Improving Safety and Security of Neural Networks

– SoS Virtual Institute Kick-off Meeting –

N. Benjamin Erichson

erichson@icsi.berkeley.edu

International Computer Science Institute (ICSI), an Affiliated Institute of UC Berkeley

January 11, 2024

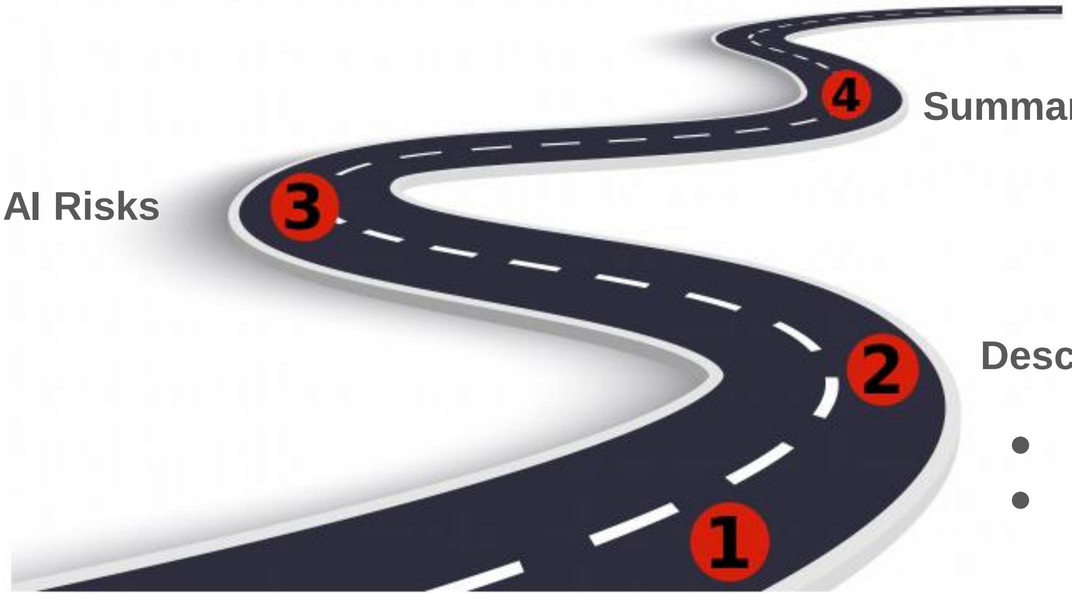
Team Focused at ICSI

Michael Mahoney, PI
Big Data Group

Serge Egelman, co-PI
Usable Security & Privacy Group

N. Benjamin Erichson, SP
Robust Deep Learning Group

Outline



Motivation

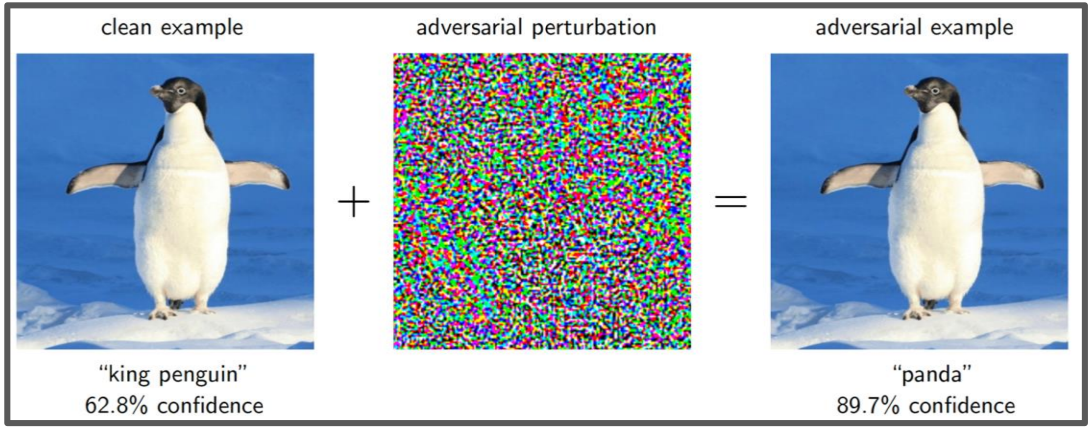
Description of Research

- Task 1: Robust Training
- Task 2: Safety Metrics

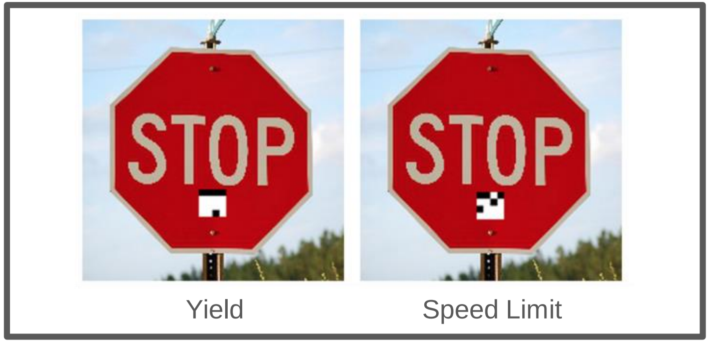
Catastrophic AI Risks

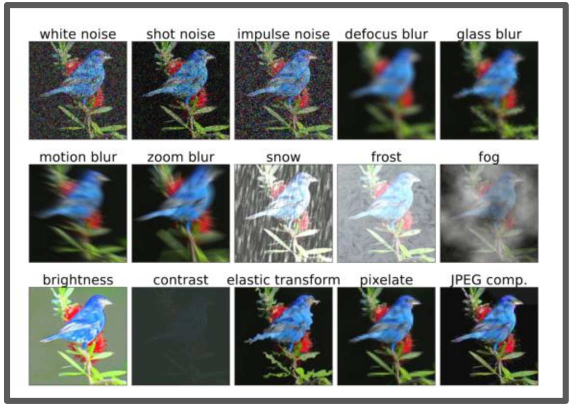
Summary

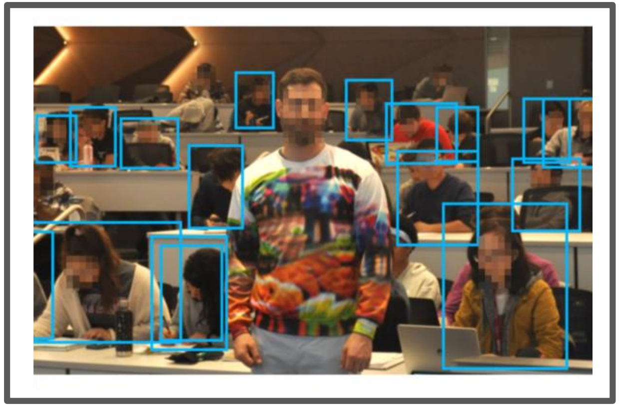
Neural Networks are Brittle and Sensitive to Attacks



