CYB-4203/6203 — Session 5 Outline (Feb 18, 2026)

I. Pre-Class Chatter

- Students discussing difficulty of the companion "Fundamentals of AI" course (taught by Lee)
- Mismatch between expected prerequisites — cyber-track students lacking data structures background that CS-track students have
- Dallas notes the AI minor and these courses are brand new/experimental; prerequisites still being figured out
- Brief check-ins on other courses (Operating Systems — exam that day; war stories about old OS finals)

II. Housekeeping

- Assignment 4 deadline extended to Friday due to Gemini difficulties and scheduling conflicts
- ~12 students already submitted; option to rework and resubmit
- Today's agenda: Sections 4.3 and 4.4 (playing catch-up from Week 4)

III. What Is Governance?

- **Etymology lesson:** Greek *kybernan* (to steer/pilot a ship) → roots shared by "government" and "cyber"
- **Norbert Wiener & Cybernetics** (mid-1900s): scientific study of control and communication in animals and machines
    - Origin story: the anti-aircraft gun aiming problem
    - Cybernetics branched into control theory, signals/systems, and eventually machine learning
    - "Cybersecurity" literally = security of control systems
- **Broad definition:** Governance = rules and processes that coordinate behavior in *any* system (biological, mechanical, societal)
    - Not just government — includes norms, policies, institutions
    - Classroom example: unwritten cultural norms governing who speaks when
- **Examples across domains:**
    - Healthcare: patient care norms, ethical standards, licensing
    - Financial services: banking laws, capital requirements, FDIC/SEC
    - Aviation: safety culture, checklists, scheduled maintenance
- Brief discussion of governance philosophies (libertarian vs. top-down)

IV. Governance Across the AI/ML Pipeline (5 stages)

A. Data Pipeline — Datasheets for Datasets

- Gebru et al. (2018) paper introducing standardized documentation for datasets
- Analogy to electronics component datasheets (spec sheets from Mouser/Digi-Key)
- Background on Timnit Gebru: Google whistleblower, founded Algorithmic Justice League
- **Live demo:** OpenML (~6,000 datasets with metadata) and Hugging Face (hundreds of thousands of datasets)
- Now a de facto global norm — not legally required, but universally adopted

B. Model Development — Model Cards

- Mitchell, Gebru, et al. — short standardized docs for each model
- Cover: intended use, performance benchmarks (especially across demographic groups), limitations
- **Live demo:** Anthropic's system card for Claude Opus 4.6 (Feb 2026) — now ~200+ pages
    - Discussion: what's the difference between a *model* and a *system*? (Model = base neural net; system = model + tools, memory, web search, etc.)
- Toured Google's model catalog (Gemma, generative models), OpenAI's developer model list, Hugging Face (2M+ models)
- Practical framing: if you're building a chatbot for a bank, this is where you'd do due diligence

C. System Design — International Standards

- **ISO/IEC 42001:** Certifiable AI management system standard
    - Comparison to other ISO certs (27001, etc.)
    - **Live demo:** OpenAI's Trust Center (trust.openai.com) — showing compliance certifications
    - Discussion of why companies pursue certification (market access, legal CYA, government contracts, trust signaling)
    - Fun tangent: searching for McDonald's trust center
- **IEEE 7000-2021:** Ethical system design process standard — blueprint for building a culture of safety within an org

D. Organizational Governance — NIST AI RMF

- Context: Came out of Biden's EO 14110 (2023), survived the rescission under Trump
- Voluntary framework, but widely adopted across industry
- Four core functions: **Govern, Map, Measure, Manage**

- Free to download (unlike ISO standards) — 33-page executive version + companion playbook
- **OECD AI Principles:** 5 principles adopted by 46 countries, closest thing to a global baseline

E. Deployment & Operations — AI Bill of Materials

- Background: Software Bill of Materials (SBOM) concept — knowing your dependencies for CVE response
- Extending SBOM to AI: tracking model provenance, training data, fine-tuning history
- **Live demo:** OWASP AI BOM Project
- NIST guidance on AI BOM and supply chain risk

V. Nuclear Weapons / AI Analogy (Student-Driven Discussion)

- Student observation: "AI feels like another arms race"
- Parallels discussed:
  - Both driven by national competition (US/Soviet → US/China)
  - Innovation pushed at the expense of safety
- **Key differences:**
  - AI is digital, ubiquitous, open-source — no physical enrichment bottleneck
  - 2M+ models on Hugging Face; no realistic "AI non-proliferation treaty"
  - AI can develop/improve *itself* — nuclear weapons can't
- Discussed GPU export controls as analog to restricting fissile material
- Mentioned Eliezer Yudkowsky's extreme proposals (melting GPUs, airstrikes on data centers)

VI. Upcoming Assignments & Midterm Discussion

- Assigned reading: Hendrycks, Chapter 8 (prior chapters: 6 and 1)
- Assignment 5 brainstorming — class split on preferences:
  - **Option A:** Code project — Python wrapping an AI to automate something personal/useful (could become iterative midterm project over 4 weeks)
  - **Option B:** In-class debate on divisive AI ethics topics (forced to argue sides you may disagree with); graded on effort
- Midterm project ~4 weeks out

VII. Outdoor Session — Informal Discussion

- **Gemini frustrations:** Multiple students reporting issues — model going off the rails, apologizing to itself, ignoring instructions

- **Hackathon initiative:** Dallas working with TU Fab Lab to host a Hurricane Hackathon (like Hacklahoma at OU); targeting ~8 weeks out; involving engineering + business colleges; 24-hour design sprint format
- **Humanoid robots:** Teleoperated robots shipping soon (~car-priced); training data collection via human operators; "teleporting labor"
- **AI & labor displacement:**
    - White-collar jobs hit first (creative, knowledge work)
    - Blue-collar/trades further out but coming
    - Tension: AI increases stock value for owners, decreases labor value for workers
    - "Who's steering the ship?" — the people with money and influence
- **Lighter moments:** Movie rec (*Good Luck, Have Fun, Don't Die*), Steel Beans one-man band, theater going AI-generated, students joking about eating the rich / "cybersecurity freedom fighters"