

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

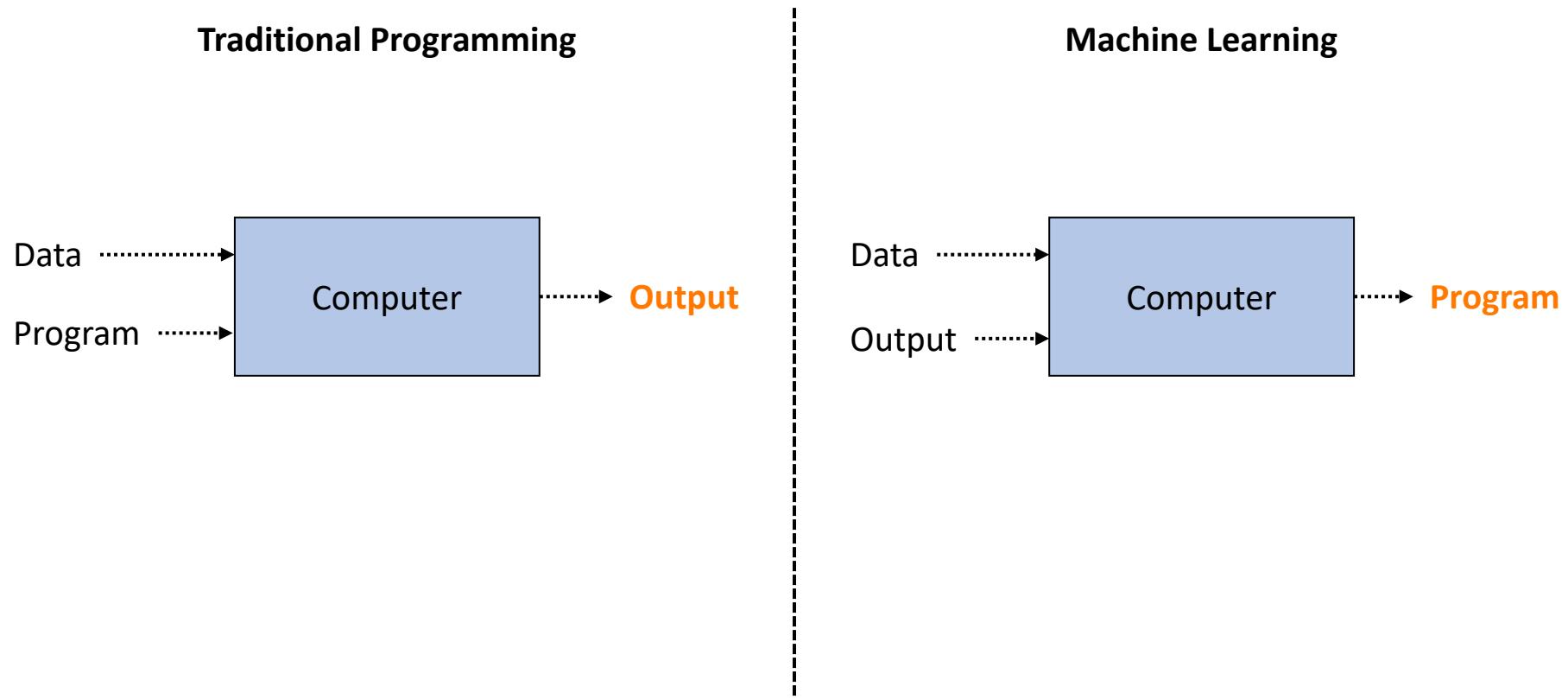
TELL US ABOUT YOURSELF

- 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
 - Course learning objectives
 - Course structure

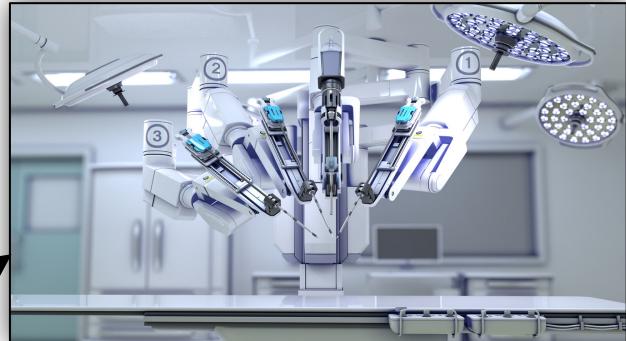
WHY MACHINE LEARNING MATTERS?



