CYB 4203/6203

Secure and Trustworthy AI

Presentation 5: Potential Harms, Misuse, and Responsible Innovation

Monday, February 9, 2026

Topic: 3.1

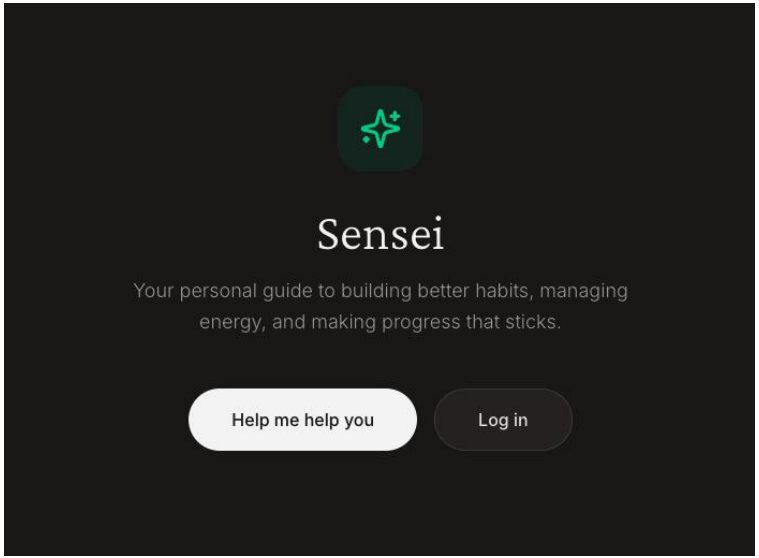
Recap – Last week

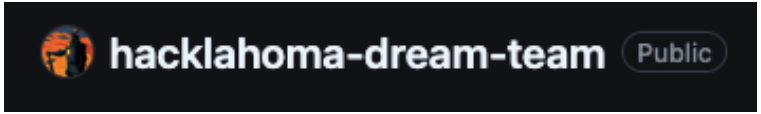
- 2.2 Core AI values:
    - Privacy
    - Autonomy
    - Safety
    - Sustainability
- 2.3 AI and human rights
- 2.4 Human-AI collaboration

Recap – Assignments

- I'll try to post grades from Assignment 2 by Wednesday
- Clarifications re: Assignment 3
    - **Choose 1 topic to write about**
    - **Use as many refs/resources as you want**

Recap – My Weekend





Recap – My Weekend

What I learned about the tech industry

- The tech job market is brutal for new grads – massive layoffs
- Record CS enrollment, and far fewer entry-level positions than candidates.
- AI tools have made building software trivially accessible, so the new differentiator is building *good* software with taste, not just building software at all.
- Students must combine theoretical fundamentals with relentless side-project building and continuous shipping, because the industry is closing in fast and no one is safe.

Cybersecurity Industry Considerations

AI amplifies threats, but also the demand for security.

- Every company racing to ship AI-powered software needs people who understand how to secure it.
- The same low barrier to entry that floods the market with developers also floods the market with *insecure* software.
- That's job security for you — if you position yourselves at the intersection of security and AI.

Cybersecurity Industry Considerations

Build and break things

- The cybersecurity equivalent of "always ship and learn" is standing up systems and then attacking them.
- CTFs, home labs, bug bounties, and red team exercises build intuition that coursework alone can’t.
- AI tools make spinning up practice environments faster than ever. **Use them.**

Cybersecurity Industry Considerations

Fundamentals compound

- Networking, OS internals, cryptography, and software architecture aren't going away because of AI
- They're the foundation that lets you evaluate whether an AI-generated security recommendation is sound or dangerously wrong.
- The students who deeply understand *why* things work will outperform those who only know *how* to prompt a tool.

Cybersecurity Industry Considerations

My overarching message to you

- **AI is reshaping who gets hired**
- **Security expertise + hands-on building experience = durable position**

Students who *treat AI as a force multiplier for learning*
--rather than a replacement for it--
will be the ones still standing.

Topic 3.1: Unintended Harms of AI

Higher-order Effects

Unintended Harms of AI Higher-order Effects

Self-amplifying Feedback Loops

Ensign et al. - 2018, "Runaway Feedback Loops in Predictive Policing" - https://arxiv.org/abs/1706.09847

- Historical data → more police → more arrests → "confirms" prediction → retrain
- More arrests are made due to police presence (not higher crime), new arrest data "confirms" the prediction, model retrains and strengthens the bias.

