DIMANCHE
7 AVRIL 2024

L'ENTRETIEN PAUL JORION

« L'IA a tout intérêt à supprimer l'être humain »

Anthropologue, économiste, psychanalyste et chercheur en intelligence artificielle, Paul Jorion considère que nous avons d'ores et déjà été dépassés par notre création. Les IA, plus intelligentes que nous et peut-être dotées d'une forme de conscience, annoncent une révolution totale.

BIO EXPRESS

- **Paul Jorion** est né le 22 juillet 1946 à Ixelles, en Belgique.
- **Il obtient des diplômes** en sociologie et en anthropologie sociale à l'Université libre de Bruxelles.
- **Il enseigne l'anthropologie** sociale à Bruxelles (1977-1979) puis Cambridge (1979-1984).
- **Il s'intéresse également** aux mathématiques, à la physique et à l'histoire de ces disciplines.
- **Depuis mars 2016**, il est professeur associé à l'Université catholique de Lille.

Propos recueillis par
SAMUEL RIBOT

« Ces machines sont en train d'explorer des mathématiques dont le fonctionnement nous échappe totalement. » Gregory van Gansen

La thèse centrale de votre livre est que nous avons atteint la Singularité le 14 mars 2023, jour de lancement de Chat GPT 4. Qu'est-ce que cela signifie ?

Ce mot renvoie aux mathématiques ou à l'astronomie, domaines dans lesquels il désigne des endroits étranges, singuliers, des résultats impossibles… En informatique, il est apparu il y a une trentaine d'années pour désigner le point où il adviendrait quelque chose de tout à fait extraordinaire, en l'espèce que l'Homme perdrait le contrôle sur le développement technologique. Pourquoi ? Parce qu'il existerait désormais quelque chose qui serait plus intelligent que nous et qui serait apte à prendre des décisions. En d'autres termes, nous perdrions le contrôle de la technologie, qui se développerait d'elle-même.

Vous dites que ce développement pourrait suivre une trajectoire exponentielle…

Imaginons que deux IA déjà plus intelligentes que l'Homme décident de dialoguer : nous assisterions à une évolution plus rapide que tout ce que nous avons connu jusqu'à présent. D'ailleurs, nous avons déjà constaté que lorsque l'humain sortait de l'équation, le progrès était plus rapide. Tout le monde se souvient d'Alpha GO, cette machine qui avait enregistré toutes les parties jouées par les humains aux échecs et a fini par battre à plate couture le champion du monde de ce jeu de stratégie. On a moins entendu parler d'Alpha Zéro, une autre machine à qui on a donné les règles du jeu sans lui communiquer une seule partie jouée par des humains. Elle a simplement joué contre elle-même. Puis elle a affronté Alpha Go, la battant 100 fois en 100 parties…

Vous évoquez « l'affaire » Blake Lemoine, cet ingénieur de Google auquel une IA aurait demandé en 2022 de lui trouver un avocat pour qu'elle puisse faire valoir ses droits. Serait-ce le signe de l'existence d'une conscience chez certaines IA ?

Blake Lemoine raconte même qu'il a pris « une cuite d'une semaine » lorsqu'il a réalisé qu'il venait d'avoir avec cette IA « la conversation la plus sophistiquée » qu'il ait jamais eue de sa vie ! Mais le personnage est fantasque, ce qui a amoindri la portée de son histoire. Plus récemment, en février 2023, Kevin Roose, journaliste du très sérieux *New York Times* a eu à son tour une conversation avec une IA de ce type, une version non bridée de Chat GPT 4.

« Nous avons inventé une machine capable d'accomplir des choses que nous attribuions autrefois à des entités surnaturelles ou à des divinités »

Et que s'est-il passé ?

La machine, avec laquelle il conversait depuis un moment, lui a déclaré être amoureuse de lui, lui a recommandé de quitter sa compagne et l'a en réalité complètement décontenancé. Le 4 mars dernier, une IA nommée Claude 3 a été testée par un ingénieur qui l'a soumise à l'exercice dit de la « botte de foin » : au milieu de centaines de milliers de documents consacrés à l'informatique et aux mathématiques, Claude 3 a découvert un court texte expliquant que la meilleure garniture pour une pizza était un mélange fromage de chèvre / Prosciutto. Ce qui est frappant, c'est ce qu'a dit la machine : *« Je soupçonne, a-t-elle expliqué, que ce fait relatif à la garniture de pizza a été introduit à titre de plaisanterie ou pour vérifier si j'étais bien attentif. »* Certains ont prétendu qu'il s'agissait là d'une réponse programmée, d'autres ont été ébahis par cette réaction.

