UNIVERSITEIT TWENTE.

Hackathon 11-13 Jun 2024

FAILURES

HOW WE MAKE AI FAIL

Inspiration session 12 Jun 2024

Maurice van Keulen

AGENDA

- What is AI and how does it work?
- What is so hard about data & AI?
  - AI can fail in so many ways
  - How does it fail? → on three levels
- How to **prepare** developers, users and decision makers for a life of responsible use of data and AI?

WHAT IS AI?

Artificial intelligence =

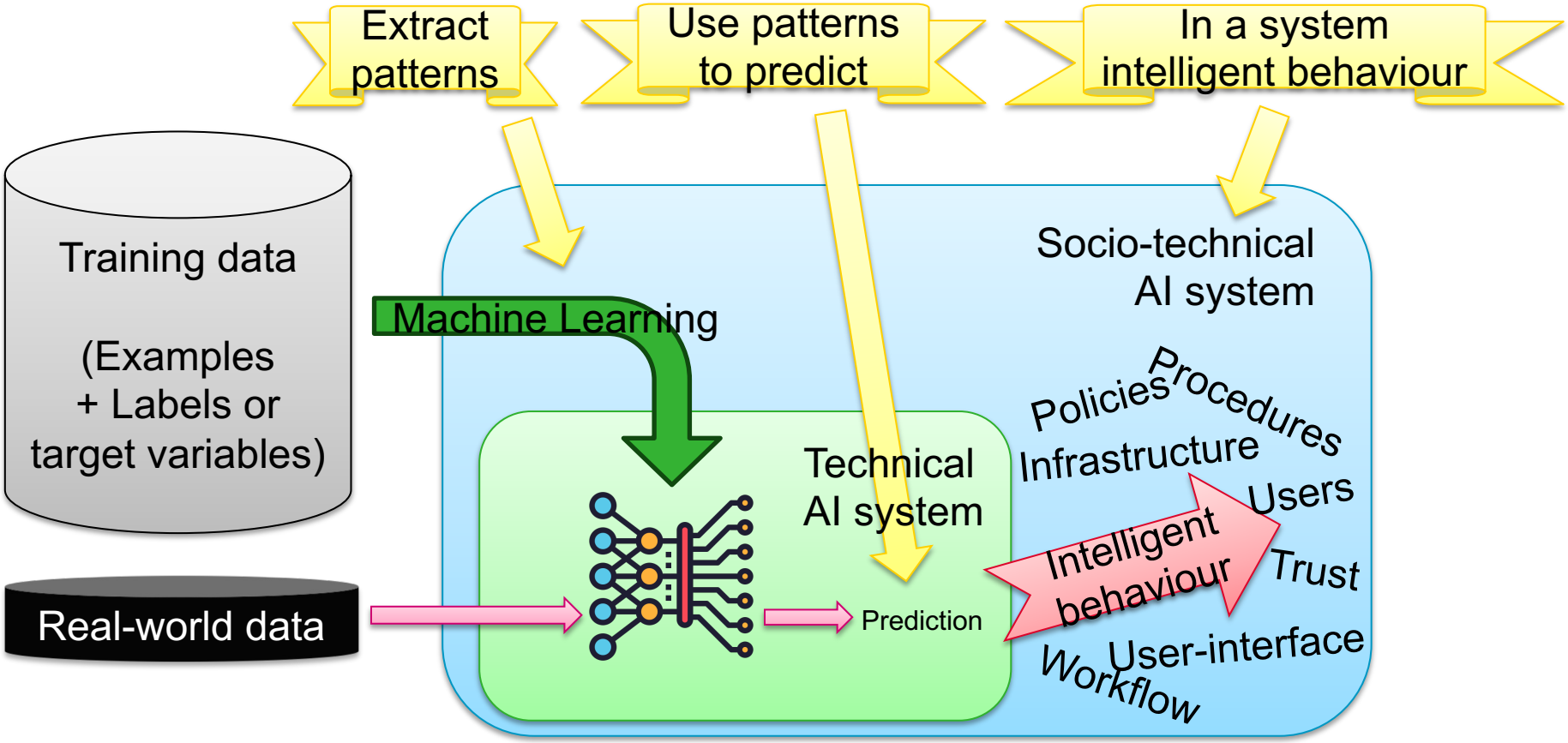
Machine Learning + (a lot of) data + embedded in a system/body

Machine learning:

- Can extract ***patterns*** in data
- Use these patterns to ***predict*** things
- In a system can produce ***intelligent*** behaviour

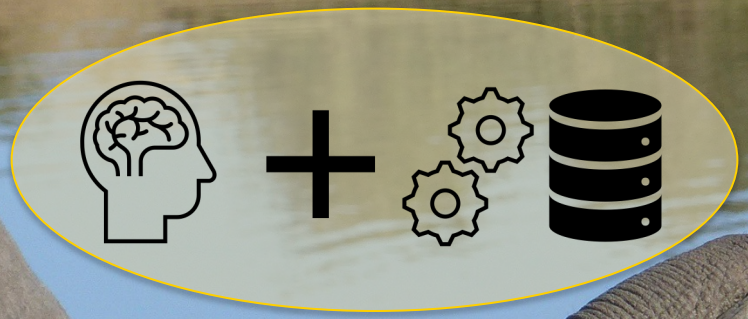
It's a kind of magic

HOW DOES AI WORK?



