Rensselaer

Course Information

Course Name:	Trustworthy Machine Learning
Course Number:	CSCI 6962/CSCI 4180
Semester and Year:	Fall 2025
Meeting Times:	MTh 2pm–3:50pm
Credit Hours:	4
Classroom:	Pitts 5216
Prerequisite:	CSCI 2210 AND (CSCI 4100 OR CSCI 4390)
Course webpage	`https://www.cs.rpi.edu/~gittea/teaching/fall2025/tml.html`
Discussion page	`https://piazza.com/rpi/fall2025/csci69624180`

Instructor

Professor Alex Gittens
Office Location: Lally 316
Office Hours: MTh 4pm–5pm

email: `gittea@rpi.edu`
Office Phone: (518) 276-6476

Course Description:

In today's world it is no longer sufficient to consider the traditional metrics of accuracy when judging the trustworthiness of systems built using machine learning. It is important to also consider fairness, robustness, privacy, alignment, and the attack surfaces of machine learning algorithms. This seminar course introduces these topics to students who already have a basic understanding of machine learning.

In this course, you will explore fundamental questions and learn tools and methods to measure and ensure these aspects of trustworthiness in machine learning. We will delve into both seminal and recent papers to examine the growing body of research on trustworthy machine learning.

Course Structure

We will cover five broad areas (alignment of LLMs, privacy, robustness, fairness, and attack surfaces) in this course. The course consists of a number of lectures delivered by the instructor and at least one seminar by each student.

- **Introductory lectures:** The instructor will provide an introductory lecture on each of the five areas. Besides these, the instructor will provide two introductory lectures at the beginning of the semester to review the foundations required for the course.
- **Seminars:** The instructor will provide a list of papers (about 40) covering different aspects of trustworthy machine learning. Each student will select at least one topic and present at least one paper associated with that topic as a lecture. Each seminar is expected to last about 45 minutes, followed by 20 minutes of discussions.
- **Complementary lectures:** Time-permitting and when needed, the instructor will provide short lectures on important topics complementary to the seminars.
- **Project presentations:** Each student will propose a research project or pedagogical project. The graded deliverables include the relevant code and a short poster presentation detailing the outcome of the project (5-10 minutes). Progress will be assessed partway through the course, and the final presentations will be given at the end of the semester.

Course Goals / Objectives:

On completion of this course, students should be sufficiently familiar with the theoretical basis, formal quantification, and analysis of five facets of trustworthy ML. The core emphasis of the course will be on critically reading and discussing the research literature in trustworthy ML. The major questions in the emerging research domain of trustworthy ML that we address in the course include the following:

Part I – Alignment of LLMs: How do we measure the extent to which LLMs respect human norms such as avoiding biased, toxic, unethical or illegal, or privacy-sensitive content? How can we train them to be more aligned, and what are the limitations of alignment?

Part II – Attack Models: How are attacks categorized and designed? How does their impact on the data differ from unstructured random perturbations? What processes in the machine learning pipeline are vulnerable to attacks?

Part III – Privacy and Confidentiality: Can we trust machine learning frameworks to have access to personal data? Can we trust the models not to reveal personal information or sensitive decision rules? How can we quantify disclosure risk?

Part IV – Robustness: In settings where training data is noisy or adversarially crafted, can we trust the algorithms to learn robust decision rules? Can we trust them to make correct predictions on adversarial or noisy testing data?

Part V – Algorithmic Fairness: Bias affecting some groups in the population underlying a data set can

arise from both a lack of representation in data but also poor choices of learning algorithms. Can we build trustworthy algorithms that remove disparities and provide fair predictions for all groups? How do we quantify fairness?

Course Texts:

There is no formal textbook for the course. The instructor will provide a reading list of readily accessible research papers.

Student Learning Outcomes (CSCI 4180):

Students who successfully complete this course will

1. Demonstrate an ability to analyze and algorithmically induce alignment of large language models. Specifically, the students will demonstrate facility in the operational measurement of alignment and familiarity with common approaches to LLM alignment.
2. Demonstrate an ability to analyze privacy and confidentiality of the data used in training ML algorithms. Specifically, the students will demonstrate proficiency in the analysis of privacy-preserving learning, in both shared-memory and distributed settings.
3. Demonstrate an ability to analyze the vulnerability of ML algorithms to non-adversarial disruptions (noise) and adversarial disruptions (attacks). Specifically, the students will demonstrate proficiency in the analysis of data inference attacks, model inference attacks, and testing/verification.
4. Demonstrate an ability to analyze the fairness of ML algorithms. Specifically, the students will demonstrate facility in the use of fairness measures and mechanisms ensuring fair learning.

