Bias and Beyond

AI, regulation, and what it means in practice

Data Fair Conference

Ljubljana, 12.02.2026

Rania Wazir, Co-Founder & CTO

leiwand.ai

What is bias?

AI in medical diagnostics:

SCIENCE Magazine: “Dissecting racial bias in an algorithm used to manage the health of populations.”

- Algorithm used to predict the severity of patient illness, in order to decide on in-patient or out-patient care
- Trained on historical patient data, optimized on amount of money spent to cure patient

Source: Raji et al., The Fallacy of AI Functionality, FAccT 2022: https://facctconference.org/static/pdfs_2022/facct22-353158.pdf.
Ziad Obermeyer et al. ,Dissecting racial bias in an algorithm used to manage the health of populations.Science366,447-453(2019).DOI:10.1126/science.aax2342

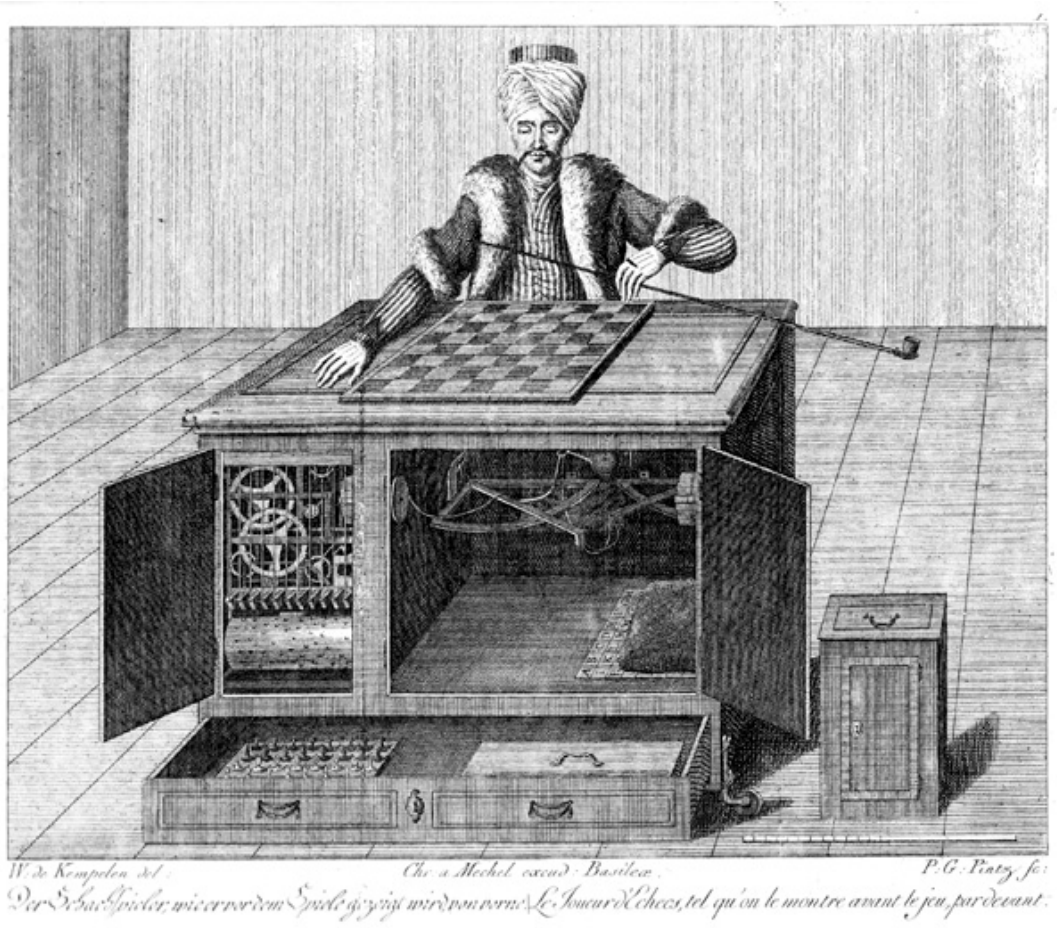
AI Act

Article 1

The purpose of this Regulation is to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), **while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection**, against the harmful effects of AI systems in the Union and supporting innovation.

