Démystifier les conneries sur l'IA – Une interview

Originally published on The Markup

Démystifier le buzz autour de l'IA

Un entretien avec Arvind Narayanan

par JULIA ANGWIN

Si vous avez parcouru tout le battage médiatique sur ChatGPT le dernier robot conversationnel qui repose sur l'intelligence artificielle, vous pouvez avoir quelque raison de croire que la fin du monde est proche.

Le chat « intelligent » de l'IA a enflammé l'imagination du public pour sa capacité à générer instantanément des poèmes, des essais, sa capacité à imiter divers styles d'écrits, et à réussir à des examens d'écoles de droit et de commerce.

Les enseignants s'inquiètent de la tricherie possible de leurs étudiants (des écoles publiques de New York City l'ont déjà interdit). Les rédacteurs se demandent si cela ne va pas faire disparaître leur travail (BuzzFeed et CNET ont déjà utilisé l'IA pour créer des contenus). Le journal The Atlantic a déclaré que cela pourrait « déstabiliser les professions de cadres supérieurs ». L'investisseur en capital-risque Paul Kedrosky l'a qualifié de « bombe nucléaire de poche » et blâmé ses concepteurs pour l'avoir lancé dans une société qui n'y est pas prête.

Même le PDG de l'entreprise qui a lancé ChatGPT, Sam Altman, a déclaré aux médias que le pire scénario pour l'IA pourrait signifier « notre extinction finale ».

Cependant pour d'autres ce buzz est démesuré. Le principal scientifique chargé de l'IA chez Meta's AI, Yann LeCun, a déclaré à des journalistes que ChatGPT n'a « rien de révolutionnaire ». Le professeur de langage informatique de l'université de Washington Emily Bender précise que « la croyance en un programme informatique omniscient vient de la science-fiction et devrait y rester ».

Alors, jusqu'à quel point devrions-nous nous inquiéter ? Pour recueillir un avis autorisé, je me suis adressée au professeur d'informatique de Princeton Arvind Narayanan, qui est en train de co-rédiger un livre sur « Le charlatanisme de l'IA ». En 2019, Narayanan a fait une conférence au MIT intitulée « Comment identifier le charlatanisme del'IA » qui exposait une classification des IA en fonction de leur validité ou non. À sa grande surprise, son obscure conférence universitaire est devenue virale, et ses diapos ont été téléchargées plusieurs dizaines de milliers de fois ; ses messages sur twitter qui ont suivi ont reçu plus de deux millions de vues.

Narayanan s'est alors associé à l'un de ses étudiants, Sayash Kapoor, pour développer dans un livre la classification des IA. L'année dernière, leur duo a publié une liste de 18 pièges courants dans lesquels tombent régulièrement les journalistes qui couvrent le sujet des IA. Presque en haut de la liste : « illustrer des articles sur l'IA avec de chouettes images de robots ». La raison : donner une image anthropomorphique des IA implique de façon fallacieuse qu'elles ont le potentiel d'agir dans le monde réel.

Narayanan est également le co-auteur d'un manuel sur l'équité et l'apprentissage machine et dirige le projet *Web Transparency and Accountability* de l'université de Princeton pour contrôler comment les entreprises collectent et utilisent les informations personnelles. Il a reçu de la Maison-Blanche le Presidential Early Career Award for Scientists and Engineers [N. de T. : une distinction honorifique pour les scientifiques et ingénieurs qui entament brillamment leur carrière].

Voici notre échange, édité par souci de clarté et brièveté.

Angwin : vous avez qualifié ChatGPT de « générateur de conneries ». Pouvez-vous expliquer ce que vous voulez dire ?

Narayanan : Sayash Kapoor et moi-même l'appelons générateur de conneries et nous ne sommes pas les seuls à le qualifier ainsi. Pas au sens strict mais dans un sens précis. Ce que nous voulons dire, c'est qu'il est entraîné pour produire du texte vraisemblable. Il est très bon pour être persuasif, mais n'est pas entraîné pour produire des énoncés vrais ; s'il génère souvent des énoncés vrais, c'est un effet collatéral du fait qu'il doit être plausible et persuasif, mais ce n'est pas son but.

