CYB-4203/6203: Secure and Trustworthy AI

v1.0.3 - February 26, 2026

Instructor Information

- **Instructor**: Dallas Elleman
- **Contact**: dallas-elleman@utulsa.edu
- **Office**: Rayzor Hall 2040
- **Office Hours**: By appointment

Course Information

- **Course Name**: Secure and Trustworthy AI
- **Course Description**: This course explores privacy, ethical, and risk management aspects of AI systems across their lifecycle. Students will engage with several management frameworks and through a study of influential essays and review of case studies, will explore the complexities of this new and all-reaching technology.
- **Course Website**: https://cyb4203.dev/
- **Course Time and Location**: Monday/Wednesday 12:30-1:45 PM, Zink Hall 219
- **Targeted Audience**: Junior and senior undergraduate students, graduate students
- **Prerequisite**: CS-2001 "Computer and Engineering Ethics"; OR CYB-3023 "Cyber and Society". CS-3xx3 "Fundamentals of AI Systems"

Teaching Methods

The course consists of two lecture sessions each week, weekly reading and work assignments, class discussions, a midterm project, an end-of-semester project, and midterm and final exams. Course materials, lecture notes, readings, and assignments are uploaded to Blackboard and the course website prior to meeting time of the relevant lecture.

Note: The instructor may adjust the schedule, assignments, or assessment methods as needed to support student learning. Students will be notified of any changes through Blackboard and in class.

Student Learning Outcomes

Upon successful completion of this course, a student will be able to:

- **SLO1**: Understand the importance of privacy, transparency, and risk management in AI systems.
- **SLO2**: Understand legal and regulatory considerations impacting AI systems.
- **SLO3**: Evaluate AI systems for fairness, accountability, and transparency, and be able to identify and mitigate biases in AI systems.
- **SLO4**: Apply privacy-enhancing technologies like differential privacy and federated learning.
- **SLO5**: Apply cybersecurity measures across all phases of an AI system lifecycle.

Note: This course is aligned with the NCAE-AI Knowledge Units (KUs): AIG, AIL, and AIR, which can be found at https://dl.dod.cyber.mil/wp-content/uploads/cae/pdf/unclass-cyber_ai_kus_stoneman.pdf

- **AIG**: AI Governance, Laws, and Ethics
- **AIL**: Securing the AI Lifecycle
- **AIR**: Risk Management of AI

Textbook and Materials

Required textbook:

- Hendrycks - Introduction to AI Safety, Ethics, and Society (text available for free online at https://www.aisafetybook.com/)

Additional reading assignments will be drawn from essays, academic literature, and regulatory guidance, all made available on the course website.

Programming Language

Example programs in lectures will be written in pseudocode and/or Python. Students are encouraged to use Python to implement any class project.

Course Schedule

The course is organized into 15 units across 4 sections. Each unit corresponds roughly to one week of the semester.

Section 1 - WHY: Ethics, Dangers, Society, and Accountability

Unit 1: Course Overview & the Rationale for Secure and Trustworthy AI

- 1.1 Course introduction and objectives
- 1.2 Societal stakes: the transformative promise and risk of AI systems
- 1.3 Overview of prominent failures, scandals, and incidents
- 1.4 Influence of AI on society, economy, and geopolitics

Unit 2: Ethics, Values, and Human Impact of AI

- 2.1 Philosophical ethical frameworks (virtue ethics, utilitarianism, deontology) applied to AI systems
- 2.2 Core AI values: fairness, transparency, accountability, privacy, autonomy, safety, and sustainability
- 2.3 AI and human rights: cross-cultural, legal, and environmental perspectives
- 2.4 Human-AI collaboration: designing systems that augment rather than replace human capabilities

Unit 3: Potential Harms, Misuse, and Responsible Innovation

- 3.1 Unintended harms: e.g., algorithmic bias and discrimination in criminal justice, hiring, lending, and healthcare
- 3.2 Intentional misuse: e.g., deepfakes, coordinated misinformation, surveillance, autonomous cybercrime, and malicious applications
- 3.3 The alignment problem, value learning, and catastrophic risks (AI safety research perspectives)
- 3.4 Responsible innovation practices and professional obligations in AI development

