Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Verantwortung der Informatik – Accountability In AI
WS 2023/2024

Kordian Gontarska, Weronika Wrazen, Felix Grzelka, Andreas Polze

about:us

Prof. Dr. Andreas Polze
Chair Representative

Andreas.Polze@hpi.de
Office: C-1.7

Kordian Gontarska
Doktorand

Kordian.Gontarska@hpi.de
Office: C-1.13

Weronika Wrazen
Doktorandin

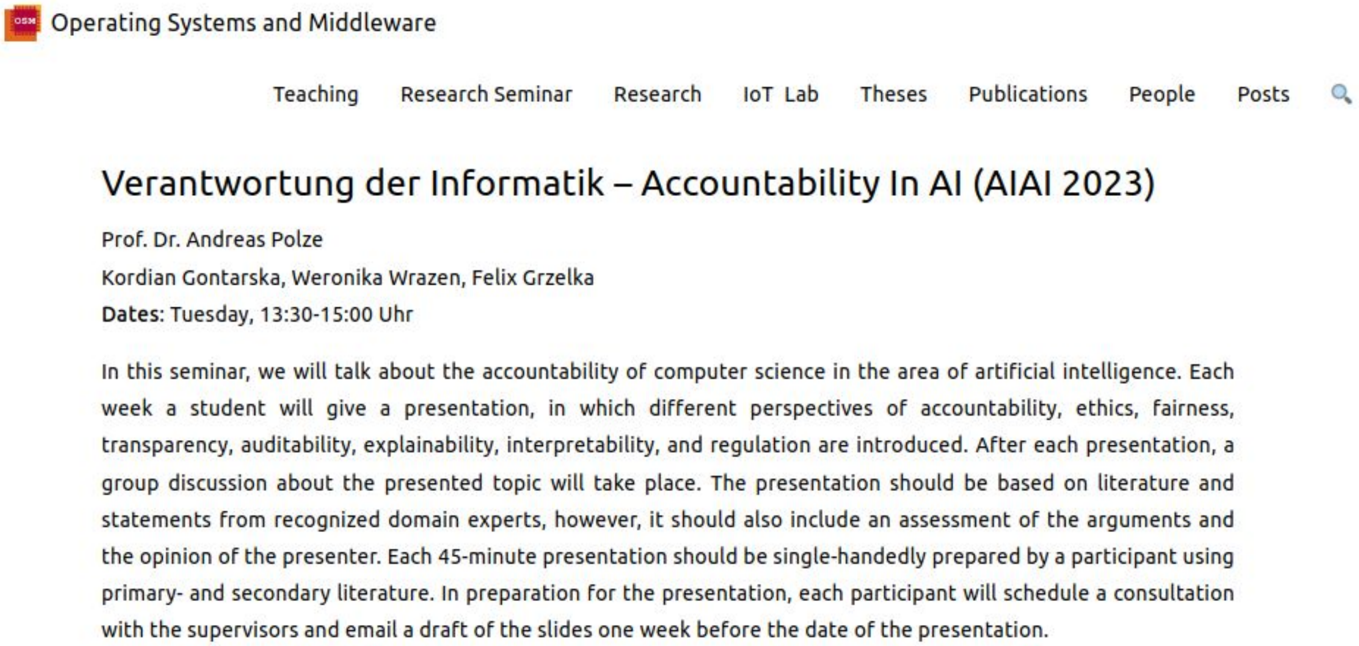
Weronika.Wrazen@hpi.de
Office: C-1.12

Felix Grzelka
Doktorand

Felix.Grzelka@hpi.de
Office: C-1.11

Seminar Website

https://osm.hpi.de/aiai



Verantwortung der Informatik – Accountability In AI (AIAI 2023)

