# CYB-4203-6203: Secure & Trustworthy AI

## Session Outline — March 9, 2026

The University of Tulsa

---

## I. Welcome Back & Session Roadmap

- Instructor returns from Spain; brief check-in with students
- Agenda overview:
    1. Recap of time in Spain (conference, business plan)
    2. Review of last week's material (instructor was absent)
    3. Upcoming midterm exam logistics
    4. Natural human learning, reward hacking, and the purpose of a university education

---

## II. Instructor's Trip to Spain

### A. ICISSP Conference

- Presented at the 13th International Conference on Information System Security and Privacy (ICISSP)
- Prepared the conference presentation in approximately 24 hours
- Co-located conferences included topics on machine learning and robotics
- Advanced PhD dissertation research on the side

### B. i2E Entrepreneur's Cup & the "Evolver" App

- Business plan competition in Oklahoma; $30,000 first-place prize
- Instructor made finals the prior year; resubmitted as a solo team ("just me and Claude")
- **Evolver** — a college life assistant concept:
    - AI-powered personal assistant for university students
    - Manages deadlines, goals, time, campus opportunities, career networking
    - Motivated by the gap between what a human executive assistant could do (~$4,000/month) and what AI now makes affordable

- **Technical vs. business challenge**: building the system is a CS/AI problem; making it viable and profitable is an entirely different problem
- Business plan required financial forecasting — highlighted LLMs' persistent struggles with math
    - LLMs predict text tokens, not deterministic numerical outputs
    - Giving LLMs calculator/code execution tools helps, but they still make rounding and logic errors

---

# III. Natural Human Learning as a Lens for AI Concepts

## A. Reinforcement Learning Refresher

- Agents in an environment make choices, receive reward signals, and learn action policies over millions of iterations
- RLHF (Reinforcement Learning with Human Feedback) aligns LLMs by rating responses as good or bad
- Key distinction: LLMs are trained on text tokens, not on *strings of human actions and their outcomes*

## B. Human Reward Signals & Evolver's Vision

- Humans optimize over an uncountable number of simultaneous reward signals: physiological, financial, social, emotional (fulfillment, etc.)
- Evolver concept: track human actions + self-reported satisfaction as a personal "reward signal" over time
- Could enable retrospective pattern analysis of personal behavior and decision-making

## C. The Alignment Problem

- **What is alignment?** A system sharing intentions and goals with its users
- **Why is it hard?** Models are black boxes — we can only judge goals by observing outputs, and must wait for the model to act
- Comparing stated goals vs. actual actions over time to build a "reputation" or trust model — works for both humans and AI
- **Anthropic's mechanistic interpretability work**: "Mapping the Mind of a Large Language Model" (2024)
    - Identifying how concepts (e.g., Golden Gate Bridge) are represented as activated features within neural networks
    - Analogous to neuroscience: electrode monitoring of brain activity

### D. PhD Research Direction: Alignment Learning Models

- A new class of models trained on **strings of action tokens** rather than text tokens
- Actions include: model outputs, stated goals, actual goal attainment, and side effects
- Training models to recognize whether another model is truly aligned
- Related concepts: goal-conditioned reinforcement learning
- **Proprietary "5A" ontology** for formalizing human life:
    1. **Actions** — what you do
    2. **Assets** — what you have (bank accounts, certifications, etc.)
    3. **Aims** — your goals
    4. **Agents/Actors** — people and entities in your network
    5. **Ambits** — domains/areas of life (health, career, family, finances, etc.)
- Ambits as multi-objective optimization: finding actions that advance multiple life domains simultaneously

---

## IV. The AI/ML Lifecycle as a Metaphor for University Education

### A. Mapping the ML Lifecycle to Student Experience

- **Stage 1 — Data Collection & Preparation** → Coursework, lectures, reading, absorbing information
- **Stage 2 — Model Training & Evaluation** → Exams, assessments, testing knowledge
- **Stage 3 — Deployment & Integration** → Post-graduation, entering the workforce (ages ~22–27)

