CYB-4203/6203

Secure and Trustworthy AI

Course Intro: Wednesday, January 21, 2026

Dallas Elleman | Zink Hall 219

Opening Discussion

What is Intelligence?

Discussion Questions

- How do you define intelligence?
- What are the characteristics of intelligent behavior?
- Is intelligence a single quality or multiple capabilities?
- How do humans demonstrate intelligence?

Opening Discussion

Intelligence: Key Considerations

- Problem-solving ability
- Learning and adaptation
- Pattern recognition
- Reasoning and decision-making
- Memory and knowledge application
- Creativity and innovation

Opening Discussion

What is Security?

Discussion Questions

- What does "security" mean to you?
- Security of what? Against what?
- Who is responsible for security?
- What are the trade-offs of security measures?

Opening Discussion

Security: Key Considerations

- Confidentiality, Integrity, Availability (CIA Triad)
- Protection against threats and vulnerabilities
- Risk management
- Access control and authentication
- Defense in depth
- Security vs. usability

Opening Discussion

What is Trust?

Discussion Questions

- What does it mean to trust something or someone?
- Can trust be measured or quantified?
- How is trust earned? How is it lost?
- What's the difference between trust and reliability?

Opening Discussion

Trust: Key Considerations

- Predictability and consistency
- Transparency and explainability
- Accountability and responsibility
- Alignment with values and expectations
- Verification vs. trust

Opening Discussion

What Dimensions Does Artificiality Add?

Discussion Questions

- How does "artificial" intelligence differ from "natural" intelligence?
- What unique challenges arise from artificial systems?
- What unique opportunities?
- How do human-AI interactions differ from human-human interactions?

Opening Discussion

Artificiality: Key Considerations

- Scale and speed (processing vast amounts of data)
- Lack of consciousness or intentionality
- Opacity and complexity ("black box" problem)
- Training data biases and limitations
- Autonomy and agency questions
- Unprecedented capabilities and risks

Synthesis

Secure and Trustworthy AI

- AI systems operate at unprecedented scale and speed
- They make consequential decisions affecting individuals and society
- Trust requires both security AND other qualities
- Building secure AI requires interdisciplinary expertise

Course Mission: To equip you with the knowledge, skills, and frameworks to understand, evaluate, and build secure and trustworthy AI systems

Course Overview

Syllabus Quick Facts

- Format: MW 12:30-1:45 PM in Zink Hall 219
- Textbook: Hendrycks - Introduction to AI Safety, Ethics, and Society (free online)
    - https://www.aisafetybook.com/
- Prerequisites: CS-2001 OR CYB-3023, AND CS-3xx3
- Instructor: Dallas Elleman (Rayzor Hall 2040)
- Email: dallas-elleman@utulsa.edu

Course Structure

Unit 1 - WHY

Ethics, Dangers, Society (Weeks 1-4)

Unit 2 - WHAT

Foundations & Vulnerabilities (Weeks 5-7)

Unit 3 - HOW

Tools, Practices, Governance (Weeks 8-12)

Unit 4 - SYNTHESIS

Industry & Professionalism (Weeks 13-15)

Course Structure

Unit 1 - WHY: Ethics, Dangers, Society (Weeks 1-4)

- Week 1: Course Overview & Rationale for Secure AI
- Week 2: Ethics, Values, and Human Impact
- Week 3: Potential Harms and Responsible Innovation
- Week 4: Regulatory and Legal Context

Unit 2 - WHAT: Foundations & Vulnerabilities (Weeks 5-7)

- Week 5: AI/ML System Components & Threat Landscape
- Week 6: AI/ML Attack Vectors
- Week 7: Privacy, Bias, Transparency, Explainability

Course Structure

Unit 3 - HOW: Tools, Practices, Governance (Weeks 8-12)

- Week 8: Privacy-Enhancing & Security Technologies
- Week 9: Testing, Evaluation, and Red-Teaming
- Week 10: Building Secure AI/ML Systems
- Week 11: Risk Management and Crisis Response
- Week 12: Auditing, Documentation, Disclosure