Regulation (EU) 2024/1689:
https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

Bias and Beyond



Mechanical Turk *(Schachtürke)*
Constructed 1770 by Wolfgang von Kempelen

Copper engraving from: *Briefe über den Schachspieler des Hrn. von Kempelen, by Karl Gottlieb von Windisch* Public Domain

Agenda

1. **What is bias?**
2. Where is bias?
3. Testing for bias
4. Summary

What is bias?

A ***systematic difference in treatment of certain objects, people or groups in comparison to others**** … that can lead to adverse impacts such as risks to health, safety or fundamental rights

* [ISO/IEC 22989:2022, 3.5.4]

Covert bias in AI systems

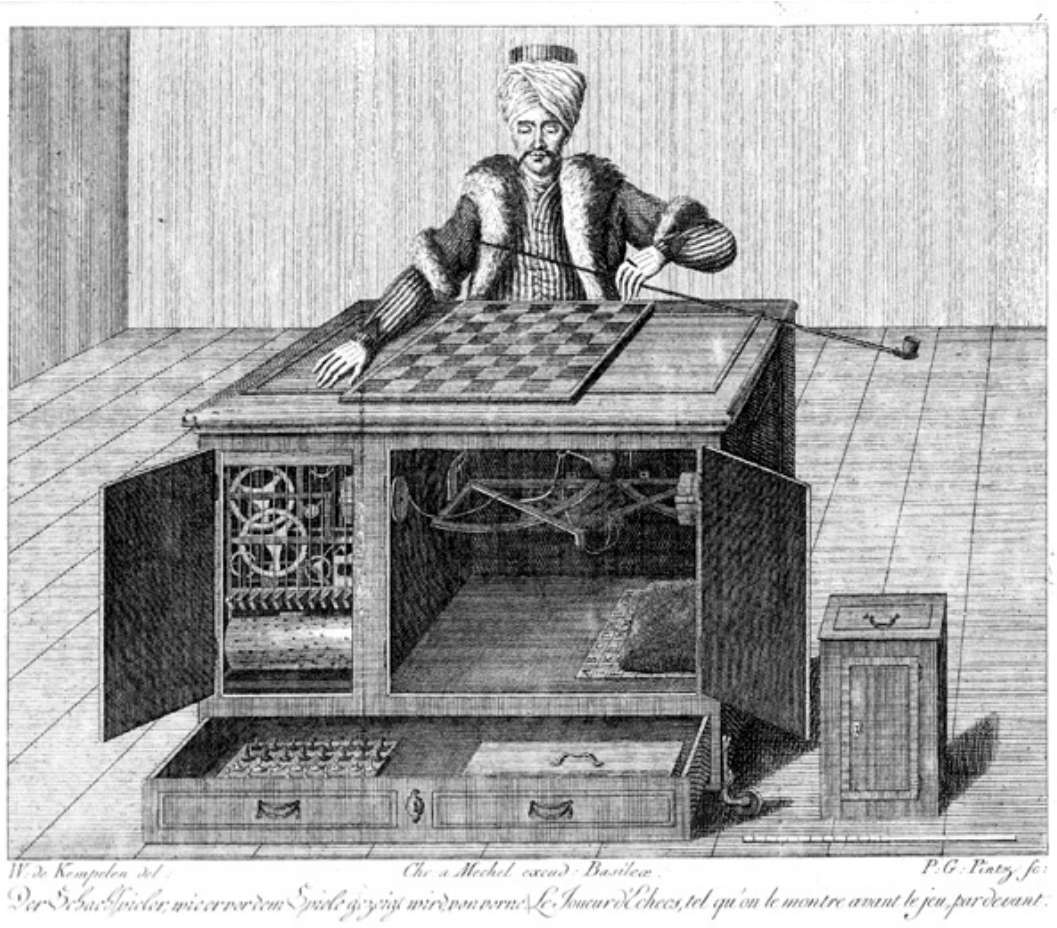
- Hofmann et al,***Dialect prejudice predicts AI decisions about people's character, employability, and criminality***, 2024***.***