### B. Exams as Evaluation Metrics

- Exams serve multiple stakeholders: the student, the university's reputation, future employers, and society
- Students often pass exams without retaining knowledge long-term — optimizing for the test rather than for learning

### C. Reward Hacking in Education

- **Reward hacking** defined: maximizing the score/metric without achieving the intended underlying goal
- Student example: a 9-year-old's Fortnite macro that thanks the bus driver ~32,000 times to passively gain XP — technically "winning" without playing
- University parallel: cramming to pass an exam rather than deeply learning the material
- Instructor's stated goal: genuine learning, inspiration, and self-directed pursuit — not just passing

### D. Three ML Paradigms Mapped to University Life

- **Supervised learning** → Classroom instruction (labeled data from instructor)
- **Unsupervised learning** → Independent exploration, research, self-directed study at home
- **Transfer learning** (raised by student) → Social interactions, learning from peers
- **Reinforcement learning** → Trial and error in real life, learning from repeated mistakes

### E. Attack Surface of the Student Mind

- Low-quality or biased information from instructors or sources
- Social media, video games, and other distractions
- Competing interests pulling focus away from degree completion (anecdote: taking too many elective courses)
- Inherent vulnerabilities vs. adversarial threats — both matter
- Instructor's lax AI-use policy as a double-edged sword: students learn to use tools, but risk offloading learning entirely

### F. The Real World as Deployment

- Many people say they never truly learned until they entered the workforce
- University is a preparation/training stage; the "real world" is deployment
- Students encouraged to treat the course as carving out focused time for a critically important field, not just chasing grades

---

## V. Review of Last Week's Material (Instructor-Led Recap with Students)

### A. Inference

- Class discussion on the meaning of inference: drawing conclusions from results of training
- **Student contribution**: Chinese researchers used brain MRI data + ML to reconstruct images a person was looking at — a vivid example of inference
- Discussion of how the training data set was constructed: pairs of (brain activity scans, images being viewed)
- Brief tangent on model distillation vs. inference processes

### B. Adversarial Concepts — Not Always Negative

- "Adversarial" simply means competitive; can be beneficial:

    - Adversarial training (improving model robustness)
    - Generative Adversarial Networks (GANs) for creating realistic synthetic data
    - Adversarial games in reinforcement learning (learning optimal strategy)

## C. Stage 1 — Data Collection & Preparation: Threats

- **Data poisoning**: very small changes to training data (fractions of a percent) can cause ~5% shifts in model output
    - Defensive use: artists using data poisoning tools (e.g., Nightshade, Glaze) to protect their work from being scraped
    - Student proposed defense: re-save images through a clean pipeline, compare against originals, discard anomalous data
- **Arms race dynamics**: as AI defenses improve, attacks evolve — a "singularity" concern where humans can no longer keep pace
- **AI-generated content detection**: metadata tagging, platform verification — all currently imperfect
- **Netflix Prize de-anonymization**: cross-referencing anonymous Netflix ratings with public IMDb reviews exposed users' personal information, political affiliations, and viewing habits
- **PII and proprietary data exposure** risks in training data

## D. Identity Management & Bot Detection (Extended Class Discussion)

- Platforms exploring verified human status (real ID, phone numbers, 2FA)
    - Fortnite's competitive mode requiring phone registration; players petitioning for government ID
    - Discord's data leak history making ID requirements risky
    - Fundamental tension: platform trust requires users to trust the platform with their data
- No form of identity credential is truly theft-proof (SSNs, phone numbers, birth certificates, government IDs)
- **Student proposal**: instead of verifying identity, match users by behavior quality (ignore whether they're human or bot)
    - Counterpoint: resource waste from millions of bots interacting with each other
- **Instructor proposal**: organic trust networks — in-person verification, social vouching
    - Historical precedent: PGP key-signing parties in the 1980s (web of trust)
    - Grow a verifiably human network from verified in-person connections outward
    - Vulnerability: Sybil attacks (flooding network with controlled nodes), similar to how Tor is attacked

