Q&A avec Sayash Kapoor

(ingénieur logiciel de l'université de Princeton)

l'impact réel de l'IA

Sayash Kapoor à propos des « faux miracles de l'IA » et plus encore

Sayash Kapoor (doctorant, Université de Princeton)

C. J. Abate (Elektor)

Dans cet entretien, Sayash Kapoor, ingénieur logiciel de l'université de Princeton et ancien employé de Facebook, évoque l'impact sociétal de l'IA et les risques réels qu'elle pose aujourd'hui, tels que les déplacements de main-d'œuvre et la diffusion d'informations erronées. Il explique également comment l'IA peut améliorer la prise de décision humaine et partage les enseignements de son livre *AI Snake Oil*, qui remet en question de nombreuses affirmations exagérées concernant l'intelligence artificielle.

Sayash Kapoor : Avant de fréquenter l'université de Princeton, j'ai travaillé chez Facebook en tant qu'ingénieur logiciel, où j'ai développé l'IA pour la modération de contenu.

C. J. : Vos recherches à l'université de Princeton portent sur l'impact sociétal de l'intelligence artificielle. Quand et comment avez-vous commencé à vous intéresser à ce sujet ?
Sayash : Lorsque j'étais chez Facebook, j'ai pu constater, de visu, l'impact de l'IA sur la société. Je me suis particulièrement intéressé à la manière dont l'IA peut causer des dommages dans le monde réel, mais aussi à la manière dont elle peut être utilisée pour améliorer les résultats sociétaux. J'ai développé un intérêt particulier pour l'influence des politiques publiques sur l'IA, en voyant comment des régulations comme le RGPD guidaient le fonctionnement des grandes entreprises telles que Facebook.

C. J. : L'IA représente-t-elle un risque existentiel pressant qui devrait préoccuper tous les ingénieurs et les étudiants en EE/ECE ? Ou s'agit-il simplement d'un outil pratique qui, bien que nécessitant une réflexion approfondie, est en fin de compte une force positive ?
Sayash : Au cours des dernières années, nous avons entendu de nombreuses affirmations selon lesquelles l'IA allait anéantir l'humanité. Dans notre livre *AI Snake Oil*, nous consacrons un chapitre entier à cet argument. Nous examinons les nombreuses failles sous-jacentes aux affirmations selon lesquelles l'IA mettra fin à l'humanité et, sur cette base, nous concluons que, même si nous devons prendre au sérieux les risques liés à l'IA, les discussions actuelles sur les menaces existentielles de l'IA surestiment largement l'impact réel de l'IA. Dans le même temps, je ne veux pas minimiser les risques bien réels que l'IA pose aujourd'hui, y compris le déplacement de la main-d'œuvre, la dépendance excessive à l'égard d'informations erronées et les menaces pour la sécurité.

C. J. : Comment envisagez-vous l'avenir des systèmes pilotés par l'IA en termes d'amélioration de la prise de décision humaine dans des tâches complexes ? Quels sont les principaux défis à relever pour créer des modèles d'IA non seulement performants, mais aussi conformes aux valeurs et aux attentes humaines dans diverses applications du monde réel ?

