CYB 4203/6203

Secure and Trustworthy AI

Presentation 3: Core AI Values, Human Rights, and Human-AI Collaboration

Monday, February 2, 2026

Topics: 2.2

Recap from last session

Key Themes: Sadness and Despair (mostly)

- 1.2: Societal stakes: AI's dual-use nature and transformative potential
- 1.3: Prominent failures: COMPAS, facial recognition, autonomous vehicles, LLMs
- 1.4: Societal influence: Social, economic, and geopolitical impacts
- 2.1: Ethical frameworks: Virtue ethics, utilitarianism, deontology

Discussion: What did you learn during Assignment 2?
(AI risk / harm incident case study with virtue / deontology / utilitarian ethical analysis)

Today’s Agenda

Key Themes: Hope and Optimism (mostly)

- Topic 2.2 - Core AI values
- Last 15 minutes of class today: Grad student huddle

Topic 2.2: Core AI Values

Fairness – Transparency – Accountability

Privacy – Autonomy – Safety – Sustainability*

*not an exhaustive list

Core AI Value - Fairness

What does ‘Fairness’ mean?

Answer: No one really knows – it’s super complicated…

- We could make this entire course just on Fairness in Machine Learning
  - Here’s the book! (*Fairness and Machine Learning: Limitations & Opportunities* - *Barocas, Hardt, Narayanan*)

- There is no one accepted definition that fits all contexts ¯_(ツ)_/¯
- Humans have been struggling with fairness since history began
- Chances are slim that we’ll figure it out before AGI cooks us all

Core AI Value - Fairness

Ok actually, it's not that hard :)

In fact, fairness is so easy that even capuchin monkeys understand it ➡

Excerpt from Frans de Waal TED Talk – 2013
https://www.youtube.com/watch?v=meiU6TxysCg

Core AI Value - Fairness

So, what might we mean when we say “Fair”?

Distributive Fairness – Allocation of benefits & burdens

- Equal shares for everyone?
- Proportional to contribution or merit?
- Prioritizing those with the greatest need?
- What are some examples?

Core AI Value - Fairness

So, what might we mean when we say “Fair”?

Procedural Fairness – Was the process fair?

- Were the rules applied consistently?
- Did all affected parties have a voice?
- Was the decision-maker impartial?
- Examples?

Core AI Value - Fairness

So, what might we mean when we say “Fair”?

Interactional Fairness – Were people treated with dignity and respect?

- Were individuals seen as individuals, and not just members of a group?
- Were stereotypes or assumptions imposed on them?
- So many examples…
- Political, religious, racial, cultural, national, professional, etc. etc.

Fairness is Contextual

Are these scenarios ‘fair’?

- A scholarship goes to the student with the highest GPA.
- A kidney transplant goes to the youngest patient on the list.
- A job goes to the candidate who interviewed best.
- Insurance premiums are higher for people with pre-existing conditions.

WAIT!

What does this have to do with AI or Machine Learning?

Maybe exploring common ML tasks

will shed some light on **fairness**?

Fairness in Machine Learning Tasks

ML Task Category: Classification

Assign input to discrete categories based on learned patterns.

- Email: Spam or Not Spam
- Medical Imaging: Malignant or Benign
- Loan Applications: Approve or Deny
- Content Moderation: Violates policy or Acceptable
- Hiring: Advance to interview or Reject

Is an AI/ML system making binary or categorical decisions that directly affect people?

Fairness in Machine Learning Tasks

ML Task Category: Regression and Prediction

Predicts a continuous value or probability based on input features

- Predicting house prices
- Estimating credit risk scores
- Forecasting patient readmission likelihood
- Predicting recidivism risk (probability of reoffending)
- Estimating insurance premiums

Are the predictions equally *accurate* across groups?
Do *errors concentrate* in particular populations?

