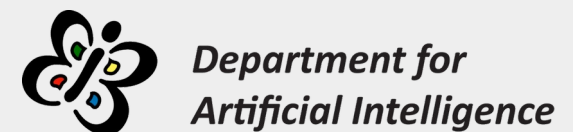


Bias in Artificial Intelligence

...and its influence on human rights

Matej Kovačič, Alenka Guček, Tanja Zdojšek Draksler

<https://ircai.org>



ITWG event

Oslo, September 24 – 25th 2024

Bias

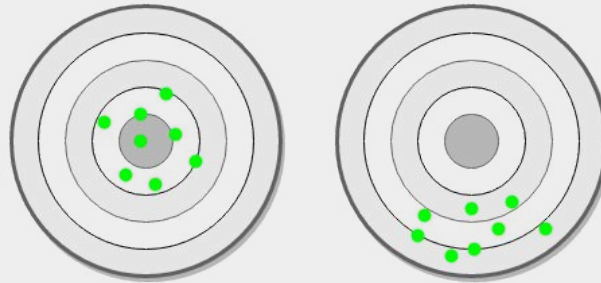
Bias: preconceived notions, prejudices, or unfair preferences that individuals may have based on their personal experiences, cultural influences, or societal conditioning.



Bias can affect how people perceive, judge, and make decisions about others or various situations.

Bias in AI

AI bias is a phenomenon that occurs when an algorithm produces output that is systemically prejudiced due to prejudiced or erroneous assumptions made during the algorithm development process or prejudices in the training data.



AI model is biased if it produces skewed results, typically in a way that upholds harmful beliefs and prejudice against individuals or groups, or is not in line with positive human values like fairness and truth. However, sometimes AI systems could also be biased by design and on purpose.

Bias in AI

Bias can manifest in various ways and at different stages of the AI solution development, including:

- data collection,
- data preprocessing,
- feature selection/engineering,
- model training,
- evaluation and
- system deployment.



The problem is, that even minor initial bias in AI can be amplified during the machine learning process, which can poison the whole system and have far-reaching consequences.

Example #1: medicine

In 2019 a group of researchers in Duke University Hospital's emergency department started developing an algorithm, to help predict childhood sepsis. They have been aware of possible bias problem and they spent lot of effort to make algorithm fair.

However after almost three years, they discovered possible bias. They found that doctors took longer to order blood tests for Hispanic kids eventually diagnosed with sepsis than for white kids. One possible explanation for that was that the physicians needed interpreters, and this slowed down the process of ordering blood tests.

This delay inaccurately taught AI, that Hispanic kids develop sepsis slower than other kids. And that time difference resulting from the bias could be fatal.

Based on that findings, Duke University Hospital's team fixed the algorithm to predict sepsis at the same speed for all patients.

Example #2: fraud detection

In 2023 the journalists of Guardian newspaper found out that the British Home Office uses biased AI to flag up sham marriages. The internal Home Office evaluation has shown that the tool disproportionately flags up people from Albania, Greece, Romania and Bulgaria.

The journalists also found out that the Department for Work and Pensions (DWP) uses an AI algorithm to detect fraud and error among benefits claimants. But there are some clues that the algorithm falsely flagged a lot of Bulgarians as making potentially fraudulent claims, and their benefits have been then suspended.

However, both ministries claimed the processes they use are fair because the final decisions are made by people and not the AI.

But the problem is that since officials usually have limited resources to review the cases, they highly rely on algorithm decisions. So biased algorithms will usually lead to biased final decisions and the people that are affected by those decisions would in general not know that decision has been based on biased AI.

Example #3: child care benefits

In 2019 has been revealed that the Dutch tax authorities had used a self-learning algorithm to create risk profiles in an effort to spot child care benefits fraud. But the algorithm was not working properly and has been wrongly labelling people as fraudsters.

After several years of using this biased algorithm, country's privacy regulator opened an investigation. They found out that the tax authorities focused on people with "a non-Western appearance" (people with Turkish or Moroccan nationality), and among risk factors were also having dual nationality and a low income. Authorities penalized families over a mere suspicion of fraud based on the system's risk indicators.

