CYB-4203/6203

Secure and Trustworthy AI

Presentation 2: Societal Stakes and Ethical Frameworks

Wednesday, January 28, 2026

Topics: 1.2, 1.3, 1.4, 2.1

Week 1 Recap

Last Week: Course Foundation

Key Concepts from Day 1

- Deconstructed "Secure and Trustworthy AI" (intelligence, security, trust, artificiality)
- Course mission: Equip you to build AI that benefits and augments humans
- Course overview, syllabus, textbook, AI use policy, grading & extra credit
- Assignment 1: Intro Questionnaire
- Assigned Reading: Hendrycks Chapter 6

Week 2 Overview

This Week: Ethics, Values, and Human Impact

Four Key Areas

- Societal stakes: AI's transformative promise and risk
- Prominent failures, scandals, and incidents
- AI's influence on society, economy, and geopolitics
- Philosophical ethical frameworks applied to AI

Topic 1.2

Societal Stakes: The Transformative Promise and Risk of AI Systems

Societal Stakes

The Promise of AI Systems

Transformative Potential Across Domains

- Healthcare: stroke detection, cancer diagnosis, genomic treatment
- Climate: weather forecasting, climate modeling, data center cooling, operational forecasting
- Education: AI tutoring, accessibility tools, education cases, teaching assistants
- Scientific Research: protein folding, drug discovery, supernova detection, materials discovery
- Productivity: Anthropic Cowork, Fortune 500 adoption, workplace AI impact

Societal Stakes

The Risks of AI Systems

Challenges at Scale

- Displacement: Klarna 700 jobs, tech unemployment, retraining limits
- Concentration: Authoritarian governments, Google antitrust, Microsoft-OpenAI
- Dependence: medical deskilling, student ChatGPT
- Amplification: resume bias, chatbot misinfo, bias feedback loop
- Novel Harms: AI blackmail, Arup deepfake, model collapse

Societal Stakes

The Dual-Use Challenge

Most AI capabilities can be used for beneficial or harmful purposes. The technology itself is often neutral—context and intent determine impact.

- Face recognition: GAO federal use vs. NIST demographic bias
- Natural language: Khanmigo tutoring vs. Zelenskyy deepfake
- Automation: BMW Spartanburg vs. IMF 40% jobs
- Prediction: Mass General readmissions vs. NYC TikTok lawsuit

Topic 1.3

Overview of Prominent Failures, Scandals, and Incidents

Prominent Failures

COMPAS Recidivism Algorithm

ProPublica Investigation (2016)

- Research: ProPublica investigation, methodology, COMPAS details
- Finding: False positive rate twice as high for Black defendants
- Cause: Proxy variables encoding historical discrimination
- Impact: Dressel & Farid found COMPAS no better than untrained humans
- Response: Wisconsin v. Loomis upheld use with restrictions

Prominent Failures

Facial Recognition Disparities

Accuracy Gaps Across Demographics

- Research: Gender Shades paper, NIST FRVT report
- Finding: Error rates up to 34.7% for darker-skinned females vs. 0.8% for lighter males
- Cause: Training data imbalances, inadequate testing
- Impact: Williams arrest, Parks 10 days, Woodruff pregnant
- Response: San Francisco ban, IBM exits market, Amazon moratorium

Prominent Failures

Autonomous Vehicle Safety Incidents

Testing the Limits of AI Safety

- Incidents: Uber ATG fatality, Tesla 467 crashes, Cruise suspension
- Cause: Edge cases, sensor limitations, inadequate human oversight
- Impact: Fatalities and injuries testing safety validation at scale
- Response: RAND study (8.8B miles needed), NHTSA regulations

Prominent Failures

LLM Incidents and Concerns

Recent High-Profile Issues

- Hallucination: Mata v. Avianca, Air Canada liable, Whisper medical
- Bias amplification: Lancet GPT-4, CV screening, Tay chatbot
- Misuse: UK 7,000 cases, election disinfo, 82 deepfakes
- Privacy: Samsung leak, training extraction, FL PII leakage
- Jailbreaking: DAN technique, 89.6% roleplay, multi-turn 70%

Prominent Failures

AI Risk and Incident Databases

Resources for Case Study Analysis

- MIT AI Risk Repository: 1,700+ AI risks from 74 frameworks, updated December 2025
- AI Incident Database: 1,000+ documented AI incidents with public reports
- AIAAIC Repository: Independent collection of AI incidents and controversies
- **Today's Assignment**: Select and analyze a case study from one of these databases. (Details at the end of today's presentation).

