DSC 180A: Responsible AI

Q1, Fall 2022 | Classes held on Zoom (https://ucsd.zoom.us/j/96921239764) on Fridays at 10AM PT.

Please note this Syllabus may be corrected, adjusted or updated throughout the course. Students will be updated if such changes occur via email, or other course communication listed.

Overview

The explosive proliferation of research around ethical and trustworthy AI during the last decade reflects a rising societal awareness and desire to address potential and actual harmful impacts and consequences of AI. We will explore the socio-technical risks of AI by examining several high-profile algorithmic discrimination debates that have sparked seminal research in this field. We will discuss common definitions and metrics, the perspectives they represent, and the interesting yet mathematically irreconcilable relationships that interweave them.

Concepts around AI-fairness will be translated into real-world applications through classroom debates from the perspectives of different disciplines and stakeholders, and by replicating the analyses from a real-world example. This replication project will inform an understanding of how issues that originate in code and math eventually affect real human beings and society as a whole.

Course Learning Outcomes

- What does AI trustworthiness and ethics mean for different stakeholders?
- What are the ways to think about whether or not a model is fair?
- What are the risks of 'black box' algorithms, and how do we mitigate them? How is AI explainability related to fairness?
- How do inherent fairness problems in AI models affect human beings?

Instructor: David Danks

David Danks is a Professor of Data Science & Philosophy at University of California, San Diego as well as affiliate faculty in UCSD's Department of Computer Science & Engineering. He serves on advisory boards for various organizations including: National AI Advisory Committee (NAIAC), Special Competitive Studies Project (SCSP), Partnership to Advance Responsible Technology (PART), Center for Advancing Safety of Machine Intelligence (CASMI), Topos Institute, AI Community of Practice (US Government), and Grefenstette Center for Ethics in Science, Tech., & Law. He has received a James S. McDonnell Foundation Scholar Award (2008) and an Andrew Carnegie Fellowship (2017). Previously, he was the L.L. Thurstone Professor of Philosophy & Psychology at Carnegie Mellon University. While at CMU, he served as the Chief Ethicist of CMU's Block Center for Technology & Society and co-director of CMU's Center for Informed Democracy and Social Cybersecurity (IDeaS).

Industry Partner: Deloitte

As one of the largest professional services organizations in the United States, Deloitte provides a vast array of information security services across 2,800 engagements in major commercial industries and 15 cabinet-level federal agencies. Our Trustworthy AI team has helped many of our clients work through the burgeoning regulatory landscape and growing awareness around ethical and trustworthy AI. For this course, the Deloitte team will consist of Rasmus Nielsen, Aritra Nath, Emma Harvey, Nandita Rahman, Meira Gilbert, and Jeffry Liu. We're excited to work with UCSD in developing this course, and we look forward to discussing these exciting topics with you.

Course Resources

Course Website: https://nanrahman.github.io/capstone-responsible-ai/

Office Hours

- Professor Danks will hold office hours Wednesdays 3-4PM in UC 302, Room 101
- Deloitte: Will hold ad-hoc office hours (TBD) for assistance with replication project materials

Communications

Please send any questions to the Responsible AI Microsoft Teams Channel (Sign up here):

- Channel link: [Course Chat, Q1] DSC-A05 Responsible AI;
- for any issues email Nandita Rahman (nanrahman@deloitte.com)

For private or personal questions, you can reach out to the course instructors privately via email (ddanks@ucsd.edu).

Assignment Submission

Participation questions should be submitted to Gradescope and will be due 24 hours prior to the start of the class for which they are assigned. Other assignments, including both one-off tasks and assignments related to the replication project, should be submitted to Gradescope will be due at the start of the class for which they are assigned.

Course Responsibilities and Assignments

Please see the Capstone Program Syllabus for a detailed description of the assignment weights and rubric.