Prof. Dr. Andreas Polze

Kordian Gontarska, Weronika Wrazen, Felix Grzelka

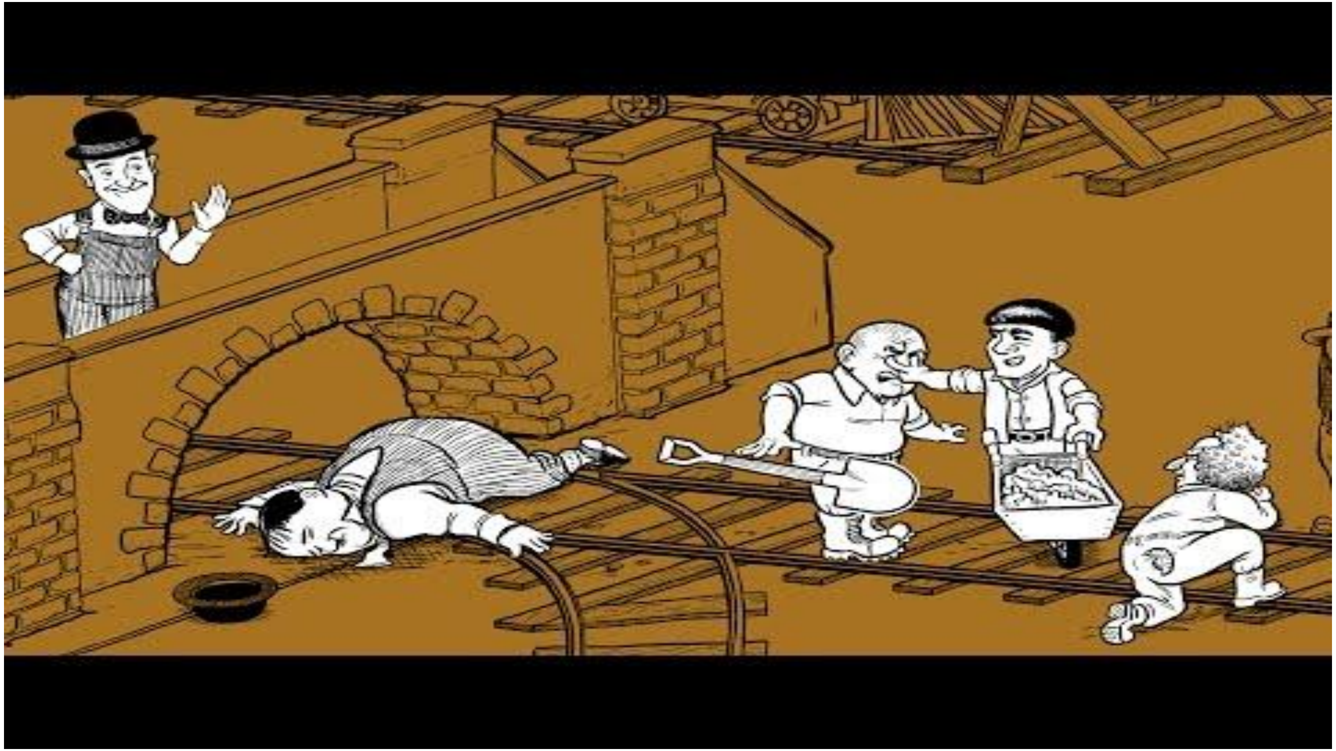
Dates: Tuesday, 13:30-15:00 Uhr

In this seminar, we will talk about the accountability of computer science in the area of artificial intelligence. Each week a student will give a presentation, in which different perspectives of accountability, ethics, fairness, transparency, auditability, explainability, interpretability, and regulation are introduced. After each presentation, a group discussion about the presented topic will take place. The presentation should be based on literature and statements from recognized domain experts, however, it should also include an assessment of the arguments and the opinion of the presenter. Each 45-minute presentation should be single-handedly prepared by a participant using primary- and secondary literature. In preparation for the presentation, each participant will schedule a consultation with the supervisors and email a draft of the slides one week before the date of the presentation.

Agenda

1. Introduction to ethical Artificial Intelligence (AI)
2. Seminar topics
3. Organization issues:
   - General organization
   - Time schedule
   - Grading
4. Questions (feel free to ask in between!)

