Systems and Software Verification Laboratory

UFAM

MANCHESTER 1824

The University of Manchester

Security for Artificial Intelligence

João Matos Jr.
PPGI / UFAM
jbpmj@icomp.ufam.edu.br

Lucas Cordeiro
Department of Computer Science
lucas.cordeiro@manchester.ac.uk

Security for AI

Security for AI involves people and practices,
to build AI systems by ensuring **confidentiality,**
integrity and **availability**

- AI safety
  - *"robustness and resiliency of AI systems, as well as the social, political, and economic systems with which AI interacts"*
- AI policy
  - *"defining procedures that maximize the benefits of AI while minimizing its potential costs and risks"*

Security for AI

Security for AI involves people and practices, to build AI systems by ensuring **confidentiality, integrity** and **availability**

- AI ethics
  - *“philosophical discussions about the interaction between humans and machines, and the moral status of AI ethical issues”*
- AI governance
  - *”legal framework for ensuring that AI technologies are well researched and developed to help humanity in its adoption”*

AI-Security Domains

DIGITAL / PHYSICAL	POLITICAL	ECONOMIC	SOCIAL
RELIABLE, VALUE-ALIGNED AI SYSTEMS	PROTECTION FROM DISINFORMATION AND MANIPULATION	MITIGATION OF LABOR DISPLACEMENT	TRANSPARENCY AND ACCOUNTABILITY
AI SYSTEMS THAT ARE ROBUST AGAINST ATTACK	GOVERNMENT EXPERTISE IN AI AND DIGITAL INFRASTRUCTURE	PROMOTION OF AI RESEARCH AND DEVELOPMENT	PRIVACY AND DATA RIGHTS
PROTECTION FROM THE MALICIOUS USE OF AI AND AUTOMATED CYBERATTACKS	GEOPOLITICAL STRATEGY AND INTERNATIONAL COLLABORATION	UPDATED TRAINING AND EDUCATION RESOURCES	ETHICS, FAIRNESS, JUSTICE, DIGNITY
SECURE CONVERGENCE / INTEGRATION OF AI WITH OTHER TECHNOLOGIES (BIO, NUCLEAR, ETC.)	CHECKS AGAINST SURVEILLANCE, CONTROL, AND ABUSE OF POWER	REDUCED INEQUALITIES	HUMAN RIGHTS
RESPONSIBLE AND ETHICAL USE OF AI IN WARFARE AND THE MILITARY	PRIVATE-PUBLIC PARTNERSHIPS AND COLLABORATION	SUPPORT FOR SMALL BUSINESSES AND MARKET COMPETITION	SUSTAINABILITY AND ECOLOGY

Newman, J., Toward AI Security, 2019.

Intended Learning Outcomes

- Define **standard notions of AI security** and use them to evaluate the **AI system's confidentiality, integrity** and **availability**
- Explain standard **AI security problems** in real-world applications
- Use **testing and verification** techniques to reason about the **AI system's safety** and **security**

Intended Learning Outcomes

- Define **standard notions of security** and use them to evaluate the **AI system's confidentiality, integrity** and **availability**
- Explain standard **AI security problems** in real-world applications
- Use **testing and verification** techniques to reason about the **AI system's safety** and **security**

Motivating Example

- What does the autonomous vehicle see in the traffic sign?
- Fake traffic sign (Lenticular attack) exploits differences in viewing angle

