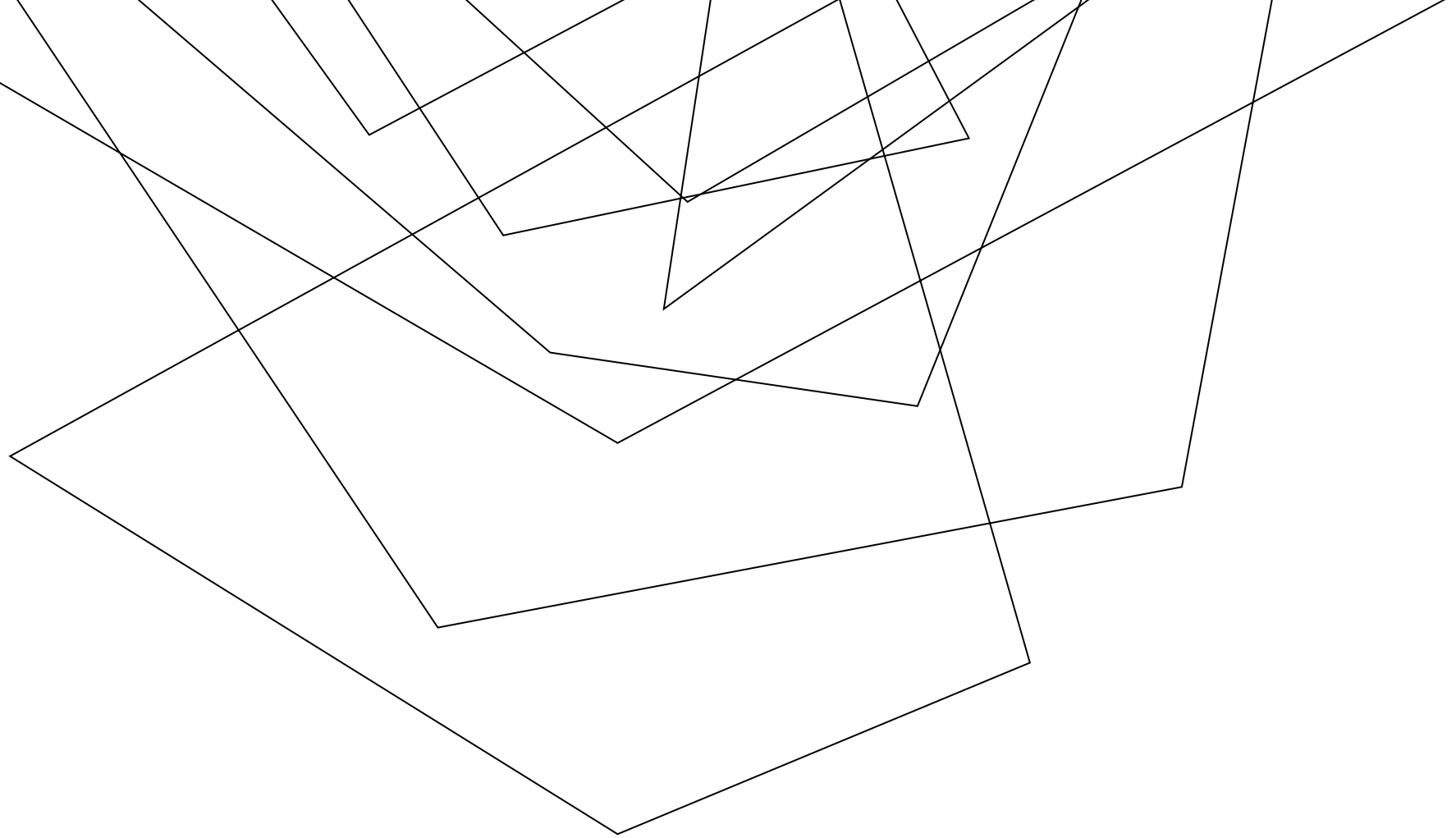


RESPONSIBLE AI

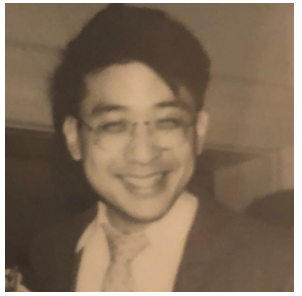
Week 1: Introduction

COURSE INSTRUCTORS

David Danks
UCSD

Rasmus Nielsen
Deloitte

Emma Harvey
Deloitte



Jeffry Liu
Deloitte

Nandita Rahman
Deloitte

Aritra Nath
Deloitte

Meira Gilbert
Deloitte

COURSE OBJECTIVES

The goal of this course is to help students understand and develop points of view on the answers to questions like:

- What are the ways to think about whether a model is fair?
- What does responsible AI mean for different stakeholders?
- How do ethical problems in AI models affect human beings?
- What are the risks of ‘black box’ algorithms, and how do we mitigate them? How is AI explainability related to fairness?

