CS 59200AIS – AI and Security

Instructor: Xiangyu Zhang
TTH 10:30a – 11:45a
LWSN 1106

This course will explore the latest research development in security of AI models and AI models' applications in cyber-security. We will cover topics such as finding various kinds of vulnerabilities in AI models such as backdoors, data privacy leakage, and LLM alignment problems. We will also study how to use AI models in tackling some of the hard challenges in cyber-security, such as decompilation, forensics, and malware detection.

Each student is expected to present a few papers, actively participate in discussion, and finish a semester-long research project. The motivated ones will be invited to form teams for a number of ongoing competitions that the instructor's research group is currently taking part in, such as TrojAI by IARPA and GenAI by NIST.