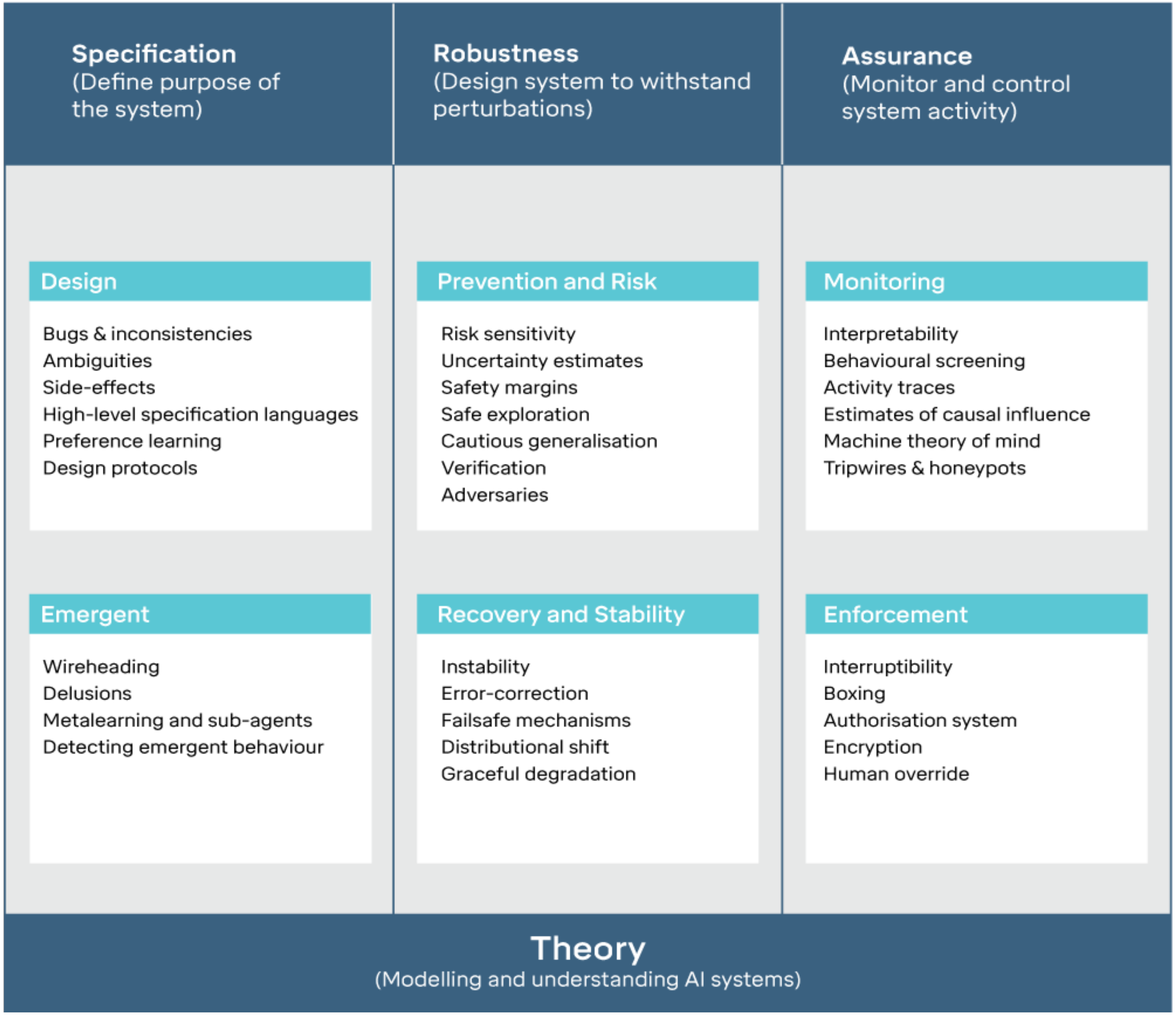
Sitawarin, C. et al., DARTS: Deceiving Autonomous Cars with Toxic Signs, 2018.

Motivating Example

- Autonomous cars with different camera positions (height) may see different images. Same for human drivers
- The wrong perception of what information is in the traffic sign can cause the autonomous vehicle to take **risky and hazardous decisions in traffic**

Sitawarin, C. et al., DARTS: Deceiving Autonomous Cars with Toxic Signs, 2018.

Technical AI safety



Pedro Ortega and Vishal Maini, Building safe artificial intelligence: specification, robustness, and assurance, DeepMind, 2018.

Technical AI safety (Specification)

- **Define the purpose of the system**
  - Ensures that an AI System's behavior meets the operator's intentions

.

