FLORIDA ATLANTIC UNIVERSITY

CAP 4623
TRUSTWORTHY ARTIFICIAL
Date:
Building: TBA Room: TBA
03 Credit(s)

Instructor Information

Fernando Koch
Email: kochf@fau.edu
Office: TBA
Office Hours: TBA
Phone: 914-309-2643

TA Name: TBA
Office: TBA
Office Hours: TBA
Telephone: Email: TBA

Course Description

This course introduces the foundational concepts of Trustworthy AI, emphasizing the importance of ethical considerations, robustness, and accountability in AI systems. Students will explore various dimensions of Trustworthy AI and learn methodologies to design, implement, and evaluate AI systems that align with ethical standards and societal values.

Imagine a scenario where the application of AI has the potential to save lives in critical medical situations but could also lead to ethical dilemmas if not properly managed. This course will delve into such scenarios, ensuring that students understand how to harness the power of AI responsibly and ethically.

The course adopts an active learning approach where presentations work as gateways to a series of practical labs. The course will immerse students in the realm of real-world AI applications, providing hands-on experience with designing and evaluating AI systems. Through these practical labs, students will tackle challenges related to fairness, transparency, and accountability, ensuring they can apply theoretical knowledge to practical situations.

The objectives of this course include:

- Understanding the principles and importance of Trustworthy AI.
- Analyzing and addressing ethical considerations in AI development.
- Understanding governance frameworks that ensure AI accountability and responsibility.
- Understanding the principles behind AI systems that are robust, safe, and accountable.
- Learning techniques for ensuring transparency, privacy, and data security in AI.
- Promoting fairness, inclusivity, and societal well-being through AI applications.
- Evaluating the impact of AI systems on society and the environment.

Prerequisites

COP 4773 and (COP 3014C or COP 3530C)

- **Introductory Knowledge of Artificial Intelligence**: students should have a basic understanding of AI concepts, including machine learning, neural networks, and natural language processing. The course will provide preparatory and self-study material prior coursework.
- **Collaborative Mindset and Teamwork Skills**: As the course involves practical labs where students will work in teams, the ability to collaborate effectively and contribute to group projects is important. Previous experience working in collaborative environments, whether in academic, professional, or personal projects, will be advantageous.
- **Readiness for Active Learning**: The course is designed with an active learning approach, where students will engage in hands-on labs and real-world scenarios. Students should be prepared to actively participate, take initiative, and immerse themselves in the learning experience.

Corequisites

Students are encouraged to concurrently enroll in or review the following corequisite subjects:

- **Introduction to Artificial Intelligence:** a course, preparation material, or module that covers the fundamental principles of AI, including machine learning, neural networks, and natural language processing.

Instructional Method

In-Person -- This course may be offered in in-person, hybrid, or fully online modes

Required Texts/Materials

Course Objectives/Student Learning Outcomes

At the end of this course, students should be able to:

- Understand Trustworthy AI Principles: articulate the foundational concepts and importance of Trustworthy AI; explain the ethical, legal, and social implications of AI technologies.
- Analyze Ethical Considerations: identify and critically assess ethical dilemmas in AI development and deployment; apply ethical theories and frameworks to real-world AI scenarios.
- Implement Accountability Frameworks: understand the principles behind accountability mechanisms for AI systems; Understand legal and regulatory perspectives on AI accountability.
- Promote Transparency in AI: utilize techniques and tools to enhance AI transparency and explainability; develop strategies for effective communication of AI processes and decisions.
- Foster Fairness and Inclusivity: learn how to identify and mitigate biases in AI models and datasets; understand the principles behind the policies and practices that promote diversity, non-discrimination, and fairness in AI systems.
- Design Robust and Safe AI Systems: understand the principles behind developing AI systems with technical robustness and safety in mind; implement fail-safe mechanisms and redundancy in AI system design.
- Empower Human Agency in AI Systems: understand the principles behind AI systems that ensure human oversight and intervention; create mechanisms to maintain human control over autonomous AI systems.
- Assess Societal and Environmental Impact: able to evaluate the societal impact of AI technologies and their potential for social good; leverage AI for promoting environmental sustainability and societal well-being.

