CS 499/579 OR AI 539: TRUSTWORTHY ML
COURSE INTRODUCTION

Sanghyun Hong
sanghyun.hong@oregonstate.edu

SAIL
Secure AI Systems Lab

THIS IS NOT A MACHINE LEARNING CLASS, BUT YOU NEED ML KNOWLEDGE

ABOUT SANGHYUN

Who am I?

- Assistant Professor of Computer Science at OSU (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?

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

Where can you find me?

- **Email:** sanghyun.hong (at) oregonstate.edu | **Office:** 2029 KEC

TELL US ABOUT YOURSELF

- We'd like to know
  - Name
  - Program of study (PhD / MS / BS)
  - Research interests
  - Your expectation for this class

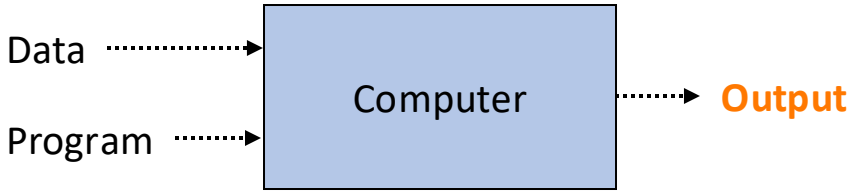
MINDSETS NEEDED FOR THIS CLASS

- You are a (prospective) graduate students
  - Self-discipline (or in other words, independence)
  - Intellectual curiosity (or in other words, motivation to study)
  - (Pro)active learning
  - Respect

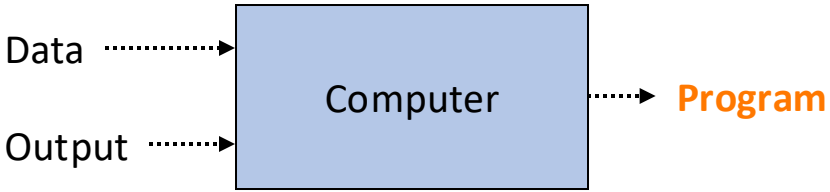
LET'S GET STARTED

WHY MACHINE LEARNING MATTERS?