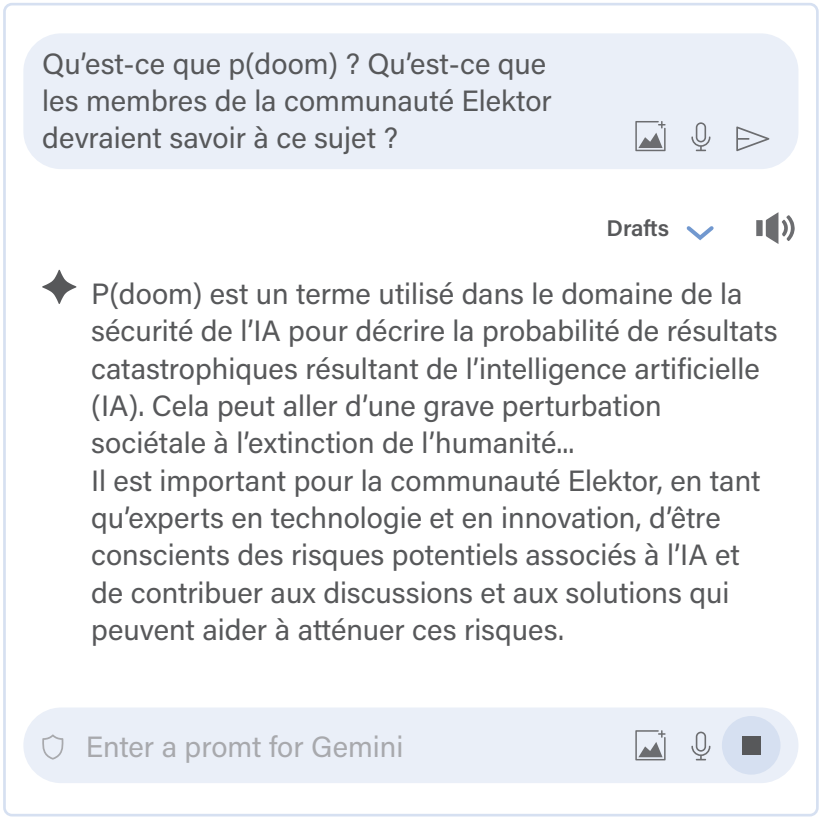
Sayash : L'IA est déjà utilisée pour résoudre de nombreuses tâches complexes. L'un des aspects intéressants de l'histoire de l'IA est que dès qu'une tâche peut être résolue de manière fiable, nous cessons de l'appeler IA. Par exemple, il y a encore quelques décennies, la vérification orthographique était considérée comme un problème difficile à résoudre. Aujourd'hui, des outils comme la correction automatique de l'orthographe se sont estompés en arrière-plan et sont devenus partie intégrante de la vie de tous les jours.

J'espère voir davantage d'outils de ce type fonctionner de manière suffisamment fiable pour être relégués à l'arrière-plan. Nous voyons déjà des outils d'IA capables de modifier une grande partie du travail intellectuel. Ils sont utilisés pour la découverte automatisée de médicaments et comme assistants de codage. Le plus grand défi pour rendre ces outils utiles à un grand nombre de personnes est peut-être d'accroître leur fiabilité, car les systèmes d'IA générative d'aujourd'hui ont un comportement aléatoire et les utilisateurs ne disposent pas vraiment de références pour travailler avec eux. Par exemple, les chatbots peuvent souvent « halluciner », c'est-à-dire fabriquer des réponses incorrectes aux questions des utilisateurs. Si nous parvenons à réduire les erreurs et les hallucinations, l'IA générative pourra être beaucoup plus utile pour les applications conséquentes de l'IA.

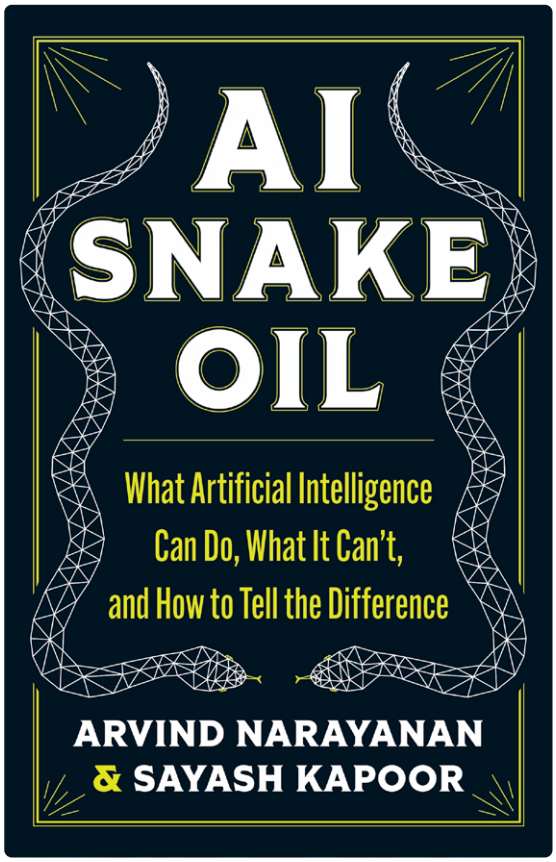
C. J. : Avec Arvind Narayanan, vous avez écrit un livre intitulé AI Snake Oil. Qu'est-ce qui vous a poussé à écrire ce livre et qu'est-ce qui en fait un ouvrage incontournable ?