Intro



Source: https://youtu.be/bOpf6KcWYyw

Ethics

1. Ethics is the conceptualization for a scientific discipline, which consists of reflecting of behaviour and judgments valid in a society.

2. Ethics is the reflection theory of morality, which investigates human actions and behaviour against certain judgments of good and bad or right and wrong in terms of morality.

(Grimm, Keber, Zöllner. 2019. Digitale Ethik.p. 9)

Ethics

Ethics	
Normative Ethics	**Descriptive Ethics**
study of ethical action	study of people’s views about moral beliefs
analyses how people ought to act	analyses people’s moral values, standards and behaviour
attempts to evaluate or create moral standards and prescribes how people ought to act	describes how people behave and what types of moral standards they claim to follow

Questions in Ethics

1. What is good and right?
2. What should people's behaviour and life look like?
3. Which actions are allowed and which are forbidden?

Ethics in AI

Ethical challenges in AI

Privacy

Explainability

Safety

Manipulation

Transparency

Fairness

To overcome mentioned challenges scientists deploy

Responsible AI.

Responsible AI

Responsible AI is a governance framework aimed at define a clear approach to using AI. It defines how an organization is addressing the challenges from both an ethical and legal point of view.

Responsible AI aims to create **accountability** for AI systems.

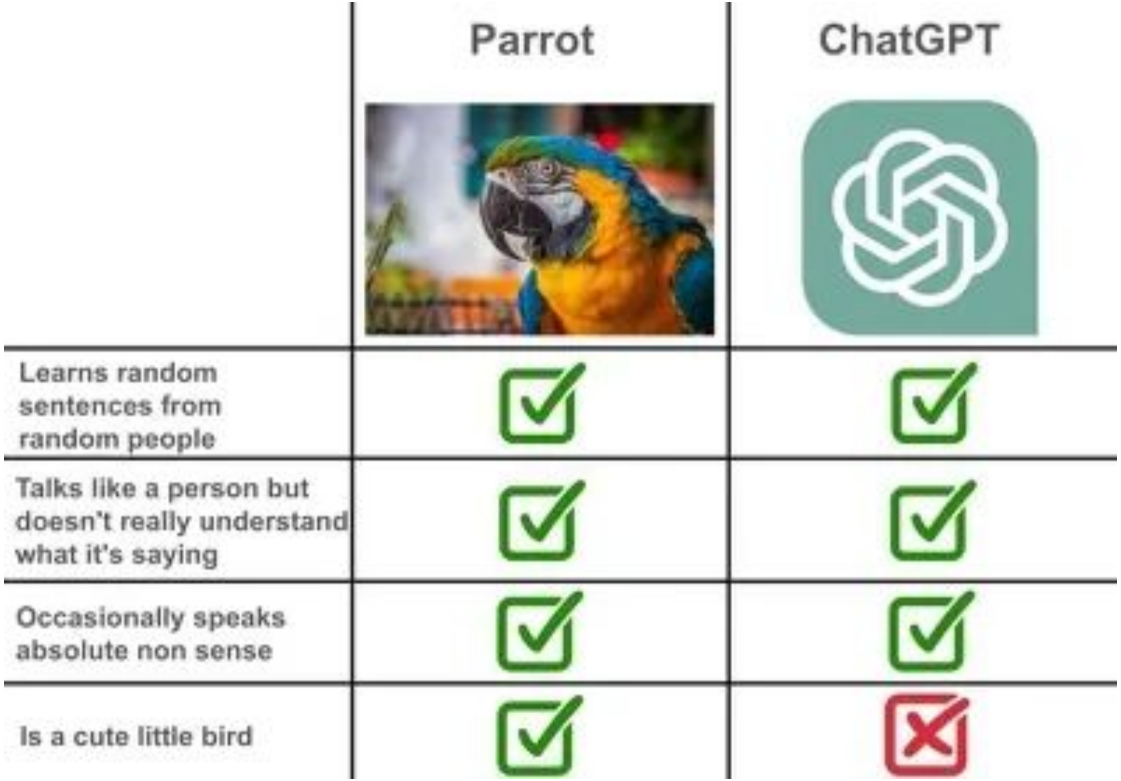
Responsible AI includes fairly and reliable practices of:

