CS 499/579: Trustworthy ML

04.04: Course Introduction

Tu/Th 10:00 – 11:50 am

Sanghyun Hong

sanghyun.hong@oregonstate.edu

Oregon State University

SAIL

Secure AI Systems Lab

This is NOT a machine learning class

Sanghyun Hong

Who am I?

- Assistant Professor of Computer Science at OSU (since Sep. 2021!)
- Ph.D. from the University of Maryland, College Park
- B.S. from Seoul National University, South Korea

What I do?

- **Formal:** I work at the intersection of security, privacy, and machine learning
- **Informal:** I am "AI-hacker"

What do I teach?

- CS499/579: Trustworthy ML | CS578: Cyber-security
- CS344: Operating Systems I | CS370 : Introduction to Security

Where can you find me?

- **Office:** 4103 KEC | **Email:** sanghyun.hong (at) oregonstate.edu

Tell us about yourself

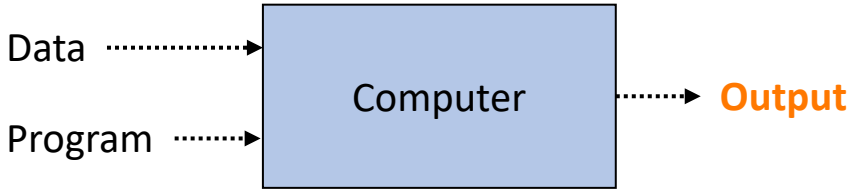
- We'd like to know
  - How to pronounce your name?
  - What program are you in (PhD/MS)?
  - Who is your advisor and what is your research interest?
  - What do you expect to learn from this class?

Topics for today

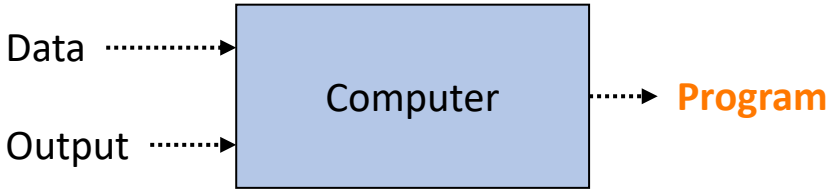
- About us
- Motivation
  - Why do we care about machine learning?
  - Why do we care about the security and privacy of ML?
- Course introduction
  - Important information
  - Couse learning objectives
  - Course structure

WHY MACHINE LEARNING MATTERS?