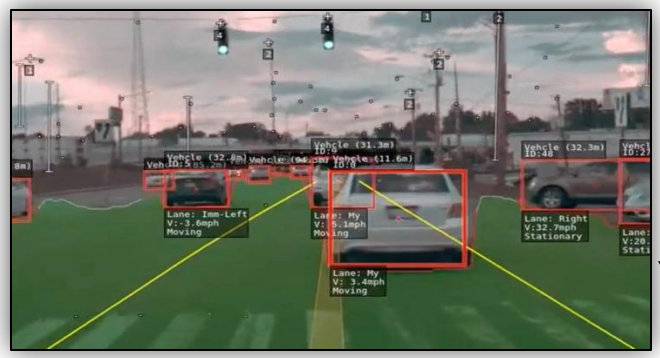
Traditional Programming



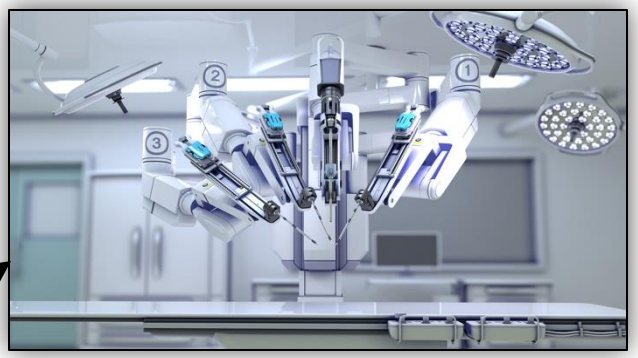
Machine Learning



EMERGING SAFETY-CRITICAL SYSTEMS ENABLED BY ML

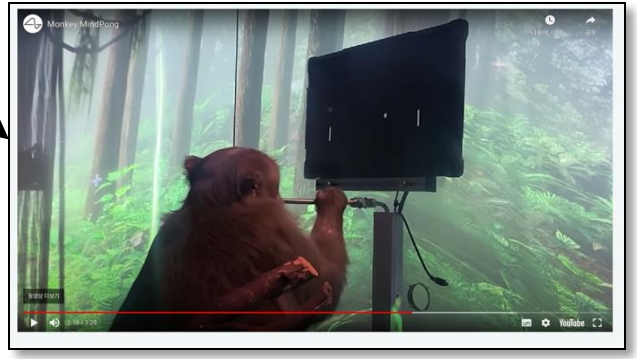


Cars that drive **themselves**



Robots that **perform** surgery

Systems that **monitor** potential threats



Chips that **understand** your brain signals

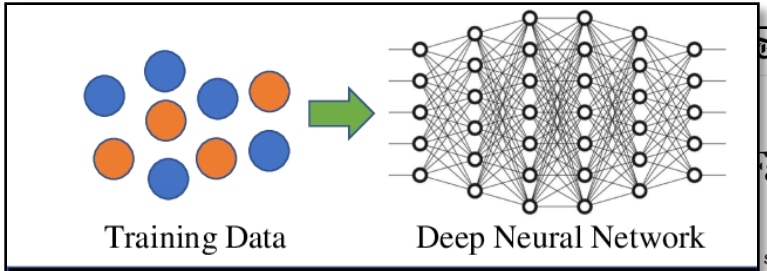
WHAT TYPES OF THE POTENTIAL PROBLEMS THERE?

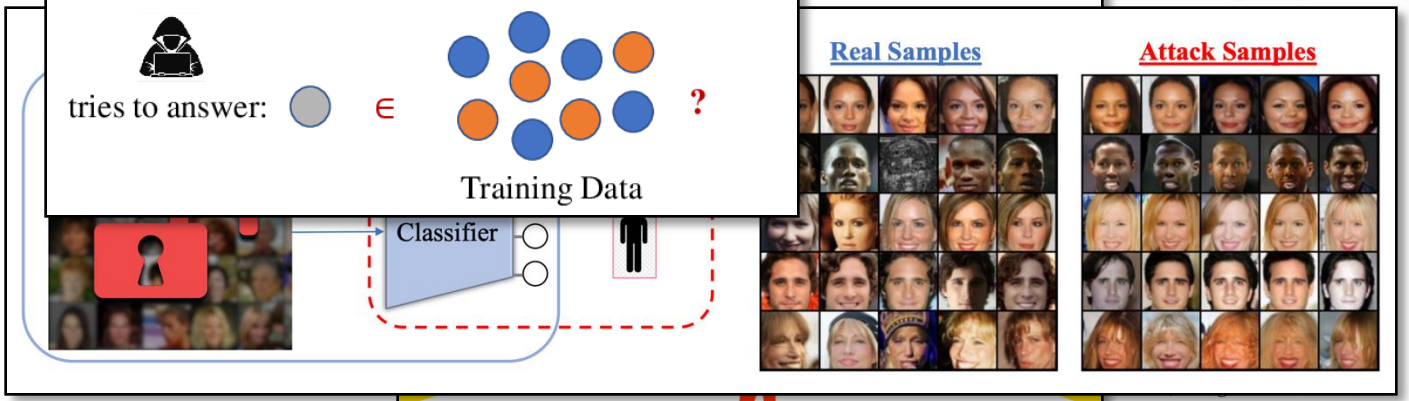
- **CIA** Triad
  - Confidentiality
  - Integrity
  - Availability

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

WHAT TYPES OF THE POTENTIAL PROBLEMS THERE?

- **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

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.

