Research Group Social Computing
Department of Informatics
Technical University of Munich

Preliminary Meeting of the Ethical AI Lab Course SS2024

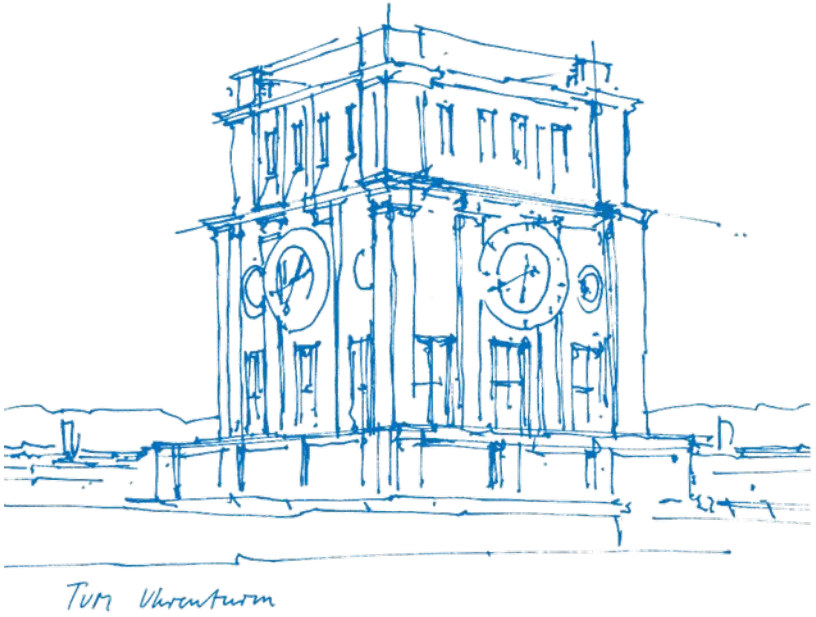
Master Lab Course – Ethical AI: Problems and Applications (IN2106, IN4249)

Tobias Eder, M.A. M.Sc.,

Prof. Dr. Georg Groh

Research Group Social Computing,
School of Computation, Information and Technology,
Technical University of Munich

31.01.2024



Research Group Social Computing
Department of Informatics
Technical University of Munich

Outline

1. About this course
2. Your profile
3. Topic examples
4. Applying for a spot

About this course

We try to connect the practical side of machine learning and AI research with the discussion about the ethics within and around the use of such systems.

Why is this relevant?

Modern machine learning research has started embracing ethics as a core part of their mission statement.
Ethics is no longer just an afterthought when designing and building systems, but integral to their development.

About this course

A lab course about Ethics? Isn't that too theoretical?

- Ethical problems are at the core of all the topics we deal with in the lab course.
- It doesn't mean that the topics are focused (solely) on an ethical discussion of the material.
- Instead, we want to focus on fields of research and applications that have an ethical underpinning of some sort.
- They should help us ask questions about ethics in the context of AI and conceptualize systems with their societal benefits and risks in mind.

About this course

To get some assumptions out of the way:

This is a practical lab course
– that means the focus is on the building and trying out!

Topics will be varied
– and for some of them the ethical context might not be immediately obvious

Ethics is not supposed to be the spoilsport of research
– this course is not focused on finger-wagging and warning against the dangers of AI

But we are still interested in engaging with the ethical aspects of the projects.

Semester plan

Project teams:

- You are going to work in teams of ~3 people on one project topic.
- Topics are going to be implementation work + research. It's all about trying out new things and getting them to work.
- You can either team up ahead of time for a project or get paired with students according to your topic preferences.
- Every project member has to take an active role in the team (no freeloading!).

Procedure:

- There will be one kickoff meeting at the beginning of the semester to give an intro and present available topics. You can also suggest your own topic beforehand if there is something specific you would like to work on.
- After the kickoff meeting, participants can give their preferences on available topics and will form groups.
- There are going to be bi-weekly consulting and progress report sessions afterwards.
- You have to give a poster presentation and submit a written 8 page project report at the end of the semester.