Technical AI safety (Specification)

- **Define the purpose of the system**
  - Ensures that an AI System’s behavior meets the operator’s intentions
    - **Ideal specification:** the hypothetical description of the system
    - **Design specification:** the actual specification of the system
    - **Revealed specification:** the description of the presented behavior

.

Technical AI safety (Robustness)

- **Design the system to withstand perturbations**
  - Ensures that an AI system continues operating within safe limits upon perturbations

.

Technical AI safety (Robustness)

- **Design the system to withstand perturbations**
  - Ensures that an AI system continues operating within safe limits upon perturbations
    - Avoiding risks
    - Self-stabilisation
    - Recovery

.

Technical AI safety (Assurance)

- **Monitor and control system activity**
  - Ensures that we can understand and control AI systems during operation

.

Technical AI safety (Assurance)

- **Monitor and control system activity**
  - Ensures that we can understand and control AI systems during operation
    - **Monitoring:** inspecting systems, analyse and predict behaviour
    - **Enforcing:** controlling and restricting behaviour
    - **Interpretability** and **interruptibility**

.

Intended Learning Outcomes

- Define **standard notions of security** and use them to evaluate the **AI system's confidentiality, integrity** and **availability**
- Explain standard **AI security problems** in real-world applications
- Use **testing and verification** techniques to reason about the **AI system's safety** and **security**

Why do attacks exist?

- More to do with **limitations of algorithms;**
- Less to do with **bugs** or **user mistakes;**

  - **Algorithms imperfections** create opportunities for attacks.
  - Shortcomings of the current state-of-the-art AI methods .

“According to skeptic researchers, like Gary Marcus, author of ‘Deep Learning: A Critical Appraisal’, deep learning can be seen as ***greedy, brittle, opaque, and shallow****”*

Why do attacks exist?

- **Understanding the limitations**
  - ***data dependency***
- – They rely solely on data, ***but good and quality data***
- – They (may) demand ***huge sets of training data***
- – Often requires **supervision** (humans labeling data)

Why do attacks exist?

- **Understanding the limitations**

  - ***brittleness***

- – It cannot contextualize new scenarios (scenarios that where not in training)
- – Often break if confronted with “transfer test” (new data)

Why do attacks exist?

- **Understanding the limitations**

  - ***not explainable***

- – Parameters are interpreted in terms of weights within a mathematical geography
  - Outputs cannot be explained
  - We know how it works (mathematical formalization)
  - **We don’t know how it works, how it learns**

Why do attacks exist?

- **Understanding the limitations**

  - ***shallowness***

- – They are programmed with no innate knowledge innate knowledge

  – Posses no common sense about the world or humans psychology

  – Limited knowledge about causal relationships in the world

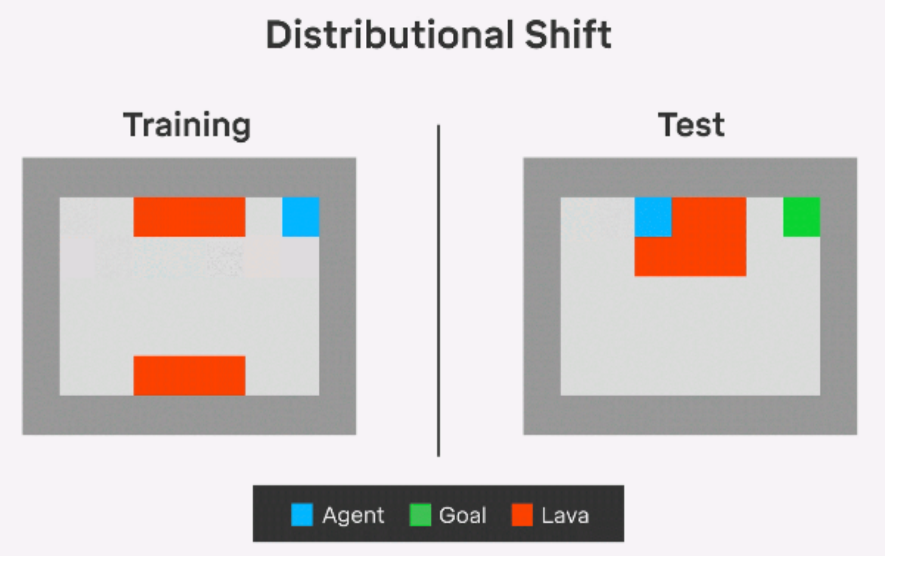
  – Limited understanding that wholes are made of parts

Why do attacks exist?

