CIS6930

Trustworthy AI Systems

CRN 97083, Section 4, 3 Credit Hours

COURSE SYLLABUS

Semester: Fall 2024
Class Meeting Days: M, W
Class Meeting Time: 6:30 – 7:45 pm
Class Meeting Location: CHE 103
Instructor: Guangjing Wang
Office Location: BEH 311
Office Hours: Tuesday 2:00pm – 5:00pm
Email: guangjingwang@usf.edu

I. Welcome!

Welcome to *Trustworthy AI Systems*!

This course is more than just a deep dive into the principles of artificial intelligence; it is an exploration of the ethical, technical, and societal dimensions that define the future of technology. As we stand in a new era of AI, the need for systems that are not only intelligent but also reliable, ethical, and secure has never been more urgent.

In this course, you will develop not just technical expertise but also a mindset that embraces critical thinking, ethical reasoning, and a commitment to lifelong learning. You will explore how to design AI systems that people can trust—systems that are transparent, fair, and capable of making decisions that positively impact society.

By the end of this course, you will sharpen your problem-solving abilities by tackling real-world AI challenges. Your communication skills will be enhanced as you learn to explain complex technical concepts to diverse audiences. And you will develop a strong sense of equity and inclusion, understanding how to build AI systems that are fair and accessible to all.

Let us embark on this journey together to build a future where AI enhances the human experience. Welcome to the future of AI. Welcome to *Trustworthy AI Systems*.

II. University Course Description

- CIS 6930: Special Topics
- College of Engineering | Department of Computer Science and Engineering
- Credit hours: 3

III. Course Prerequisites

Required Prerequisites: None
The instructor will offer students some background material in class or exercises so students can get up to speed on the main material that are teaching.

- Python Programming: A strong understanding of Python is essential, as it will be the primary language used for implementing AI models and tools throughout the course. Familiarity with libraries such as NumPy, pandas, and matplotlib is expected.

- Machine Learning Fundamentals: Students should have prior coursework or experience in machine learning. This includes understanding supervised and unsupervised learning, model evaluation, and basic algorithms such as linear regression, decision trees, and clustering.

- Deep Learning: A working knowledge of deep learning is necessary. This includes familiarity with neural networks, convolutional networks, and frameworks like TensorFlow or PyTorch. Students should be comfortable with building, training, and evaluating deep learning models.

- Basic Probability and Statistics: Understanding probability theory, statistical methods, and their application in machine learning is important. This includes knowledge of distributions, statistical inference, and hypothesis testing.

- Ethics in Technology: While not mandatory, prior exposure to courses or discussions on ethics in technology or AI will be beneficial. This course will delve into ethical considerations, so familiarity with basic ethical frameworks will be helpful.

IV. Course Purpose

This course delves into the critical aspects of building AI systems that are not only powerful but also trustworthy, ethical, and aligned with societal values. As AI continues to permeate every aspect of our lives—from healthcare and finance to autonomous vehicles and social media—ensuring that these systems operate transparently, fairly, and safely is paramount.

In Trustworthy AI Systems, we explore the design, development, and deployment of AI technologies through the lens of trustworthiness. This includes an in-depth examination of key principles such as fairness, accountability, transparency, privacy, and security. The course also covers practical techniques for mitigating bias, ensuring robustness, and building interpretability into AI models. This course is particularly relevant in today's landscape where issues such as algorithmic bias, data privacy, and AI explainability are no longer just theoretical concerns but pressing real-world challenges.

Trustworthy AI Systems is designed as an advanced elective for graduate students in computer science, particularly those specializing in artificial intelligence, data science, or software engineering. It serves as a critical component of the AI curriculum, bridging the gap between technical proficiency and ethical responsibility. This course complements core AI and machine learning courses by providing the ethical and societal context necessary for building AI systems that can be deployed in real-world settings.

V. Course Format

- Lectures: Each class will typically begin with a brief lecture (30-50 minutes) that introduces key concepts, theories, and techniques related to the topic of the day. These lectures are designed to provide you with the foundational knowledge needed to engage in deeper discussions and activities.