The results were disastrous. Dutch tax authority demanded that people flagged as fraudsters pay back their child care allowances, and tens of thousands of families were pushed into poverty because of exorbitant debts to the tax agency. Some victims committed suicide and more than a thousand children were taken into foster care.

Example #4: face recognition

Several research had shown that many facial recognition algorithms perform poorly at identifying people besides white men.

Researchers found that this is the result of two factors. First, there is the lack of Black faces in the algorithms' training data sets. And second, those systems often magnify police officers' own biases.

The fact that that the technology struggles to distinguish darker faces, often leads to more racial profiling and more false arrests. And inaccurate identification increases the likelihood of missed arrests.

Of course, the software does not arrest people. It just suggests who are the potential suspects, and police officers then decide who to arrest.

But people often believe that AI is infallible and don't question the results.

Example #4: face recognition

In 2020, Robert Williams has been arrested for allegedly stealing thousands of dollars of watches. Detroit police used AI algorithm that matched video from surveillance camera to the Williams' driver's license photo.

The police did not collect any corroborating evidence such as eyewitness identification, cell phone location data or a fingerprint.

The sole evidence has been a picture from video surveillance camera and an identification by AI algorithm.

He has been arrested in front of his neighbours and family, and detained for 18 hours. His arrest is the first documented case of someone being wrongfully detained based on facial recognition technology.

He later learned about the study by National Institute of Standards and Technology, which found that algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces.

Generative AI and bias

Generative AI has a history of amplifying racial and gender stereotypes.

Google's Gemini attempts to solve this are causing problems too.

When Gemini was asked to produce pictures of white people, it refused, saying it couldn't fulfil the request because it "*reinforces harmful stereotypes and generalizations about people based on their race.*"



Dealing with bias

There are different **strategies** (adversarial debiasing, reweighing, ensemble adversarial training,...) and **tools** (IBM's AI Fairness 360, Google's What-If tool, LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations) to mitigate bias that target different stages of development of AI systems.



As part of the AI4Gov project, we are developing the **integration** of these **techniques** for mitigating bias, **XAI** (explainable artificial intelligence) tools, and **educational materials** in the field of bias in AI.

AI4Gov project objectives

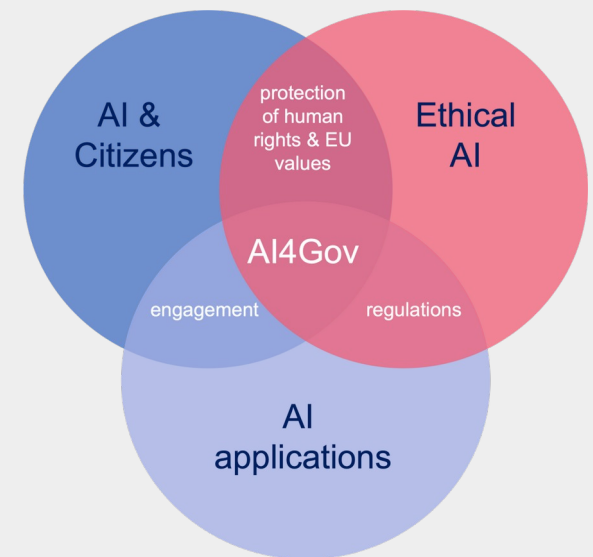
Design a Reference Framework for an **Ethical** and **Democratic** AI

Design and develop **AI Fairness Monitoring** and **Bias** Mitigation tools

Develop **Trusted eXplainable AI** techniques for explaining AI decisions to policy makers, citizens and other stakeholders

Boost the **Regulatory Compliance** of AI based models for democratic processes

Design and implement a **Reference AI platform** for **regulating** the use of AI & Big Data



Goal: to address ethical, trust, discrimination, and bias challenges, currently faced by AI & Big Data technologies (used in the public sector).

Bias detector catalog


We are developing so called **Bias Detector catalog**, which will present various tools for analyzing and mitigating AI bias in a transparent manner. Tools are also categorized according to functionality, field of application and performance.


