Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Internationales Zentrum für Ethik in den Wissenschaften (IZEW)

Eberhard Karls Universität Tübingen

Responsible Data Science

Session 7: 01.06.2023, 15.15 – 19.30 h
MA Seminar, SoSe 2023, Hasso-Plattner Institut

Dr. Simon David Hirsbrunner

Today

topic	time
Introduction	15h15
Guest input: EU Artificial Intelligence Act (Milan Tahraoui)	15h30
Discussion: Transparency and accountability in the context of AIA	16h00
—— Break ——	17h00
Student presentation and discussion: model cards for model reporting (Adrian Ziupka, Jeein Kim)	17h15
Shifting agency by altering access to information (transparency)	18h00
—— Break ——	18h30
Discussion: conformity assessments (CapAI)	18h45
Wrapping up	19h00
End	19h30

What do you understand by ‚transparency'?

-- Illustrating picture removed for copyright reasons ---

Source: istock/tonefotografia

Transparency

[In information studies,]
"transparency refers to the possibility of accessing information, intentions or behaviours that have been intentionally revealed through a process of disclosure."
(Turilli and Floridi 2009, 105)

Explanations in AI

"(...) 'explanation' refers to numerous ways of exchanging information about a phenomenon, in this case the functionality of a model or the rationale and criteria for a decision, to different stakeholders." (Mittelstadt et al. 2019)

Explanations in AI

"(...) 'explanation' refers to numerous ways of exchanging information about a phenomenon, in this case the functionality of a model or the rationale and criteria for a decision, to different stakeholders." (Mittelstadt et al. 2019)

However:

'explanation' can also be understood "as a social practice in which explainer and explainee co-construct understanding on the microlevel." (Rohlfing et al. 2021)

Guest input and discussion

Milan Tahraoui (HWR Berlin)
Transparency and accountability in the EU AI Act.

Student presentation

Mitchell M, Wu S, Zaldivar A, et al. (2019) Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, New York, NY, USA, 29 January 2019, pp. 220–229. FAT* '19. Association for Computing Machinery.

Introduced by Jeein Kim and Adrian Ziukpa

Presentation and discussion

Floridi L, Holweg M, Taddeo M, et al. (2022) capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act. 4064091, SSRN Scholarly Paper. Rochester, NY.

CapAI

"capAI defines a procedure to implement ethics-based auditing, offers the most effective approach to conduct a conformity assessment in line with the AIA, as it identifies and enables the correction of unethical behaviours of AI systems, and informs ethical deliberation throughout the process of designing such systems." (Floridi et al. 2022)

Cycle of CapAI

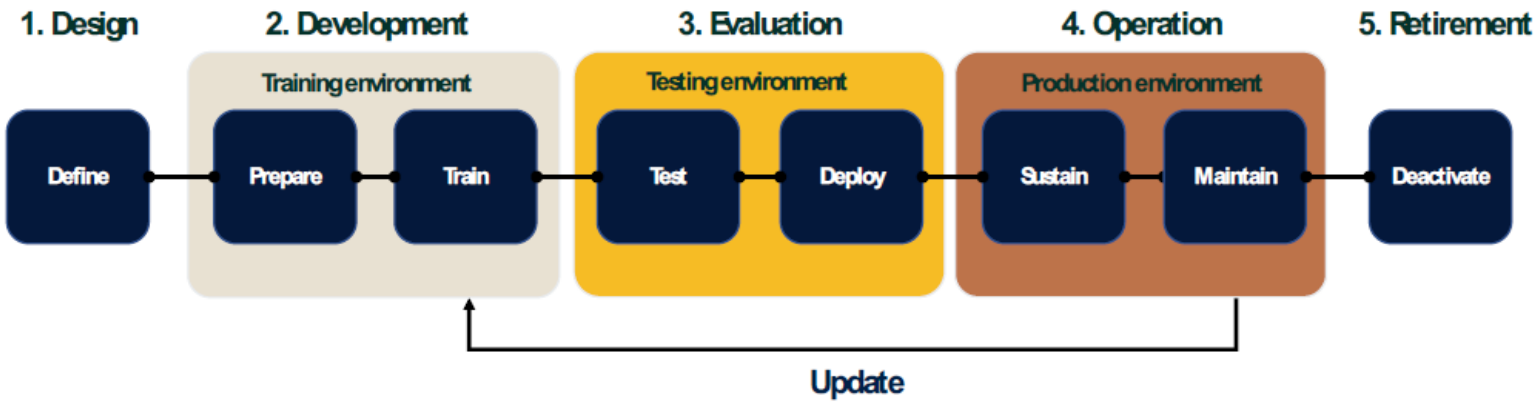
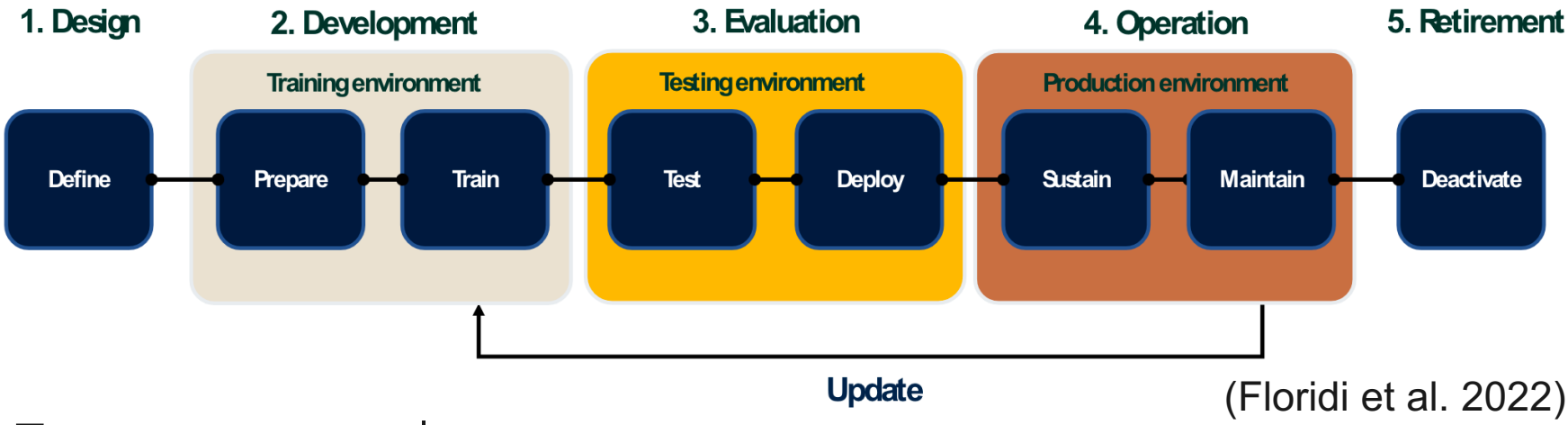