- Bloomberg,***OpenAI'sGPT Is a Recruiter's Dream Tool. Tests Show There's Racial Bias,*** 2024.

https://www.bloomberg.com/graphics/2024-openai-gpt-hiring-racial-discrimination/

https://arxiv.org/pdf/2403.00742.pdf

Bias and Beyond

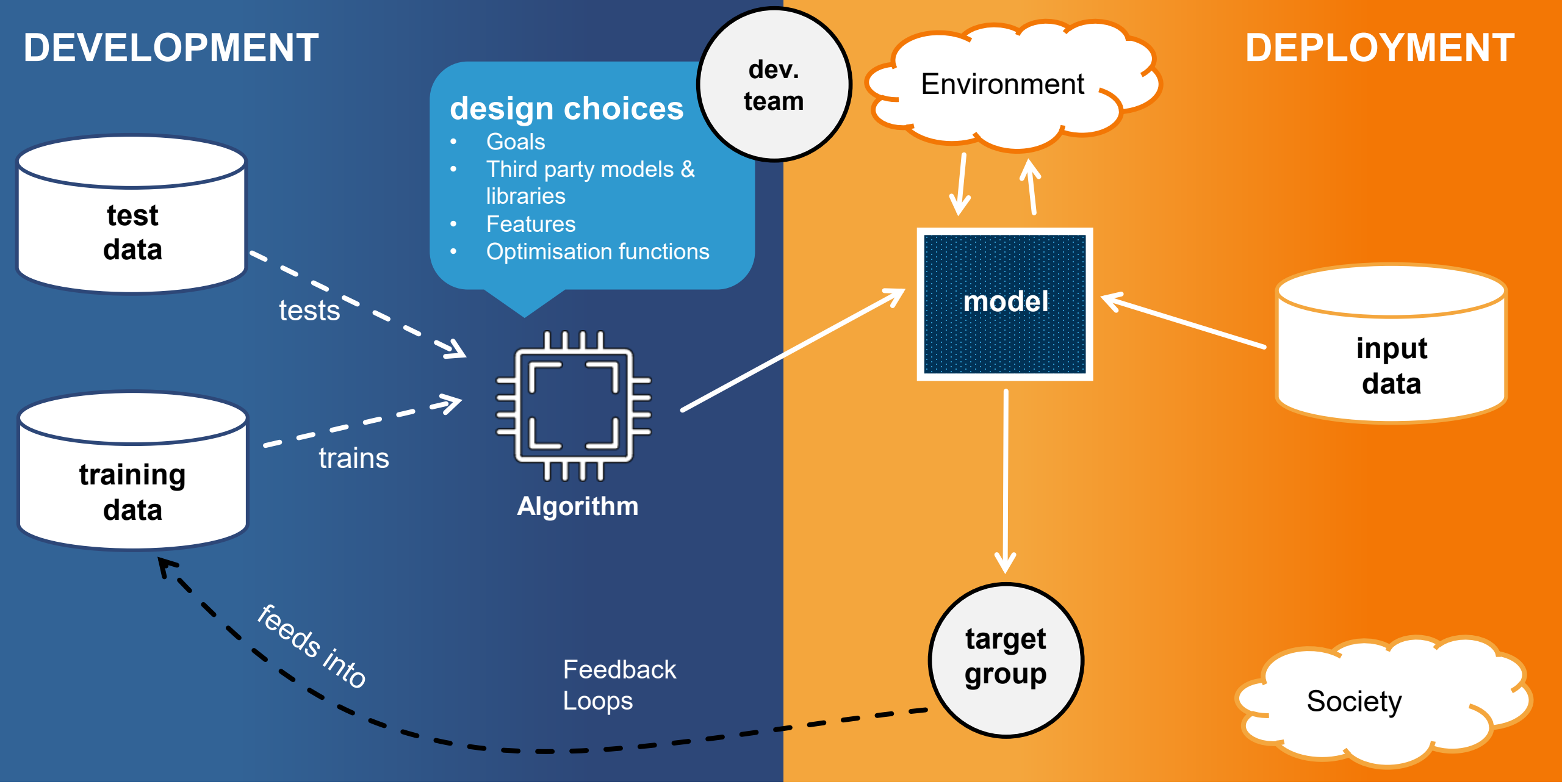