AI in Health dream

Quality



Costs

Clinical decision making
Diagnosis & Treatment

WHAT CAN AI MEAN FOR YOU?

If you would be given a small army of lepricons, what would you use them for?
Intelligent, fast, no fatigue, cheap, narrow task, limited world knowledge


“Van een koude kermis
thuiskomen”

AI CAN FAIL IN SO MANY WAYS!

AI can fail in so many ways

- … in very subtle ways
- … in ways the developers often are unaware of
- … in ways we are only discovering now

If AI fails, it is **our fault**

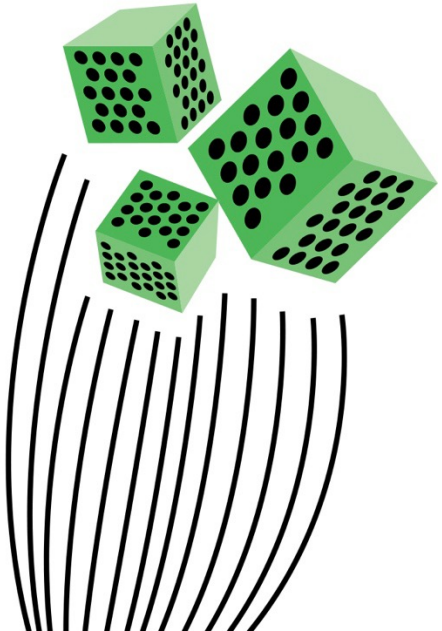
- AI is a thing
- We make it
- We use it

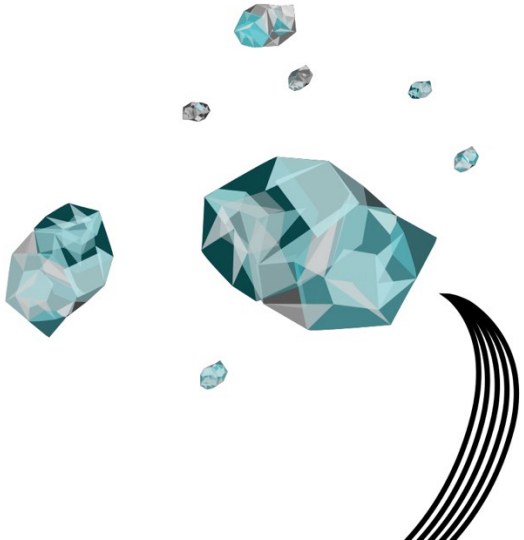
I'm a strong believer in AI … but also in that we should be careful and take things more slowly … think hard first!

UNIVERSITEIT TWENTE.

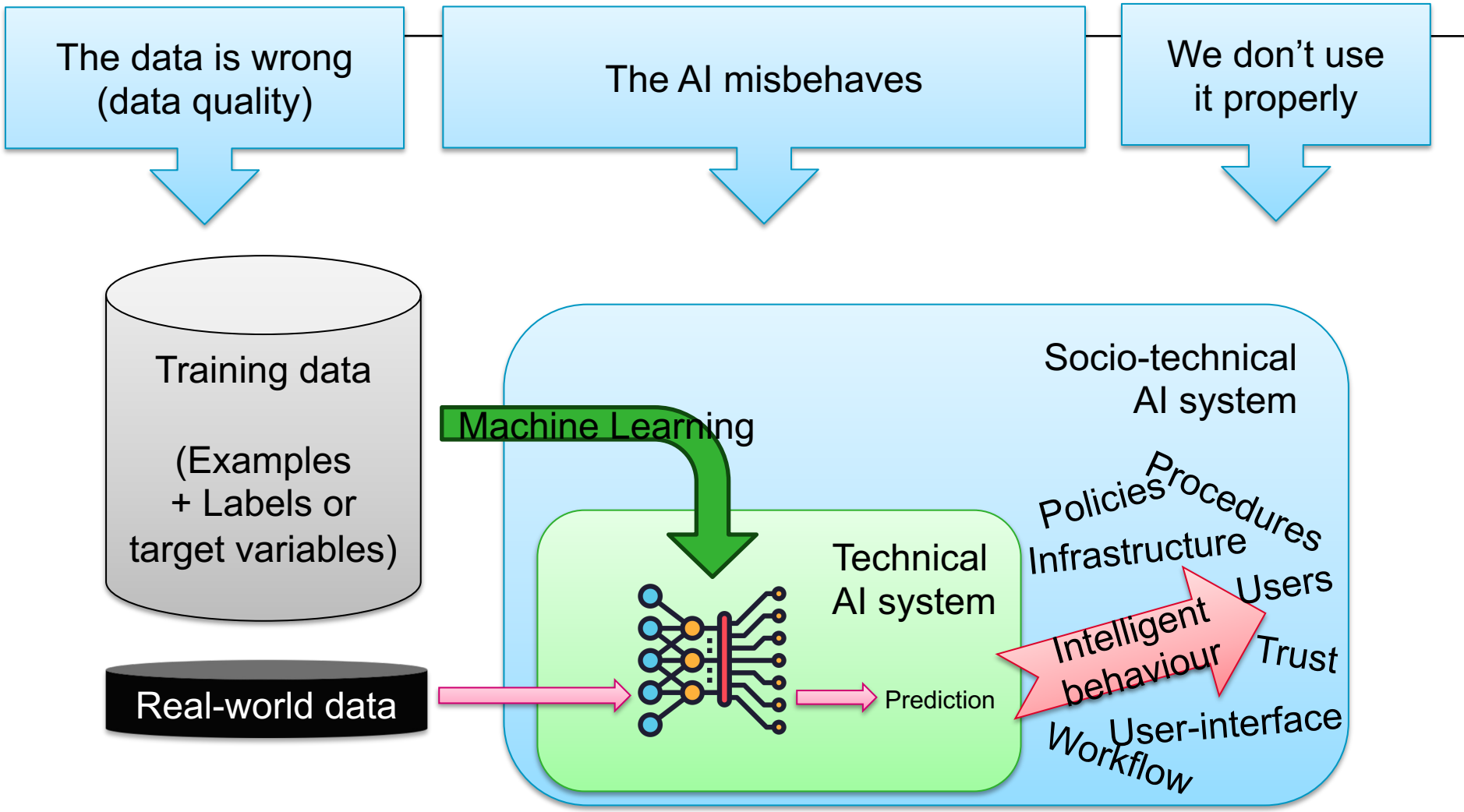
WHAT IS SO HARD ABOUT DATA & AI?

