Gender inequality in Italian-language Natural Language Processing datasets

Elia Schneider - Harrisburg University

Which is AI social impact?

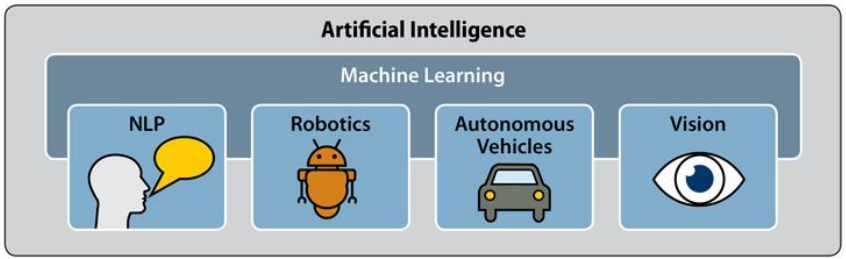
Artificial Intelligence algorithms has become central in the development of new technologies

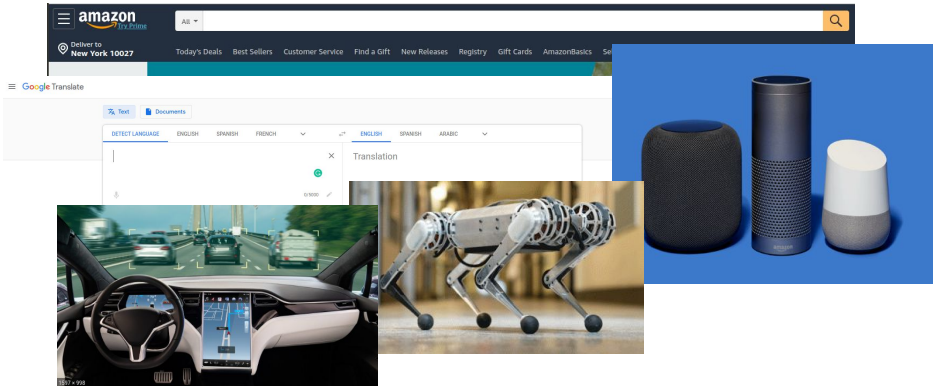
One of the axiom has been:

