Droits et devoirs

Encadrer les biais algorithmiques : un impératif

Les systèmes d'intelligence artificielle (SIA) peuvent intégrer intentionnellement ou non des partis pris discriminatoires difficiles à détecter et à corriger. Des études récentes sur les biais algorithmiques permettent d'entrevoir des pistes pour en limiter les risques.

Les algorithmes sont utilisés pour prendre des décisions souvent importantes, comme l'admission à une université, l'octroi d'un prêt ou l'attribution d'un emploi ou d'une aide sociale. Or, ils peuvent être influencés par les données et les préjugés de ceux qui les ont conçus et entraînés.

Alain Bensoussan.

Les biais dans les SIA d'assistance à la prise de décision

L'Organisation internationale de normalisation a publié fin 2021, une norme[1] qui définit le biais comme une « *différence systématique dans le traitement de certains objets, personnes ou groupes par rapport à d'autres* », le traitement s'entendant de « tout type d'action, y compris la perception, l'observation, la représentation, la prédiction ou la décision ».

Les biais algorithmiques peuvent se manifester de différentes manières[2]. Ils peuvent affecter les données initialement employées. Une donnée n'est jamais un fait totalement « objectif ». Un facteur de discrimination peut naître de façon indirecte si une donnée cible une caractéristique qui peut se révéler discriminatoire.

Ainsi, une simple adresse postale peut indiquer avec une certaine probabilité que la personne est issue d'une catégorie de population (par exemple, de l'immigration), en fonction de la concentration de cette catégorie dans certaines zones géographiques.

Une IA autoapprenante pourra ainsi déceler une corrélation entre l'adresse et le sens de la décision, et aboutir à des résultats discriminatoires, alors même que l'origine des personnes n'est pas renseignée.

Les risques de discrimination

Les biais algorithmiques soulèvent, on le voit, des préoccupations considérables en termes de droits des individus concernés par une prise de décision algorithmique. Ils devraient être détectés, signalés et neutralisés au stade le plus précoce possible.

Dans son dernier rapport paru en décembre 2022[3], l'Agence des droits fondamentaux de l'Union européenne a développé deux études de cas qui analysent la façon dont les biais apparaissent dans les algorithmes et comment ils affectent la vie des personnes.
De leur côté, le Défenseur des droits et la Cnil ont dès 2020, mis l'accent sur les risques de discrimination pouvant résulter des biais algorithmiques[4].
Les risques sont bien réels. On se souvient des derniers scandales, particulièrement chez les GAFAM, à propos de certains algorithmes de recommandation proposant un contenu reflétant les préférences et les croyances des utilisateurs, plutôt que de présenter de nouvelles idées et perspectives.

Les pistes à exploiter pour limiter les risques

La norme ISO précitée analyse les pratiques actuelles pour détecter et traiter les biais algorithmiques, quelle qu'en soit la source.
Elle aborde l'ensemble des phases du cycle de vie du système d'IA, y compris, mais sans s'y limiter, la collecte de données, la formation, l'apprentissage continu, la conception, les tests, l'évaluation et l'utilisation, afin que les biais soient éliminés à chacun de ces stades.
De son côté, le Conseil de l'Europe[5] a fixé des orientations à l'intention des développeurs, fabricants et prestataires de service en IA concernant les biais : ceux-ci devraient *« à tous les stades du traitement des données, y compris lors de la collecte, adopter une approche des droits de l'Homme dès la conception (by design) et éviter tout biais potentiel, y compris les biais non intentionnels ou cachés, ainsi que les risques de discrimination ou d'autres effets négatifs sur les droits de l'Homme et libertés fondamentales des personnes concernées »*.
Il est essentiel que l'humain reste au centre des décisions lorsqu'il s'agit de sa protection.
Pour qu'ils puissent prendre des décisions équitables et cohérentes, exemptes de préjugés et de discrimination, il est indispensable que les algorithmes reçoivent des informations fiables et neutres.
Une analyse approfondie des biais et de leur impact sur les applications au monde réel devrait précéder le déploiement des outils d'automatisation. Tout comme il est obligatoire d'effectuer une analyse d'impact (AIPD) pour certaines opérations de traitement de données susceptibles d'engendrer un risque élevé pour les droits et libertés des personnes (RGPD, art. 35).

▶ **Alain Bensoussan**

1 Norme ISO/CEI TR 24027:2021 Biais dans les systèmes d'IA et prise de décision assistée par l'IA, novembre 2021.
2 Voir notre ouvrage, *Algorithmes et droit*, Editions Lexing, Mars 2023.
3 *Bias in algorithms – Artificial intelligence and discrimination*, FRA Report 2022.
4 *Algorithmes et discriminations*, *Defenseurdesdroits.fr*, 31 mai 2020.
5 Comité de la Convention 108+ du Conseil de l'Europe, Lignes directrices sur l'IA et la protection des données, 25 janvier 2019.