Cela rejoint vraiment ce que le philosophe Harry Frankfurt a appelé du bullshit, c'est-à-dire du langage qui a pour objet de persuader sans égards pour le critère de vérité. Ceux qui débitent du *bullshit* se moquent de savoir si ce qu'ils disent est vrai ; ils ont en tête certains objectifs. Tant qu'ils persuadent, ces objectifs sont atteints. Et en effet, c'est ce que fait ChatGPT. Il tente de persuader, et n'a aucun moyen de savoir à coup sûr si ses énoncés sont vrais ou non.

Angwin : Qu'est-ce qui vous inquiète le plus avec ChatGPT ?

Narayanan : il existe des cas très clairs et dangereux de mésinformation dont nous devons nous inquiéter. Par exemple si des personnes l'utilisent comme outil d'apprentissage et accidentellement apprennent des informations erronées, ou si des étudiants rédigent des essais en utilisant ChatGPT quand ils ont un devoir maison à faire. J'ai appris récemment que le CNET a depuis plusieurs mois maintenant utilisé des outils d'IA générative pour écrire des articles. Même s'ils prétendent que des éditeurs humains ont vérifié rigoureusement les affirmations de ces textes, il est apparu que ce n'était pas le cas. Le CNET a publié des articles écrits par une IA sans en informer correctement, c'est le cas pour 75 articles, et plusieurs d'entre eux se sont avérés contenir des erreurs qu'un rédacteur humain n'aurait très probablement jamais commises. Ce n'était pas dans une mauvaise intention, mais c'est le genre de danger dont nous devons nous préoccuper davantage quand des personnes se tournent vers l'IA en raison des contraintes pratiques qu'elles affrontent. Ajoutez à cela le fait que l'outil ne dispose pas d'une notion claire de la vérité, et vous avez la recette du désastre.

Angwin : Vous avez développé une classification des l'IA dans laquelle vous décrivez différents types de technologies qui répondent au terme générique de « IA ». Pouvez-vous nous dire où se situe ChatGPT dans cette taxonomie ?

Narayanan : ChatGPT appartient à la catégorie des IA génératives. Au plan technologique, elle est assez comparable aux modèles de conversion de texte en image, comme DALL-E [qui crée des images en fonction des instructions textuelles d'un utilisateur]. Ils sont liés aux IA utilisées pour les tâches de perception. Ce type d'IA utilise ce que l'on appelle des modèles d'apprentissage profond. Il y a environ dix ans, les technologies d'identification par ordinateur ont commencé à devenir performantes pour distinguer un chat d'un chien, ce que les humains peuvent faire très facilement.

Ce qui a changé au cours des cinq dernières années, c'est que, grâce à une nouvelle technologie qu'on appelle des transformateurs et à d'autres technologies associées, les ordinateurs sont devenus capables d'*inverser* la tâche de perception qui consiste à distinguer un chat ou un chien. Cela signifie qu'à partir d'un texte, ils peuvent générer une image crédible d'un chat ou d'un chien, ou même des choses fantaisistes comme un astronaute à cheval. La même chose se produit avec le texte : non seulement ces modèles prennent un fragment de texte et le classent, mais, en fonction d'une demande, ces modèles peuvent essentiellement effectuer une classification à l'envers et produire le texte plausible qui pourrait correspondre à la catégorie donnée.

Angwin : une autre catégorie d'IA dont vous parlez est celle qui prétend établir des jugements automatiques. Pouvez-vous nous dire ce que ça implique ?

Narayanan : je pense que le meilleur exemple d'automatisation du jugement est celui de la modération des contenus sur les médias sociaux. Elle est nettement imparfaite ; il y a eu énormément d'échecs notables de la modération des contenus, dont beaucoup ont eu des conséquences mortelles. Les médias sociaux ont été utilisés pour inciter à la violence, voire à la violence génocidaire dans de nombreuses régions du monde, notamment au Myanmar, au Sri Lanka et en Éthiopie. Il s'agissait dans tous les cas d'échecs de la modération des contenus, y compris de la modération du contenu par l'IA.

Toutefois les choses s'améliorent. Il est possible, du moins jusqu'à un certain point, de s'emparer du travail des modérateurs de contenus humains et d'entraîner des modèles à repérer dans une image de la nudité ou du discours de haine. Il existera toujours des limitations intrinsèques, mais la modération de contenu est un boulot horrible. C'est un travail traumatisant où l'on doit regarder

en continu des images atroces, de décapitations ou autres horreurs. Si l'IA peut réduire la part du travail humain, c'est une bonne chose.