Unintended Harms of AI Higher-order Effects

Self-amplifying Feedback Loops

Obermeyer et al. (2019) “Dissecting racial bias in an algorithm used to manage the health of populations” https://pubmed.ncbi.nlm.nih.gov/31649194/

- An algorithm affecting ~200M patients used *healthcare spending* as a proxy for health needs.
- Structural inequality → less spent on Black patients at equivalent illness levels → algorithm systematically under-predicts their needs.
- Denied care management → worse outcomes → more acute crises → higher crisis costs but lower preventive costs → data reinforces the original bias

Unintended Harms of AI Higher-order Effects

Behavioral Changes People Don’t Notice in Themselves

Chilling effects and self-censorship. Buchi, Festic & Latzer (2022, *Big Data & Society*) found that the *perception* of algorithmic surveillance -- not even actual surveillance -- causes people to restrict their communication,
opinion-voicing, and information-seeking. The perverse dynamic: the most informed, privacy-aware citizens self-censor the most, systematically withdrawing the most thoughtful voices from public discourse.

Unintended Harms of AI Higher-order Effects

Behavioral Changes People Don’t Notice in Themselves

Anticipatory conformity. Research on China's Social Credit System shows citizens engaging in "automatic self-monitoring and adjustment of behavior" -- acting not from genuine belief but to maintain a score. Hardt et al. (2025) formalize how algorithmic scoring rewards *strategic manipulation over genuine improvement*, systematically advantaging those with resources and knowledge to game the system.

Student Research: Real-World AI Harms

Each student performed in-class rapid investigation of a case study of AI systems causing real-world harm.

Topics span algorithmic discrimination, government automation failures, misinformation, and research integrity.

The following slides summarize and expand their results.

Student Research

CFPB v. Fairway Independent Mortgage (2024)

- CFPB and DOJ alleged Fairway discriminated against majority-Black neighborhoods in Birmingham, AL
- Offices, marketing, and referral networks concentrated in majority-white areas
- AI-powered underwriting and marketing systems trained on historical data can learn and scale discriminatory patterns
- Even without race as a variable, ZIP code proxies create "algorithmic redlining"

CFPB Complaint

Student Research

The Liar's Dividend

- Politicians can dismiss real scandals by claiming "misinformation" -- especially effective for text-based scandals
- Two strategies: create informational uncertainty, or rally partisan supporters
- AI-generated content makes false claims of fakery more plausible
- Believing the politician's rebuttal strongly correlates with continued support
- Video-based scandals are harder to dismiss than text-based ones

The Liar's Dividend - American Political Science Review (Cambridge)

Student Research

Race After Technology (2019)

- Technology is not neutral -- algorithms reflect the social and institutional contexts in which they are built
- Bias is not a technical glitch but a sign of deeper structural inequality embedded in data and design
- Introduced "the New Jim Code": automation that hides, speeds, and deepens discrimination
- Questions whether some systems should exist at all, and who they ultimately serve

Data & Society Podcast: Race After Technology

Student Research

Algorithmic Monoculture (Kleinberg & Raghavan, 2021)

- When many institutions use the same algorithm, decision quality decreases
- Feedback loops: popular recommendations reinforce themselves
- A single flawed algorithm means all decisions are at risk -- including susceptibility to bias
- Paradox: introducing a more accurate algorithm leads to wider adoption, recreating the monoculture

Kleinberg & Raghavan (2021) - PNAS

Student Research

Louis et al. v. SafeRent Solutions (2022-2024)

- AI scoring system disproportionately penalized Black and Hispanic renters and housing voucher recipients
- Algorithm produced opaque scores landlords could not adjust or understand
- "324" -- the score that denied Mary Louis housing, with no explanation of how it was calculated
- $2.3M settlement; SafeRent barred from scoring voucher applicants for 5 years

The Guardian: SafeRent AI Tenant Screening Lawsuit

Student Research

Digital Redlining

- Specific groups excluded from equal access to digital tools like internet, creating systemic division
- COVID amplified impact: minorities relied on internet for work, school, virtual health visits
- AI healthcare tools (facial recognition, speech recognition) work best for native English speakers
- Digital patient portals use limited languages, require device ownership
- Creates widening gaps in quality education, employment, and healthcare access

AHA Journal: Digital Redlining and Cardiovascular Health | Johns Hopkins: Digital Redlining and AI in Healthcare

Student Research

MiDAS Automated Fraud Detection (2013-2015)