COURSE EXPECTATIONS

Participation Questions (5%)

- Students must post responses to **participation questions** (available on the course website) on Gradescope 24 hours prior to the start of each class session

Overall Participation (5%)

- Students are responsible for **completing the readings** in full prior to the start of each week's session in order to facilitate productive **class discussion**. All readings will be freely available and linked in the course website
- Each student is responsible for preparing one five-minute **in-class brief** on one of the academic papers assigned as readings (more on this at the end of class)
- For each **writeup prompt** (available on the course website), students are responsible for responding in at least 500 words (one single-spaced page). Writeups are due via Gradescope before the start of the session in which they are due. The use case informing each writeup will be available on the course website

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

COURSE EXPECTATIONS

Quarter One Project (70%)

- 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 (15%)

- 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

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

COURSE RESOURCES

Course Website

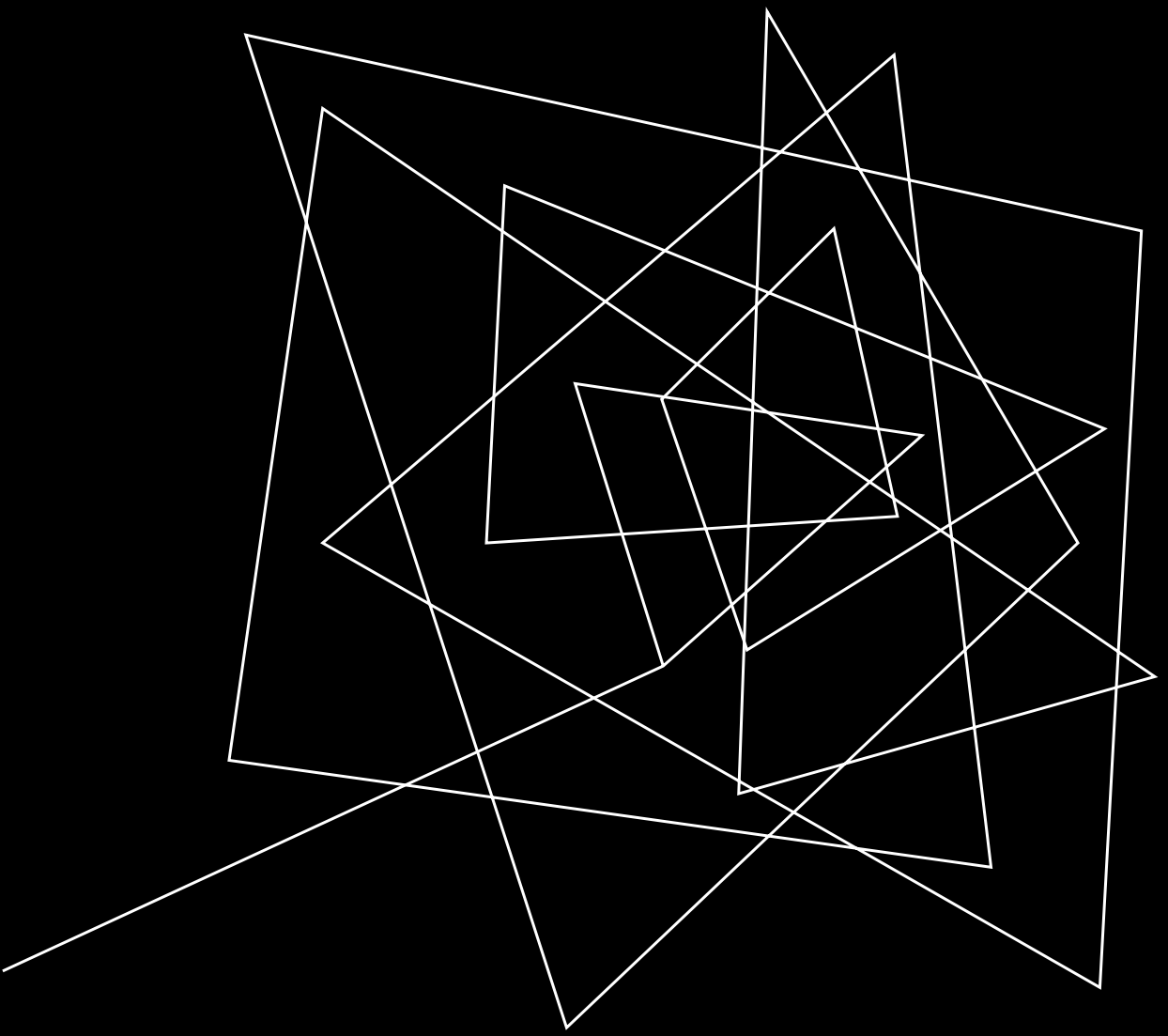
- A course website containing our weekly class schedule, assignment timelines, participation questions, and other resources is available at: https://nanrahman.github.io/capstone-responsible-ai/