Faculty Rights and Responsibilities

Florida Atlantic University respects the rights of instructors to teach and students to learn. Maintenance of these rights requires classroom conditions that do not impede their exercise. To ensure these rights, faculty members have the prerogative to:

- Establish and implement academic standards.
- Establish and enforce reasonable behavior standards in each class.
- Recommend disciplinary action for students whose behavior may be judged as disruptive under the Student Code of Conduct University Regulation 4.007.

Disability Policy

In compliance with the Americans with Disabilities Act Amendments Act (ADAAA), students who require reasonable accommodations due to a disability to properly execute coursework must register with Student Accessibility Services (SAS) and follow all SAS procedures. SAS has offices across three of FAU's campuses – Boca Raton, Davie and Jupiter – however disability services are available for students on all campuses. For more information, please visit the SAS website at www.fau.edu/sas/.

Course Evaluation Method

Your grade in the class will be broken into the following components:

- **Lab Exercises: 30%**
  - Participation and performance in the lab sessions are crucial as they provide hands-on experience with Trustworthy AI. Each lab will be graded based on completeness, accuracy, and innovation in solving the given problems.

- **Midterm Exam: 30%**
  - The midterm exam will test your understanding of the fundamental concepts and applications of Trustworthy AI. The exam will include multiple-choice questions, short answers, and problem-solving questions.

- **Final Exam: 40%**
  - The final exam will be comprehensive, covering all topics discussed in the course. It will assess your ability to apply theoretical knowledge to practical scenarios, including case studies and problem-solving exercises. The exam will consist of multiple-choice questions, short answers, and essay questions.

Code of Academic Integrity

Students at Florida Atlantic University are expected to maintain the highest ethical standards. Academic dishonesty is considered a serious breach of these ethical standards, because it interferes with the university mission to provide a high quality education in which no student enjoys an unfair advantage over any other. Academic dishonesty is also destructive of the university community, which is grounded in a system of mutual trust and places high value on personal integrity and individual responsibility. Harsh penalties are associated with academic dishonesty. For more information, see University Regulation 4.001.

Attendance Policy Statement

Students are expected to attend all their scheduled University classes and to satisfy all academic objectives as outlined by the instructor. The effect of absences upon grades is determined by the

instructor, and the University reserves the right to deal at any time with individual cases of non-attendance. Students are responsible for arranging to make up work missed because of legitimate class absence, such as illness, family emergencies, military obligation, court-imposed legal obligations, or participation in University-approved activities. Examples of University-approved reasons for absences include participating on an athletic or scholastic team, musical and theatrical performances, and debate activities. It is the student's responsibility to give the instructor notice prior to any anticipated absences and within a reasonable amount of time after an unanticipated absence, ordinarily by the next scheduled class meeting. Instructors must allow each student who is absent for a University-approved reason the opportunity to make up work missed without any reduction in the student's final course grade as a direct result of such absence.

Religious Accommodation Policy Statement

In accordance with the rules of the Florida Board of Education and Florida law, students have the right to reasonable accommodations from the University in order to observe religious practices and beliefs regarding admissions, registration, class attendance, and the scheduling of examinations and work assignments. University Regulation 2.007, Religious Observances, sets forth this policy for FAU and may be accessed on the FAU website at www.fau.edu/regulations.

Any student who feels aggrieved regarding religious accommodations may present a grievance to the executive director of The Office of Civil Rights and Title IX. Any such grievances will follow Florida Atlantic University's established grievance procedure regarding alleged discrimination.

Time Commitment Per Credit Hour

For traditionally delivered courses, not less than one (1) hour of classroom or direct faculty instruction each week for fifteen (15) weeks per Fall or Spring semester, and a minimum of two (2) hours of out- of-class student work for each credit hour. Equivalent time and effort are required for Summer Semesters, which usually have a shortened timeframe. Fully Online courses, hybrid, shortened, intensive format courses, and other non-traditional modes of delivery will demonstrate equivalent time and effort.