Section Participation: Participation Questions and Reading Presentations

Each week, students will be assigned a set of readings. There is no textbook for this course; all readings are freely available, and links will be provided. Each student is responsible for reading all content in full prior to the start of each week's course, as the readings will guide the week's discussion. In addition, students will be responsible for engaging with the readings in the following ways:

- Posting responses to **participation questions** to Gradescope. These responses will be due 24 hours prior to the start of each session.
- Preparing **reading presentations** for selected meetings. Following the first session, students will have the opportunity to sign up to present on one of the course readings. Students are responsible for creating a PowerPoint presentation summarizing the reading, including its background, methodology, argument/key contributions, and their thoughts on the implications/impact of the article. Reading presentations are due via Gradescope prior to the start of the session for which the reading is assigned; students are also responsible for presenting to the rest of the class during the session. Presentations should take five minutes.

Quarter One Project

Students will complete coding tasks related to the replication project and are also responsible for creating a final writeup.

- Full details of the requirements for the Q1 project can be found in the Capstone Program Syllabus.

Quarter Two Project Proposal

Finally, students will develop a project proposal for Q2 based on their learnings and interests from the course readings and the replication project.

- Full details of the requirements for the project proposal can be found in the Capstone Program Syllabus.

Course Outline

Week 1 (9/30): Introduction to Responsible AI

Class Coverage	Pre-Class Readings	Assigned	Due
Course overview and expectations, classroom introductions. Brief introduction to responsible AI – what is it? What ethical considerations are present in various everyday AI use cases? *Participation Questions:* Introduce yourself. Who are you and why did you select this domain within the capstone? • Talk about a time when AI/ML had a real-world impact on the life of you or someone you know. • What characteristics, in your opinion, does an AI system need to have in order to be “ethical”?	Syllabus Deloitte’s Trustworthy AI Introduction *Mitigating Bias in Machine Learning for Medicine* (Vokinger et al.)	**Writeup #1:** Identify places where ethical AI issues or harms could be present. Outline what the consequences of those harms could be and propose ways to identify and mitigate those harms.	-

Week 2 (10/7): A Multi-Stakeholder Perspective on Ethical AI

Class Coverage	**Pre-Class Readings**	**Assigned**	**Due**
What are the main doctrines and frameworks of AI fairness? How does responsible AI affect different fields? How do ethical considerations affect different stakeholders? What are some limitations to ethical AI? *Participation Questions:* • Provide an example (other than COMPAS) of a classification use case where a false negative and a false positive have different impacts on different groups. Describe the impacts. • Select one of the use cases discussed in AI Now's 2019 report. Describe as many stakeholders as you can think of (they may not be mentioned in the report) and what impact the AI use case may have on them.	*Inherent Trade-Offs in the Fair Determination of Risk Scores* (Kleinberg et al.) *Machine Bias* (ProPublica) Response to ProPublica (Northpointe) *Who Audits the Auditors?* (Costanza-Chock et al.) *2019 Report* (AI Now)* ***Select one section** (you don't have to read the whole thing)	**Writeup #2**: Analyze the potential impacts (benefits, risks, limitations, etc.) from the perspective of as many different stakeholders as you can think of. Point out where the interests of different stakeholders may be aligned and where they may be at odds.	Week 2 Reading Presentations Writeup #1

Week 3 (10/14): Replication Project 0: Introduction

Class Coverage	Pre-Class Readings	Assigned	Due
An overview of the Q1 Project objectives, AIF 360, and the Medical Expenditure Tutorial. A brief overview of EDA. *Workshop:* Walk through project logistics, including datasets, code repo, and setup. *Participation Questions:* • What are some ways that data can introduce bias into an AI/ML model? List as many as you can think of. • Get familiar with the AIF360 API and functionality. What do you think of it? What is good about it? Is it missing any features that you think are important?	*AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias* (Bellamy et al.) *Datasheets for Datasets* (Gebru et al.)	**Replication Part #0**: Set up notebooks, fill out team persona survey	Week 3 Reading Presentations Writeup #2

Week 4 (10/21): Replication Project 1: Running Data Science Teams and Projects

Class Coverage	Pre-Class Readings	Assigned	Due
How do we run a data science project? What skills are necessary on AI teams? How can/do data scientists work effectively with other roles? *Workshop:* Initiate Q1 Project team groupings and discuss team roles. *Participation Questions:* • What role do you see yourself taking on in a data science team, and why? • Describe a time when you struggled to complete an assignment (career or academic). What was missing (resources, support, etc.)? How did you resolve it?	*Taxonomy of AI Risk* (NIST) *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?* (Holstein et al.)	**Replication Part #1:** Notebook and writeup on data cleaning & EDA	Week 4 Reading Presentations Replication Part #0

