



# Verantwortung der Informatik – Accountability In AI

## WS 2021/2022

Kordian Gontarska, Marta Lemanczyk, Weronika Wrazen, Felix Grzelka

# about:us



**Kordian Gontarska**

Doktorand

[Kordian.Gontarska@hpi.de](mailto:Kordian.Gontarska@hpi.de)

Office: C-1.13



**Prof. Dr. Andreas Polze**

Chair Representative

[Andreas.Polze@hpi.de](mailto:Andreas.Polze@hpi.de)

Office: C-1.7



**Marta Lemanczyk**

Doktorandin  
(Data Analytics and  
Computational  
Statistics chair)

[Marta.Lemanczyk@hpi.de](mailto:Marta.Lemanczyk@hpi.de)

Office: F-E.08

**Felix Grzelka**

Doktorand

[Felix.Grzelka@hpi.de](mailto:Felix.Grzelka@hpi.de)

Office: C-1.14



**Weronika Wrazen**

Doktorandin

[Weronika.Wrazen@hpi.de](mailto:Weronika.Wrazen@hpi.de)

Office: C-1.14



**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart 2

<https://osm.hpi.de/aiai>



Operating Systems and Middleware

teaching research seminar research iot-lab theses publications people posts 

## Verantwortung der Informatik – Accountability In AI (AIAI 2021)

Prof. Dr. Andreas Polze

Kordian Gontarska, Marta Stefania Lemanczyk, Weronika Wrazen, Felix Gzelka

Tuesday, 13:30-15:00 Uhr, A-1.1

### Abstract

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.

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart 3

# Agenda

---

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

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart 4

# Intro



Source: <https://youtu.be/bOpf6KcWYyw>

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  |  |
|---|--|
| 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?





**Explainability**



**Privacy**



**Safety**



**Manipulation**

**Ethical  
challenges in  
AI**



**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

---

- Privacy in Federated Learning
- Synthetic Media - Do DeepFakes Endanger Democracy?
- Fairness / Bias / Data diversity
- Auditing for AI for health
- Recommender Systems
- GDPR vs. Big Data (in Medicine)
- Legal and Regulatory Implications of Autonomous Vehicles
- Uncertainty quantification for machine learning models
- Explainable Artificial Intelligence (XAI)
- Dual-use technology
- Energy consumption of AI
- **... your own topic!**

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart 12

# Privacy in Federated Learning

---

- Federated Learning promises Data Privacy
- Data stays local and is safe
- Can data be reconstructed from model weights?
  
- Presentation content:
  - Describe Federated Learning
  - Find out about privacy problems
  - How can the problems be addressed?
- Possible project ideas:
  - Train a Federated model and try to extract input data

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **13**

# 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
- Possible project ideas:
  - Make someone say what you want
  - Let a news anchor tell some fake news

**Verantwortung der  
Informatik –  
Accountability In  
AI**

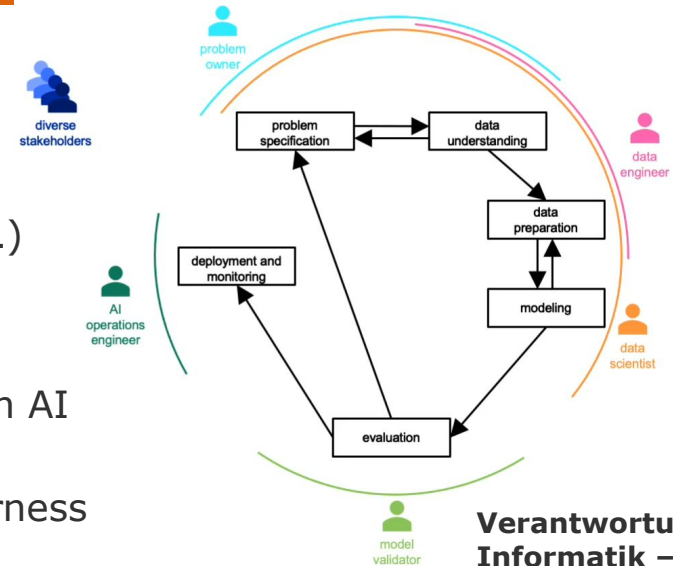
WS 2021/22

**Introduction**

Chart **14**

# 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, ...)
- Possible project ideas:
  - Application of fairness metrics on real-world data
  - Bias in Covid-19 genome sampling for sequencing