Course Grading Scale

Letter Grade	Letter Grade
A	94 - 100%
A-	90 - 93%
B+	87 - 89%
B	83 - 86%
B-	80 - 82%
C+	77 - 79%
C	73 - 76%

C-	70 - 72%
D+	67 - 69%
D	63 - 66%
D-	60 - 62%
Letter Grade	**Letter Grade**
F	Below 60

Grade Appeal Process

You may request a review of the final course grade when you believe that one of the following conditions apply:

- There was a computational or recording error in the grading.
- The grading process used non-academic criteria.
- There was a gross violation of the instructor's own grading system.

University Regulation 4.002 of the University Regulations contains information on the grade appeals process

Policy on Make-up Tests, Late work, and Incompletes

Late submissions will not be accepted or graded. No makeup exams will be offered.

Throughout the semester, multiple homework assignments will be posted via Canvas. For each homework assignment, you will have about a week to complete and submit your solution via Canvas. Allow enough time to submit your work since once the system is closed there will not be other possibilities to submit (don't send your work via email). Please note that the due date for homework assignments will not be updated after the assignment is posted.

Policy on the Recording of Lectures

Students enrolled in this course may record video or audio of class lectures for their own personal educational use. A class lecture is defined as a formal or methodical oral presentation as part of a university course intended to present information or teach students about a particular subject.

Recording class activities other than class lectures, including but not limited to student presentations (whether individually or as part of a group), class discussion (except when incidental to and incorporated within a class lecture), labs, clinical presentations such as patient history, academic

exercises involving student participation, test or examination administrations, field trips, and private conversations between students in the class or between a student and the lecturer, is prohibited.

Recordings may not be used as a substitute for class participation or class attendance and may not be published or shared without the written consent of the faculty member. Failure to adhere to these

requirements may constitute a violation of the University's Student Code of Conduct and/or the Code of Academic Integrity.

Counseling and Psychological Services (CAPS) Center

Life as a university student can be challenging physically, mentally and emotionally. Students who find stress negatively affecting their ability to achieve academic or personal goals may wish to consider utilizing FAU's Counseling and Psychological Services (CAPS) Center. CAPS provides FAU students a range of services – individual counseling, support meetings, and psychiatric services, to name a few – offered to help improve and maintain emotional well-being. For more information, go to http://www.fau.edu/counseling/

Student Support Services and Online Resources

- Center for Learning and Student Success (CLASS) Counseling and Psychological Services (CAPS) FAU Libraries
- Math Learning Center
- Office of Information Technology Helpdesk Center for Global Engagement
- Office of Undergraduate Research and Inquiry (OURI)
- Science Learning Center Speaking Center
- Student Accessibility Services
- Student Athlete Success Center (SASC) Testing and Certification
- Test Preparation
- University Academic Advising Services
- University Center for Excellence in Writing (UCEW) Writing Across the Curriculum (WAC)

Course Topical Outline

- **Lesson 1: Basic Concepts**
    - Defining Trustworthy AI
    - Foundations of Trustworthy AI
    - Key Requirements
    - Balancing AI benefits with ethical responsibilities
    - Basic view on Ethical considerations in AI development
    - DEMO: Trustworthy AI in Practice; do you think AI is fair? Examples and Discussion.

- **Lesson 2: Requirements of Trustworthy AI**
    - Introduction to various ethical frameworks relevant to AI
    - Identifying common ethical challenges in AI (e.g., bias, privacy, consent).
    - Understanding why transparency is crucial for Trustworthy AI.
    - Techniques and practices for ensuring transparency in AI systems.
    - What is traceability in the context of AI?
    - Techniques for ensuring traceability in AI development and deployment.
    - Defining explainability and its significance in AI.
    - Methods for making AI decisions understandable to different stakeholders.
    - Importance of clear and effective communication in AI interactions.
    - Designing user interfaces that facilitate understanding and trust.
    - LAB 1: Practical Applications of Trustworthy AI Requirements; apply the concepts of ethics, transparency, traceability, explainability, and communication to a practical AI project.

