Combattez les biais algorithmiques

Pendant longtemps nous avons cru - ou espéré - que remplacer des tâches humaines par des algorithmes allait nous affranchir de tous les défauts des humains. Et notamment de nos **biais** : fini le racisme, le sexisme, l'homophobie… Après tout, un algorithme c'est des mathématiques, alors c'est forcément **juste et impartial**, non?

Eh bien non, pas du tout. Ces dernières années des algorithmes discriminants sont apparus, entraînant des conséquences humaines et économiques **désastreuses**.

Mais d'abord, ça veut dire quoi qu'un algorithme "discrimine"?

La discrimination est définie dans le Droit français et européen comme toute distinction opérée entre les personnes sur le fondement de leurs différences, celles-ci étant définies par 21 critères (art 225-1 du Code Pénal) dits "**sensibles**".

Pour un algorithme, cette distinction peut se décliner de plusieurs façons :

- Lorsqu'un critère sensible influer de façon disproportionnée sur la décision de l'algorithme - on parle alors de *disparate impact* ;
- Lorsque le taux d'erreur de l'algorithme est significativement plus élevé pour un groupe d'individus ;
- Lorsque les taux de faux positifs et/ou de faux négatifs sont disproportionnés selon les groupes d'individus.

La discrimination par les algorithmes est un concept défini **mathématiquement**. Il est donc possible de détecter ce type de biais dans des intelligences artificielles.

Il existe deux **sources** de discrimination :

- Lorsqu'une injustice est déjà présente dans la société : en reproduisant un phénomène, l'algorithme va également en reproduire l'injustice, et dans la plupart des cas aggraver le biais sociétal ;
- Lorsque les données d'apprentissage ne reflètent pas la diversité de la société : l'algorithme ne sera pertinent que pour la population représentée, au détriment des autres.

Les données, seules responsables.

Que la discrimination soit préexistante ou induite par l'apprentissage, l'algorithme lui-même n'est pas en cause. Réseaux de neurones, *deep learning*, forêts aléatoires… quel que soit le modèle utilisé, l'algorithme ne fait que reconnaître et reproduire ce qu'on lui a donné à observer. La discrimination est contenue dans les données d'apprentissage.

On a des exemples?

Malheureusement on n'en manque pas.

En 2016, le journal d'investigation ProPublica met à jour une défaillance dans l'automatisation du système judiciaire américain. Le logiciel **COMPAS** était alors utilisé pour donner un score de risque de récidive aux détenus et ainsi aider les juges dans leurs prises de décision pour accorder la liberté conditionnelle. Cet algorithme s'est révélé être **raciste**, en étant bien plus sévère pour les détenus de couleur que pour les caucasiens.

Raciste mais légal

L'algorithme n'est pas discriminant selon la première définition, le *disparate impact*. Or cette notion mathématique est la seule pour laquelle il existe une **jurisprudence** aux Etats-Unis (*Ricci v DeStefano, 2009*). Avec l'algorithme COMPAS les détenus de couleur avaient un plus fort risque d'être injustement taggué à fort risque de récidive, tandis que les caucasiens à risque avaient tendance à être considérés comme inoffensifs.

Mathématiquement cela signifie que le taux de faux positifs était bien plus élevé pour les personnes de couleur, tandis que le taux de faux négatifs était bien plus faible pour les caucasiens. L'algorithme, toujours utilisé aujourd'hui, est donc raciste mais pas illégal aux Etats-Unis. Northpointe a tout de même fait face à un **scandale** de grande envergure et ouvert la voie à de nombreux travaux sur les biais algorithmiques.

En 2017 le MIT publie un rapport sur les performances des algorithmes de reconnaissance faciale et surtout leur incapacité à reconnaître des femmes de couleur. La conférence TED de Joy Buolamwini illustre magistralement la situation : l'algorithme de reconnaissance faciale ne détectait son visage que si… elle mettait un **masque blanc** !

Comment ça se fait?

L'algorithme a été entraîné à reconnaître des visages en apprenant à partir d'une base de données de photos. Si la base de données n'est pas diversifiée, l'algorithme ne saura reconnaître que le type de visage qu'il a déjà observé !

