**Responsible AI (CS 4390/5390)**

Tuesday/Thursday from 3:00 pm to 4:20 pm MST (Classroom Building C305)

Spring 2024

Dr. Saeid Tizpaz Niari

saeid@utep.edu

## COURSE DESCRIPTION

Responsible AI is a cross-list special topic course. The primary goal of this course is to study safe, responsible, and trustworthy AI. In previous courses, students have studied computer security, software systems, and machine learning/data mining/AI. In this class, we will explore safety, security, and ethical concerns in AI. After covering the basics such as trust, safety, data, AI software lifecycle, supervised learning, causal logic, generative AI, and large language models; we study responsible AI topics such as privacy, adversarial robustness, distribution shifts, fairness, risks of foundational models (e.g., ChatGPT), Interpretability, Explainability, Transparency, and AI Alignment problem.

## COURSE OBJECTIVES

Upon the completion of the course:

- Students have a clear understanding of computations in data-driven and AI-driven software solutions,
- Students can evaluate the security, privacy, and fairness of prevalent AI systems such as deep neural networks and large language models,
- Students understand the limitations of AI-enable software systems, and
- Students can apply AI techniques for social good.

## PREREQUISITES

This course requires no prior experience in security and privacy but assumes the willingness to seek out and read background material as needed. Although it is not a requirement, knowledge of the core topics of machine learning and familiarity with Python and Numpy would be very helpful.

## Course Outlines

- Establishing AI Trust
- Machine Learning Lifecycle
- AI Safety

- Data Sources, Biases, and Detection Theory
- Supervised Learning (KNN, Linear Classifiers, Optimization, Backpropagations, and deep neural networks)
- Generative AI and Large Language Models
- Adversarial Robustness and Data Poisoning
- Causality (Graphical Models, Causal Discovery, Interventions, and Counterfactuals)
- Privacy
- Distribution Shift
- Fairness
- Explainable AI
- AI Alignment

## REQUIRED MATERIALS

The primary book for this course is Trustworthy Machine Learning by Kush R. Varshney: http://www.trustworthymachinelearning.com/.

## COURSE ASSIGNMENTS AND GRADING

This is a research-oriented and discussion-based course, which also includes hands-on exercises. The students are required to write a review for assigned chapters and papers prior to the class so that they can participate in class discussions. Every student needs to present one of the paper or chapter in the class syllabus and lead the class discussion. Students will also work on a major project in groups of 1 to 2 students and deliver in phases. While it is not required, the ideal result of a major project is a complete research paper draft, submittable in a well-respected AI/Security/SE venue.

| Category | Percentage |
|---|---|
| Code Assignments | 10% |
| Paper/Chapter Assignments (Summary) | 25% |
| Class Presentations | 10% |
| Discussion Participations | 10% |
| Final Project (Write-up, code, and presentation) | 45% |

*Code Assignment (10% of the grade)*

There will be 2-3 code assignments in the first half of the course.

*Paper Assignment (25% of the grade)*

There will be paper assignments for each class. Students are required to write the paper summary and submit it before the class. The format and deadline for the assignments will be announced in class.

*Paper Presentation (10% of the grade)*

Students are required to sign-up and present one of major paper from the list of assigned papers.

*Class Discussion Participations (10 % of the grade)*

Since the course is discussion-based, participation in class discussion (and online forum) is required.

*Final Project (45 % of the grade)*

The final project is the most important component for the course. Students need to form a group of 1-2 and deliver materials in phases. Deliveries include write-ups, code, and presentations.

Your grade is independent of anyone else's grade in this class; that is, we do not grade on a curve. Everyone can get an A in this class.

The instructor reserves the right to adjust these criteria, e.g., so that 88% or higher represents an A, based on overall class performance.

## ATTENDANCE POLICY

This is a discussion-based senior/grad level class. Participation in class is absolutely required and counts as extra points up to 5% (in addition to the class discussion participations).

## TECHNOLOGY REQUIREMENTS

Course contents such as submission are delivered via the Blackboard learning management system (LMS). Ensure your UTEP e-mail account is working and that you have access to the Web and a stable web browser. Mozilla Firefox and Google Chrome are the most supported browsers for Blackboard; other browsers may cause complications with the LMS. When having technical difficulties, update your browser, clear your cache, or try switching to another browser. You will need to have or have access to a computer/laptop, a webcam, and a microphone. If you encounter technical difficulties beyond your scope of troubleshooting, please contact the Help Desk as they are trained specifically in assisting with technological needs of students.