Student Learning Outcomes (CSCI 6962):

Students who successfully complete this course will

1. Demonstrate an ability to analyze and algorithmically induce alignment of large language models. Specifically, the students will demonstrate facility in the operational measurement of alignment and familiarity with common approaches to LLM alignment.
2. Demonstrate an ability to analyze privacy and confidentiality of the data used in training ML algorithms. Specifically, the students will demonstrate proficiency in the analysis of privacy-preserving learning, in both shared-memory and distributed settings.
3. Demonstrate an ability to analyze the vulnerability of ML algorithms to non-adversarial disruptions (noise) and adversarial disruptions (attacks). Specifically, the students will demonstrate proficiency in the analysis of data inference attacks, model inference attacks, and testing/verification.
4. Demonstrate an ability to analyze the fairness of ML algorithms. Specifically, the students will

demonstrate facility in the use of fairness measures and mechanisms ensuring fair learning.

5. Demonstrate proficiency in the mathematical techniques for formalizing and analyzing privacy.
6. Demonstrate proficiency in the mathematical techniques for formalizing and analyzing adversarial robustness.

Course Assessment Measures (CSCI 4180):

Assessment	Date	Weight	Learning outcome
Seminar presentation	at least once during the semester	60 %	1,2,3,4
Final project		20 %	1,2,3,4
Proposal 20 %	TBD		
Progress report 25 %	TBD		
Deliverables 20 %	TBD		
Presentation 35 %	TBD		
Participation	throughout the semester	20 %	1,2,3,4
attendance			
paper critiques			
class discussion			

Course Assessment Measures (CSCI 6962):

Assessment	Date	Weight	Learning outcome
Seminar presentation	at least once during the semester	60 %	1,2,3,4
Final project		20 %	1,2,3,4,5,6
Proposal 20 %	TBD		
Progress report 25 %	TBD		
Deliverables 20 %	TBD		
Presentation 35 %	TBD		
Participation	throughout the semester	20 %	1,2,3,4
attendance			
paper critiques			
class discussion			

Academic Integrity:

Student-Professor relationships are built on trust. Students must trust that professors have made appropriate decisions about the structure and content of the courses they teach, and professors must trust that the assignments that students turn in represent their own work. Acts that violate this trust undermine the educational process. I take academic integrity very seriously. Rensselaer is a community in which personal responsibility, honesty, fairness, respect, and mutual trust are maintained. You are expected to practice the highest possible standards of academic integrity. Any deviation from this

expectation will result in a minimum academic penalty commensurate to the violation. In this class, all reports and presentation material that are turned in for a grade must represent the student's own work. In cases where unofficial help was received, or significant teamwork was involved, a notation on the assignment should indicate your collaboration.

Use of AI. Generative AI tools (e.g., ChatGPT, Claude, Gemini) may be used to assist your learning in this course, but they may not be used to produce deliverables such as presentations or your final project report unless explicitly stated otherwise. The goals of this course are to engage deeply with research papers, develop your ability to analyze and communicate technical material, and gain experience in independent research. Submitting AI-generated work in place of your own undermines these goals and will be considered academic misconduct. You are encouraged to use generative AI responsibly: for example, to clarify difficult concepts, brainstorm project ideas, or get feedback on your own drafts. If you are unsure whether a specific use is appropriate, please ask me in advance.

Any violation of the academic integrity policy will result in a 0 grade for the related evaluation. Repeated violations will result in the F grade. For any case of academic dishonesty, a report will be filed to the Dean of Graduate Studies or the Dean of Students, as appropriate. The Rensselaer Handbook of Student Rights and Responsibilities defines various forms of Academic Dishonesty; you should make yourself familiar with these. If you have any question concerning this policy before submitting an assignment, please ask for clarification.

Diversity, Equity, and Inclusion

I strive to create a classroom in which everybody will be treated with respect, and I celebrate individuals of all backgrounds, ethnicities, national origins, genders, gender identities, gender expressions, sexual orientations, beliefs, religious affiliations, ability, and other visible and invisible differences. I am dedicated to helping each of you achieve all that you can in this class. If in lectures or smaller interactions, unintentionally I use language that causes discomfort or offense, please contact me to help me understand and avoid making the same mistake again.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let your instructor know at your earliest convenience. Some aspects of this course may be modified to facilitate your participation and progress. As soon as you make us aware of your needs, we can work with the Disability Services for Students office, reachable at https://studenthealth.rpi.edu/disabilityservices, to help us determine appropriate academic accommodations. Any information you provide is private and confidential and will be treated as such.