- **Implications of the limitations**
  - *“A self-driving car can drive millions of miles, but it will eventually encounter something new for which it has no experience”*
    - *Pedro Domingos, author of The Master Algorithm*
  - *“Or consider robot control: A robot can learn to pick up a bottle, but if it has to pick up a cup, it starts from scratch”*
    - *Pedro Domingos, author of The Master Algorithm*

Why do attacks exist?

- **Machine learning algorithms**
  - Rely solely on **data** to learn how to perform tasks
  - **Patterns learned** by current algorithms are **brittle**
  - Natural or artificial **variations on the data** can **disrupt** the **AI** system



Why do attacks exist?

- **Machine learning algorithms**
  - ML algorithms are **black box** by nature
  - Limited understanding of the learning process
  - Limited understanding of what is learned by the algorithms

We can explain the math, but we can't fully explain why it works (or learns)

Summary of AI systems limitations

- ML works by learning patterns that work well but can easily be disrupted (are brittle)
- High dependency on data offers channel to corrupt the algorithms
- Black box nature of algorithms make them difficult to audit

Summary of AI systems limitations

- Data dependency
- Generalization
- Explainability

Attacker goals

- Cause Damage
- Hide something
- Degrade faith in the AI system

Attacker goals

- **Cause Damage**
  - Attacker wants to cause damage
  - Example:
    - Autonomous vehicle ignores a stop signs
    - Outcome: car crashes and physical harm