Week 5 (10/28): AI Regulations

Class Coverage	Pre-Class Readings	Assigned	Due
What does the impossibility theorem mean for organizations employing AI? Should we just avoid using AI completely? What regulatory efforts help ascertain the definitions of fairness applicable to the use case? *Participation Questions:* • How effective do you think NYC's hiring law would be? Should future regulation use this as a template? Is it over-broad? Under-broad? Just right? What is wrong with how the law is written? What do you like about the law?	*Big Data's Disparate Impact (Barocas and Selbst)* *Why New York City is cracking down on AI in hiring (Brookings)*	**Writeup #3:** Research regulatory considerations that apply to this use case. How do they map (or not) to assessments you would conduct to identify/mitigate the harms that you identified in Writeup #1?	Week 5 Reading Presentations Replication #1

Week 6 (11/4): Replication Project 2: Data

Class Coverage	Pre-Class Readings	Assigned	Due
How can we develop responsible models? What data-specific ethical AI issues are there to consider? What metrics are commonly used to measure fairness? What can't be captured in data? *Workshop:* Q1 Project office hours *Participation Questions:* • "Individual fairness" refers to the practice of treating similar individuals similarly. What fairness metrics can you use to measure individual fairness? • "Group fairness" refers to the process of equalizing some output metric(s) (positive predictions, error rates, etc.) across demographic subgroups. What fairness metrics can you use to measure group fairness? • Can you think of any types of fairness other than individual and group? How would you measure them?	*Ethical Machine Learning in Health Care* *(Chen et al. 2021)* *Millions of black people affected by racial bias in health-care algorithms* *(Nature)*	**Replication Part #2:** Notebook and writeup on model development and fairness metric assessments	Week 6 Reading Presentations Writeup #3

Week 7 (11/11): Replication Project 3: Fairness Assessments and Improvements*

Class Coverage	Pre-Class Readings	Assigned	Due
Mitigating fairness issues in AI through pre-, in-, and post-processing. Re-learning and re-deploying models. *Workshop:* Q1 Project office hours *Participation Questions:* • In what situations do you think "fairness through awareness" are most appropriate? Are there any situations where "fairness through unawareness" is necessary instead? • Describe a situation in which a pre-, in-, or post-processing technique might be appropriate for mitigating model bias and explain what technique you would use.	*Model Cards for Model Reporting (Mitchell et al.)* *Fairness Through Awareness (Dwork et al.)* *Demo: Explaining model behavior using LIME*	**Replication Part #3:** Notebook and writeup on post-processing and bias mitigation.	Week 7 Reading Presentations Replication #2

Week 8 (11/18): Capstone Planning: Avoiding Techno-Solutionism

Class Coverage	Pre-Class Readings	Assigned	Due
How do implementation and other non-technical considerations affect the ethical impacts of AI? How do we design proposals and analysis plans for data science projects? *Workshop:* Planning the Q2 Project proposal (30 minutes) *Participation Questions:* • Describe a situation (not unemployment fraud detection) in which the same algorithm could have different impacts based on how it is deployed. Describe the criteria that you think would lead to best- and worst-case outcomes.	*Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models (Black et al.)* *Government's Use of Algorithm Serves Up False Fraud Charges (Undark)*	Prepare final Q1 project presentations Develop a proposal and plan for Q2 project	Week 8 Reading Presentations Replication #3

Week 9 (12/2): Replication Project 4: Presentations

Class Coverage	Pre-Class Readings	Assigned	Due
Students present their Q1 project reports and their Q2 project proposals for discussion *Participation Questions:* • Revisit your answer to this Participation Question from Week 1: "What characteristics, in your opinion, does an AI system need to have in order to be 'ethical'?" Has your answer changed? What characteristics do you think are necessary now?	-	-	Q1 Project Presentations Q2 Project Proposals