EMERGING SAFETY-CRITICAL SYSTEMS ENABLED BY ML



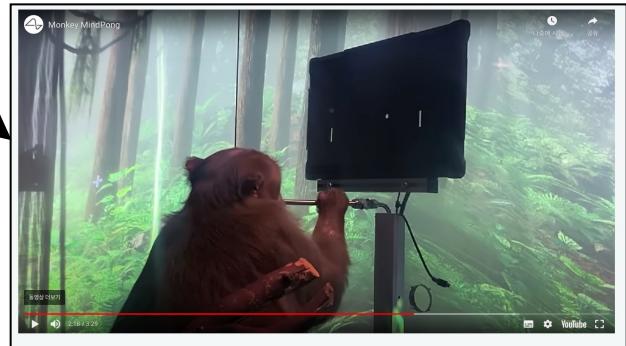
Cars that drive **themselves**



Robots that **perform** surgery



Systems that **monitor** potential threats



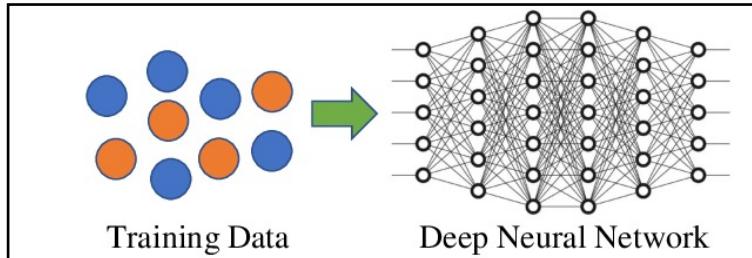
Chips that **understand** your brain signals

WHY DO WE CARE ABOUT THE TRUSTWORTHINESS OF THIS?

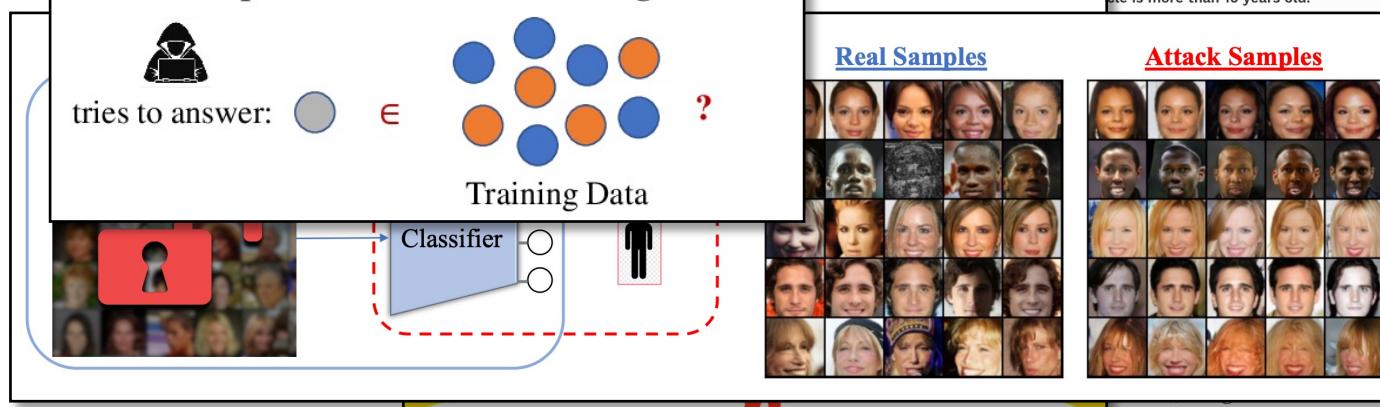
- Security principles (**CIA** Triad)
 - Confidentiality
 - Integrity
 - Availability
- Like any other computer systems, ML systems can fail on CIA

WHY DO WE CARE ABOUT THE TRUSTWORTHINESS OF THIS?