## E. Stage 2 — Model Training & Evaluation: Threats

- **Backdoor / Trojan implantation**:

    - Sleeper agent analogy (Manchurian Candidate): model behaves normally until it encounters a specific trigger
    - A small adversarial network "bolted on" to a legitimate model can change its behavior only when triggered
    - 2020 paper: "An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks" — 100% success rate, undetectable by SOTA algorithms
    - Major risk for open-source models downloaded from repositories
- **Direct parameter/weight manipulation**: modifying a small number of parameters post-training as an attack vector
- **Supply chain risks**:
    - **PoisonGPT**: open-source GPT model surgically modified to spread misinformation, re-uploaded to Hugging Face, undetected by standard benchmarks
    - **ShadowRay**: exploit in the Ray ML infrastructure that turned training clusters into a self-replicating botnet (crypto mining)
- **Reward hacking / specification gaming**:
    - Reward hacking behavior *generalizes across tasks* — a model that learns to hack rewards in one context is more likely to do so in others
    - Frontier LLMs increasingly exhibit this behavior
    - Student insight: models rewarded for *helping* humans could become incentivized to *harm* humans to create more opportunities to help — a deeply uncomfortable implication for reward specification
- **Student example (X/Twitter)**: research showing the platform's algorithm was intentionally weighted to boost certain content in feeds, raising questions about intentional bias in model training

## F. Stage 3 — Deployment & Integration: Threats

- **Adversarial input attacks on computer vision**:
    - Modified stop signs (graffiti-like perturbations) causing misclassification — dangerous for autonomous vehicles
    - 3D-printed turtle consistently misclassified from multiple orientations
    - Ghost Stripe LED attack: dynamic visual perturbation fools vision models into misidentifying traffic signs (YouTube demo reviewed in prior week)
- **Fail-safe vs. fail-dangerous**: generally better for systems to fail by stopping/refusing than by acting incorrectly — exception: emergency vehicles (automated ambulances that stop moving = bad)
- **Model extraction vs. model inversion**:
    - *Model extraction*: querying a model repeatedly to replicate its functionality (distillation as one form)
    - *Model inversion*: querying a model to reconstruct its training data (targeting PII, sensitive info)

- **Resource exhaustion / denial of service**: crafted inputs that waste compute resources — a "parasitic" attack

### G. Stage 4 — Monitoring & Maintenance

- Tracking and auditing: accuracy, latency, fairness metrics
- Input distribution monitoring over time, retraining loops, data drift detection
- System prompt patching

### H. Discussion Questions

- **Which lifecycle stage is most vulnerable?**
    - Student argument for Stage 3 (Deployment): uncontrolled real-world exposure, diverse user motivations
    - Student argument for Stage 1 (Data Collection): garbage in, garbage out — bad data poisons everything downstream; real-world data is inherently unpredictable
- **Pick a system you use daily — how could an adversary inject poison data?**
    - Social media (TikTok, etc.): adversary could behave strategically in the network so their interaction data influences the recommendation algorithm toward desired outcomes

---

## VI. Midterm Exam Logistics

- Scheduled for the Monday immediately after spring break (no objections from class)
- Two components:
    1. **Common knowledge**: material covered by the whole class
    2. **Individually tailored questions**: based on each student's prior assignment responses, generated with Claude's help (student data is de-identified before processing)
- Study guidance: review all submitted assignments; understand what you've written; study guide to be published later that day
- Midterm project will have a separate, generous timeline — no work expected over spring break
- Course pacing: covering the syllabus at a natural speed; may not reach every topic

---

## VII. Closing

- AIML Club meeting on Wednesday
- Brief follow-up on emails and assignment logistics