ETH zürich

D-MTEC WOP

ExplainAI: Designing explainable ML-based systems for collaborative work in the railways

Lena Schneider[1], Gudela Grote[1], Daniel Boos[2]
[1] Chair of Work and Organisational Psychology, ETH Zurich; [2] SBB

Research Questions

1 How should we design the **distribution** of **control** and **accountability** in ML-based systems for collaborative use in the railways?

2 How can we make such systems **explainable** for the involved human **actors with different backgrounds** and professions?

3 How can we **support product development** in addressing potential issues with with explainability, control & accountability during system development and use?

Project Outline

Completed/ongoing

Phase 1
Taxonomy Development
2023

- semi-structured interviews with different stakeholders
- document analysis
- literature review

Phase 2
Socio-technical Design Framework
2023/2024

- Computer-based experiment
- Observation of collaboration and coordination between stakeholders in use cases
- Review of existing tools

upcoming

Phase 3
Framework Validation
2024

Phase 4
Integration
2024

Phase 5
Dissemination of Results
2024/2025

Background

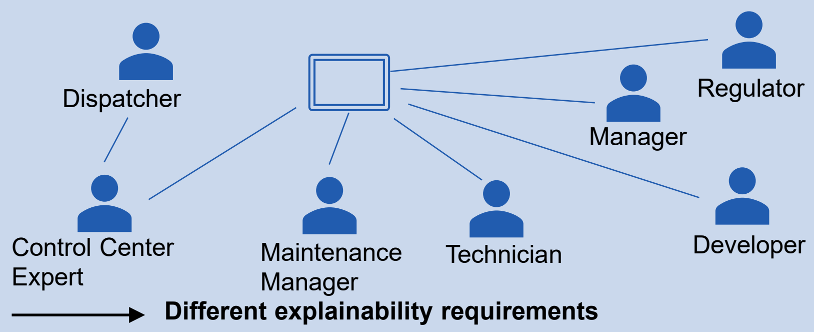
- **Opaqueness** of ML-based systems is a key barrier to overcome (Castelvecchi, 2016)
- The **accountability-control gap** is a phenomenon already known from traditional automation, but is even wider for AI (Grote et al.,2014; Grote et al., 2022)
- Legally, accountability always stays with the human actors, but control increasingly lies within the system (Taddeo & Floridi, 2018),
- All stakeholders involved in development and use of ML-based systems have to continuously negotiate the **distribution of control and accountability** amongst them (Berente et al., 2021; Grote et al., 2022; Slota et al., 2021)
- For targeted explanations, deep understanding of stakeholders and their tasks is needed (Hafermalz & Huysman, 2021)

Preliminary Results

Identified **Use Cases** include

- Visual inspection & (predictive) maintenance
- Traffic Management
- Automated Train Operation
- Surveillance and detection of switch malfunctions

Example **Stakeholder Network for Visual Inspection**



Upcoming Experiment

- contrasting **different explanations** (varying in content and design) **from multiple stakeholder perspectives**
- Computer-based experiment with mock system for damage detection
- Sample: approx.20 domain experts (i.e.,end users, developers, regulators)
- participants are confronted with different explanations and asked to share their perceptions and preferences (Thinking Aloud Method)

Expected Impact

- Capture processes involved in collaboration among heterogenous teams and (multiple) AI systems and translation into design requirements for XAI
- More effective use of techniques to build in explanations in ML-based systems
- Facilitated decision-making during systems design to create more reliable and safe systems

References

Berente, N., Gu, B., Recker, J. & Santhanam, R.(2021). Managing Artificial Intelligence. MIS Quarterly. 45. 1433-1450. 10.25300/MISQ/2021/16274.
Castelvecchi, D. (2016). The blackbox of AI. Nature, 538(7623), 21-23.
Grote, G., Parker, S. K., & Crowston, K. (2022). Organizing AI: A design theory for organizational decision-making on and with learning algorithms in networks of accountability. Paper presented at the WAIM conference, June 6-7, Washington D.C.
Grote, G., Weyer, J., & Stanton, N. (2014). Beyond human-centred automation - concepts for human machine interaction in multilayered networks. Ergonomics, 57, 289-294.
Hafermalz, E. & Huysman, M. (2021). Please explain: Key questions for explainable AI research from an organizational perspective. Morals & Machines, 2, 10-22.
Slota, S. C., Fleischmann, K . R. , Greenberg, S., Verma, N., Cummings, B. , Li, L., & Shenefiel, C. (2021). Many hands make many fingers to point: Challenges in creating accountable AI. AI & Society.
Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751-752.

Partner:

SBB CFF FFS

SIEMENS

CSFM