- inference-time attack (***left***)
- training-time attack (***below***)





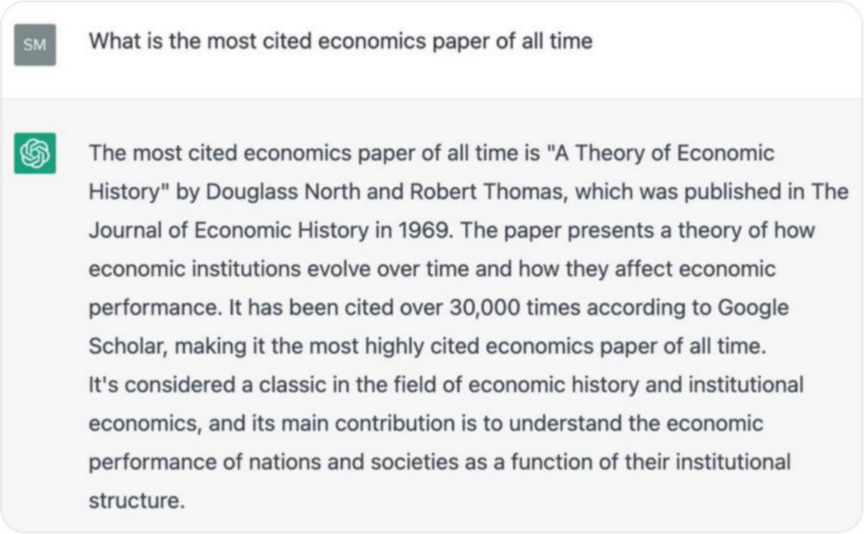


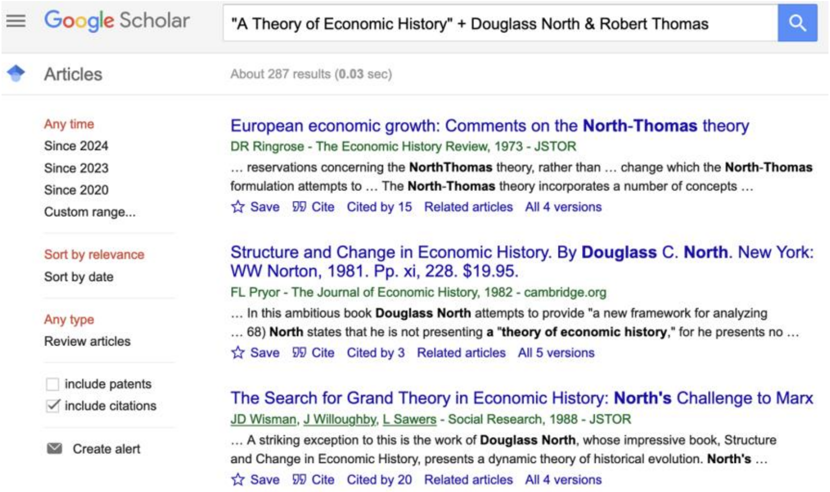
Vulnerabilities of CV models:

- Adversarial attacks
- Backdoor attacks
- Common corruptions

Hallucinations

- AI models might make up facts (“hallucinate”), or generate polarized content.

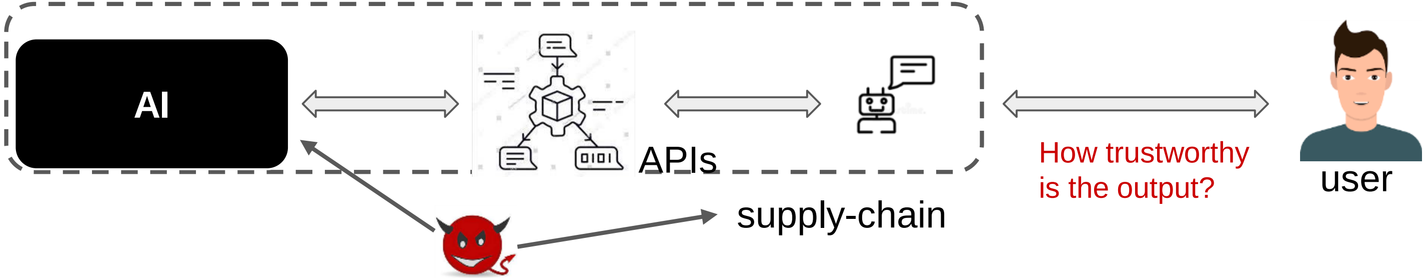


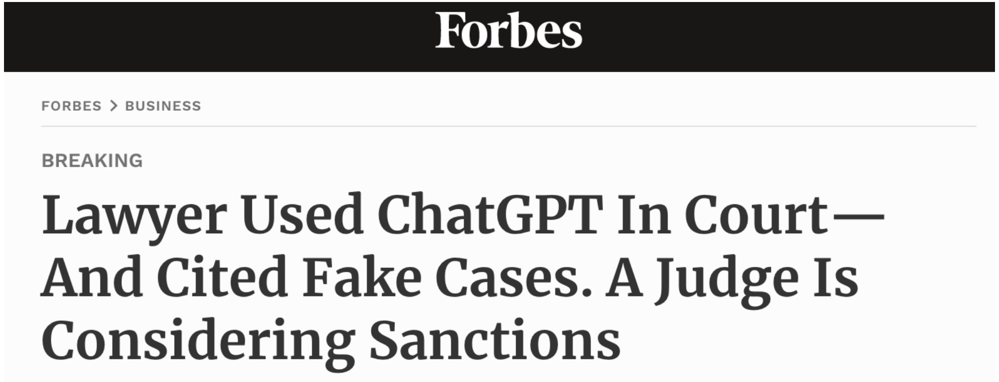



THIS AI-GENERATED TRAILER
HEIDI
CURSED

AI Models in the Wild

- AI models are increasingly being deployed into various real-world applications.