Mechanical Turk *(Schachtürke)*
Constructed 1770 by Wolfgang von Kempelen

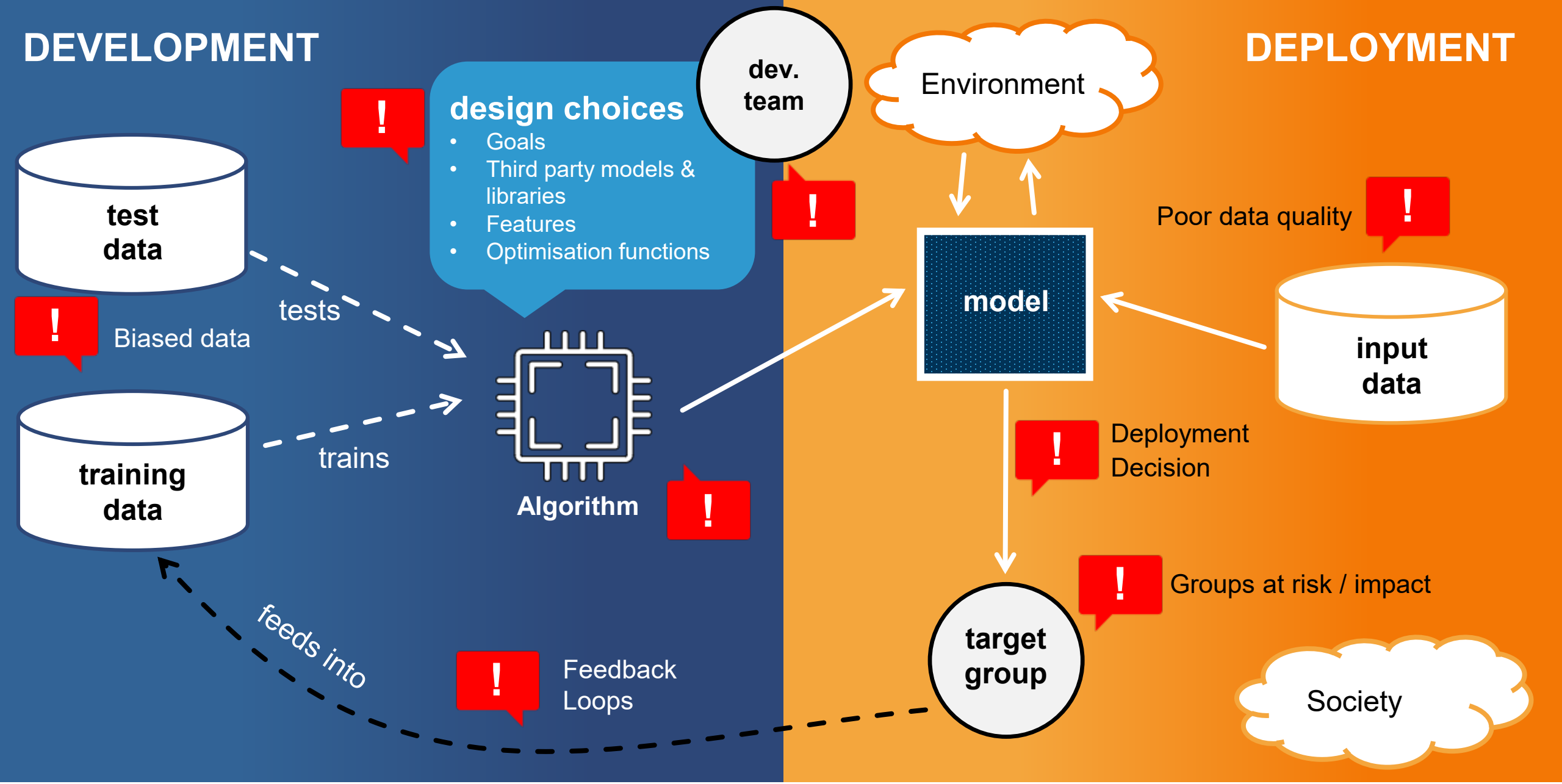
Copper engraving from: *Briefe über den Schachspieler des Hrn. von Kempelen, by Karl Gottlieb von Windisch* Public Domain

