Où en sommes-nous exactement en termes de biais discriminatoires et d'IA ?

Rechercher des schémas, traiter des données, développer des idées et prendre des décisions en conséquence, c'est ce que nous faisons tous chaque jour. L'intelligence artificielle (IA) étant de plus en plus omniprésente dans notre société, nous attendons d'elle non seulement qu'elle fasse la même chose, mais en mieux. De incidents récents ont cependant montré que les modèles d'IA, de quelque type que ce soit, peuvent intégrer des biais, en traitant différemment des personnes d'origine raciale ou de sexe différents. Où en sommes-nous exactement en termes de biais discriminatoires et d'IA ? Les machines sont-elles les seules responsables ?

Au cours de l'automne 2019, la carte de crédit d'Apple a rencontré de gros problèmes lorsque les utilisateurs ont remarqué que son algorithme d'apprentissage des machines semblait proposer des limites de crédit plus basses aux femmes qu'aux hommes. Ces accusations sont apparues sur les médias sociaux après que l'entrepreneur technologique David Heinemeier Hansson a tweeté que la carte Apple lui offrait 20 fois la limite de crédit de sa femme, alors même qu'ils avaient les mêmes actifs en commun et qu'elle bénéficiait d'un score de crédit plus élevé.

Dans le domaine du traitement de l'image, un autre domaine important pour l'IA, Zoom et Twitter ont récemment fait parler d'eux à des occasions semblables. En effet, lors de l'utilisation de la plateforme de vidéoconférence, un enseignant noir était « décapité » chaque fois qu'il activait la fonction d'arrière-plan virtuel, l'algorithme de segmentation de l'image ne reconnaissant pas sa tête comme telle. En discutant de ce sujet sur Twitter, un étudiant de la même école a également été confronté au problème : l'application de réseau social coupait automatiquement l'image du professeur en mode de prévisualisation.

En matière de reconnaissance vocale, des recherches ont montré que le degré de précision de certains systèmes de reconnaissance vocale très performants était supérieur de 13 % pour les hommes par rapport aux femmes. Compte tenu de la prévalence croissante de ces interfaces, cela ne peut que conduire à des situations problématiques.

Enfin, dans le domaine du traitement du langage naturel (TLN), les modèles d'intégration des mots - éléments de base de nombreuses tâches de TLN - ont également tendance à encoder des biais. Un exemple couramment cité est une analogie dérivée de ces modèles selon laquelle le programmeur informatique est à la femme au foyer ce que l'homme est à la femme.

Il existe de multiples raisons pour lesquelles de tels biais apparaissent dans les résultats des modèles d'IA. Dans le cas des cartes de crédit, l'algorithme avait apparemment été vérifié par une tierce partie pour détecter d'éventuels biais et n'utilisait pas le sexe comme facteur. Alors comment pouvait-il faire de la discrimination si personne ne lui avait jamais indiqué comment distinguer un client homme d'un client femme ? En fait, il est tout à fait possible pour les algorithmes d'apprentissage automatique de faire des discriminations en fonction du sexe, même lorsqu'ils sont programmés pour être « aveugles » sur cette variable. En fait, c'est précisément ce à quoi ces algorithmes excellent : trouver des caractéristiques latentes dans les données, c'est-à-dire des caractéristiques qui ne sont pas utilisées pour entraîner les modèles, mais qui peuvent être dérivées des données utilisées pour les entraîner.

Une autre source commune de biais réside dans le fait que les modèles tirent des enseignements d'une série d'exemples, mais peuvent les utiliser dans un contexte différent. Serait-il judicieux d'emprunter un modèle de reconnaissance faciale conçu pour le personnel de l'armée en Suède et de l'utiliser pour un magasin de détail en Espagne ?

Dans le domaine du TLN, l'un des coupables semble être les représentations sous-jacentes des mots (« plongement lexical »). Bien que ces modèles soient généralement formés à partir de corpus relativement génériques et représentatifs (les « datasets » dans le jargon du TLN) tels que Wikipédia, ils n'en absorbent pas moins les préjugés existants de la société qui sont encore implicitement présents dans notre langue. Des études en sciences sociales ont longtemps souligné la manière dont l'idéologie de genre est ancrée dans le texte (et dans d'autres artefacts sociétaux tels que les films ou les images). Ainsi, dans les publications d'entreprises, les hommes sont mentionnés dix fois plus souvent que les femmes. Dans ces cas, le biais ne résulte pas tant de la sélection des données de formation que des préjugés de la société, qui sont simplement captés par le processus de formation.

