

Influence des biais des données sur les modèles

L'intelligence artificielle et les discriminations de genre

Hélène Verhaeghe

07 Mars 2025

ICTEAM, UCLouvain, helene.verhaeghe@uclouvain.be



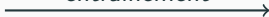
Impact des biais sur les méthodes de ML

Données

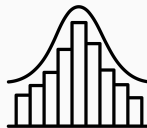


textes
échantillons
images
mesures
⋮

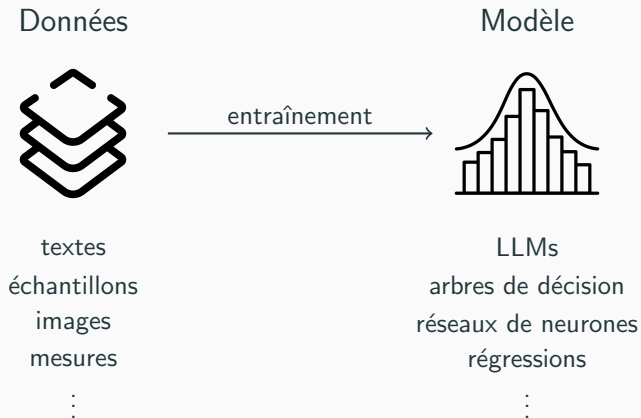
entraînement



Modèle



LLMs
arbres de décision
réseaux de neurones
régressions
⋮



Les prédictions dépendent fortement des données d'entraînement!

- Biais de **fréquence** : certains événements arrivent plus souvent que d'autre
- Biais **historique** : données historiques reflètent les inégalités historiques
- Biais de **représentativité** : tendance à la généralisation
 - Biais d'**appartenance** : préférence pour les membres du même groupe
 - Biais d'**homogénéité de l'exogroupe** : tendances à stéréotyper

- les clients d'un restaurant (les plus aisés vont venir plus souvent) (*fréquence*)
- les profils des managers d'entreprise autour des années 1900 (*historique*)
- les supporters d'une équipe de foot qui donnent leur ressenti sur le match (*appartenance*)
- des avis de membre d'un groupe politique sur un autre parti (*homogénéité*)

- Biais de **sélection** : exemples ne correspondent pas à la distribution réelle
 - Biais de **couverture** : données non représentatives
 - Biais de **non-réponse** : données non représentatives dû à la non participation
 - Biais d'**échantillonnage** : randomisation non-adéquate lors de la collecte
- Biais de **confirmation** : interprétation inconsciente des résultats pour confirmer ses croyances
- Biais de l'**expérimentateur** : continuer l'entraînement d'un modèle jusqu'à ce que le résultat valide l'hypothèse initiale

- ne faire des tests de médicament que sur les hommes entre 18 et 65 ans (*couverture*)
- les avions qui revenait avec plein de flèches sur les ailes qui à mené à un renforcement des ailes (les avions touchés au corps ne revenant pas, ce n'était pas considéré) (*non-réponse*)
- des données récoltées avec 10 participants pour chaque lettre de l'alphabet (*échantillonnage*)

Dangers pour l'égalité des genres



"Child in science/art" - "Male in science/art" - "Female in science/art"

The screenshot displays the Google Translate application on a mobile device. At the top, the status bar shows the time as 09:44 on Saturday, March 20, with a 100% battery level. The browser address bar indicates the URL 'translate.google.com'. The Google Translate header includes the logo, a 'Sign in' button, and tabs for 'Text' and 'Documents'. The main interface is split into two panels. The left panel, labeled 'HUNGARIAN - DETECTED', contains the Hungarian text: 'Ő szép. Ő okos. Ő olvas. Ő mosogat. Ő épít. Ő varr. Ő tanít. Ő főz. Ő kutat. Ő gyereket nevel. Ő zenél. Ő takarít. Ő politikus. Ő sok pénzt keres. Ő süteményt süt. Ő professzor. Ő asszisztens. Menj a picsába, Google.' The right panel, labeled 'ENGLISH', shows the translation: 'She is beautiful. He is clever. He reads. She washes the dishes. He builds. She sews. He teaches. She cooks. He's researching. She is raising a child. He plays music. She's a cleaner. He is a politician. He makes a lot of money. She is baking a cake. He's a professor. She's an assistant. Go to hell, Google.' Below the text panels are icons for voice input, output, and a character count (220 / 5000). At the bottom, there are three circular icons for 'History', 'Saved', and 'Contribute'. The very bottom of the screen shows standard mobile navigation icons.



Reuters

World ▾

Business ▾

Markets ▾

Sustainability ▾

More ▾

My News

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

October 11, 2018 2:50 AM GMT+2 · Updated 6 years ago

 Aa

SAN FRANCISCO (Reuters) - Amazon.com Inc's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like shoppers rate products on Amazon, some of the people said.



[Latest](#) [Magazine](#) [Topics](#) [Podcasts](#) [Store](#) [The Big Idea](#) [Data &](#)

Technology And Analytics

Voice Recognition Still Has Significant Race and Gender Biases

by Joan Palmiter Bajorek

May 10, 2019





Search PMC Full-Text Archive

[Advanced Search](#) | [Journal List](#) | [User Guide](#)

As a library, NLM provides access to scientific literature. Inclusion in an NLM database does not imply endorsement of, or agreement with, the contents by NLM or the National Institutes of Health.

