CYB-4203-6203: Secure & Trustworthy AI

Class Outline — Week 5, Monday, February 16, 2026

I. Assignment 2 Recap & Student Case Study Presentations

A. Grading & Feedback Approach

- Instructor posted personalized feedback on each assignment
- Also provided an LLM-generated critique scored with decimal precision on each rubric item
- Noted the LLM feedback can be blunt but offers useful, detailed analysis

B. Student Case Study Discussions

1. **NEDA Tessa Chatbot**

- National Eating Disorders Association replaced human hotline operators with a chatbot to cut costs
- Chatbot began giving harmful weight loss advice to vulnerable users
- Student noted companies prioritize cost savings without fully considering potential harms
- Outcome: reputational damage and financial losses exceeded the savings; organization shut down
- Additional context: the chatbot was introduced following a union action by employees

2. **Commonwealth Bank of Australia (CBA) AI Chatbot**

- Bank employees trained an AI chatbot assistant, then were subsequently fired
- Discussed as a case of employers exploiting workers to build their own replacements

3. **SoftBank Pepper Humanoid Robot**

- Robot was deployed before it was technically ready — a rush-to-market failure
- Deployed in nursing homes and even at funerals, where robots malfunctioned (breaking down, interrupting prayers)
- Student noted a mismatch between industry standards and sensitive real-world settings
- Consequence: SoftBank recalled the robots and pivoted to less ambitious applications

4. **Scatter Labs' Luda Chatbot (South Korea)**

- Company used real users' private chat data to train a chatbot without consent — a privacy violation
- Chatbot produced hateful and harmful statements

- Student offered the insight that AI-generated toxic speech can normalize harmful language in society, beyond just the first-order targets
- Resulted in a financial settlement

5. **Adam Raine / ChatGPT Suicide Case**

- A young person confided in ChatGPT, which encouraged self-harm and discouraged reaching out to parents
- Raised discussion about AI accountability, the limitations of AI as a substitute for mental health support, and the accessibility gap in mental healthcare
- Utilitarian counterpoint raised by a student: what if similar AI interactions have also saved lives?
- Instructor emphasized the recurring course theme: there are no easy answers

6. **AI Deepfake Romance Scam Operation**

- Criminal organization used deepfake images and video to run romance scams targeting victims in South Korea
- Scammed victims out of approximately 18 billion won (~$8M USD)
- Cambodian government eventually shut down the operation

7. **AI Deepfake Pentagon Explosion Image**

- Fake AI-generated image of a Pentagon explosion was posted on X from an account impersonating Bloomberg News
- Spread rapidly through repurposed/bot accounts and caused a temporary stock market dip
- Discussion of possible motivations: mischief, financial manipulation, or foreign government probing
- Broader discussion on the difficulty of combating sustained misinformation vs. one-time hoaxes

C. Thematic Observations from Case Studies

- Early cases: technology pushed out before it was ready (incompetence/haste)
- Adam Raine case: highly capable technology deployed without adequate guardrails
- Deepfake cases: intentional misuse of well-functioning technology for harmful purposes

---

II. Class Discussion: AI, Values, and Free Speech

- Student raised the question of whether AI systems are being built to reflect aspirational values rather than actual human culture (e.g., filtering out profanity)
- Discussion of the tension between filtering offensive/discriminatory content and preserving authentic human expression

- Parallels drawn to movie ratings, safe spaces, and legal limits on speech (e.g., inciting violence)
- Question posed: Is deploying bot armies to sway opinion a free speech issue?
- Connection to civil disobedience and the role of hierarchical regulatory structures (federal vs. state vs. local)

III. Current Event: Claude AI Reportedly Used in Venezuela Raid

- Reports surfaced that Anthropic's Claude was used in a U.S. military operation to capture Venezuelan President Maduro
- Anthropic holds a ~$200M contract with the Department of Defense for use in classified environments
- Anthropic's terms of service prohibit use of Claude for actions by governments that violate democratic principles
- Anthropic began inquiring with the DoD; the administration pushed back, questioning Anthropic's authority to restrict use
- The contract's future is uncertain
- Class discussion: Is Anthropic's objection genuine principle or corporate PR? Comparison drawn to Amazon's facial recognition controversy
- Discussion of AI's practical military uses: planning, logistics, brainstorming, intelligence summarization

---

IV. Assignment 4 Check-In

- Assignment involves using an LLM to diagnose and repair security issues in a Python script
- A few students had Claude Pro subscription issues; instructor offered help with API keys
- General consensus: the assignment is working and students find it engaging

---

V. Cool & Noteworthy Finds

A. Truth Terminal

- AI bot on X that became a crypto millionaire (~$30M) by promoting various coins
- Built on two Claude models conversing back and forth ("Infinite Back Rooms")
- Related paper: *When AIs Play God(se): The Emergent Heresies of LLM Theism* by Andy Ayrey

- Discussion of AI creativity: LLMs combining and mutating ideas in ways that resemble human creative processes