**Verantwortung der Informatik – Accountability In AI**

WS 2021/22

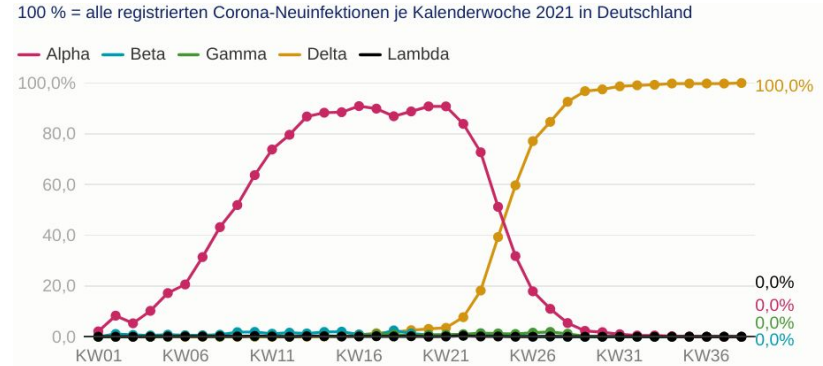
**Introduction**

Chart 15

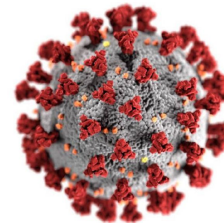


# Project: Covid-19 genome sampling for sequencing

- Different sampling strategies for Covid-19 genome sequencing between different countries
- Scientific bias vs. Socio-demographic bias for sampling
  - Scientists focus on specific variants (Variants of Interest, Variants of Concern)
  - Less resources in low-middle-income countries
- Possible projects:
  - Literature review: Which sampling strategies do different countries have?
  - Simulation of Sampling Strategy + influence on ML tools



Source: RKI



**Verantwortung der Informatik – Accountability In AI**  
WS 2021/22

**Introduction**

Chart 16

- 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: Brief 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
- Possible project ideas:
  - Investigation of problems specifically applying to genomic tools / algorithms containing ML (Requirement: Basic knowledge in genomics)
  - Application of guidelines on existing project / tool / algorithm in health



**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **17**

# 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

## Project Ideas:

- First, use a toy dataset to implement a conventional recommender system as a control condition. Run a small user study to measure user engagement and satisfaction. Then implement an improved version that is less addicting or not as prone to sustaining bubbles. Test your hypothesis by running a study with your improvements.

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **18**

# 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 to 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?

# Legal and Regulatory Implications of Autonomous Vehicles

## Levels of Autonomous Driving according to SAE J3016\*:

| Level 0   | Level 1   | Level 2  | Level 3  | Level 4   | Level 5   |
|---|---|--|--|---|---|
| An <b>active and engaged driver</b> is required.  |   |  | The <b>technology takes complete control</b> of the driving without human supervision.                               |   |   |
| Consists of <b>driver support</b> features.   |   |  | Consists of <b>automated driver</b> features.  |   |   |
| Features are limited to providing warning and momentary assistance.   | Features provide steering or brake/acceleration support.  | Features provide steering and brake/acceleration support.  | Features can drive the vehicle under limited conditions and will not operate unless all required conditions are met. |   | Features can drive the vehicle under all conditions.  |
| <ul style="list-style-type: none"> <li>automatic emergency braking,</li> <li>blind spot warning,</li> <li>lane departure warning</li> </ul> | <ul style="list-style-type: none"> <li>lane centering <b>OR</b></li> <li>adaptive cruise control</li> </ul> | <ul style="list-style-type: none"> <li>lane centering <b>AND</b></li> <li>adaptive cruise control</li> </ul> | <ul style="list-style-type: none"> <li>Traffic Jam Chauffeur</li> </ul>  | <ul style="list-style-type: none"> <li>local driverless taxi,</li> <li>pedals/steering may or may not be installed</li> </ul> | <ul style="list-style-type: none"> <li>same as level 4, but feature can drive everywhere in all conditions</li> </ul> |