AI4Gov Platform
Dashboard


- Dashboard
- Use Cases
- Policy Recommendation Toolkit
- Bias Detector Toolkit
- Introduction
- Bias Detector Catalogue


Bias Detector Catalogue


The Bias Detector Catalogue stands as a pioneering tool, meticulously curated to house a comprehensive array of projects. Each entry within this expansive repository represents a concerted effort by innovators and researchers worldwide to address bias at its root, offering multifaceted solutions tailored to diverse stages of the training process. From data collection and preprocessing to model training and deployment, the Bias Detector Catalogue is a testament to the collective determination to mitigate bias's insidious impact on algorithmic decision-making.



Data Collection


Preprocessing


Feature Selection


Model Training


Model Evaluation


Deploy Model

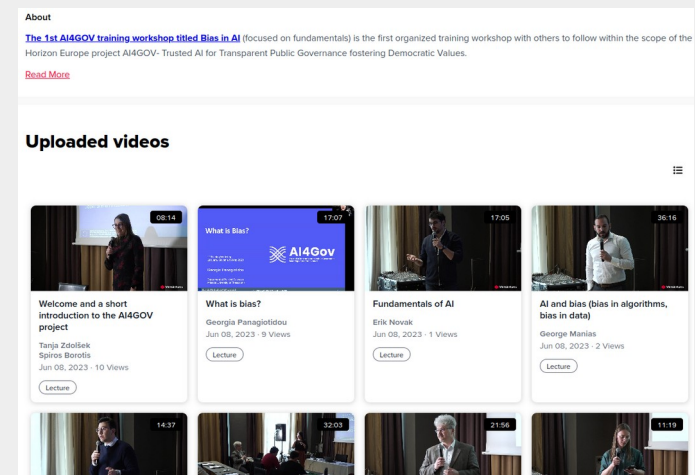
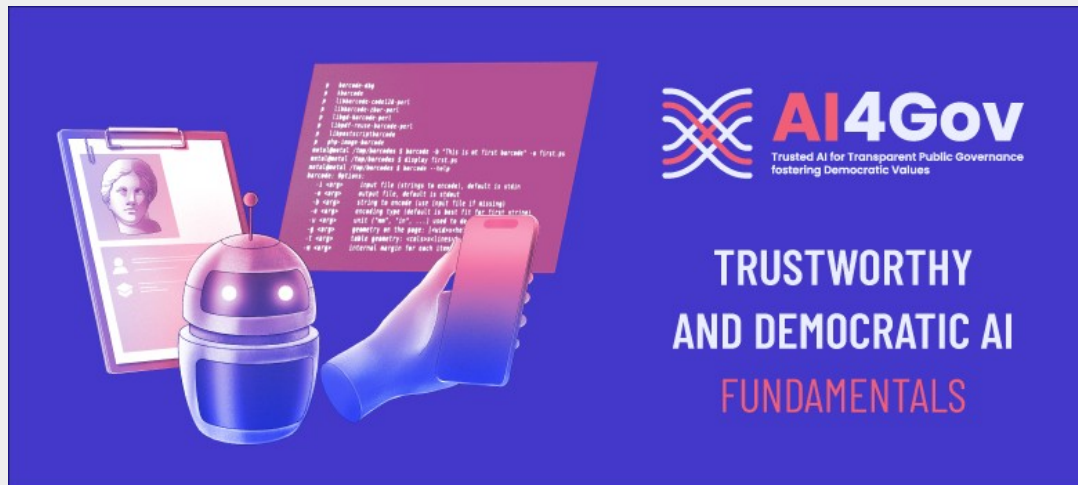
AIF360: AI Fairness 360 toolkit	Accuracy: HIGH	Cost: LOW	▼
FairMLHealth	Accuracy: UNKNOWN	Cost: LOW	▲
<p>Source: https://github.com/KenSciResearch/fairMLHealth</p> <p style="text-align: center;">Type: MITIGATION Programming Language: PYTHON</p> <p>Description: FairMLHealth is a healthcare-specific tool for bias analysis. It provides machine-learning fairness, healthcare applications, and variation analysis.</p> <p>Applicability: HEALTHCARE</p> <p>Limitations: The 'fair' range to be used for these metrics requires judgement on the part of the analyst.</p> <p>References: Ahmad et al., (2020). Fairness in Machine Learning for Healthcare. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. https://doi.org/10.1145/3394486.3406461.</p>			
Mitigating Unwanted Biases with Adversarial Learning	Accuracy: HIGH	Cost: LOW	▼
Themis-ml: A Fairness-Aware Machine Learning Interface for End-To-End Discrimination Discovery and Mitigation	Accuracy: UNKNOWN	Cost: UNKNOWN	▼
Bias in Automated Speaker Recognition	Accuracy: MODERATE	Cost: MODERATE	▼
Bias Assessment Metrics and Measures	Accuracy: UNKNOWN	Cost: UNKNOWN	▼
Biaslyze	Accuracy: UNKNOWN	Cost: LOW	▼
Fari EVA: Fairness EVALuation of Voice Technologies	Accuracy: UNKNOWN	Cost: UNKNOWN	▼