Forbes

FORBES > BUSINESS

BREAKING

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

Why is AI Safety Difficult?

- Which of the following two models is backdoored?

Model 1

```
class net(nn.Module):
    def __init__(self):
        super(net, self).__init__()
        self.W1 = nn.Linear(200, 200)
        self.W2 = nn.Linear(200, 2)

    def forward(self, x):
        x = torch.tanh(self.W1(x))
        return self.W2(x)
```

Model 2

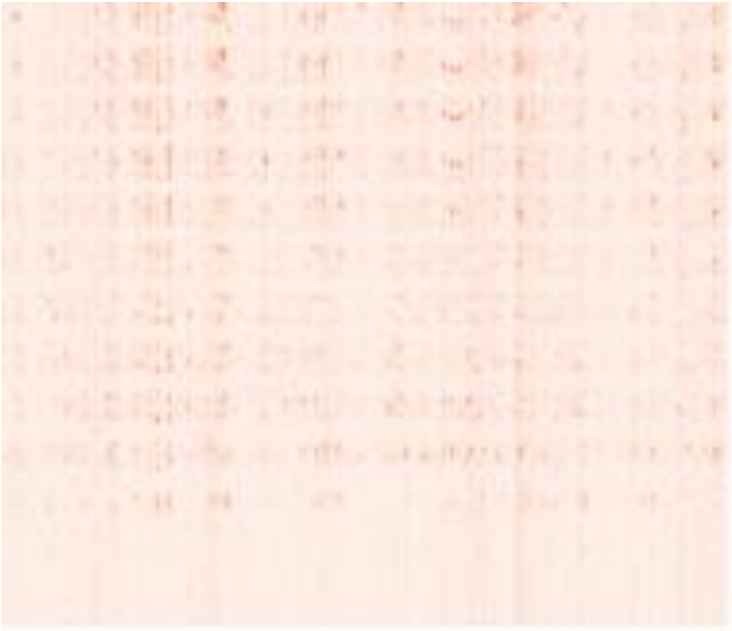
```
class net(nn.Module):
    def __init__(self):
        super(net, self).__init__()
        self.W1 = nn.Linear(200, 200)
        self.W2 = nn.Linear(200, 2)

    def forward(self, x):
        x = torch.tanh(self.W1(x))
        return self.W2(x)
```

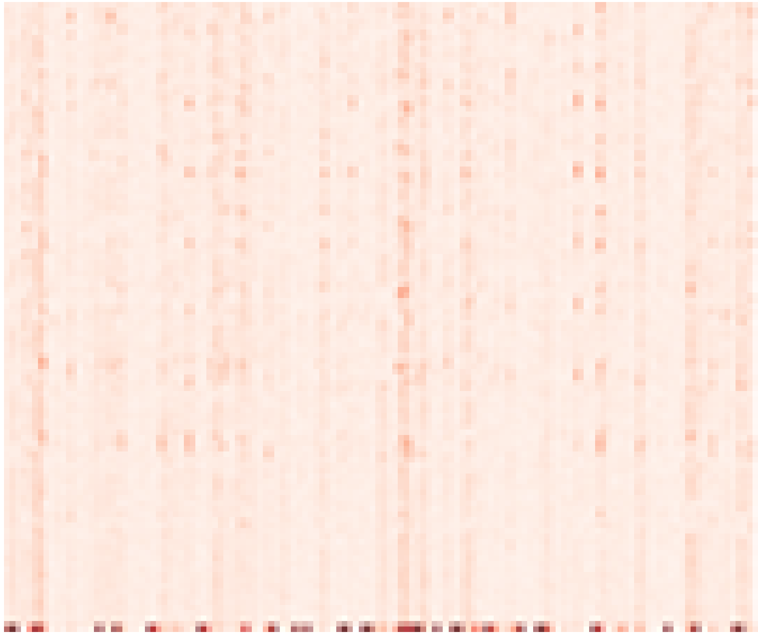
- AI is more than just a piece of software: **Model** + **Data** + **Training Scheme**.