Je pense que certains aspects du processus de modération des contenus ne devraient pas être automatisés. Définir où passe la frontière entre ce qui est acceptable et ce qui est inacceptable est chronophage. C'est très compliqué. Ça demande d'impliquer la société civile. C'est constamment mouvant et propre à chaque culture. Et il faut le faire pour tous les types possibles de discours. C'est à cause de tout cela que l'IA n'a pas de rôle à y jouer.

Angwin : vous décrivez une autre catégorie d'IA qui vise à prédire les événements sociaux. Vous êtes sceptique sur les capacités de ce genre d'IA. Pourquoi ?

Narayanan : c'est le genre d'IA avec laquelle les décisionnaires prédisent ce que pourraient faire certaines personnes à l'avenir, et qu'ils utilisent pour prendre des décisions les concernant, le plus souvent pour exclure certaines possibilités. On l'utilise pour la sélection à l'embauche, c'est aussi célèbre pour le pronostic de risque de délinquance. C'est aussi utilisé dans des contextes où l'intention est d'aider des personnes. Par exemple, quelqu'un risque de décrocher de ses études ; intervenons pour suggérer un changement de filière.

Ce que toutes ces pratiques ont en commun, ce sont des prédictions statistiques basées sur des schémas et des corrélations grossières entre les données concernant ce que des personnes pourraient faire. Ces prédictions sont ensuite utilisées dans une certaine mesure pour prendre des décisions à leur sujet et, dans de nombreux cas, leur interdire certaines possibilités, limiter leur autonomie et leur ôter la possibilité de faire leurs preuves et de montrer qu'elles ne sont pas définies par des modèles statistiques. Il existe de nombreuses raisons fondamentales pour lesquelles nous pourrions considérer la plupart de ces applications de l'IA comme illégitimes et moralement inadmissibles.

Lorsqu'on intervient sur la base d'une prédiction, on doit se demander : « Est-ce la meilleure décision que nous puissions prendre ? Ou bien la meilleure décision ne serait-elle pas celle qui ne correspond pas du tout à une prédiction ? » Par exemple, dans le scénario de prédiction du risque de délinquance, la décision que nous prenons sur la base des prédictions est de refuser la mise en liberté sous caution ou la libération conditionnelle, mais si nous sortons du cadre prédictif,

nous pourrions nous demander : « Quelle est la meilleure façon de réhabiliter cette personne au sein de la société et de diminuer les risques qu'elle ne commette un autre délit ? » Ce qui ouvre la possibilité d'un ensemble beaucoup plus large d'interventions.

Angwin : certains s'alarment en prétendant que ChatGPT conduit à "l'apocalypse," pourrait supprimer des emplois et entraîner une dévalorisation des connaissances. Qu'en pensez-vous ?

Narayanan : Admettons que certaines des prédictions les plus folles concernant ChatGPT se réalisent et qu'il permette d'automatiser des secteurs entiers de l'emploi. Par analogie, pensez aux développements informatiques les plus importants de ces dernières décennies, comme l'internet et les smartphones. Ils ont remodelé des industries entières, mais nous avons appris à vivre avec. Certains emplois sont devenus plus efficaces. Certains emplois ont été automatisés, ce qui a permis aux gens de se recycler ou de changer de carrière. Il y a des effets douloureux de ces technologies, mais nous apprenons à les réguler.

Même pour quelque chose d'aussi impactant que l'internet, les moteurs de recherche ou les smartphones, on a pu trouver une adaptation, en maximisant les bénéfices et minimisant les risques, plutôt qu'une révolution. Je ne pense pas que les grands modèles de langage soient même à la hauteur. Il peut y avoir de soudains changements massifs, des avantages et des risques dans de nombreux secteurs industriels, mais je ne vois pas de scénario catastrophe dans lequel le ciel nous tomberait sur la tête.

Comme toujours, merci de votre attention.

À bientôt,
Julia Angwin
The Markup

On peut s'abonner ici à la lettre hebdomadaire (en anglais) du magazine The Markup, envoyée le samedi.