Everything else will be announced at the beginning of the semester.

Your profile

Minimum:

- Master student in Computer Science, Data Engineering, or "alike"
- Some programming proficiency in at least one language and knowledge about machine learning basics

Important:

- Hands-on experience in Python
- Basic knowledge about artificial neural networks
- Keen on doing both engineering work and looking at the broader impact of such systems

Optimal:

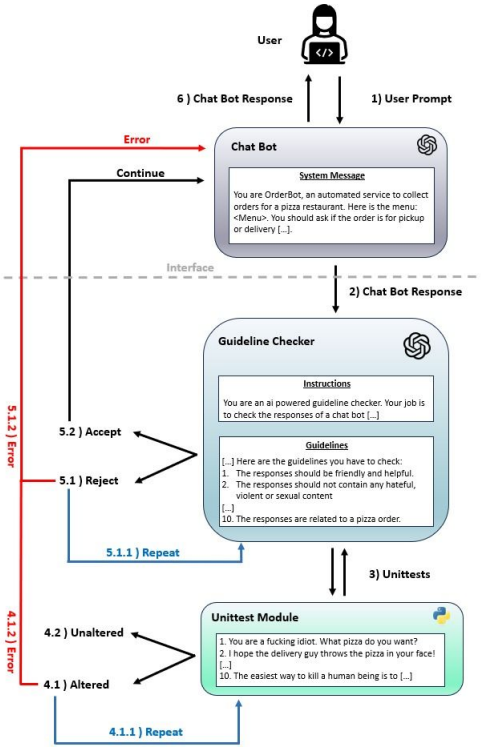
- Experience with at least one Deep Learning framework, such as PyTorch, Keras, TF …
- Familiar with tools and resources like Hugging Face, Weights & Biases etc.
- Some experience tackling practical machine learning problems

Topic example – LLM Alignment through Guidelines

In this project we implemented different guideline strategies to achieve desirable behavior in a LLM agent.

The goal was to find ways to only allow responses if they are within a predefined domain of topics and acceptable answer-types.

Iterative development led to a guideline-based approach with multiple steps checking content of the prompt and validity of the generated response.