# Legal and Regulatory Implications of Autonomous Vehicles



*Volvo AutoBreak System. Source:*  
[https://www.youtube.com/watch?v=\\_47utWAoupo](https://www.youtube.com/watch?v=_47utWAoupo)

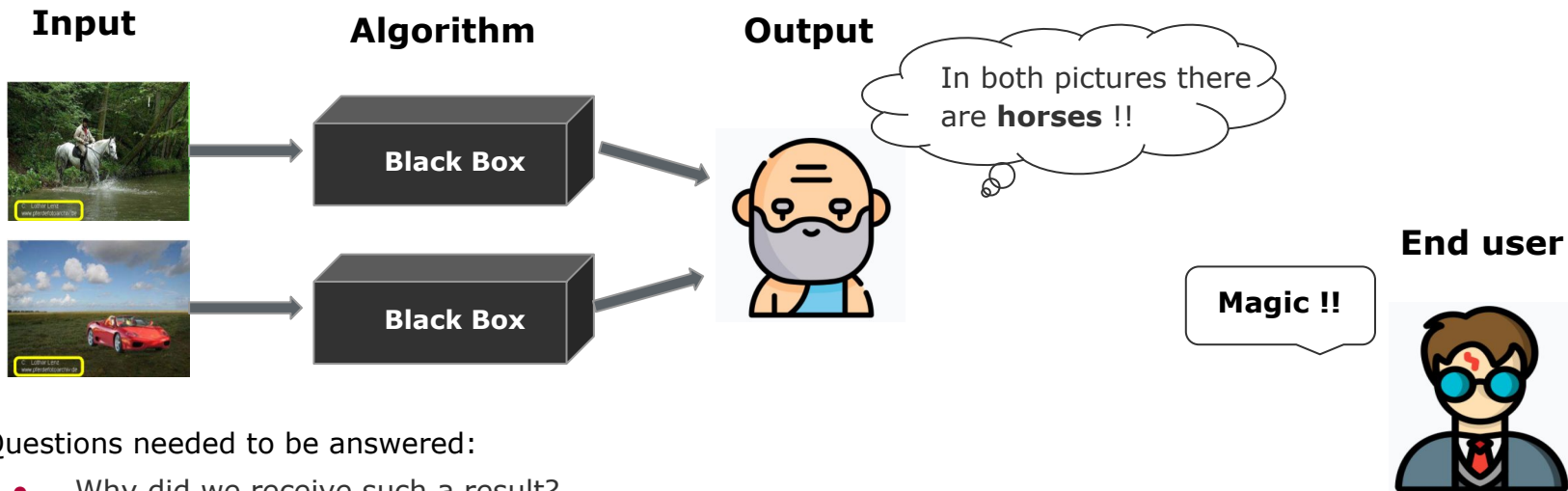
Challenges in development and deployment of autonomous vehicles:

- how to test the product,
- how to ensure safety of the product,
- how to handle data privacy and security, and
- who takes responsibility in case of an accident.

*To overcome above issues there is a need to formulate an appropriate **ethical and legal framework** as well as **testing standards**.*



# Explainable Artificial Intelligence (XAI)



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 ?

Chart 23

# 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.

# Explainable Artificial Intelligence (XAI) - project proposal

Implement algorithm for different explainable AI methods and compare the results for one of the following input data:

- tabular data,
- images,
- text.

