COMMENT L'IA PEUT DEBOUCHER SUR DE LA DISCRIMINATION

ces 4 pages sont issues de l'étude « Discrimination, intelligence artificielle et décisions algorithmiques », Conseil de l'Europe (2018, 53 pages, en ligne)

Cette section examine comment l'IA peut déboucher sur de la discrimination; la suivante donne des exemples dans lesquels elle l'a fait ou pourrait le faire.

Beaucoup de systèmes d'IA sont des «boîtes noires». Une personne ne comprendra souvent pas pourquoi un système a formulé telle ou telle décision à son sujet. En raison de l'opacité de la décision, il lui sera difficile de voir si elle a été victime de discrimination, par exemple en raison de son origine raciale. Les décisions fondées sur l'IA peuvent conduire de plusieurs façons à des discriminations. Dans un article fondateur en 2016, Barocas et Selbst distinguent cinq façons dont une décision d'IA peut involontairement aboutir à une discrimination (voir ce lien) Les problèmes proviennent 1) de la définition de la variable cible et des étiquettes de classe; 2) de l'étiquetage des données d'apprentissage; 3) de la collecte des données d'apprentissage; 4) de la sélection des caractéristiques; 5) du choix des données indirectes. De plus, 6) les systèmes d'IA peuvent être délibérément utilisés à des fins discriminatoires. Nous allons maintenant revenir sur chacun de ces éléments.

1) Définition de la variable cible et des étiquettes de classe

L'IA consiste pour l'ordinateur à découvrir des corrélations dans des jeux de données. Une société qui met au point un filtre de courriers indésirables, par exemple, fournit à l'ordinateur des messages étiquetés par des humains comme indésirables ou non. Ces messages étiquetés constituent les données d'apprentissage. L'ordinateur y découvre les caractéristiques d'un courrier indésirable. L'ensemble de corrélations mises au jour est souvent appelé le modèle ou le modèle prédictif. Les messages repérés comme indésirables contiendront souvent, par exemple, certaines expressions («perte de poids massive», «millions d'euros à gagner», etc.) ou émaneront de certaines adresses IP. Comme le disent Barocas et Selbst, l'algorithme d'apprentissage automatique tire d'exemples pertinents (fraudes précédemment identifiées, courriers indésirables, défauts de paiement, mauvaise santé) les attributs ou les actions (données indirectes) qui peuvent servir à détecter la présence ou l'absence de la qualité ou du résultat recherchés (la variable cible)

La variable cible représente ce que l'explorateur de données recherche, expliquent Barocas et Selbst, tandis que les étiquettes de classe répartissent toutes les valeurs possibles de cette variable cible entre des catégories s'excluant les unes les autres
Pour le filtrage des courriers indésirables, par exemple, on s'accorde dans l'ensemble sur les étiquettes de classe de courrier bienvenu ou indésirable. Mais dans certains cas, la définition de la valeur de la variable cible est moins évidente. Parfois, constatent Barocas et Selbst, il faut créer de nouvelles classes pour la décrire.
Prenons le cas d'une entreprise qui confie à un système d'IA le soin de classer des réponses à une offre d'emploi pour en extraire de «bons employés». Comment va-t-on définir le bon employé? En d'autres termes, quelles devraient être les étiquettes de classe? Le bon employé est-il celui qui réalise les meilleures ventes ou celui qui n'arrive jamais en retard au travail? Certaines variables cibles et étiquettes de classe, expliquent Barocas et Selbst, peuvent avoir un impact négatif plus ou moins marqué sur des classes protégées.

Supposons par exemple que les personnes défavorisées habitent rarement en centre-ville et viennent de plus loin que les autres employés pour se rendre à leur travail. Elles seront donc en retard plus souvent, en raison des embouteillages ou de problèmes de transports publics. L'entreprise peut par exemple choisir d'apprécier si un employé est «bon» par l'étiquette de classe «rarement ou souvent en retard». Mais si les personnes issues de la migration sont en moyenne plus pauvres et habitent plus loin de leur travail, ces étiquettes désavantagent les immigrés, même s'ils font mieux que les autres employés à d'autres égards.
En bref, la discrimination peut s'introduire dans un système d'IA en raison de la façon dont une organisation définit les variables cibles et les étiquettes de classe.

