CYB-4203-6203: Secure & Trustworthy AI — Class Outline

Wednesday, February 25, 2026

I. Pre-Class Discussion: Current Events in AI Policy

- **U.S. Defense Department and Anthropic** — Discussion of reports that the Trump administration may invoke the Defense Production Act to force Anthropic to allow military use of its AI models, overriding Anthropic's terms of service
- **Risks of AI in military applications** — Student raised concerns about automation bias, misidentification, and the real-world cost of AI errors in lethal contexts (e.g., Israel's use of AI to identify Hamas targets)
- **Impact on industry and innovation** — Class discussed potential reputational harm to companies compelled into military use, the chilling effect on the broader AI industry, and tension with the administration's stated goal of promoting AI innovation
- **Anthropic's identity and values** — Instructor noted Anthropic's founding story: key engineers left OpenAI over safety concerns to build a safety-first competitor
- **Geopolitical dimensions** — Student raised geopolitical complexity (Venezuela, China/Taiwan 2027 timeline, Panama Canal) as context for why the executive branch may feel urgency
- **Legitimacy of the Defense Production Act** — Class debated whether the current threat level justifies invoking wartime powers; suggestion that a negotiated defense contract might be more ethical than compulsion
- **Anthropic's specific objections** — Student noted Anthropic's opposition centers on mass surveillance of Americans and certain weapons-related use cases

II. Housekeeping & Announcements

- Updated course schedule: more time allocated for midterm projects (two additional weeks after spring break)
- Guest instructors for next week: Dr. Weiping Pei (Monday) and Dr. Yi Ting Chua (Wednesday)
- Midterm exam date: Wednesday, March 11 — comprehensive through Unit 6
- Exam format: short-answer/definition style (not multiple choice), with a personalized section based on each student's prior assignment topics; study guide forthcoming

- Course website updates: renamed sections/units for clarity, updated weekly materials, expanded resources page (threat intelligence, frameworks, case studies, academic papers, books)

III. Review of Monday's Session & Emergence

- **Recap of 3Blue1Brown Transformer videos** — Acknowledged the density of the material; Transformers as foundational to secure/trustworthy AI development
- **"Are LLMs just glorified autocomplete?"** — Revisited a student's question from Monday; mechanically yes, but this is reductionist — neurons are just electrochemical switches, computers just flip bits; what matters is **emergence** at scale
- **Emergence defined** — Complex behaviors arising from simple components interacting; examples: snowflakes, ant colonies, neural activity in the brain
- **Swarm behavior and BOIDs** — Craig Reynolds' 1986 model: three simple rules (cohesion, separation, alignment) produce realistic flocking behavior; demonstrated via video; presented as a form of artificial intelligence distinct from LLMs, and a valid direction for Assignment 5

IV. Reinforcement Learning, Reflection, and "Shower Thoughts"

- **Reinforcement learning basics** — Student described positive/negative feedback loops; instructor drew parallel to human learning (e.g., touching a hot stove)
- **"Shower thoughts" and AI reflection** — Followed up on a student's question from Monday about AI that reflects on its own reasoning; connected to concepts of default mode network, incubation, and hippocampal replay in neuroscience
- **Google DeepMind's Atari research** — Deep Q-Networks combining deep neural networks with reinforcement learning; Breakout example where the agent discovered a tunneling strategy the developers hadn't anticipated — AI teaching humans
- **AlphaGo documentary** — Recommended as an Assignment 5 direction; millions of self-play games with no human input led to superhuman Go performance; noted this event spurred China's national AI push

V. Continual Learning & Catastrophic Forgetting

- Neural networks, once trained, are essentially frozen; further training on new data can overwrite previously learned knowledge (catastrophic forgetting)
- Contrasted with human learning — learning to tie shoes doesn't erase the ability to read
- Continual learning as an active research frontier: retaining existing knowledge while acquiring new capabilities
- Referenced a well-cited paper and a 45-minute YouTube overview
- **Class discussion** — Students drew analogies to human experience: the "backwards brain bicycle" experiment, muscle memory decay, age-related knowledge loss

