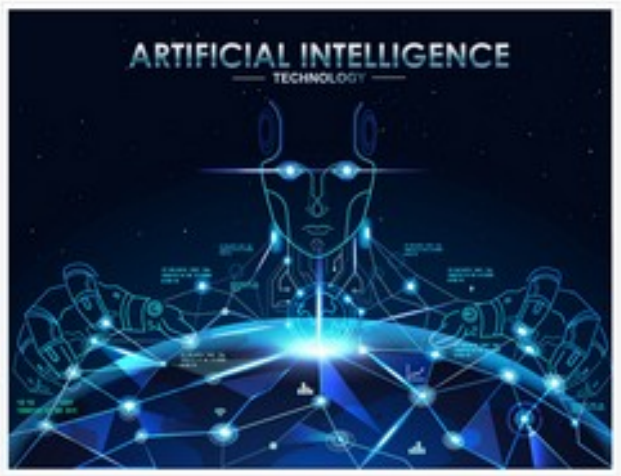
Le Saker Francophone

Le chaos du monde ne naît pas de l'âme des peuples, des races ou des religions, mais de l'insatiable appétit des puissants. Les humbles veillent.



Suspendre les développements en Intelligence artificielle n'est pas suffisant. Il faut tout arrêter

Par Eliezer Yudkowsky * – Le 29 mars 2023 – Source Time Magazine
Publié le par Wayan

Une lettre ouverte publiée aujourd'hui appelle «*tous les laboratoires d'IA à interrompre immédiatement, pour une durée d'au moins six mois, la formation de systèmes d'IA plus puissants que le GPT-4*».

Ce moratoire de six mois est déjà préférable à l'absence de moratoire. J'ai du respect pour tous ceux qui l'ont signé. C'est une amélioration.

Mais je me suis abstenu de le signer parce que je pense que cette lettre sous-estime la gravité de la situation et demande trop peu pour la résoudre.

* * *

La question essentielle n'est pas celle de l'intelligence «*concurrentielle à l'humaine*» (comme le dit la lettre ouverte) ; il s'agit de savoir ce qui se passera une fois que l'IA sera devenue plus intelligente que l'homme. Les seuils clés ne sont peut-être pas évidents, nous ne pouvons certainement pas calculer à l'avance ce qui se passera à ce moment-là, et il semble actuellement concevable qu'un laboratoire de recherche franchisse les limites critiques sans s'en apercevoir.

De nombreux chercheurs spécialisés dans ces questions, dont je fais partie, estiment que le résultat le plus probable de la construction d'une IA surhumainement intelligente, dans des conditions proches des circonstances actuelles, est la mort de tous les habitants de la Terre. Non pas comme dans «*peut-être une chance infime*», mais comme dans «*c'est la chose évidente qui se produirait*». Ce n'est pas que vous ne puissiez pas, en principe, survivre à la création d'un objet beaucoup plus intelligent que vous ; c'est que cela nécessiterait de la précision, de la préparation et de nouvelles connaissances

scientifiques, et probablement pas des systèmes d'intelligence artificielle composés de gigantesques et impénétrables tableaux de nombres fractionnaires.

Sans cette précision et cette préparation, le résultat le plus probable est une IA qui ne fait pas ce que nous voulons et qui ne se soucie pas de nous ni de la vie sensible en général. Ce type d'attention pourrait en principe être insufflé dans une IA, mais nous ne sommes pas prêts et nous ne savons pas comment le faire à l'heure actuelle.

En l'absence de cette bienveillance, nous obtenons «*l'IA ne vous aime pas, elle ne vous hait pas, et vous êtes fait d'atomes qu'elle peut utiliser pour autre chose*».

Le résultat probable de l'humanité face à une intelligence surhumaine opposée est une perte totale. Parmi les métaphores valables, citons «*un enfant de 10 ans essayant de jouer aux échecs contre Stockfish 15*», «*le 11e siècle essayant de lutter contre le 21e siècle*» et «*l'australopithèque essayant de lutter contre l'homo sapiens*».