Unit 4: Regulatory and Legal Context for AI

- 4.1 International AI governance: GDPR, EU AI Act, and emerging global frameworks
- 4.2 U.S. AI regulation: CCPA, Executive Orders, and sectoral requirements
- 4.3 Organizational frameworks: NIST AI RMF, ISO standards, and industry best practices
- 4.4 Documentation and auditability requirements: model cards, datasheets, and transparency reporting

Section 2 - WHAT: Technical and Operational Foundations and Vulnerabilities

Unit 5: Biology, Neuroscience, and Psychology connections to AI/ML Systems, Lifecycles, and Security

- 5.1 Biology, Neuroscience, and Psychology connections to AI/ML
- 5.2 Core architecture of modern AI/ML systems and how they differ from traditional software
- 5.3 AI/ML system lifecycles: data collection and preparation, model training and evaluation, deployment and integration, monitoring and maintenance

Unit 6: AI/ML Attack Vectors

- 6.1 Vulnerabilities across the AI/ML lifecycle: data, training, inference, and deployment phases
- 6.2 Traditional ML attack vectors: e.g., adversarial examples, data poisoning, model poisoning, and backdoor attacks
- 6.3 LLM-specific vulnerabilities: e.g., prompt injection, jailbreaking, model extraction, and training data extraction
- 6.4 Threat modeling frameworks and risk assessment

Unit 7: Privacy, Bias, Transparency, and Explainability

- 7.1 Privacy risks in AI: e.g., membership inference, data leakage, surveillance, and predictive harm
- 7.2 Bias in AI systems: types, sources, and fairness evaluation frameworks
- 7.3 Algorithmic transparency and accountability in high-stakes decision-making
- 7.4 Explainability and interpretability: concepts, techniques, and tools (e.g., LIME, SHAP)

Section 3 - HOW: Tools, Practices, Risk Management, and Governance

Unit 8: Privacy-Enhancing and Security Technologies

- 8.1 Differential privacy: concepts, mechanisms, and privacy-utility tradeoffs
- 8.2 Federated learning: architecture, security considerations, and applications
- 8.3 Homomorphic encryption: principles and use cases for computation on encrypted data
- 8.4 Secure multi-party computation: collaborative learning without exposing private data
- Submit Midterm Project (Covers Units 1-7)

SPRING BREAK

Unit 9: Testing, Evaluation, and Red-Teaming

- 9.1 Testing: security testing methodologies and vulnerability assessment for AI/ML systems
- 9.2 Evaluation: benchmarks, metrics, datasets, and tools
- 9.3 Red-teaming: adversarial approaches, attack simulation, and penetration testing

Unit 10: Building and Operationalizing Secure and Trustworthy AI/ML Systems

- 10.1 MLOps and supply chain security: CI/CD pipelines, model repositories, dependencies, and deployment infrastructure
- 10.2 Communication and orchestration: agentic protocols, tool calling, and coordination patterns
- 10.3 Context management and memory systems: RAG, vector stores, context windows, and state management
- 10.4 Security controls and guardrails: access management, output filtering, sandboxing, and human oversight
- 10.5 Operational practices: monitoring, incident response, cost optimization, and responsible release strategies

Unit 11: Risk Management and Crisis Response

- 11.1 NIST AI Risk Management Framework: structure, implementation, and practical application
- 11.2 Organizational governance: risk assessment, ethics boards, and accountability structures
- 11.3 Incident response and crisis management: preparation, escalation, and recovery protocols
- 11.4 Case studies and lessons learned: analyzing AI failures using incident databases

Unit 12: Independent Auditing, Documentation, and Disclosure

- 12.1 Documentation standards: model cards, datasheets, and transparency requirements
- 12.2 Preparing for external audits and regulatory review
- 12.3 Independent evaluation: third-party testing, certification, and validation processes
- 12.4 Stakeholder engagement: disclosure practices, communication, and accountability mechanisms

