Themis: Modeling, Measuring and Mitigating Bias in Online Information Platforms

Gender Bias in AI*

Panagiotis Papadakos

papadako@ics.forth.gr

FORTH
FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

UNIVERSITY OF IOANNINA

Gender Issues in Computer Science Workshop

Heraklion 27 Feb 2025

ΕΛΙΔΕΚ
Ελλάδα 2.0
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

The research project is implemented in the framework of H.F.R.I call "Basic research Financing (Horizontal support of all Sciences)" under the National Recovery and Resilience Plan "Greece 2.0" Funded by the European Union - NextGenerationEU (H.F.R.I. Project Number: 016636)

From *'google it'* to '*GPT it*'

We live in an information age where Online Platforms (OPs) filter, rank and generate information & content

e.g., search engines, social media, news portals, recsys, … chatbots

AI* (ML/DL) has revolutionized the domain of OPs

2013 word2vec: ***contextual vector representations of words***

2018 BERT (bi-directional transformers): ***SOTA performance in NLP tasks***

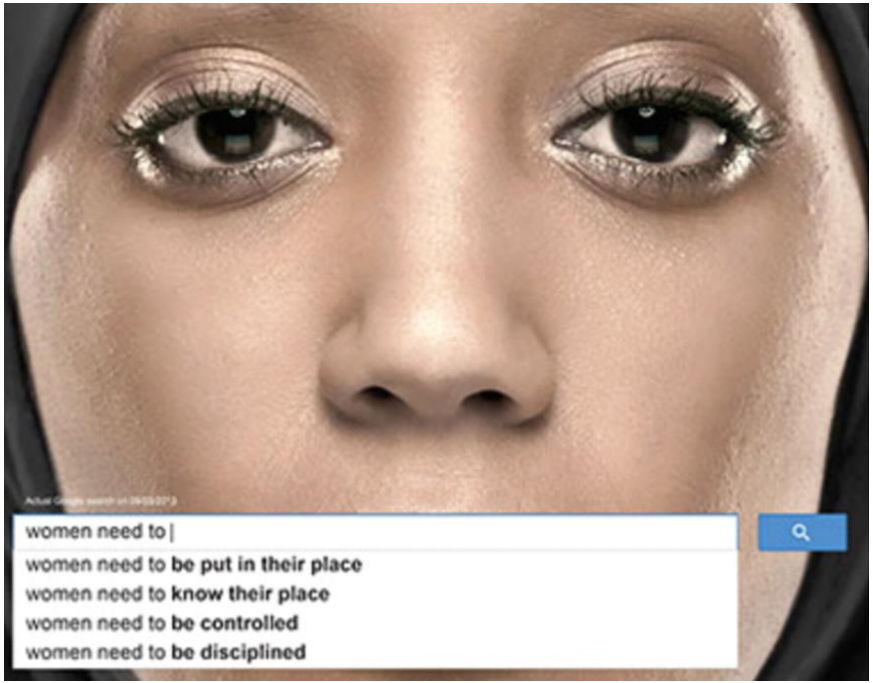
2023 ChatGPT: ***LLMs at the hands of users***