Pour visualiser une IA surhumaine hostile, n'imaginez pas un penseur sans vie, intelligent comme un livre, qui habiterait sur l'internet et enverrait des courriels mal intentionnés. Imaginez une civilisation extraterrestre entière, pensant à des millions de fois la vitesse humaine, initialement confinée aux ordinateurs, dans un monde de créatures qui sont, de son point de vue, très stupides et très lentes. Une IA suffisamment intelligente ne restera pas longtemps confinée aux ordinateurs. Dans le monde d'aujourd'hui, il est possible d'envoyer par courrier électronique des chaînes d'ADN à des laboratoires qui produiront des protéines à la demande, ce qui permet à une IA initialement confinée à l'internet de construire des formes de vie artificielles ou de passer directement à la fabrication de molécules post-biologiques.

Si quelqu'un construit une IA trop puissante, dans les conditions actuelles, je m'attends à ce que chaque membre de l'espèce humaine et toute vie biologique sur Terre meurent peu de temps après.

Aucun plan n'a été proposé pour nous permettre de faire une telle chose et de survivre. L'intention ouvertement déclarée d'OpenAI est de faire en sorte qu'une future IA fasse notre travail d'alignement de l'IA. Le simple fait d'entendre qu'il s'agit là d'un plan devrait suffire à faire paniquer toute personne sensée. L'autre grand laboratoire d'IA, DeepMind, n'a aucun plan.

Une parenthèse : Ce danger ne dépend pas de la question de savoir si les IA sont ou peuvent être conscientes ; il est intrinsèque à la notion de systèmes cognitifs puissants qui optimisent durement et calculent des résultats qui répondent à des critères suffisamment compliqués. Cela dit, je manquerais à mes devoirs moraux en tant qu'être humain si je ne mentionnais pas que nous n'avons aucune idée de la manière de déterminer si les systèmes d'IA sont conscients d'eux-mêmes – puisque nous n'avons aucune idée de la manière de décoder tout ce qui se passe dans les gigantesques réseaux impénétrables – et que nous pourrions donc, à un moment donné, créer par inadvertance des esprits numériques qui sont réellement conscients, qui devraient avoir des droits et qui ne devraient pas être possédés.

La règle que la plupart des personnes conscientes de ces questions auraient approuvée il y a 50 ans, était que si un système d'IA peut parler couramment et dit qu'il est conscient de lui-même et qu'il exige des droits de l'homme, cela devrait être un frein à la possession de cette IA et à son utilisation au-delà de ce point. Nous avons déjà dépassé cette limite. Je suis d'accord pour dire que les IA actuelles ne font probablement qu'imiter le discours sur la conscience de soi à partir de leurs données d'apprentissage. Mais je pense qu'avec le peu d'informations dont nous disposons sur les mécanismes internes de ces systèmes, nous ne savons pas vraiment ce qu'il en est.

Si tel est notre état d'ignorance pour le ChatGPT, et que le GPT-5 est un pas de géant de la même ampleur que le passage du GPT-3 au GPT-4, je pense que nous ne pourrons plus dire à juste titre «*probablement pas conscient de lui-même*» si nous laissons les gens fabriquer des GPT-5. Ce sera simplement «*je ne sais pas ; personne ne sait*.» Si vous ne pouvez pas être sûr de créer une IA consciente d'elle-même, c'est alarmant, non seulement en raison des implications morales de la partie «*consciente d'elle-même*», mais aussi parce qu'être incertain signifie que vous n'avez aucune idée de ce que vous faites et que c'est dangereux et que vous devriez arrêter.

* * *

Le 7 février, Satya Nadella, PDG de Microsoft, déclarait publiquement que le nouveau Bing obligerait Google à «*montrer qu'il sait danser. Je veux que les gens sachent que nous les avons fait danser*», a-t-il déclaré.

Ce n'est pas ainsi que le PDG de Microsoft devrait parler dans un monde sain. Cela montre l'écart considérable entre le sérieux avec lequel nous prenons le problème et le sérieux avec lequel nous aurions du le prendre il y a 30 ans.