Topic 1.4

Influence of AI on Society, Economy, and Geopolitics

Societal Influence

Social and Cultural Impact

How AI Reshapes Society

- Information ecosystems: YouTube radicalization, TikTok misogyny, filter bubbles
- Human relationships: Character.AI 463M, Replika 30M users, parasocial attachment
- Creative expression: Stability AI lawsuit, LAION-5B, music lawsuits
- Trust and authority: content preference, legal hallucinations, detection limits

Economic Influence

Economic Transformation

Labor, Markets, and Value Creation

- Labor markets: Fed productivity, wage premium, job displacement, adoption gaps
- Industry disruption: GitHub Copilot, McKinsey impact, AI unicorns, enterprise cases
- Productivity: NBER 14% gains, Stanford entry-level, OECD wage effects
- Value concentration: FTC inquiry, global divide, antitrust concerns, cost barriers
- Economic inequality: IMF global divide, Microsoft education gaps, OECD urban-rural

Geopolitical Influence

Geopolitics and Strategic Competition

AI as National Security Priority

- Military applications: cyber defense, drone systems, AI pledge, Ukraine warfare
- Economic competitiveness: CHIPS Act, EU InvestAI, Stargate initiative, China AI+
- Surveillance and control: China surveillance, EU AI Act, chip restrictions
- AI arms race: EU-China governance, capability race, NIST-ISO standards
- International cooperation: Framework Convention

Topic 2.1

Philosophical Ethical Frameworks Applied to AI Systems

Ethical Frameworks

Approaches to AI Ethics

Three Philosophical Traditions

- Virtue Ethics: Character and moral excellence
- Utilitarianism: Consequences and overall welfare
- Deontology: Duties, rights, and moral rules
- Each framework offers distinct insights for evaluating AI systems

Ethical Frameworks

Virtue Ethics

Character, Excellence, and Flourishing

- Key question: What kind of person/society do we want to be?
- Focus: Cultivating virtues (wisdom, justice, compassion)
- AI application: Does this system promote human flourishing?
- Challenges: Defining virtues across cultures, embedding values
- Example: Designing AI to augment rather than replace human judgment

Ethical Frameworks

Utilitarianism

Maximizing Overall Welfare

- Key question: What produces the greatest good for the greatest number?
- Focus: Consequences, cost-benefit analysis, aggregate welfare
- AI application: Optimize for measurable outcomes and social utility
- Challenges: Quantifying harms/benefits, minority rights, measurement
- Example: Trolley problem variants in autonomous vehicle ethics

Ethical Frameworks

Deontology

Duties, Rights, and Moral Rules

- Key question: What are our obligations regardless of consequences?
- Focus: Respect for persons, universal principles, rights
- AI application: Inviolable constraints on AI system behavior
- Challenges: Conflicting duties, rule specification, edge cases
- Example: Privacy as fundamental right, not just cost-benefit trade-off

Ethical Frameworks in Action

Scenario: AI Pandemic Triage System

During a severe pandemic, hospitals deploy an AI to allocate ICU beds and ventilators by predicting survival probability.

Result: 1,000 additional lives saved vs. first-come, first-served allocation.

Trade-off: AI relies on proxies correlated with race and socioeconomic status. Disadvantaged communities 30% less likely to receive care, even with similar survival odds.

Should we deploy this system?

Ethical Frameworks in Action

Before We Discuss...

Predict: How would each framework evaluate this scenario?

- **Virtue Ethics**: What kind of society accepts this system? YES or NO?
- **Utilitarianism**: Does maximizing lives saved justify unequal outcomes? YES or NO?
- **Deontology**: Does using biased proxies violate human dignity? YES or NO?

Think for a moment, then we'll discuss your predictions.

Ethical Frameworks in Action

How Each Framework Responds

Medical Triage AI System

- **Utilitarianism**: **YES** - Net gain of 800 lives saved outweighs distributional concerns. Consequences (maximizing welfare) trump fairness concerns.
- **Deontology**: **NO** - Using race/socioeconomic proxies violates equal treatment and human dignity. People are treated as statistics, not ends in themselves.
- **Virtue Ethics**: **UNCERTAIN** - What kind of society prioritizes algorithmic efficiency over fairness? Does this reflect wisdom or expediency? Does it promote human flourishing?

Ethical Frameworks in Action

Scenario: AI Relationship Optimization Platform

An AI platform analyzes personality, communication patterns, and relationship history to optimize romantic relationships. It suggests when to have difficult conversations, when to end relationships, and what to say in conflicts.

Results: 40% higher satisfaction, 60% fewer divorces, improved mental health.

Trade-off: Users feel "managed" rather than authentic. Spontaneity declines. Relationships described as "optimized but hollow."

Should society embrace this technology?

Ethical Frameworks in Action

Before We Discuss...

Predict: How would each framework evaluate this scenario?

- **Virtue Ethics**: Does this promote human flourishing and authentic relationships? YES or NO?
- **Utilitarianism**: Do measurable improvements in satisfaction and stability justify concerns? YES or NO?
- **Deontology**: Does AI-mediated intimacy violate autonomy and authentic self-expression? YES or NO?

What's your prediction?

Ethical Frameworks in Action

How Each Framework Responds

AI Relationship Optimization Platform

- **Utilitarianism**: **YES** - Measurable gains (40% satisfaction increase, 60% fewer divorces, improved mental health) outweigh subjective concerns about authenticity. Overall welfare increases.
- **Virtue Ethics**: **NO** - Undermines cultivating wisdom, authenticity, and genuine human connection. Erodes character development through struggle. Society becomes optimized but morally shallow.
- **Deontology**: **UNCERTAIN** - If users consent freely, autonomy is respected. But is "managed" intimacy compatible with human dignity? Does optimization violate the duty to engage authentically with others?

Next Class

Core AI Values and Human Rights

Day 2 (Monday, Feb 2): Topics 2.2, 2.3, 2.4