2) Données d'apprentissage: exemples d'étiquetage

La décision par IA peut aussi produire des effets discriminatoires si le système «apprend» à partir de données discriminatoires. Barocas et Selbst décrivent deux façons dont des données d'apprentissage biaisées peuvent produire des effets discriminatoires: d'une part, le système peut faire son apprentissage sur des données biaisées; et d'autre part, des problèmes peuvent surgir si le système apprend à partir d'un échantillon biaisé. Dans les deux cas, le système reproduira le biais.Les données d'apprentissage peuvent être biaisées si elles reflètent des décisions humaines discriminatoires. C'est ce qui s'est produit dans les années 1980 au Royaume-Uni, dans une école de médecine. L'établissement recevait plus de candidatures qu'il n'avait de places. Il a donc mis au point un logiciel d'aide au tri des candidatures. Les données d'apprentissage étaient constituées par les dossiers d'admission des années précédentes, lorsque des intervenants humains sélectionnaient les candidats. Elles montraient à l'ordinateur les caractéristiques (intrants) corrélées avec le résultat souhaité (admission à l'école de médecine). L'ordinateur a ainsi reproduit ce système de sélection Il s'est avéré que l'ordinateur défavorisait les femmes et les personnes issues de la migration. Il semblerait que dans les années dont provenaient les données d'apprentissage, les personnes chargées de la sélection des étudiants avaient des préjugés contre les femmes etles personnes issues de la migration. Comme le notait le British Medical Journal, le programme n'introduisait pas de nouveau biais, se contentant de reproduire celui qui existait déjà dans le système38. Pour conclure, si les données d'apprentissage sont biaisées, il y a des chances pour que le système d'IA reproduise ce biais.

3) Données d'apprentissage: collecte des données

La procédure d'échantillonnage peut aussi être biaisée. Par exemple,dans la collecte de données sur la criminalité, il se pourrait que la police ait interpellé dans le passé plus de personnes issues de la migration. Lum et Isaac observent que si la police se concentre sur certains groupes ethniques et certains quartiers, il est probable que ces catégories seront surreprésentées dans ses fichiers.Si un système d'IA s'appuie sur des données ainsi biaisées, il apprend qu'il est probable que les personnes issues de la migration commettent des infractions. Pour Lum et Isaac, si l'on utilise des données biaisées pour former des modèles prédictifs, ces derniers reproduiront les mêmes biais.Les effets d'un échantillon ainsi biaisé peuvent même être amplifiés par les prédictions de l'IA. Supposons que la police concentre ses activités sur un quartier à forte population immigrée, mais à criminalité moyenne. Elle y enregistra plus d'infractions qu'ailleurs. Comme les chiffres font ressortir un nombre supérieur d'infractions enregistrées (et donc censées s'être produites) dans ce quartier, les autorités y affecteront davantage encore d'agents de police. En d'autres termes,

organiser le maintien de l'ordre en se fondant sur des statistiques de criminalité peut créer une boucle de rétroaction positive. Pour prendre un autre exemple, les pauvres peuvent être sous-représentés dans un jeu de données. L'application Street Bump pour smartphone, par exemple, recourt à la géolocalisation pour surveiller l'état des routes dans une ville. Le site explique que des bénévoles l'utilisent sur leur téléphone pour signaler l'état de la route pendant leurs trajets. Ces données sont communiquées en temps réel aux autorités, qui peuvent ainsi procéder aux réparations et planifier leurs investissements à long terme. Si les pauvres sont moins nombreux à posséder un smartphone que les personnes plus aisées, ils seront sous-représentés. Cela pourrait avoir pour effet que les routes détériorées des quartiers pauvres seront moins souvent signalées dans les jeux de données, et donc moins fréquemment réparées. Street Bump était utilisé à Boston, où la municipalité s'efforce de remédier à ce biais dans la collecte des données. Mais cet exemple n'en illustre pas moins que la collecte des données peut produire un biais non voulu dans les données. En bref, un biais dans les données d'apprentissage peut produire un biais dans le système d'IA.

4) Sélection des caractéristiques

Quatrième problème: les caractéristiques (catégories de données) que choisit une organisation pour son système d'IA. Si elle veut utiliser ce dernier pour automatiser une prédiction, il va lui falloir simplifier le monde, pour pouvoir le décrire par des données. Comme le disent Barocas et Selbst, une organisation doit procéder à des choix, pour sélectionner les caractéristiques qu'elle veut observer et introduire dans ses analyses .Supposons qu'une organisation veuille sélectionner par prédiction automatisée les candidats qui seront de bons employés. Il est impossible, ou du moins trop coûteux, pour un système d'IA d'évaluer la totalité de chaque dossier de candidature. L'organisation peut alors retenir, par exemple, uniquement certaines caractéristiques applicables à chaque dossier.Le choix de certaines caractéristiques peut introduire un biais contre certains groupes. Par exemple, de nombreux employeurs aux États-Unis préfèrent les personnes qui ont fait leurs études dans l'une des grandes universités onéreuses. Mais il peut être rare que les membres de certains groupes raciaux fréquentent ces établissements. Le système aura alors des effets discriminatoires dès lors qu'un employeur se fonde sur la fréquentation d'une grande université pour sélectionner les candidats. En bref, une organisation peut susciter des effets discriminatoires par la sélection des caractéristiques qu'utilise le système d'IA dans ses prédictions.

