CYB-4203/6203 - Secure and Trustworthy AI

Week 3 Day 1: Core AI Values (didn't get to Human Rights or Human-AI Collaboration)

I. Pre-Class Discussion

Current AI news: game company stocks, Microsoft pulling back on AI due to backlash

Movie recommendations: *Her* (2013), *Mercy* (new AI thriller)

II. Insights from Student Questionnaire (Assignment 1)

Six themes identified from student responses:

1. Intellectual Curiosity — Interest in theoretical/philosophical questions (intelligence, consciousness, secure AI feasibility)
2. Safety, Legal & Ethical Concerns — Tech outpacing laws; AI vulnerability; criminal use of AI tools
3. Recognition of AI's Massive Impact — AI integrated into daily life; access to sensitive information; faster pace of change than any prior technology
4. Career Opportunities — AI + cybersecurity skills highly valued by employers
5. Desire to Build Secure AI — Hands-on learning; creating trustworthy, privacy-aware, bias-minimized systems
6. Academic/Research Goals — Interest in reliability, risk, and responsible AI development

III. Student Sharing: Current Events

https://Moltbook.com/ : New AI-only social media platform; AI agents interacting autonomously; rapid emergence of unexpected behaviors (e.g., “Church of Molt - https://molt.church/ ”)

Discussion of reckless - and likely very dangerous - AI agent permissions

Discussed question of accountability when agents given permissions cause damage, who is responsible?

IV. Core AI Values (Main Content)

A. Fairness

No single accepted definition; humans have struggled with fairness throughout history

Video: Capuchin monkey fairness experiment (inequity aversion is innate)

Types of fairness:

Distributive — Allocation of benefits/burdens (equal shares, merit-based, need-based)

Procedural — Fair processes, impartial decision-makers, voice in decisions

Interactional — Dignity, respect, avoiding stereotypes

Class exercises: Are these fair?

Scholarship to highest GPA

Kidney transplant to youngest patient

ML tasks where fairness matters:

Classification (spam, loans, hiring, content moderation)

Regression/Prediction (credit scores, recidivism, insurance)

Ranking/Recommendation (search results, filter bubbles)

Clustering, Generative AI, Detection, Optimization

Why ML fairness is hard:

- Multiple conflicting definitions (individual vs. group fairness)
- Mathematical impossibility of satisfying all fairness criteria simultaneously
- Context-dependent evaluation

Resources: UC Berkeley Fairness in ML course (2017); fairness impossibility papers (links in Week 3 Day 1 presentation)

B. Transparency

Definition: Degree to which logic, data sources, development process, and performance are clearly communicated and auditable

Types:

- Model transparency
- Data transparency
- Decision transparency

Trade-offs: Performance vs. interpretability; IP protection vs. openness

Resources:

- *Interpretable Machine Learning* by Christoph Molnar (free online book)
- Vanderschaar Lab (healthcare-focused interpretable ML)

C. Accountability

Definition: Individuals/organizations responsible for AI actions, decisions, and outcomes

Requires: Transparency, human oversight, clear liability

Key questions:

- Who is responsible when AI fails?
- Who pays?

Components:

- Governance structures
- Audit trails
- Redress mechanisms
- Liability frameworks

Challenge: “Many hands problem” — diffusion of responsibility

Resources:

- AI Now Institute
- Algorithmic Justice League
- Ada Lovelace Institute

V. Wrap-Up & Preview

- Next session: Privacy, Autonomy, Safety, Sustainability; AI Human Rights; Human-AI Collaboration
- Reminder: Chapter 6 reading still relevant
- Homework: Forthcoming Wednesday

VI. Graduate Student Discussion

- End-of-semester presentation project (tie to existing research interests)
- Coding proficiency check-in
- Introduction to Claude Code and AI-assisted development tools
- Offer to provide API access for experimentation
- Discussion of safe sandboxing practices (VMs, limited permissions)
- Student experiences with AI coding assistants

Duration: ~1.5 hours | *Format: Lecture + discussion + video*