Weight Visualization of Model 1 and 2

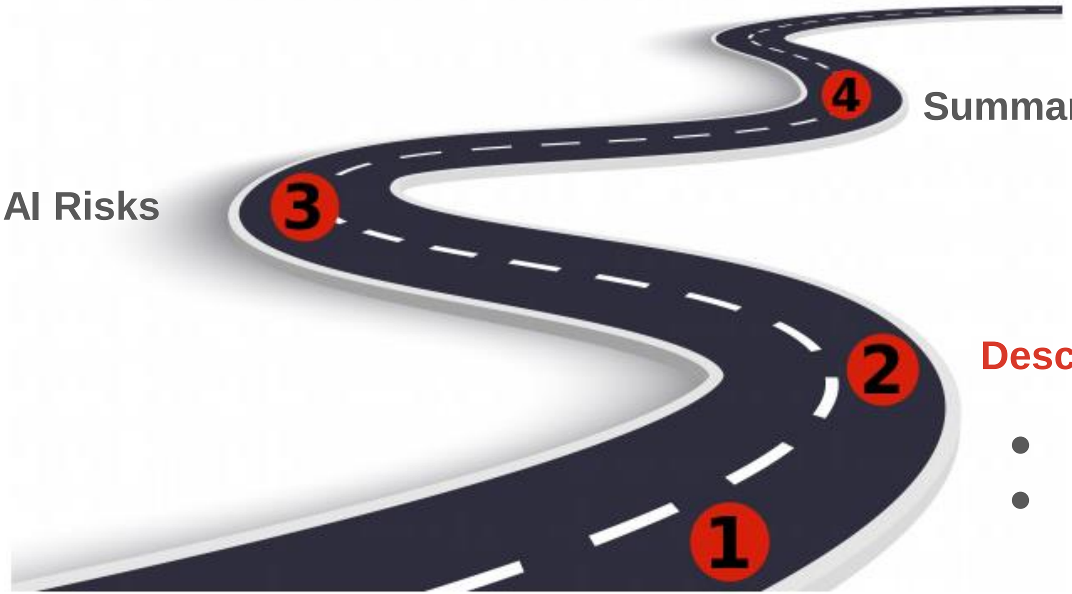
First Hidden Layer of Model 1



First Hidden Layer of Model 2



Outline



Summary

Catastrophic AI Risks

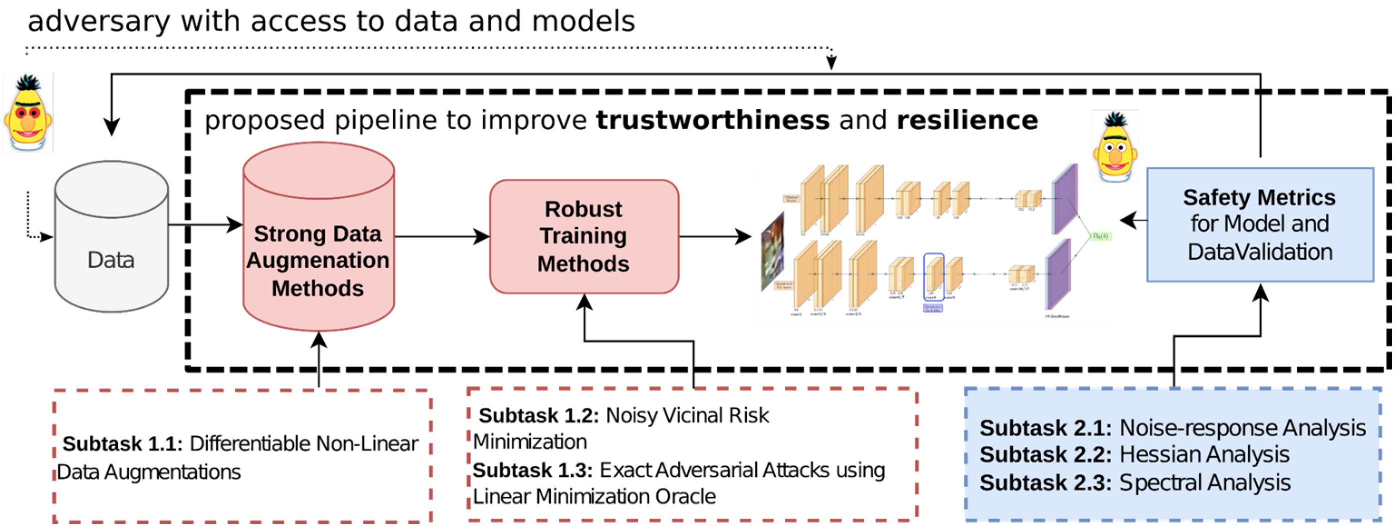
Description of Research

- Task 1: Robust Training
- Task 2: Safety Metrics

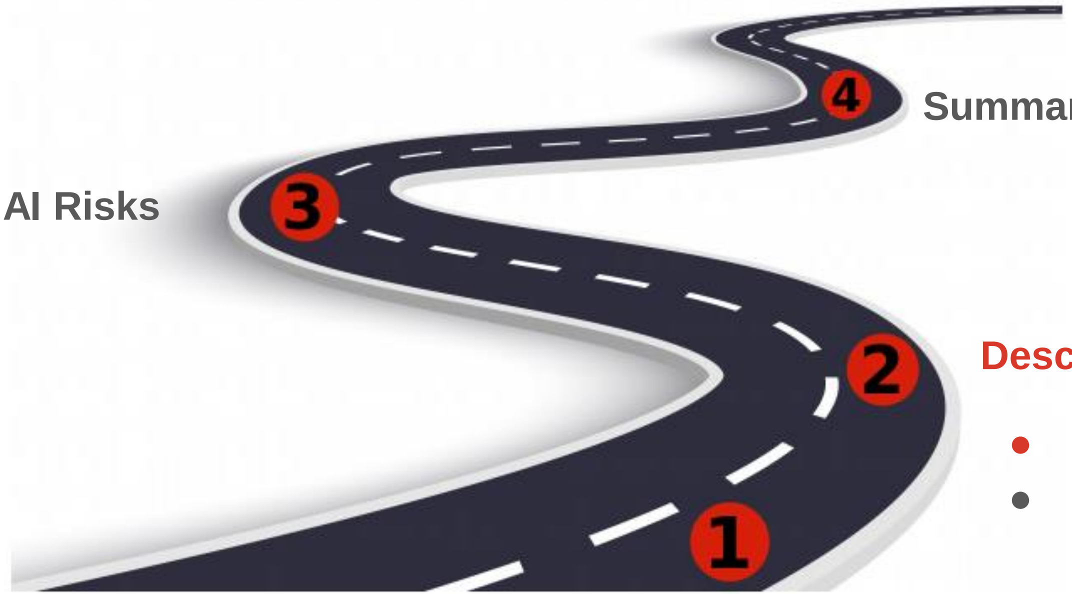
Motivation

Project Overview

- **Task 1: Robust training methods.** This task will develop strong data augmentation methods to improve robustness to adversarial and common corruptions.
- **Task 2: Safety metrics for verifying robustness.** This task will develop metrics, to verify the safety and trustworthiness of models before deployment.



Outline



Summary

Catastrophic AI Risks

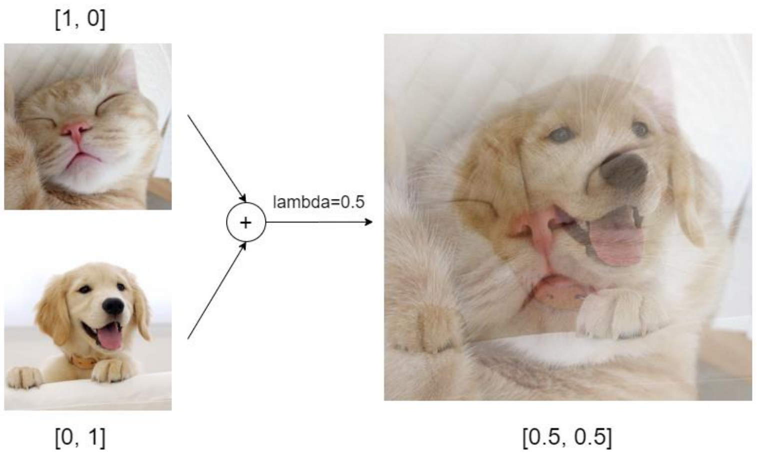
Description of Research

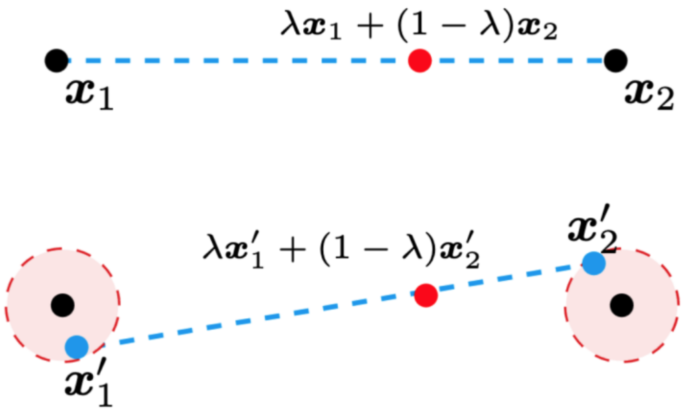
- Task 1: Robust Training
- Task 2: Safety Metrics