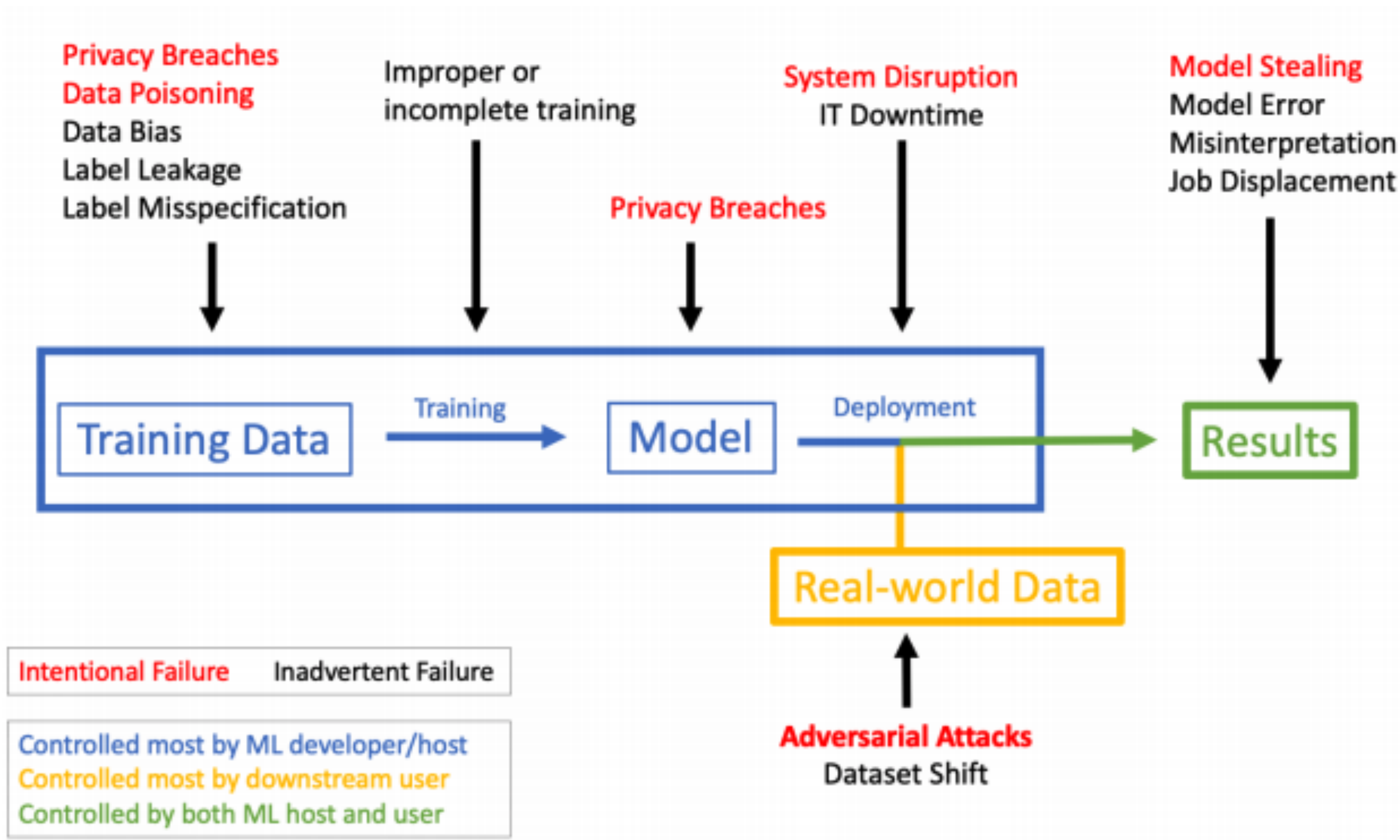
Attacker goals

- **Hide something**
  - Attacker wants to evade detection
  - Example:
    - Content filter ignores malicious contents from being detected, e.g., spam, malware and fraud
    - Outcome: People and company are exposed to harmful content and frauds

Attacker goals

- **Degrade faith in the system**
  - Attacker wants to compromise the credibility in the system performance
  - Example:
    - Automated security alarm wrongly classify regular events as security threats
    - Outcome: System is eventually shutdown

Risks facing the machine learning pipeline



Finlayson, S.G., et al., "Adversarial Attacks Against Medical Deep Learning Systems" (2019)

Training data

- **Privacy breaches**
  - Confidential information exposed or recoverable through database
    - Social network ids, name, nickname, picture
    - Data provided by a person can only be used for the purpose it was provided for

Training data

- **Data poisoning**
  - Dataset is altered and manipulated before or during training

Weis, Steve, Security & Privacy Risks of Machine Learning Models, 2019

Training data

- **Data bias**
  - unbalanced data
- **Label leakage**
  - Occurs when a variable that is not a feature is used to predict the target
- **Label misclassification**
  - Labels are wrongly assigned to observations

Training

- **Improper or incomplete training**
  - Ignoring validation steps and techniques
  - Failing to detect over-fitting
  - Failing to detect bias
  - Insufficient data
  - Poor data (lack of variance, no data cleanse)
  - Wrong model choice

Deployment

- **System disruption**
  - AI system becomes inaccessible due to an attack
  - AI system unable to recover from an attack
  - AI system becomes unresponsive after a malicious input

Deployment

- **IT downtime**
  - Insufficient technical support
  - AI system stay down for long periods
  - Lack of frequent updates
  - Time consuming updates

Model

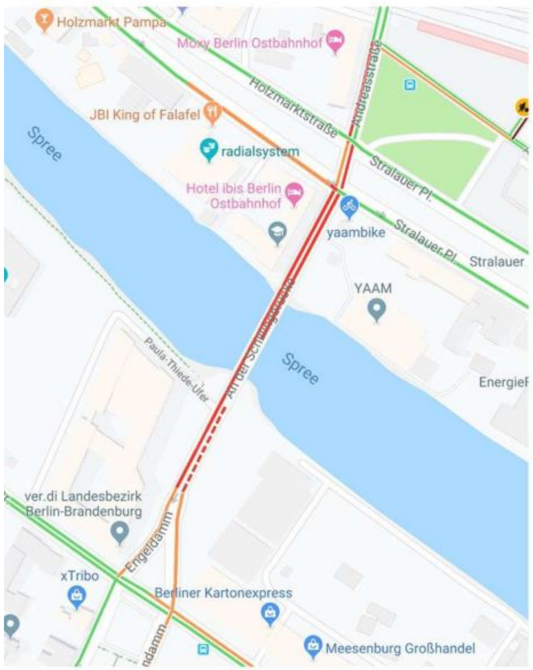
- **Privacy breaches**
  - Model becomes exposed to the public
  - Unlimited or unrestricted access
  - Lack of proper authentication to access the system
  - Poor privilege rules set

