Trustworthy Machine Learning

CIS 6261 Section 1XYZ
Class Periods: M,W,F | Period 6 (12:50 PM - 1:40 PM)
Location: MALA 1000
Academic Term: Fall 2025

Instructor:
Vincent Bindschaedler
vbindsch (at) cise (dot) ufl (dot) edu

Office Hours: (before class) M | 12:00 PM – 12:45 PM [Online] or by appointment.

Teaching Assistant:
Please contact through the Canvas website

- Amal Hashky

Course Description
Machine Learning (ML) is increasingly integrated into our daily lives and promises advances in many applications domains including autonomous driving, facial recognition, and medical diagnosis. At the same time, machine learning techniques are surprisingly brittle and easy to misuse or abuse which highlights the potential dangers of this technology. Complex models can be fooled by tiny perturbations of their inputs; they can unintentionally memorize their training data; and they make decisions that are often inexplicable.

Students will read, analyze, and discuss recent research on trustworthy machine learning and undertake a semester-long research project.

Credit Hours: 3

Course Pre-Requisites / Co-Requisites
A course in "programming fundamentals" (e.g., COP 3502/COP 3503 or similar) is required. Some assignments involve Python programming.

No formal pre-requisites for machine learning. But familiarity with machine learning concepts is a plus.

Course Objectives
Students will learn about foundational concepts at the intersection of machine learning with security and privacy and acquire a firm grasp on recent developments in this area.

By the end of the semester, successful students will:
(1) be able to identify, classify, and explain threats on machine learning models and defense methods to train or make models robust;
(2) be able to present and critically analyze research papers in this space, and
(3) have demonstrated the ability to formulate a research idea or hypothesis and gather evidence to support it through their course project.

Materials and Supply Fees
None

Required Textbooks and Software
No required textbooks. PDFs of reading material and academic papers will be provided.
Required software and libraries: Python 3, NumPy, PyTorch.

Recommended Materials
None

Required Computer
Recommended Computer Specifications: https://it.ufl.edu/get-help/student-computer-recommendations/
HWCOE Computer Requirements: https://www.eng.ufl.edu/students/advising/fall-semester-checklist/computer-requirements/
The computer needs to be able to run Python3 and jupyter notebook in an environment with the dependencies stated above are installed.

(Tentative) Course Schedule
Week 1: Syllabus, Course Overview, Introduction
Week 2: Background & Refresher: ML & probability and statistics
Week 3: Robustness & Adversarial ML
Week 4: Adversarial Examples / Discussion #1
Week 5: Adversarial Robustness (2) / Discussion #2
Week 6: Poisoning & Attacks on Models / Discussion #3
Week 7: Data Privacy & Privacy Metrics / Discussion #4
Week 8: Privacy Attacks / Discussion #5
Week 9: Differential Privacy
Week 10: Differentially Private ML / Discussion #6
Week 11: Privacy Auditing / Discussion #7
Week 12: Explainable & Interpretable ML / Discussion #8
Week 13: Explainable & Interpretable ML (2) / Discussion #9
Week 14: Safety & Alignment / Discussion #10
Week 15: Review / Project

Important Dates
- ***12/3/2025*** ***Project Due***
- ***12/3/2025*** ***Final Exam (in class)***

Attendance Policy, Class Expectations, and Make-Up Policy
Attendance is ***strongly recommended*** but not mandatory. Due to the course format, students who miss many lectures will be at a ***significant*** disadvantage.

Students are expected to have done the reading before class and actively participate during lectures and discussions (e.g., by asking questions or by volunteering their opinions).

Students will be assigned written, hands-on assignments related to course topics and the course research project. Assignments will be announced in class and will be handled through the E-learning platform (elearning.ufl.edu). Assignments turned in late will incur a lateness penalty of 15% per day, up to a maximum of 3 days (after which the grade will be 0).

Requirements for class attendance and make-up exams, assignments, and other work in this course are consistent with university policies. Click here to read the university attendance policies:
https://catalog.ufl.edu/UGRD/academic-regulations/attendance-policies/

Instruction Format & Evaluation of Grades

Instruction format will be a blend of traditional lecture-style instruction and student-led seminar-style learning through paper reading and discussion. Students will be expected to read several research articles every week and discuss them in class. The research project will require that students execute research alone or in a group.

Students will be evaluated based on the following breakdown:

Assignment	Total Points	Percentage of Final Grade
Course Research Project	100	20%
Homework & Assignments	100 each	20%
Paper Discussions & Reviews	100 each	20%
Final Exam	100	25%
Quizzes	100	15%

Grading Policy

Percent	Grade	Grade Points
93.0 - 100	A	4.00
88.0 – 92.9	A-	3.67
83.0 - 87.9	B+	3.33
78.0 - 82.9	B	3.00
74.0 - 77.9	B-	2.67
70.0 - 73.9	C+	2.33
66.0 – 69.9	C	2.00
62.0 - 65.9	C-	1.67
58.0 - 61.9	D+	1.33
54.0 – 57.9	D	1.00
50.0 - 53.9	D-	0.67
0 - 49.9	E	0.00

Academic Integrity

Students are required to follow the university guidelines on academic conduct and the student honor code (https://sccr.dso.ufl.edu/policies/student-honor-code-student-conduct-code/) at all times. Students failing to meet these standards will be reported to the Dean of Students, ***which can result in the student receiving an 'E' for the semester***. In particular, students are explicitly forbidden from copying anything off of the Internet (e.g., source code, text, slides) without proper attribution or citation. This includes *unauthorized use of AI tools to produce text or code*. Students are also forbidden from copying code/answers from each other for the purposes of completing any assignment.

CISE Department Academic Integrity Policy

Academic integrity violations (i.e., cheating, plagiarism) will be reported to SCCR! The CISE department policy for such offenses is a course grade of E. But additional sanctions may be imposed by SCCR.

Reminder of the Honor Pledge: On all work submitted for credit by Students at the University of Florida, the following pledge is either required or implied: "On my honor, I have neither given nor received unauthorized aid in doing this assignment."

Academic Policies & Resources

To support consistent and accessible communication of university-wide student resources, instructors must include this link to academic policies and campus resources: https://go.ufl.edu/syllabuspolicies. Instructor-specific guidelines for courses must accommodate these policies.

Commitment to a Positive Learning Environment

The Herbert Wertheim College of Engineering values varied perspectives and lived experiences within our community and is committed to supporting the University's core values.

If you feel like your performance in class is being impacted by discrimination or harassment of any kind, please contact your instructor or any of the following:

- Your academic advisor or Graduate Coordinator
- HWCOE Human Resources, 352-392-0904, student-support-hr@eng.ufl.edu
- Pam Dickrell, Associate Dean of Student Affairs, 352-392-2177, pld@ufl.edu