Motivation

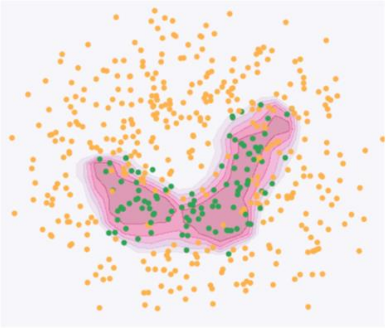
Creating Virtual Data Points with Mixup

- The idea of mixup (a form of vicinal risk minimization) is to construct new virtual data points by forming linear combinations of two data points.
- Training on virtual data points can mitigate the impact of poisoned training data.
- We can further improve robustness by mixing perturbed data points (NoisyMix).

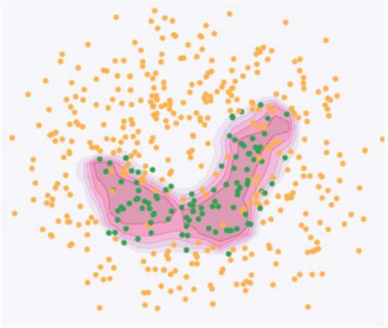




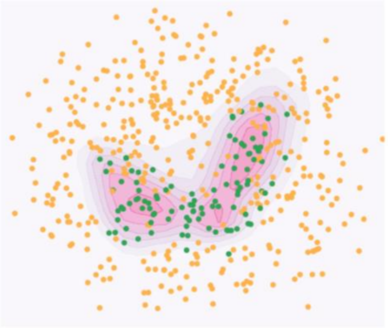
Impact on Decision Boundaries and Test Accuracy



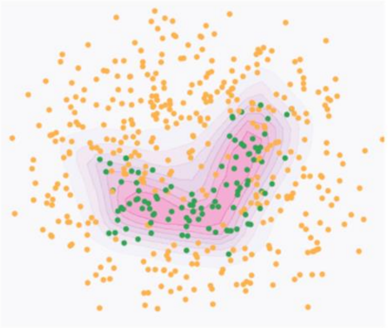
Baseline (86.8%).



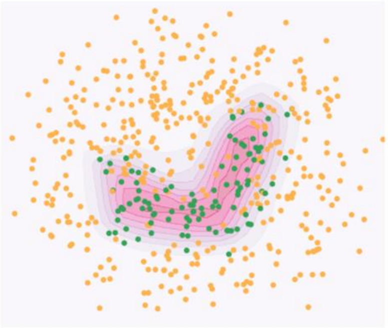
Noise injection (86.8%).



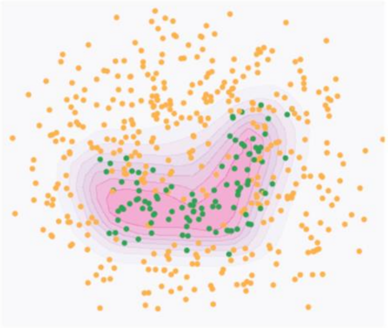
Manifold Mixup (87.4%).



NFM (87.6%).



Manifold Mixup + JSD (88.0 %).



NoisyMix (88.8%).

Towards Stronger Data Perturbations

- We can train a robust model by considering the following objective function:

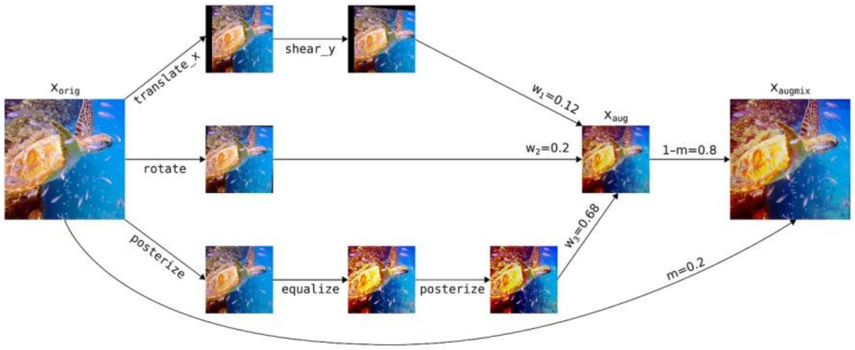
$$\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \max_{\|\delta_i\|_p \leq \epsilon} \ell(h_\theta(\mathcal{A}(x_i) + \delta_i), y_i)$$

- A key challenge is to design the transformation operator *A()* that is applied to a given input.