Pour résumer, la validité des algorithmes dépend de celle des données à partir desquelles ils sont créés. Les êtres humains ont une forte prédisposition pour les préjugés,

LÉA DELERIS & RIM TEHRAOUI

qu'ils soient conscients ou inconscients. Ces biais cognitifs sont profondément enracinés dans de nombreux aspects de notre environnement économique et social : lieu de travail, soins de santé, bourses d'études, application de la loi, marketing, pour n'en citer que quelques-uns. Il n'est donc pas surprenant que le biais s'infiltre également dans chaque octet de données produites par notre société et, par conséquent, dans les algorithmes constitués à partir de ces données.

Naturellement, plus nous les utilisons, en les incorporant dans les processus de décision, plus ils propagent, voire amplifient, nos biais sociaux et cognitifs implicites, même lorsque nous-mêmes, en tant que société, œuvrons activement à les supprimer.

Il existe déjà une variété d'approches techniques destinées à réduire les biais dans les modèles d'IA. Les solutions proposées sont généralement spécifiques à un certain domaine ou à un certain type de modèles. Les approches de suppression des biais, quelles qu'elles soient, constituent nécessairement des contraintes pour les modèles et peuvent donc entraîner une diminution de leur performance brute. Un tel compromis n'est pas nouveau, ni nécessairement problématique en soi, mais il doit être reconnu et surtout accepté.

Cependant, au-delà des solutions techniques, une première étape évidente pour traiter les biais de l'IA est d'assurer une supervision humaine adéquate, en sensibilisant et en formant les gens sur l'importance d'identifier et d'atténuer les biais lorsque les modèles intègrent des variables pouvant conduire à des décisions discriminatoires. Les spécialistes des données (mais aussi les propriétaires des modèles en cours d'élaboration) devraient avoir pour mission d'analyser le contexte de l'utilisation du modèle et les données qu'il utilise comme intrants.

Dans ce contexte, une mesure importante consiste à prendre en compte le problème de la diversité sur le lieu de travail, afin de garantir l'inclusion d'un éventail de perspectives critiques dans le développement des modèles. Dans une enquête du Women's Forum qui sera bientôt publiée, 77 % des personnes interrogées (provenant de nombreux pays du monde) estiment que si les femmes avaient un meilleur accès à l'emploi dans le domaine des STEM (science, technologie, ingénierie et mathématiques) et de l'IA, cela conduirait à des applications numériques et à des outils d'IA bénéfiques pour tous. De même, 75 % pensent que cela réduirait le risque que les applications technologiques et les outils d'IA génèrent des inégalités entre hommes et femmes.

Il convient toutefois d'étudier le fond du problème. Les conclusions de l'étude de l'UNESCO « Je rougirais si je pouvais » mettent en évidence la persistance de la fracture numérique entre les genres, mais aussi l'existence du « paradoxe de l'égalité des genres dans le domaine des technologies de l'information et de la communication (TIC) », c'est-à-dire le fait que les pays où l'égalité des sexes est la plus élevée ont aussi les plus faibles pourcentages de femmes qui poursuivent des études supérieures en informatique et dans des domaines connexes.

Ne vous découragez pas trop ! La diversité au sein des équipes d'IA ne signifie pas simplement l'inclusion de profils de scientifiques plus diversifiés. Il faut également s'efforcer d'inclure différents profils techniques et non techniques au sein des équipes : analystes commerciaux, ingénieurs en logiciels, concepteurs d'UX (expérience utilisateur), spécialistes des sciences sociales, mais aussi une variété d'origines culturelles et de parcours professionnels. La clé est d'avoir un environnement où les gens se sentent libres, et même habilités, à remettre en question les théories.

Globalement, si l'IA peut jouer un rôle dans la propagation des préjugés, il n'y a pas de fatalité et c'est aux êtres humains de mettre en place les garde-fous appropriés en plaçant la diversité au centre !

LÉA DELERIS & RIM TEHRAOUI