https://divergences.be/spip.php?article3694

Interrompre les expériences géantes en matière d'IA

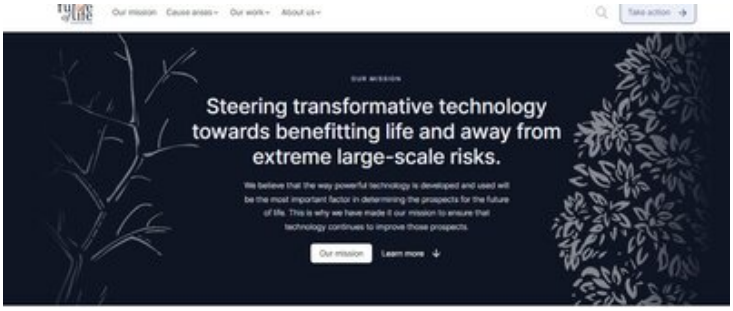
Lettre ouverte

- Aujourd'hui - 2023 - Juin -

Date de mise en ligne : mardi 30 mai 2023

Nous demandons à tous les laboratoires d'IA d'interrompre immédiatement, pour une durée d'au moins six mois, la formation de systèmes d'IA plus puissants que le GPT-4.

Origine



Les systèmes d'IA dotés d'une intelligence compétitive avec celle de l'homme peuvent présenter des risques profonds pour la société et l'humanité, comme le montrent des recherches approfondies[1] et comme le reconnaissent les principaux laboratoires d'IA[2]. Comme l'indiquent les principes d'IA d'Asilomar, largement approuvés, l'IA avancée pourrait représenter un changement profond dans l'histoire de la vie sur Terre, et devrait être planifiée et gérée avec l'attention et les ressources nécessaires. Malheureusement, ce niveau de planification et de gestion n'existe pas, même si les derniers mois ont vu les laboratoires d'IA s'enfermer dans une course incontrôlée pour développer et déployer des esprits numériques toujours plus puissants que personne - pas même leurs créateurs - ne peut comprendre, prédire ou contrôler de manière fiable.

nous demandons à tous les laboratoires d'IA d'interrompre immédiatement.

Les systèmes d'IA contemporains deviennent aujourd'hui compétitifs sur le plan humain pour des tâches générales[3], et nous devons nous interroger : Devons-nous laisser les machines inonder nos canaux d'information de propagande et de mensonges ? Devrions-nous automatiser tous les emplois, y compris ceux qui sont gratifiants ? Devons-nous développer des esprits non humains qui pourraient un jour être plus nombreux, plus intelligents, plus obsolètes et nous remplacer ? Devons-nous risquer de perdre le contrôle de notre civilisation ? Ces décisions ne doivent pas être déléguées à des leaders technologiques non élus. Les systèmes d'IA puissants ne doivent être développés que lorsque nous sommes certains que leurs effets seront positifs et que leurs risques seront gérables. Cette confiance doit être bien justifiée et augmenter avec l'ampleur des effets potentiels d'un système. La récente déclaration de l'OpenAI concernant l'intelligence artificielle générale indique qu'"à un moment donné, il pourrait être important d'obtenir un examen indépendant avant de commencer à former les futurs systèmes, et pour les efforts les plus avancés d'accepter de limiter le taux de croissance du calcul utilisé pour créer de nouveaux modèles". Nous sommes d'accord. C'est maintenant qu'il faut agir.

C'est pourquoi nous demandons à tous les laboratoires d'IA d'interrompre immédiatement, pendant au moins six mois, la formation de systèmes d'IA plus puissants que le GPT-4. Cette pause devrait être publique et vérifiable, et inclure tous les acteurs clés. Si une telle pause ne peut être mise en place rapidement, les gouvernements devraient intervenir et instituer un moratoire.

Les laboratoires d'IA et les experts indépendants devraient profiter de cette pause pour élaborer et mettre en œuvre conjointement un ensemble de protocoles de sécurité communs pour la conception et le développement de l'IA

avancée, rigoureusement contrôlés 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[4], ce qui ne signifie pas une pause dans le développement de l'IA en général, mais simplement un recul par rapport à la course dangereuse vers des modèles de boîte noire toujours plus grands et imprévisibles, dotés de capacités émergentes.

La recherche et le développement dans le domaine de l'IA devraient être recentrés sur l'amélioration de la précision, de la sécurité, de l'interprétabilité, de la transparence, de la robustesse, de l'alignement, de la fiabilité et de la loyauté des systèmes puissants et modernes d'aujourd'hui.

Parallèlement, les développeurs d'IA doivent collaborer avec les décideurs politiques pour accélérer considérablement le développement de systèmes robustes de gouvernance de l'IA. Ceux-ci devraient au minimum inclure : de nouvelles autorités réglementaires compétentes dédiées à l'IA ; la surveillance et le suivi des systèmes d'IA hautement performants et des grands pools de capacité de calcul ; des systèmes de provenance et de filigrane pour aider à distinguer le réel du synthétique et pour suivre les fuites de modèles ; un écosystème robuste d'audit et de certification ; la responsabilité pour les dommages causés par l'IA ; un financement public robuste pour la recherche technique sur la sécurité de l'IA ; et des institutions dotées de ressources suffisantes pour faire face aux perturbations économiques et politiques dramatiques (en particulier pour la démocratie) que l'IA provoquera.

L'humanité peut jouir d'un avenir florissant grâce à l'IA. Ayant réussi à créer des systèmes d'IA puissants, nous pouvons maintenant profiter d'un "été de l'IA" au cours duquel nous récolterons les fruits de nos efforts, concevrons ces systèmes pour le plus grand bénéfice de tous et donnerons à la société une chance de s'adapter. La société a mis en pause d'autres technologies aux effets potentiellement catastrophiques[5], et nous pouvons faire de même ici. Profitons d'un long été de l'IA et ne nous précipitons pas sans préparation vers l'automne.

Traduit avec www.DeepL.com/Translator (version gratuite)