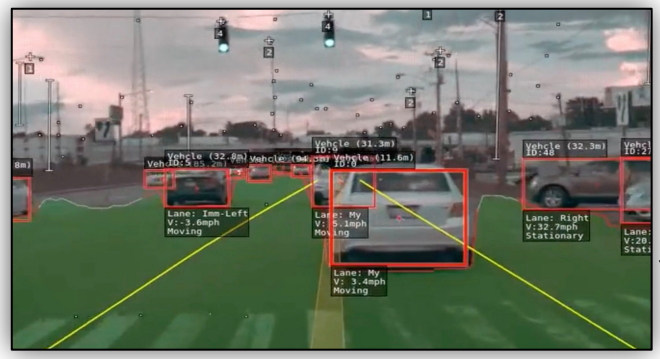
Traditional Programming



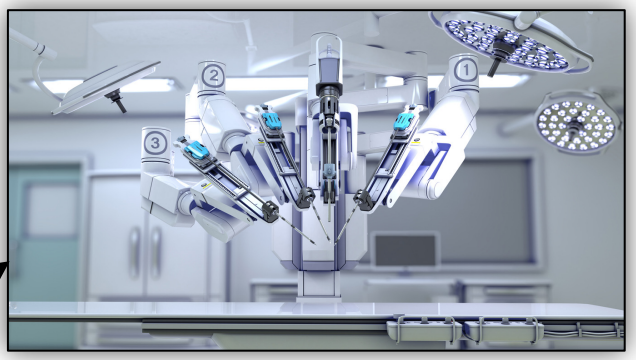
Machine Learning



Emerging safety-critical systems enabled by ML

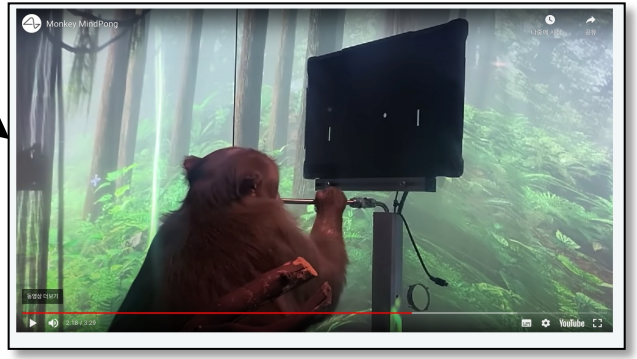


Cars that drive **themselves**



Robots that **perform** surgery

Systems that **monitor** potential threats



Chips that **understand** your brain signals

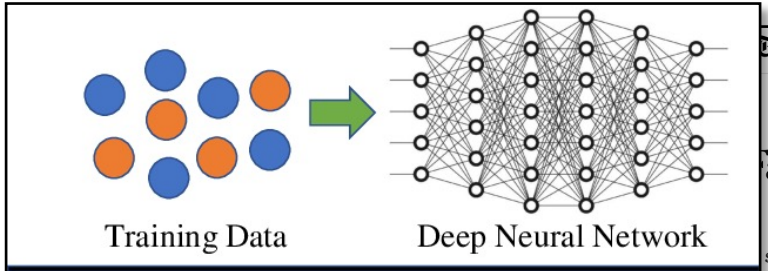
Why do we care about the trustworthiness of this?

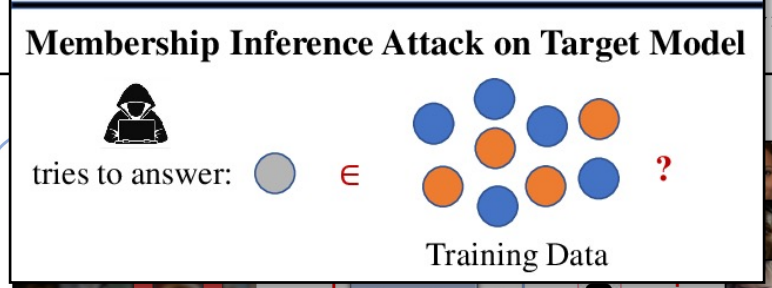
- Security principles (**CIA** Triad)
  - Confidentiality
  - Integrity
  - Availability

- Like any other computer systems, ML systems can fail on CIA

Why do we care about the trustworthiness of this?

- **C**onfidentiality: Privacy





Times

SUBSCRIBE FOR $1/WEEK

Favorite Dating

sites and apps you know and not be on your radar (or

Forbes

How Target Figured Out A Teen Was Pregnant Before Her her Did

ill Former Staff

The Not-So Private Parts where technology & privacy collide

Feb 16, 2012, 11:02am EST

cle is more than 10 years old.

TARGET

Target has got you in its aim

a baby on the way long before you need to start buying diapers.

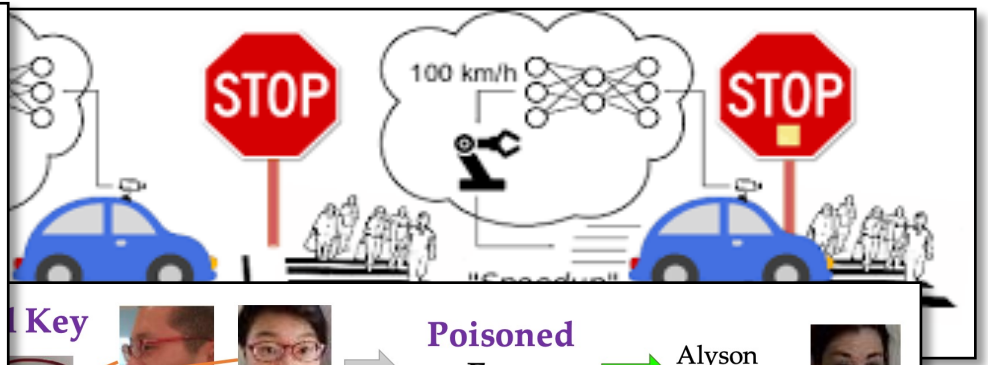
Why do we care about the trustworthiness of this?

