Communiqué de presse de la FRA
Vienne, le 8 décembre 2022

EMBARGO : 8 décembre 2022, 06 h 00 HEC

Tester les algorithmes pour biais avant application pour éviter la discrimination

L'intelligence artificielle est partout et concerne tout le monde. Un nouveau rapport de l'Agence des droits fondamentaux de l'UE (FRA) examine l'utilisation de l'intelligence artificielle dans la police prédictive et la détection des discours offensants. Le rapport démontre comment les préjugés dans les algorithmes apparaissent et comment ils peuvent affecter la vie de la population. C'est la première fois que la FRA fournit des éléments concrets sur la manière dont les biais se forment. L'Agence invite les décideurs politiques à veiller à ce que l'Intelligence Artificielle (IA) soit testée pour détecter les préjugés qui pourraient conduire à la discrimination.

« *Des algorithmes bien conçus et testés peuvent apporter de nombreuses améliorations. Mais sans vérifications appropriées, les développeurs et les utilisateurs ont de grandes chances d'avoir une incidence négative sur la vie des citoyens*, déclare le directeur de la FRA, Michael O'Flaherty. *Il n'existe pas de solution rapide. Nous avons besoin d'un système pour évaluer et atténuer les biais avant et pendant l'utilisation d'algorithmes afin de protéger les personnes contre la discrimination.* »

Pour son nouveau rapport intitulé « Bias in algorithms – Artificial intelligence and discrimination » (« Les biais dans les algorithmes - intelligence artificielle et discriminations »), la FRA a élaboré deux études de cas pour rechercher de potentiels biais dans les algorithmes :

1) La **police prédictive** montre comment les biais peuvent s'amplifier au fil du temps, ce qui peut conduire à des mesures de police discriminatoires. Si la police ne se rend que dans une seule zone en se basant sur des prédictions influencées par des casiers judiciaires biaisés, elle détectera des infractions essentiellement dans cette zone. Ce processus crée ce qu'on appelle une boucle de rétroaction. Dans ce cas, les algorithmes influencent les algorithmes, en renforçant ou en créant des pratiques discriminatoires susceptibles de cibler de manière disproportionnée des minorités ethniques.
2) La **détection des discours offensants** analyse les biais ethniques et de genre dans les systèmes de détection des discours offensants. Elle montre que les outils utilisés pour détecter les discours de haine en ligne peuvent conduire à des résultats biaisés. Les algorithmes peuvent même signaler comme étant des propos offensants des expressions inoffensives telles que « Je suis musulman » ou « Je suis juif ». Il existe également des biais de genre dans des langues genrées comme l'allemand ou l'italien, qui peuvent conduire à une inégalité d'accès à des services en ligne pour des motifs potentiellement discriminatoires.

Ces résultats nécessitent une évaluation complète des algorithmes. La FRA invite donc les institutions et les pays de l'UE à :

- **Rechercher les biais** - les algorithmes peuvent être biaisés ou bien développer des biais au fil du temps, ce qui peut conduire à des discriminations. Rechercher les

biais avant et pendant l'utilisation, surtout dans le cadre de prises de décisions automatisées, réduit ce risque.

- **Fournir des indications en matière de données sensibles** - afin d'évaluer d'éventuelles discriminations, des données sur les caractéristiques protégées (comme l'ethnicité ou le genre) peuvent être nécessaires. Cela nécessite des orientations sur les cas dans lesquels la collecte de telles données est autorisée. Elle doit être justifiée, nécessaire et assortie de garanties efficaces.
- **Évaluer les biais ethniques et de genre** - les biais ethniques et de genre dans la détection des discours et les modèles de prédiction sont très présents. Ils doivent être évalués au cas par cas. De telles évaluations doivent se fonder sur des preuves et être mises à disposition des organes de contrôle et du public.
- **Prendre en considération tous les motifs de discrimination** - les biais sont présents dans de nombreux domaines. Par conséquent, tous les motifs de discrimination interdits, comme le sexe, la religion ou l'origine ethnique, doivent être évalués. Différentes réglementations de l'UE, existantes ou en projet, sont nécessaires pour lutter contre les discriminations dues aux algorithmes, y compris la proposition de directive sur l'égalité de traitement.
- **S'efforcer d'accroître la diversité linguistique** - les modèles de détection des discours tendent à se concentrer sur l'anglais. Il est nécessaire de promouvoir et de financer la recherche sur d'autres langues afin de favoriser l'utilisation d'outils linguistiques dûment testés, documentés et actualisés pour toutes les langues officielles de l'UE.
- **Améliorer l'accès à une surveillance fondée sur des preuves** - ce qui se trouve derrière les systèmes d'intelligence artificielle peut être largement inconnu. Une surveillance efficace nécessite un meilleur accès aux données et aux infrastructures de données afin d'identifier et de combattre le risque des biais dans les algorithmes.

Ces conclusions visent à contribuer aux évolutions réglementaires en cours en informant les décideurs politiques, les professionnels des droits de l'homme, l'industrie technologique et le public sur le risque de biais dans l'intelligence artificielle.

.

Elles s'inscrivent dans le cadre des travaux de la FRA sur l'intelligence artificielle et les mégadonnées. Les recherches précédentes ont identifié des failles dans l'utilisation de l'IA et ont appelé l'UE et les États membres à s'assurer que l'IA protège tous les droits fondamentaux.

Pour de plus amples informations, veuillez contacter : media@fra.europa.eu / Tél. : +43 1 580 30 653