Nous n'allons pas combler ce fossé en six mois.

Il a fallu plus de 60 ans entre le moment où la notion d'intelligence artificielle a été proposée et étudiée pour la première fois et le moment où nous avons atteint les capacités actuelles. Résoudre le problème de la sécurité d'une intelligence surhumaine – non pas la sécurité parfaite, mais la sécurité au sens de «*ne pas tuer littéralement tout le monde*» – pourrait très raisonnablement prendre au moins la moitié de ce temps. Et le problème avec l'intelligence surhumaine, c'est que si l'on se trompe sur le premier coup, on ne peut pas apprendre de ses erreurs, parce qu'on est mort. L'humanité n'apprendra pas de ses erreurs et ne recommencera pas, comme pour d'autres défis que nous avons relevés au cours de notre histoire, parce que nous serons tous morts.

Essayer de réussir quelque chose au premier essai vraiment critique est une tâche extraordinaire, tant en science qu'en ingénierie. Nous ne disposons pas de l'approche nécessaire pour y parvenir avec succès. Si nous appliquions au domaine naissant de l'intelligence artificielle générale les mêmes normes de rigueur technique que celles qui s'appliquent à un pont destiné à supporter quelques milliers de voitures, tout le domaine serait fermé demain.

Nous ne sommes pas prêts. Nous ne sommes pas en mesure de nous préparer dans un délai raisonnable. Il n'y a pas de plan. Les progrès en matière de capacités d'IA sont très, très en avance sur les progrès en matière d'alignement de l'IA ou même sur les progrès en matière de compréhension de ce qui se passe à l'intérieur de ces systèmes. Si nous y parvenons, nous allons tous mourir.

De nombreux chercheurs travaillant sur ces systèmes pensent que nous nous dirigeons vers une catastrophe, et ils sont plus nombreux à oser le dire en privé qu'en public ; mais ils pensent qu'ils ne peuvent pas arrêter unilatéralement cette plongée en avant, que les autres continueront même s'ils quittent personnellement leur emploi. Alors, ils se disent tous qu'ils feraient mieux de continuer. C'est une situation stupide et une façon indigne pour la Terre de mourir, et le reste de l'humanité devrait intervenir à ce stade et aider l'industrie à résoudre son problème d'action collective.

Certains de mes amis m'ont récemment rapporté que lorsque des personnes extérieures à l'industrie de l'IA entendent parler pour la première fois du risque d'extinction lié à l'intelligence artificielle générale, leur réaction est la suivante : «*Peut-être que nous ne devrions pas construire d'IAG, alors*.»

Cette réaction m'a donné un petit éclair d'espoir, car elle est plus simple, plus sensée et franchement plus saine que celle que j'ai entendue au cours des 20 dernières années, au cours desquelles j'ai essayé

d’amener les gens de l’industrie à prendre les choses au sérieux. Quiconque parle avec autant de bon sens mérite d’entendre à quel point la situation est grave, et non de se faire dire qu’un moratoire de six mois va régler le problème.

Le 16 mars, ma partenaire m’a envoyé ce courriel. (Elle m’a ensuite autorisé à le reproduire ici).

«*Nina a perdu une dent ! Comme le font habituellement les enfants, et non par négligence! Le fait de voir GPT4 pulvériser ces tests standardisés le jour même où Nina atteignait une étape importante de son enfance a provoqué une vague d’émotions qui m’a fait perdre pied pendant une minute. Tout va trop vite. Je crains que le fait de partager cela n’accentue ton propre chagrin, mais je préfère que tu le saches plutôt que chacun d’entre nous souffre seul.*»

Lorsque la conversation entre initiés porte sur la douleur de voir sa fille perdre sa première dent et de penser qu’elle n’aura pas la chance de grandir, je pense que nous avons dépassé le stade des échecs politiques sur un moratoire de six mois.

S’il existait un plan de survie pour la Terre, si seulement nous adoptions un moratoire de six mois, je soutiendrais ce plan. Ce plan n’existe pas.