## STANDARDS of CONDUCT

You are expected to conduct yourself in a professional and courteous manner, as prescribed by the Handbook of Operating Procedures: Student Conduct and Discipline. All graded work (except the final project with your classmate) is to be completed independently and should be unmistakably your own work, although you may discuss your work with others in a general way. You may not represent as your own work material that is transcribed or copied from another source, including persons, books, or Web pages. Plagiarism is a serious violation of university policy and will not be tolerated. All cases of suspected plagiarism will be reported to the Dean of Students for further review. You are welcome and encouraged to work together in learning the material. However, whatever you submit must be your own. In other words, cutting and pasting or copying verbatim from another source be it a classmate, an online source or even something that the TA/instructor showed you is strictly forbidden.

- The use of generative AI tools such as ChatGPT is permitted in this course for the following activities, which must be noted or cited: learning the concepts, finding alternatives to the proposed solutions by students. However, you should not use AI tools to ask for the exact questions. Students must cite any borrowed content sources to comply with all applicable citation guidelines, copyright law, and avoid plagiarism. Instances that violate these guidelines will be referred to the Office of Student Conduct and Conflict Resolution.
- Cite Your Sources: If you worked with someone on an assignment, or if your submission includes quotes from a book, a paper, or a web site, you should

clearly acknowledge the source. **Bottom line: feel free to use resources that are available to you as long as the use is reasonable, and you cite them in your submission**. However, copying answers directly or indirectly from solution manuals, web pages, or your peers is certainly forbidden.

- Inspiration is free: you may discuss homework assignments with anyone. You are especially encouraged to discuss in black board with your instructor and your classmates.
- Plagiarism is forbidden: the assignments and code that you turn in should be written entirely on your own. You should not need to consult sources beyond your textbook, class notes, posted lecture slides and notebooks, programming language documentation, and online sources for basic techniques. Copying/soliciting a solution to a problem from the internet or another classmate constitutes a violation of the course's collaboration policy and the honor code and will result in an F in the course and a trip to the honor council.
- Do not search for a solution online: You may not actively search for a solution to the problem from the internet. This includes posting to sources like StackExchange, Reddit, Chegg, etc.
- StackExchange Clarification: Searching for basic techniques in Python/Pandas/Numpy is totally fine. If you want to post and ask "How do I group by two columns, then do something, then group by a third column" that's fine. What you cannot do is post "Here's the problem my professor gave me. I need to convert Age in Earth years to Martian years and then predict the person's favorite color. Give me code!" That's cheating.
- When in doubt, ask: We have tried to lay down some rules and the spirit of the collaboration policy above. However, we cannot be comprehensive. If you have doubts about this policy or would like to discuss specific cases, please ask the instructor. If it has not been described above, you should discuss it with us first

Please also pay attention to the following netiquettes:

- Always consider audience. Remember that members of the class and the instructor will be reading any postings.

- Respect and courtesy must be provided to classmates and to instructor at all times. No harassment or inappropriate postings will be tolerated.
- Blackboard is not a public internet venue; all postings to it should be considered private and confidential. Whatever is posted on in these online spaces is intended for classmates and professor only. Please do not copy documents and paste them to a publicly accessible website, blog, or other space. If students wish to do so, they have the ethical obligation to first request the permission of the writer(s).

## ACCOMMODATIONS POLICY

The University is committed to providing reasonable accommodations and auxiliary services to students, staff, faculty, job applicants, applicants for admissions, and other beneficiaries of University programs, services and activities with documented disabilities in order to provide them with equal opportunities to participate in programs, services, and activities in compliance with sections 503 and 504 of the Rehabilitation Act of 1973, as amended, and the Americans with Disabilities Act (ADA) of 1990 and the Americans with Disabilities Act Amendments Act (ADAAA) of 2008. Reasonable accommodations will be made unless it is determined that doing so would cause undue hardship on the University. Students requesting an accommodation based on a disability must register with the UTEP Center for Accommodations and Support Services.

## State Law, Borrowed from American Association of University Professors (AAUP)

Texas Senate Bill 17, the recent law that outlaws diversity, equity, and inclusion programs at public colleges and universities in Texas, does not in any way affect content, instruction or discussion in a course at public colleges and universities in Texas. Expectations and academic freedom for teaching and class discussion have not been altered post-SB 17, and students should not feel the need to censor their speech pertaining to topics including race and racism, structural inequality, LGBTQ+ issues, or diversity, equity, and inclusion.