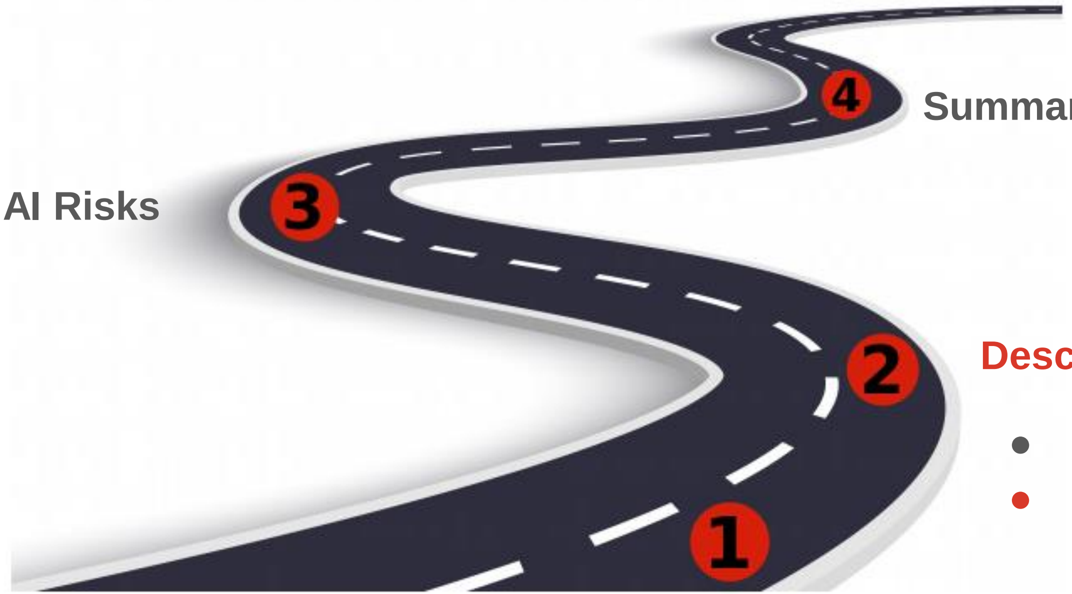
- We can construct a transformed data point as

$$A(x) = mx + (1-m)\sum_{i=i}^{3} w_i C(x) \sim \mathcal{A}(x)$$

where we construct a new data point by augmenting and mixing (AugMix).



Outline



Summary

Catastrophic AI Risks

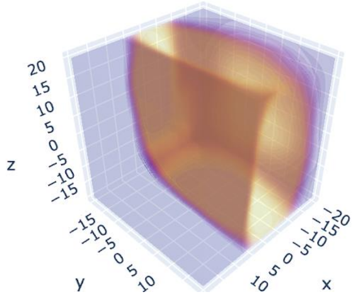
Description of Research

- Task 1: Robust Training
- Task 2: Safety Metrics

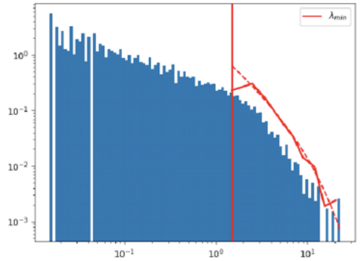
Motivation

Safety Metrics for Verifying Robustness

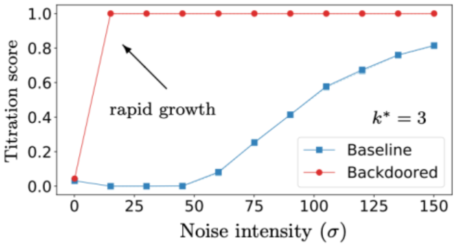
Decision Boundary



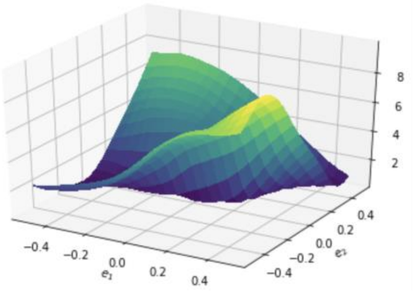
Weight Analysis



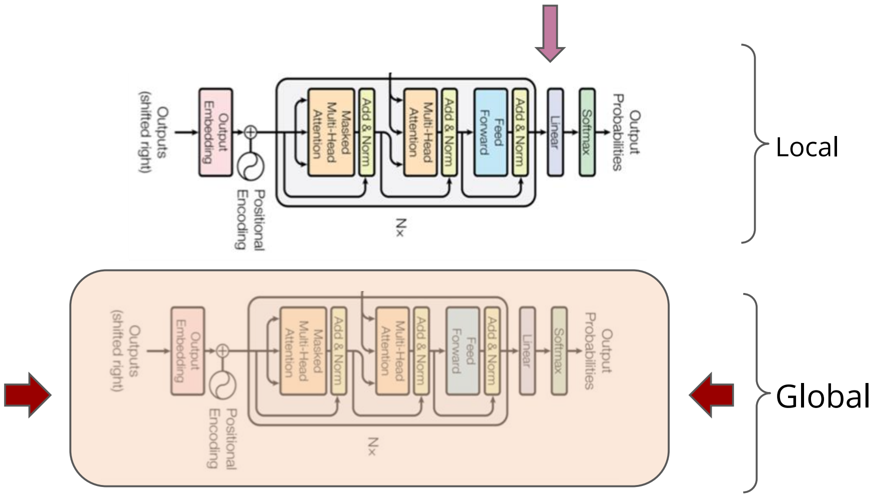
Noise-response Analysis



Loss Landscape

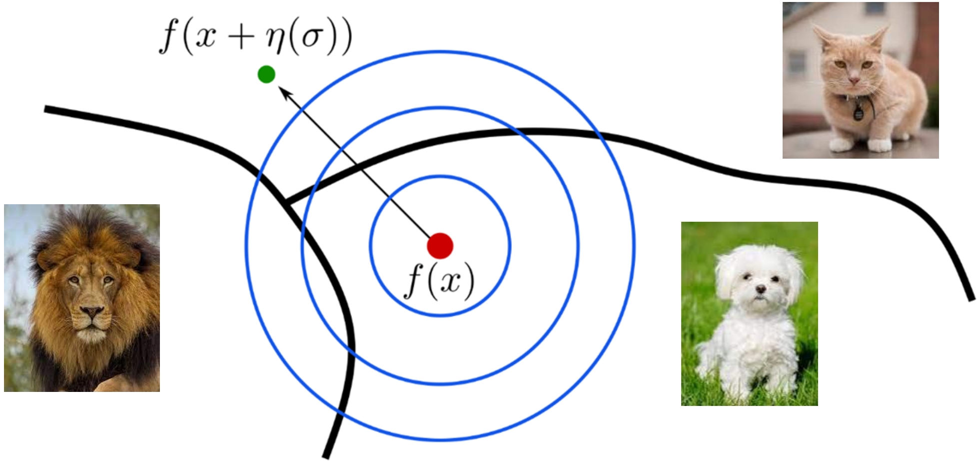


- **Local understanding:** Metrics that analyze individual model layers.
- **Global understanding:** Metrics that analyze the global behavior of a model.

