AI and bias (bias in algorithms, bias in data)

Presenter: George Manias (gmanias@unipi.gr)
Department of Digital Systems, University of Piraeus, Piraeus, Greece
OUTLINE

- AI in modern societies
- What is AI bias?
- Examples of AI bias in real world
- What are the types of bias in AI?
- Impact of bias in AI
- Common forms of bias in AI
- Widely used tools to mitigate bias
- How to better tackle bias in AI?
- How AI4Gov identifies and removes bias in AI?
- How AI4Gov creates Responsible and Trustworthy AI?
AI IN MODERN SOCIETIES

AI and bias (bias in algorithms, bias in data)
AI
Data makes the world go around...

...BUT
AI and bias (bias in algorithms, bias in data)

**Germany's Ruling Party Boss Warn Against Bias in AI Data Sets**

Germany's Social Democrat Leader Saskia Esken called for a closer examination of the discrimination inherent within data sets used to train artificial intelligence models.

---

**AI poses existential threat and risk to health of millions, experts warn**

BMJ Global Health article calls for halt to ‘development of self-improving artificial general intelligence’ until regulation in place.

---

**Danger in the Machine: The Perils of Political and Demographic Biases Embedded in AI Systems**

---

**The politics of AI: ChatGPT and political bias**

WHAT IS AI BIAS?

AI bias is an anomaly in the output of AI algorithms. This could be due to the prejudiced assumptions made during the algorithm development process or prejudices in the training data.
EXAMPLES OF AI BIAS IN REAL WORLD

- Racism embedded in US healthcare *(due to features/variables used)*
- Amazon’s Recruiting Engine gender bias *(due to training data)*
- COMPAS algorithm race bias with reoffending rates *(due to training data and algorithm process)*
- Google Photos Algorithm race bias *(due to features/variables used)*
- PredPol Algorithm biased against minorities *(due to training data used and human bias)*
WHAT ARE THE TYPES OF BIAS IN AI?

❑ Statistical/Computational Bias
  - Stem from the datasets and algorithmic processes used (e.g., heterogeneous data, undersampling in data, algorithmic biases such as over- and under-fitting, erroneous treatment of outliers, selection of features and data cleaning)

❑ Human Bias
  - Present in the institutional, group, and individual decision-making processes across the AI lifecycle and in the use of AI applications once deployed (e.g., social assumptions and norms create blind spots and expectations in thinking)

❑ Systemic Bias
  - Historical, societal, institutional (e.g., institutional racism and sexism, smart-homes that exclude people with disabilities)

IMPACT OF BIAS IN TECH LANDSCAPE

Concerns about AI bias among tech leaders

- 56% loss of customer trust
- 50% compromised brand reputation
- 43% increased regulatory scrutiny
- 42% loss of employee trust
- 37% mismatch with personal ethics
- 25% lawsuits
- 22% eroding shareholder value

Source: DataRobot

Impact of data bias on business

- 62% lost revenue
- 61% lost customers
- 43% lost employees
- 35% incurred legal fees due to lawsuit
- 6% lost customer trust

Source: DataRobot

* “State of AI Bias” report, DataRobot, 2022

AI and bias (bias in algorithms, bias in data)
COMMON FORMS OF BIAS IN AI

- Historical Bias
- Sample/Selection Bias
- Feature/Label Bias
- Aggregation Bias
- Confirmation Bias
- Interaction bias
- Evaluation Bias
WIDELY USED TOOLS TO MITIGATE BIAS IN AI

- Google’s “What-If Tool”
- IBM’s AI Fairness 360
- Academic fairness-in-AI projects are under progress funded by the National Science Foundation (NSF) in collaboration with Amazon.
- Facebook AI is using a new “radioactive data” technique to detect if a dataset that was used to train a ML model underlines bias.
- audit-AI Toolkit for bias testing in ML applications
- FairLens
- PwC’s Responsible AI
- ...

AI and bias (bias in algorithms, bias in data)
HOW TO BETTER TACKLE BIAS IN AI?

The challenge of bias in AI is complex and multi-faceted. While there are many approaches and tools for mitigating this challenge there is no quick fix.

Employ human-in-the-loop technology for constant improvement

Attention on data and algorithms for constant feedback
- Underlying data
- User-generated data
- Deploying suitable tools to identify bias and inaccuracies

Diverse and inclusive team of experts and extensive research in bias

Algorithm test in a real-world setting
HOW AI4GOV MITIGATES BIAS?

1. **Joint Effort**
   - Of policy makers, public organizations, legal, social scientists & AI experts following a multidisciplinary approach for understanding bias

2. **Regulatory Frameworks**
   - Introduces a Data Governance Framework and Holistic Regulatory Framework to govern the whole data and policy-making cycles

3. **Bias Detection Tools**
   - To effectively evaluate and mitigate common bias of AI models in real-world data

4. **Promotes Transparency**
   - Explainable AI (XAI) tools for explaining and understanding the underlying AI mechanisms and derived policies

5. **Involves Stakeholders**
   - Involving and training stakeholders to build trust and allow for easy identification and correction of any bias

In-line with applicable laws, protocols, and regulations (i.e., the GDPR), but also with ethical recommendations for AI, e.g., the recommendations of the HLEG and the EU AI Act.

Raising the awareness of stakeholders will reassure them that AI is being leveraged responsibly and ethically

AI and bias (bias in algorithms, bias in data)
HOW AI4GOV CREATES RESPONSIBLE AND TRUSTWORTHY AI?

- **Research on AI and Rule of Law**: Examine the effectiveness of monitoring and control protocols of established legislation and non-regulatory measures over AI and Big Data (e.g., ALTAI, EU AI ACT)

- **AI Fairness and Bias Mitigation**: AI4Gov seeks to introduce a set of tools and solutions covering the issue of bias in AI across the complete AI chain. (Bias Detector Toolkit, HRF and Virtual Unbiased Framework)

- **Trusted, Transparent, and Interpretable AI solutions**: Enhance trustworthiness, fairness and explainability, by enabling humans to reason about the outcomes of AI-based models (eXplainable AI (XAI) & Situation-Aware Explainability (SAX))

- **Identification and Modelling of Bias and Discrimination and their Sources**: Methods and techniques to identify and mitigate bias (VUF, Bias Detector). XAI Toolkit will be combined with VUF to provide bias removal recommendations.

- **AI applications in the context of Citizen Engagement and Participation**: Establish hybrid citizen-centric training approaches and utilize AI techniques to facilitate the interaction between citizens and organizations.
If not now, then when?
THANK YOU FOR YOUR ATTENTION