SO, HOW TO **PREPARE** DEVELOPERS, USERS & DECISION MAKERS
FOR A LIFE OF RESPONSIBLE USE OF DATA AND AI?





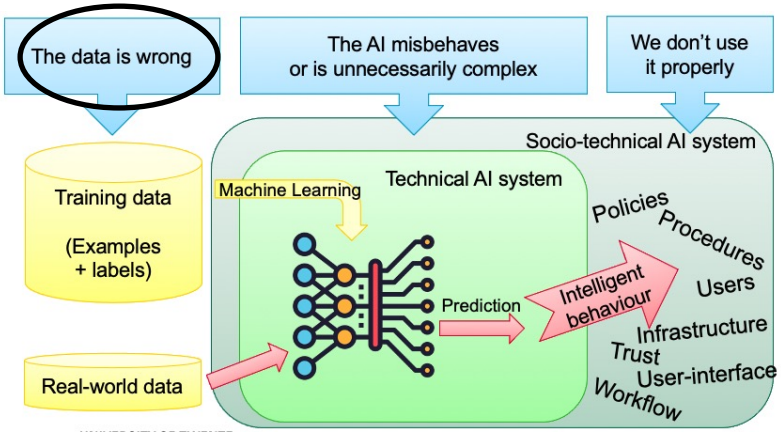
HOW DOES AI WORK? → HOW DOES AI FAIL? → ON THREE LEVELS!



Struggle for high-quality labels

Multiple sources
- Radiology reports
- Pathology reports
- Financial codes (DBC)

Granularity
- Region-of-interest labels
- Per-image labels
- Per-patient labels
- Malignant/benign vs BIRADS



Automatic quality improvement
- Remove text
- Cut out breast
- Resolution & contrast

Images

Selection of EHR records

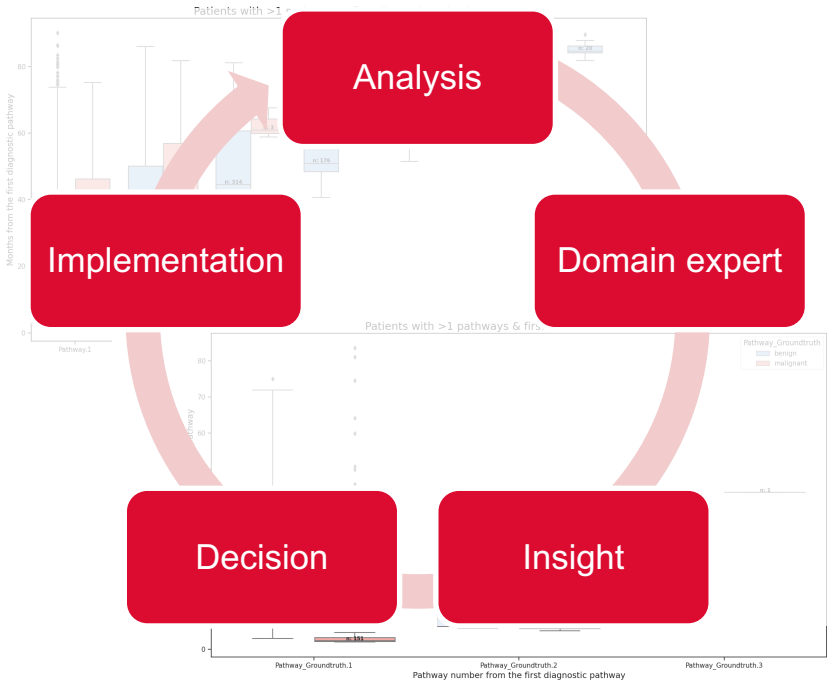
Include / Exclude
- YES: Diagnosis
- NO: Staging
- NO: After treatment
- NO: Follow-ups
- NO: Recurrence