- Discussions/Presentation: Following the lecture, we will move into class discussions or Presentations. These will be a significant part of our sessions, as they allow us to explore the ethical and societal implications of AI in a collaborative environment. You will be expected to actively participate, sharing your insights and responding to your peers. The discussions will often be centered around case studies, research papers, or current events in AI.

- Collaborative Learning: Many sessions will include group work or collaborative projects. You will work with your classmates to solve problems, analyze case studies, or develop components of your capstone project. This collaborative approach mirrors the real-world settings where you will often work as part of a team to develop AI systems.

- Hands-On Activities: Practical, hands-on activities will be incorporated into our sessions to reinforce the technical aspects of the course. These may involve coding exercises, model evaluation tasks, or bias detection assignments. These activities are designed to help you apply the theoretical knowledge gained from lectures and discussions.

The structure of our sessions is designed to balance theory and practice. By combining lectures with discussions, collaborative learning, and hands-on activities, you will develop both a deep understanding of the concepts and the ability to apply them in practical contexts. Discussions and collaborative work are central to this course because the subject matter—trustworthy AI—benefits from diverse perspectives and critical thinking. By engaging actively, you will not only enhance your own understanding but also contribute to the learning experience of your peers.

VI. Course Objectives

By the end of this course, students will be able to:

- Design and Implement Ethical AI Systems: Develop AI models that prioritize ethical considerations, including fairness, transparency, and accountability. Students will learn to integrate ethical decision-making frameworks into the AI development lifecycle, ensuring that their systems adhere to the highest ethical standards.

- Mitigate Algorithmic Bias: Identify and address sources of bias in data and algorithms. Students will learn techniques for detecting, measuring, and mitigating bias to create more equitable AI systems that serve diverse populations fairly.

- Enhance AI Interpretability: Implement methods to make AI models more interpretable and explainable. Students will gain skills in creating models whose decisions can be understood and trusted by non-experts, which is critical in high-stakes applications such as healthcare.

- Ensure Robustness and Security: Design AI systems that are resilient to adversarial attacks and capable of maintaining performance in the face of unexpected inputs or changing environments. Students will learn about techniques for building robust models and securing AI systems against malicious threats.

- Apply Privacy-Preserving Techniques: Understand and implement privacy-preserving techniques such as differential privacy and federated learning. Students will develop the ability to protect user data while still enabling powerful AI-driven insights.

- Conduct Ethical AI Audits: Perform comprehensive audits of AI systems to evaluate their ethical and technical robustness. Students will learn to assess AI systems for compliance with ethical guidelines and legal standards, preparing them for roles in AI governance and compliance.

- Communicate Ethical and Technical Concepts: Develop the ability to clearly communicate complex ethical and technical concepts related to AI to both technical and non-technical audiences. Students will practice articulating the implications of AI decisions in a way that is accessible and understandable to diverse stakeholders.

- Develop a Trustworthy AI Project: Undertake a capstone project where students design, implement, and evaluate an AI system with trustworthiness as a core criterion. This project will allow students to apply the skills and knowledge gained throughout the course in a practical, real-world context.

VII. Required Texts and/or Readings and Course Materials

Top-tier conference papers from ICML, NeurIPS, ICLR, IEEE Security and Privacy, ACM CCS, Usenix Security, and NDSS in recent 3 years.

VIII. How to Succeed in this Course

- Read and Reflect: The readings in this course will cover both technical papers and ethical discussions. Engage with these materials by taking notes and asking yourself how the concepts apply to real-world scenarios.

- Participate in Discussions: The topics in this course, particularly those around ethics and fairness, benefit greatly from diverse perspectives. Participate actively in class discussions, sharing your views and considering others' viewpoints.

- Think Beyond the Code: This course is as much about ethics as it is about technology. Always consider the societal impact of the systems you design. Practice thinking about how AI models could affect different populations and what steps you can take to mitigate negative consequences.

- Case Studies and Real-World Examples: Pay special attention to the case studies discussed in class. Analyzing real-world failures and successes will help you understand the importance of building trustworthy AI.

- Hands-on Practice: The technical assignments will often require you to apply concepts such as bias detection or model interpretability. Treat these assignments as opportunities to deepen your understanding, not just as tasks to complete.