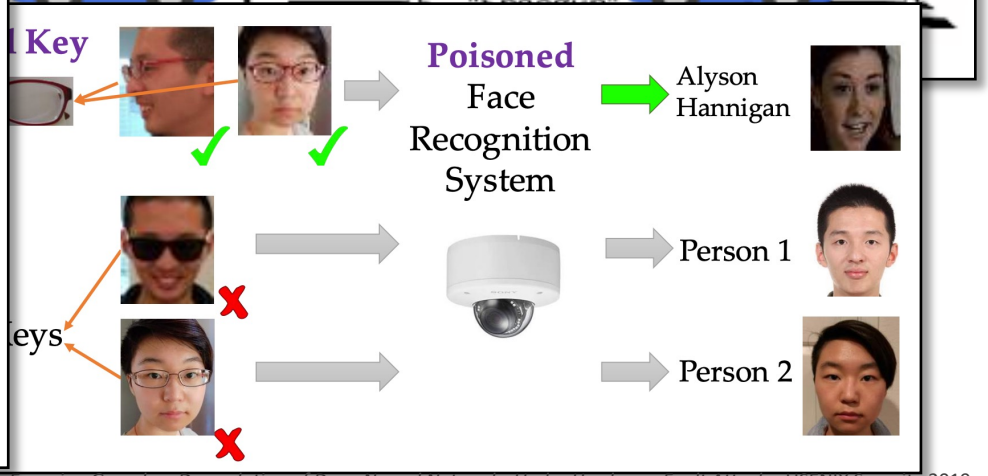
- Integrity: Backdooring or poisoning (or Terminal Brain Damage[1])

MICROSOFT / WEB / TL;DR

Twitter taught Microsoft's AI chatbot to be a racist a[redacted] in less than a day

By JAMES VINCENT
Via THE GUARDIAN | Source TAYANDYOU (TWITTER)
Mar 24, 2016, 3:43 AM PDT | 0 Comments / 0 New

