17-416/17-716/19-416/19-716: AI Governance: Identifying and Mitigating Risks in the Design, Development and Operation of AI Solutions
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 the methodologies and techniques available to identify and mitigate these risks. The course introduces students to the 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 the 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.

6-Unit vs 9-Unit Sections & Grading:

Students enrolled in the 9-unit section are expected to work on a team project

- **Grading 6-Unit Section:**
  - Midterm: 25%
  - Final: 25%
  - HW Assignments: 2 x 25% each
- **Grading 9-Unit Section:**
  - Midterm: 20%
  - Final: 20%
  - HW Assignments: 2 x 15% each
  - Team Project: 30%

Lecture Topics

Week 1: Welcome & Overall Context

Week 2: Trustworthy AI: Ethical Principles

Week 3: AI & Data Governance/Privacy: A First Example

Week 4: Data Governance, Privacy Threat Identification & Mitigation

Week 5: NIST AI Threat Modeling Framework: A Deep Dive

Week 6: AI Legal and Regulatory Landscape: EU AI Act

Week 7: MIDTERM

SPRING BREAK

Week 8: Transparency, Explainability, Interpretability, and Agency

Week 9: Model Alignment, including RLHF and red teaming

Week 10: Fairness and Bias

Week 11: AI/ML Security

Week 12: AI Safety - Autonomous Driving

Week 13: Copyright Issues; Military Uses of AI

Week 14: PROJECT FAIR

Project Fair (each team presents their poster)