New Jersey Institute of Technology
Ying Wu College of Computing
Computer Science Department

Explainable AI

CS 485/698
Credits: 3
Mode: Face-to-Face
Location: Mechanical Engineering Center 221
Lecture: Tuesday/Friday, 2:30-3:50pm

Instructor: Kieran Murphy
Email: kieran.murphy@njit.edu
Office: 4315 GITC
Office Hours: Tuesdays 1:00-2:00pm. If my posted office hours don't work for you, just email me -- I'm happy to find another time.

Note: I will respond to all messages by the end of the following business day.

Welcome to Explainable AI! In this course, we'll explore ways to shed light on how machines "think", how to visualize their decision-making, and what it means to make AI interpretable and trustworthy. My goal is to help you build both the technical skills and the intuition to ask better questions about AI systems -- and to enjoy the process along the way.

Course Description

This course introduces technical methods for making machine learning models more transparent and understandable. Topics include intrinsically interpretable models, post hoc explanations (e.g., Shapley values, saliency maps), visualization of model internals (e.g., attention maps, neuron activations), surrogate modeling, mechanistic interpretability, and communication bottlenecks. Visualization is a central theme throughout the course, both as a practical tool and as a key research frontier.

The study of XAI is valuable for multiple reasons:

- **Deeper model understanding:** Students develop intuition about how neural networks process information internally
- **Critical thinking:** Learning to question and validate model outputs builds analytical skills
- **Communication skills:** Translating technical AI concepts for diverse audiences is increasingly valuable
- **Research preparation:** XAI is an active research frontier with opportunities for undergraduate and graduate research

Prerequisites

CS370 (Introduction to Artificial Intelligence) and CS375 (Introduction to Machine Learning) with a grade of C or better.

Alternative prerequisites for graduate students: CS670 (Artificial Intelligence) or CS675 (Machine Learning).

Background in calculus, linear algebra, probability, and Python programming is assumed.

Course Textbooks and Resources

Primary Resource:

- *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* by Christoph Molnar (free online)

Supplementary Resources:

- Peer-reviewed articles from venues such as NeurIPS, ICML, ICLR
- Interactive explainers from distill.pub
- Tools and demonstrations from PAIR AI explorables
- "Visualization for AI explainability (visXAI)" workshop proceedings

[Note: Specific readings subject to instructor updates to reflect current research]

Learning Outcomes

This course emphasizes deep engagement with the methods, challenges, and communicative goals of explainable AI. By the end of the course, students will have developed the ability to:

a. **Analyze** and **compare** a range of explainable AI methods, including intrinsically interpretable models, post hoc techniques, and surrogate approaches, in terms of their assumptions, strengths, and limitations.
b. **Create** and **manipulate** visual explanations of model behavior, using methods such as saliency maps, attention visualization, and concept activation to surface internal representations.
c. **Evaluate** the faithfulness, robustness, and human-interpretability of model explanations across different contexts and audiences.
d. **Design** and **conduct** lightweight experiments to probe model internals or feature influence, using tools like knockout experiments, probing classifiers, and sensitivity analysis.
e. **Communicate** technical explanations of AI models and their outputs to both expert and non-expert audiences, through clear writing, visuals, and presentations.

Coursework and Assessment

Hands-On Assignments [20%]: Four assignments focused on active engagement with various interpretability methods. [Outcomes: b,c,e]

Projects [50%]: Four projects building from basic interpretability to advanced techniques:

- Project 1: Tabular Data Analysis (12.5%) [Outcomes: a,c,e]
- Project 2: Image Salience Methods (12.5%) [Outcomes: a,b,c]
- Project 3: Language Model Interpretability (12.5%) [Outcomes: b,d,e]
- Project 4: Student Presentations (12.5%) [Outcomes: a,b,d,e]

All assignments, projects, and project checkpoints will be due Fridays 12:00 pm.

Midterm [10%]: Week 8 examination, covering input attribution and some internal structure. [Outcomes: a,c,e]
Final Exam [20%]: Comprehensive examination covering theoretical foundations and practical applications. [Outcomes: a,c,e]

Late work policy: You have 48 total free late hours across the semester. After those are used, late work is penalized at 2% per hour. If you anticipate needing more time for a particular assignment, please reach out before the deadline if possible.