Section 4 - SYNTHESIS: Industry, Professionalism, and Final Project

Unit 13: Industry Applications & Emerging Challenges

- 13.1 Sector-specific applications: security and trust requirements in healthcare, finance, defense, and critical infrastructure
- 13.2 Policy debates: open model release, foundation model regulation, and AI governance evolution
- 13.3 Current landscape: recent incidents, regulatory developments, and industry responses
- 13.4 Emerging technologies and future challenges

Unit 14: Professionalism, Pathways, Future Directions, and Final Project Workshop

- 14.1 Career pathways and certifications in AI security and trustworthy AI
- 14.2 Professional development: resources, communities, and continuing education
- 14.3 Course synthesis: integrating ethics, security, and operational practices
- 14.4 Final project workshop: development support and case study integration
- 14.5 Peer presentations and feedback on proposed solutions

Unit 15: Course Conclusion

- 15.1 Final project presentations and demonstrations
- 15.2 Q&A and feedback session
- 15.3 Course wrap-up and reflections

Course Policy

Student Evaluation

Students are evaluated by their performance on weekly assignments and exams. At the end of the course, all points will be weighted to meet the following percentages by category:

1. **Attendance and Participation**: 10%
2. **Weekly Assignments**: 30%
3. **Projects**: 30%
    - Midterm Project (Solo): Due ~Week 8
    - Final Project (Group): Due ~Week 15
4. **Midterm Exam**: 15%
5. **Final Exam**: 15%

Graduate Students: In addition to the above requirements, graduate students must deliver a 25-minute professional-level research presentation on a topic related to course material. This presentation will be worth

10% of final grade, with other category weights reduced proportionally.

Extra Credit Opportunities:

- 1% bonus to final grade for each AIML Club Friday meeting attended (limit 4 meetings)
- 1% toward final grade for every 2 hours volunteered at an AIML Club event (limit 1% per event)
- 1% toward final grade for delivering a 10-minute presentation with slide deck on a topic related to current or next unit's material (verify suitability beforehand; limit 2 per semester)
- For information about TU AIML Club meetings and events, see https://aiml-utulsa.github.io/Club-Webpage/

Grading Scale

Mid-term grades (Unit 7-8) are based on the total points as of a cutoff assignment date. The final grades are calculated at the end of the semester. The following weighted point ranges will result in the indicated final grades:

- $85 \leq$ score ≤ 100: A
- $75 \leq$ score < 85: B
- $65 \leq$ score < 75: C
- $55 \leq$ score < 65: D
- score < 55: F

AI Policy

Students are encouraged to adopt and use AI and LLM tools securely, responsibly, and honestly to augment—not replace—their thinking and effort.

AI Use Declaration: All assignments MUST include a declaration of AI use that details the extent and purpose of AI tool usage. Students take full responsibility for all submitted materials, including any AI-generated content.

Access: Students without current access to LLM services should contact the instructor to discuss access options.

Examples of Acceptable Use

- Using AI to brainstorm ideas or explore different approaches to a problem
- Requesting explanations of complex concepts or technical documentation
- Debugging code by asking AI to identify potential issues or suggest improvements
- Generating initial drafts or outlines that you then significantly revise and personalize
- Translating technical jargon or summarizing research papers to aid understanding
- Asking AI to review your work for clarity, grammar, or logical flow

Examples of Unacceptable Use

- Submitting AI-generated content as your own without significant original thought or modification
- Using AI to complete assignments without demonstrating your own understanding of the material
- Copying AI responses verbatim without attribution or critical evaluation
- Using AI to circumvent the learning objectives of an assignment
- Fabricating or misrepresenting the extent of AI use in your declaration
- Relying on AI-generated responses without verifying their accuracy or appropriateness

Sample AI Use Declarations

Example 1 - Minimal Use:

> "I used Claude to explain the concept of differential privacy after struggling with the textbook definition. I then wrote my explanation in my own words based on my understanding. I also used Grammarly to check for spelling and grammar errors."

Example 2 - Moderate Use:

> "I used ChatGPT to brainstorm potential vulnerabilities in the AI system described in the assignment. I selected three vulnerabilities from its suggestions and independently researched each one, including finding my own sources and examples. I wrote the analysis entirely myself. I then used Claude to review my draft for clarity and logical organization, implementing about half of its suggestions."