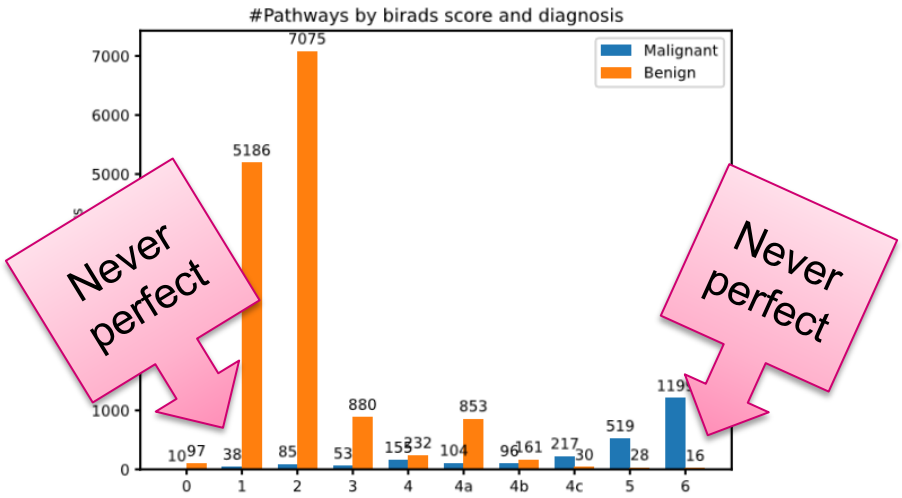
Example: Breast cancer diagnosis

Shreyasi Pathak, MSc

Data is wrong

STRUGGLE FOR HIGH QUALITY LABELS





UNIVERSITEIT TWENTE.

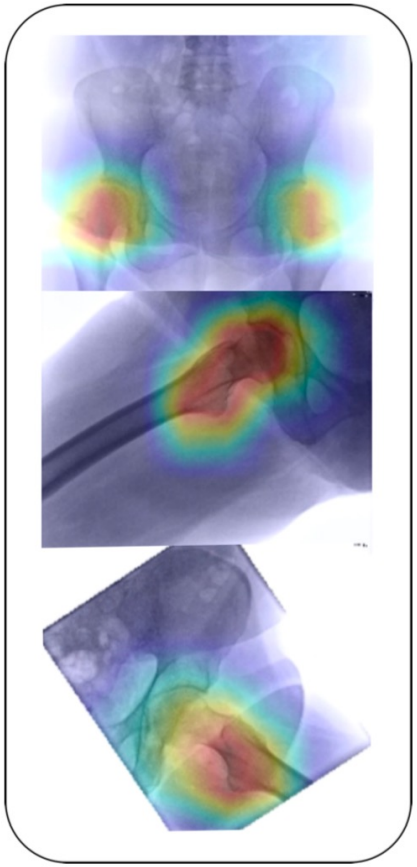
AI is misbehaving

SHORT CUT LEARNING
EXAMPLE: CONTRACT RESEARCH WITH ZGT HOSTPITAL

Goal: Check literature with own data

- Hip-fracture detection
- Literature: accuracy 95%
- ZGT X-rays: accuracy 93% (different machines, zoom levels, viewpoints, implants, etc.)
- Standard k-fold cross validation

Hurray! It works!



(b) No fracture images

ADDITIONAL VALIDATION

But does it really always detect the fractures?

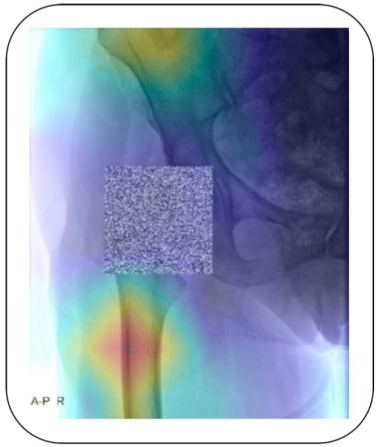
Additional validation approach

- Erase fracture from image + put image back into model => 25% model still detects a fracture

➢ What is going on here?

Further investigation with attention techniques

- Technique shows model focuses on skin and pubic area ?!?



DOMAIN EXPERT (RADIOLOGIST)

Input: X-Ray; Label: Patient ***has*** fracture.
≠ fracture is well visible in the X-ray

Domain expert suspects the following model reasoning

- Wrinkled skin, stool in diaper => highly elderly person
- Elderly person is getting an X-Ray of hip taken
  - Guess what this person is in for?

Short-cut

- Model ***cheats***: it uses **proxies** for high likelihood of fracture even though there is none to be seen
  - Will give high scores on accuracy, even on test set

