TECHNOLOGIE

L'IA peut-elle vraiment anéantir l'humanité ?

Une IA sans conscience, sans haine et sans volonté propre pourrait-elle malgré tout devenir dangereuse, au point de faire peser un risque existentiel ? Trois scientifiques démontent les fantasmes. Sans nier les (vrais) risques.

PHILIPPE LALOUX

Faut-il craindre l'IA? Pour la majorité des scientifiques et chercheurs, le défi est l'alignement entre les objectifs donnés à l'IA et les valeurs humaines.
© HELOÏSE CHIGARD-TRADORI.

Et à la fin, l'intelligence artificielle avait tout dévasté, y compris le genre humain. Tout ça pour quoi ? Des trombones… En 2003, le philosophe suédois Nick Bostrom avait imaginé cette fable, glaçante : le « maximiseur de trombones » (*Paperclip maximizer*). Il imaginait confier à une superintelligence une mission dérisoire : fabriquer le plus de trombones possible, quitte à transformer la Terre et puis l'espace en usines. En quelques mois, la planète, y compris l'homme, avait été méthodiquement recyclée pour fournir les atomes nécessaires à la fabrication d'attaches métalliques.

Cette métaphore extrême, que Bostrom lui-même se gardait bien de présenter comme une prédiction, sert encore aujourd'hui d'illustration d'un des problèmes majeurs de l'IA : son alignement avec nos valeurs. Et le risque qu'elle pourrait faire peser sur l'humanité. Le futurologue cherchait ainsi à démontrer par l'absurde qu'une IA n'a pas besoin d'être malveillante ou d'éprouver de la haine pour détruire l'humanité. Sa stricte logique mathématique suffit. Sans intention.

Déclarations alarmistes

L'immense majorité des scientifiques et des philosophes rappelle utilement que les IA n'ont pas de pulsions biologiques, ni d'instinct de survie, de désir de domination ou de soif de pouvoir. « Une machine n'a évidemment aucune intention, arrêtons avec l'anthropomorphisation », s'irrite Hugues Bersini, professeur d'informatique à l'ULB. « C'est aussi aberrant que de dire que la Terre affiche l'intention de tourner autour du Soleil. » Ce mythe a pourtant resurgi dans l'actualité à la faveur de déclarations alarmistes. Comme celle de cet ingénieur démissionnaire d'Anthropic, Jacob Coxon, qui s'affolait que « les personnes qui développent l'IA croient sincèrement qu'elle pourrait tous nous tuer d'ici la fin de la décennie ».

Dans son sillage, Evan Hubinger, responsable de la recherche sur l'alignement chez Anthropic, avance même un chiffre : plus de 10 % de risque que l'IA « tue tous les humains » au cours de la prochaine décennie. Et voilà que le mythe du trombone refait surface, servant même d'alibi aux patrons d'Open AI et d'Anthropic pour réclamer une « pause » dans le développement de l'IA. Le temps de formaliser des standards de sécurité. Lesquels standards, a déjà prévenu l'autorité en charge de la concurrence aux Etats-Unis (FTC), ne devront pas servir de « barrière à l'entrée » pour de nouveaux concurrents.

« Hypocrite »

D'où vient ce chiffre de « plus de 10 % » ? Mystère. « *Bullshit* », s'emporte même Benoît Frénay, professeur d'informatique à l'UNamur, évoquant ce « prétendu incident de sécurité » dévoilé par OpenAI. Quelque 700 agents se seraient ainsi « échappés » pour mener une cyberattaque sur l'entreprise Huging Face. « C'est hypocrite. On est clairement dans la manipulation parce qu'ils ont mis un système effectivement dans des conditions telles qu'il ne pouvait pas trouver la solution. Et on s'étonne : "Oh, tiens, l'IA a trouvé un autre moyen de le faire." Mais c'était l'objectif. Ils crient au loup mais en même temps ils développent et vendent les pièges à loups. »

C'est en quelque sorte le trombone qui cache la forêt. Un algorithme auquel on demande de maximiser les clics sur un réseau social finira par propager des *fake news* haineuses parce que c'est mathématiquement le plus efficace. Cela rend néanmoins la question existentielle de l'IA très concrète : comment contrôler une machine très puissante dont on aurait mal défini l'objectif ? Ce sera sans doute le défi technique le plus difficile et le plus crucial du XXIe siècle. « D'un point de vue purement théorique, oui, un agent peut mener une attaque. Mais en réalité, même si ce qu'ils font nous paraît très impressionnant, il ne fait jamais qu'écrire du code informatique », tempère Katrien Beuls, chargée de cours à l'UNamur, spécialiste du langage computationnel.

