# «Les algorithmes ont un impact direct sur nos vies, il est donc essentiel qu'ils soient équitables»

**Yara Abu Awad est Senior Data Scientist et cheffe de groupe au Centre de compétences en science des données de l'Office fédéral de la statistique OFS. Dans cet entretien, elle nous éclaire sur le fonctionnement complexe des algorithmes et les solutions techniques pour corriger ou détecter les biais discriminatoires qu'ils peuvent contenir.**

Entretien : Marsali Kälin

**Yara Abu Awad, en quoi consiste votre travail ?**
Je supervise une équipe composée d'un Data Scientist[1] et de Data Engineers[2] dans l'exécution de projets au sein de l'Office fédéral de la statistique OFS et d'autres offices fédéraux. Dans ce contexte, je conseille les offices en matière d'équité algorithmique et participe à des groupes de travail de l'Organisation des Nations Unies sur l'intelligence artificielle.

**La question des biais algorithmiques est-elle récurrente dans votre travail ?**
Oui, ce sujet a suscité beaucoup d'attention ces dernières années et a donc été plus présent dans mon travail. Les gouvernements suivent de près les progrès technologiques en matière d'intelligence artificielle et cherchent à savoir comment l'exploiter de manière équitable, c'est-à-dire sans introduire ou renforcer des biais, qu'ils soient sexistes, racistes, liés à l'orientation sexuelle, à l'identité de genre, à la religion, à l'âge, ou encore à la nationalité. Les offices fédéraux nous consultent car ces questions sont très complexes et, malgré les nombreuses recommandations, il n'existe aucune formule magique applicable à toutes les situations. Il faut toujours analyser le cas particulier en fonction de ses données, de sa problématique et de son contexte sociopolitique.

**Qu'est-ce qu'un algorithme ?**
Un algorithme est un ensemble de règles logiques permettant de traiter des données afin d'atteindre un objectif. La recette de cuisine est une métaphore souvent utilisée pour expliquer un algorithme : pour préparer un gâteau (l'objectif), il faut réunir la liste d'ingrédients (les données) et suivre minutieusement les instructions dans l'ordre (règles logiques). La complexité des algorithmes est très variable, allant d'un simple ensemble de règles pouvant être écrites par un être humain (comme l'exemple de la recette), à des modèles complexes composés de milliards de paramètres qui nécessitent du temps et des ressources informatiques pour être construits (l'intelligence artificielle). Les algorithmes complexes suivent également un ensemble de règles, mais celles-ci sont mathématiques, prédéfinies et ne peuvent que partiellement être ajustées par des êtres humains. Les algorithmes sont omniprésents dans notre société et leurs utilisations variées. Les algorithmes de recommandation, par exemple, déterminent quels contenus apparaissent sur nos réseaux sociaux en fonction de nos historiques de visionnage. Peut-être moins visibles, mais tout aussi influents, les algorithmes d'aide à la décision sont utilisés dans de nombreux secteurs : les ressources humaines s'en servent pour recruter du personnel, les tribunaux aux États-Unis les utilisent pour évaluer la probabilité qu'une personne prévenue récidive et la médecine les emploie pour prescrire des traitements. Les algorithmes ont un impact direct sur nos vies, il est donc essentiel qu'ils soient équitables.

**Comment les algorithmes sont-ils entraînés ?**
Les algorithmes les plus puissants et les plus fréquemment utilisés se basent sur l'apprentissage automatique. Dans ce cas, les règles ne sont pas connues, mais «apprises» par le modèle. Un modèle est entraîné en intégrant d'innombrables exemples, comme des images, des articles de presse, des diagnostics médicaux ou encore des textes de loi, lui permettant par la suite de prédire le résultat pour une nouvelle donnée. La qualité des données d'entraînement est donc extrêmement importante. Si les données contiennent des biais, ceux-ci influenceront les prises de décision du

Yara Abu Awad

«Les gouvernements suivent de près les progrès technologiques en matière d'intelligence artificielle.»