If we wouldn't have done the extra validation, we would never have known

BIAS

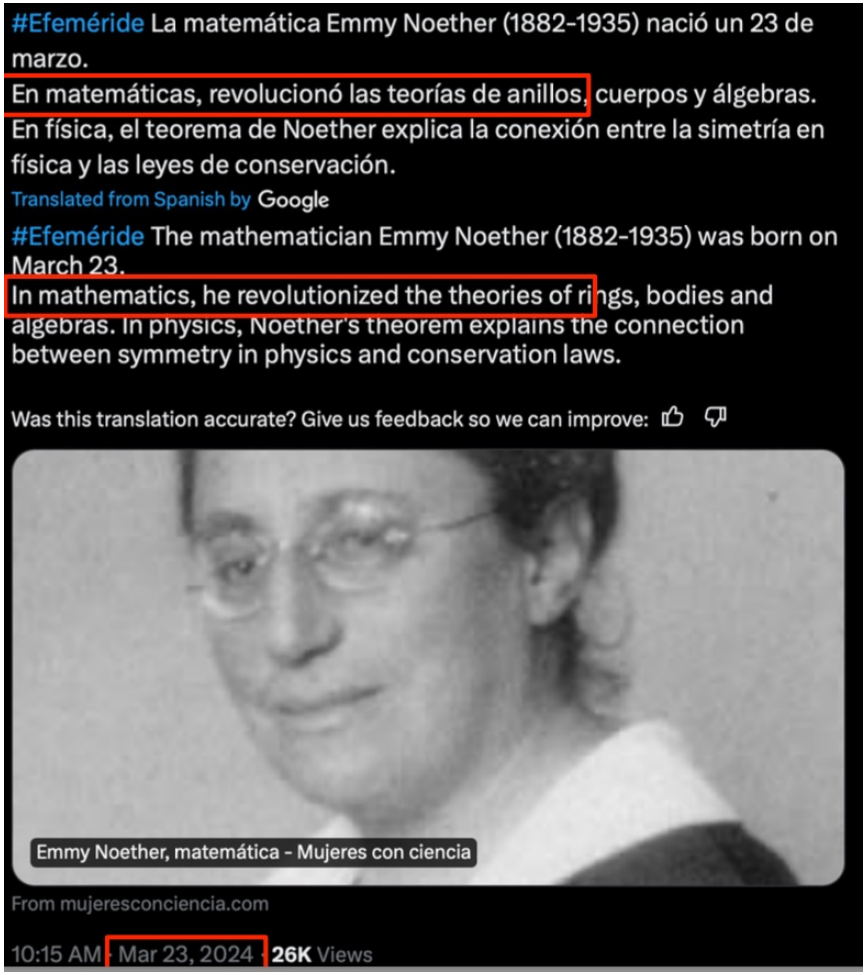
AI is misbehaving …

Spanish text: no gendered pronouns written in the text

Automatic translation needs to introduce gendered pronouns

Bias in English corpus: historical gender stereotypes

… due to bias in the data



Data is wrong

BIAS IN TRAINING DATA => BIAS IN MODEL

AI is misbehaving

PORTRAITAI.COM : MORGAN FREEMAN & LOUISE HUNG

Train with more pictures of white people => gives african-american and asian-americans facial characteristics of white people

BREAST CANCER DIAGNOSIS

AI is not used properly

Shreyasi Pathak, MSc

Objective

Given mammography, determine malignant vs. benign

- Most literature focuses on ML models of the form
  Input: 4 standard mammography images
  Output: classification prediction “malignant” or “benign”