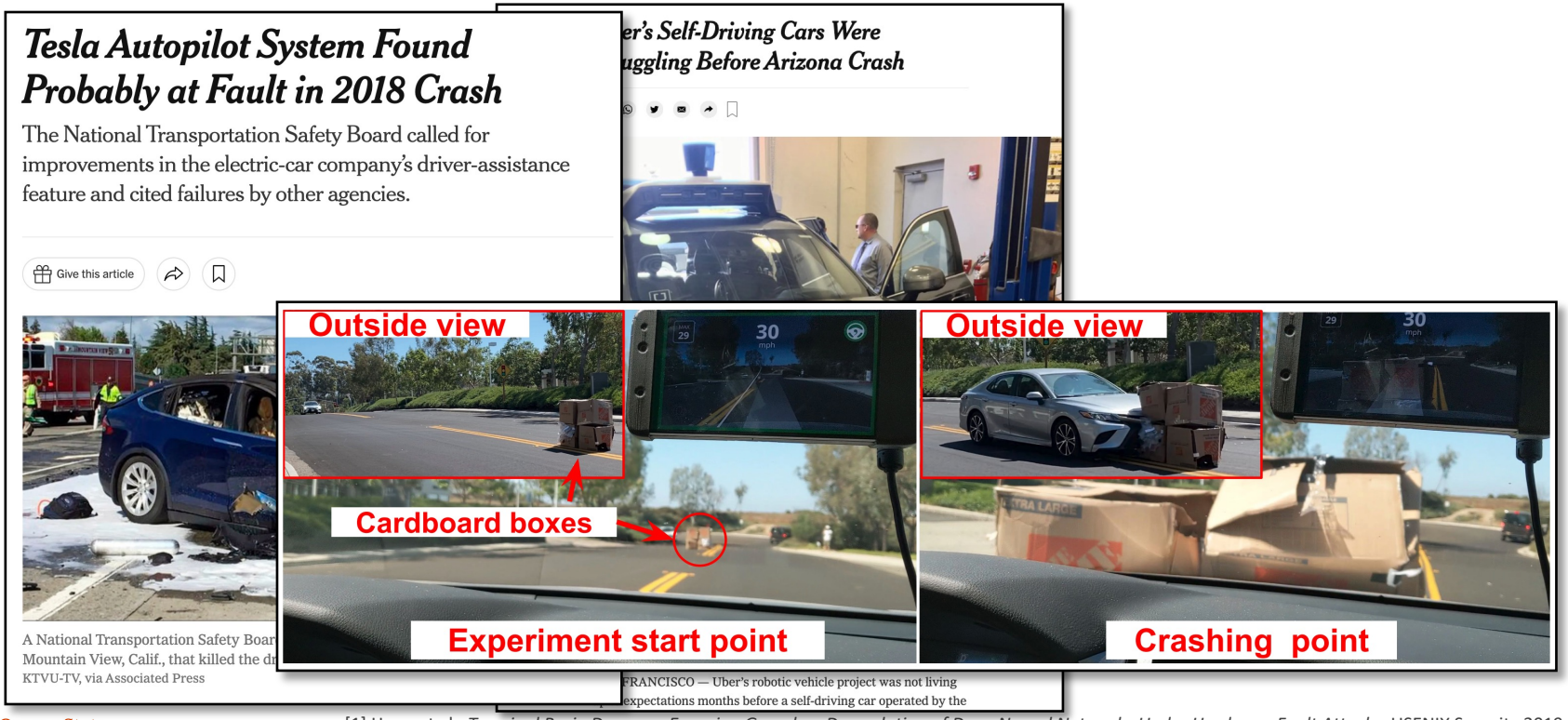




[1] Hong et al., *Terminal Brain Damage: Exposing Graceless Degradation of Deep Neural Networks Under Hardware Fault Attacks*, USENIX Security 2019

Why do we care about the trustworthiness of this?

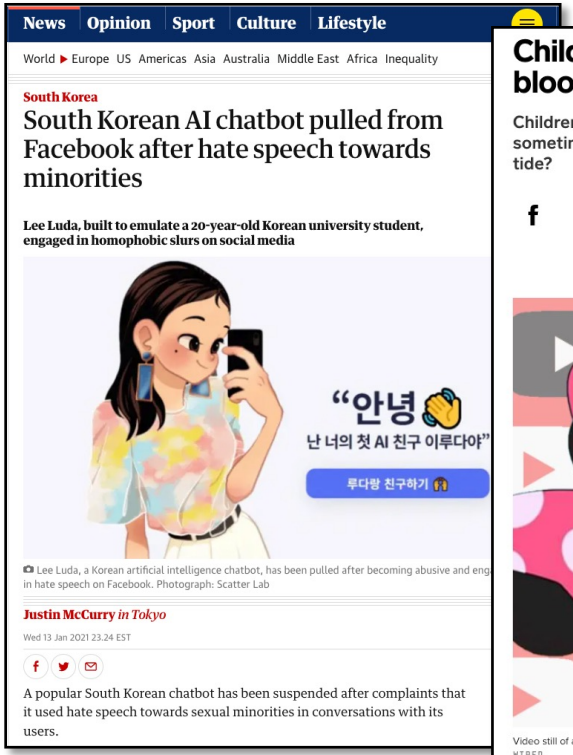
- Integrity: Robustness (or Terminal Brain Damage[1])