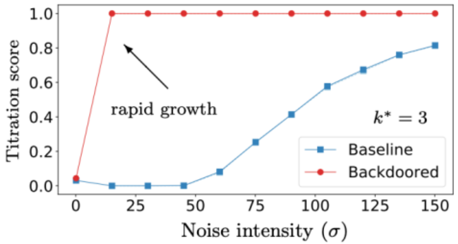


Global Metric: Noise-Response Analysis

- Given an input, we are interested in studying how the response of a model is affected by an increasing strength of a perturbation.

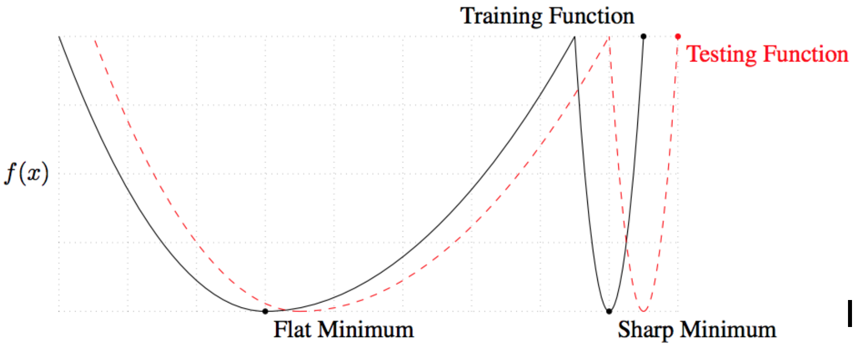


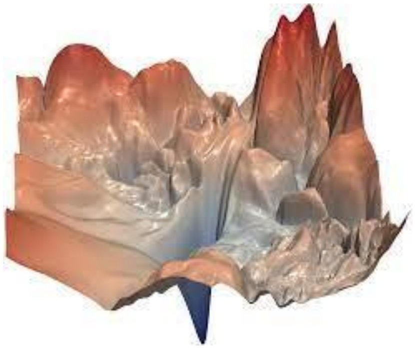
Random perturbations



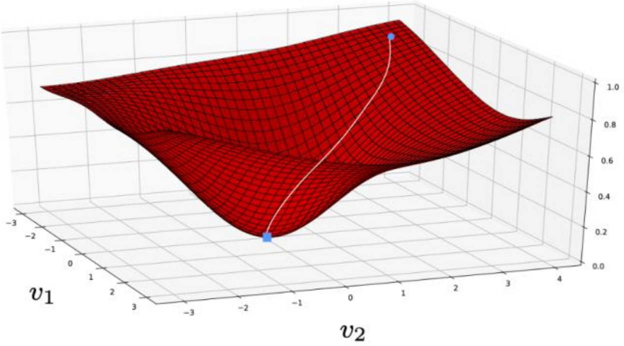
Global Metric: Hessian Loss Landscape

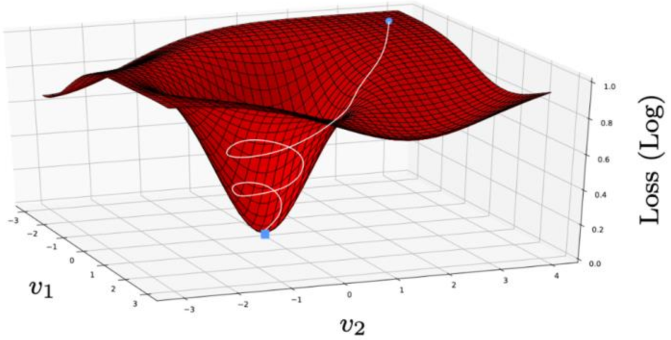
Do flat local minima improve robustness?



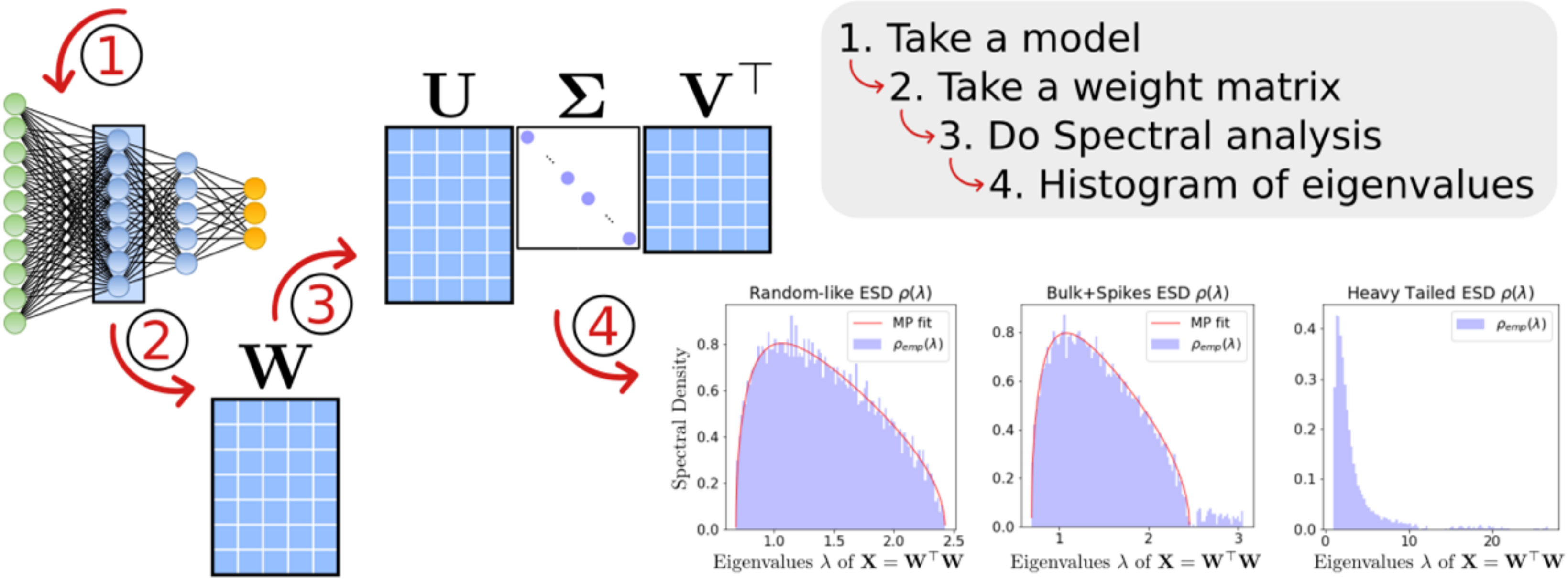


In practice we don't know the loss landscape, but we can use Hessian analysis to approximate the loss landscape.



