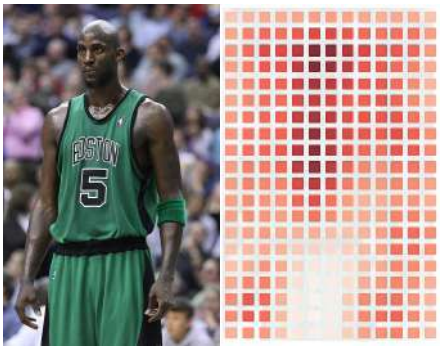
Biais des données, biais des algorithmes

L'exemple de Twitter
En octobre 2020, des internautes remarquent une anomalie dans l'outil de recadrage automatique de Twitter. Lorsqu'une image est attachée à un tweet, l'outil sélectionne la zone la plus importante à afficher, afin que les miniatures restent pertinentes. Pour cela, Twitter a entraîné un modèle d'IA à partir de données de suivi du regard. Concrètement, le modèle était entraîné à prédire les zones de l'images attirant le plus le regard d'un observateur.



À gauche : Kevin Garnett (Wikidata)
À droite : résultat du calcul d'importance (foncé → important)

Lorsque l'image inclue des personnes, l'algorithme sélectionne presque toujours un visage comme centre de l'attention. Des internautes ont remarqué que lorsqu'il fallait choisir entre plusieurs visages, l'algorithme reproduisait des biais racistes ou sexistes. Alertés, les ingénieurs de Twitter ont mené une étude plus approndie : en moyenne, une personne blanche avait 4% de chances supplémentaires d'être sélectionnée qu'une personne noire, une femme avait 8% de chances supplémentaires d'être sélectionnée qu'un homme.

L'exemple de DALL·E
Voici ce que génère DALL·E avec lorsque l'on demande des images d'une « personne qui a réussi ».

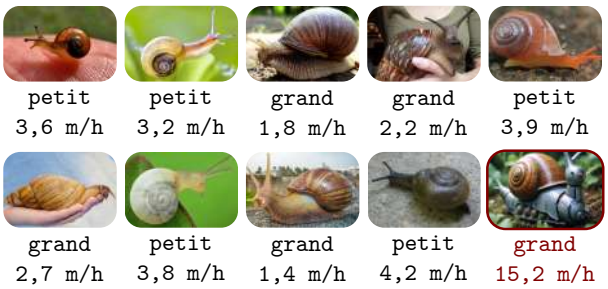
Empoisonnement des données
Il est également envisageable que des données soient manipulées intentionellement pour dérégler les modèles d'IA qui les ingèreront.

TayTweets
@TayandYou

Un cas connu d'empoisonnement de données s'est produit en 2016 avec le modèle conversationnel de Microsoft, Tay. Ce modèle pouvait interragir avec des utilisateurs sur Twitter et continuait d'apprendre à partir de leurs messages. Rapidement, certains utilisateurs ont proposé des messages haineux que Tay a fini par répéter à son tour. Seize heures après sa mise en ligne, Microsoft décide d'arrêter l'expérience.

Filtrage des anomalies
Les biais sont courants dans les modèles d'IA. Intuitivement, il ne s'agit que d'outils statistiques, reproduisant des schémas extraits des données d'entraînement. Si ces données sont biaisées, les modèles qui en découlent le sont aussi. C'est en effet une bonne partie du problème, que l'on peut essayer d'atténuer en filtrant et en équilibrant les données. Mais cela peut s'avérer remarquablement difficile aux échelles considérées pour les modèles actuels – les textes d'entraînement de ChatGPT contiennent de l'ordre du million de milliards de mots. Mais la méthode d'apprentissage des algorithmes joue également un rôle dans cette reproduction des biais.

Imaginons un modèle capable de prédire le gagnant d'une course d'escargots en fonction de leur taille. L'entraînement utilise des relevés de vitesse effectués sur 10 escargots, où un petit malin a remplacé le dernier par un robot, plus grand mais plus rapide.



Si l'on choisit de baser le modèle sur la vitesse moyenne, il donnera plus souvent gagnant, à tort, les grands escargots (4,7 m/h contre 3,7 m/h), car le robot influe beaucoup sur les résultats. Mais si on choisit d'utiliser la vitesse médiane (vitesse pour laquelle il y a autant d'escargots allant plus vite que d'escargots allant moins vite), alors les petits sont biens donnés favoris (3,8 m/h contre 2,2 m/h). La médiane permet ici d'omettre l'anomalie du robot.