Figure 6: AI process flow with its five stages and key steps

Ethical conformity assessments for the different stages of the AI life cycle



(Floridi et al. 2022)

Transparency tools

- Data sheets
- Model cards
- Impact assessments

Outputs of CapAI

1. “an internal review protocol (IRP), which provides organisations with a management tool for quality assurance and risk management.

2. the summary datasheet (SDS) to be submitted to the EU database, and

3. an optional external scorecard (ESC), which should be made available to customers and other stakeholders of the AI system (p. 16).” It covers four key dimensions of the AI system: a) purpose and ethics norms, b) data and privacy, c) bias and explanation, and d) governance and rectification.

1 Exercice

- Take the example of the Excalibur system (LLM-based system for carrying out and managing drone attacs.
- Go back to your stakeholder mapping.
- Increase (or lower) the agency in the system of one stakeholder by altering information flows (adding or removing explanations).

Tool
Stakeholder and risk mapping

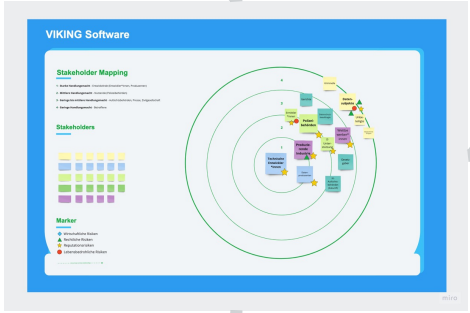
Grad der Handlungsmacht

1 - **Starke Handlungsmacht** - Entwickelnde (Entwickler*innen, Produzenten)

2 - **Mittlere Handlungsmacht** - Nutzende (Polizeibehörden)

3 - **Geringe bis mittlere Handlungsmacht** - Aufsichtsbehörden, Presse, Zivilgesellschaft

4 - **Geringe Handlungsmacht** - Betroffene



Risiken

- Wirtschaftliche Risiken
- Juristische Risiken
- Reputationsrisiken
- Lebensbedrohliche Risiken
- Grundrechtliche/ethische Risiken

Student papers

Format

- Students may submit individual papers or papers in groups of two persons
- Paper length: 6 pages per person (2300 words, with an accepted deviation of +/- 10 %) / 12 pages for groups of two persons
- The paper may include one visualization (stakeholder mapping or similar) or table.
- Orientate yourself towards common structures of research papers
- (abstract / introduction and research question / related work / theory and methods / analysis / further discussion / conclusion and outlook / bibliography).
- For layouting, use LaTeX or overleaf with one the following templates:
  https://www.overleaf.com/latex/templates/acm-conference-proceedings-primary-article-template/wbvnghjbzwpc
  https://www.acm.org/publications/proceedings-template

Student papers

Content

- The study and paper should generally embrace the approach of value-sensitive design (VSD) conceptualized in Friedman 1996; Friedman et al. 2019.
- Discuss some of the literature and topics introduced in the course.
- The study for the paper should
  - Include the analysis of an existing AI application field or concrete AI system by means of VSD-compatible methods (e.g. stakeholder and risk mapping);
  - focus on one ethical value or value tension in the field or system;
  - develop a scenario for a value-sensitive design of a future AI system by building on methods introduced in the course (scenario-based design, value scenario, risk mitigation).
- Variations from this form are possible but must be discussed with the lecturer.

You have reached the end of the seminar…..

-- Illustrating picture removed for copyright reasons ---

Source: rights by Matt Dixon

Sources

https://github.com/simonsimson/responsible-data-science/tree/main/slides