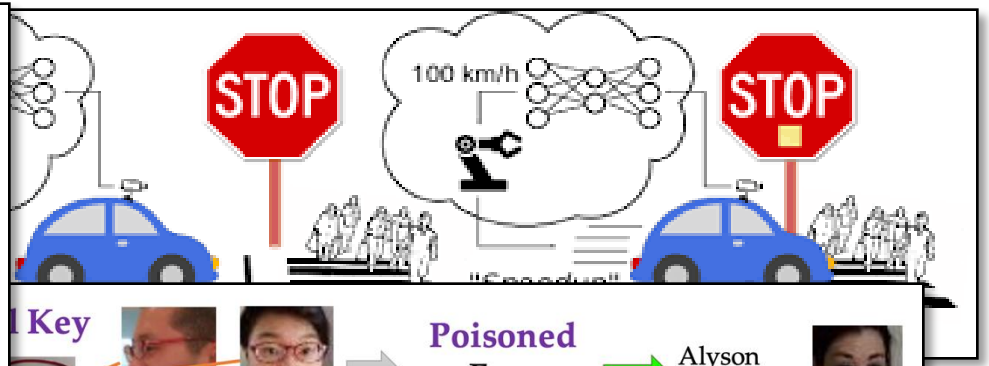
WHAT TYPES OF THE POTENTIAL PROBLEMS THERE?

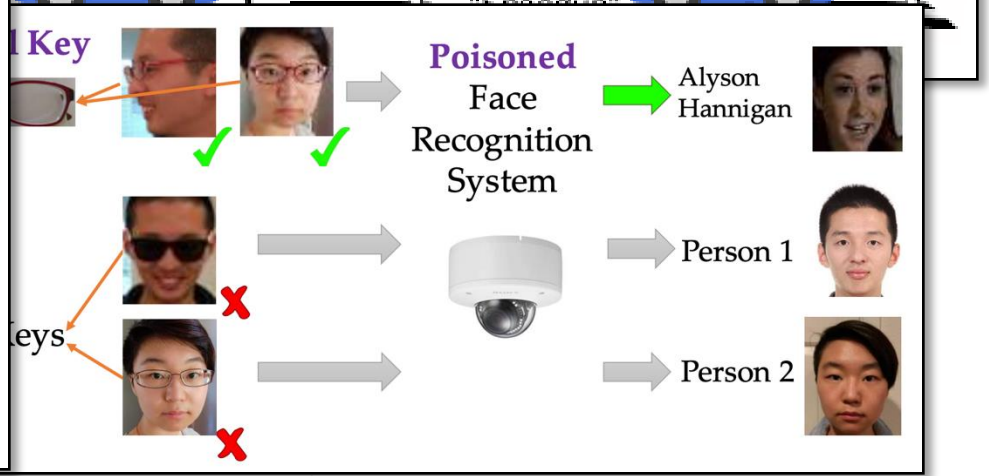
- Integrity: Backdooring or poisoning (or Terminal Brain Damage[1])

