Metadata

- Id: EU.AI4T.O1.M3.3.3t
- Title:
- Type: text
- Description:
- Subject: Artificial Intelligence for and by Teachers
- Authors:
  - AI4T
- Licence: CC BY 4.0
- Date: 2022-11-15

D'OÙ VIENT LE RISQUE ?

Dans son étude sur l'intelligence artificielle [1], le Service de Recherche du Parlement Européen (Scientific Foresight Unit - STOA) a déclaré : "*Il est important de noter que les algorithmes d'IA ne peuvent pas être objectifs parce que, tout comme les gens, au cours de leur formation, ils développent une façon de donner un sens à ce qu'ils ont vu auparavant, et utilisent cette "vision du monde" pour catégoriser les nouvelles situations qui leur sont présentées.*"

Voyons d'où vient la subjectivité d'une IA et quels sont les risques associés.

LE BIAIS DANS LES DONNÉES ET DANS LES ALGORITHMES

Comme pour tout système numérique, les données utilisées dans les plateformes basées sur de l'IA proviennent de différentes sources et ont des formats multiples. Elles sont porteuses de différents types de biais[2]. Les biais des données sont principalement d'ordre statistique. Nous allons en énumérer quelques-uns.

- **Le biais de données** est généralement présent dans les valeurs des données. Par exemple, c'est le cas d'un algorithme de recrutement entraîné sur une base de données dans laquelle les hommes sont surreprésentés qui exclura les femmes.
- **Le biais de stéréotype** est une tendance à agir en référence au groupe social auquel on appartient. Par exemple, une étude montre que les femmes ont tendance à cliquer sur les offres d'emploi qu'elles pensent être plus faciles à obtenir en tant que femme.
- **Le biais de la variable omise** (biais de modélisation ou de codage) est un biais dû à la difficulté de représenter ou de coder un facteur dans les données. Par exemple, comme il est difficile de trouver des critères factuels pour mesurer l'intelligence émotionnelle, cette dimension est absente des algorithmes de recrutement.

- **Le biais de sélection** est dû aux caractéristiques de l'échantillon sélectionné pour tirer des conclusions. Par exemple, une banque utilisera des données internes pour établir un score de crédit, en se concentrant sur les personnes qui ont obtenu ou non un prêt, mais en ignorant celles qui n'ont jamais eu besoin d'emprunter, etc.

Le biais algorithmique est principalement une question de raisonnement. Ce biais est introduit par les ingénieurs en IA, délibérément ou non.

L'étude du Service de recherche du Parlement européen, mentionnée précédemment, donne deux exemples concrets : "*Considérez un algorithme d'IA symbolique pour examiner les demandes d'emploi. Il pourrait évaluer les candidats en leur attribuant des notes uniquement sur la base de leur formation et de leur expérience. Pourtant, s'il ne parvient pas à prendre en compte des facteurs tels que le congé de maternité ou à reconnaître de manière appropriée les études suivies dans des établissements étrangers, comme le feraient des comités de sélection humains, l'algorithme pourrait pratiquer une discrimination à l'encontre des femmes et des candidats étrangers.*"

"*Considérons maintenant un outil d'IA similaire dans le cadre du paradigme ML (Machine Learning). Ces algorithmes trouvent leurs propres moyens d'identifier les types de candidats sélectionnés dans leurs données d'apprentissage. Lorsqu'il existe un historique de biais structurels dans ces sélections - par exemple la discrimination raciale - l'algorithme peut les apprendre. Même lorsque les données relatives à la nationalité ou à l'origine ethnique sont supprimées des données, le ML est capable de trouver des substituts pour des modèles sous-jacents dans d'autres données telles que les langues, les codes postaux ou les écoles qui peuvent être de bons prédicteurs de l'origine ethnique.*"

LES TROIS FACETTES DU RISQUE ALGORITHMIQUE

Le risque algorithmique peut être caractérisé de trois façons[3].

- Tout d'abord, il y a l' **enfermement algorithmique**, qui peut également concerner les opinions, les connaissances culturelles ou encore les pratiques commerciales. En effet, les algorithmes confrontent l'internaute au même contenu, en fonction de son profil et des paramètres intégrés, malgré le respect du principe d'équité. C'est le cas sur les sites de recommandation d'actualités de certains réseaux sociaux ou de produits comme ceux des entreprises de vente en ligne.
- La deuxième facette du risque algorithmique est liée à **la maîtrise de tous les aspects de la vie d'un individu**, de la régulation de l'information à destination des investisseurs jusqu'à ses habitudes alimentaires, ses hobbies, ou encore son état de santé. Ce traçage de l'individu suggère une forme de surveillance qui contrevient à l'essence même de la liberté individuelle.
- La troisième est liée à la **potentielle violation des droits fondamentaux**. En particulier, la discrimination algorithmique définie comme un traitement défavorable ou inégal, par

rapport à d'autres personnes ou d'autres situations égales ou similaires, fondé sur un motif expressément interdit par la loi. Cela englobe l'étude de l'équité (*fairness*) des algorithmes de classement (tri de personnes cherchant un emploi en ligne), de recommandation et d'apprentissage prédictif. Le problème des biais discriminatoires induits par les algorithmes concerne plusieurs domaines tels que l'embauche en ligne, les décisions de justice, les décisions des patrouilles de police, ou les admissions scolaires.

COMMENT GÉRER LES RISQUES LIÉS AUX DONNÉES ET AUX ALGORITHMES ?

Pour R. Schwartz & al.[4], "*Les biais ne sont ni nouveaux ni propres à l'IA et il n'est pas possible d'atteindre un risque zéro de biais dans un système d'IA*".
En attendant, reconnaître que les agents d'IA sont intrinsèquement subjectifs est une condition préalable cruciale pour garantir qu'ils ne sont appliqués qu'à des tâches pour lesquelles ils sont bien adaptés.

L'étude de l'EPRS se conclut par plusieurs recommandations lors de l'utilisation d'applications basées sur l'IA :

- Comprendre les préjugés et la subjectivité
- Éviter les applications qui dépassent les capacités de l'IA
- Éviter les applications ayant des effets indésirables
- Maintenir l'autonomie humaine
- Rechercher des solutions aux problèmes et non des problèmes aux solutions
- Réfléchir à ce que nous voulons vraiment de l'IA

---

1. Etude en anglais: Artificial intelligence: How does it work, why does it matter, and what can we do about it ? - Philip Boucher, Scientific Foresight Unit (STOA) - ISBN: 978-92-846-6770-3 - Union Européenne, 2020 ↩
2. Algorithmes, données et biais : quelles politiques publiques ?, Anne Bouverot, Thierry Delaporte, 2019 ↩
3. D'où vient le risque ? Des données et des algorithmes - Serge Abiteboul, Thierry Viéville, 2020 ↩
4. Article en anglais : "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" - Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, NIST Special Publication 1270 , 2022 ↩