
UHON 351: **AI Safety and Society**

Semester: Spring 2026

Time: 01:00 PM – 04:00 PM

Days: Monday

Location: Morris Library 0610 A

Credits: 3

Instructor Information

Professor: **Dr. Abdur Rahman Bin Shahid**

Office Location: EGRA 409C

Office Hours: Tuesday and Wednesday 11 AM – 02:00 PM

Email: shahid@cs.siu.edu

Course Description

This course examines the societal and ethical challenges of ensuring AI safety in a rapidly advancing technological landscape. Students will explore catastrophic risks, ethical frameworks, and governance solutions while engaging in collaborative projects and discussions. The course encourages students to think critically about the impacts of AI on society and reflect on how responsible AI governance can empower individuals and communities.

Student Learning Objectives

- Identify and explain major AI-related risks, including societal risks, ethical concerns, and unintended consequences of advanced AI systems.
- Evaluate and apply ethical principles and frameworks relevant to AI development, understanding how these frameworks can guide responsible AI practices.
- Reflect on the societal impact of AI and discuss how AI governance and regulation can affect individual rights, collective well-being, and societal norms.
- Develop practical and ethical governance proposals to address AI safety challenges.

Outline of Topics

Topic	Overview and Key Questions
Introduction to AI and Its Societal Impact	Overview: Basics of AI systems, including key technologies and their applications in various fields. Key Questions: What is AI, and how is it shaping our society? Why are AI advancements raising ethical and safety concerns?

Catastrophic AI Risks	Overview: Exploring catastrophic risks, such as rogue AI, malicious use, AI race dynamics, and organizational risks. Key Questions: What are the potential large-scale harms associated with advanced AI? How can specific risks like AI races or loss of control lead to societal disruption?
Fundamentals of Machine Learning for AI Safety	Overview: Introduction to machine learning fundamentals, with a focus on neural networks, large language models (LLMs), and scaling laws. Key Questions: How do machine learning models work, and what technical challenges make ensuring their safety difficult?
Challenges in Ensuring AI Safety	Overview: Technical safety challenges, including transparency issues, adversarial attacks, and unintended AI behavior. Key Questions: What are the limitations of current AI systems in terms of robustness and control? How do adversarial attacks compromise AI safety?
Ethics in AI: Machine Ethics and Beneficial AI	Overview: Examination of ethical principles for AI, such as balancing user preferences, law, and ethics in decision-making. Key Questions: How should AI be guided by ethical principles? What happens when ethical, legal, and user preferences conflict?
Collective Action and Global AI Governance	Overview: Addressing the need for cooperative action in AI safety across different nations and organizations. Key Questions: How do global competitive pressures impact AI safety? What role does international cooperation play in ensuring AI safety?
Policy and Governance for Responsible AI	Overview: Exploring various approaches to AI governance, including national policies, safety standards, and international frameworks. Key Questions: What governance models could effectively balance the risks and benefits of AI? How do different governance structures impact safety?

List of Texts and Other Reading Materials

- Dan Hendrycks. *Introduction to AI Safety, Ethics and Society*. Taylor & Francis, (forthcoming). ISBN: 9781032798028. URL: www.aisafetybook.com
- National Institute of Standards and Technology (NIST). (2024). *NIST AI Risk Management Framework (AI RMF) 1.0*. Retrieved from <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf>
- Advanced Artificial Intelligence Safety Initiative (AISI). (2024). *Early Lessons from Evaluating Frontier AI Systems*. Retrieved from <https://www.aisi.gov.uk/work/early-lessons-from-evaluating-frontier-ai-systems>

Brief description of learning experiences, projects, nature of assignments.