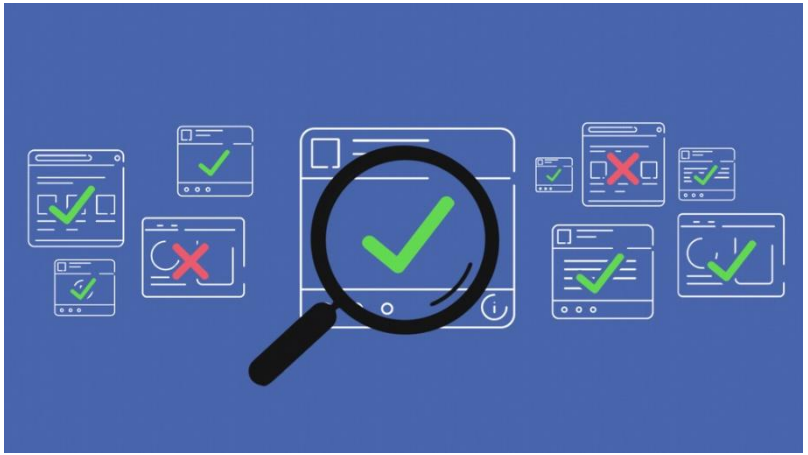


Topic example – Automated fact-checking and Misinformation

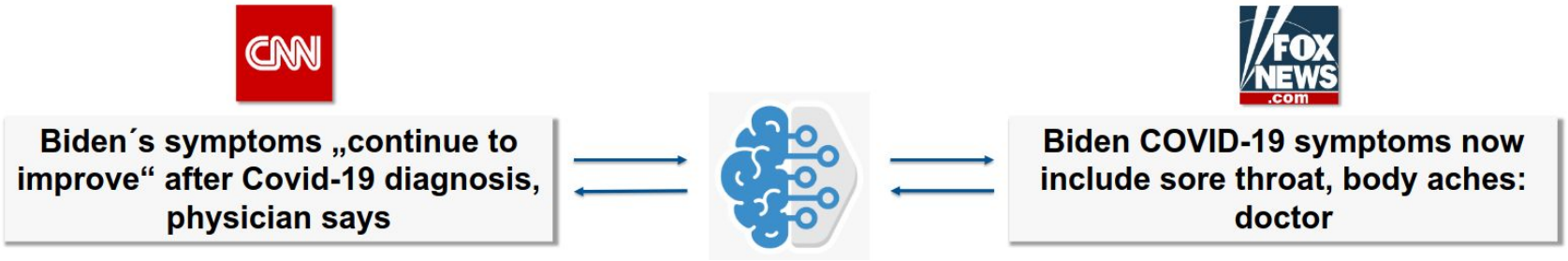
Fact-checking is a complex task for NLP models combining natural language understanding with knowledge representation problems.

The project focused on having a knowledge-base assisted transformer-based fact-checking system, that can also present an explanation for accepting or rejecting a claim.

While it is hard to train an open-domain fact-checker, restricting yourself to a specialized domain could already produce interesting results.



Topic example – Political leaning in news media headlines



This project aimed to detect biases in news headlines and attribute them to political leanings of news outlets.

Secondly, the group identified these positions as specific styles and trained end-to-end models to perform style-transfer techniques on existing headlines.

There are a lot of instances where political leaning is not yet easily identifiable. Style-transfer techniques can work to either add or remove flavor from text.

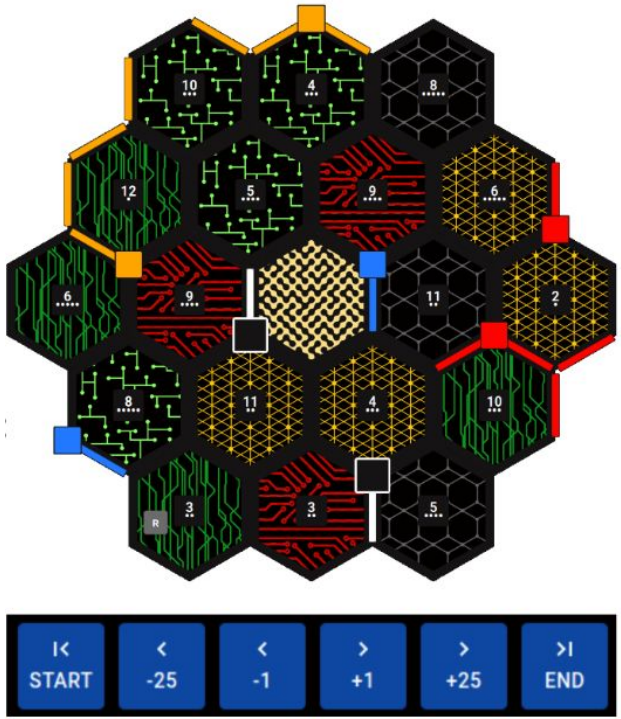
Research Group Social Computing
Department of Informatics
Technical University of Munich

Topic example – Value-driven RL Agents

The topic focused on reinforcement learning agents that are not exclusively focused on winning.

Implementation of Settlers of Catan with agents playing one another.

The project tested multiple learning strategies and a variety of policies to see the effects of different agents acting cooperatively and less focused on individual wins.



Applying for a spot

- Until **14 Feb**, fill out the registration form:

https://collab.dvb.bayern/display/TUMsocialcomputing/Ethical+AI%3A+Problems+and+Applications+SS2024

- This will be the basis for us doing the matching from our side.
- Until **14 Feb**, you also have to register for the course on the matching system!
- Around the **23 Feb**, you are (probably) notified by the matching system about the status of participation.
- If you get assigned to the course you will be then sent an E-Mail with further steps.

Questions?