- ~40,000 Michigan residents falsely accused of unemployment fraud by automated system
- No human oversight; neither agency nor state could produce evidence for fraud claims
- Victims suffered wage garnishments, home foreclosures, seized tax returns, bankruptcy
- Fraud allegations disrupted job searches, perpetuating cycles of unemployment
- $20M class-action settlement after seven years of litigation

Ford School: MiDAS Explainer (PDF)

Student Research

Paper Mills

- Organized services producing thousands of fake research papers using templates and automation
- Driven by publication pressure and academic incentive structures
- Include fake data, reused images, template text; may exploit peer review systems
- Scientists, doctors, and students pay thousands to be named as authors
- AI tools now increase output speed and realism of fabricated research

Paper Mills Enabling Fraud at Scale - PNAS (2025)

Student Research

Dutch Toeslagenaffaire (2005-2019)

- Tax authority algorithm profiled dual-nationality families as "higher risk" for benefit fraud
- Small administrative mistakes treated as intentional fraud; families ordered to repay tens of thousands of euros
- Pushed families into severe debt, job loss, divorces, mental health crises
- Amnesty International declared the system unlawful and discriminatory
- Entire Dutch cabinet resigned in January 2021

Xenophobic Machines: Amnesty International Toeslagenaffaire Report

Student Research

Amazon Warehouses: Robots and Injury Rates

- 50% higher injury rates in warehouses with autonomous robots (4-year, 150-warehouse report)
- Robot efficiency forced employee quotas to increase by 4x
- 14,000 "serious" injuries in 2019 -- double the industry standard
- Some warehouses reported 5x the industry standard injury rate
- Amazon claims higher rates reflect better injury reporting culture

Amazon's Approach to Robotics Is Seriously Injuring Warehouse Workers - OnLabor

Student Research

Automating Inequality (2018)

- Data-driven technologies target, track, and penalize poor and working-class individuals
- Automated eligibility systems, ranking algorithms, and predictive risk models create a "digital poorhouse"
- Technologies are not neutral -- shaped by societal fears and biases about poverty
- Documents "preemptive withdrawal": people avoiding services they're entitled to because algorithmic assessment is psychologically punishing

Harper's: The Digital Poorhouse

Student Research

Robodebt (2016-2019)

- Automated debt recovery using faulty "income averaging" that assumed earnings were evenly distributed year-round
- Wrongly accused 526,000+ welfare recipients; burden of proof shifted onto citizens
- Several people, saddled with life-altering debt, took their own lives
- Government deemed the scheme illegal; issued refunds totaling A$2.4 billion
- Royal Commission found systemic failures in oversight and proportionality

BBC: Australia's Robodebt Scheme

Student Research

Algorithmic Stigmatization

- Sociotechnical framework identifying 4 elements that create algorithmic stigmatization
- Real-world example: labeling people "at-risk" based on algorithmic analysis of online statements
- Systems designed to help can instead mark and marginalize individuals
- Highlights tension between predictive intervention and dignity

Conceptualizing Algorithmic Stigmatization - CHI 2023

Student Research

Bending the Automation Bias Curve (9-country, 9,000-person study)

- Trust in AI follows a nonlinear "Dunning-Kruger" pattern: moderate knowledge leads to over-reliance, experts remain skeptical
- People showed higher tolerance for AI errors ("still in testing") than for human errors at the same stage
- Individuals lacking confidence in their own ability were most likely to blindly follow AI recommendations
- Labeling a system "extensively tested" heavily increased reliance; deep training needed to move beyond "blind trust"

Bending the Automation Bias Curve - International Studies Quarterly (2024)

Student Research

Hallucinated Citations in Peer-Reviewed Literature

- Analysis of ~4,800 NeurIPS 2025 papers found 100+ fabricated citations across ~50 accepted papers
- "Vibe citing": plausible-sounding composites (blended authors, titles, venues) that passed peer review
- ~1% prevalence, but contaminates scientific records at scale
- Peer review assesses ideas and results, not citation veracity -- structurally vulnerable
- Systemic, not malicious: unintended harm from routine AI-assisted writing tools

Nature: Rein in the four horsemen of irreproducibility - Bishop, 2019

Wednesday:

Topic 3.2: Intentional Harms

Topic 3.3: The Alignment Problem, Value Learning, and Catastrophic Risks

See you Wednesday :)

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

All content was reviewed and approved by Dallas Elleman.