Winter Term 21/22

Artificial Intelligence, Ethics and Engineering

Org & Introduction

Prof. Dr. Holger Giese (holger.giese@hpi.uni-potsdam.de)

Christian Medeiros Adriano (christian.adriano@hpi.de) - **"Chris"**

Christian Zöllner (Christian.zoellner@hpi.de)

Key Facts

- Weekly Hours: **4**
- Credit Points: **6**
- Teaching Form: **Project Seminar**
- Enrolment Type: **Compulsory Elective Module** (“Wahlpflichtmodul”)
- Course Language: **English**
- Study Programs and Modules:
  - **IT-Systems Engineering MA**
    - Specialization module(s): *„Software Architecture & Modeling Technology“ (SAMT)*
    - Specialization module(s): *„Operating Systems & Information Systems” (OSIS)*
  - **Data Engineering MA**
  - **Digital Health MA**
    - Specialization module(s): *„Acquisition, Processing and Analysis of Health Data“ (APAD)*

Dates

- Enrollment deadline: **22.10.2021**
  - Cancellation deadline for enrollment: **30.01.2022**

- Introductory meeting: **02.11.2021** **[NOW]**

- Meetings:
  - *Lectures - scheduled*
  - *Update meetings – on demand, usually weekly*

- Final Presentations at end of the semester: **To be decided**
  - *Presentations will be at the lecture room and the participants will be able to join via Zoom.*

Communicantion Plan

Motive	Content	Medium
Artifacts	Source code, Data Documentation, Wiki	Github - https://github.com/orgs/hpi-sam/
Papers	Copyrighted material	Zotero
Messaging ad hoc	Questions, Suggestions, Sharing	Our Slack group: **https://aiethicsengineering.slack.com**
Official communications	Schedule, Orientations, Administrative issues	Email christian.adriano@hpi.de
Meetings	Lectures, Status, Work meetings	Zoom, Skype
Emergency	Call, SMS, messaging	Chris mobile number (check Chris' Slack profile)

Project Proposal

Team size: up to four (preferred)

Project proposal in two stages:

1- State-of-art (1 page, double column) – ~ in approx. 6 weeks

- covering at least 5 well-selected papers per person

2- Proposal - first draft ~ in approx. 8 weeks (ideally before New Years break)

- Detail the problem (what is it? why should I care?, why is it challenging?)
- Describe the scenario (source, size, main features, cite any papers that used it)
- Determine the methods that you plan to use (preliminary insights, it might change)
- Discuss how you will evaluate your results (benchmarks, baselines, null-models)

Roadmap (1/2)

- **Project Phase 1: Learn fundamentals - Lectures**
  - Goal: learn fundamentals
  - Deadline: Mid-End of December
  - Two lectures per week (Tuesday 9:15 and Wednesday 17:00)
- **Project Phase 2: Present Proposal - Reading and Writing**
  - Goal: learn about the state of art of one application area
- **Project Phase 3: Apply a method - Coding and Evaluation**
  - Goal: learn to apply and evaluate a method
  - Present update in weekly meetings (either Tuesday 9:15 or Wednesday 17:00)

- **Final Presentations** in one session in late **January or February 2022**
- **Submission of final report** one week after the presentation

Road Map (2/2) Topics of Lectures

1. Intro and Course Organization
2. Responsible AI (Fairness, Explainability, Trustworthiness)

Overview and Team Building

3. Ethics Foundations (Normative Ethics and Normative Argumentation, Ethical Questions in the Design of Technology and their Risks)
4. Regulations and Governance (IEEE Ethically Aligned Design, EU regulation, Governance)

Ethics theory, laws, and best practices

5. Experimental Methods for Ethical AI (Simulators, Experimental philosophy)
6. Requirements Engineering for Ethical AI (feature engineering, specification for fairness, human-in-the-loop)

Methods for the elicitation of Ethical AI requirements

7. Validation & Verification Ethical AI (testing, algorithm recourse, model comparison, risks, AI Alignment)
8. Implementing Ethical AI (logic programming, argumentative methods, safety methods)

Methods and Models to build Ethical AI

Seminar Work, Deliverables and Grading

- Seminar work **alone or in groups** on **one selected topic/project**.
- Each team is supervised individually by a teaching assistant.

Project Execution: [60% of final grade]

- Weekly update meeting
- Intermediary Presentations

Written deliverables: [30% of final grade]

- Final report on findings
  - □ Length: 6 to 10 pages ACM Format per team
  - □ Some parts must be attributable to each individual author

Final Presentations: [10% of final grade]

- Presentation on findings
- Questions and feedback for other students' presentation

Suggested Domains

- Autonomous Lethal Weapons
- Autonomous Driving
- Autonomous Operation (assembly lines, mining, farming)
- Autonomous Recommender Systems (shopping, dating)
- Autonomous Identification Systems (security, vigilance, medical diagnostics)
- Autonomous Support Administrative Decisions (justice, granting parole)

Ethical Judgements and Dilemmas (examples)

- How autonomous should a given AI-based system be?
  - Which tasks should be restricted to humans?
  - When not, under which conditions?

- Which biases and mistakes made by autonomous agents are morally less acceptable?
  - For these mistakes, what are the possible mitigation actions?

- How is blame and credit attributed?
  - when complex engineered system rely on multiple AI components
  - when humans and autonomous systems collaborate to achieve a common goal

- Are current judgements about autonomous agents atemporal or they might change?
  - If they change, who should decide about their acceptability?

Techniques

Software engineering techniques

- Requirements Engineering
- Analysis & Modeling
- Coding
- Verification & Validation (testing)

Surveys

- Focus groups
- Crowdsourcing

Tools

- Prototyping
- Goal-based argumentative techniques
- Ethics Simulators

Some tools

Simulators

- MIT Troley case https://www.moralmachine.net/
- Board game https://www.simulationtrainingsystems.com/corporate/products/where-do-you-draw-the-line/
- Allen Delphy - https://delphi.allenai.org/
- DeepBlue - https://businessethicssimulation.com/

Examples of Projects

- Prototype a simulator for ethical dilemmas
- Analyze various aspects of an ethical dilemma using argumentative logic
- Prototype a tool to validate ethical decisions using logic programming
- Specify and design an architecture to act as an ethics supervisor
- ?

END