# de Paris 1 Panthéon-Sorbonne Avril 203

### Les biais de l'intelligence artificielle



Désignent les **préjugés** ou les **distorsions** dans les résultats produits par les systèmes d'IA.



## Biais de données

Les données utilisées pour entraîner l'IA sont déjà biaisées (erreurs, stéréotypes, etc).

Exemple: un système de reconnaissance faciale entraîné surtout sur des visages blancs fonctionne mal pour les visages noirs.

# Biais de sélection

Les données ne représentent pas toute la population cible

Exemple : une IA médicale moins performante sur les femmes si les données viennent majoritairement de patients masculins.



### Biais d'interaction

L'IA apprend de comportements biaisés des utilisateurs.

Exemple : un Chatbot qui adopte un langage inapproprié en apprenant des conversations en ligne non modérées

# Biais de confirmation

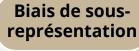
L'IA renforce les opinions existantes de l'utilisateur.

Exemple : un fil d'actualités qui montre toujours les mêmes opinions politiques, enfermant l'utilisateur dans une bulle.

### Biais d'algorithme

L'algorithme peut amplifier les biais en privilégiant certaines variables ou objectifs sans tenir compte de l'équité.

Exemple : un système de notation de crédit favorise les profils issus de quartiers riches.





Certains groupes sont trop peu présents dans les données.

Exemple : une IA qui détecte mal les symptômes de maladies rares chez les enfants car elle a peu de données pédiatriques.

Conséquences: discrimination, inégalités, décisions injustes dans certains domaines (emploi, justice pénale, services financiers...)

Solution : diversifier les données, tester l'équité, concevoir de façon éthique, surveiller en continu