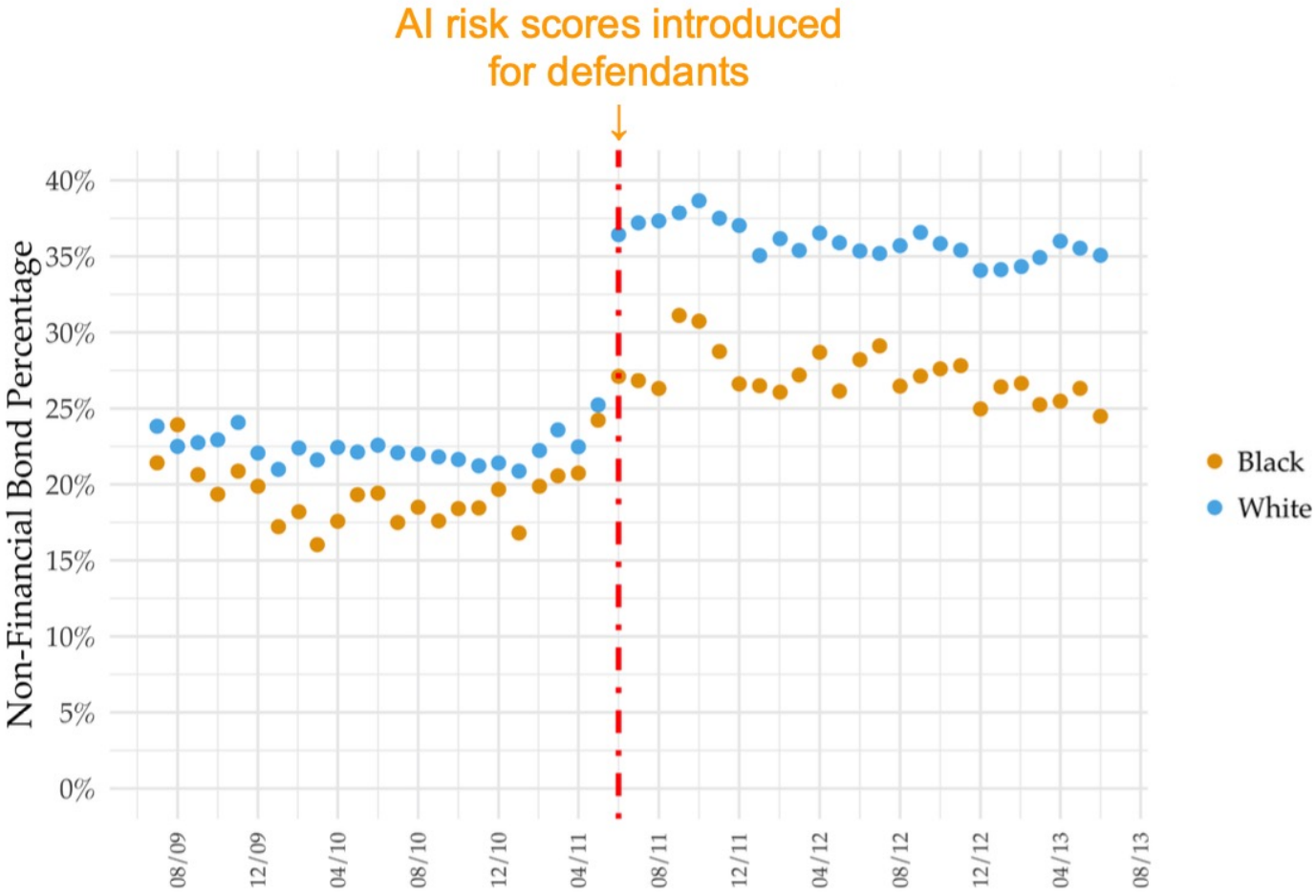
But

- Real hospital ZGT: 7% patients do not have 4 images
- Clinician is not interested in “malignant” or “benign”, but more/stronger evidence fore or against malignancy

➢ Everyone seems to build the wrong model!

IF YOU GIVE A JUDGE A RISK SCORE

AI is not used properly



AI risk score introduced a new bias

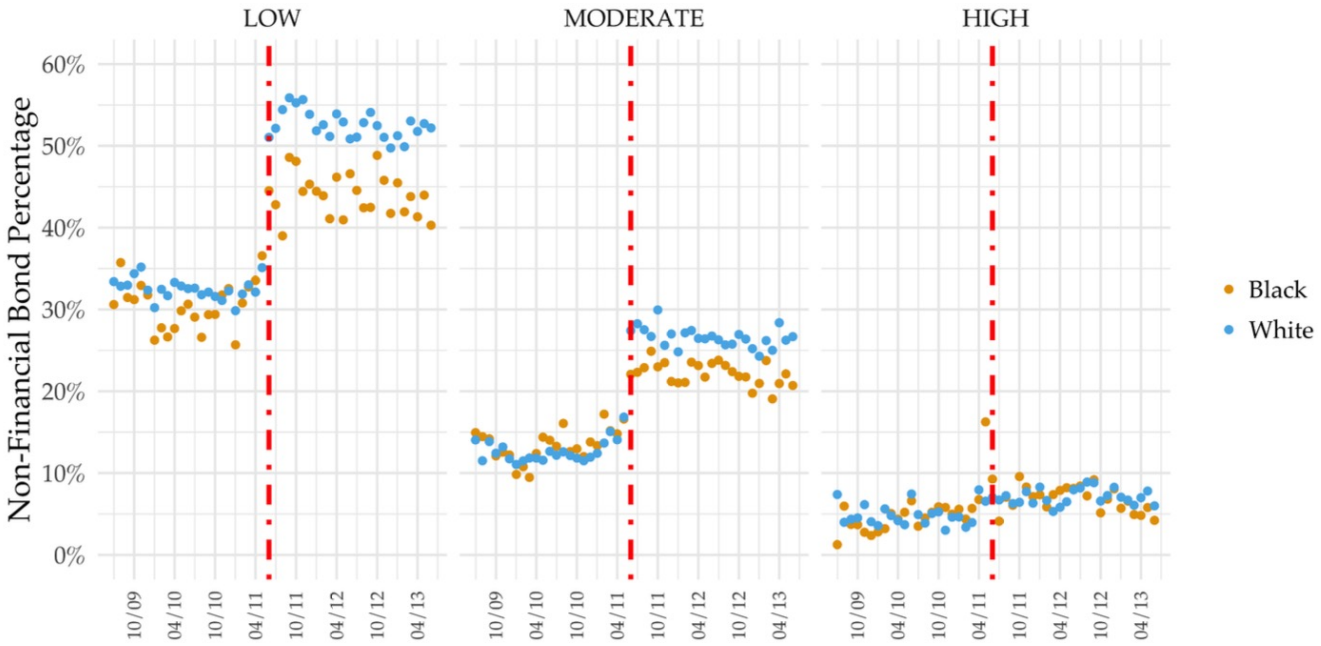
Is this bias in the AI? (in the AI's risk score?)