This course will engage students in a variety of interactive and thought-provoking activities designed to deepen their understanding of AI safety and its societal impact. The learning experiences include:

1. **Case Studies:** Analysis of real-world scenarios to explore the ethical dilemmas and risks posed by AI. Students will work in teams to present their findings and lead discussions on possible solutions and strategies for mitigation.
2. **Reflective Essays:** Individual assignments where students reflect on their perspectives about AI's role in society, ethical considerations, and personal insights gained from class discussions.
3. **Collaborative Projects:** Group projects where students develop policy proposals or ethical frameworks for managing AI risks. These projects focus on teamwork, research, and communication, allowing students to apply concepts in practical, relatable ways without requiring technical expertise.
4. **Quizzes:** Short, periodic quizzes designed to assess students' understanding of key concepts, terminology, frameworks, and ethical principles related to AI and society.

Assessment Component	Description	Percentage
Case Studies	Team-based analysis, presentations, and discussions	20%
Reflective Essays	Individual written reflections on AI and society	20%
Collaborative Project	Group policy or ethical framework development project	30%
Quizzes	Short, periodic conceptual assessments	30%
Total		100%

Honors Policy

Students must earn a C or above in UHON seminars to get Honors credit for the course.

Description of how this course provides a unique educational experience appropriate for Honors students.

“AI Safety and Society” offers a unique educational experience by emphasizing the national significance of responsible AI development. By bringing together ideas from ethics, public policy, and social sciences, this course helps students explore how AI safety impacts society, national security, and public well-being. Students will dive into real-world case studies and participate in discussions and group projects. This hands-on approach allows them to analyze and discuss policies and strategies that can protect national interests and public safety. By understanding the broader implications of AI, students will be equipped to engage in conversations and contribute to solutions that promote responsible and secure AI practices.

GRADING SCALE

A = 93–100	C = 70-74
A– = 90–92	C– = 65-69
B+ = 88–89	D+ = 60-64
B = 85–87	D = 55-59

B- = 80-84 F = 0-54
C+ = 75-79

MAKE-UP POLICY, LATE WORK

- Late assignments and projects will be assessed at a 20% penalty for each day it is late after the submission time.
- Missed quizzes cannot be made up unless you have a valid reason.

Email Policy

Please note that all email communication from Professor to students will be through the SIUC student accounts. Students sending emails to the Professor can expect an answer within 24 hours. (This does not include weekends) Between Friday at 5:00pm and Sunday at 5:00pm a response is NOT guaranteed. In many cases, a weekend response may be given to the student, but the student should not count on their email being answered during these times. Students also should understand that although responses to emails may be given very late at night, this may not always be possible.

Attendance Policy

Regular attendance is mandatory for all class sessions. Arrive on time to avoid disruptions. A maximum of 4 days of absence is allowed throughout the semester. Exceeding this limit will result in zero attendance points. If you miss more than 35% of scheduled classes, you will receive a letter grade of F for the course.

Plagiarism Policy

Plagiarism is presenting another existing work, original ideas, or creative expressions as one's own without proper attribution. Any ideas or materials taken from another source, including one's own work, must be fully acknowledged unless the information is common knowledge. What is considered "common knowledge" may differ from subject to subject. To avoid plagiarizing, one must not adopt or reproduce material from existing work without acknowledging the original source. Existing work includes but is not limited to ideas, opinions, theories, formulas, graphics, and pictures. Examples of plagiarism, subject to interpretation, include but are not limited to directly quoting another's actual words, whether oral or written; using another's ideas, opinions, or theories; paraphrasing the words, ideas, opinions, or theories of others, whether oral or written; borrowing facts, statistics, or illustrative material; and offering materials assembled or collected by others in the form of projects or collections without acknowledgment.

Cheating Policy

Cheating of any kind will not be tolerated and will result in a grade of zero for the exam or assignment in question for all parties involved. A second offense will result in a grade of F for the course. The instructor will be the sole judge of when cheating has occurred. The instructor reserves the right to review an assignment orally with a student, with the option to adjust the grade based on the results of the review. All incidents of cheating will be reported to the Academic Integrity Committee at the university, which may impose further sanctions.

Note: Data extracted from this course may be used for research purposes.

Note: This course outline is subject to change