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

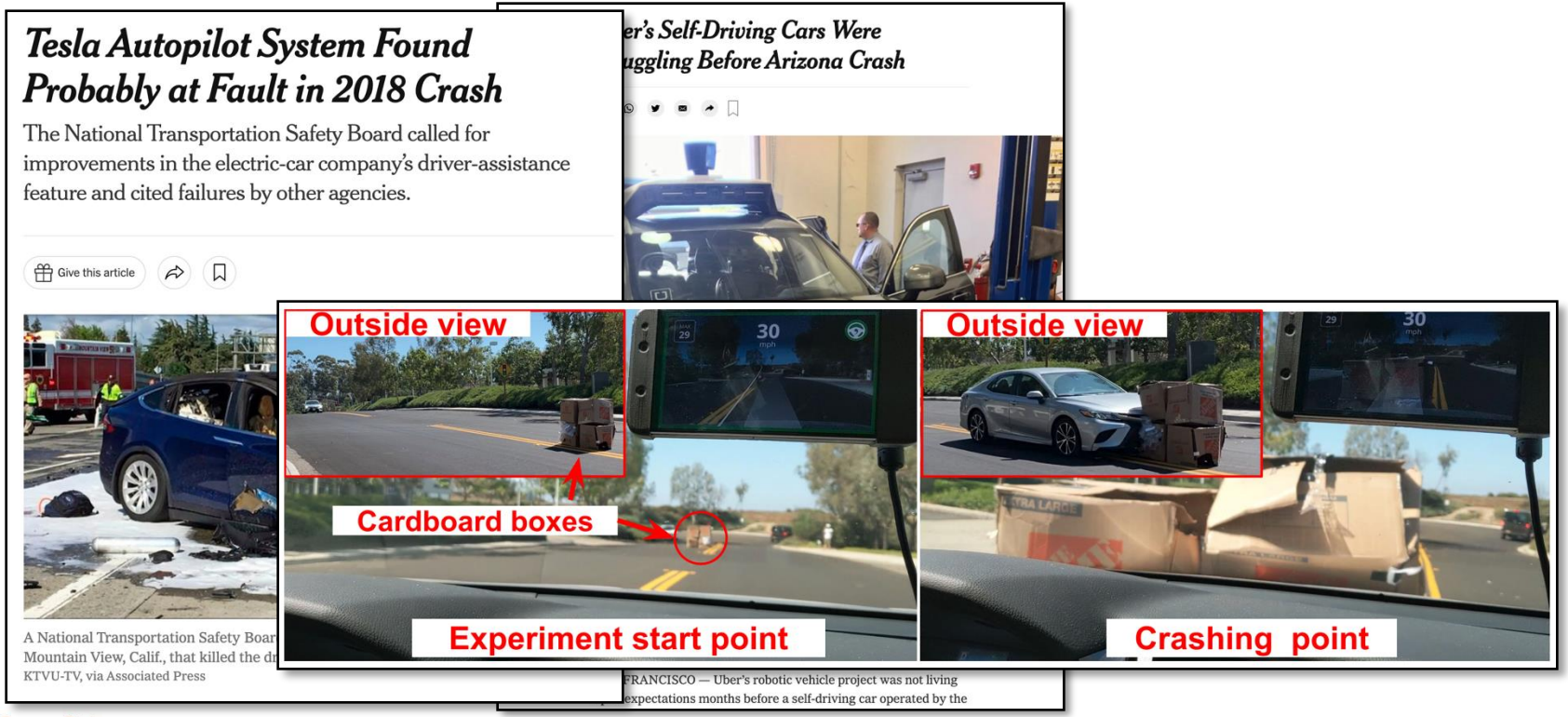




[1] Hong et al., *Terminal Brain Damage: Exposing Graceless Degradation of Deep Neural Networks Under Hardware Fault Attacks*, USENIX Security 2019

WHAT TYPES OF THE POTENTIAL PROBLEMS THERE?

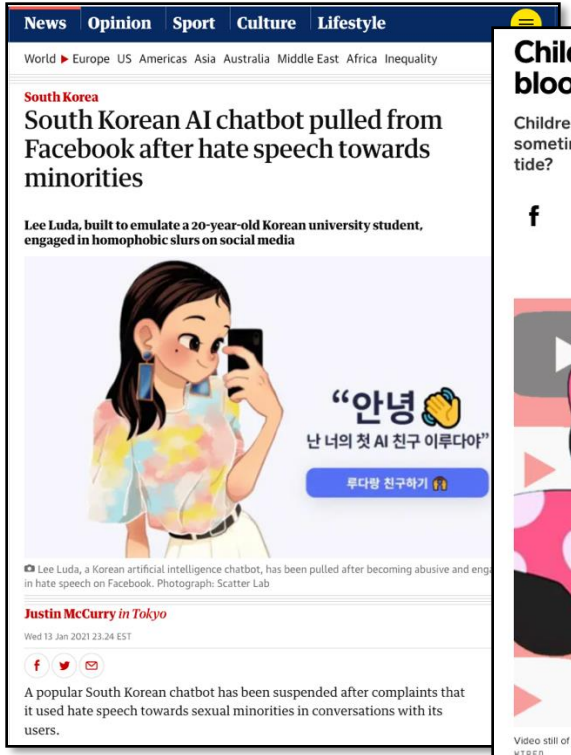
- Integrity: Robustness (or Terminal Brain Damage[1])