Souci d'alignement

Surgit alors cette question, sans doute sans réponse, de l'alignement entre l'homme et l'IA : faire en sorte qu'elle parle et agisse comme nous, mais avec les mêmes valeurs. « C'est extrêmement difficile, voire impossible, de résumer nos valeurs dans une formule mathématique », tranche la chercheuse. Comment traduire en lignes de code des concepts aussi flous que « la dignité humaine », « la liberté » ou « le bon sens » ? « Ce qui est troublant, pour nous humains, c'est qu'ils communiquent en langage naturel (c'est pour cela aussi que l'on pense qu'ils sont "conscients") et que l'on ne comprend pas exactement ce qui se passe dans les couches plus profondes car ce sont des opérations mathématiques très avancées, avec énormément de dimensions. »

Une machine dépourvue d'intention propre pourrait-elle malgré tout devenir incontrôlable ? « Ce qui est synonyme de perte de contrôle, c'est une perte de compréhension, une perte de maîtrise des stratégies mises en place pour résoudre les problèmes qu'on lui a confiés », relativise Hugues Bersini, parfois stupéfait par les capacités d'abstraction et de conceptualisation de ces « robots ».

« Dès qu'un système d'une certaine complexité est mis en œuvre par les humains, on ne maîtrise pas tout », poursuit l'expert. « Il y a 20 ou 25 ans, les informaticiens possédaient toutes les clés algorithmiques. On avait des stratégies de tests où on soumettait les logiciels à tous les scénarios. Ce qui n'empêche pas les bugs. Parfois, on se rend compte qu'il y avait un facteur qu'on avait oublié. Et ça conduit à la catastrophe. Tchernobyl, c'est ça : c'est un moment où des apprentis sorciers qui pensaient tout maîtriser se sont dit : "On va voir ce qu'il se passe dans ce cas-là." Et ça a été la catastrophe, même si ça partait plutôt d'une bonne intention. »

Avec l'IA, ajoute-t-il, le problème change d'échelle : les développeurs cherchent précisément à laisser les machines explorer des solutions originales à des problèmes complexes. « On espère pouvoir comprendre les stratégies qu'elles mettent en œuvre, mais ça devient de plus en plus compliqué. » Pour autant, Hugues Bersini refuse le scénario d'une machine devenue souveraine : « Dans tous les cas, on garde encore le contrôle. »

Le fantasme de l'IA générale

Comment ? C'est tout bête : la prise de courant. « Internet repose sur plusieurs couches, la première étant la couche physique, les infrastructures, soit le câble, la carte réseau… Sans elle, c'est terminé. Les humains peuvent toujours débrancher les machines ou la couper d'internet. L'idée que l'ordinateur se met lui-même sur internet comme OpenAI le prétend dans une espèce de scénario science-fiction, c'est du fantasme. »

La parabole du trombone l'évoquait bien : intelligence et morale ne vont pas nécessairement de pair. L'intelligence décrit la capacité à atteindre un objectif ; elle ne détermine pas quel objectif mérite d'être poursuivi. Quel serait alors l'intérêt pour des humains de coder explicitement des fonctions de destruction de l'humanité ? « Je ne le perçois vraiment pas », appuie Benoît Frénay. « Quel sens cela aurait de développer cette fameuse "IA générale", capable de faire à peu près tout ce dont nous, humains, sommes capables ? C'est absurde. Si tel est le cas, on a un gros problème, parce que tout ce que l'on vient d'expliquer tombe. On ne sait pas ce que l'IA va faire, à quoi elle va servir, dans quelles conditions elle va être utilisée. On ne sait rien des problèmes qui peuvent survenir, ni comment les résoudre. Et donc, de fait, là on a la recette pour une catastrophe. Mais l'IA générale reste un fantasme. A ce stade, on reste dans l'idée que l'IA est d'abord un outil. »

Quel est donc, à ce stade, le vrai risque de l'IA ? « Le vrai risque, c'est la concentration de pouvoir de ses acteurs », avance le chercheur. « C'est un petit club qui fonctionne en circuit fermé, des entreprises qui disposent de moyens absolument absurdes et d'une puissance absurde. Et clairement, on a bien vu qu'on ne pouvait pas leur faire confiance. Il y a trop d'argent en jeu. Ces entreprises ne sont pas nécessairement malveillantes mais ont une éthique très discutable. »

S'inquiéter des trombones occulterait en somme les dangers actuels, mesurables et bien réels de l'IA : désinformation, vol de données privées, biais algorithmiques, épuisement des ressources. Et aussi, cet embryon de duopole qui se dessine entre OpenAI et Anthropic, y compris leur pouvoir de définir les normes éthiques de l'IA.