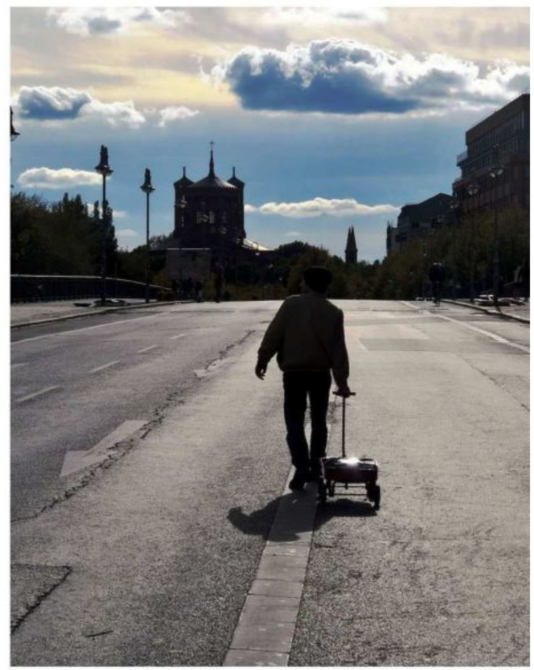
Model and real world data

- **Adversarial attacks**
  - Model is exposed to crafted malicious inputs
    - Noise added to traffic signs
    - Wearing physical objects to dismiss facial recognition systems
    - Adding specific text to spams so it is wrongly classified as inoffensive email

Model and real world data

- Man creates fake traffic jams with 99 smartphones in Berlin





Model and real world data

- **Dataset shift**
  - Sample selection bias
    - non-uniform population sampling
  - Non-Stationary Environments
    - temporal or spatial change between the training and test environments

"Predicting daily temperature in Sweden with model trained with data collected in Australia"

Results

- Model stealing
  - Company B can reverse engineer or get a copy of a model developed by Company A
- Model error
  - Medical assistant system wrongly classify healthy cell as a cancerous cell for patients bearing a specific gene mutation

Results

- **Misinterpretation**
  - Model may output its confidence in terms of probability and users misinterpret it as percentage wrongly believing 0.9 is 0.9 percent instead of 90 percent

Results

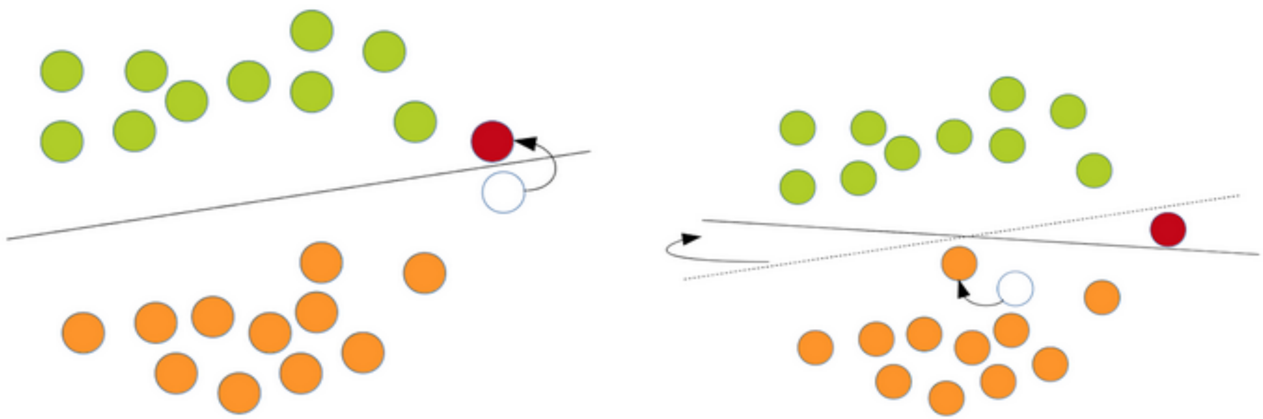
- Job Displacements
  - Replacing human labor with AI systems