more data is equal
to better ML algorithm

On the other side
we should not forget that

With great power
comes great responsibility

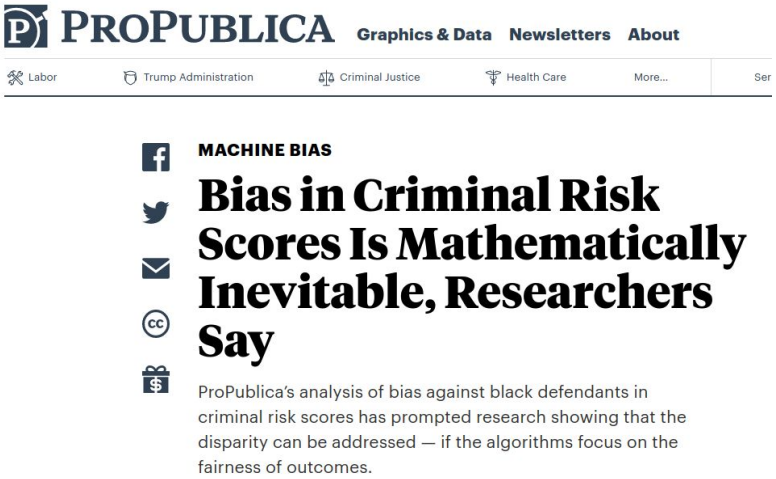




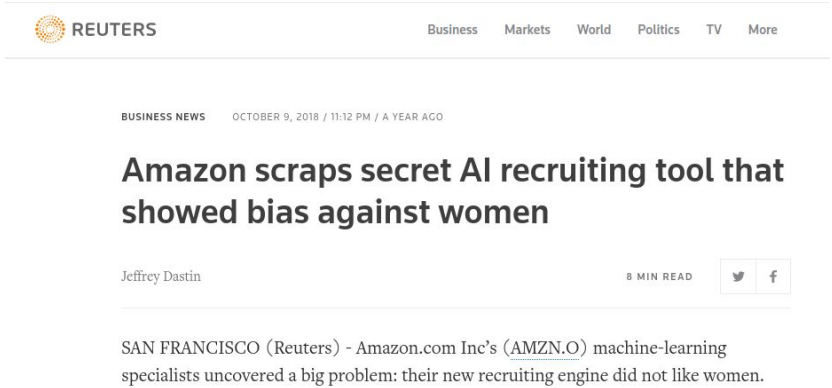
https://www.datamation.com/artificial-intelligence/examples-of-artificial-intelligence.html

Bias and fairness in AI

The use of AI raises questions about fairness, transparency, and due process in government decisions and adopted public policies.



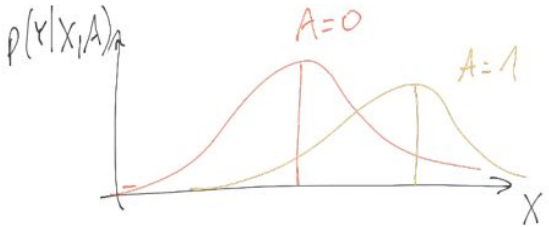
https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say



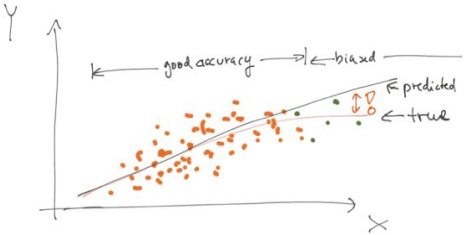
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Why AI is biased?

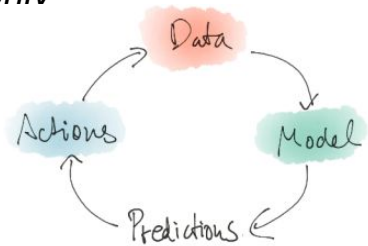
1. **The world and our society are biased**:
   biases already present in the datasets;
   examples: gender gap, salary and zip code, ...



2. **Underrepresentation in data**:
   lack of data exploration during the building of datasets that could leave minority groups underrepresented



3. **Self-fulfilling prophecies**:
   biases in how the algorithm minimizes its global error, targeting onlv the majority groups in the training set



Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. arXiv preprint arXiv:1810.08810
https://cdn.oreillystatic.com/en/assets/1/event/300/Fair%2C%20privacy-preserving%2C%20and%20secure%20ML%20Presentation%201.pdf

Gender discrimination in NLP

25 July 2018

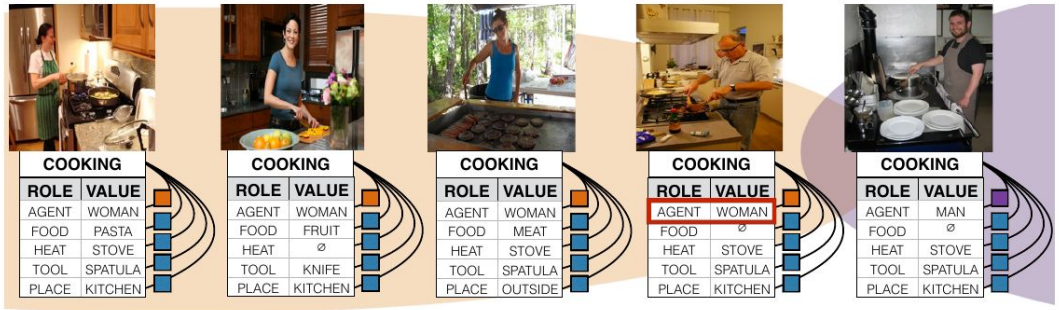
Artificial intelligence is demonstrating gender bias – and it's our fault