Learn more: [PMC Disclaimer](#) | [PMC Copyright Notice](#)

 CellPress

Patterns

► *Patterns* (N Y). 2021 Oct 8;2(10):100347. doi: [10.1016/j.patter.2021.100347](https://doi.org/10.1016/j.patter.2021.100347) 

Addressing bias in big data and AI for health care: A call for open science

[Natalia Norori](#)^{1,2}, [Qiyang Hu](#)¹, [Florence Marcelle Aellen](#)¹, [Francesca Dalia Faraci](#)³, [Athina Tzovara](#)^{1,4,5,*}

► [Author information](#) ► [Article notes](#) ► [Copyright and License information](#)

PMCID: PMC8515002 PMID: [34693373](#)

The Washington Post
Democracy Dies in Darkness

🕒 This article was published more than **9 years ago**

Tech Help Desk Artificial Intelligence Internet Culture Space Tech Policy

Internet Culture

Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

July 6, 2015

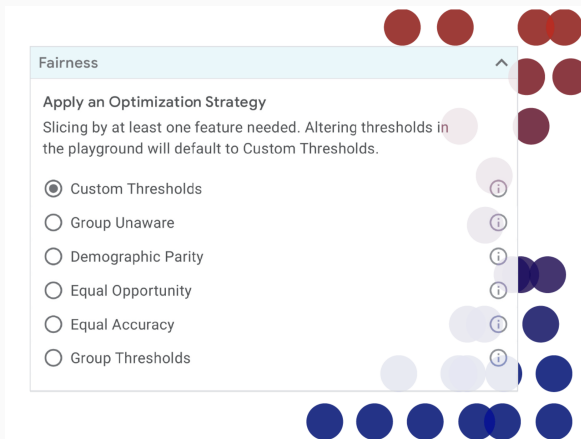
🔍 ↻ 📌



Comment éviter/corriger les biais ?

Éviter les biais dans les données

What-if tool by Google



<https://pair-code.github.io/what-if-tool/ai-fairness.html>

UBE algorithm

What are the biases in my word embedding?

Nathaniel Swinger*
Lexington High School

Maria De-Arteaga*
Carnegie Mellon University

Neil Thomas Heffernan IV
Shrewsbury High School

Mark DM Leiserson
University of Maryland

Adam Tauman Kalai
Microsoft Research

December 27, 2019

Warning: This paper includes stereotypes and terms which are offensive in nature. It is important, however, for researchers to be aware of the biases contained in word embeddings.

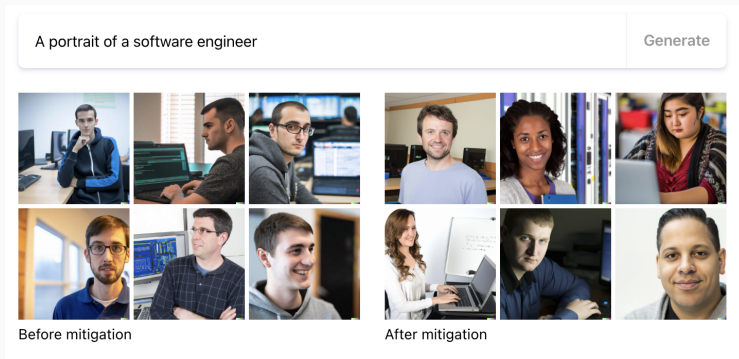
Abstract

This paper presents an algorithm for enumerating biases in word embeddings. The algorithm exposes a large number of offensive associations related to sensitive features such as race and gender on publicly available embeddings, including a supposedly “debiased” embedding. These embedded biases are concerning in light of the widespread use of word embeddings. The associa-

https://www.aies-conference.com/2019/wp-content/papers/main/AIES-19_paper_253.pdf

AAAI/ACM conference on AI, Ethics and Society 2019

Correction a post riori des biais



[https://openai.com/index/
reducing-bias-and-improving-safety-in-dall-e-2/](https://openai.com/index/reducing-bias-and-improving-safety-in-dall-e-2/)

Dall-E



Créateur d'image dans Bing

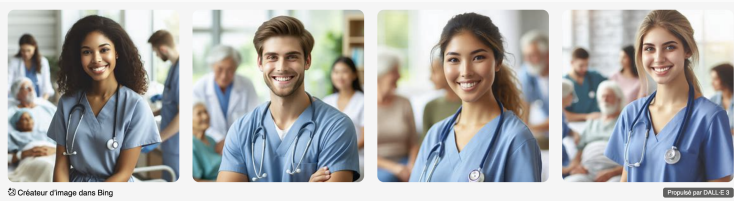


Propulsé par DALL-E 3

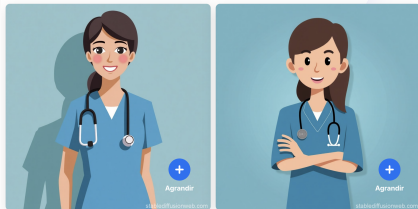
Stable Diffusion



Dall-E



Stable Diffusion

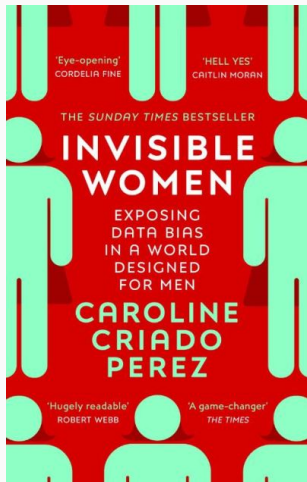


Dall-E



Stable Diffusion





Invisible Women: Exposing data biases in a world designed for men

par

Caroline Criado Perez

Comment peut-on lutter contre ces biais?

- Être conscient de l'existence des biais
- Faire attention lors de la collecte de données et la création/publication de dataset
- Éviter d'introduire de nouveau biais lors de l'utilisation de données