- **Lesson 3: Governance and Accountability**
    - Accountability in AI Systems
    - Mechanisms for AI accountability and responsibility
    - Legal and regulatory perspectives on AI accountability
    - Developing and enforcing accountability frameworks
    - Data governance frameworks and best practices
    - Quality and integrity of data:
    - Methods for ensuring data security and user consent
    - LAB 2: The lab session will provide students with hands-on experience in designing and implementing an AI accountability framework. Given an hypothetical AI project scenario, students will design an accountability framework, including internal and external accountability mechanisms, legal and regulatory compliance; methods for governance; methods for user consent and transparency measures, and others.

- **Lesson 4: Transparency**
    - Promoting Transparency in AI Systems
    - Understanding AI transparency and its significance
    - Techniques for explaining AI decisions
    - Transparency tools and frameworks

- LAB 3: The lab session will provide students with practical experience in applying transparency techniques to AI systems. Students will work to implement transparency measures in a given AI project, ensuring that the AI processes and decisions are clear and understandable to various stakeholders. Students will identify areas in the AI project where transparency can be improved. Students will develop explanations tailored to different stakeholders (e.g., users, developers, regulators). At the end of the lab, students will summarize key takeaways from the lab and their application to real-world AI projects.

- **Lesson 5: Robustness and Safety**
  - Technically Robust and Safe AI
  - Technical robustness: Definition and importance
  - Safety protocols in AI system design
  - Fail-safe mechanisms and redundancy in AI systems
  - Resilience to attack and security
  - Fall-back plan and general safety
  - LAB 4: The lab session will provide students with hands-on experience in implementing robustness and safety measures for AI systems. Given an hypothetical AI project scenario, students will be tasked with ensuring that the AI system is technically robust, safe, and secure against potential failures and attacks. At the end of the lab, students will summarize key takeaways from the lab and their application to real-world AI projects.

- **Lesson 6: Fairness and Inclusivity**
  - Ensuring Diversity, Non-Discrimination, and Fairness
  - Defining fairness in AI: Challenges and opportunities
  - Mitigating bias in AI models and datasets
  - Policies and practices for fostering diversity and inclusivity
  - LAB 5: The lab session will provide students with hands-on experience in implementing fairness and inclusivity measures for AI systems. Given a hypothetical AI project scenario, students will be tasked with ensuring that the AI system is fair and inclusive, and that measures to mitigate bias are in place.

- **Lesson 7: Empowering Humans in AI Systems**
  - Determining the appropriate level of human control for specific AI systems and use cases.
  - Establishing mechanisms and measures to ensure human control or oversight.
  - Enabling audits and remedies for issues related to governing AI autonomy.
  - Transparency in AI Interactions

- Mechanisms to detect and respond to potential issues in AI operations.
- LAB 6: The lab session will provide students with practical experience in designing and implementing mechanisms for human oversight and control in AI systems. Students will work on a hypothetical AI project scenario to ensure effective human control, transparency, and response mechanisms.

- **Lesson 8: Regulatory Agencies and Regulations**
  - Overview of the need for regulations in AI.
  - Historical development of AI regulations.
  - Ethical and societal considerations driving regulatory frameworks.
  - Major AI Regulations and Frameworks.

- **Lesson 9: Societal and Environmental Good**
  - Societal and Environmental Well-Being through AI
  - Assessing the societal impact of AI technologies
  - Leveraging AI for environmental sustainability
  - Ethical AI for social good: Case studies and initiatives
  - LAB 7: The lab session will provide students with practical experience in designing AI solutions aimed at promoting societal and environmental well-being. Students will work on a hypothetical project to develop an AI system that addresses a specific social or environmental challenge.

- **Lesson 10: Future Trends and Innovations in Trustworthy AI**
  - Current research and developments in AI ethics.
  - New approaches to addressing ethical challenges in AI.
  - Advances in explainable AI and transparency tools.
  - Innovative techniques for enhancing transparency in AI systems.
  - New frameworks and standards for AI accountability.
  - Technological advancements in monitoring and enforcement mechanisms.
  - The role of AI in augmenting human capabilities.
  - LAB 8: Students will investigate a specific emerging trend or innovation in Trustworthy AI and develop a proposal for its implementation. Develop a proposal outlining how this innovation could be integrated into a real-world AI project. Present the proposal and discuss potential benefits and challenges.