CYB-4203/6203 Secure and Trustworthy AI

Week 3, Day 2 - Class Presentation and Discussion Outline

Date: February 4, 2026 (Wednesday)

Pre-Class (~before official start)

- Informal discussion about Claude vs ChatGPT interfaces
- Technical discussion about screen sharing setup for future project presentations
- Attendance check

I. Recap of Monday's Content

- **Fairness types covered**: distributive, procedural, interactional
- **Tensions**: personal vs. collective fairness
- **Challenges**: context dependence, mathematical impossibility theorems
- **Technical introduction**: core ML tasks, transparency, accountability

II. Today's Agenda (Remaining 4 AI Values)

1. Privacy
2. Autonomy
3. Safety
4. Sustainability

Plus: AI and human rights, human-AI collaboration, Assignment 3 preview, assigned reading

III. Privacy

New AI-Specific Privacy Threats:

- **Deepfake identity fraud** - AI can now convincingly impersonate people using voice, face, behavior patterns
- **Automated phishing** - Lower barrier for mass, personalized phishing campaigns; intelligence can be "multiplied infinitely"
- **Terms of Service concerns** - Most users don't read ToS; chatbots receive highly personal information (therapeutic conversations, personal questions)

Resources Discussed:

1. **Stanford study** - Exposes privacy risks of AI chatbot conversations
2. **NCSL database** - National database of state AI legislation
    - 2025: **1,262 different AI-related laws** proposed/passed across US states
    - Topics: data centers, energy bills, community protection
    - Key point: No good federal framework; states "going it alone"
3. **Machine Unlearning** - GitHub repository with papers on removing private data from trained models without full retraining

IV. Autonomy

Opening: Rush lyrics from "Freewill" - "If you choose not to decide, you still have made a choice"

Framework (Carina Frunkl's definition):

A. Internal Autonomy (Authenticity)

- Are my beliefs truly mine?
- Adaptive preference formation from limited options

Student Discussion - "Decisions you'd make differently with more information":

- Job interview answers
- Undergraduate major choice
- Relationships / time spent with people
- Big purchases (timing, reviews, sales)

B. External Autonomy (Agency)

Three dimensions:

1. **Competency** - Capacity for good decision-making
2. **Freedom** - Legal freedom to act
3. **Opportunity** - Cultural/social enablers and constraints

Class Discussion - How is AI impacting autonomy?

Internal (beliefs/knowledge):

- **Sycophancy** (student contribution) - ChatGPT tendency to agree with users, creating confirmation feedback loops

- **AI-generated content on social media** - "AI swarms" where bots argue with each other convincingly
- **Filter bubbles/echo chambers** - Algorithm-driven engagement optimization

External (competency/skills) - Negative impacts:

- Depends on usage - can be tool for efficiency or crutch for laziness
- **Over-reliance** (student) - People citing ChatGPT as authoritative when it's confidently wrong
- Critical thinking burden shifts from "finding information" to "validating AI output"
- Wikipedia comparison: same trust issues, but AI "sounds like an expert every time"

External - Positive impacts:

- **Research acceleration** (student) - Faster than digging through Google Scholar
- **Writing assistance** (student) - Getting expansion suggestions, then fleshing out in own words
- Frees humans for "big picture" thinking; less nitty-gritty detail work
- Conversational brainstorming partner when humans unavailable

Instructor reflection: Pattern emerging of "independently working with your bot swarm" - danger is we stop talking to each other

Resource: Springer Nature paper - "Human Autonomy at Risk: An Analysis of the Challenges from AI"

V. Safety

Resources Discussed:

1. **International AI Safety Report** (Yoshua Bengio, first author)
    - ~228 pages, free download
    - Covers: AI capability improvements, domain impacts, safety concerns

Student Discussion - Scaling Laws:

- Like Moore's Law for AI/LLMs
- Performance improves predictably with: more data, more compute
- "No theoretical limit" - if scaling continues, performance keeps improving
- **Investment implications:** Holy grail = AGI; once achieved, "sky's the limit"
- **AI bubble concern:** Unless arbitrary theoretical wall is hit

2. **Future of Life Institute Winter 2025 Safety Index**
    - Company rankings: Anthropic (C+, top), OpenAI, Google DeepMind

- DeepSeek (D)
- Discussion of Z.AI (Chinese company, makes GLM models)
- Student clarified: "ZAI is out of China and they make the GLM models"

VI. Sustainability

Three resources presented: (pro, balanced, con perspectives)

- Left: IBM promotional video (AI solving climate problems)
- Middle: Balanced analysis of both sides
- Right: Medium article (AI will "wreck the planet")

Class Discussion - Environmental Challenges:

- Data center land footprint ("visible from space")
- Energy and water use
- Noise pollution
- Rent/housing price increases near data centers
- **Oklahoma legislation** - Proposed law requiring data centers to pay upfront for grid impact
- **Zoning changes** - Agricultural land rezoned to commercial increases area taxes