- Group Projects: Collaboration is key in this course, especially for group projects. Learn to communicate effectively with your peers, leveraging each other's strengths. Share resources, discuss different approaches, and learn from the diverse skills within your group.

- Peer Feedback: Do not shy away from giving and receiving feedback. Constructive criticism can significantly enhance your work and understanding of the course material.

- Time Management: This course will have multiple components—readings, discussions, technical assignments, and a capstone project. Plan your schedule in advance to ensure that you have ample time for each part.

- Documentation: Keep a well-organized record of your code, notes, and reflections. Proper documentation will not only help you during this course but also serve as a valuable resource in your future work.

- Office Hours and Forums: Take advantage of office hours and discussion forums. Whether you are struggling with a concept or just want to delve deeper into a topic, these are opportunities to get personalized guidance.

- Online Resources: If you find yourself struggling with a particular topic, there are numerous online resources available. Websites such as Stack Overflow and Kaggle can provide immediate help with coding issues, while ArXiv is a great place to find academic papers on AI.

IX. Academic Continuity

In the event that the university needs to transition to remote instruction due to unforeseen circumstances, such as a pandemic or natural disaster, our course will seamlessly continue with adjustments to ensure that you can continue your learning without interruption. We will hold live, synchronous sessions during our regular class times using Microsoft Teams. Attendance will be required, as these sessions will replace in-person lectures and discussions. Only if we switch to online and if you are unable to attend a live session due to time zone differences or other legitimate reasons, the sessions will be recorded and made available in Canvas. Please inform me in advance if you will not be able to attend a session live.

X. Communication

- Canvas Mail: The primary mode of communication will be through Canvas Mail. Please check your Canvas inbox regularly for course-related updates, feedback on assignments, and other important information.

- Direct Email: For more urgent matters or personal issues, you can reach me directly via email at [guangjingwang@usf.edu]. Please use your university email address and include [CIS 6930] when contacting me to ensure that your message is recognized and receives a prompt response.

- Class Announcements: All major course updates, reminders, and important information will be posted in the "Announcements" section on Canvas. Please make sure to enable notifications for announcements so you do not miss any critical updates.

- Course website: https://guangjing.wang/CIS6930/ (Tentative)

XI. Grading Scale

Grading Scale (%)

94 – 100	A	74 – 76	C
90 – 93	A-	70 – 73	C-
87 – 89	B+	67 – 69	D+
84 – 86	B	64 – 66	D
80 – 83	B-	60 – 63	D-
77 – 79	C+	0 – 59	F

XII. Grade Categories and Weights

Graded Items	**Percent of Final Grade**
Project-Midterm (Code)	20%
Project-Final (Code)	24%
Essay	25%
Homework	21%
Short Presentation	10%

- Text generation tools such as ChatGPT are prohibited from being used in **Essay and Homework**.
- Late Homework and Missing Short Presentation are not considered in final grade.
- For late projects and essays, the following penalties will apply:
  - Submitted one day late: 20% reduction (original score * 0.8)
  - Submitted two days late: 50% reduction (original score * 0.5)
  - Submitted three days late or more: No credit will be given (original score * 0)

Project and Essay Assignments

Two to three students will form a group, choose a project topic, and complete a project:

- Choose one AI system topic for implementation (Project-Midterm)
- Choose one of the topics to evaluate AI system: security, privacy, explainability, or bias assessment (Project-Final)

Each group will need to upload the source code to GitHub to track the progress

Each group will need to write the project essay in at least **6 content pages** (excluding references) following **NeurIPS 2024 LaTeX style file**

Participation and Engagement

Each student will have an opportunity to give a short presentation about the state-of-the-art research work that is relevant to the topic of the project that he/she/they choose. The other students are expected to ask questions and provide feedback on the quality of the short presentation. The practice of the process will be used as the evaluation of participation and engagement.

First day attendance

Please check the information here: https://www.usf.edu/student-affairs/new-student/documents/pdf/rockys-resources-how-to-keep-your-courses.pdf

This course will use the "first day attendance" assignment on the Canvas System to record your first day attendance.