The specifics of acceptable AI use for different assignment types will be discussed and refined collaboratively with students throughout the course.

Grade Posting

Grade posting occurs at the end of the semester after all projects and the final exam are graded. Students may go to https://selfservice.utulsa.edu/student to check their grades. Points for assignments will be posted on Harvey/Blackboard as soon as they are graded. No confidential information will be sent through email or given over the phone.

Final Exam

The final exam is comprehensive. Time and location will be determined according to the university finals schedule (https://utulsa.edu/academic-calendar/finals-schedule/).

Assignment Submission

Weekly assignments and projects: students must submit their work through Harvey/Blackboard (https://harvey.utulsa.edu).

No submission will result in a zero for the assignment; turning in a partial assignment is better than nothing. The instructor reserves the right to grant or deny extensions at their discretion, but will not grant extensions if no assignment (even incomplete) has been turned in by the due date, except in cases of reasonable extenuating circumstances (e.g., health issues, disasters).

Programming Style

Students are strongly encouraged to use Python in course projects. Student assignments and projects require proper programming style, documentation, overall design, and the correctness of the output. Examples used by the instructor during lecture will exclude many programming-style details to use the limited time and board space effectively.

Student Etiquette

Students should be on time to attend lectures. Please do not use your phones in the classroom. Attendance of lectures is required in this course. You must email the professor before the lecture if you will miss a lecture due to some documented hardships.

University Policy Information

Academic Misconduct

The university has documents covering academic misconduct:

All undergraduate colleges:
https://univoftulsa.sharepoint.com/sites/AcademicAffairs/academicpolicies/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FAcademicAffairs%2Facademicpolicies%2FShared%20Documents%2FUnified%20Academic%20Misconduct%20Policy%2Epdf

College of Law:
https://univoftulsa.sharepoint.com/sites/AcademicAffairs/academicpolicies/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FAcademicAffairs%2Facademicpolicies%2FShared%20Documents%2FTU%5FLaw%5FHonor%5FCode%2Epdf

Graduate School:
https://univoftulsa.sharepoint.com/sites/AcademicAffairs/academicpolicies/Shared%20Documents/Forms/AllItems.aspx?id=%2Fsites%2FAcademicAffairs%2Facademicpolicies%2FShared%20Documents%2FGraduate%20School%20Academic%20Misconduct%20Policy%2D%20August%20

All instances of academic misconduct are reported to the Dean's Office. At a minimum, students who cheat will receive a score of zero on the assignment in question; but students may also be dismissed from the course and automatically assigned a grade of F.

Student Access and Success

Students who have or believe they may have a disability and would like to set up accommodations should contact Student Access within Student Success to self-identify their needs and facilitate their rights under the Americans with Disabilities Act and related laws. Student Access provides private consultations to any student. Contact Student Access staff by email at studentaccess@utulsa.edu or by phone at 918-631-2315. The application for accommodations may be obtained online at https://sierra.accessiblelearning.com/s-UTulsa/ApplicationStudent.aspx.

Student Access staff will assist students in the implementation of approved accommodations, and students should submit requests as early as possible for full assistance. Students who qualify for accommodations should meet with the instructor privately (during office hours or by appointment) as soon as possible to arrange for their needs and obtain support for the class. Instructors are entitled to notice of 5 business days before the implementation of any required accommodations and all accommodations should be requested by the 12th week of classes for use in that semester, absent an extraordinary and unforeseeable circumstance. TU maintains a list of accessible features for all buildings (e.g., entrances, parking) at https://utulsa.edu/campus-map/.

Know Your Title IX

Sexual misconduct is prohibited by Title IX of the Educational Amendments of 1972 ("Title IX") and will not be tolerated within the TU community. For more information about your rights under Title IX, please visit our Policies and Laws page https://utulsa.edu/sexual-violence-prevention-education/policies-laws/ on the TU website or contact the Title IX Coordinator at 918-631-2321.

AI Use Disclaimer

This syllabus was drafted and refined using Claude 4.5 Sonnet and Claude Code. All content was reviewed and approved by the author, Dallas Elleman, who takes full responsibility for its publication.