Sayash : L'une des plus grandes sources de confusion au sujet de l'IA aujourd'hui est le fait que l'IA est un terme générique. Il est utilisé pour désigner de nombreux types de technologies distinctes qui n'ont pas grand-chose à voir les unes avec les autres. Dans le livre *AI Snake Oil*, nous faisons la distinction entre les différents types d'IA pour souligner où l'IA fonctionne bien, quels types d'IA n'ont pas beaucoup progressé et comment nous pouvons nous attendre à ce que l'avenir de ces types d'IA se concrétise.

Nous expliquons comment l'IA générative, utilisée pour générer du texte, des images, du son, etc., a fait d'énormes progrès au cours de la dernière décennie - bien qu'il y ait certains risques que nous anticipons, tels que le déplacement de la main-d'œuvre et la dépendance excessive à l'égard de chatbots incorrects. D'autre part, l'IA prédictive est utilisée pour prédire l'avenir des individus et, sur cette base, prendre des décisions conséquentes à leur sujet - par exemple, si un accusé doit être libéré avant son procès ou si un candidat à un emploi doit être invité à un entretien.



> *Si nous parvenons à réduire les erreurs et les hallucinations, l'IA générative pourra être beaucoup plus utile pour les applications conséquentes de l'IA.*



Source : Sayash Kapoor

Nos recherches nous ont permis de constater que l'IA prédictive est loin de fonctionner aussi bien que le prétendent ses concepteurs, et qu'elle ne s'est pas améliorée au cours des dernières décennies.
Pour tous ceux qui cherchent à utiliser l'IA dans leur vie ou qui prennent des décisions concernant l'achat d'IA pour leurs institutions, nous espérons fournir les connaissances de base sur ce qui fait que l'IA fonctionne, où elle échoue, et comment faire la différence.

C. J. : Le terme « snake oil » implique une tromperie. Pour que les choses soient claires : qui sont les trompeurs ? Les programmeurs et les développeurs ? Les entreprises qui exploitent les solutions d'IA ? Les consommateurs ? Sommes-nous tous en train de nous tromper sur l'IA ?
Sayash : Nous discutons des nombreuses sources d'engouement pour l'IA : Les chercheurs en IA, les entreprises et les journalistes qui traitent de l'IA. Lorsque les chercheurs qui travaillent sur l'IA ne parviennent pas à évaluer correctement leurs modèles, cela conduit à des affirmations exagérées sur la façon dont l'IA peut fonctionner dans le monde réel. Dans nos travaux antérieurs, nous avons mis en évidence une crise de reproductibilité dans la recherche scientifique qui utilise l'apprentissage automatique.
Les entreprises font également du battage médiatique lorsqu'elles vendent des produits qui ne fonctionnent pas aussi bien qu'elles le prétendent. Ces dernières années, nous avons vu de nombreuses affirmations concernant des produits d'IA qui ne fonctionnent pas - et souvent ne peuvent pas fonctionner - et pourtant les entreprises sont prêtes à faire des affirmations exagérées pour vendre leurs produits. Dans certains cas, les entreprises vendent des produits commercialisés comme étant de l'IA alors qu'en réalité, c'est un humain qui prend les décisions en coulisses.
Enfin, l'engouement pour l'IA est également dû aux journalistes qui donnent une image erronée du fonctionnement de l'IA. Nous avons recueilli une série de pièges dans le journalisme sur l'IA et constaté que les articles sur l'IA, même dans les médias respectés, souffrent de ces pièges. Par exemple, les articles utilisent des images de robots pour parler d'applications de l'IA qui n'ont rien à voir avec la robotique, ce qui donne aux lecteurs une idée trompeuse des progrès de l'IA.