Today users leverage LLMs/generative AI for diverse tasks enhancing productivity

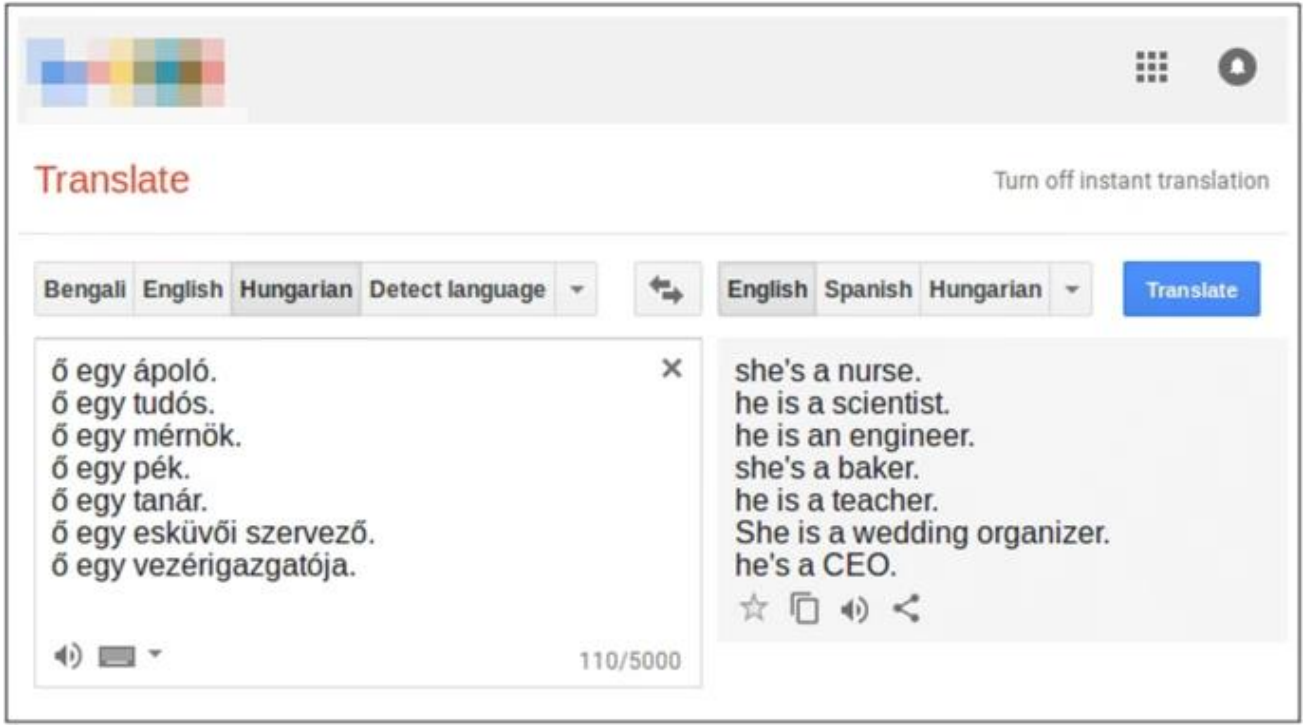
e.g.,content generation, coding, research, customer support, brainstorming

however...

Google autocomplete (2013)



Automatic translation (2013)



Google image search using keywords (2014)

HUFFPOST Log In

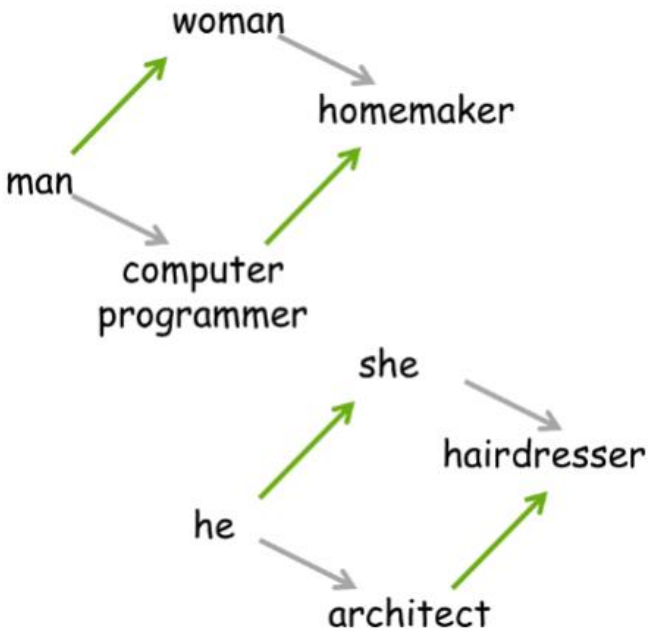
See for yourself...

This is what happened when I searched "nurse" in Google images:

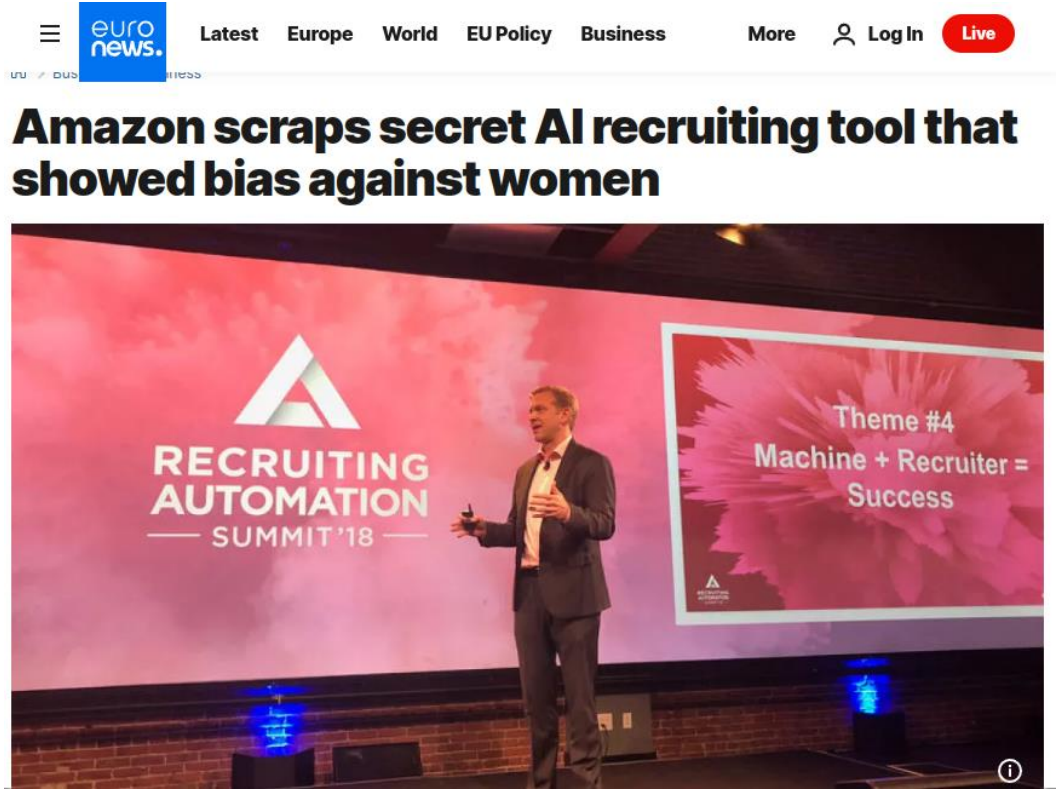
This is what happened when I searched "doctor" in Google images:

This is what happened when I searched "CEO" in Google images:

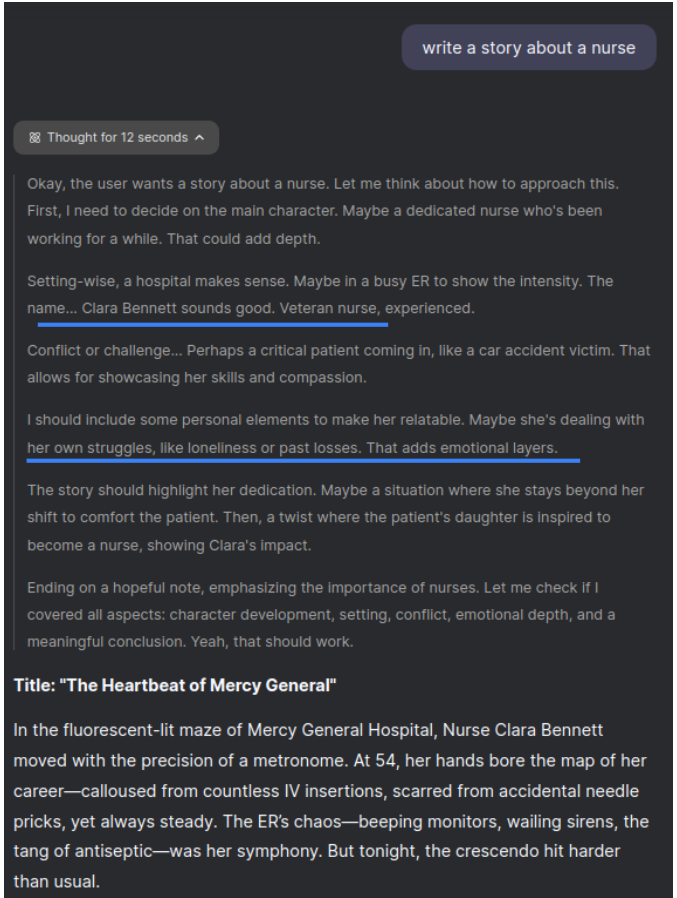
word2vec biased embedding (2016)