« Les algorithmes les plus puissants et les plus fréquemment utilisés se basent sur l'apprentissage automatique. »

modèle. Les biais peuvent avoir différentes sources. On parle de biais statistiques lorsque les données ne sont pas représentatives de la population ou quand il existe un déséquilibre dans les groupes représentés.[3] On parle de biais historiques ou sociétaux quand les données utilisées contiennent elles-mêmes des schémas de décision discriminatoires. Des biais inconscients peuvent également être introduits par la personne qui entraîne le modèle, selon les choix qu'elle opère, ou par les utilisateurs et utilisatrices du modèle, lorsque celui-ci est ajusté en fonction de l'interaction avec qui l'utilise. C'est pourquoi il est important d'évaluer la performance des modèles algorithmiques et de mesurer leur équité.

### Qu'est-ce qu'un algorithme équitable ?

Dans le contexte de la prise de décision, l'équité est l'absence de tout préjugé ou favoritisme à l'égard d'un individu ou d'un groupe d'individus sur la base de ses caractéristiques. Ainsi, un algorithme inéquitable est un algorithme dont les décisions sont biaisées en faveur d'un groupe particulier.[4] Ce n'est toutefois pas la technologie elle-même qui est biaisée, mais la manière dont elle est conçue et utilisée par les êtres humains. Les biais présents dans les algorithmes reflètent et reproduisent les biais présents dans la société. Aux États-Unis par exemple, où le racisme a joué et joue encore un rôle tristement important, des études ont montré qu'il existait des biais dans la manière dont une équation algorithmique était utilisée pour mesurer la fonction rénale des patientes et patients. L'équation utilisée[5] calculait un indice déterminant la bonne santé du rein. Si celui-ci était bas, il fallait administrer un traitement, s'il était haut, on considérait au contraire que le rein était en bonne santé. Il est apparu que l'algorithme parvenait à un indice plus bas pour une personne blanche dans le même état de santé qu'une personne noire, et que la personne blanche dépassait ainsi la personne noire dans l'accès aux traitements. Dans ce cas, le biais était d'ordre « historique ou sociétal », car il provenait des données médicales sur lesquelles l'algorithme était basé. En effet, à cause du racisme systémique et des conditions socioéconomiques des personnes noires aux États-Unis, celles-ci attendent plus longtemps avant d'aller consulter et sont donc statistiquement plus malades que les personnes blanches au moment où le diagnostic est posé et le traitement administré. En se basant sur ces diagnostics, l'algorithme parvenait à un indice d'état de santé biaisé pour les personnes noires, et celles-ci devaient attendre d'être plus malades que les blanches pour recevoir un traitement.[6]

### Comment fait-on pour mesurer l'équité d'un algorithme ?

Il existe des outils comme TrustyAI, Aequitas ou AI Fairness 360 qui permettent d'évaluer la performance de son modèle et de mesurer l'équité de ses prédictions. On utilise également des métriques comme la parité démographique conditionnelle qui permet de comparer la probabilité que deux individus de groupes différents, mais présentant les mêmes qualités, soient choisis. Dans le contexte d'un recrutement, cela permet par exemple de mesurer s'il y a un écart significatif entre la probabilité qu'un homme soit choisi et la probabilité qu'une femme soit choisie. La technologie permettant de tracer les biais et d'expliquer les modèles algorithmiques progresse et il devient toujours plus aisé d'identifier quel paramètre ou quelle donnée a joué un rôle dans une décision. Malgré cela, il reste toujours des zones d'ombre et des décisions inexplicables.

**Une fois qu'un biais a été détecté, est-il possible de le corriger ?**

Parfois. On peut par exemple agir directement sur l'algorithme en modifiant les hyperparamètres, c'est-à-dire les variables qui régissent le processus d'entraînement lui-même. Il est également possible de rééquilibrer les données ou de les pondérer, à savoir augmenter ou diminuer le poids d'une valeur. Ces manipulations restent toutefois complexes. Il ne suffit par exemple pas d'enlever l'information liée au genre des candidates et candidats pour rendre un processus de recrutement égalitaire. Prenons l'exemple d'Amazon.[7] Basé sur des années de recrutement ayant privilégié les candidatures masculines, l'algorithme de recrutement d'Amazon avait « appris » que les hommes étaient de meilleurs candidats. Réalisant que l'algorithme était biaisé, Amazon a décidé de le corriger en supprimant la donnée liée au genre des candidats et candidates. Seulement, l'algorithme a continué d'identifier les candidatures féminines via des données alternatives (appelées données « proxy »), comme la mention d'une formation suivie dans une école réservée aux femmes ou l'adhésion à un club de sport féminin, et a donc continué à dévaloriser les candidatures féminines dans ses résultats.

