

SP2023 Week 09 • 2023-03-23

AI Hacking I

Anusha



Announcements

- We're playing LINE CTF!
 - It starts tomorrow, please come play with us!
- No meeting next Thursday
 - We're off to CypherCon :)
- Fill out feedback form for our research paper please
 - <https://forms.gle/kYg16ZJicwuVwTca6>



ctf.sigpwny.com

sigpwny{dan_says_hello}



Overview

- What is "AI security"?
 - using AI for security purposes
 - exploring the security of AI itself
 - AI ethics/security around AI
- All of them are awesome fields, but be aware that "AI security" can depend on who you're talking to!



AI for Security



AI for Malware

- Using AI to find malware and shut down execution
 - scans code before execution to determine malware potential
- AI for network monitoring
 - scans packets to monitor for attacks
- AI for anything else
 - pretty much anything that requires human foresight/monitoring



Pros/Cons

- Good things!
 - potential to find undiscovered malware and patch it
 - discover bugs before they can be exploited
 - continuous monitoring without continuous human involvement
- Bad things :(
 - potential to harmfully misclassify
 - black box
 - does accuracy generalize to the real world?



Security of AI



Security of AI

- How to find ways to make models behave incorrectly on purpose
 - key word here: adversarial
- Many different ways to get models to misbehave
 - small differences in input can completely change output!



LLM Attacks

- Prompt injection/other attacks popularized by open access to models like ChatGPT
- How can we make LLMs misbehave? Can we gain access to unsanitized output? What are the harms of doing so?



GAN Attacks

- Are there ways for images included in a dataset to “poison” the outputs of a GAN?
 - Might be able to enforce ethical collection of art/images for training datasets
 - See [GLAZE](#) for more info!
- Can we find ways to produce unsanitized output on GANs? What would that look like? How is this dangerous?



Detector/Classifier Attacks

- Can we change the input to a detection/classification model to make it misclassify outputs?
 - [FAWKES](#) does it for facial recognition models
- Many papers have shown how adversarial inputs can produce altered output



Privacy Preserving AI



Privacy Preserving AI/ML

- Data collection side
 - How do you select proper datasets?
- Data processing side
 - Can we make training data more private?
- Architecture side
 - Model federation
 - How to train on decentralized data
 - Model unlearning



Career Paths

- Research
 - Industrial research with a company
 - Staying in academia
- AI Engineering
- AI red teaming



AMA



Next Meetings

2023-03-24 - Tomorrow

Come play LINE CTF with us! We will be playing both virtually and in person.

2023-03-26 - This Sunday

- AI Hacking II
- A more in-depth look at the technology behind today's talk

2023-03-30 - Next Thursday

- No meeting because of CypherCon



```
sigpwny{dan_says_hello}
```



SIGPwny