- designing,
- developing, and
- deploying of AI

Possible Topics

- Effects of Chat GPT & other LLMS
- Generative AI vs. Artists
- Panic as a Service
- State of Regulation
- Synthetic Media - Do DeepFakes Endanger Democracy?
- Fairness / Bias / Data diversity
- Auditing for AI for health
- Recommender Systems
- GDPR vs. Big Data (in Medicine)
- Uncertainty quantification for machine learning models
- Explainable Artificial Intelligence (XAI)
- Dual-use technology
- Energy consumption of AI
- **… your own topic!**

Effects of Chat GPT & other LLMS



	Parrot	ChatGPT
Learns random sentences from random people	☑	☑
Talks like a person but doesn't really understand what it's saying	☑	☑
Occasionally speaks absolute non sense	☑	☑
Is a cute little bird	☑	☒

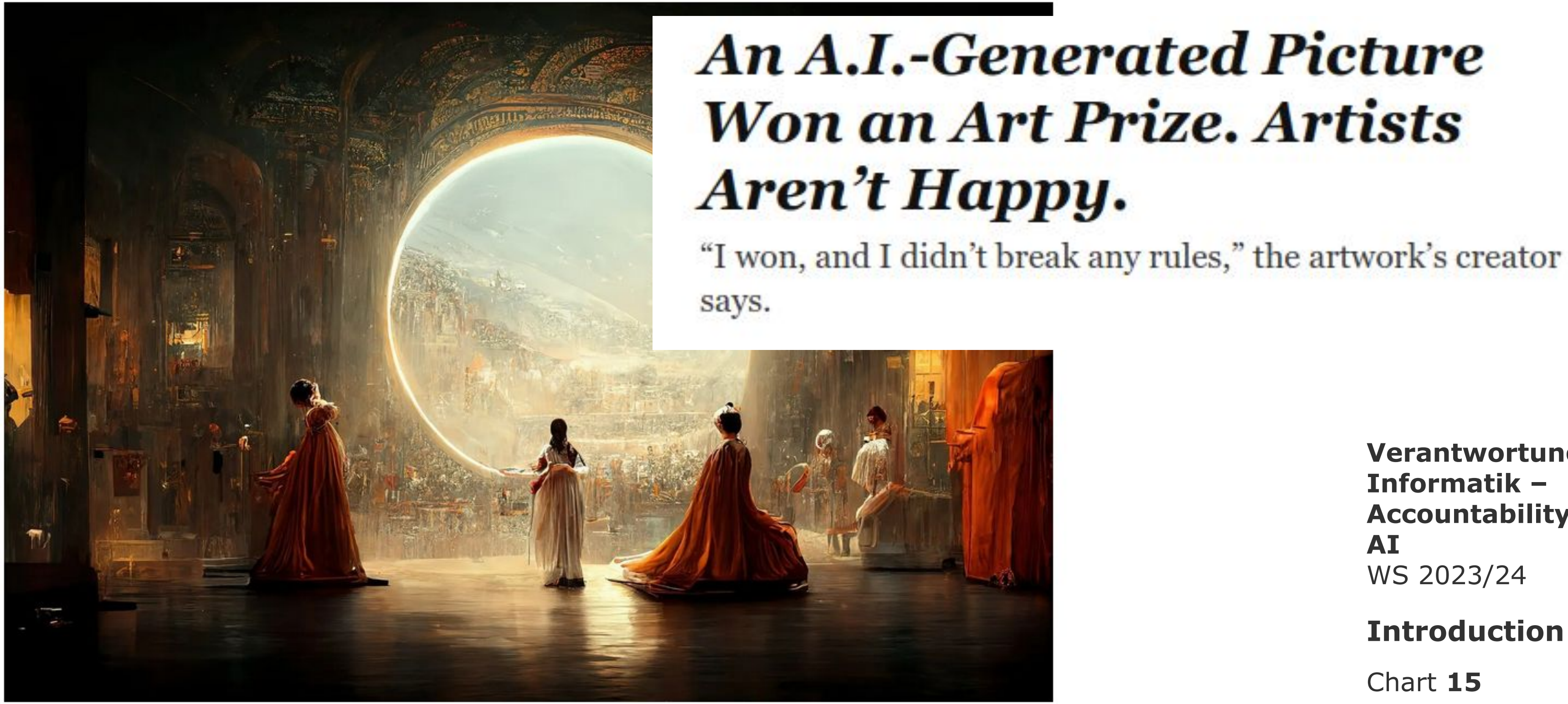
source: https://www.reddit.com/r/ChatGPT/comments/138j7a3/chatgpt_vs_parrot/

