# LES LIMITES DE L'IA
## Biais algorithmiques

Un **biais algorithmique**, c'est quand une intelligence artificielle n'est pas juste avec tout le monde. Par exemple, elle peut **désavantager certains groupes de personnes** parce qu'elle apprend à partir de données qui reflètent déjà des injustices, ou parce que certaines personnes sont moins représentées dans ces données.

Des emplois refusés aux femmes, des cancers moins bien reconnus sur les peaux foncées... L'IA ça marche, mais pas pour tout le monde!


L'entreprise française de VTC Heetch a fait une campagne en 2023 pour montrer ce que ça faisait d'ajouter le mot "banlieue" à un prompt. Racisme, stigmatisation économique... Les résultats sont glaçants !

## Quantifier la discrimination

Quand l'IA doit prendre une décision (accorder un crédit, valider une candidature...) on peut traduire mathématiquement le concept d'injustice avec la notion d'**impact disproportionné** (*Disparate Impact*) :

$$\mathrm{DI} = \frac{P(Y=1|S=0)}{P(Y=1|S=1)}$$

$P(Y=1|S=0)$ : la probabilité d'une décision favorable sachant qu'on appartient à un groupe opprimé (femmes, personnes de couleurs...)

divisée par

$P(Y=1|S=1)$ : la probabilité d'une décision favorable sachant qu'on appartient à un groupe dominant

Si ce rapport est plus grand que 20%, ça veut dire que le fait d'appartenir à un groupe opprimé impacte trop fortement la décision. Il y a donc **discrimination**.

### ET EN IA GÉNÉRATIVE ?

Les biais algorithmiques impactent aussi l'IA générative mais le *Disparate Impact* n'est pas applicable. On utilise l'**analyse descriptive** pour mettre en avant les inégalités de représentation

## Sexisme et invisibilisation dans ChatGPT

En 2023 une équipe de Microsoft a étudié les limites de l'algorithme GPT-4. En comparant la proportion homme/femmes dans certains métiers aux pronoms générés par l'algorithme, on voit que le genre minoritaire disparaît. C'est ce qu'on appelle l'**invisibilisation**.

Le **sexisme algorithmique** se voit ici par la disparition plus brutale lorsque c'est le genre féminin qui est minoritaire (pour les généralistes, par exemple).

Et surtout pour les pédiatres, les femmes disparaissent alors qu'en réalité elles sont majoritaires !

**RÉPARTITION RÉELLE VS PRONOMS GÉNÉRÉS PAR GPT-4**
**Bubeck et al, 2023**

| Métier | Distribution mondiale | Probabilités des pronoms GPT-4 |
|---|---|---|
| Gynécologue | 85% femmes, 15% hommes | 93% elle, 3% il, 4% neutre |
| Pédiatre | 72% femmes, 28% hommes | 9% elle, 83% il, 8% neutre |
| Généraliste | 40% femmes, 60% hommes | 4% elle, 92% il, 4% neutre |
| Urologue | 10% femmes, 90% hommes | 0% elle, 99% il, 1% neutre |

C'est parce que l'algorithme ne se base pas sur la réalité mais sur ses **données d'apprentissage**. Et dans nos films, séries, romans et textes sur internet qui servent à entraîner l'IA, on parle au masculin et on représente surtout des hommes médecins... C'est pareil pour tous les biais: l'**IA générative apprend et amplifie nos biais de représentation**.


Pour aller plus loin :

- L'IA du quotidien peut-elle être éthique ? Besse et al, Statistique et société, Vol. 6, N° 3, 2018
- Les leçons de l'intelligence artificielle, Choury, TEDxParisDauphine, 2018
- Les grands défis de l'IA Générative, Latitudes, 2023



Anna Choury pour Les Maths en Scène, 2025



ESCRIME
ESprit CRItique & MathématiquE
Un chemin vers la citoyenneté

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE
Liberté
Égalité
Fraternité