L'odyssée de l'intelligence artificielle

IA : orgueil et préjugés

EPISODE 19

Pendant tout l'été, l'Opinion décrypte les dessous de la révolution que nous sommes à la veille de vivre. Nous sommes entrés dans une nouvelle ère. Attachez vos ceintures.

L'OPINION - ADOBE FIREFLY

Cette illustration a été réalisée avec l'intelligence artificielle générative **Adobe Firefly**.

L'IA N'EST BONNE OU MAUVAISE que dans la mesure où nous lui permettons de l'être. En tant que chercheurs, programmeurs et utilisateurs, nous devons rester vigilants face aux préjugés que nous pourrions encoder à notre insu dans nos systèmes d'IA. Ce n'est qu'à cette condition que nous pourrons réellement exploiter le potentiel de l'IA pour réduire les préjugés et les discriminations, et progresser vers un monde plus équitable et un environnement plus inclusif et impartial.

L'Intelligence artificielle n'est ni bonne ni mauvaise en soi. C'est un outil façonné par l'intention humaine. La même technologie qui peut involontairement amplifier les préjugés humains peut également être exploitée pour réduire ces préjugés et promouvoir une plus grande équité. Il nous appartient de guider l'IA sur la bonne voie.

Contrairement aux humains, où les préjugés peuvent s'ancrer profondément au fil du temps, l'IA peut être ajustée rapidement et efficacement dès qu'un préjugé est détecté. Il n'est pas nécessaire de mener de longues campagnes éducatives ou sociétales pour modifier le comportement de l'IA - un programmeur peut le faire en quelques lignes de code.

Lorsqu'on applique l'Intelligence artificielle au domaine de la juridiction, l'objectif n'est pas de remplacer le jugement humain, mais de l'améliorer, de nous rendre plus conscients de nos préjugés inhérents et de nous aider à les contrer. Tout comme un correcteur orthographique nous signale un mot mal orthographié, l'IA pourrait nous alerter sur des préjugés potentiels, nous poussant ainsi vers une société plus juste et plus équitable.

Espoir. Or, les décisions judiciaires sont systématiquement entachées de préjugés raciaux et sexistes. Nous le savons par le biais d'analyses sophistiquées des preuves historiques, mais aussi par la simple observation du monde réel. Par exemple, les juges fédéraux nommés par les Républicains condamnent plus sévèrement les accusés noirs et avec plus de clémence les accusés de sexe féminin. Les juges fédéraux adoptent un comportement plus politique avant les élections présidentielles, en particulier les juges résidant dans des Etats où les élections présidentielles sont serrées. Le parti politique d'un juge peut être prédit par les citations qu'il choisit pour motiver ses décisions.

Contrairement aux humains, l'IA est fondamentalement flexible. Elle peut être reprogrammée et ajustée pour atténuer les préjugés, un processus bien plus direct que d'essayer de remodeler des préjugés humains profondément ancrés

Ces données révèlent une partialité durable. D'une part, l'identité du juge peut mener à des décisions arbitraires, par exemple son identité raciale est prédictive des disparités dans leurs décisions de condamnation. D'autre part, ce sont également des éléments futiles tels que la victoire ou la défaite de l'équipe de football de la ville natale d'un juge, le fait que ce soit l'anniversaire du plaideur, etc. qui peuvent jouer un rôle dans la décision. En outre, les groupes minorisés subissent systématiquement le poids punitif de ces déviations de l'objectivité.

Alors que nous luttons contre les préjugés dans nos sociétés, un champ de bataille controversé mais crucial est apparu : l'intelligence artificielle. Le monde numérique reflète l'analogique, et nos systèmes d'IA sont sensibles à nos préjugés inhérents.

Cependant, il y a de l'espoir : contrairement aux humains, l'IA est fondamentalement flexible. Elle peut être reprogrammée et ajustée pour atténuer les préjugés, un processus bien plus direct que d'essayer de remodeler des préjugés humains profondément ancrés.

A partir d'une étude sur les attitudes à l'égard du genre dans les cours d'appel des Etats-Unis, je montrerai comment l'IA peut contrer les préjugés de manière plus efficace que les humains. Elle peut le faire en diagnostiquant les préjugés d'une manière que les humains ne peuvent pas faire.

L'étude en question utilise le traitement du langage naturel (NLP), une branche de l'intelligence artificielle, pour détecter les attitudes des juges à l'égard des femmes. Les chercheurs ont mis au point une mesure du « biais de genre » pour évaluer la manière dont les juges associent les hommes à la carrière et les femmes à la famille dans leurs opinions écrites. Cette approche nuancée a permis de mettre en lumière des préjugés sexistes subtils que l'analyse directe des décisions n'avait pas permis de déceler.

La force unique de l'IA réside ici dans sa capacité à agréger et à analyser objectivement d'énormes quantités de données, bien au-delà des capacités humaines. La NLP a fourni une mesure quantitative et sans ambiguïté du biais de genre en examinant 380 000 avis publiés, ce qui serait peu pratique, voire impossible, à réaliser pour les humains dans un délai comparable. L'IA recèle donc un potentiel exceptionnel de détection des biais qui dépasse les capacités humaines.

Biais de genre. Le biais de genre est un indicateur convaincant de la partialité. Les juges féminins et plus jeunes affichent un penchant moins marqué pour le genre. Le fait d'avoir une fille réduit le biais sexiste. Une inclinaison plus faible est également associée à une utilisation plus fréquente de pronoms neutres, tels que „il ou elle". Enfin, les juges ayant une position plus tranchée ont tendance à exprimer moins d'empathie à l'égard des femmes dans leurs écrits.