Fairness in Machine Learning Tasks

ML Task Category: Ranking and Recommendation

Order items by predicted relevance, quality, or fit for a particular user/query

- Search engine results
- Social media feeds
- Product recommendations (ads, online marketplaces)
- Resume ranking for recruiters
- News article prioritization

Ranking creates *visibility hierarchies* with real consequences for people.
Are groups systematically ranked lower? Do *filter bubbles* limit exposure?

Fairness in Machine Learning Tasks

ML Task Category: Clustering and Segmentation

Discover data structure and group similar items together without predefined labels

ML Task Category: Generation

Create new content – text, images, audio, video, code – based on learned patterns

ML Task Category: Detection and Recognition

Identify specific patterns, objects, faces, speech, or events within input data

ML Task Category: Optimization and Resource Allocation

Find the best solution given constraints – allocating limited resources

Core AI Value - Fairness

Multiple Definitions; Persistent Challenges

- Individual fairness: Similar individuals are treated similarly
- Group fairness: Equitable outcomes across demographic groups
- Substantive fairness: Concerned with end results, regardless of process
- Mathematical tension: Definitions can be mutually exclusive, i.e.: under some conditions, it is provably impossible to satisfy all definitions*
- Challenge: Fairness is context-dependent and value-laden

* Or is it?

Core AI Anti-Value - Unfairness

YOU'VE been a victim!

Think of a time you were affected by an 'unfair' decision

- What would have been 'fair'?
- What caused the unfair decision?
- Could an algorithm / human have done better?
    - **Prove it!**

Why else is ML Fairness so hard?

Because Algorithms!

When Algorithms enter the picture

- Mathematical / logical algorithms need explicit definition
    - Definition is hard
- Defining groups / protected classes requires measurement
    - Measurement is hard
- Data reflects the world we have, not the world we want
- Optimization creates trade-offs

* Or is it?

Core AI Value - Transparency

Easier to explain than Fairness, but still difficult to achieve

The degree to which an AI / ML system's logic, data sources, development process, and performance are clearly communicated, understood, and auditable

- Model transparency: Understanding system architecture and logic
- Data transparency: Knowing training data sources and characteristics
- Decision transparency: Explaining individual predictions or actions
- Trade-offs: Performance vs. interpretability, IP vs. openness
- Stakeholder-appropriate: Different audiences need different explanations

Transparency (AKA Explainability / Interpretability)

Resources

Interpretable Machine Learning: A Guide for Making Black Box Models Explainable

Christoph Molnar - https://christophm.github.io/interpretable-ml-book/

Van Der Schaar Lab – ML in Healthcare focus

https://www.vanderschaar-lab.com/interpretable-machine-learning/

Core AI Value - Accountability

Easy to define - Difficult to enforce

The degree to which individuals and organizations are responsible for the actions, decisions, impacts, and outcomes of their AI systems. Requires transparency, human oversight, and clear liability.

- Clear ownership: Who is responsible when AI systems fail?
- Governance structures: Oversight, audit trails, documentation
- Redress mechanisms: Processes for appeals and corrections
- Liability frameworks: Legal responsibility for AI harms
- Challenge: ‘Many hands’ problem, diffusion of responsibility

AI Accountability Resources

AI Now Institute – New York University

Founded in 2017 – produces diagnosis and policy research on artificial intelligence, public accountability of AI industry for social consequences. https://ainowinstitute.org/

Algorithmic Justice League

Combines research, art, and advocacy to expose bias in AI systems and push for accountability mechanisms. https://www.ajl.org/

Ada Lovelace Institute

An independent research institute with a mission to ensure data and AI work for people and society. https://www.adalovelaceinstitute.org/

TO BE CONTINUED…

Next time:

- **Privacy**
- **Autonomy**
- **Safety**
- **Sustainability**

2.3: AI and Human Rights

2.4: Human-AI Collaboration

Grad student huddle time

This presentation was drafted and refined using Claude 4.5 Sonnet and Claude Code.

All content was reviewed and approved by Dallas Elleman.