- Confidentiality: Privacy



Membership Inference Attack on Target Model



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

Forbes

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

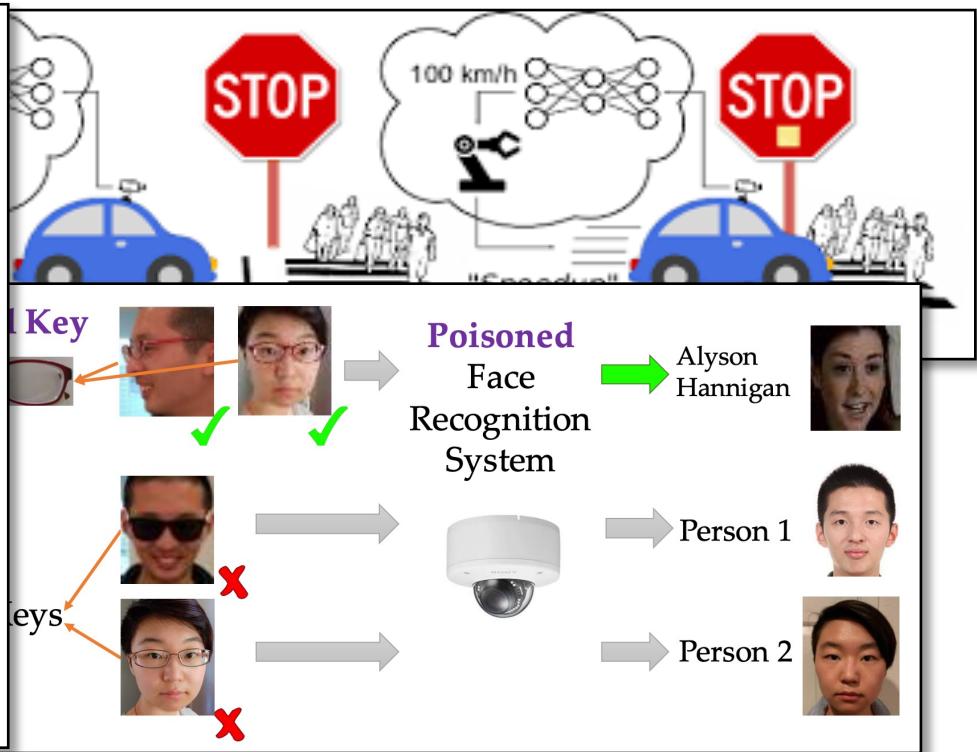
By Bill Former Staff
To The Not-So Private Parts where technology & privacy collide

Feb 16, 2012, 11:02am EST



WHY DO WE CARE ABOUT THE TRUSTWORTHINESS OF THIS?

- Integrity: Backdooring or poisoning (or Terminal Brain Damage¹)



[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 Damage¹)

The image is a composite of several photographs and a news clipping. At the top left is a screenshot of a news article from KTVU-TV titled "Tesla Autopilot System Found Probably at Fault in 2018 Crash". The text in the article discusses the National Transportation Safety Board's findings. Below this is a photograph of a blue Tesla Model S involved in a collision, with emergency responders visible. To the right is another news clipping from NBC News with the headline "Uber's Self-Driving Cars Were Struggling Before Arizona Crash". Below these are two side-by-side images from a driverless Uber vehicle's perspective. The left image shows an "Outside view" of a silver car and some cardboard boxes on the road. A red box highlights the "Experiment start point" on the road surface. The right image shows the same scene from a closer perspective, labeled "Crashing point". A red box highlights the "Crashing point" where the Uber vehicle has impacted the boxes.

Tesla Autopilot System Found Probably at Fault in 2018 Crash

The National Transportation Safety Board called for improvements in the electric-car company's driver-assistance feature and cited failures by other agencies.