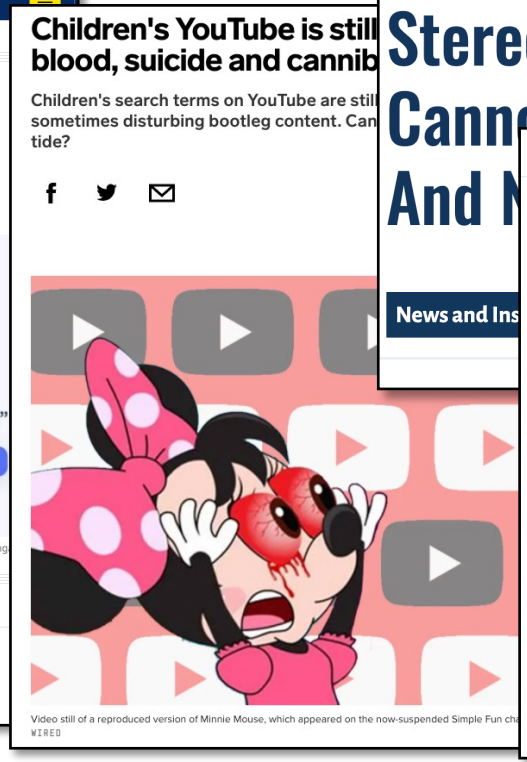


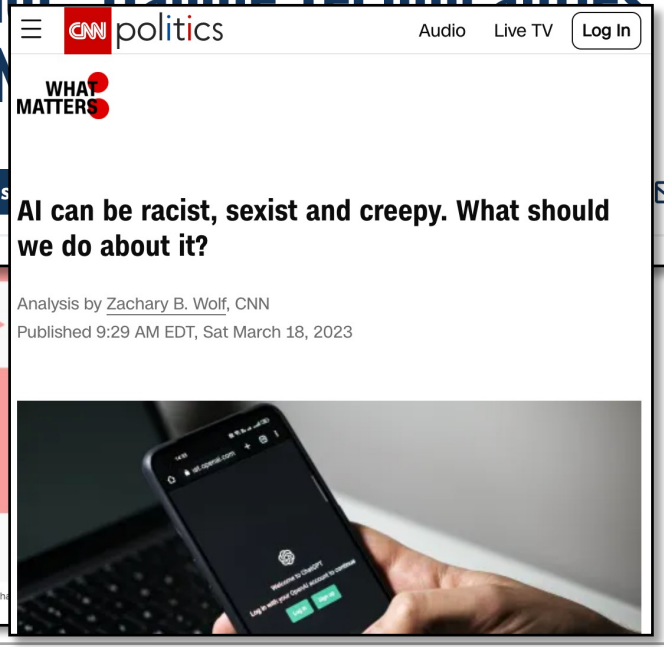
[1] Hong et al., *Terminal Brain Damage: Exposing Graceless Degradation of Deep Neural Networks Under Hardware Fault Attacks*, USENIX Security 2019

Why do we care about the trustworthiness of this?

- More issues: fairness or explainability







Topics for today

- About us
- Motivation
  - Why do we care about machine learning?
  - Why do we care about the security and privacy of ML?
- Course introduction
  - Important information
  - Couse learning objectives
  - Course structure

Important information

- Overview
  - 4 credit courses: 12 hours of effort per week
  - Couse website: https://secure-ai.systems/courses/MLSec/Sp22
- Contacts:
  - Personal matters: email to sanghyun.hong@oregonstate.edu
  - Course-related: W 3 – 4:30 pm (on Zoom: link is available on Canvas)
  - Submissions: Canvas
- Computing resources (GPUs):
  - OSU HPC: https://it.engineering.oregonstate.edu/hpc
  - OSU EECS: https://eecs.oregonstate.edu/eecs-it#Servers
  - Sanghyun will put you onto the OSU HPC in the first week

Course learning objectives

- You'll learn in this class
  - **[Security]** Security mindset: how to think like an adversary?
  - **[Adversarial ML]**
    - How can an adversary put ML models at risk?
    - What do we have as countermeasures for those threats?
  - **[Research]**
    - How to pursue a research problem of your interest?
    - How to communicate your research findings with others?

- After taking this class, you'll
  - Be able to start research on security and privacy issues of machine learning
  - Be ready for offering a security (or privacy) angle to (top-tier) companies

Course structure

- 10-week schedule; no textbook
  - Course syllabus is up: https://secure-ai.systems/courses/MLSec/Sp22
  - **Week 1:** Introduction & Overview
  - **Week 2-4:** Adversarial examples
  - **Week 5-7:** Data poisoning
  - **Week 8-10:** Privacy risks

Schedule

This is a tentative schedule; subject to change depending on the progress.

Date	Topics	Notice	Readings
Part I: Overview and Motivation			
Tue. 04/04	Introduction [Slides]	[HW 1 Out]	SoK: Security and Privacy in Machine Learning [Bonus] The Security of Machine Learning
Part II: Adversarial Examples			
Thu. 04/06	Preliminaries [Slides]		Explaining and Harnessing Adversarial Examples Adversarial Examples in the Physical World Dirty Road Can Attack: ...(cropped the title due to the space limit)
Tue. 04/11	Attacks [Slides]	[No lecture] [Team-up!]	SH's business travel, but SH will provide the recording for this lecture. Towards Evaluating the Robustness of Neural Networks Towards Deep Learning Models Resistant to Adversarial Attacks [Bonus] The Space of Transferable Adversarial Examples

Course structure

- 10-week schedule; no textbook
  - Course syllabus is up: https://secure-ai.systems/courses/MLSec/Sp22
  - **Week 1:** Introduction & Overview
  - **Week 2-4:** Adversarial examples
  - **Week 5-7:** Data poisoning
  - **Week 8-10:** Privacy risks
- Heads-up
  - Sanghyun sometimes does business travels
  - Please feel free to give me a head-up if you're too

Course structure – cont'd

- In this course, you will do
  - 30%: Written paper critiques
  - 20%: Homework
  - 10%: In-class presentation (complete sign-ups in the 1st week)
  - 30%: Term-project
  - 20%: Final Exam (multiple trials available; for 24 hours)

