CYB-4203/6203 - Secure and Trustworthy AI

Week 2 Day 1: Societal Stakes and Ethical Frameworks - Wednesday, January 28, 2026

Annotated Class Discussion Transcript Outline (started late - about 15 minutes into class)

I. Introduction & Housekeeping

- Recording consent notice for transcript purposes
- Transition from previous discussion ("the good stuff") to risks ("the bad stuff")

II. Major AI Risks: Worker Displacement

- **Job displacement concerns**
    - Impact on tech sector and software development
    - Uncertain future for current students' career paths
- **Class discussion: Cybersecurity sector impacts**
    - Automation of entry-level tasks reducing junior positions
    - Internship and early-career roles being replaced
    - Long-term concern: elimination of experience pipeline for future leadership
- **Systemic challenges**
    - Pace of AI advancement vs. society's ability to create policy solutions
    - Corporate competitive pressure as an "arms race"
    - Monetization as the primary driver of problem-solving
    - Counterpoint: Large gaps create new opportunities for those who adapt
- **Implications for education and career preparation**
    - Training may shift back to universities and self-directed learning
    - Importance of proactive engagement with AI tools
    - "Skate to where the puck is heading" (Gretzky reference)

III. The Singularity and Exponential Growth

- **Student question on "liftoff" and exponential AI progress**

- **Discussion of the singularity concept**

    - Recursive self-improvement leading to intelligence explosion
    - Event horizon metaphor: unpredictability beyond a certain point

- **Ray Kurzweil reference**

    - Instructor's personal meeting with Kurzweil
    - ~80% prediction accuracy rate
    - Founded Singularity Institute
    - **NOTE**: Kurzweil's predictions have been criticized for being overly optimistic and inaccurate. The Singularity Institute has faced criticism for its focus on AI safety and its lack of concrete solutions.

- **Human cognitive limitations**

    - Brains evolved for linear prediction (intercepting prey)
    - Difficulty comprehending exponential relationships

- **Class consensus**: Likely to occur unless catastrophic disruption; uncertain timeline (5–20 years)

- **Technology as cumulative catalyst**

    - Each invention accelerates the next
    - Combinatorial effects exceed sum of parts

IV. Additional AI Risks

- **Cognitive and skill degradation**

    - Student over-reliance on ChatGPT
    - Medical deskilling (e.g., radiology AI reducing need for trained specialists)
    - Historical parallel: blacksmithing/horseshoes becoming obsolete

- **Bias amplification**
    - AI trained on biased data reinforces societal biases
- **Misinformation**
    - Political misinformation affecting elections
    - Swaying public opinion
- **Emerging threats**
    - AI-powered blackmail schemes
    - Deepfakes (example: $25 million bank heist via video call impersonation)
    - Model collapse from training on synthetic data

V. The Dual-Use Challenge

- **Core principle**: AI is a neutral tool; ethics depend on application
- **Facial recognition examples**
    - Positive: Tracking Boston Marathon bombers; Global Entry convenience
    - Negative: Authoritarian surveillance; monitoring political demonstrations
- **Natural language processing**
    - Positive: Khan Academy's Khanmigo AI tutoring
    - Negative: Zelensky deepfake and election manipulation examples
- **Automation**
    - Benefits: Reducing dangerous/expensive processes
    - Costs: Job losses with real economic impacts despite GDP growth

---

VI. Prominent AI Failures – Case Studies

A. COMPAS Recidivism Algorithm

- Purpose: Predict criminal recidivism for sentencing/bail decisions
- Problems: Racial bias in predictions; no more accurate than untrained humans guessing
- Impact: Real harm to individuals' lives based on flawed predictions

B. Gender Shades Study (Facial Recognition)

- Finding: Higher misidentification rates for darker-skinned females than lighter-skinned males.
- Cause: Biased training datasets
- **NOTE**: More info here: http://gendershades.org/overview.html

C. Autonomous Vehicle Safety Incidents

- Companies affected: Tesla, Uber, Cruise
- Issue: Premature deployment before safety was established
- **Class discussion on utilitarian perspective**
    - Student point: Autonomous vehicles have lower incident rates than humans per mile
    - Media amplification of individual incidents vs. statistical safety
    - Waymo statistics: <10% of human driver incident rate, no fatalities **NOTE**: THIS STATISTIC IS WRONG LOL
    - Instructor counterpoint: Potential for catastrophic systemic failures (e.g., simultaneous crashes)
    - Unresolved legal questions: liability for accidents, compensation for victims **GOOD PAPER HERE**: https://pubmed.ncbi.nlm.nih.gov/39485678/

VII. LLM-Specific Failures

- **Hallucination and bias amplification**

- **Cyber attacks using LLMs**

    - Anthropic report (November 2024): Chinese threat actor using Claude for cyber espionage **ARTICLE**: https://www.anthropic.com/news/disrupting-AI-espionage
    - Targeted ~30 global entities (government, tech, financial, chemical, agencies)
    - Student note: Nearly 100% of major cyber attacks now incorporate AI

- **Microsoft Tay chatbot**

    - Deployed on Twitter; quickly trained by users to express Nazi sympathies
      WIKIPEDIA: https://en.wikipedia.org/wiki/Tay_(chatbot)

