L’intelligence artificielle reproduit nos préjugés

Une nouvelle étude montre que les programmes sont aussi biaisés que les êtres humains

Le Monde · 18 apr 2017 · MORGANE TUAL

En 2016, le jury, programmé par algorithme, d’un concours de beauté a éliminé la plupart des candidats noirs

Les femmes associées aux arts et au foyer, les hommes aux professions scientifiques… Ces stéréotypes ont tellement la vie dure qu’ils se retrouvent reproduits dans des programmes d’intelligence artificielle (IA).

Dans un article publié vendredi 14 avril par la revue Science, Joanna Bryson, Aylin Caliskan et Arvind Narayanan, des chercheurs des universités de Princeton (New Jersey) et de Bath (Royaume-Uni), montrent comment une technologie de machine

learning (apprentissage machine) reproduit les biais humains, pour le meilleur et pour le pire.

La technologie sur laquelle ces scientifiques ont basé leurs travaux s’appelle GloVe. Développée par l’université Stanford (Californie), elle calcule les associations entre les mots. Ce type de programme équipe par exemple des outils de traduction automatique qui ont considérablement progressé ces dernières années.

Pour corréler des mots entre eux, GloVe doit se baser sur des exemples. D’innombrables données à partir desquelles le programme va s’entraîner pour détecter les associations les plus logiques. Une version de GloVe, fournie préentraînée, s’est basée sur Common Crawl, une base de données de milliards de textes issus du Web, sur une période de sept ans.

Reproduction de stéréotypes

Le résultat est efficace. Des mots relevant du domaine lexical des fleurs sont associés à des termes liés au bonheur et au plaisir (liberté, amour, paix, joie, paradis, etc.). Les mots relatifs aux insectes sont, à l'inverse, rapprochés de termes négatifs (mort, haine, laid, maladie, douleur, etc.).

Mais l'efficacité de cette technologie reflète aussi des associations bien plus problématiques. Des stéréotypes sexistes sont ainsi reproduits, mais aussi racistes: les qualificatifs statistiquement les plus donnés aux Noirs américains sont davantage liés à un champ lexical négatif que ceux attribués aux Blancs.

Ces résultats correspondent à ceux d'une expérimentation célèbre en psychologie, le test d'association implicite, qui étudie les associations d'idées des humains.

Conclusion : « Nos résultats suggèrent que si nous fabriquons un système intelligent qui apprenne suffisamment sur les propriétés du langage pour être capable de le comprendre et de le produire, il va aussi acquérir, dans ce processus, des associations culturelles historiques, dont certaines peuvent être problématiques. »

Or, les grands leaders du secteur comme Google ou Facebook travaillent tous à créer ce genre de système. Et les biais de l'IA sont déjà apparus au grand jour dans d'autres applications. L'une des plus spectaculaires était sans doute Tay, une IA de Microsoft lancée en 2016, censée incarner une adolescente sur Twitter.

Las, en quelques heures seulement, le programme, apprenant de ses échanges avec des humains, s'est mis à tenir des propos racistes et négationnistes, avant d'être suspendu par Microsoft en catastrophe.

Et le problème ne se situe pas seulement au niveau du langage. Quand un programme d'IA est devenu jury d'un concours de beauté, en septembre 2016, il a éliminé la plupart des candidats noirs. « Beaucoup de gens nous disent que cela montre que l'IA a des préjugés, souligne Joanna Bryson

dans le Guardian. Mais non. Cela montre que nous avons des préjugés, et que l'IA les apprend. » Le 19 janvier au Sénat, Serge Abiteboul, directeur de recherche à l'Institut national de recherche en informatique et en automatique (Inria), expliquait que ces biais pouvaient avoir des conséquences bien plus graves. « Prenez l'exemple d'une demande de prêt. [Pour le banquier], c'est une tâche relativement répétitive : à partir des données d'une personne, on décide de lui accorder le prêt ou pas. Par nature, un humain va avoir des biais et accorder ces prêts de façon injuste », rappelait-il. «On pourrait penser que l'algorithme serait beaucoup plus juste. Mais ce n'est pas si simple. Si on utilise le machine learning, alors on se base sur des données créées par les humains pendant dix ans, et l'algorithme va reproduire les préjugés que ces humains ont exprimés. »

La même logique vaut pour la sélection de CV, le prix des assurances, mais aussi pour la justice – certaines villes américaines utilisent des programmes d'IA pour tenter de prévoir les crimes. « Ces logiciels peuvent avoir des effets considérables sur nos sociétés, ils doivent donc se comporter de façon responsable », poursuit le chercheur. Une inquiétude partagée par une grande partie des acteurs de l'IA comme Google et Microsoft.

Plusieurs pistes de réflexion sont évoquées. Des chercheurs de l'université d'Oxford (RoyaumeUni) prônent dans un article la mise en place d'une autorité chargée d'auditer ces algorithmes et d'enquêter quand un citoyen s'estime victime de discrimination de la part de l'algorithme.

L'Union européenne travaille sur un « droit à l'explication », qui imposerait aux entreprises utilisant ces programmes d'être capables d'expliquer les décisions qu'ils prennent aux personnes concernées.

Des milliards d'éléments

Problème, beaucoup de ces programmes sont basés sur le deep

learning, une technologie d'apprentissage très efficace, mais très opaque.

D'autres pistes évoquées pour lutter contre les biais des IA sont aussi, par exemple, que celles-ci soient conçues par une population plus hétéroclite, à savoir moins masculine et blanche. Enfin, le plus évident pourrait être de s'attaquer aux données.

Mais comment intervenir sur des corpus comprenant des millions, voire des milliards d'éléments ? Qui plus est, si l'objectif est in

fine de permettre aux programmes d'IA de comprendre le langage en se basant sur l'interprétation du monde par les êtres humains, modifier les données pourrait fausser l'ensemble. Finalement, s'il fallait agir sur la source du problème, ce serait les humains qu'il faudrait modifier…

Ce pourrait être l'occasion de mettre au jour leurs préjugés et leurs pratiques discriminatoires, comme le suggère Sandra Wachter, chercheuse en éthique des données et des algorithmes à Oxford, dans le Guardian : «Les humains, eux, peuvent mentir sur les raisons pour lesquelles ils n'embauchent pas quelqu'un. A l'inverse, les algorithmes ne mentent pas et ne nous trompent pas. »