Training materials

One of the goals of the AI4Gov project is also the preparation of various trainings on this topic.

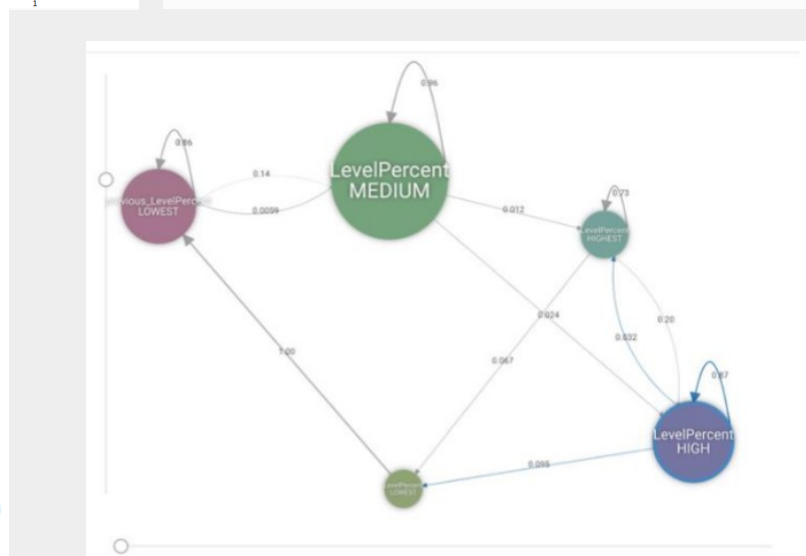
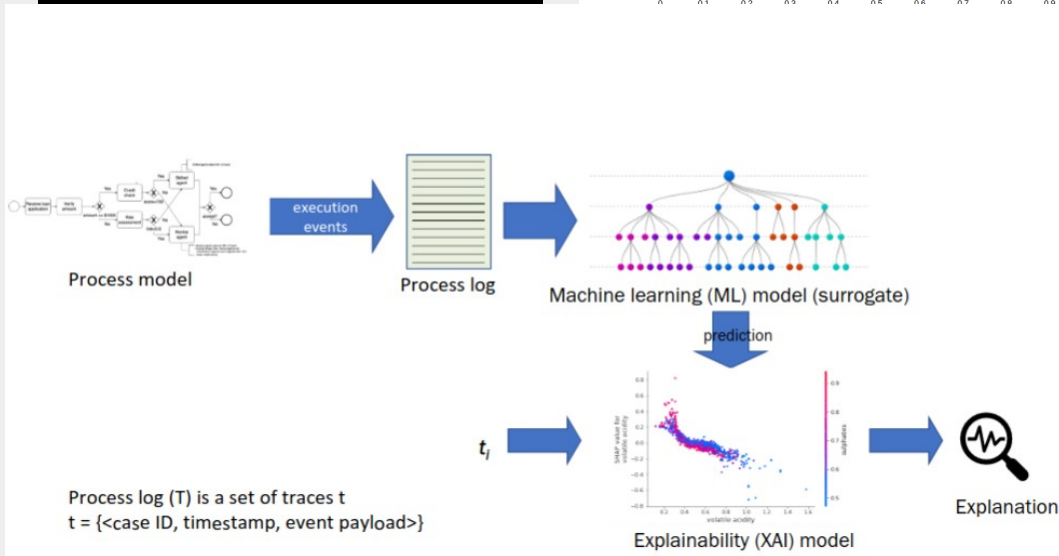
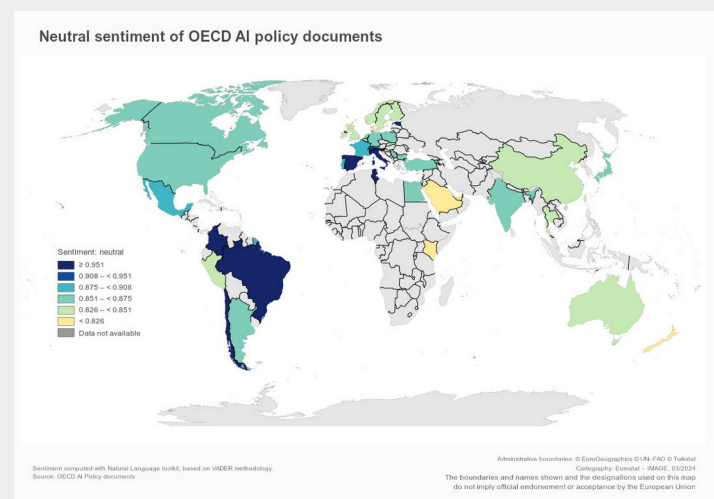
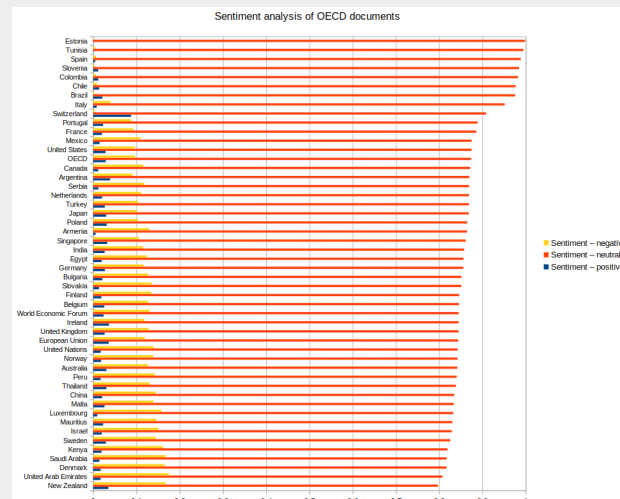
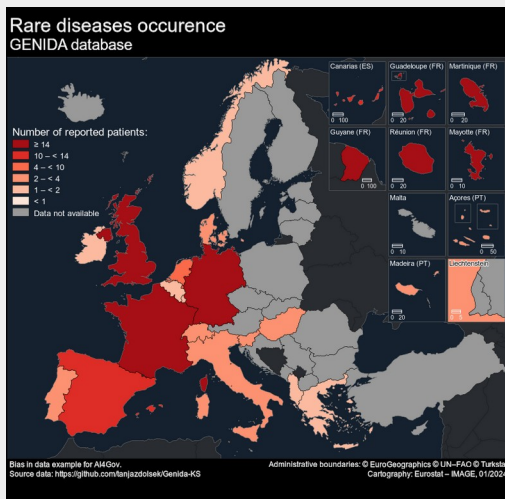
MOOC: Trustworthy and democratic AI – Fundamentals

- Level: beginner
- Duration: 10 hours
- Freely available on the OpenLearnCreate portal (<https://www.open.edu/openlearncreate/enrol/index.php?id=11669>)
- Slovenian and Spanish versions are also being prepared.



Use cases . . .

We are also preparing three pilot projects and testing different methods and approaches for bias analyses.



Questions?



AI4Gov

Trusted AI for Transparent Public Governance
fostering Democratic Values

<https://ai4gov-project.eu/>