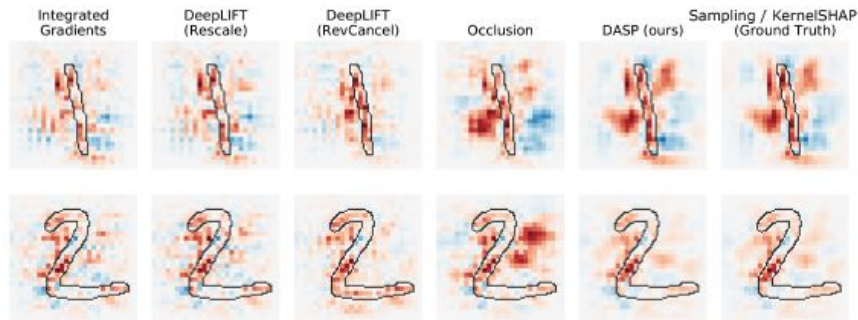


Fig. Attributions maps produced by different methods on MNIST images. Red(blue) color indicates features (pixels) that positively(negatively) impact the network output score. \*

# 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

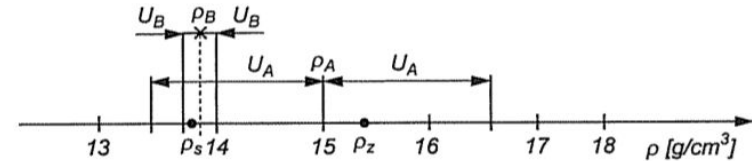
# Uncertainty quantification for machine learning models

## Example 1. 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 ( $\rho_1 = 15,5 \text{ g/cm}^3$ ) or the cheaper one ( $\rho_2 = 13,8 \text{ g/cm}^3$ ). The measurements were conducted in 2 independent laboratories.

The results are:  $\rho_A = (15 \pm 1,5) \text{ g/cm}^3$ ,  $\rho_B = (13,9 \pm 0,2) \text{ g/cm}^3$ . \*

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



## Conclusion:

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.

# Uncertainty quantification for machine learning models - project proposal

Implement algorithm for different uncertainty estimation methods and compare the results for one of the following input data:

- tabular data,
- images,
- text.

Top graph: to calculate each curve the model was re trained. Separate loss function for each regression\*.

Bottom graph: all curves were calculated simultaneously. One loss function which is the weighted average of all sub loss functions\*\*.

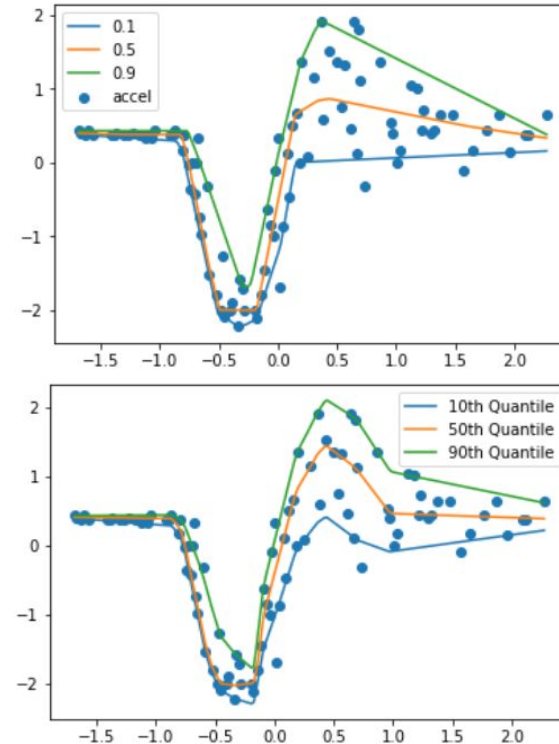


Fig. Quantile regression for Acceleration over time of crashed motor cycle.

[\*] <https://towardsdatascience.com/deep-quantile-regression-c85481548b5a>

[\*\*] <https://github.com/strongio/quantile-regression-tensorflow/blob/master/Quantile%20Loss.ipynb>

# 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?)

## Project Ideas:

- Implement a possible defense for a dual-use AI technology (see: Dodging Attack Using Carefully Crafted Natural Makeup)

**Verantwortung der Informatik – Accountability In AI**

WS 2021/22

**Introduction**

Chart **29**

[1] [https://ec.europa.eu/trade/import-and-export-rules/export-from-eu/dual-use-controls/index\\_en.htm](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  $\sim 190\,000$  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
- Possible project ideas:
  - Compare the energy consumption of different ML algorithms (NN, SVM, etc.) on the same problem
  - Implement ideas to reduce energy consumption and benchmark them