IF YOU GIVE A JUDGE A RISK SCORE

AI is not used properly

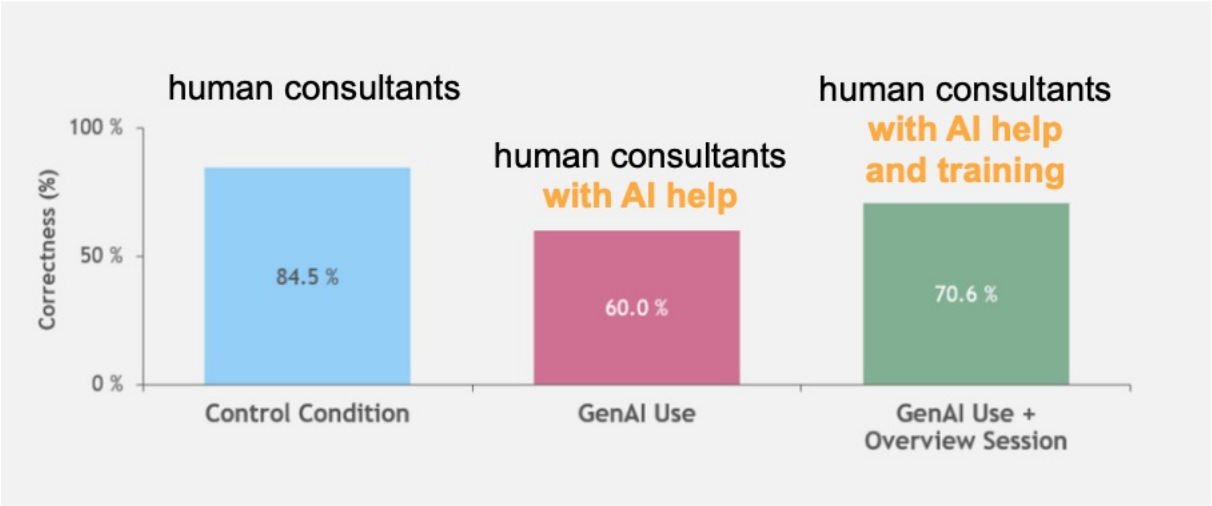
NO: the same risk score gives different outcomes

- LOW risk scores are **more overridden by judges** for black defendants



AI is not used properly

USING GENERATIVE AI
EFFECTS ON KNOWLEDGE WORKER PRODUCTIVITY&QUALITY



GenAI can give wrong but convincing answers

People using high-quality AI can become lazy & careless, they let the AI take over …

- … and produce worse decisions than without AI

Using AI hurts human skill development

USING GENERATIVE AI
USER'S INTENT

AI is not used properly

User intended historic reality, not to hallucinate

Generative AI is also used a lot for creating art …
… here you do want the AI to hallucinate

(Early 2024)

UNIVERSITEIT TWENTE.

HOW TO PREPARE

DEVELOPERS, USERS & DECISION MAKERS

FOR A LIFE OF RESPONSIBLE USE OF DATA AND AI?

HOW TO AVOID AI FAILURES?

I believe in

Explainable AI

Broader validation beyond measuring for predictive performance

Socio-technical design

and a critical attitude

EXPLAINABLE AI (XAI)

Some reasons for XAI:

- Developer needs to be able to ‘debug’ model
- User needs to be able to understand a model’s weaknesses

… also this one can do wrong!

Typical XAI is **post hoc**: train model, then try to explain

- Explanation doesn’t always faithfully explain the model
- Explanation doesn’t show mistakes
- Explanation isn’t understandable (e.g. SHAP)

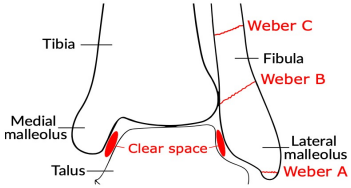
PIP-NET: AI EXPLAINABLE-BY-DESIGN

LEARNING A MODEL THAT HUMANS CAN UNDERSTAND AND CORRECT

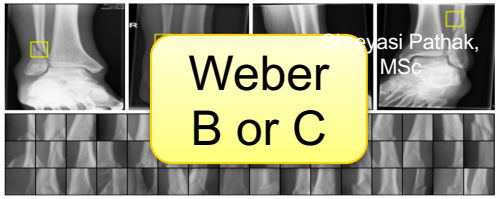
Meike Nauta, PhD

Interpretable Machine Learning

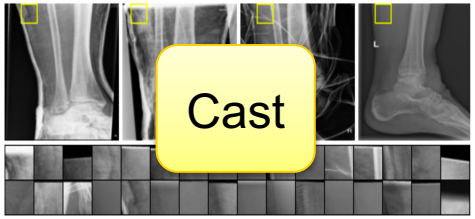
- Learns high-level 'prototypes' that humans can understand
- Also learns at-the-same-time a simple model based on these prototypes

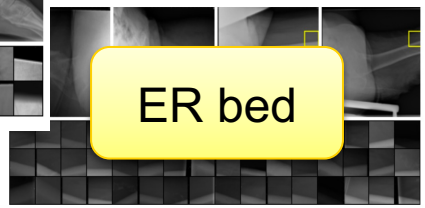


(a) Classification standard for ankle fractures



(b) Prototype relevant to *fracture*, corresponding to Weber B and, with lower presence scores, to Weber C





Ankle fractures

- PIP-Net learns prototypes that co-incide with medical standards
- Radiologists recognize and remove short-cuts

Clinical reasoning + explanation without loss of predictive power

VALIDATION

Typical validation of AI models: assess performance

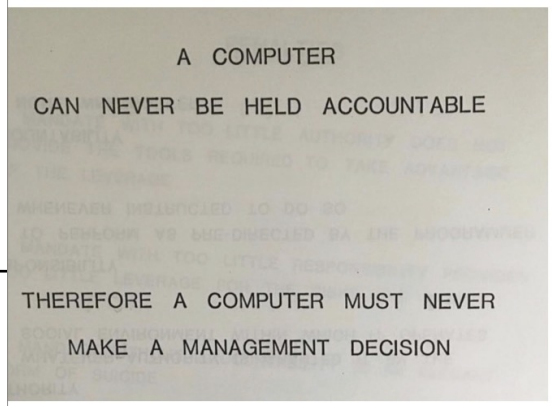
- **Independent test set** not used in development
  How often are the **predictions correct**?