VIII. AI Incident Databases (Homework Resource)

- MIT AI Risk Initiative - https://airisk.mit.edu/
- AI Incident Database - https://incidentdatabase.ai/
- AIAAIC - https://www.aiaaic.org/home
- **Assignment preview**: Select an incident, write one-page executive report, analyze through ethical frameworks (virtue, utilitarianism, deontology)

IX. Social and Cultural Impacts

- Slide noted but largely skipped due to time

- **Student observation**: The overwhelm itself demonstrates the social impact

- **Class discussion on AI dependency**

    - Desire for AI to handle complexity and reduce stress
    - Danger of reduced capacity to handle difficulty
    - Example: High schoolers unable to read multiple paragraphs
    - *Dune* series reference: Society outlawed AI because people became "slaves to it"
    - Balance needed: AI expands processing capacity but risks replacing thinking

- **Educational implications**

    - Need for proper instruction on effective AI use
    - Current pitfall: Using AI to avoid engagement rather than enhance understanding
    - "AIs talking to each other" (AI-written emails being read by AI)

- **Instructor guidance on AI use for coursework**

    - Encourages AI to accelerate work
    - Warns against replacing thinking with AI shortcuts

    - Prefers honest incomplete work over polished work without understanding

- **Note for future**: Include transhumanism as future discussion topic

---

X. Geopolitics and Strategic Competition

- **Military applications**

    - Autonomous weapons systems
    - Drone warfare
    - "Terminator" scenarios

- **Economic competitiveness**

    - CHIPS Act: U.S. restrictions on AI hardware sales to China
    - Geopolitical positioning for advantage

- **Stargate Initiative**

    - Trump administration AI infrastructure investment
    - Hundreds of billions in data centers and energy infrastructure
    - Tax-funded acceleration of U.S. AI development

- **Surveillance and control concerns**

- **AI arms race dynamics**

XI. Philosophical and Ethical Frameworks

A. Virtue Ethics

- **Core question**: What kind of person/society do we want to be?
- **Focus**: Wisdom, justice, compassion, human flourishing
- **Application to AI**: Does this system promote virtuous outcomes?
- **Challenge**: Cultural variation in values

B. Utilitarianism

- **Core principle**: Maximize overall welfare; greatest good for greatest number

- **Focus**: Consequences, cost-benefit analysis, aggregate welfare
- **Application to AI**: Quantify harms and benefits mathematically
- **Challenge**: Difficulty quantifying harm; individual vs. collective trade-offs
- **Classic example**: Trolley problem

C. Deontology

- **Core principle**: Duties, rights, and moral rules regardless of consequences
- **Focus**: Obligations to each other; universal human rights
- **Application to AI**: Inviolable lines AI must never cross
- **Challenge**: Defining and enforcing universal rights

XII. Ethical Framework Application – Scenario 1: Pandemic ICU Allocation AI

Scenario

- AI allocates ICU beds/ventilators based on survival probability predictions
- Result: 1,000 additional lives saved vs. first-come-first-served
- Problem: Racial/socioeconomic bias causes disadvantaged communities to be 30% less likely to receive care with similar survival odds

Analysis by Framework

Framework	**Verdict**	**Reasoning**
Virtue Ethics	Unclear/No	Prioritizing algorithmic efficiency over fairness may not promote human flourishing
Utilitarianism	Yes	Net gain of lives saved outweighs distributional concerns
Deontology	No	Using biased proxies violates equal treatment and human dignity

---

XIII. Ethical Framework Application – Scenario 2: AI Relationship Optimization Platform

Scenario

- AI analyzes personality, communication, and history to optimize romantic relationships
- Suggests timing for difficult conversations, when to end relationships, conflict responses
- Results: 40% higher satisfaction, 60% fewer divorces, improved mental health
- Trade-off: Users feel "managed," reduced spontaneity, relationships feel "optimized but hollow"

Class Discussion

- *Black Mirror* comparison
- Current analog: People already use ChatGPT for relationship advice, wedding vows
- Questions raised:
    - Is following AI advice different from following a therapist's advice?
    - Does the medium (machine vs. human) affect authenticity?
    - Would you want a partner making decisions based on AI recommendations?
    - Are we already "manipulated" by culture, therapy-speak, individualism?

Analysis by Framework

Framework	**Question Posed**	**Verdict**
Virtue Ethics	Does this promote human flourishing and authentic relationships?	Unclear
Utilitarianism	Do measurable improvements justify concerns about authenticity?	Possibly yes
Deontology	Does AI-mediated intimacy violate autonomy and authentic self-expression?	Possibly no

---

XIV. Closing

- **Key takeaway**: These frameworks provide different lenses for evaluating AI systems; no single "right" answer

- **Legal approaches** (from textbook Chapter 6) also face limitations

- **Instructor goal**: Increase class discussion and participation

- **Homework assignment**: Available on Blackboard

    - Select an AI incident from databases
    - Write one-page report
    - Analyze through three ethical frameworks

- **Next class preview**: Core AI values and human rights; Section 2 of syllabus