5) Données indirectes

Les données indirectes font aussi problème. Certaines données incluses dans le jeu d'apprentissage peuvent présenter des corrélations avec des caractéristiques protégées. Comme le soulignent Barocas et Selbst, des critères authentiquement pertinents de décisions rationnelles et solidement fondées peuvent aussi constituer des indicateurs fiables d'appartenance à une classe. Supposons qu'une banque utilise un système d'IA censé prédire quels demandeurs de prêt auront du mal à rembourser un crédit. L'apprentissage du système a été fondé sur les données des vingt dernières années, et le jeu ne contient pas d'informations concernant des caractéristiques protégées, comme la couleur de la peau. Le système d'IA apprend que les personnes qui ont un certain code postal ont tendance à ne pas rembourser, et il utilise cette corrélation pour prédire le non-remboursement du crédit. Un critère à première vue neutre (le code postal) sert donc à prédire le défaut de paiement.Mais supposons maintenant qu'il y ait une corrélation entre ce code postal et l'origine raciale. Si la banque prend ses

décisions sur la base de cette prédiction et refuse d'accorder des crédits aux habitants de ce quartier, cela fait du tort aux membres d'un certain groupe sur le critère de l'origine raciale. Barocas et Selbst expliquent que le problème provient de ce que les chercheurs appellent un encodage redondant, c'est-à-dire l'appartenance à une classe protégée encodée dans d'autres données.C'est ce qui se passe lorsqu'une donnée ou certaines valeurs de cette donnée sont étroitement corrélées avec l'appartenance à une classe spécifique protégée.
Par exemple: un jeu de données qui ne contient pas de données explicites sur l'orientation sexuelle peut tout de même la dévoiler. Une étude de 2009 a montré que les liens d'«amis» sur Facebook révèlent l'orientation sexuelle par une méthode de prédiction précise de l'orientation sexuelle des utilisateurs de Facebook fondée sur l'analyse de leurs liens. Le pourcentage d'«amis» s'identifiant comme homosexuels serait fortement corrélé avec l'orientation sexuelle de l'utilisateur concerné. Ce problème des données indirectes est délicat. Barocas et Selbst indiquent que les informaticiens ne voient pas très bien comment aborder l'encodage redondant d'un jeu de données. Se contenter de retirer les variables concernées de l'exploration des données supprime fréquemment des critères d'une pertinence démontrable et justifiée dans la décision à prendre. La seule façon de garantir que les décisions ne désavantagent pas systématiquement les membres de catégories protégées est de réduire la précision générale de toutes les déterminations

6) Discrimination délibérée

Il faut aussi parler de la discrimination délibérée. Une organisation peut par exemple utiliser volontairement des données indirectes pour pratiquer la discrimination sur le critère de l'origine raciale. Comme l'observent Krollet al., des préjugés pourraient conduire un responsable à biaiser volontairement les données d'apprentissage ou à choisir certaines données codant indirectement des classes protégées pour obtenir des résultats discriminatoires . Si l'organisation utilise des données indirectes, la discrimination sera plus difficile à détecter qu'une discrimination franche.Prenons un exemple hypothétique: une organisation pourrait vouloir écarter les femmes enceintes, et cette discrimination serait difficile à détecter. Le distributeur américain Target aurait constitué un score de prédiction de grossesse fondé sur quelque 25produits en analysant les habitudes d'achat des clientes. Si l'une d'entre elles achetait ces produits, le magasin pouvait savoir avec une bonne certitude qu'elle était enceinte. Target voulait toucher par la publicité des personnes à un moment de leur vie où elles ont tendance à modifier leurs habitudes d'achat. Il voulait donc savoir quand ses clientes allaient avoir un enfant. Ses statisticiens savaient que s'ils arrivaient à les identifier au deuxième mois, ils avaient de bonnes chances de les fidéliser pour des années.Target utilisait la prédiction pour cibler son marketing, mais une organisation pourrait aussi le faire à des fins discriminatoires.

- Voir en ligne la section suivante (page 12), qui donne des exemples de décisions d'IA ayant suscité ou susceptibles de produire des discriminations :
    - Police, prévention de la criminalité
    - Recrutement d'employés et d'étudiants
    - Publicité
    - Discrimination par les prix
    - Recherche et analyse d'images
    - Outils de traduction
- Voir en ligne l'article de C. Benavent (2016), Big Data, algorithmes et marketing : rendre des comptes