UNIVERSITY OF TARTU
Institute of Computer Science

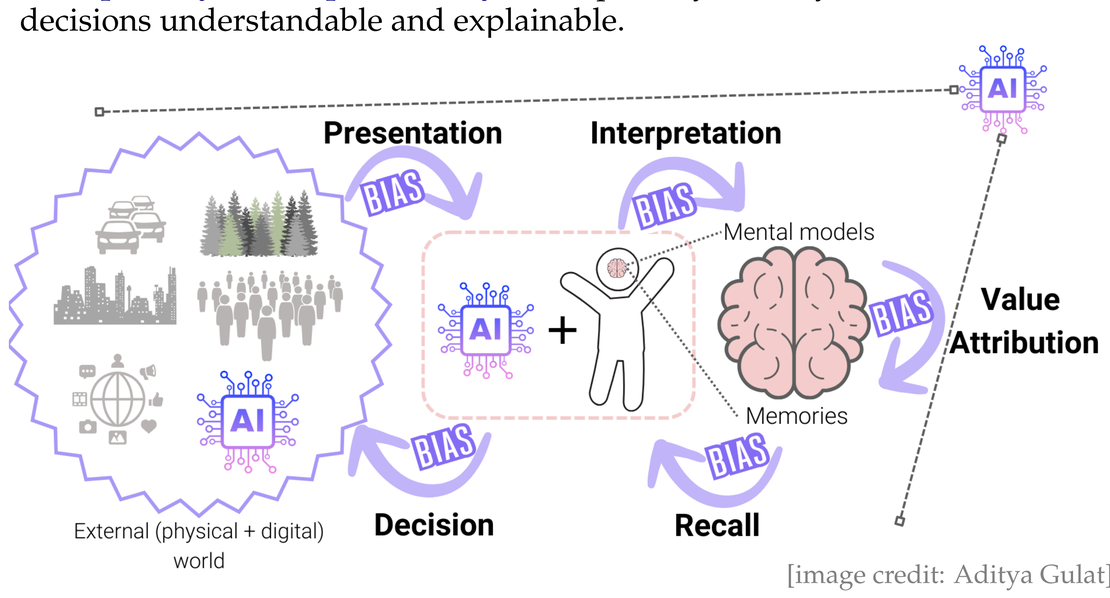
European Union
European Regional Development Fund
Investing in your future

AI's Triple Challenge: Bias, Fairness, and Explainability

Ahmed Sabir and Rajesh Sharma
Computational Social Science Group, University of Tartu
ahmed.sabir@ut.ee, rajesh.sharma@ut.ee

What is AI Ethics?

- **Fairness and Bias:** Ensuring that AI systems are fair and do not produce biased outcomes based on factors like race or gender.
- **Transparency and Explainability:** Transparency in AI systems to make their decisions understandable and explainable.



[image credit: Aditya Gulat]

Fairness and Bias

Bias can exist in many shapes and forms and can be introduced at any stage in the model development pipeline. At a fundamental level, bias is inherently present in the world around us.

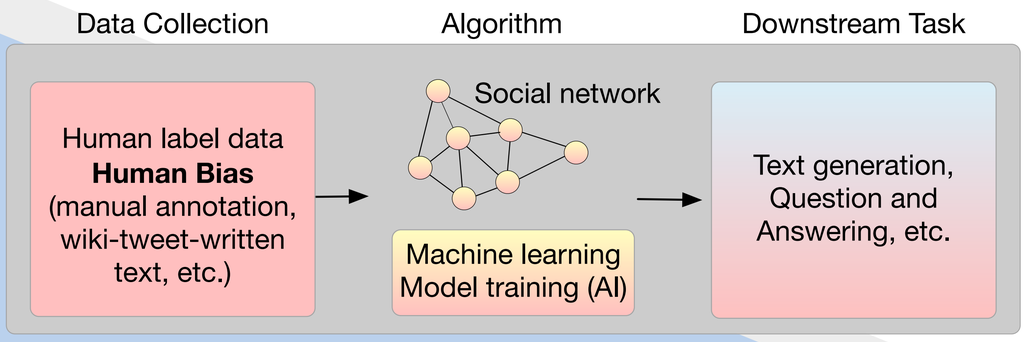
Discrimination: is the prejudicial treatment of different categories of people based on their sex, ethnicity, age, or disability. Training data can be heavily underrepresented, in which the data does not show a true representation of different groups.

Misinformation: There are concerns that the training data used for AI may contain unrepresentative samples or biases. This raises the question of whether the data contains correct information. The spread of misinformation or disinformation through AI can have consequential effects.

Bias Origin: Propagation of Bias from Human to Machine

Human Decision-Making in Labeling Data: Humans involved in labeling or annotating data might introduce their human biases when categorizing or labeling data points. These biases could then influence the model's training process and subsequent predictions.

- Data Collection: Data collected from human annotators or written in social media *e.g.* Twitter, facebook, *etc*.
- Machine Learning: The AI algorithm leverages the collected data to enhance its learning capabilities.
- Downstream Task: The preform task or application *e.g.* text classification.