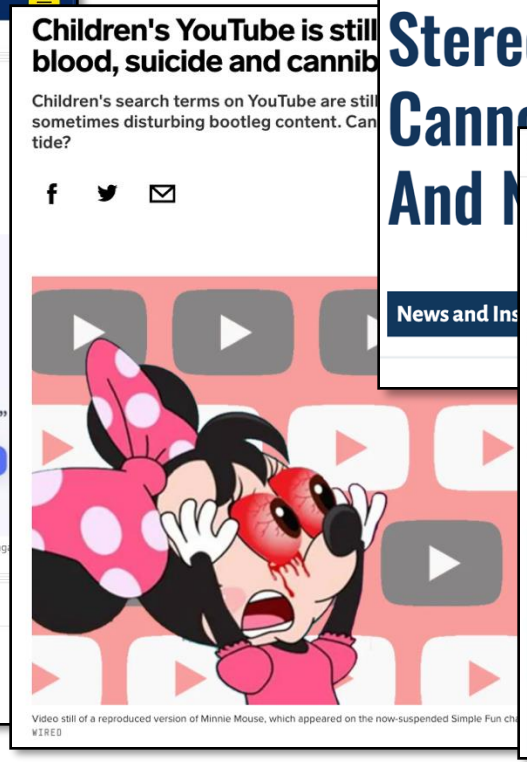


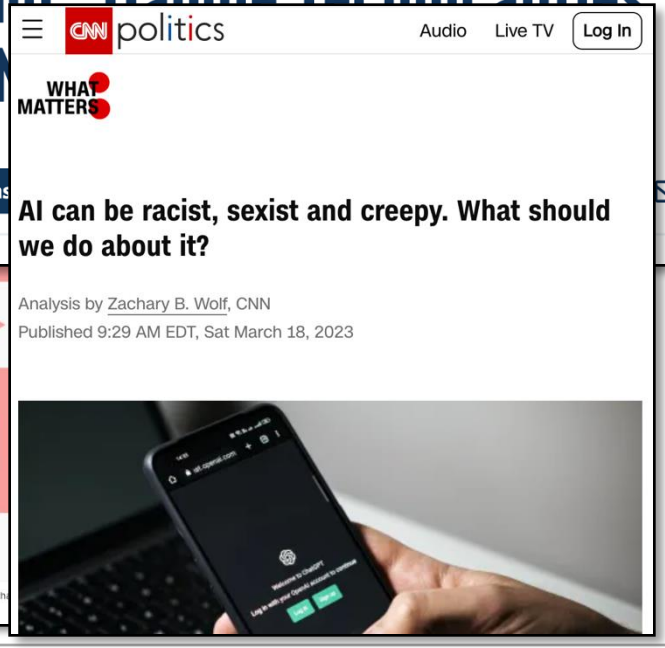
[1] Hong et al., *Terminal Brain Damage: Exposing Graceless Degradation of Deep Neural Networks Under Hardware Fault Attacks*, USENIX Security 2019

WHAT TYPES OF THE POTENTIAL PROBLEMS THERE?

- More issues: fairness or explainability







HERE IS HOW YOU'LL LEARN

OVERVIEW

- Course overview:
  - 4 credit courses: 12 hours of effort per week
  - Couse website: https://secure-ai.systems/courses/MLSec/current
- Contacts:
  - Personal matters: email to sanghyun.hong@oregonstate.edu
  - Course-related: F 3 – 3:50 pm (on Zoom)
  - Class submissions: HotCRP and Canvas
- Computing resources (GPUs):
  - OSU HPC: https://it.engineering.oregonstate.edu/hpc
  - OSU EECS: https://eecs.oregonstate.edu/eecs-it#Servers
  - **[Required]** Email Sanghyun by Thursday if you don't have access to the cluster

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 companies

COURSE STRUCTURE

- 10-week schedule; no textbook
  - Course syllabus is up: https://secure-ai.systems/courses/MLSec/current
  - **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/current
  - **Week 1:** Introduction & Overview
  - **Week 2-4:** Adversarial examples
  - **Week 5-7:** Data poisoning
  - **Week 8-10:** Privacy risks
- Heads-up
  - A few classes will be on Zoom
  - Please check the syllabus or the Canvas announcements

COURSE STRUCTURE – CONT'D

- In this course, you will do
  - 30%: 15-16 written paper critiques
  - 20%: 4 homework
  - 10%: 1 in-class presentation (must complete sign-ups in the 1st week)
  - 30%: 1 term-project (must complete team-ups in the 1st week)
  - 20%: 1 final Exam (multiple trials available; for 24 hours)