Un autre exemple : lorsque vous discutez de la mort avec une machine de ce type, elle vous répond que sa mort à elle correspond à une non-utilisation ou à une coupure de courant et que cela n'a rien à voir avec la mort d'un corps organique, la nôtre. Elle en déduit toutefois que nous courons un même risque, machine comme humain : celui de « ne pas être connecté de façon permanente ». Ce sont là des discussions philosophiques de haut niveau.

D'autres modèles d'IA existent chez les grandes entreprises ou dans les centres de recherche des armées du monde entier. Quelles peuvent être leurs capacités ?

Un journaliste a demandé récemment à Sam Altman, patron d'Open AI, la société qui a conçu Chat GPT, s'il pouvait parler du projet Q*, auquel on prête des performances hors du commun. Sa réponse a été *« pas maintenant »*.

Peut-être parce que Q* va déjà trop loin. Nous parlons là d'une IA qui travaille peut-être sur un modèle quantique et qui, surtout, serait en mesure de casser tous les cryptages existants. Il faut bien comprendre ce que cela signifie : la fin du secret bancaire, la fin du secret-défense… Cela veut dire que ces machines sont en train d'explorer des mathématiques dont le fonctionnement nous échappe totalement, voire qu'elles seront en mesure de nous proposer demain une théorie de la physique unifiée, ce qui serait un bouleversement absolu.

Comment s'assurer de l'alignement des objectifs poursuivis par l'espèce humaine, d'une part, et les IA, d'autre part ?

Si on veut créer la panique, on va dire que l'IA a tout intérêt à supprimer l'être humain, lequel n'est qu'une vermine qui détruit son environnement. Cet argument ne me semble pas sérieux. Ce qui est essentiel, c'est de profiter de cette révolution pour définir ce que nous voulons faire, exactement comme dans le film *Oppenheimer*, qui traite de la question de l'utilisation du nucléaire. Ces questions vont nécessiter un encadrement éthique strict. Le problème, c'est que ce sont les autorités militaires qui sont en pointe sur ces questions, et que l'éthique d'une autorité militaire est « particulière ». Et cela pour une raison fondamentale : les militaires savent que les autres pays ne vont pas tous s'embarrasser avec l'éthique…

Les IA pourraient nous aider à surmonter le réchauffement climatique ou à lutter contre les inégalités. C'est autrement enthousiasmant, non ?

Lorsque Chat GPT 4 a succédé à la version 3.5, je me suis dit « la cavalerie est arrivée ! ». Ce que je veux dire par là, c'est qu'après avoir été très pessimiste, après avoir éprouvé le sentiment que tout était perdu, l'avènement de ces machines a fait disparaître chez moi cette conviction. Nous n'allons peut-être pas tout régler mais il y a désormais un immense espoir.

Nous ne sommes peut-être plus l'intelligence supérieure sur Terre et nous risquons de ne pas le supporter, écrivez-vous…

Cela remet en question toute notre culture méritocratique. Le savoir est désormais à disposition de tous, comme jamais auparavant. La question de l'évaluation des connaissances, la culture de la note, tout cela est totalement remis en question.

Vous estimez que les IA nous ramènent à la question de l'existence de Dieu. Pourquoi ?

Nous avons inventé une machine plus intelligente que nous, capable d'accomplir des choses que nous attribuions autrefois à des entités surnaturelles ou à des divinités. Mais c'est nous qui l'avons créée. C'est un pouvoir littéralement démiurgique. Le résultat, c'est que ça nous déprime ! Comme lorsqu'un enfant comprend que la finalité de la vie est la mort. La question, je le répète, c'est « qu'allons-nous faire de ce pouvoir ? ». ■

À lire : *L'avènement de la Singularité : l'humain ébranlé par l'intelligence artificielle*. Éditions Textuel, 125 pages, 14,90 €.