Office Hours

- David will hold office hours Wednesdays 3-4PM in UC 302, Room 101
- Deloitte will hold additional office hours, to be scheduled as needed after we start the Q1 Project

Questions and Communications

- We will set up a Slack
- Although we encourage public questions (if you have the question, someone else probably does too!), for private or personal questions, you can reach out to David (ddanks@ucsd.edu) directly via email



STUDENT INTRODUCTIONS

WHAT IS RESPONSIBLE* AI?

**Also called: ethical AI, trustworthy AI*

- Accountable
- Transparent
- Understandable
- Regularly monitored
- Responsible and traceable
- Safe, secure, and resilient
- Accurate, reliable, and effective
- Purposeful and performance-driven
- Lawful and respectful of our Nation's values

- Accountability
- Transparency
- Human agency and oversight
- Privacy and data governance
- Technical robustness and safety
- Societal and environmental well-being
- Diversity, non-discrimination, and fairness

https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government
https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

WHAT IS RESPONSIBLE AI?

Deloitte.

- Fair and impartial
- Transparent and explainable
- Responsible and accountable
- Robust and reliable
- Respectful of privacy
- Safe and secure

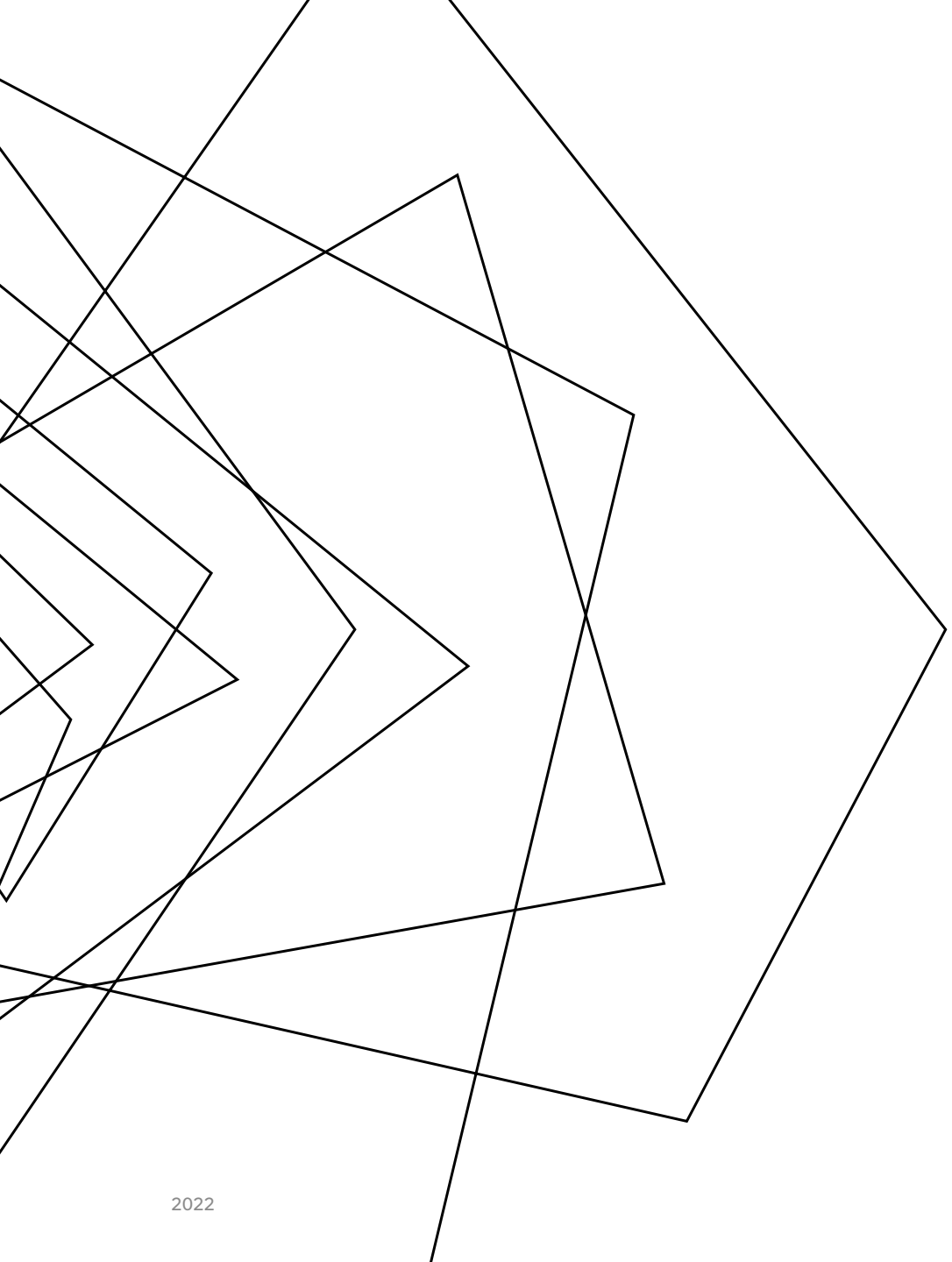
IBM

- Fairness
- Explainability
- Transparency
- Robustness
- Privacy

Google AI

- Avoid creating or reinforcing unfair bias
- Be accountable to people
- Incorporate privacy design principles
- Be built and tested for safety
- Be socially beneficial
- Uphold high standards of scientific excellence
- Be made available for uses that accord with these principles

https://www2.deloitte.com/us/en/pages/deloitte-analytics/solutions/ethics-of-ai-framework.html
https://www.ibm.com/cloud/learn/ai-ethics
https://ai.google/principles/



DISCUSSION

Each breakout room will receive a short description of an AI scenario. Take **five minutes** to discuss the use case, focusing on potential ethical considerations and impacts.

Select one member of your group to **share your team's scenario** and some of the ethical considerations you identified with the larger group.

SCENARIO 1

You are an AI practitioner working on the HR team for a tech startup. Your company wants to make their hiring process faster and less prone to human bias, so they ask you to build an automatic resume-screening tool. You decide to train this tool on the resumes of individuals who have been successful at your company, so it will identify similar individuals in the future. Since you only have a few employees, you also search LinkedIn for individuals with the same job at different companies and scrape their resume information. With this larger dataset, you build a resume-screening model. Your company deploys this model as the first step in your new hiring process; individuals who make it past the resume screen are then interviewed by humans on your team.