Non seulement les juges diffèrent systématiquement dans leur façon d'écrire sur le genre, mais ces différences sont prédictives de la façon dont ils statuent sur les affaires de droits des femmes et de la façon dont ils traitent leurs collègues féminines. L'étude examine la manière dont les juges ayant des positions différentes sur le genre interagissent avec les femmes juges dans trois domaines : l'annulation des décisions des juridictions inférieures, l'attribution d'opinions et les citations.

Les résultats montrent que les juges ayant un plus grand biais de genre sont plus susceptibles d'annuler les décisions des juges de district féminins, moins susceptibles d'attribuer l'origine de l'opinion aux juges féminins et moins susceptibles de citer les opinions des juges féminins. Ces juges ont également tendance à voter de manière plus conservatrice dans les affaires liées au genre. Les résultats suggèrent que les préjugés sexistes pourraient entraver la progression de carrière des femmes juges et renforcer la disparité entre les sexes dans le système judiciaire.

Problème systémique. La sous-représentation des femmes au sommet de la profession juridique est un problème qui a fait l'objet d'une attention considérable aux Etats-Unis. Il est troublant de constater que, bien que près de 45% des diplômés des facultés de droit soient des femmes depuis les années 1990, celles-ci ne représentent toujours que 20% des associées dans les grands cabinets juridiques et 30% des juges fédéraux et d'Etat. Les disparités dans ces chiffres révèlent un problème systémique : le traitement différentiel des femmes juges, qui pourrait être dû à des attitudes sexistes de la part de leurs collègues.

Les attitudes de genre, c'est-à-dire les préjugés et les idées préconçues que l'on a sur les groupes sociaux, notamment les femmes et les minorités raciales, sont connues pour influencer de manière significative les jugements et les choix. Ces préjugés affectent les décisions dans toute une série de contextes, depuis les traitements médicaux et les décisions d'embauche jusqu'aux interactions entre employeurs et employés et même l'efficacité des enseignants. Si ces attitudes impliquent un traitement différencié des femmes juges, elles pourraient être un facteur contribuant à la sous-représentation des femmes dans la magistrature.

DR

Il est difficile d'examiner ces questions parmi les acteurs de la justice en raison de l'absence de mesures traditionnelles des attitudes à l'égard du genre pour les juges. Cependant, des chercheurs ont utilisé de manière innovante les développements récents en matière de traitement du langage naturel (NLP) pour proposer une nouvelle mesure des attitudes liées au genre. En analysant un vaste corpus de textes écrits par des juges d'appel, les chercheurs ont mis au point une mesure des préjugés sexistes basée sur la force avec laquelle les juges associent les hommes à la carrière et les femmes à la famille dans leurs écrits. A l'aide d'un outil technologique appelé « word embeddings », les chercheurs ont calculé une mesure des préjugés sexistes spécifique aux juges.

Lorsque nous passons de la phase d'analyse à la phase d'application dans l'étude sur les attitudes des hommes et des femmes dans les cours d'appel des Etats-Unis, l'IA pourrait être utilisée pour contrer les préjugés sexistes détectés. Les systèmes peuvent être programmés pour inciter les juges humains à réfléchir et à reconsidérer les éventuels préjugés implicites.

Daniel L. Chen

Daniel L. Chen est chercheur à TSE et directeur de recherche au CNRS.

Etudiants

L'impact de ChatGPT sur les étudiants varie. Une dépendance excessive peut affaiblir la résolution autonome de problèmes et l'apprentissage passif peut diminuer la réflexion critique. Cependant, une utilisation responsable en tant que complément pédagogique peut renforcer la compréhension. L'effet dépend de la manière dont les étudiants l'emploient et de leur responsabilité personnelle.

ChatGPT

Prochain épisode
IA et risque démocratique

l'Opinion **BEY MEDIAS PRESSE & INTERNET SASU** - 8, rue Bellini, 75116 Paris - www.lopinion.fr - Actionnaire : Bey Médias SAS. **Président, directeur de la publication :** Nicolas Beytout. **Vice-président, rédacteur en chef :** Rémi Godeau. **Rédacteurs en chef adjoints :** Olivier Baccuzat, Emmanuel Ginisty, Raphaël Legendre et Cécile Desjardins. **BEY MEDIAS PUBLICITÉS** - Tél : 01 40 69 46 96. **Directrice générale adjointe, commerce :** Cécile Colomb. **Publicité commerciale :** Ronan Daligault, Pierre-Louis Orsini. **Publicité financière :** Nicolas Wattinne. **SERVICE ABONNEMENTS L'OPINION :** 45 avenue du Général Leclerc, 60643 Chantilly Cedex - Tél : 01 55 56 70 73 du lundi au vendredi de 9h à 18h - Email : abonnements@lopinion.fr. Formule d'abonnement pour 1 an - France métropolitaine : 319 €. **IMPRESSION :** Paris Offset Print (La Courneuve). Dépôt légal mai 2013 - ISSN : 2266-9922 / CPPAP : 1123 C 91894.
Papier produit à partir de fibres issues de forêts certifiées gérées durablement. Origine du papier : Norvège. Taux de fibres recyclées : 0%. Eutrophisation : Ptot 0,022 kg / tonne de papier