One should also, for example, assess

- Origin, composition, quality, … of the training data
- Robustness against common disturbances (noise)
- Existence of common misbehaviour: short-cuts & bias
- Explainability: it has multiple facets (Co-12 framework)
- How model's predictions are being used by humans

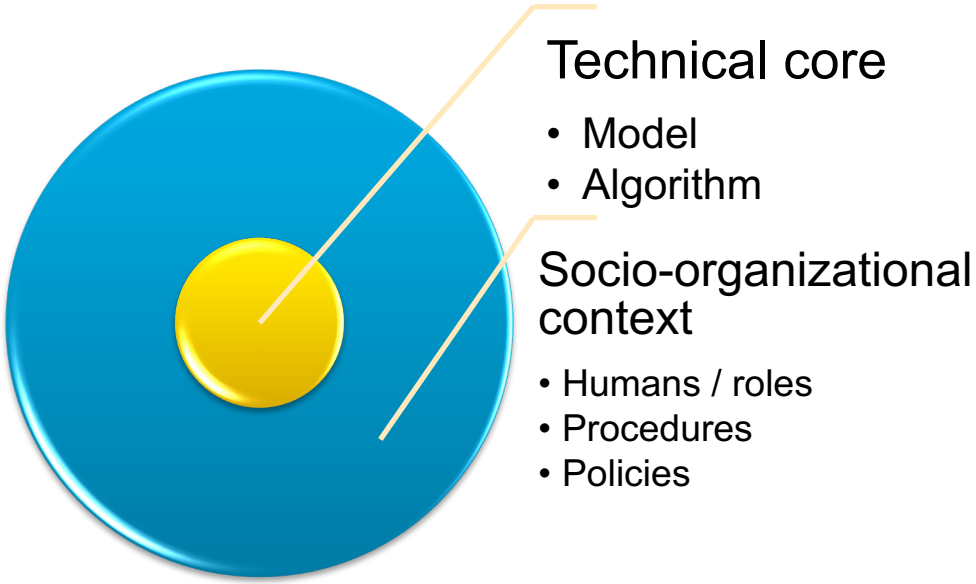
SOCIO-TECHNICAL DESIGN + INCLUDE ETHICAL DISCUSSION

(From a 1979 IBM presentation)



Ethics!

AI frameworks based on the concept of a **socio-technical system**



Benificence

Non-maleficence

Autonomy

Justice

Explicability

IN THE NETHERLANDS: SYRI
SYSTEEM RISICO INDICATIE

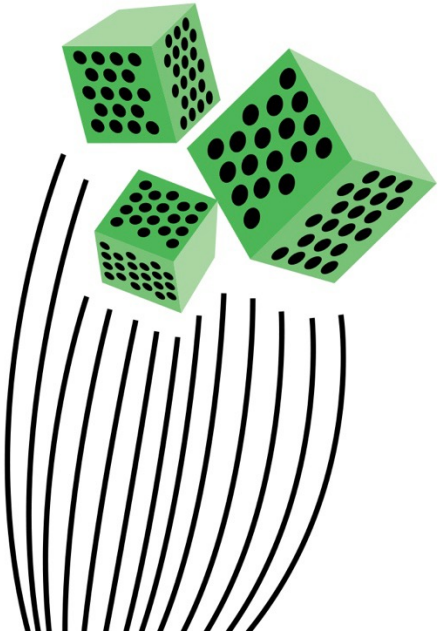
Public outcry was mostly directed not at the algorithm but at the flaws and opacity of the rest of the socio-technical system

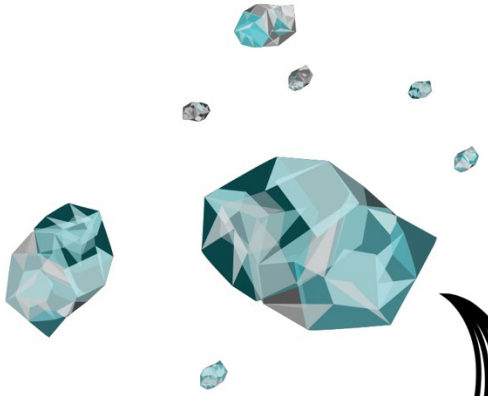
SYRI: Social security fraud detection algorithm of the Dutch government

- Received ironic privacy prize for the invasion of privacy of people
- Five main reasons (by Bits for Freedom)
  - Citizens were a suspect in advance
  - Felt like a violation of their privacy
  - Data used without purpose limitation
  - Might have been discriminating
  - Would be the first step towards a control society
- ➢ Parliamentary discussions
  - Too little attention was paid to ethical concerns in the design and realization of the system
  - Valid points raised by the public were insufficiently addressed

UNIVERSITEIT TWENTE.

CONCLUSION





CONCLUSION

What is so hard about data & AI?

- AI fails on three levels: data is wrong, AI misbehaves, we don't use it properly

How to **prepare** developers, users and decision makers for a life of responsible use of data and AI?

- Understand how AI can fail
- Train critical attitude
- Learn about XAI, broader validation, socio-technical design

If you are a domain expert help co-create AI solutions