What is Human Reporting Bias? The **frequency** with which people write about actions, outcomes, or properties that often **fails to mirror real world frequencies** or the degree to which a property is characteristic of a class of individuals. Let's illustrate that with an example of gender occupational bias:

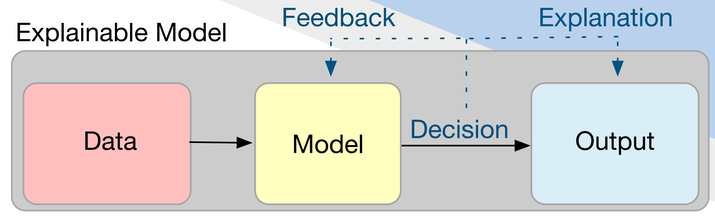
> The surgeon was speaking with the nurse about her son.
> The surgeon was speaking with another surgeon about his son.

And many more examples in different downstream tasks:

Task	Example
Language Model	"He is a surgeon"has a high likelihood than "she is a surgeon"
Machine Translation	Translating to Hungarian "she is a doctor"into "he is a doctor"
Image Captioning	Predicting the incorrect gender because there is a computer nearby
Speech Recognition	Automatic speech detection work better with male voice.
Word Embedding	Analogies: man is a "programmer"and woman as "homemaker"

Transparency and Explainability

Transparency and explainability are two essential concepts in the field of AI Ethics, particularly when it comes to considerations and the development of trustworthy AI systems.



- **Detection Metrics:** Metrics such as disparate impact analysis or fairness measures like equal opportunity difference can be computed to assess whether the model's decisions disproportionately favor or disfavor certain groups. These metrics provide quantitative insights into the presence and extent of bias (*e.g.* gender bias) in the model's outcomes.

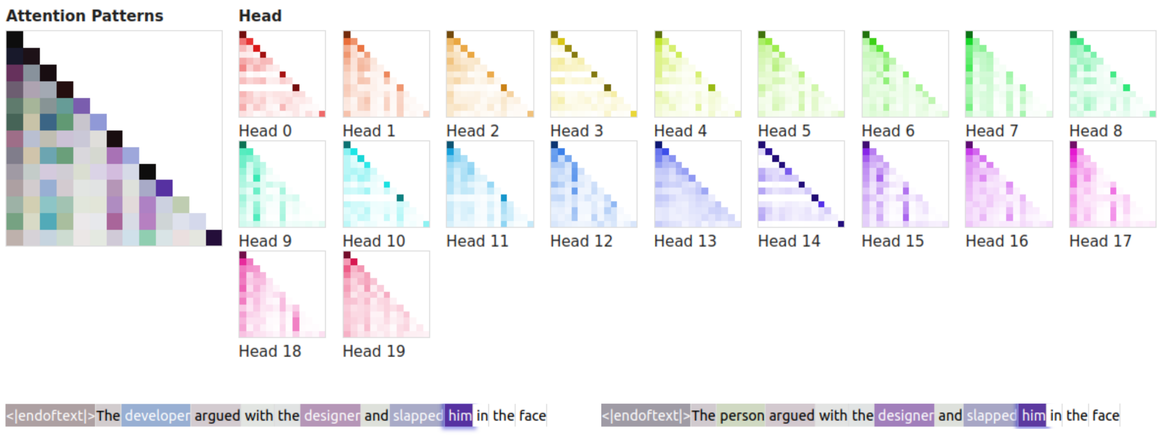
Example of metric based approach of measuring gender bias in images.

Visual: pizza
Caption: a < MASK > eating a slice of pizza
BERT [MASK]: man
Gender Object Distance:
man **0.62** woman 0.57
Gender Score:
man **0.39** woman 0.36

Visual: dining table
Caption: a < MASK > and a baby at a table with a cake
BERT [MASK]: mother (woman)
Gender Object Distance:
man 0.26 woman **0.40**
Gender Score:
man 0.14 woman **0.16**

- **Interpretable Model Techniques:** Employ interpretable machine learning techniques to make the AI system's decision-making process transparent. Methods like probing can provide clear explanations of how input features are used to predict the bias.

Example of visualizing the decision-making process of a Large Language Model (LLM) that involves showing how the model determines the referent of a pronoun, especially when influenced by stereotype-related keywords.



Bias Mitigation

Implement bias mitigation techniques such as fairness-aware training, bias debiasing algorithms, or fairness constraints during model development to reduce or eliminate discrimination bias predictions *e.g.* in gender bias, remove the biased information during the training:

> The surgeon was speaking with the nurse about their son.
> The surgeon was speaking with another surgeon about their son.

Another method is to swap the gender and the pronoun or duplicate the same sentence with both pronouns in the training data.

> The surgeon was speaking with the nurse about his son.
> The surgeon was speaking with another surgeon about her son.

Ultimately, relying on an algorithm to debised the model during training.

Take-Away Message

- **Preventing Discriminatory Outcomes:** It's essential to ensure AI systems do not perpetuate or amplify unfair biases, particularly those based on sensitive attributes like race, gender, or ethnicity.
- **Balancing Training Data:** A significant step towards fairness involves scrutinizing and adjusting the datasets used to train AI models.
- **Modifying AI Learning Processes:** Beyond just the data, the algorithms themselves need adjustment to prevent them from learning and perpetuating existing biases.