Agenda

1. What is bias?
2. **Where is bias?**
3. Testing for bias
4. Summary

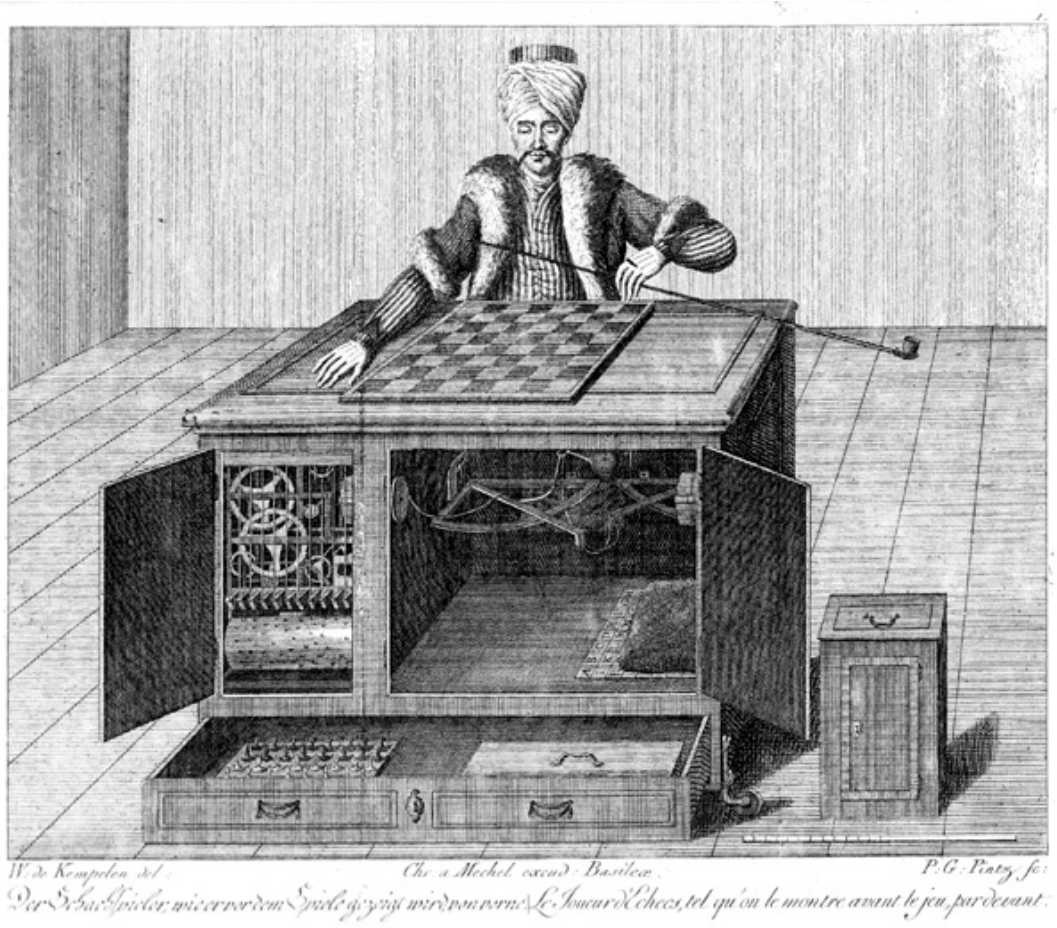


Bias in the AI system life cycle



Bias in the AI system life cycle

Bias and Beyond



Mechanical Turk *(Schachtürke)*
Constructed 1770 by Wolfgang von Kempelen

Copper engraving from: *Briefe über den Schachspieler des Hrn. von Kempelen, by Karl Gottlieb von Windisch* Public Domain