- **Model rot / model drift** — Instructor introduced the concept: a model's performance degrades over time as real-world conditions shift away from its training data, even if the model itself is unchanged; requires periodic retraining

VI. Connections Between Biology and Artificial Intelligence

- **Core theme** — Intelligence is orthogonal to substrate; biological and artificial intelligence share mechanisms, and insights flow both directions between neuroscience/psychology and AI research
- **The Perceptron (1957)** — Modeled on biological neurons (dendrites → inputs, axon → output); foundational to all neural network architectures
- **Convolutional neural networks and cat vision** — Hubel & Wiesel's Nobel Prize-winning 1950s experiments discovering edge detection in cat visual cortex; directly informed modern computer vision
- **Deep neural networks as brain models** — DNNs now make better predictions about brain function than traditional neuroscience models
- **In Silico neuroscience research** — Simulating brain functionality with computers as a current research frontier
- **"A Brief History of Intelligence" (book recommendation)** — Evolutionary survey of intelligence breakthroughs: chemical detection in bacteria, dopamine-driven reinforcement learning in early vertebrates, mental simulation in mammals, theory of mind in primates, symbolic language in humans; available as audiobook on Spotify; Medium article summary also linked
- **Why this matters for security** — Understanding AI's biological roots and analogs informs how we think about vulnerabilities and defenses; e.g., prompt injection as analogous to social engineering/lying to a person; immune systems and fight-or-flight as models for AI defense mechanisms

VII. AI/ML Systems vs. Traditional Software

- **Key contrasts** — Traditional software is deterministic, rule-based, written line-by-line, and provably correct; AI/ML systems are probabilistic, trained/grown, and carry inherent statistical error
- **Development processes differ** — Software: requirements → design → code → test; AI/ML: data collection → training → evaluation → fine-tuning → monitoring
- **Complementary strengths** — Software excels at well-defined logic; AI/ML excels at tasks too complex for explicit rules; secure system design requires knowing where to use each
- **AI/ML system lifecycle (4 stages)**: data collection & preparation, model training & evaluation, deployment & integration, monitoring & maintenance

VIII. Testing, Verification & Assignment 5 Preview

- Traditional software testing vs. AI/ML evaluation challenges
- **Assignment 5** — Choose a topic from the day's presentation (or another biology–AI connection); write an executive report covering components, architecture, the AI/ML pipeline as applicable, and security implications/attack surfaces; will also feed into the midterm project

IX. Model Inversion & Anthropic's Distillation Report

- **Model inversion** — Student question; defined as reconstructing input data or extracting sensitive information from a trained model by exploiting its parameters and outputs
- **Anthropic's February 23 report on model distillation attacks** — Three Chinese AI labs (DeepSeek, Moonshot/Kimi K2, MiniMax) ran industrial-scale IP extraction campaigns against Claude using ~24,000 fraudulent accounts and over 16 million exchanges; used proxy access to circumvent regional restrictions
- **Security implications of distillation** — Capabilities transfer, but safety guardrails and alignment work do not; distilled models inherit power without the responsible safeguards
- **Connection to opening discussion** — Student tied this back to the Trump administration's innovation push and the competitive geopolitical landscape; regulation vs. competitiveness tension

X. The Moloch Problem & Competitive AI Dynamics

- **Moloch concept** — Named for an ancient deity; describes game-theoretic pressures that drive collectively harmful competitive behavior (originated with Daniel Schmachtenberger)
- **Prisoner's dilemma analogy** — Each AI company's individually optimal strategy (race ahead) produces the globally worst outcome (unsafe AI); cooperation would be better for all, but trust and communication barriers prevent it
- **Relevance** — Framed as a potential Assignment 5 topic: game-theoretic dynamics in competitive AI development

XI. Closing Q&A

- Student question connecting catastrophic forgetting to human memory and muscle memory decay — instructor affirmed the analogy and connected it to model rot
- Discussion of photographic memory as a possible analog to continual learning without forgetting, and potential psychological trade-offs
- Reminder about API access, upcoming study guide, and exam logistics