Effects of Chat GPT & other LLMS

- Chat GPT & co. have gained a lot of attention
- They show promising results for many tasks
- but fail in unexpected ways

- presentation content:
  - energy consumption
  - dual use concerns
  - GDPR compliance
  - plagiarism
  - communication shifts
  - hallucination

Generative AI vs. Artists



An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy.

"I won, and I didn't break any rules," the artwork's creator says.

Jason Allen's A.I.-generated work, "Théâtre D'opéra Spatial," took first place in the digital category at the Colorado State Fair. via Jason Allen

source: https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html

Generative AI vs. Artists

- Images generated by Midjourney and Stable Diffusion are becoming more prevalent
- They have won art contests
- Models were trained on millions of images, without explicit consent

- presentation content:
  - copyright infringement (even watermarks are reproduced)
  - copyright of generated images
  - is this art?
  - who pays the artists?
  - is it just a tool (like photoshop)?

Panic as a Service

- Panic (as in AI will take over the world) is used to
  - sell products/increase engagement
  - distract from the real harms

- presentation content:
  - proposed ai moratorium
  - theoretical threats vs. real harm
  - use of AI to amplify and spread panic

State of Regulation

- To mitigate risks of AI:
- many governments plan to regulations

- presentation content:
  - overview of existing and planned AI regulations
    - Germany
    - EU
    - Worldwide
  - aspects of regulations
  - pro & cons
  - impact on ai research and development
  - stakeholders involved in regulation attempts

Synthetic Media - Do DeepFakes Endanger Democracy?

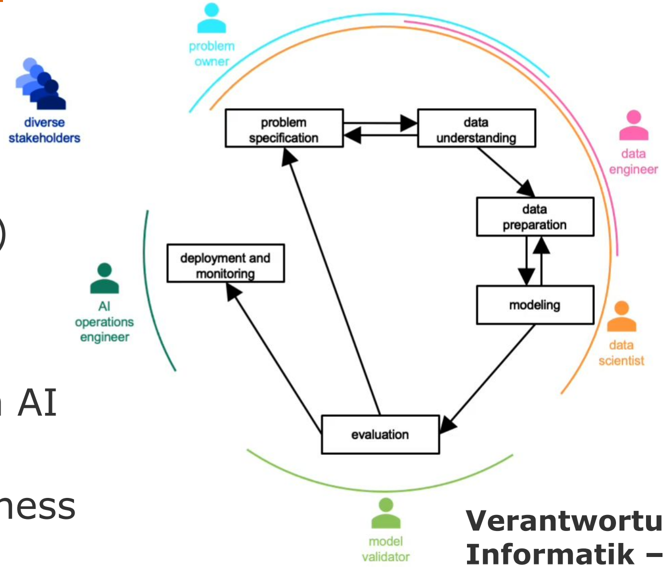
- Artificial production, manipulation, and modification of data and media
- DeepFakes, music synthesis, text generation, speech synthesis, etc.
- Technology is getting more accessible
- Is it dangerous?

- Presentation content:
  - Explain functionality of approaches
  - Find malicious examples

Bias / Fairness / Data Diversity

- Bias can appear in each step of ML pipeline
- Affects often minorities or disadvantaged subpopulations based on socio-demographic affiliations (ethnicity, age, sex, religion, income,...)