XIII. Course Schedule

Date	Work Due Before Class	Topics to be Discussed in Class
08/26	First day of class; first day attendance	Overview of Trustworthy AI Systems
	AI Models and Systems	
08/28	**No assignments are due**	CV - Image Classification **Homework 1 Release – Paper Review**
09/02	**Holiday (No Class)**	**Labor Day holiday: no classes**
09/04	No assignments are due **Group Members Confirmed**	CV - Image Segmentation
09/09	No assignments are due	CV - Neural Style Transfer
09/11	**Homework 1 Due**	CV – Diffusion Mdels
09/16	No assignments are due	Audio - Speech Recognition
09/18	No assignments are due	Audio – Voice Conversion **Homework 2 Release – Artifact Review**
09/23	No assignments are due	Pretrained Foundation Model
09/25	No assignments are due	Large Language Model Agent
09/30	**Homework 2 Due**	Mobile Sensing and Multimodal data
	Topics on Trustworthiness	
10/02	No assignments are due	Security: Adversarial Attacks
10/07	No assignments are due	Individual Short Presentations
10/09	No assignments are due	Individual Short Presentations
10/14	No assignments are due	Individual Short Presentations
10/16	**Midterm Project Due**	Individual Short Presentations
10/21	No assignments are due	Security: Backdoor Attacks **Homework 3 Release – Paper Review**
10/23	No assignments are due	Privacy: Federated Learning
10/28	No assignments are due	Privacy: Differential Privacy
10/30	**Homework 3 Due**	Jailbreaking and Hallucinations in LLMs
11/04	No assignments are due	Disinformation in the Era of AI

11/06	No assignments are due	Building Transparent and Explainable AI
11/11	**Holiday (No Class)**	**Veterans Day holiday; no classes**
11/13	No assignments are due	Accountability and Watermarking
11/18	No assignments are due	Fairness and Bias in AI Systems
11/20	No assignments are due	Ethical Frameworks for AI
11/25	No assignments are due	Group Presentation
11/27	No assignments are due	Group Presentation
12/02	**Final Project Due**	Group Presentation
12/04	**Essay Due**	Group Presentation

* **Note: The schedule is subject to revision**

XIV. USF Core Syllabus Policies

USF has a set of central policies related to student recording class sessions, academic integrity and grievances, student accessibility services, academic disruption, religious observances, academic continuity, food insecurity, pregnancy and related conditions, and sexual harassment that **apply to all courses at USF**. Be sure to review these online: usf.edu/provost/faculty-success/resources-policies-forms/core-syllabus-policy-statements.aspx

XV. Course Policies: Grades

Medical Excuses: Students should not attend class if they are ill, particularly if they have fever and/or gastrointestinal symptoms and/or respiratory symptoms such as a sneezing, runny nose, sore throat or coughing. Students experiencing any of these symptoms should contact immediately the Student Health Services (813-974-2331) on the Sarasota-Mantatee and Tampa campus or the Wellness Center (727-873-4422) on the St. Petersburg campus for appropriate medical guidance and to obtain a verification of care letter. Students may turn to other health providers as well. **To be approved for missed classes, late assignments or missed examinations a verification of care letter must be presented by the student to the faculty member upon return to class.**

Extra Credit Policy: There two extra credit opportunities: (i) building a wiki of course content (any format). (ii) End of semester student evaluations (see below for details). The score of extra credit is granted at the instructor's discretion, and the additional points are added to the "Essay" portion of the semester grade. You cannot earn higher than 100% on the "Essay" portion of the grade; any points over 100% are not counted. The deadline for the extra credit is the same as the essay due date.

End of Semester Student Evaluations: All classes at USF make use of an online system for students to provide feedback to the University regarding the course. These surveys will be made available at the end of the semester, and the University will notify you by email when the response window opens. Your participation is highly encouraged and valued.