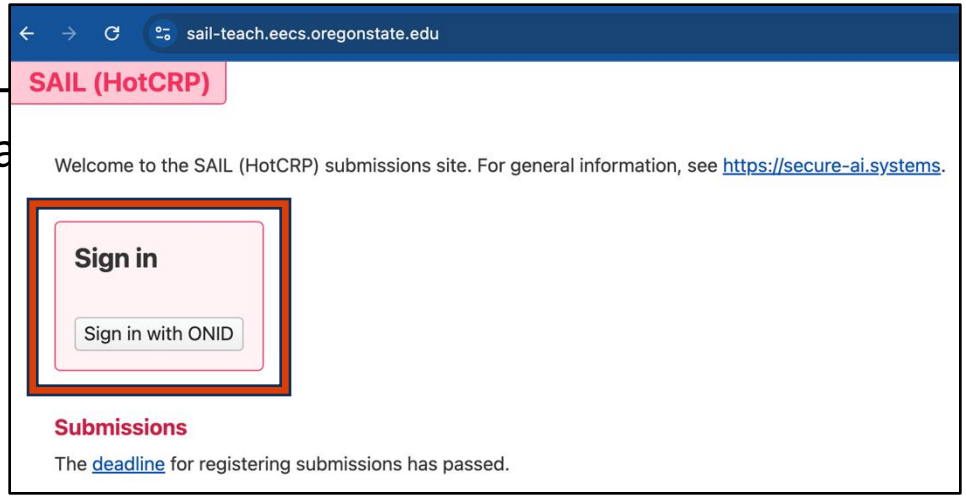
- [Bonus] You will also have extra points opportunities
  - + 5%: Outstanding project work
  - + 5%: Submitting the final report to workshops
  - … (will be more)

30%: WRITTEN PAPER CRITIQUES

- **[Due]** Before each class (hard deadline)
- You need to:
  - Pick a paper
  - Submit your review on **HotCRP**

30%: WRITTEN PAPER CRITIQUES

- **[Due]** Before each class (hard dea
- You need to:
  - Pick a paper
  - Submit your review on **HotCRP**



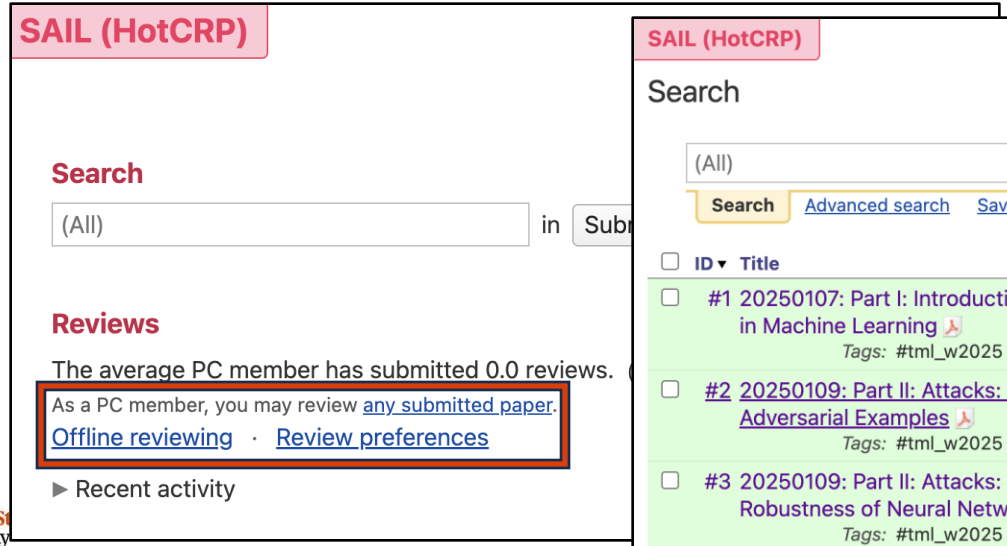
- **HotCRP!**
  - https://sail-teach.eecs.oregonstate.edu (only accessible on Campus / via VPN)
  - You must register this system now!
    (Sanghyun will assign papers to you tomorrow)