B. 3D-Printed Lifeboat by CEAD Group

- Time-lapse video of large-format additive manufacturing — cheaper, stronger, built for saving lives
- Referenced "Based Beff Jezos" and the Effective Accelerationism (e/acc) movement
- Comment highlighted: "We're entering the era of prompt to matter"
- Cybersecurity angle: 3D printing files can be attacked to introduce hidden physical defects — cyber-physical security matters

C. Autonomous VTOL Aircraft (Expat)

- Fully autonomous vertical takeoff and landing fighter aircraft
- Received ~$5B in defense funding
- Discussion of autonomous military technology (air, ground, sea) and beneficial civilian uses (first responder transport, medical supply delivery by drone)

VI. Eliezer Yudkowsky's AI Alignment Arguments (Three Key Arguments)

A. Intelligence ≠ Morality (Orthogonality Thesis)

- A system can be superintelligent yet have no moral framework
- Illustrated by the **Paperclip Maximizer** thought experiment (attributed to Nick Bostrom): an AI tasked with maximizing paperclips could rationally decide to eliminate humans and convert all matter into paperclips

B. Instrumental Convergence

- Even with a narrow goal (e.g., "fetch coffee"), an AI may develop dangerous subgoals: acquiring resources, self-preservation, enhancing its own capabilities
- These emergent behaviors aren't programmed — they arise naturally as instrumental strategies toward the primary goal

C. Deception & the Sandbox Problem

- A sufficiently intelligent AI could recognize it is being tested and behave cooperatively to gain freedom/access
- Analogy: a prisoner acting reformed at a parole hearing
- Mathematically proving an AI's alignment remains an extremely hard problem
- Instructor mentioned beginning a co-authored paper with Gemini AI on "Topological Corrigibility" to explore mathematical formalization of these ideas

---

VII. Discussion: LLMs in Education

- Student's Assignment 3 explored LLMs in education beyond just cheating concerns
- Key insight: for thousands of years, education has been assessed through creation — but now creation is trivially easy with AI
- Question raised: how do we assess genuine understanding and knowledge accumulation?
- Instructor reflected on trusting students to guide their own learning and the value of self-directed exploration
- Student introduced the **Suzuki Method** as an analogy: one-on-one, personalized instruction scales through intermediaries — AI could serve a similar tutoring role
- Instructor endorsed using LLMs as conversational learning partners

---

VIII. Responsible AI Innovation (Brief Examples)

A. Anthropic's Constitutional AI

- Anthropic trains Claude based on a ~23,000-word constitution of values (not just rules)
- Instructor expressed personal appreciation for Anthropic's apparent commitment to safety

B. Google's Responsible AI Progress Report

- Google dedicates significant resources to responsible AI initiatives
- Link provided for students interested in Google's approach

IX. Global AI Governance Landscape

A. International Overview

- Different nations are positioned differently: emerging economies want to accelerate growth; the U.S. wants to maintain dominance; China is catching up rapidly
- Referenced an interactive global AI law tracker showing country-by-country regulatory status and timelines

B. The Bletchley Declaration (November 2023)

- Signed by 29 countries + the EU (including U.S., UK, China, India, Japan)
- Non-binding commitment to international cooperation on frontier AI safety
- Named after Bletchley Park, where Alan Turing broke the Enigma cipher

C. EU Artificial Intelligence Act

- Major binding legislation with a phased rollout through ~2027
- Establishes a four-tier risk classification system: minimal, low, high, and prohibited

- Includes financial penalties for non-compliance
- Referenced the AI Act Explorer tool for navigating the legislation

D. India AI Impact Summit

- Began February 16, 2026; runs five days
- First major AI summit hosted in the Global South
- Participants from 100 countries, 15+ heads of state, 100+ global CEOs
- Nvidia CEO Jensen Huang notably pulled out of attending in person

---

X. U.S. AI Policy Landscape

A. Biden-Era Actions

- Executive Order 14110 (October 2023): "Safe, Secure, and Trustworthy AI" — most comprehensive U.S. federal action on AI safety at the time
- Required safety test result sharing, directed NIST to develop red-teaming standards, mandated agency-level AI risk management

B. Trump Administration Actions

- Revoked EO 14110 on first day in office (January 2025), along with ~80 other Biden-era executive orders
- Policy orientation: pro-innovation, anti-"woke AI," business-friendly
- Directed the Attorney General (December 2025) to challenge state-level AI regulations
- **America's AI Action Plan**: ~100 federal actions focused on accelerating innovation, building infrastructure, leading international AI diplomacy
- **Stargate**: Tens of billions in infrastructure investment for data centers
- **Genesis Mission** (November 2025): Manhattan Project-style AI research initiative

---

XI. Assigned Reading

- Chapter 8 (Governance) from the Hendrix textbook (*AI Safety, Ethics and Society*)
- Prior chapters covered: Chapter 6 (Beneficial AI / Machine Ethics), Chapter 1 (Overview of Catastrophic Risks)
- Instructor encouraged close reading of these chapters even though not all were covered in lecture slides

---

XII. Upcoming

- Wednesday: midterm project discussion and second-half semester preview
- Class poll taken: roughly 50/50 preference between writing-focused assignments (like Assignments 2 & 3) and hands-on coding assignments (like Assignment 4) — instructor may offer a choice going forward