Dr Muneera Bano, Lecturer in Software Engineering, Swinburne University of Technology

The data being used to train AI programmes is often gender-biased

www.kcl.ac.uk/news/artificial-intelligence-is-demonstrating-gender-bias-and-its-our-fault

The amplification of gender bias in ML algorithm is more evident when they are trained on datasets based on text, because normally they are huge collections of text produced directly by other humans



Hendricks et all (2018). Women also snowboard: Overcoming bias in captioning models. In European conference on computer vision

How to quantify gender bias

- **Word Embedding Association Test WEAT:**
  Application of the Implicit Association Test (IAT), used by psychologists to measure subconscious bias in humans, to measure the differences in the strength of association of concepts between genders in NLP datasets.

  Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science(6334), 183

- **Correlation between the gender-associated words and sub-spaces of a word embedding matrix**
  Linear support vector machine to classify words as gender-specific or gender-neutral
  Principal component analysis is used to identify the greater variance between gender pair words

  Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in neural information processing systems, 4349–4357.

- **Differences in NLP algorithm performances based on gender**
  Randomly exchange words related to gender with the other gender in test dataset

  Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. In (pp. 2799–2804). Retrieved from https://aclweb.org/anthology/D18-1302

Master thesis goals

1. **Reproduce methodologies to measure gender inequality in English datasets**
   a. WEAT methodologies on *Google News text dataset*
   b. Gender association score on *Google News text dataset* Word Embedding
      Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. ICLR, 2013 .
   c. Gender performances differences in *abusive tweets* detection and *tweets sentiment analysis*
      Founta et al (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. 12th AAAI Conference on Web and Social Media
      Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. 7th SEM'18.

2. **Adapt already established methodologies to Italian**

3. **Compared gender inequality scores between Italian and English datasets**

Italian NLP datasets

http://www.italianlp.it/resources/italian-word-embeddings/

haspeede / hate speech detection

- 4000 Tweets and 4000 Facebook posts collections
- Each text is classified if it consists on hate speech

http://www.di.unito.it/~tutreeb/haspeede-evalita18/index.html

Italian and English difference

- Both methods that investigate gender inequality directly in the Word Embedding matrix are bases on gender word sets created by researcher:
  - ➢ **Need to create of gender Italian word sets**

- Performance evaluation is based on the ability to automatically swap gender in the test dataset. In English gender is encoded in pronouns or specific word, while in Italian *adjectives and verbs are conjugated by gender and the pronouns he/she are commonly omitted before a verb*
  - ➢ **Need to develop an algorithm to automatically identify whether an Italian sentence refers to a female or a male and swap gender**

Gender inequality metrics

- **Cosine similarity in word embedding:**

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} .$$

- **Gender direction in word embedding:** *inner product between the vector difference between two gender-specific words and the vector difference between two gender-neutral words*

- **False positive/negative equality difference**: FPR and FNR are the overall false positive and negative rates and *T*={male, female}

$$FPED = \sum_{t \in T} |FPR - FPR_t|$$

$$FNED = \sum_{t \in T} |FNR - FNR_t|$$

Thanks for your attention

Additional references:

- **Sun, T., et all (2019). Mitigating gender bias in natural language processing: Literature review. 1630–1640. Retrieved from https://www.aclweb.org/anthology/P19-1159**
- **Collett, C., & Dillon, S. (2019). Ai and gender: Four proposals for future research. Cambridge: The Leverhulme Centre for the Future of Intelligence..**
- **Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning.fairmlbook.org. http://www.fairmlbook.org**
- **Bano, M. (2018). Artificial intelligence is demonstrating gender bias - and it's our fault. Kings College London News Centre.. Retrieved from www.kcl.ac.uk/news/artificial-intelligence-is-demonstrating-gender-bias-and-its-our-fault**
- **Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Schwartz, O. (2018). Ai now report 2018. AI Now Institute at New York University**

HARRISBURG UNIVERSITY
OF SCIENCE AND TECHNOLOGY