Gaspard Koenig, *philosophe et essayiste*

Elon Musk, ainsi que des milliers de chercheurs et penseurs d'envergure mondiale - Yoshua Bengio, Stuart Russel, Yuval Noah Harari, pour ne citer qu'eux - ont rédigé une pétition exigeant un moratoire sur le développement de l'intelligence artificielle (IA) au nom de la survie de la civilisation humaine. L'autorité numérique italienne (l'équivalent de notre CNIL) a semblé exaucer leur souhait en interdisant ChatGPT. De quoi faut-il avoir peur ?

Les pétitionnaires redoutent surtout l'émergence d'une IA « générale » ou « dure », c'est-à-dire dotée d'une forme de conscience qui lui permette de s'affranchir de ses créateurs et, à terme, de détruire toute vie sur terre. Ils s'appuient sur les travaux de Nick Bostrom ou Max Tegmark (cités en référence) sur la superintelligence.

C'est un vieux fantasme qui fait depuis longtemps les délices de la science-fiction. Or il faut rappeler que Bostrom et Tegmark sont des astrophysiciens amateurs de scénarios extrêmes, et dont la définition de l'intelligence se résume à une capacité de calcul.

En chair et en os

Pour les réfuter (et se détendre un peu), je recommande la lecture d'un neuroscientifique comme Antonio Damasio, qui met en lumière le rôle que joue notre organisme biologique dans notre fonctionnement cognitif à travers des phénomènes comme l'homéostasie. Si nous pouvons donner du sens et prendre des décisions, c'est que nous sommes en chair et en os.

Privée de finalité, une superintelligence ne pourrait « vouloir » ni le mal ni le bien. Damasio a réhabilité Spinoza contre Descartes. Au fond, Bostrom et Tegmark sont les nouveaux dualistes qui croient en la toute-puissance de l'esprit : des métaphysiciens !

Si les craintes de Musk et de sa bande me paraissent ainsi exagérées, celles de la CNIL italienne demeurent en revanche trop modestes. Certes, ChatGPT exploite nos données personnelles, mais ni plus ni moins que la vaste majorité des plateformes.

Le problème est inhérent à l'économie numérique aujourd'hui (je plaide, pour ma part, en faveur d'une propriété privée sur les données personnelles). On voit mal comment un tel argument permettrait de condamner ChatGPT sans débrancher l'Internet mondial.

Notre rapport à la vérité

La question posée par les IA génératives est plutôt celle du rapport à la vérité. J'ai déjà mentionné dans ces colonnes la tragique disparition des sources. Il est vrai que

ChatGPT, ou désormais Bing, peut citer ses sources à la demande. Mais ce sont des sources (probables), non les sources (exactes).

Des chercheurs comme Laurence Devillers ont récemment attiré l'attention sur ce point. Par construction, la « boîte noire » de l'apprentissage automatique interdit en effet de décomposer analytiquement le chemin suivi par le robot.

Le résultat produit par ChatGPT n'est donc ni démontrable, ni réfutable, ni explicable. C'est ce qui fait dire à Henry Kissinger (99 ans !), coauteur d'un remarquable article sur le sujet avec Eric Schmidt et un spécialiste de l'IA, que ChatGPT et consorts tournent le dos à notre conception de la connaissance, héritée des Lumières.

Les IA génératives ne s'appuient pas sur des vérités consolidées mais sur des « ambiguïtés cumulatives ». J'ai pu le constater à mon détriment en demandant à ChatGPT : « Qui est Gaspard Koenig ? » Le résultat n'est pas entièrement faux mais constamment approximatif et, bien sûr, impossible à dénoncer. Voilà ce qui le rend si dangereux.

Démêler le vrai du faux

Il est incontestable que ces IA, comme nous le rappellent tous les jours les bons apôtres de la tech, peuvent constituer des outils précieux, et que des esprits avertis sauront trier le bon grain de l'ivraie. Mais comment ignorer, quand on connaît la manière dont fleurit le complotisme sur les réseaux sociaux, que nous nous trouvons face à une bombe nucléaire cognitive ?

ChatGPT compte déjà 100 millions d'utilisateurs actifs. Combien d'entre eux s'interrogent sur l'origine et la valeur d'une information dont la mise en forme est conçue pour faire illusion ? On peut s'amuser des montages si réalistes obtenus avec Midjourney, une IA générative produisant des images. Sauf que bientôt, plus personne ne pourra démêler le vrai du faux.

Les menaces qui pèsent sur notre civilisation consistent donc moins en l'extermination de la vie par une IA malveillante qu'en la désintégration de l'idée de vérité. Comme si la connexion désormais universelle entre les êtres humains finissait paradoxalement par détruire la possibilité même d'un savoir commun.

Faut-il intervenir ? Evidemment, et vite. Mais les gouvernements restent prisonniers de la logique de l'ère industrielle (les « gains de productivité » !) et plus éloignés que jamais d'une capacité de coordination mondiale. Deux carences majeures qui rappellent hélas l'impasse de la question écologique.