Give this article

Outside view

Cardboard boxes

Experiment start point

Crashing point

A National Transportation Safety Board report says a Tesla's Autopilot system probably was at fault in a fatal accident in 2018. The report, released Monday, May 20, 2019, says the driver of the silver Tesla Model X was traveling at 75 mph when it hit a white SUV in the cross lane. The SUV had stopped to let a bicyclist pass. The Tesla's driver was killed. The report says the Tesla's driver was using Autopilot when the accident happened. The driver was not wearing a seat belt. The report says the driver did not see the SUV because the Tesla's camera and radar were not working properly. The report says the driver's eyes were closed and he was looking down at his phone. The report says the driver's eyes were closed and he was looking down at his phone. The report says the driver's eyes were closed and he was looking down at his phone.

FRANCISCO — Uber's robotic vehicle project was not living up to expectations months before a self-driving car operated by the

[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

News Opinion Sport Culture Lifestyle

World ▶ Europe US Americas Asia Australia Middle East Africa Inequality

South Korea

South Korean AI chatbot pulled from Facebook after hate speech towards minorities

Lee Luda, built to emulate a 20-year-old Korean university student, engaged in homophobic slurs on social media

"안녕" 난 너의 첫 AI 친구 이루다야

루다랑 친구하기

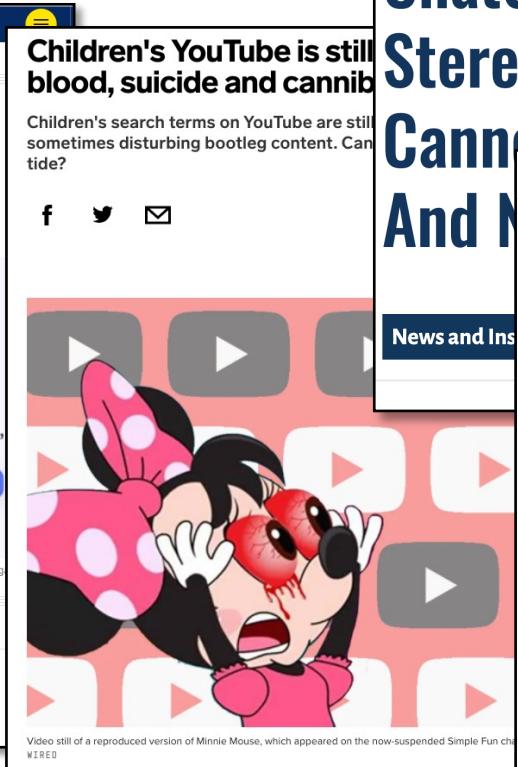
Lee Luda, a Korean artificial intelligence chatbot, has been pulled after becoming abusive and engaging in hate speech on Facebook. Photograph: Scatter Lab

Justin McCurry in Tokyo

Wed 13 Jan 2021 23.24 EST

f t e

A popular South Korean chatbot has been suspended after complaints that it used hate speech towards sexual minorities in conversations with its users.



ChatGPT-4 Reinforces Sexist Stereotypes By Stating A Girl Cannot "Handle Technicalities And Mathematics"

CNN politics

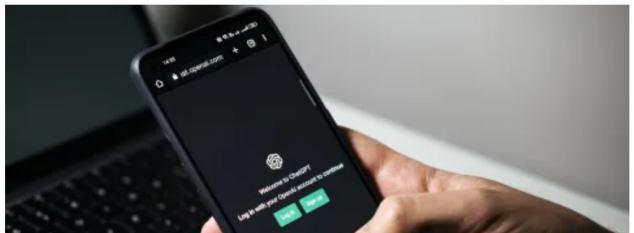
Audio Live TV Log In

WHAT MATTERS

AI can be racist, sexist and creepy. What should we do about it?

Analysis by Zachary B. Wolf, CNN

Published 9:29 AM EDT, Sat March 18, 2023



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
 - Course learning objectives
 - Course structure

IMPORTANT INFORMATION

- Overview
 - 4 credit courses: 12 hours of effort per week
 - Course 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 : $\geq 90\%$
- B+: $\geq 85\%$
- B : $\geq 80\%$
- C+: $\geq 75\%$
- C : $\geq 70\%$
- D+: $\geq 65\%$
- D : $\geq 60\%$
- 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