Unit 4 - SYNTHESIS: Industry & Professionalism (Weeks 13-15)

- Week 13: Industry Applications & Emerging Challenges
- Week 14: Professionalism, Pathways, Future Directions
- Week 15: Course Conclusion

Student Learning Outcomes

Upon completion, you will be able to:

- Understand privacy, transparency, risk management in AI
- Understand legal and regulatory considerations for AI
- Evaluate AI for fairness, accountability, transparency
- Apply privacy-enhancing technologies
- Apply cybersecurity across AI system lifecycle

NCAE-AI Aligned: Governance, Lifecycle, Risk Management

Grading & Evaluation

Grade Breakdown

Category	Weight
Attendance & Participation	10%
Weekly Assignments (12)	30%
Projects (Midterm + Final)	30%
Midterm Exam	15%
Final Exam	15%

Grading Scale

A: 85-100 | B: 75-84 | C: 65-74 | D: 55-64 | F: <55

Extra Credit Opportunities

6%+ bonus available

- 1% per AIML Club Friday meeting attended (limit 4)
- 1% per 2 hours volunteered at AIML Club events
- 1% per 10-minute presentation on course topic (limit 2)

Assignments & Policies

Weekly Assignments

- 12 total assignments
- Assigned Wednesdays, due following Mondays
- Partial submission better than nothing

Projects & Exams

- Midterm Project (Solo): Due Week 8 (March 11)
- Final Project (Group): Due Week 15 (May 4)
- Midterm Exam: Week 8 | Final Exam: Finals period

Extensions

- Instructor discretion with partial submission
- No extension granted without partial submission by assignment deadline
- Exceptions: documented extenuating circumstances

AI Use Policy

Core Principle: Use AI to augment—not replace—your thinking and effort

- AI is a transformative technology that can:
- Level the playing field in many ways
- Create new opportunities for learning
- Introduce new challenges and inequalities

In this course, you're encouraged to use AI tools securely, responsibly, and honestly.

AI Use: Examples & Requirements

Acceptable Use Examples

- 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

Unacceptable Use Examples

- 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

AI Use: Examples & Requirements

Mandatory AI Use Declaration - All assignments MUST include a declaration of AI use that details the extent and purpose of AI tool usage

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."

AI Access: Don't have access to AI/LLM tools? Contact me to discuss access options - everyone will have equitable access to AI tools.

Collaborative Refinement

The specifics of acceptable AI use will be discussed and refined collaboratively with you throughout the course.

- Different assignment types may have different AI use guidelines
- We'll learn together what works best
- Your feedback will shape our policies

This is an evolving conversation, not a fixed rulebook.

Taking While Teaching: An Experiment in Transparency

What Does This Mean?

- I'll complete all assignments
- I'll participate in discussions
- I'll share my work and thinking process
- You'll grade my work

Your Role

- Provide honest feedback on my work
- Hold me to the same standards
- Challenge my thinking
- Help me learn too

Why?

- Transparency: You see the standards I'm holding myself to and understand what "good work" looks like
- Empathy: I experience the assignments as you do and understand the time and effort required
- Learning Together: AI security is a rapidly evolving field. We're all continuous learners, and I want to model professional learning practices

This is a collaborative learning environment.

Assignment 1: Intro Questionnaire

Assigned: Today | Due: Monday, Jan 26, 12:29 PM

Requirements

- Answer all 11 questions thoughtfully
- Include AI Use Declaration
- Submit via Blackboard/Harvey
- No right or wrong answers—be honest

Next Steps

Before Next Class (Monday, Jan 27)

- Complete Assignment 1 (due Monday 12:29 PM)
- Review the full syllabus
- Set up AI tool access if needed
- Read Hendrycks Chapter 6: Beneficial AI
- https://www.aisafetybook.com/textbook/beneficial-ai

Next Class: Week 1 Content Begins

Office Hours: By appointment | dallas-elleman@utulsa.edu | Rayzor 2040

Welcome to the Course!

I'm looking forward to learning with you this semester

This presentation was drafted and refined using Claude 4.5 Sonnet and Claude Code.

All content was reviewed and approved by Dallas Elleman.