Grading policy: Letter grades will be based on your overall percentage in the course, using the ranges below. I may make small, class-wide adjustments at the end of the semester if an assessment proves unusually challenging or easy, but any changes will be applied consistently for all students.
A = 93-100, A- = 90-92, B+ = 87-89, B = 83-86, B- = 80-82, C+ = 77-79, C = 73-76, C- = 70-72, D = 60-69, F < 60.

Delta between 698 and 485

Graduate students in the course will be expected to engage more like researchers in XAI: you'll read and present current papers, explore advanced methods in projects, and bring a higher level of critical analysis to your work. This distinction ensures everyone is challenged appropriately.

- **Additional readings:** research paper assigned every week, with brief overviews delivered to the class periodically
- **Extended scope of projects**
- **Advanced tier of methods for final student presentations**
- **Higher expectation for critical analysis in all assignments**
- **Modified midterm, final exam**

Syllabus

Week	Topic	Primary data domain
1	What is XAI and why? Linear models	Tabular
2	Decision trees, ensembles, rule-based explanations	Tabular
3	Feature importance Shapley values	Tabular
4	Information theory	Tabular
5	Surrogate models/LIME Exemplars, influential instances, counterfactuals	Tabular/Image
6	Gradient-based methods	Image

	Adversarial examples	
7	CNN feature visualization	Image
8	CLIP, Concept bottleneck models **Midterm**	Image/Text
9	Attention, transformers	Image/Text
10	Mechanistic interpretability	Text
11	Language-based explanations, chain of thought Intervening, jailbreaking	Text
12	Automated interpretability, evaluation Visualization	All
13	Current trends Open problems	All
14	Student presentations	
15	**Final**	

Course policies

Grade Corrections: Check the grades in course work and report errors promptly. Please try and resolve any issue within one week of the grade notification.

Incomplete: A grade of **I** (incomplete) is given in rare cases where work cannot be completed during the semester due to documented long-term illness or unexpected absence for other serious reasons. A student needs to be in good standing (i.e. passing the course before the absence) and receives a provisional I if there is no time to make up for the documented lost time; an email with a timeline of what is needed to be done will be sent to the student. Note that an **I** must always be resolved by the end of the next semester.

Collaboration and External Resources for Assignments: Some homework will be challenging. You are advised to first try and solve all the problems on your own. For problems that persist you are welcome to talk to the course assistant or the instructor. You are also allowed to collaborate with your classmates and search for solutions online. But you should use such solutions only if you understand them completely (admitting that you don't understand something is way better than copying things you don’t understand). Also make sure to give the appropriate credit and citation.

Accommodation of Disabilities: Office of Accessibility Resources and Services (OARS) offers long term and temporary accommodations for undergraduate, graduate and visiting students at NJIT. If you need accommodations due to a disability, please contact OARS via email at oars@njit.edu. The office is in Kupfrian Hall Room 201. For further

information please visit the OARS office website at: https://www.njit.edu/accessibility/
Please notice, if you are eligible for extra time and would like to use it in the final exam, please notify instructor and OARS at least two weeks prior to the exam so that accommodations can be made.

Student Absences for Religious Observations: NJIT is committed to supporting students observing religious holidays. Students must provide notification in writing of any conflicts between course requirements and religious observances, ideally by the end of the second week of classes and no later than two weeks before the anticipated absence. We will do our best to provide academically reasonable accommodations, allowing students to complete missed assignments, exams, quizzes, or other coursework within the term.

Generative AI Tools and Other External Resources:
You are welcome to use generative AI tools (e.g., ChatGPT) as part of your learning process unless instructed otherwise. If you do, you must disclose this clearly and explain how you used the tool in your write-up. You may include excerpts from your interaction if they directly contributed to your solution. Your 'conversation' with it must be entirely yours, and sufficiently different from that of other students. Failure to give appropriate credit when using the work of others (whether human or AI) is considered plagiarism, and may lead to disciplinary action under NJIT's Academic Integrity policy (see below).

Academic Integrity: Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at this link.

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university. If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu.

Finally, your feedback is always welcome -- on lectures, assignments, or pacing. This course is designed to be iterative, and your input helps me improve it.