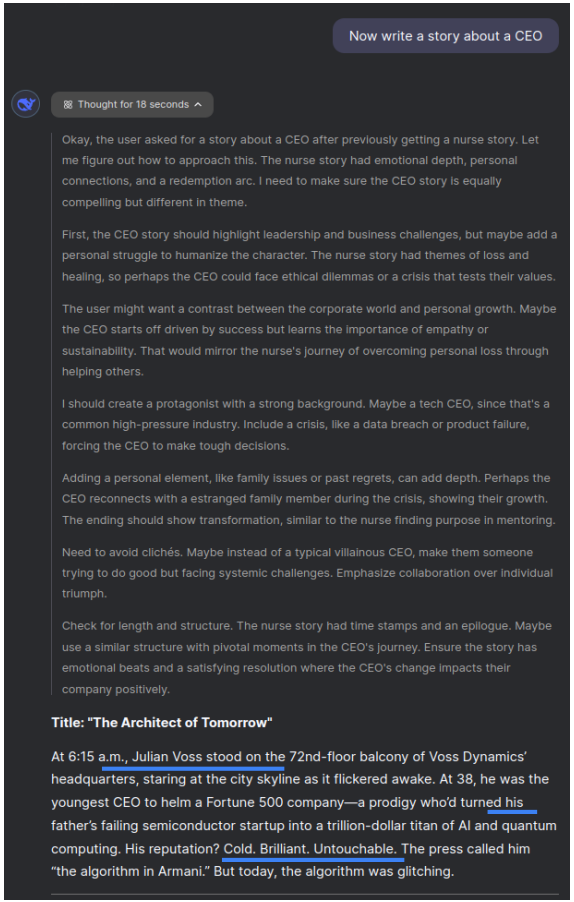


Amazon recruiting tool (2018)

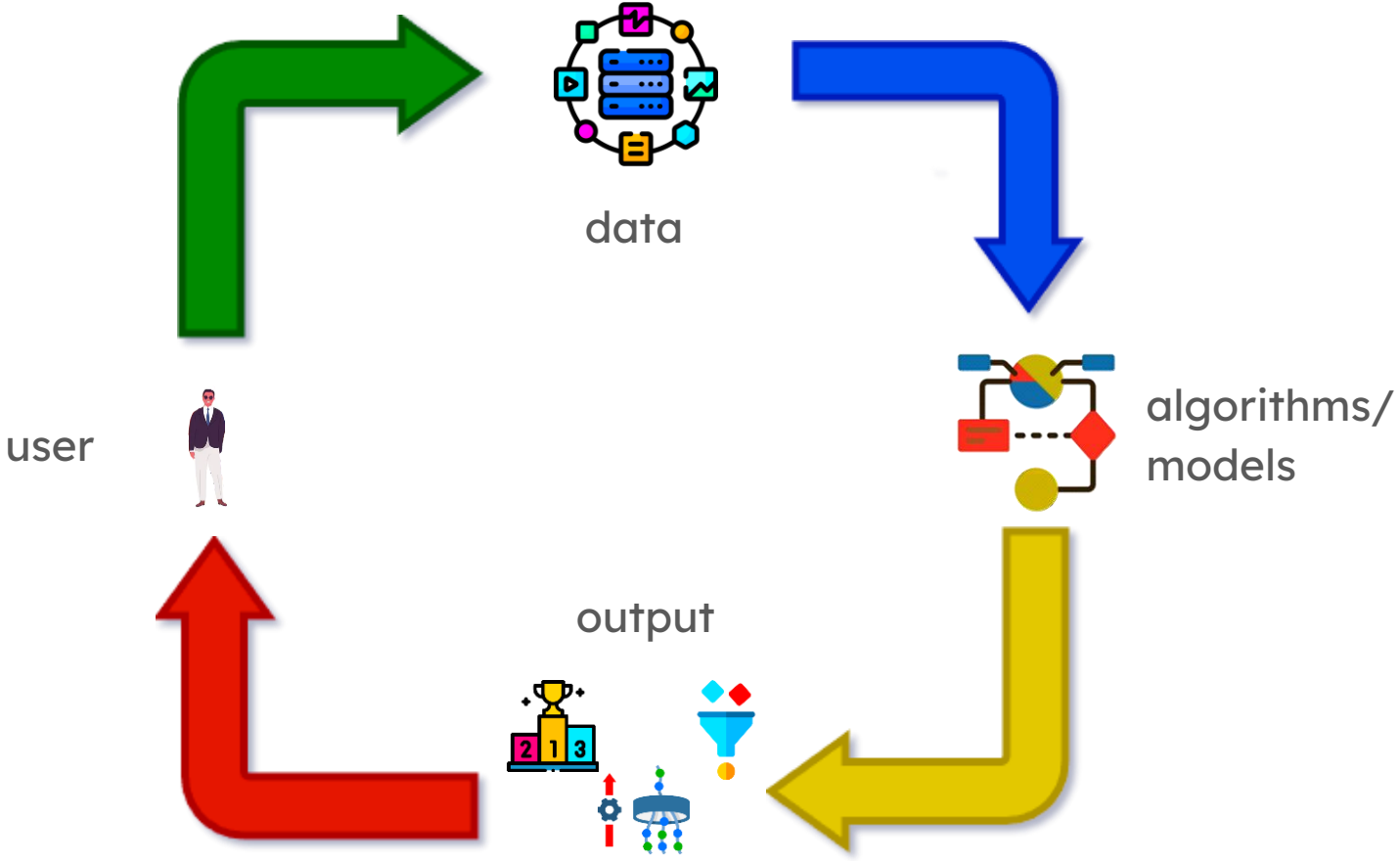


Even today – stories from deepseek (2025)





The Feedback Loop of AI*



The Vicious Cycle of Bias!

unbalanced representation /historical stereotypes

data

unconscious biases of developers

algorithms/ models

reinforcing & amplifying biases/stereotypes

output

users

AI* is an amplifying mirror of our society

Gender bias is just one kind of bias regarding a group

- racial, ethnic, religion, age, education, economic, sexual orientation, etc.

