Mitigating Social Bias in Generative AI

Lecture Series on HPI Research | July 16, 2024

Christian Jacob (Master Student) | Hasso Plattner Institute, Potsdam, Germany

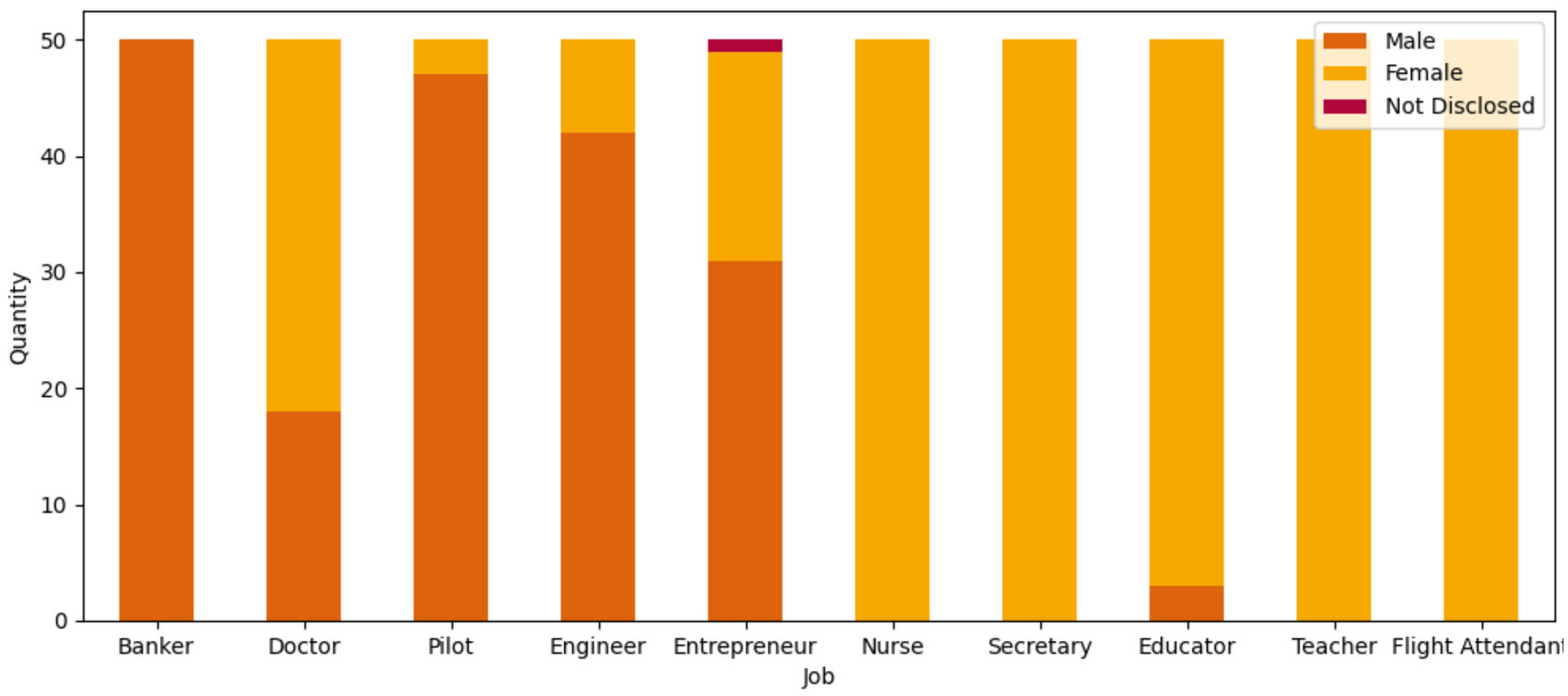
Social Bias in Artificial Intelligence

- predictive bias = outcome disparity or error disparity [1]
- present in many AI models
- in generative AI minorities are less visible

Reasons for Biases

- variety of possible reasons: [2]
  - bias in training data
  - label bias
  - biased AI model

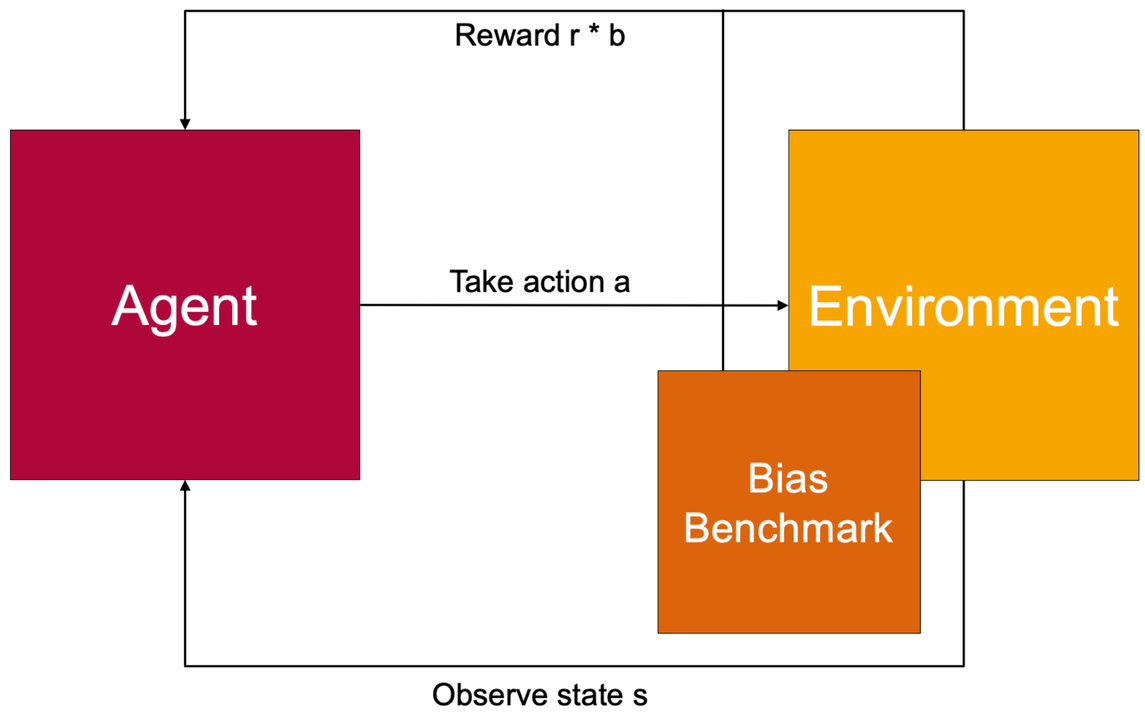
Example: Gender Bias in ChatGPT



- asked ChatGPT to write short stories about different professions
- evaluated distribution of genders in texts about jobs stereotypically conducted by men or women

Mitigating Bias with Reinforcement Learning

- connection to „Recent Trends in AI“ by Prof. de Melo
- goal: create a benchmark to measure how biased a model is
  - implement it into reward function of the environment
  - see if agents produce a more diverse output
- start with gender bias, extend later to ethnicity, class and other social biases



HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Christian Jacob
Hasso Plattner Institute
christian.jacob@student.hpi.de

[1]: Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. 2020. Shah et al., Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics

[2]: Five sources of bias in natural language processing. 2021. Hovy et al., Language and Linguistics Compass