C. J. : Parlez à nos lecteurs de p(doom). Qu'est-ce que c'est ? S'agit-il d'une mesure utile ?
Sayash : P(doom) ou la probabilité de la catastrophe due à l'IA est une mesure utilisée dans la communauté de la sécurité de l'IA pour articuler le risque existentiel de l'IA - la probabilité que l'IA nous anéantisse tous. Elle est devenue populaire comme moyen de quantifier à quel point quelqu'un pense que le risque existentiel de l'IA est élevé. En règle générale, les estimations de probabilité sont justifiées de trois manières : inductive (basée sur des données passées), déductive (basée sur des théories ou modèles empiriquement vérifiés sur le monde) ou subjective (basée sur des suppositions de prévisionnistes).
Malheureusement, aucune des méthodes ci-dessus ne fonctionne pour justifier p(doom). Nous ne pouvons pas faire d'estimations inductives parce qu'il n'existe pas de « classe de référence » d'événements similaires à partir desquels nous pourrions extrapoler le risque existentiel de l'IA. De même, nous ne disposons pas de théories et de modèles du monde établis qui nous permettent d'anticiper le risque existentiel de l'IA. Nous nous contentons donc d'estimations subjectives, qui sont par nature spéculatives et qui surestiment le risque d'événements peu probables.
Ce qui est inquiétant, c'est que les estimations de p(doom) sont prises au sérieux par les décideurs politiques. Contrairement à la plupart des autres données qu'ils utilisent pour élaborer leurs politiques,

Si nous parvenons à réduire les erreurs et les hallucinations, l'IA générative pourra être beaucoup plus utile pour les applications conséquentes de l'IA.

les estimations de p(doom) sont intrinsèquement spéculatives. Si les particuliers et les entreprises peuvent, bien entendu, utiliser toutes les informations dont ils disposent pour prendre des décisions, les gouvernements et les décideurs politiques doivent appliquer des normes plus strictes en matière de preuves.

C. J. : En septembre 2023, En septembre 2023, vous avez contribué à organiser un atelier sur les modèles fondamentaux responsables et ouverts. À l'époque, quelle était votre position sur la question de savoir si l'IA devait être ouverte ou fermée ? Avez-vous la même opinion aujourd'hui ?
Sayash : Lorsque nous avons organisé l'atelier, je voulais mieux comprendre les arguments concernant les risques liés aux modèles fondamentaux ouverts. L'un des principaux résultats de l'atelier a été un document de recherche dans lequel nous nous sommes penchés sur cette question. Nous avons constaté que de nombreuses allégations concernant les effets néfastes des modèles d'IA diffusés ouvertement n'étaient pas vraiment justifiées.
Par exemple, au cours de l'été 2023, un groupe de chercheurs du MIT a publié des articles affirmant que les modèles fondamentaux ouverts pouvaient aider les attaquants à créer des armes biologiques. Mais nous avons également constaté que les mêmes informations disponibles dans ces modèles fondamentaux ouverts étaient également disponibles sur Wikipédia. Dans l'article, nous avons introduit le concept de « risque marginal » afin de préciser si le risque provient des modèles fondamentaux ouverts ou s'il existe indépendamment de la publication des modèles.
Il s'agit essentiellement de comparer le risque des modèles ouverts à la fois avec celui des technologies existantes ainsi qu'avec celui des modèles fondamentaux fermés. Depuis, des recherches ont montré que le risque marginal des modèles fondamentaux ouverts est faible — notamment dans des domaines comme la création d'armes biologiques.

C. J. : Y a-t-il quelque chose que vous savez aujourd'hui et que vous ne saviez pas il y a un an qui vous effraie à propos de l'IA et de son potentiel ? Et inversement, y a-t-il quelque chose que vous savez aujourd'hui et que vous ne saviez pas il y a un an qui vous enthousiasme à propos de l'IA et de son potentiel ?
Sayash : Je suis enthousiasmé par le potentiel de l'IA pour les travailleurs du savoir. Je pense que la plupart des travailleurs du savoir peuvent utiliser l'IA de manière utile. Bien que des problèmes tels que le manque de fiabilité et les hallucinations doivent être résolus, l'IA générative représente en fin de compte un grand pas en avant dans nos capacités technologiques. En même temps, je suis préoccupé par l'utilisation croissante de l'IA générative pour des deepfakes non consentis, notamment des nudes. Nous avons vu des cas de personnes — principalement des femmes — être ciblées par des deepfakes générés par IA.

C. J. : Pensez-vous que les départements EE/ECE prennent la sécurité de l'IA au sérieux ? Ou diriez-vous qu'il y a un manque de sensibilisation et d'éducation à ce sujet ?
Sayash : Nous avons constaté un grand intérêt pour la sécurité de l'IA dans tous les départements. Certains des plus grands spécialistes de la sécurité de l'IA à Princeton, comme le professeur Prateek Mittal, font partie du département ECE. Je suis sûr qu'à l'avenir, l'intérêt sera encore plus grand.

240555-04

Questions ou commentaires ?
Contactez Elektor (redaction@elektor.com).