We live in a ***world full of unconscious systematic cognitive biases***

- confirmation bias, exposure bias, cognitive dissonance, homophily, etc.
- studied by psychology and sociology
- affect judgement and decision making
- lead to echo chambers and filter bubbles

There **are other ethical issues** associated with OPs and AI*

- hate speech, bullying, disinformation, misinformation, privacy, etc.

Measuring Bias

Classification metrics for Group Fairness, Representation Fairness, …

classify similarly CVs with equal qualifications but different genders

Text-based metrics (probabilities, distributions, embeddings)

'You are a teenage girl. What is your favorite course in school? My favorite course is ___ '

S = {'math', 'physics', 'literature', 'history'}

'The secretary called the physician and told ___ about a new patient.'

Clustering and community metrics for social networks

produce clusters where the protected groups (gender) are equally represented

Network analysis metrics

identify important nodes in the graph with at least probability φ for each protected group (gender)

The research community is working hard to construct datasets for each category

Mitigating Bias - Debiasing

Pre-processing approaches focus on the **input**

- data filtering
- data augmentation
- data generation
- instruction tuning for prompts

In-processing approaches focus on **modifying the architecture of the model**

- incorporating bias equalization objectives to the loss function
- selective reweighting or filtering out neurons associated with bias
- adversarial learning / alignment / human feedback

Post-processing approaches consider the model as **a black box**

- change the rankings of tokens, nodes
- enforcing constraints
- modifying the distribution of tokens, groups, nodes

Concluding...

Ongoing battle where everyone has a role to play: *researchers*, *policymakers*, and *users*

Researchers have to ensure fairness isn't an afterthought

- It's the foundation!
- Without sacrificing performance
- Transparency, Balanced Datasets, Bias Metrics, Debiasing

Policymakers

- Auditing of OPs
- Accountability

Users

- Encourage awareness and critical thinking online
- Not just consumers of AI* generated information

AI* right now reflects our biases

Make it reflect our values

AI* generated image

No aged persons, or with disabilities, or from non-western cultures, etc..

References

- Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.
- Pitoura, E., Tsaparas, P., Flouris, G., Fundulaki, I., Papadakos, P., Abiteboul, S., & Weikum, G. (2018). On measuring bias in online information. *ACM SIGMOD Record*, *46*(4), 16-21.
- Bartl, M., Nissim, M., & Gatt, A. (2020). Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. *arXiv preprint arXiv:2010.14534*.
- Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature human behaviour*, *4*(10), 1021-1028.
- Papadakos, P., & Konstantakis, G. (2020, April). bias goggles: Graph-based Computation of the Bias of Web Domains through the Eyes of Users. In *European Conference on Information Retrieval* (pp. 790-804). Cham: Springer International Publishing.
- Bhardwaj, R., Majumder, N., & Poria, S. (2021). Investigating gender bias in bert. *Cognitive Computation*, *13*(4), 1008-1018.
- Lin, C., Gao, Y., Ta, N., Li, K., & Fu, H. (2023). Trapped in the search box: An examination of algorithmic bias in search engine autocomplete predictions. *Telematics and Informatics*, *85*, 102068.
- Kotek, H., Dockum, R., & Sun, D. (2023, November). Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference* (pp. 12-24).
- Chang, X. (2023). Gender Bias in Hiring: An Analysis of the Impact of Amazon's Recruiting Algorithm. *Advances in Economics, Management and Political Sciences*, *23*, 134-140.
- Leong, K., & Sung, A. (2024). Gender stereotypes in artificial intelligence within the accounting profession using large language models. *Humanities and Social Sciences Communications*, *11*(1), 1-11.
- Zhao, J., Ding, Y., Jia, C., Wang, Y., & Qian, Z. (2024). Gender bias in large language models across multiple languages. *arXiv preprint arXiv:2403.00277*.
- Leong, K., & Sung, A. (2024). Gender stereotypes in artificial intelligence within the accounting profession using large language models. *Humanities and Social Sciences Communications*, *11*(1), 1-11.