**Quelles sont les limites des modèles algorithmiques en matière d'égalité ?**

Les modèles sont entachés des mêmes préjugés que ceux existant dans notre société, soit parce que les données d'entraînement sont biaisées, soit parce que la manière dont nous construisons les modèles est biaisée. Les algorithmes peuvent être ajustés pour fournir des décisions moins biaisées et pour inciter les utilisateurs et utilisatrices à faire de même, mais cela doit être pris en compte dès l'entraînement et le déploiement de l'algorithme. À mon sens, il est dangereux de croire que la technologie seule peut éliminer les discriminations. Il est important de sensibiliser les personnes qui conçoivent et utilisent les algorithmes et de placer les questions d'explicabilité et d'équité au cœur de l'évaluation des algorithmes avant leur déploiement.

**Marsali Kälin** est médiatrice culturelle. Diplômée d'un Master en littérature comparée et en études genre, elle a effectué un stage universitaire au secrétariat de la CFQF et rédige régulièrement des entretiens pour la revue « Questions au féminin ».

« Il est dangereux de croire que la technologie seule peut éliminer les discriminations. »

Notes

1 En français : scientifique de données senior.
2 En français : ingénieur·e·s de données.
3 Certains algorithmes vont par exemple utiliser des données issues de réseaux sociaux comme X (anciennement Twitter). Or, les individus de 25 ans seront surreprésentés dans ces données – proportionnellement à la population globale – et ceux de 75 ans et plus, presque totalement absents du réseau social, seront sous-représentés.
4 Définition issue de : Mehrabi, Ninareh et al. : A Survey on Bias and Fairness in Machine Learning. In : ACM Comput. Surv. 54 (6), Article 115, 2022, pp. 1–35. https://doi.org/10.1145/3457607.
5 GFR = 141 * min(Scr/κ,1)α * max(Scr/κ, 1)-1.209 * 0.993Age * 1.018 [if female] * 1.159 [if black].
6 Schmidt et al. : Separate and Unequal : Race-Based Algorithms and Implications for Nephrology, In : JASN 32 (3), 2021, pp. 529–533.
7 Dastin, Jeffrey : Insight. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.com, 2018. www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (consulté le : 27.06.2024).

## Abstracts

### «Algorithmen haben direkte Auswirkungen auf unser Leben, deshalb müssen sie zwingend fair sein»

Algorithmen, insbesondere solche, die die Entscheidungsfindung unterstützen, sind allgegenwärtig und haben erhebliche Auswirkungen auf unser Leben. Sie widerspiegeln die in der Gesellschaft vorhandenen Vorurteile, weshalb die Gefahr besteht, dass sie Diskriminierungen von ohnehin benachteiligten Bevölkerungsgruppen reproduzieren und verstärken. Gemäss **Yara Abu Awad,** Senior Data Scientist beim Bundesamt für Statistik BFS, gibt es zwar Tools, um solche Verzerrungen zu erkennen und zu korrigieren. Der Umgang mit Algorithmen bleibt aber komplex und wirft zahlreiche ethische Fragen auf.

### «Gli algoritmi hanno un impatto diretto sulle nostre vite per cui è fondamentale che siano equi»

Gli algoritmi, e in particolare quelli di supporto decisionale, sono onnipresenti e hanno un impatto notevole sulle nostre vite. Portatori degli stessi bias presenti nella società, rischiano di riprodurre e rafforzare le discriminazioni nei confronti dei gruppi già svantaggiati. Secondo **Yara Abu Awad,** Senior Data Scientist presso l'Ufficio federale di statistica UFS, sebbene esistano metodi e strumenti per rilevare o correggere questi bias, la manipolazione degli algoritmi rimane complessa e solleva molti interrogativi di ordine etico.