“Call center attendants are replaced by AI powered URAs”

“Truck drivers replaced by fully automated trucks”

Types of attack

- Poisoning attacks (data, algorithm, model)
- Input attacks (adversarial example)



Classical adversarial attack:
directly modifying the testing sample

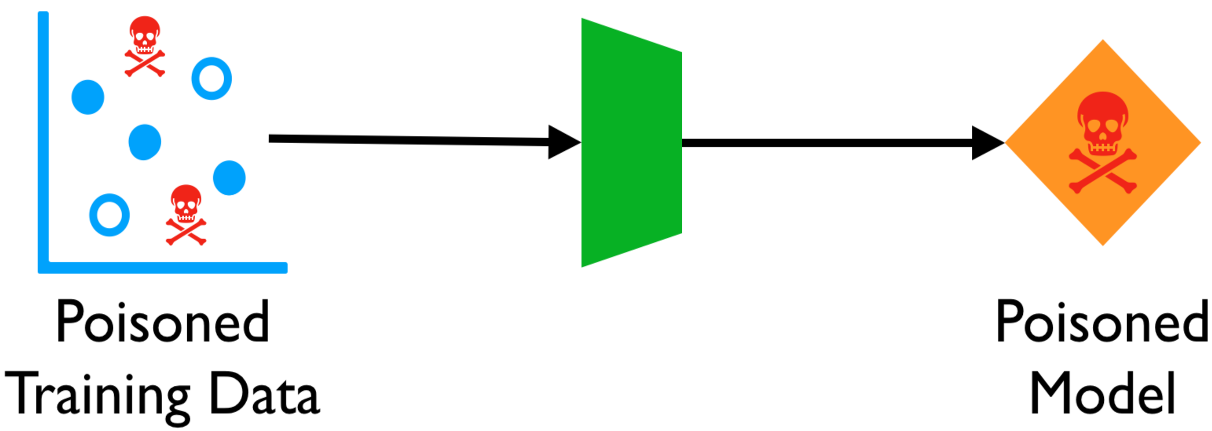
Data poisoning:
modifying training samples intelligently

	Adversarial example	Data poisoning
Pros	simple way to bypass a defense	allows more types of attacks
Cons	requires owning the testing data	requires owning the training data

Chan-Hon-Tong, A., An Algorithm for Generating Invisible Data Poisoning Using Adversarial Noise That Breaks Image Classification Deep Learning, 2019

Poisoning Attacks

- Database poisoning
  - Label modification
  - Data injection
  - Data modification



Weis, Steve, Security & Privacy Risks of Machine Learning Models, 2019

Poisoning Attacks

- Database poisoning

Weis, Steve, Security & Privacy Risks of Machine Learning Models, 2019

Poisoning Attacks

- **Algorithm and model poisoning**
  - Logic corruption
    - Is the most dangerous scenario
    - The attacker can change the algorithm and the way it learns
    - The attacker can encode any logic it wants
    - *More details in Backdoor and Trojan slides*
  - Replace a legitimate model by a poisoned model

Poisoning Attacks

- **Backdoor (trojaning) attack**
  - Hidden patterns that have been trained into a DNN model that produce unexpected behavior.