Quantitative Perspective (instructor research):

- GenAI data centers: **7-8x more energy intensive** than traditional workloads
- GPT-3 training: ~equivalent to 120 median US homes for a year
- **2026 projection:** Global AI = ~1,000 terawatt hours (5th largest "country" between Japan and Russia)
- **2030 projection:**
    - CO2: equivalent to 7.5 million cars (<1% of Earth's 1.5 billion cars)
    - Water: ~1 billion cubic meters (<0.5% of accessible freshwater)

Instructor's take: Numbers are huge in absolute terms but proportionally "doesn't seem like that big of a deal"

Additional concerns: Mining/manufacturing impacts (cobalt, lithium, copper)

- Counter: AI-enhanced exploration might enable less destructive mining

Key message: Not black or white; "walk the tightrope" between AI optimism and rejection

VII. AI and Human Rights

A. AI Bias and Representation

1. **Amazon hiring discrimination** - Tool unfairly favored males; scrapped
2. **Google Vision racism** - Labeled African Americans as animals; retracted quickly

B. Creative Labor and Displacement

1. **SAG-AFTRA strike** - Hollywood actors/writers; ongoing implications
2. **Anderson v. Stability AI** - Class action against generative AI companies; artists won
3. **Barts v. Anthropic** - Authors sued over books used for training; judgment against Anthropic; had to compensate authors
    - Instructor: Decision seemed "balanced"; implications for entire LLM industry

C. Political and Cultural Impacts

1. **Recorded Future** - Political deepfakes: targets, objectives, emerging tactics
2. **AI slop journalism** - Information environment pollution; impediment to finding useful information

D. Invisible Labor and Exploitation

1. **OpenAI Kenyan workers** - $2/hour to filter traumatic content (CSAM, violence)
2. **"Empire of AI" by Karen Hao** - Book recommendation
    - Insider view of Facebook's early days and moral decline
    - Inside look at OpenAI and Sam Altman
    - Available on Spotify audiobook; "really, really good book"

E. Language and Cultural Preservation

- Most models trained on English; gap growing for non-English speakers
- **Positive:** AI helping preserve endangered languages (Cherokee, Osage in Oklahoma)

VIII. Human-AI Collaboration

Resources:

1. **Ethan Mollick's "One Useful Thing"** - "Centaurs and Cyborgs on the Jagged Frontier"

    - AI strengths vs. weaknesses; human complementarity
    - Transhumanism themes; good infographics

2. **Real-world collaboration examples** (website)

    - Healthcare, data-driven decision making, idea generation
    - Instructor enthusiasm about team brainstorming; AI can fill gap when teams unavailable

3. **Nature AI article** - "Human-AI Teaming in Healthcare"

    - AI provably helps; some strategies more effective than others

Potential future assignment: Creative brainstorming with AI (instructor noted for consideration)

IX. Assignment 3 Preview

Format: Similar to Assignment 2 (case study)

- Choose positive OR negative AI impact
- Related to core values discussed OR legal/environmental/cultural perspectives
- Executive summary + analysis from multiple perspectives
- Will be released after Assignment 2 deadline (tomorrow, 12:29)

Student question: Due date?

- **Answer:** Moving to Wednesday-to-Wednesday rhythm (7 days)
- This assignment: Released Thursday → Due following Wednesday

X. Next Week Preview (Week 3 in Syllabus)

Topics:

1. Unintended harms (brief recap)
2. **Intentional misuse** - Cyberattacks, weaponized AI
3. **The Alignment Problem**
    - "Baby tiger" analogy: Controllable when small, dangerous when grown
    - Can AI lie? Have hidden motivations? Become misaligned enough to cause massive harm?
    - **Eliezer Yudkowsky** - "Chief AI doomer"; book: "If Anyone Builds It, Everyone Dies"
4. Responsible AI innovation practices

Assigned Reading: Hendrycks, Chapter 1 - "Overview of Catastrophic AI Risks"

- AI races between governments and companies
- Various catastrophic risk scenarios

XI. Post-Class Interactions

1. **Student question about Assignment 2 format** - Should combine responses into solid document rather than Q&A format for easier grading

2. **Student discussion** - Planning email about Claude Code tokens; possible 10-15 minute pre-class discussion Monday

3. **Brief discussion** - Oklahoma/federal AI legislation hearings; student mentioned watching at 3.5x speed

Key Themes and Takeaways

1. **Balanced perspective encouraged** - Neither pure AI optimism nor rejection; reality is nuanced
2. **Student engagement** - Multiple student contributions throughout on sycophancy, scaling laws, over-reliance, research benefits
3. **Resource-rich** - Numerous vetted sources provided for upcoming assignment
4. **Practical concerns** - ToS, state legislation fragmentation, environmental proportionality
5. **Transition point** - Moving from "values" to "risks and alignment" next week