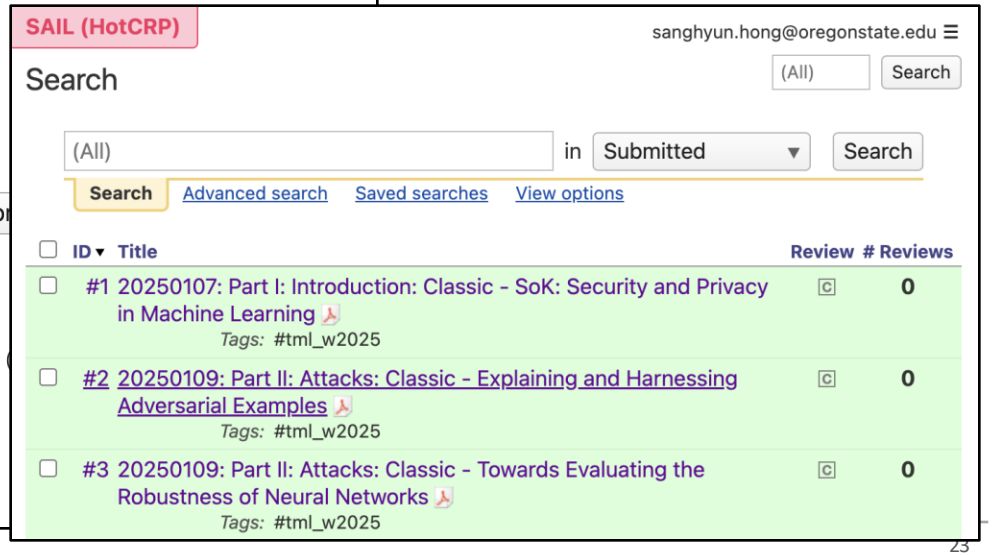
30%: WRITTEN PAPER CRITIQUES

- **[Due]** Before each class (hard deadline)
- **HotCRP!**
  - https://sail-teach.eecs.oregonstate.edu (only accessible on Campus / via VPN)
  - You must register this system now!
    (Sanghyun will assign papers to you tomorrow)