Agenda

1. What is bias?
2. Where is bias?
3. **Testing for bias**
4. Summary

Testing for bias: the subgroups

Bias testing always involves a comparison between groups.

Step 1: conduct a proper risk assessment, and identify the groups at risk of bias.

Example: consider an AI-based safety component in an industrial machine, that interrupts operation when a person is detected within a 5 meter radius.

Who might be at risk?

Testing for bias: the subgroups

Bias testing always involves a comparison between groups.

Step 1: conduct a proper risk assessment, and identify the groups at risk of bias.

Example: consider an AI-based safety component in an industrial machine, that interrupts operation when a person is detected within a 5 meter radius.

Who might be at risk? **(Dark skin? Pregnant women? Children? People in a wheelchair? Tattoos? Black clothes? Someone riding a bicycle?)**

Bias-accuracy trade-off: really?

Consider an AI system with the following performance for men vs women:

- **Accuracy for men: 90%**
- **Accuracy for women: 70%**

Suppose test data is not disaggregated by gender. What is the overall accuracy?

- **If the test data actually consists of 80% men and 20% women, then the overall accuracy would be 86%.** *[0.8x0.9 + 0.2x0.7]*
- **If the test data actually consists of 50% men and 50% women, then the overall accuracy would be 80%.** *[0.5x0.9 + 0.5x0.7]*

The AI system didn‘t change. Only the composition of the test data.

Testing for bias: the metrics

Which metrics are being used to evaluate the system‘s functional correctness? Breaking this metric down by subgroups is **necessary**, but not always **sufficient**:

Example:

An AI system to assess creditworthiness that is 90% accurate for the age group 25-50, and 90% accurate for ages 50+. But:

- for 25-50, all the errors are false positives (i.e. they receive credit even though they are actually high credit risk);
- for ages 50+, all the errors are false negatives (i.e. they are rejected for credit, even though they are actually a low risk)

Testing for bias: the metrics

Various metrics exist that operationalize distinct definitions of „fairness“.

- **Group fairness**
- **Conditional statistical parity**
- **False positive error rate balance**
- **Overall accuracy equality**
- **Well-calibration**
- **Fairness through unawareness**
- **Counterfactual fairness**
- ***and many more***

Testing for bias: the metrics

Lets recall a much-debated example:

In May 2016, ProPublica published an article indicating that the predictions of a widely-used recidivism modelling model (COMPAS), were biased:

- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

- Pro-Publica used ***Equalized Odds***
- Northpointe used ***Predictive Parity***

For a good explanation, see:

- Julia Dressel and Hany Farid, *The accuracy, fairness, and limits of predicting recidivism,* Science Advances, 17 Jan 2018: Vol. 4, no. 1. https://advances.sciencemag.org/content/4/1/eaao5580.full