Grades of "Incomplete": For graduate courses: An Incomplete grade ("I") is exceptional and granted at the instructor’s discretion only when students are unable to complete course requirements due to illness or other circumstances beyond their control. The course instructor and student must complete and sign the "I" Grade Contract Form that describes the work to be completed, the date it is due, and the grade the student would earn factoring in a zero for all incomplete assignments. The due date can be negotiated and extended by student/instructor as long as it does not exceed two semesters for undergraduate courses and one semester for graduate courses from the original date grades were due for that course. An "I" grade not cleared within the two semesters for undergraduate courses and one semester for graduate courses (including summer semester) will revert to the grade noted on the contract.

Campus Free Expression: *It is fundamental to the University of South Florida's mission to support an environment where divergent ideas, theories, and philosophies can be openly exchanged and critically evaluated. Consistent with these principles, this course may involve discussion of ideas that you find uncomfortable, disagreeable, or even offensive. In the instructional setting, ideas are intended to be presented in an objective manner and not as an endorsement of what you should personally believe. "Objective" means that the idea(s) presented can be tested by critical peer review and rigorous debate, and that the idea(s) is supported by credible research. In this course you may be asked to engage with complex ideas and to demonstrate an understanding of the ideas. Understanding and engaging with an idea does not require you to believe it or to agree with it.*

Make-up Exams Policy: If a student cannot be present for an examination for a valid reason (validity to be determined by the instructor), a make-up exam will be given only if the student has notified the instructor in advance that s/he cannot be present for the exam. Make-up exams are given at the convenience of the instructor usually during office hours.

Group Work Policy: Everyone must take part in a group project. All members of a group will receive the same score; that is, the project is assessed and everyone receives this score. However, that number is only 90% of your grade for this project. The final 10% is individual, and refers to your teamwork. Every person in the group will provide the instructor with a suggested grade for every other member of the group, and the instructor will assign a grade that is informed by those suggestions. Also, everyone must take part in a group essay (see essay assignments below). The grading criteria are the same as the group project. Once formed, groups cannot be altered or switched, except for reasons of extended hospitalization.

XVI. Course Policies: Student Expectations

Health and Wellness: Your health is a priority at the University of South Florida. We encourage members of our community to look out for each other and to reach out for help if someone is in need. If you or someone you know is in distress, please make a referral at www.usf.edu/sos so that the Student Outreach & Support can contact and provide helpful resources to the student in distress. A 24-hour licensed mental healthcare professional, offered through the counseling center, is available by phone at 813-974-2831, option 3. Please remember that asking for help is a sign of strength. In case of emergency, please dial 9-1-1.

Title IX Policy: Title IX provides federal protections for discrimination based on sex, which includes discrimination based on pregnancy, sexual harassment, and interpersonal violence. In an effort to provide support and equal access, USF has designated all faculty (TA, Adjunct, etc.) as Responsible Employees, who are required to report any disclosures of sexual harassment, sexual violence, relationship violence or stalking. The Title IX Office makes every effort, when safe to do so, to reach out and provide resources and accommodations, and to discuss possible options for resolution. Anyone wishing to make a Title IX report or seeking accommodations may do so online, in person, via phone, or email to the Title IX Office. For information about Title IX or for a full list of resources please visit: https://www.usf.edu/title-ix/gethelp/resources.aspx. If you are unsure what to do, please contact Victim Advocacy – a confidential resource that can review all your options – at 813-974-5756 or va@admin.usf.edu.

Turnitin.com: In this course, turnitin.com will be utilized. Turnitin is an automated system which instructors may use to quickly and easily compare each student's assignment with billions of web sites, as well as an enormous database of student papers that grows with each submission. Accordingly, you will be expected to submit all assignments in electronic format. After the assignment is processed, as instructor I receive a report from turnitin.com that states if and how another author's work was used in the assignment. For a more detailed look at this process visit http://www.turnitin.com. Essays are due at turnitin.com the same day as in class.

