17-416/17-716: AI Governance: Identifying and Mitigating Risks in the Design and Development of AI Solutions

Spring 2024 - 6 units

Instructor: Norman Sadeh (https://normsadeh.org)

Overall Description

With AI and ML finding their way into an increasingly broad range of products and services, it is important to identify and mitigate the risks associated with the adoption of these technologies. This course reviews the different types of risks associated with AI and discusses methodologies and techniques available to identify and mitigate these risks. The course introduces students to ethical frameworks available to identify and analyze risks. It also examines best practices emerging from both government and industry efforts in this area. This includes looking at new regulations such as the EU AI Act as well as emerging frameworks such as the one developed by NIST. The course also examines frameworks developed by leading companies and how these frameworks combine both technical and non-technical approaches. It further discusses changes that need to be enacted by organizations to adopt more systematic approaches to AI governance.

This course combines a mix of technical, policy, and management discussions.

Objective

This course is intended for a broad cross-section of students, both advanced undergrads and graduate students, planning to work on the design, development and deployment of AI-based solutions. The course is designed to introduce students to key concepts, challenges, principles, methodologies, techniques, best practices, legal requirements and trends associated with the responsible design, development and deployment of AI technologies.

Prerequisites

The course does not assume a deep technical understanding of AI/ML techniques. Instead, gentle introductions to relevant techniques and concepts will be provided over the course of the semester, as required to follow discussions of different topics. Material and discussions are designed to enable people with diverse technical backgrounds to benefit from topics discussed in the lectures. Students will however be expected to have a basic understanding of probability and statistics. Because AI governance is emerging as an activity that has to involve a broad set of roles within the enterprise (e.g., product managers, AI/ML engineers, legal & compliance, UX/UI designers, security engineers, privacy engineers, safety engineers, software architects, software engineers), the

course is designed to take a broad, multi-faceted view of relevant topics and aims to appeal to a broad cross-section of students.

Format:

The class will meet once a week. It will combine lectures, class discussions, and work on group projects. Project teams will present their work at a poster fair at the end of the semester. Grading will be based on a midterm, final and a team project.

Overall Outline of Topics Covered:

Week 1	Introduction - Overall Context
Week 2	What is AI & What is AI Governance
Week 3	AI: An Ethical Perspective
Week 4	Regulating AI: Overall Landscape, Challenges and Trends
Week 5	Self-Regulation & Best Practices: A Historical Perspective
Week 6	AI Threat Modeling Framework (e.g. NIST and beyond)
Week 7	AI Transparency, incl. Explainability & Interpretability
Week 8	AI and Agency
Week 9	AI Governance today (e.g., Microsoft, META, Amazon, Open AI)
Week 10	ML Governance: A Deeper Dive (e.g. model drift, bias, etc.)
Week 11	Human AI Interactions & Trustworthy AI Principles
Week 12	Gen AI: New Threats & New Governance Challenges - I
Week 13	Gen AI: New Threats & New Governance Challenges - II
Week 14	AI Governance and the Military
Week 14	Project Fair

Grading (tentative):

Class participation: 10%
Midterm & Final: 40%
Project Fair Presentation & Final Project Report: 50%