Les Echos 29.03.2022 Lettre ouverte des 1100 personnalités

« Les systèmes d'IA dotés d'une intelligence capables de concurrencer celle de l'homme posent de graves risques pour la société et l'humanité, comme le montrent des recherches approfondies et reconnus par les meilleurs laboratoires d'IA. Comme l'affirment les « principes d'IA d'Asilomar », largement reconnus, **l'IA avancée pourrait représenter un changement majeur dans l'histoire de la vie sur Terre,** et devrait être planifiée pour, et gérée avec, un soin et des ressources proportionnés.

Malheureusement, ce niveau de planification et de gestion est absent, alors que ces derniers mois ont vu les laboratoires d'IA lancés dans une course effrénée pour développer et déployer des systèmes numériques toujours plus puissants que personne - pas même leurs créateurs - ne peut comprendre, prédire ou contrôler de manière fiable.

Les systèmes d'IA contemporains deviennent désormais capables de concurrencer l'homme dans des tâches de portée générale et nous devons nous poser des questions : devrions-nous laisser les machines inonder nos canaux d'information de propagande et de contrevérités ? Devrions-nous automatiser tous les emplois, y compris ceux qui sont épanouissants ?

Devrions-nous développer des systèmes non humains qui pourraient éventuellement nous dépasser en nombre et en intelligence, nous rendre obsolètes et nous remplacer ? Devrions-nous risquer de perdre le contrôle de notre civilisation ? De telles décisions ne doivent pas être laissées à des leaders technologiques non élus. Des systèmes d'IA puissants ne devraient être développés qu'une fois acquise la certitude que leurs effets sont positifs et leurs risques gérables.

Cette certitude doit être solidement étayée, et renforcée au fur et à mesure de la progression des effets potentiels de ces systèmes. La récente déclaration d' OpenAI concernant l'intelligence artificielle indique que, « à un certain moment, il deviendra essentiel de procéder à une expertise indépendante avant de commencer à entraîner de futurs systèmes, et d'accepter de limiter la croissance de la puissance de calcul employée pour créer de nouveaux modèles dans les travaux les plus en pointe. » Nous sommes d'accord. Et ce moment, c'est maintenant.

Par conséquent, nous appelons tous les laboratoires d'IA à suspendre immédiatement l'entraînement des systèmes d'IA plus puissants que GPT-4 pour au moins six mois. Cette pause doit être publique, vérifiable, et inclure tous les acteurs clés. Si une telle pause ne peut pas être décrétée rapidement, les gouvernements devraient intervenir et décréter un moratoire.

Les laboratoires d'IA et les experts indépendants devraient profiter de cette pause pour développer et mettre en œuvre conjointement un ensemble de protocoles partagés de sécurité pour la conception et le développement d'IA rigoureusement audités et supervisés par des experts externes indépendants.

Ces protocoles devraient garantir que les systèmes qui y adhèrent sont sûrs au-delà de tout doute raisonnable. Cela ne signifie pas une pause dans le développement de l'IA en général, mais plutôt la

descente d'un cran dans la course dangereuse vers d'imprévisibles boîtes noires, toujours plus grandes, développant de nouvelles capacités.

La recherche et le développement de l'IA devraient être recentrés sur la fabrication de systèmes puissants et à la pointe de la technologie, plus précis, plus sûrs, interprétables, transparents, robustes, alignés, dignes de confiance et loyaux. En parallèle, les développeurs d'IA doivent travailler avec les décideurs politiques pour accélérer considérablement le développement de systèmes robustes de gouvernance de l'IA.

Ces dispositifs devraient au moins comprendre de nouvelles autorités de réglementation compétentes dédiées à l'IA, la surveillance et le suivi des systèmes d'IA hautement performants et des grosses capacités de calcul, [la connaissance de] l'origine et la trame des systèmes pour aider à distinguer le réel du synthétique et détecter les failles des modèles ; un écosystème robuste d'audit et de certification ; une responsabilité pour les dommages causés par l' IA ; un financement public solide pour la recherche technique sur la sécurité de l'IA ; et des institutions bien dotées en ressources pour faire face aux disruptions économiques et politiques majeures (en particulier pour la démocratie) que l'IA entraînera.

L'humanité peut profiter d'un avenir florissant avec l'IA. Ayant réussi à créer de puissants systèmes d'IA, nous pouvons maintenant profiter d'un « été de l'IA » dans lequel nous récoltons les fruits, concevons ces systèmes pour le bénéfice de tous et donnons à la société une chance de s'adapter.

La société a fait une pause pour d'autres technologies susceptibles d'avoir des effets potentiellement catastrophiques pour elle. Nous pouvons en faire de même ici. Profitons d'un long été d'IA, au lieu de nous précipiter sans préparation dans la chute. »

Texte paru sur le site du Future of Life Institute signée par plus de mille personnalités dont **Elon Musk** (PDG de SpaceX, Tesla et Twitter), **Steve Wozniak** (cofondateur d'Apple), **Yuval Noah Harari** (essayiste et professeur à l'université hébraïque de Jérusalem).