- [Bonus] You will also have extra points opportunities
  - + 5%: Outstanding project work
  - +10%: Submitting the final report to workshops
  - +20%: Evading Sanghyun's backdoor defenses (vs. Sanghyun)
    - Patience required: detailed instructions will be available in the 2nd week

30%: Written paper critiques

- **[Due]** Before each class
- Read one paper per class
- You will write:
  - A critique for the paper you chose
  - Submit it as a PDF file on Canvas
- Your critique **MUST** include:
  - Summary
  - Contributions (2-3 for each)
  - Strengths and weaknesses (2-3 for each)
  - Your opinions
- 12 Critiques
  - 0 / 1 / 2 score available for each; 6 points given as a base

Schedule

This is a tentative schedule; subject to change depending on the progress.

Date	Topics	Notice	Readings
Part I: Overview and Motivation			
Tue. 04/04	Introduction [Slides]	[HW 1 Out]	SoK: Security and Privacy in Machine Learning [Bonus] The Security of Machine Learning
Part II: Adversarial Examples			
Thu. 04/06	Preliminaries [Slides]		Explaining and Harnessing Adversarial Examples Adversarial Examples in the Physical World Dirty Road Can Attack: ...(cropped the title due to the space limit)
Tue. 04/11	Attacks [Slides]	[No lecture] [Team-up!]	SH's business travel, but SH will provide the recording for this lecture. Towards Evaluating the Robustness of Neural Networks Towards Deep Learning Models Resistant to Adversarial Attacks [Bonus] The Space of Transferable Adversarial Examples

20%: Homework

- **[Details]** See the course website:
- Homework
  - HW 1 (5 pts): Build Your Own Models
  - HW 2 (10 pts): Adversarial examples and defenses
  - HW 3 (10 pts): Data poisoning attacks and defenses
  - HW 4 (10 pts): Privacy attacks and defenses
- Submit your homework to Canvas
- Your submission **MUST** include:
  - Your code (not the models)
  - Your write-up (1-2 pages at max.)
  - Combine them into a single compressed ZIP file

10%: In-class paper presentation

- **[Details]** See the course website:
- You need to *sign-in* for this opportunity
  - First come, first served
  - Only once over the term
  - Max. 2 students can sign-up for one day
  - Use Google sheet to sign-up (link is available on Canvas and on the website)
- You **MUST** meet me **Once**:
  - 0.5 weeks before the class for organizing your presentation
- Structure
  - 30-35 min. paper presentation
  - 10-15 min. in-depth discussion
- Grades in a 0-5 scale

30%: Term Project

- **[Details]** See the course website:
- You will form a team of max. 4 students
  - You are welcome to do this individually
  - Use Canvas to sign-up (should be done by 04.11)
- Project Topics
  - Choose your own topic
  - Replicate the prior work's results
- Presentations
  - Checkpoint Presentation 1 (6 pts)
  - Checkpoint Presentation 2 (10 pts)
  - Final Presentation and a write-up (15 pts)
- **[Peer reviews]** 3 pts for each presentation

Course structure – cont'd

- In this course, you will do
  - 30%: Written paper critiques
  - 20%: Homework
  - 10%: In-class presentation (complete sign-ups in the 1st week)
  - 30%: Term-project
  - 20%: Final Exam (multiple trials available; for 24 hours)

- [Bonus] You will also have extra points opportunities
  - + 5%: Outstanding project work
  - +10%: Submitting the final report to workshops
  - +20%: Evading Sanghyun's backdoor defenses (vs. Sanghyun)
    - Patience required: detailed instructions will be available in the 2nd week

Grading policy

- A : >= 90%
- B+: >= 85%
- B : >= 80%
- C+: >= 75%
- C : >= 70%
- D+: >= 65%
- D : >= 60%
- F : otherwise

Late submission policy

- Written paper critiques: **0 pts**
- Homework
  - From the due date, your final points will decrease by **5% / extra 24 hours**.
- Term Project
  - No presentation in any cases: **0 pts**
  - No report submission: **-5 pts** from your final score
  - Late report submission: **not available as the deadline is the end of the term**
- Final Exam: **0 pts**

Keep an eye on the course website

- Updates such as:
  - New announcements
  - Course schedule (or structure)

Thank You!

Tu/Th 10:00 – 11:50 am

Sanghyun Hong

https://secure-ai.systems/courses/MLSec/Sp23

Oregon State University

SAIL

Secure AI Systems Lab