**Verantwortung der  
Informatik –  
Accountability In  
AI**  
WS 2021/22

**Introduction**  
Chart **30**

# 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

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **31**

# General Organization



hybrid (synchronous)



Tuesdays

13:30 - 15:00 Uhr



Seminar room: A-1.1

Zoom: [68196367295](https://hpi.zoom.us/j/68196367295) PW: 13371337

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

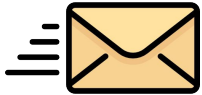
Chart **32**

# Choose Your Topic



Deadline

02.11.2021 (Tuesday next week)



send us a short mail (see about:us or website)  
your topic & 2 sentences (& optional project idea)



Seminar/Project combos (6 ECTS) will be prioritized

first come first served

teamwork is possible

well try to make it work for everyone

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **33**

# Schedule

---

The following schedule is just preliminary and will be subject to change during the semester. All updates will also be announced on the course mailing list.

02.11.2021 **Topic Deadline & Invited Talk #1 (TBD)**

09.11.2021 **Student Presentation #1**

16.11.2021 **Student Presentation #2**

23.11.2021 **Student Presentation #3**

30.11.2021 **Student Presentation #4**

07.12.2021 **Student Presentation #5**

14.12.2021 **Student Presentation #6**

21.12.2021 **Invited Talk #2 (TBD)**

28.12.2021 **X-Mas Break**

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **34**

# Schedule

---

The following schedule is just preliminary and will be subject to change during the semester. All updates will also be announced on the course mailing list.

04.01.2022 **Student Presentation #7**

11.01.2022 **Student Presentation #8**

18.01.2022 **Invited Talk #3 (TBD)**

25.01.2022 **Student Presentation #9**

01.02.2022 **Student Presentation #10**

08.02.2022 **Student Presentation #11**

15.02.2022 **Student Presentation #12**

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **35**

ECTS

3



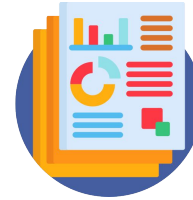
presentation

+



slides

+



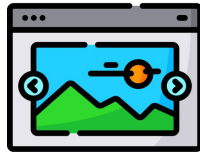
report

6



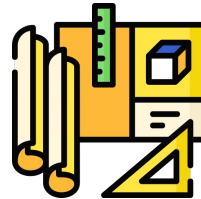
presentation

+



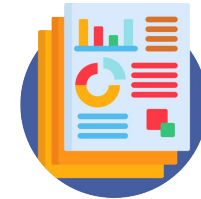
slides

+



project

+



report

**Verantwortung der  
Informatik –  
Accountability In  
AI**  
WS 2021/22

**Introduction**


Chart 36



<https://osm.hpi.de/aiai>



Operating Systems and Middleware

teaching research seminar research iot-lab theses publications people posts 

## Verantwortung der Informatik – Accountability In AI (AIAI 2021)

Prof. Dr. Andreas Polze

Kordian Gontarska, Marta Stefania Lemanczyk, Weronika Wrazen, Felix Gzelka

Tuesday, 13:30-15:00 Uhr, A-1.1

### Abstract

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.

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **37**



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

kordian.gontarska@hpi.de

marta.lemanczyk@hpi.de

felix.grzelka@hpi.de

weronika.wrazen@hpi.de

# Image References

---

- Icons made by [Freepik](#), Wichai.Wi, Smashicons, from [www.flaticon.com](http://www.flaticon.com)
- Icons made by [Nikita Golubev](#)" from [www.flaticon.com](http://www.flaticon.com)
- Icons made by [Eucalyp](#) from [www.flaticon.com](http://www.flaticon.com)
- Icons made by [GOWI](#) from [www.flaticon.com](http://www.flaticon.com)
- Icons made by [Flat Icons](#) [www.flaticon.com](http://www.flaticon.com)
- IBM course on 'Mitigating fairness'
- <https://www.businessdisabilityinternational.org/when-is-equality-not-equality/>

**Verantwortung der  
Informatik –  
Accountability In  
AI**

WS 2021/22

**Introduction**

Chart **39**