- Presentation content:
  - Explain where and how bias can appear in an AI pipeline (eg. categories of bias)
  - Provide examples how to measure bias / fairness (statistics, metrics, ...)



Auditing for ML for health

- Certifications ensure quality and standardization of products
  - Obtained through audit from a notified certification body (eg. TÜV)
- Trust can be provided through certificates, eg. masks during pandemic

- Presentation content: <u>Brief</u> description of
  - Software as a medical device (SaMD) + ISO norm 13485
  - What does have to be fulfilled to pass an audit?
- Discussion: Problems regarding ML in SaMD → Why is it difficult to regulate?
  - Focus on evaluation of ML

Recommender Systems

- RS are used by most popular (social media) websites
- they provide personalized content suggestions and optimize engagement
- they can lead to filter bubbles and excessive use (addiction)

Presentation content:

- intro to RS and dangers
- propose possible solutions and discuss their pros & cons

GDPR vs. Big Data (in Medicine)

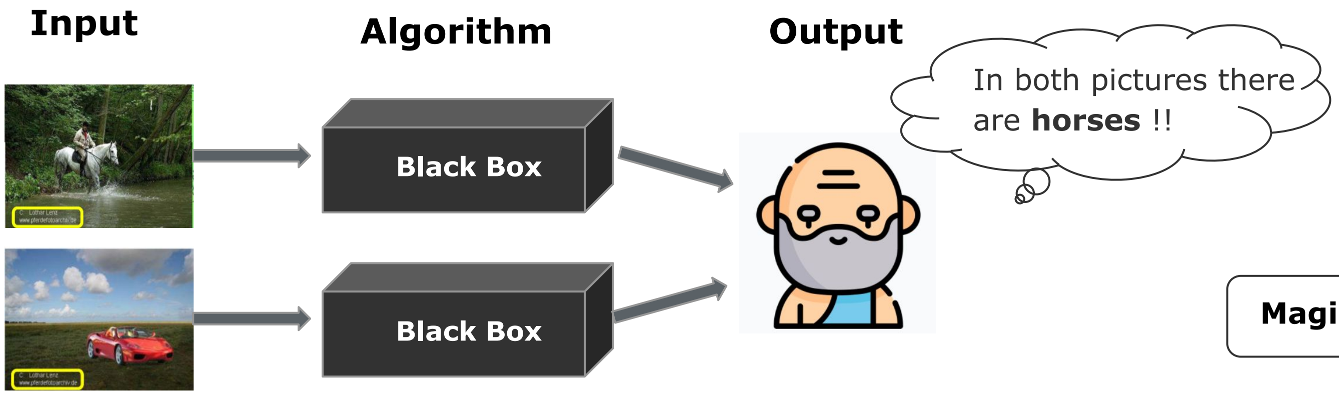
- big datasets are necessary to train big, state of the art models
- especially in medicine the protection of personal data (GDPR) is crucial

- Presentation content:
- intro to GDPR and how it relates ot ai in medicine
- how to implement data economy and the right to be forgotten in AI?
- how to protect the privacy of uninvoled thrid parties such as relatives, which share parts of their genome

Discussion:

- is it ok to give up individuals privacy to be able to train a model, which benefits all? how much privacy should we give up?

Explainable Artificial Intelligence (XAI)


Input
Algorithm
Output
© Lothar Lenz www.pferdefotoarchiv.de
Black Box
Black Box
In both pictures there are **horses** !!

Questions needed to be answered:

- Why did we receive such a result?
- Why did not we receive other result?
- How can we change the input to receive other outcome?
- When can I trust the model?
- How can we correct the errors ?

Source: S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, 'Unmasking Clever Hans predictors and assessing what machines really learn', Nat Commun, vol. 10, no. 1, p. 1096, Dec. 2019, doi: 10.1038/s41467-019-08987-4.

Explainable Artificial Intelligence (XAI)

Explainable artificial intelligence (XAI) is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms*.

