Class Discussion Outline

March 30, 2026 - Week 10, Day 1

I. Opening / Preliminary Items

- Brief look at "Rent a Human" — a platform where humans are hired to perform tasks for AI agents, including personal services and companionship
- Announcements: final project options discussed (red teaming, risk analysis, mitigation); check on who watched the pre-recorded instructional video

---

II. LLM-Assisted Development Cycle (Video Walkthrough)

The instructor played a ~30-minute pre-recorded demo walking through a structured workflow for using LLMs in development projects.

A. Overview of the Workflow

A five-phase iterative cycle:

1. **Review** — Load context into the LLM (files, assignments, project requirements)
2. **Brainstorm** — Explore creative ideas; prompt the LLM to ask clarifying questions
3. **Research** — Investigate technical implementation options and validate the tech stack
4. **Plan** — Create a detailed, step-by-step checklist with acceptance criteria
5. **Execute** — Build the artifact; then loop back to Review for the next iteration

B. Demo Project: Concept Drift Interactive App

- **Topic chosen:** Model robustness and concept drift (an attack vector not selected by any student)
- **Tech stack selected:** Streamlit (a Python library for building interactive ML demos); validated via external search before proceeding
- **App functionality:**
    - Trains a simple logistic regression classifier on synthetic 2D data
    - Users manipulate sliders to simulate natural drift and adversarial drift
    - Visualizes how decision boundary and model accuracy degrade in real time
    - Includes a retrain button and reset button

C. Key Lessons Demonstrated

- The importance of having the LLM ask clarifying questions during brainstorming
- Saving phase outputs as markdown files (review.md, brainstorm.md, research.md, plan.md, execute.md) for documentation and report-writing later
- Setting acceptance criteria before building — then verifying each criterion at the end
- Staying actively engaged: an adversarial drift slider appeared in the plan without being explicitly discussed, illustrating that you cannot simply let the LLM run on autopilot
- Debugging: encountered code errors during execution; LLM explained its changes, though the instructor noted LLM explanations of its own behavior should not always be taken at face value
- After completing one cycle, the loop restarts at Review — noting what worked, what needs improvement, and what's missing

---

III. Practical Notes on LLM Usage / Token Limits

- Some students have been hitting usage limits more frequently, particularly with heavy file processing or code generation
- Discussion of Claude subscription tiers and token costs
- Mention of recent context window expansion (200K → 1M tokens) for Claude's larger models
- Instructor framed LLM tool investment in terms of career development and portfolio building

---

IV. Course Pacing Update

- Only ~6 of the planned units covered through week 10
- Midterm project due the following Wednesday
- Plan: cover Units 7, 8, and 9 over the next few weeks at normal pace, then accelerate through the final units with lighter homework loads
- Today's focus: Units 7.1 and 7.2 — Privacy Risks and Bias

V. Privacy Risks in AI/ML

A. Overview

- AI systems consume large amounts of personal and sensitive data
- Traditional privacy frameworks (terms of service, consent agreements) are poorly suited to the AI era
- Trade-off between model performance and privacy protection is a central challenge

B. Membership Inference

- Can an attacker determine whether a specific individual's data was used in training?
- Can personal information be recovered through normal model queries?
- Mitigation approach discussed: synthetic data generation — modifying the dataset so the model trains on synthesized rather than real personal records
- Limitation: partially synthetic data still allowed ~80% of original individuals to be re-identified; full synthesis improves privacy but degrades model accuracy

C. Sensitive Data Domains

- Genomic studies, mental health records, criminal databases
- High-value use cases (e.g., identifying disease biomarkers) exist alongside serious privacy risks (e.g., health insurance discrimination)

D. Machine Unlearning

- The challenge of removing specific data points or associations from a trained neural network
- Targeted removal of private information from model weights is technically very difficult and can degrade overall model performance

E. Model Inversion

- Referenced a foundational 2015 paper on facial recognition models
- Demonstrated that training data (e.g., people's faces) can be reconstructed from model weights even when no image files are stored
- Student shared a real-world example: a woman in Tennessee was arrested for crimes in North Dakota based on AI-assisted facial recognition; she spent approximately five months in jail before alibi evidence exonerated her
- Discussion of accountability, legal consequences, and precedent-setting in wrongful AI-assisted identification cases

F. Surveillance and Predictive Harm

- AI dramatically lowers the cost of large-scale surveillance (facial recognition, movement tracking)
- Predictive policing: targeting individuals based on predicted behavior, not actual actions
- Social scoring systems; location tracking via mobile data
- Secondary harm: chilling effect on behavior — people act differently when they believe they are being watched, raising questions about genuine autonomy

G. Self-Fulfilling Predictions

- Example: predictive healthcare models determining whether to resuscitate patients
- If a model predicts poor outcomes and treatment is withheld, the negative outcome is recorded and reinforces the model — a feedback loop
- Applies also to predictive policing, credit scoring, and health insurance risk models
- Key framing: "Privacy harm is not just data exposure — it's the use of data to constrain futures"

H. Mitigation Approaches

- **Differential Privacy:** Adding calibrated mathematical noise to datasets so no individual record significantly influences model outputs
- **Federated Learning:** Training on decentralized local data rather than one central dataset; models are merged afterward
- **Homomorphic Encryption:** Performing computations on encrypted data and decrypting only the final result; very secure but computationally expensive

I. Regulatory Frameworks

- **GDPR (EU):** Right to explanation, right to data erasure, purpose limitation; AI is straining existing interpretations
- **HIPAA:** Health data privacy standards; tension between compliance and beneficial use of health data for automation and accessibility

---

VI. Bias in AI/ML (Introduction — continued next session)

A. Types of Bias

- Overview of bias categories: reporting bias, historical bias, automation bias, selection bias, sampling bias, group attribution bias, and others (via Google Developers reference)

- Distinction between statistical/mathematical bias and human bias embedded in data collection and labeling

B. Not All Bias is Harmful

- Example: a spam filter is intentionally biased — that is its function
- Concern arises when bias produces unfair or discriminatory outcomes for protected groups

C. Bias Throughout the AI/ML Lifecycle

- Bias can enter at any stage: data collection, labeling, model training, evaluation, deployment

D. Facial Recognition and Intersectionality

- Referenced the MIT Media Lab "Gender Shades" paper (2018)
- Found dramatically worse model performance on darker-skinned women compared to lighter-skinned men — in some cases 30–40x worse accuracy
- Similar bias observed in generative image models: prompts like "show me a doctor" or "show me a criminal" produced racially and gender-skewed outputs in early systems

Note: Fairness evaluation frameworks and remaining bias topics to be covered next session.

VII. Closing Discussion

- Open reflection on the long-term trajectory of privacy: will future generations share current concerns about surveillance and data collection, or will norms shift entirely?
- Historical note: marketing companies were purchasing Wi-Fi and Bluetooth signal logs from retail chains ~10 years ago to build movement and relationship models — AI now makes such analysis far cheaper and more powerful
- Brief tangential discussion about data collection obligations for platforms like Craigslist
- Instructor note: midterm project due next Wednesday; class continues Wednesday with fairness frameworks and remaining bias content