En 2018 Amazon, un des GAFA et une des plus grosses boîtes de tech au monde, admet avoir passé deux ans à développer un algorithme de recrutement qui s'est avéré **sexiste**. L'objectif était de sélectionner automatiquement les meilleurs CV des candidats. Or l'algorithme **disqualifiait automatiquement** toutes les candidatures féminines.

Cette liste n'est malheureusement **pas exhaustive**. Les exemples se multiplient dans tous les domaines, de l'élection de miss à l'obtention de prêt bancaire, en passant par la santé.

Qu'est ce qu'on peut faire ?

On entend dire que les algorithmes sont biaisés parce qu'ils sont codés par des hommes blancs ; que l'intelligence artificielle est le reflet de ceux qui la développent. **C'est faux !** Ce n'est pas parce que vous êtes un homme blanc que votre algorithme sera automatiquement raciste. Et à l'inverse ce n'est pas parce que vous êtes une femme que vous ne pouvez pas développer un algorithme sexiste. Vigilance, donc, mais pas désespoir !

Responsabilisons nos données

On l'a vu, les biais algorithmiques viennent des données d'apprentissage. A vous donc de vous assurer que ces données sont justes avant de les utiliser. Ne sautez pas l'étape d'**analyse descriptive** de vos données. Vérifiez qu'elles sont équilibrées et qu'elles ne contiennent pas déjà des biais.

Concrètement pour un jeu de données d'apprentissage supervisé vous pouvez **tester** les trois notions de biais : le disparate impact, le taux d'erreur par population, les taux de faux positifs et de faux négatifs par population. Travailler avec un jeu de données propre est la première étape pour des algorithmes justes et responsables.

Repensons la notion de performance

Cela peut choquer mais il faut savoir qu'un algorithme raciste est **performant** ! La notion de base de performance algorithmique se calcule de la façon suivante : on compte le nombre de fois où l'algorithme a bien reproduit le phénomène et on divise par le nombre total d'essais. Donc si l'algorithme se trompe, même **systématiquement**, pour une minorité, sa performance ne va pas significativement baisser.

Il faut donc repenser la notion de performance pour y intégrer ce qui fait les biais algorithmiques : le disparate impact, le taux d'erreur par population, les taux de faux positifs et de faux négatifs par population… et plus encore peut-être ?

Plutôt que de simple performance on parle aujourd'hui de **loyauté des algorithmes**. Il faut se poser la question : qu'est ce qu'on veut réellement de cet algorithme ? Et une fois qu'on a une réponse il faut pouvoir s'assurer que c'est bien ce que fait l'algorithme.

L'homme dans la boucle.

On dit souvent que l'intelligence artificielle n'est pas dangereuse parce qu'il y a "l'homme dans la boucle". "L'homme dans la boucle" c'est cette notion qui consiste à dire qu'il n'y a pas à responsabiliser l'innovation tant qu'il reste de l'humain dans le processus. "C'est un humain qui appuie sur le bouton pour envoyer le missile", "c'est un humain qui décide d'envoyer les gens en prison"... Mais quelle est réellement la place de l'homme lorsqu'une machine qu'on dit plus puissante, plus performante, intime une vérité sans nuance?

Comprenons nos algorithmes

La plupart des algorithmes de Machine Learning sont des systèmes dits "boîtes noires". C'est à dire qu'on se sait pas comment ils sont arrivés au résultat obtenu. Un domaine de recherche consiste à assurer la **transparence** et l'**explicabilité** des décisions. En cherchant à savoir comment l'algorithme a pris ses décisions, on espère pouvoir s'assurer qu'il n'y a pas de biais.

En attendant, peut être que les boîtes noires du machine learning ne sont pas des solutions acceptables pour certaines utilisations, notamment pour les algorithmes qui impactent la vie des individus.

En résumé

L'algorithme apprend à partir de données qu'on lui fournit; il ne fait qu'apprendre ce qu'on lui donne à voir du monde. Donc si il y a de la discrimination dans ses données d'apprentissage, ou si une population est sous-représentée, l'algorithme va **reproduire**, **amplifier** et parfois **créer** de toutes pièces de la **discrimination**.