It helps to define:

- accuracy of an algorithm,
- fairness of an algorithm,
- transparency of an algorithm, and
- reveal unknown relationships among inputs as well as between inputs and output.

Source: https://www.ibm.com/watson/explainable-ai

Uncertainty quantification for machine learning models

Machine learning models give as an output only **point estimation** (usually the average value).

As a result we **do not know** how the model is **certain** about the test result.

To overcome that limitation we need to estimate **confidence interval** for the prediction. Confidence interval tells us with defined probability that real value of the prediction lies into specified range.

Why it is important?

- it enlarge the trust in the prediction,
- it makes the model safer.

Sources of uncertainty:

- noisy data (measurement imprecision),
- too few observations in training dataset

Uncertainty quantification for machine learning models

<u>Example 1.</u> Why there is a need to add confidence level to the output?

The task is to define if the product is made from 18-carat gold (ρ1 = 15,5 g/cm3) or the cheaper one (ρ2 = 13,8 g/cm3). The measurements were conducted in 2 independent laboratories.

The results are: ρA = (15 ±1,5) g/cm3, ρB = (13,9 ±0,2) g/cm3. *

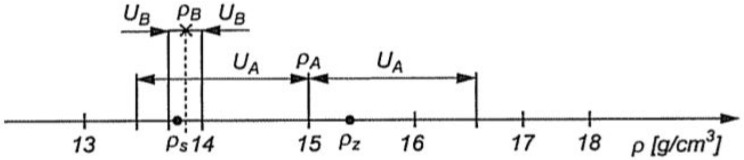


*Fig. Test results with confidence interval received at 2 laboratories.**

<u>Conclusion:</u>

Both results are reliable but only one - lab B - gives us **useful** results, based on which we can define the real class of the input.

[*] Arendarski, J. (2013). Niepewność pomiarowa. OWPW. Ed 3.

Dual-use technology

- "Dual use goods are products and technologies normally used for civilian purposes but which may have military applications." [1]
- AI can be used by the military (killer drones, facial recognition, spreading misinformation, etc.)

Presentation content:

- present some AI products/technologies that fall under dual use, show how they can be used in civil and in military settings
- argue how or if the military use of each technology can or should be avoided (regulation, ban on certain fields of research?)

[1] https://ec.europa.eu/trade/import-and-export-rules/export-from-eu/dual-use-controls/index_en.htm

Energy consumption of AI

- Training deep neural networks is energy-intensive.
- training GPT-3 required ~ 190 000 kWh

- Presentation content:
  - select a few state of the art NN, try to find estimates on their energy consumption, compare them to previous NNs and classical methods
  - argue whether or not the gains in accuracy are worth the increase in energy consumption

Presentation Structure

1. Introduction
   - What is the topic?
   - How is it defined?
   - Are there multiple, different definitions?
   - Why is it important?

2. Main Part
   - Present a method/paper/tool which addresses the problem
   - Check topic description/literature section for some suggestions
   - Explain the main idea
   - Highlight benefits and potential shortcomings

3. Discussion
   - Provide 2-3 points/questions to start the interactive discussion

General Organization

Tuesdays

13:30 - 15:00 Uhr

Seminar room: K-1.03

Choose Your Topic

Deadline

24.10.2023 (Tuesday next week)

send us a short mail (see about:us or website)