30%: WRITTEN PAPER CRITIQUES

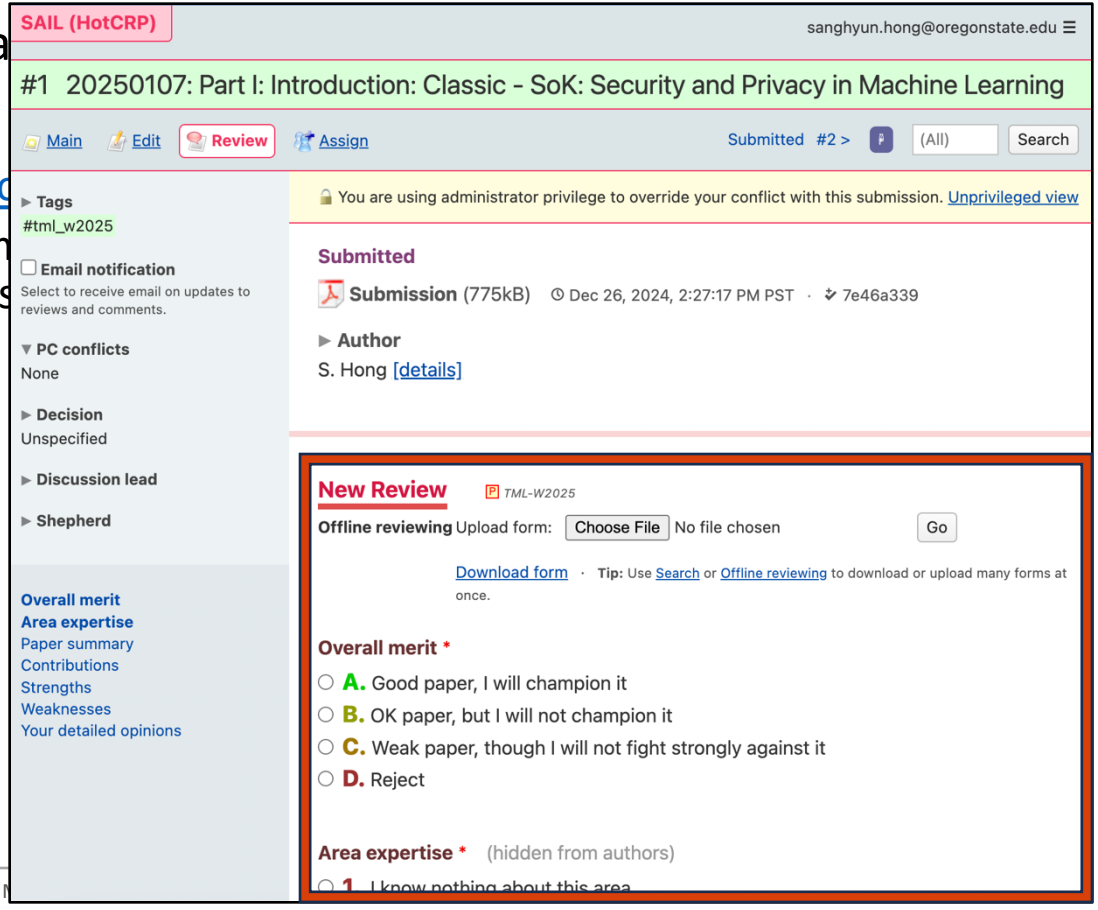
- **[Due]** Before each class (hard deadline)
- **HotCRP!**
  - https://sail-teach.eecs.oregonstate.edu (only accessible on Campus / via VPN)
  - You must register this system now!
    (Sanghyun will assign papers to you tomorrow)





30%: WRITTEN PAPER CRITIQUES

- **[Due]** Before each class (ha
- **HotCRP!**
  - https://sail-teach.eecs.oregc
  - You must register this systen (Sanghyun will assign papers
  - Your review should include
    - Merit / expertise
    - Summary
    - Contributions
    - Weaknesses
    - Strengths
    - Your opinions



30%: WRITTEN PAPER CRITIQUES

- **[Due]** Before each class (hard deadline)
- **HotCRP!**
  - https://sail-teach.eecs.oregonstate.edu (only accessible on Campus / via VPN)
  - You must register this system now!
    (Sanghyun will assign papers to you tomorrow)
  - Your review should include
    - Merit / expertise
    - Summary, contributions, weaknesses, strengths, your opinions
  - **[Must]**
    - This is **not** a pleasant reading
    - Must look at an example at: https://secure-ai.systems/courses/MLSec/current/critiques.html
  - Grades: 0 / 1 / 2

20%: HOMEWORK

- 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 file

10%: IN-CLASS PAPER PRESENTATION

- 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

- You will form a team of max. 4 students
  - You are welcome to do this alone
  - Use Canvas to sign-up (should be done in the first week)
- 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: HotCRP]**

COURSE STRUCTURE – CONT'D

- In this course, you will do
  - 30%: 15-16 written paper critiques
  - 20%: 4 homework
  - 10%: 1 in-class presentation (must complete sign-ups in the 1st week)
  - 30%: 1 term-project (must complete team-ups in the 1st week)
  - 20%: 1 final Exam (multiple trials available; for 24 hours)

- [Bonus] You will also have extra points opportunities
  - + 5%: Outstanding project work
  - + 5%: Writing a critique using ChatGPT
  - +10%: Submitting the final report to workshops

"GENEROUS" GRADING POLICY

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

LATE SUBMISSION POLICY

- Written paper critiques:
  - No submission in any case: 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
- Final Exam:
  - No submission in any case: 0 pts

KEEP AN EYE ON THE COURSE WEBSITE AND CANVAS

- You will find updates such as:
  - New announcements
  - Changes in our course schedule (or structure)

Thank You!

Sanghyun Hong

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

SAIL
Secure AI Systems Lab