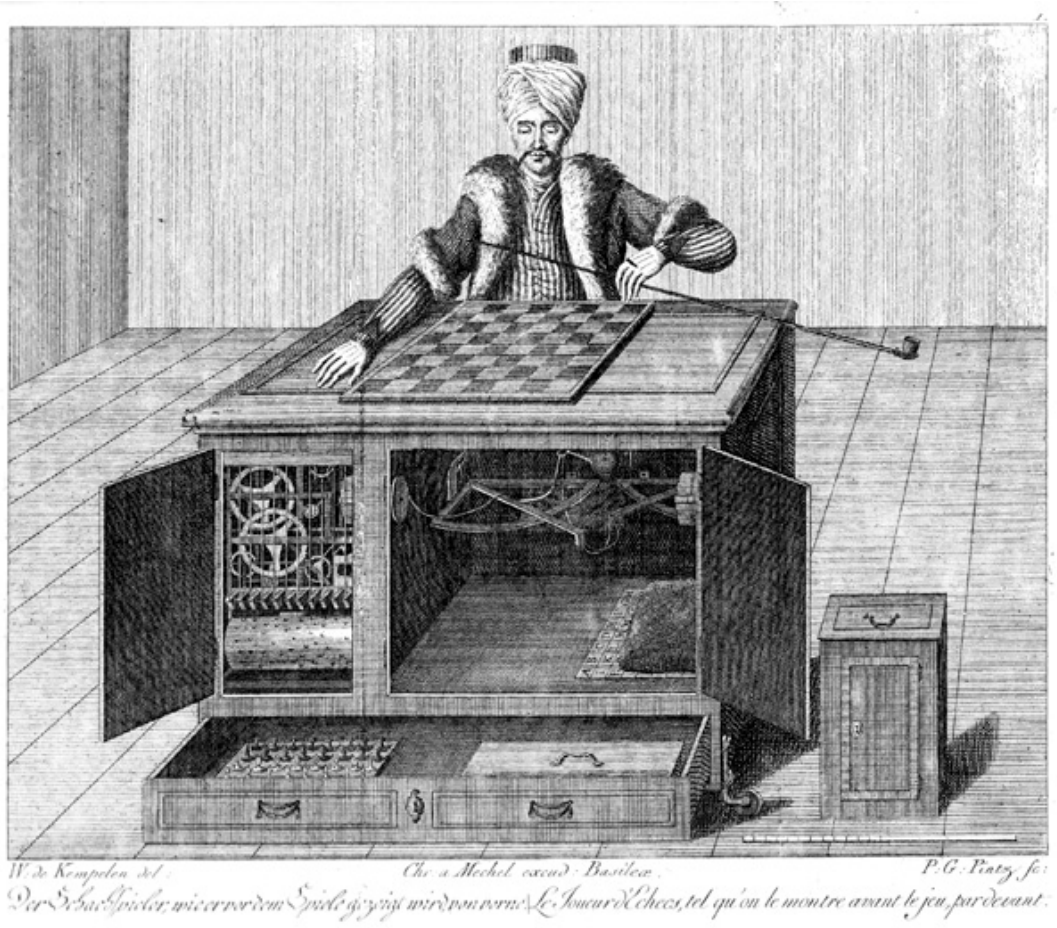
Testing for bias: the metrics

Various metrics exist that operationalize distinct definitions of „fairness“.

- **Group fairness**
- **Conditional statistical parity**
- **False positive error rate balance**
- **Overall accuracy equality**
- **Well-calibration**
- **Fairness through unawareness**
- **Counterfactual fairness**
- ***and many more***

prEN 18283 will provide an overview of available metrics, and the process by which they should be chosen

Bias and Beyond



Mechanical Turk *(Schachtürke)*
Constructed 1770 by Wolfgang von Kempelen

Copper engraving from: *Briefe über den Schachspieler des Hrn. von Kempelen, by Karl Gottlieb von Windisch* Public Domain

Agenda

1. What is bias?
2. Where is bias?
3. Testing for bias
4. **Summary**

Bias testing: Key Takeaways

- Have all groups potentially at risk from bias been identified?
- Are those groups properly represented in the test data?
- Are all functional correctness metrics broken down by groups?
- What bias metrics and thresholds were used, and is the choice justified?
- ... and what about bias mitigation?

Bias testing: Key Takeaways

- Did the risk management process identify all groups potentially at risk from bias?
- Are those groups properly represented in the test data?
- Are all functional correctness metrics broken down by groups?
- What bias metrics and thresholds were used, and is the choice justified?
- ... and what about bias mitigation? **Another topic, another talk ...**

Thank You

Let’s keep the conversation going!

rania.wazir@leiwand.ai
www.leiwand.ai