your topic & 2 sentences

first come first served

teamwork is possible

well try to make it work for everyone

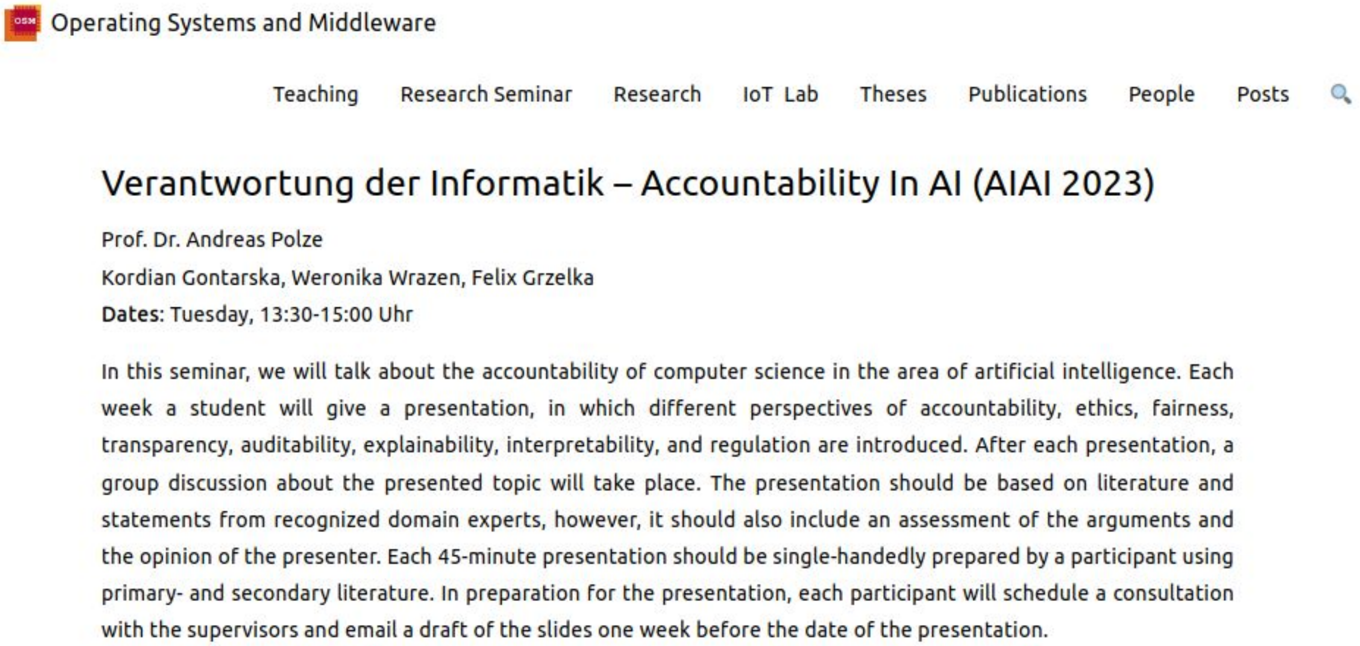
Grading

3 ECTS

presentation + slides

Seminar Website

https://osm.hpi.de/aiai



Operating Systems and Middleware

Verantwortung der Informatik – Accountability In AI (AIAI 2023)

Prof. Dr. Andreas Polze

Kordian Gontarska, Weronika Wrazen, Felix Grzelka

Dates: Tuesday, 13:30-15:00 Uhr

In this seminar, we will talk about the accountability of computer science in the area of artificial intelligence. Each week a student will give a presentation, in which different perspectives of accountability, ethics, fairness, transparency, auditability, explainability, interpretability, and regulation are introduced. After each presentation, a group discussion about the presented topic will take place. The presentation should be based on literature and statements from recognized domain experts, however, it should also include an assessment of the arguments and the opinion of the presenter. Each 45-minute presentation should be single-handedly prepared by a participant using primary- and secondary literature. In preparation for the presentation, each participant will schedule a consultation with the supervisors and email a draft of the slides one week before the date of the presentation.

Thanks for your attention!
Website: https://osm.hpi.de/aiai

kordian.gontarska@hpi.de

felix.grzelka@hpi.de

weronika.wrazen@hpi.de

Image References

- Icons made by Freepik, Wichai.Wi, Smashicons, from www.flaticon.com
- Icons made by Nikita Golubev" from www.flaticon.com
- Icons made by Eucalyp from www.flaticon.com
- Icons made by GOWI from www.flaticon.com
- Icons made by Flat Icons www.flaticon.com
- IBM course on ‘Mitigating fairness’
- https://www.businessdisabilityinternational.org/when-is-equality-not-equality/