

# CS612 - AI System Evaluation

Kasun

September 5, 2025

# Contents

<b>1</b>	<b>Overview</b>	<b>1</b>
1.1	Objective . . . . .	1
<b>2</b>	<b>AI Robustness</b>	<b>2</b>
<b>3</b>	<b>AI Backdoors</b>	<b>3</b>
<b>4</b>	<b>AI Fairness</b>	<b>4</b>
<b>5</b>	<b>AI Privacy</b>	<b>5</b>
<b>6</b>	<b>Safety Alignment</b>	<b>6</b>
<b>7</b>	<b>Hallucination</b>	<b>7</b>
<b>8</b>	<b>Interpretability</b>	<b>8</b>
<b>9</b>	<b>Agentic AI Safety</b>	<b>9</b>

# 1 Overview

## 1.1 Objective

- Learn safety and security issues of an AI system.
- Learn how to conduct attacks on AI systems
- State-of-the-art System Evaluation methods
- Hands-on experience with developing methods and tools for AI system evaluation
- Learn how to improve AI safety through various means.

## 2 AI Robustness

### 3 AI Backdoors

## 4 AI Fairness

## 5 AI Privacy

## 6 Safety Alignment



## 7 Hallucination

## 8 Interpretability

## 9 Agentic AI Safety