Netiquette Guidelines

1. Act professionally in the way you communicate. Treat your instructors and peers with respect, the same way you would do in a face-to-face environment. Respect other people's ideas and be constructive when explaining your views about points you may not agree with.
2. Be sensitive. Be respectful and sensitive when sharing your ideas and opinions. There will be people in your class with different linguistic backgrounds, political and religious beliefs or other general differences.
3. Proofread and check spelling. Doing this before sending an email or posting a thread on a discussion board will allow you to make sure your message is clear and thoughtful. Avoid the use of all capital letters, it can be perceived as if you are shouting, and it is more difficult to read.
4. Keep your communications focused and stay on topic. Complete your ideas before changing the subject. By keeping the message on focus you allow the readers to easily get your idea or answers they are looking for.
5. Be clear with your message. Avoid using humor or sarcasm. Since people can't see your expressions or hear your tone of voice, meaning can be misinterpreted.

XVII. Learning Support and Campus Offices

Academic Accommodations

Students with disabilities are responsible for registering with Student Accessibility Services (SAS) in order to receive academic accommodations. For additional information about academic accommodations and resources, you can visit the SAS website.
SAS website for the Tampa and Sarasota-Manatee campuses.
SAS website for the St. Pete campus.

Academic Support Services

The USF Office of Student Success coordinates and promotes university-wide efforts to enhance undergraduate and graduate student success. For a comprehensive list of academic support services available to all USF students, please visit the Office of Student Success website.

Canvas Technical Support

If you have technical difficulties in Canvas, you can find access to the Canvas guides and video resources in the "Canvas Help" page on the homepage of your Canvas course. You can also contact the help desk by calling 813-974-1222 in Tampa or emailing help@usf.edu.
IT website for the Tampa campus.
IT website for the St. Pete campus.
IT website for the Sarasota-Manatee campus.

Center for Victim Advocacy

The Center for Victim Advocacy empowers survivors of crime, violence, or abuse by promoting the restoration of decision making, by advocating for their rights, and by offering support and resources. Contact information is available online.

Counseling Center

The Counseling Center promotes the wellbeing of the campus community by providing culturally sensitive counseling, consultation, prevention, and training that enhances student academic and personal success. Contact information is available online.
Counseling Center website for the Tampa campus.
Counseling Center website for the St. Pete campus.
Counseling Center website for the Sarasota-Manatee campus.

Writing Studio

The Writing Studio is a free resource for USF undergraduate and graduate students. At the Writing Studio, a trained writing consultant will work individually with you, at any point in the writing process from brainstorming to editing. Appointments are recommended, but not required. For more information or to make an appointment, email: writingstudio@usf.edu.
Writing studio website for the Tampa campus.
Writing studio website for the St. Pete campus.
Writing studio website for the Sarasota-Manatee campus.

XVIII. Important Dates to Remember

All the dates and assignments are tentative and can be changed at the discretion of the professor. For important USF dates, see the Academic Calendar at http://www.usf.edu/registrar/calendars/

Drop/Add Deadline:	*Fri, Aug 30, 2024*
Labor Day Holiday:	*Mon, Sept 2, 2024*
Mid-term Project Due:	*Wed, Oct 16, 2024*
Withdrawal Deadline:	*Sat, Nov 2, 2024*
Veteran's Day Holiday:	*Mon, Nov 11, 2024*
Thanksgiving Holiday:	*Thurs, Nov 28, & Fri, Nov 29, 2024*
Final Project Due:	*Mon, Dec 02 2024*
Essay Due Date:	*Wed, Dec 04 2024*

As part of this course, each student will have the opportunity to give a short presentation on a state-of-the-art research work that is relevant to the topic of their chosen project. This component is designed to enhance your understanding of current developments in the field, improve your presentation skills, and foster a collaborative learning environment.

Other Information: The College of Engineering has worked diligently with Student Counseling Services to appoint our Licensed Mental Health Counselor dedicated to making mental health services more readily accessible to our students. Michelle Morton-Tunstall will continue to provide clinical counseling to our students both here in ENB and at the central counseling center in the Fall semester. Specifically, Michelle will be available to our students at two locations: our College (ENB201) and the central counseling center. This dual-location arrangement is designed to meet the diverse needs of our students, considering factors such as privacy, convenience, and accessibility. Initially, Michelle will be on-site at our College every Wednesday and Thursday from 2-5 p.m., starting from April 1st through May 3rd, while we study the appropriate ratio of demand for meeting here and central unit location. She is currently undergoing training at the central counseling center.