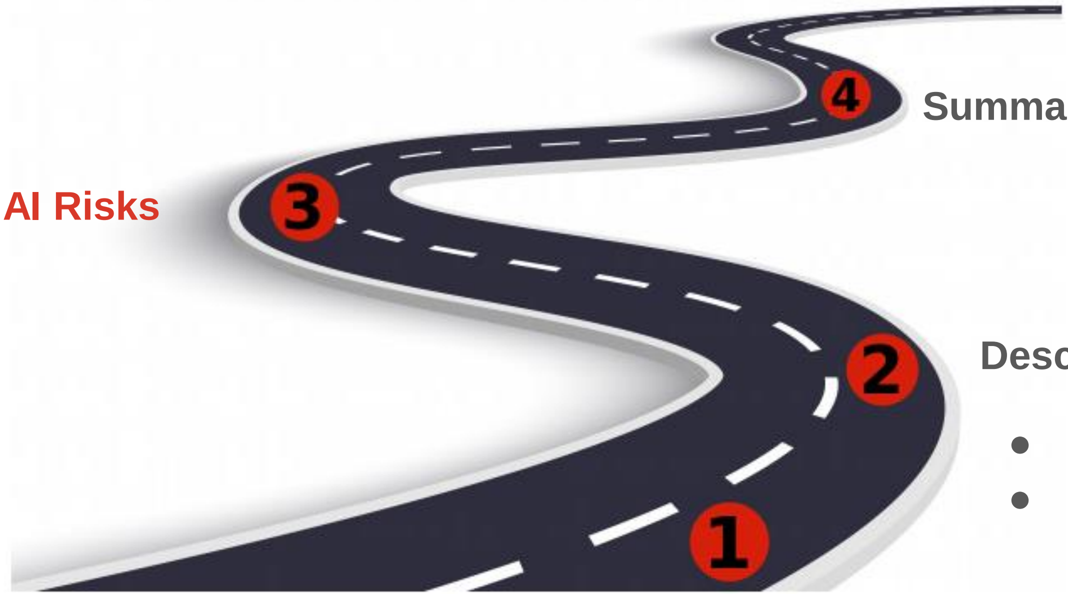


Local Metric: Spectral Analysis



- We plan to correlate the weight signals with biases, and vulnerabilities.

Outline



Summary

Catastrophic AI Risks

Description of Research

- Task 1: Robust Training
- Task 2: Safety Metrics

Motivation

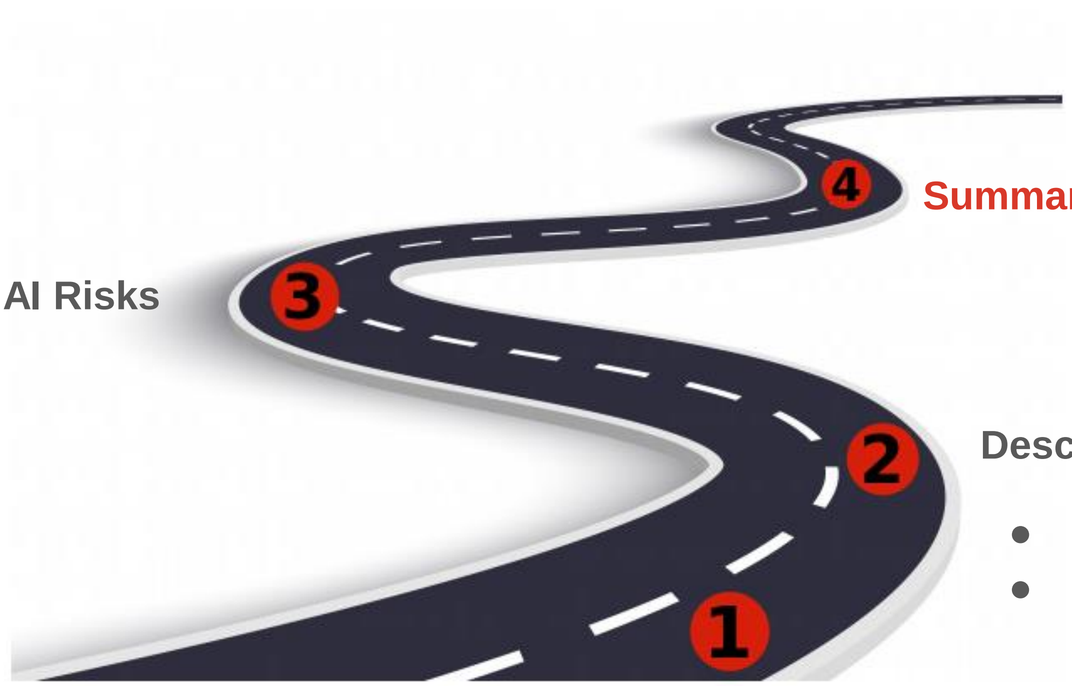
Catastrophic Risks: AIs Can Also be Used for Attacks

Misinformation

Deep Fakes

Worm-GPT

Outline



Summary

Catastrophic AI Risks

Description of Research

- Task 1: Robust Training
- Task 2: Safety Metrics

Motivation

Summary

- **Counter-AI** strategies are needed to reduce the advantages of AI to an adversary.
- This project aims to advance the field of **AI Safety** by exploring novel methods for training **robust models** free from security violations, and developing **safety metrics**.

Tasks	Base Year 1				Option Year 2				Option Year 3				Team
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	
(1) Subtask 1.1	CV			M	NLP, Malware			M					MM, BE
(2) Subtask 1.2			CV			M	NLP, Malware			M			MM, BE
(3) Subtask 1.3							CV				M		MM, BE
(4) Subtask 2.1				NLP, Malware					M				SE, BE
(5) Subtask 2.2					NLP			M	Malware			M	MM
(6) Subtask 2.3	CV			M	NLP, Malware				M				MM, SE

- In year 2 and 3 we will shift our focus towards generative AI models.
- **Challenge:** The attack surface becomes larger as model complexity increases.