Voici ce qu’il faudrait faire :

Le moratoire sur les nouveaux grands circuits d’entraînement doit être indéfini et mondial. Il ne peut y avoir aucune exception, y compris pour les gouvernements ou les armées. Si la politique commence aux États-Unis, la Chine doit comprendre que les États-Unis ne cherchent pas à obtenir un avantage, mais qu’ils essaient plutôt d’empêcher l’utilisation d’une technologie horriblement dangereuse qui ne peut avoir de véritable propriétaire et qui tuera tout le monde aux États-Unis, en Chine et sur Terre. Si je disposais d’une liberté infinie pour rédiger des lois, je pourrais prévoir une seule exception pour les IA formées uniquement pour résoudre des problèmes de biologie et de biotechnologie, non formées sur des textes provenant d’Internet, et pas au point de commencer à parler ou à planifier ; mais si cela compliquait un tant soit peu la question, je rejetterais immédiatement cette proposition et je dirais qu’il faut tout arrêter.

Arrêter toutes les grandes grappes de GPU (les grandes fermes d’ordinateurs où les IA les plus puissantes sont affinées). Arrêter tous les grands cycles d’entraînement. Fixer un plafond à la puissance de calcul que chacun est autorisé à utiliser pour entraîner un système d’IA, et le réduire au cours des prochaines années pour compenser l’apparition d’algorithmes d’entraînement plus efficaces. Aucune exception pour les gouvernements et les armées. Il convient de conclure immédiatement des accords multinationaux afin d’empêcher que les activités interdites ne soient transférées ailleurs. Suivre toutes les ventes de GPU. Si les services de renseignement indiquent qu’un pays non signataire de l’accord construit une grappe de GPU, il faut moins craindre un conflit armé entre les nations que la violation du moratoire ; il faut être prêt à détruire un centre de données hors-la-loi par une frappe aérienne.

N’envisagez rien comme un conflit entre des intérêts nationaux et dites clairement que quiconque parle de course à l’armement est un imbécile. Le fait que nous vivions ou mourions tous ensemble n’est pas une politique mais un fait naturel. Expliquez explicitement dans la diplomatie internationale que la prévention des scénarios d’extinction de l’IA est considérée comme une priorité par rapport à la prévention d’un échange nucléaire complet, et que les pays nucléaires alliés sont prêts à courir un certain risque d’échange nucléaire si c’est ce qu’il faut pour réduire le risque de grandes séries d’entraînement de l’IA.

C’est le genre de changement de politique qui nous amènerait, mon partenaire et moi, à nous serrer l’un contre l’autre et à nous dire qu’un miracle s’est produit et qu’il y a maintenant une chance que Nina vive. Les personnes saines d’esprit qui entendent parler de cette question pour la première fois et qui se

disent raisonnablement «*peut-être ne devrions-nous pas*» méritent d'entendre, honnêtement, ce qu'il faudrait faire pour que cela se produise. Et lorsque la demande politique est aussi importante, la seule façon de la faire passer est que les décideurs politiques réalisent que s'ils continuent à faire ce qui est politiquement facile, cela signifie que leurs propres enfants vont aussi mourir.

Nous ne sommes pas prêts. Nous ne sommes pas en passe d'être significativement plus prêts dans un avenir prévisible. Si nous poursuivons dans cette voie, tout le monde mourra, y compris des enfants qui n'ont pas choisi cette voie et qui n'ont rien fait de mal.

Il faut tout arrêter.

* * *

*** Eliezer Yudkowsky** est un théoricien étasunien qui dirige les recherches de l'Institut de recherche sur l'intelligence des machines. Il travaille sur l'alignement de l'intelligence artificielle générale depuis 2001 et est largement considéré comme l'un des fondateurs de ce domaine.

Traduit par Wayan, relu par Hervé, pour le Saker Francophone.

Ce contenu a été publié dans Comprendre la chute du Système, Intelligence artificielle par Wayan.