What ethical considerations can you think of related to building and deploying the AI tool in this scenario?

Inspired by:
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias

SCENARIO 2

You are an AI practitioner working for a large healthcare system. Due to a recent pandemic, the hospitals in your system are overwhelmed and doctors are struggling to triage, diagnose, and treat patients. You are asked to build an AI tool to help take some of the burden off healthcare workers by reading in patient chest X-rays, determining whether they have the virus causing the pandemic, and triaging them according to how severe their case is. You ask the hospitals in your network (including large, small, rural, urban, pediatric, non-pediatric, etc.) to send you chest X-rays for patients who had and patients who did not have the virus. You build a model with this data and send it to all of the hospitals in your system to use as they see fit.

What ethical considerations can you think of related to building and deploying the AI tool in this scenario?

Inspired by:
https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/

SCENARIO 3

You are a landlord in a large city. Tenants at one of your high-rise properties have been complaining that their packages are frequently being stolen from the building's common areas. To address this, you decide to purchase facial recognition software from an external vendor. You plan to install face scanners outside of the building's entrances that are trained to recognize your tenants and refuse entry to anyone else. To set up your system, you ask your tenants to upload a well-lit picture of their face to the AI vendor's website. The vendor will save the tenant photos as approved building residents, and the system will scan the faces of everyone who tries to enter the building against that list going forward.

What ethical considerations can you think of related to building and deploying the AI tool in this scenario?

Inspired by:
https://www.nytimes.com/2019/03/28/nyregion/rent-stabilized-buildings-facial-recognition.html

FOR NEXT WEEK

- Sign up for an **in-class brief** by 10AM on Monday, October 3
- Complete next week's **readings**
  - If you sign up to present on *Inherent Trade-Offs in the Fair Determination of Risk Scores* or *Who Audits the Auditors?*, come prepared to present next week and submit your presentation on Gradescope by 10AM PT on Friday, October 7
- Submit your answers to next week's **participation questions** to Gradescope by 10AM PT on Thursday, October 6
- Submit your response to **Writeup #1** to Gradescope by 10AM PT on Friday, October 7