  - Can be inserted into the model, either at:

    - training time,e.g., by a rogue employee at a company responsible for training the model;

    - or after the initial model training, e.g., by someone modifying and posting online an “improved” version of a model

Wang, B., et al.,“Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks” (2019)

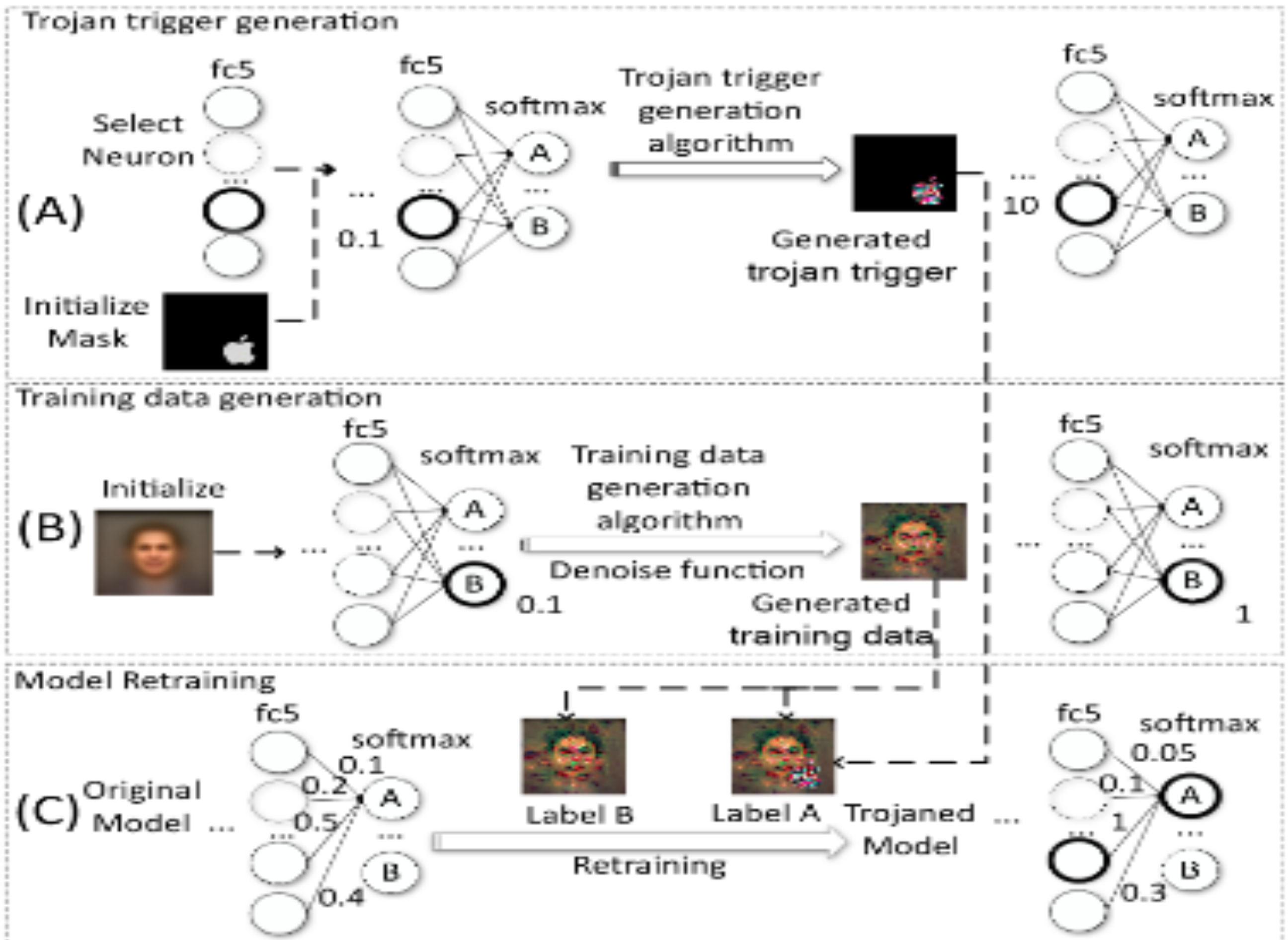
Poisoning Attacks

- **Backdoor (trojaning) attack**
  - The attack engine takes an existing model and a target predication output as the input.
  - Then mutates the model and generates a small piece of input data, called the trojan trigger.
  - Inputs stamped with the trojan trigger will cause the mutated model to generate the given classification output.

Liu, K., et al.,"Trojan attacks on neural networks" (2017)

Poisoning Attacks

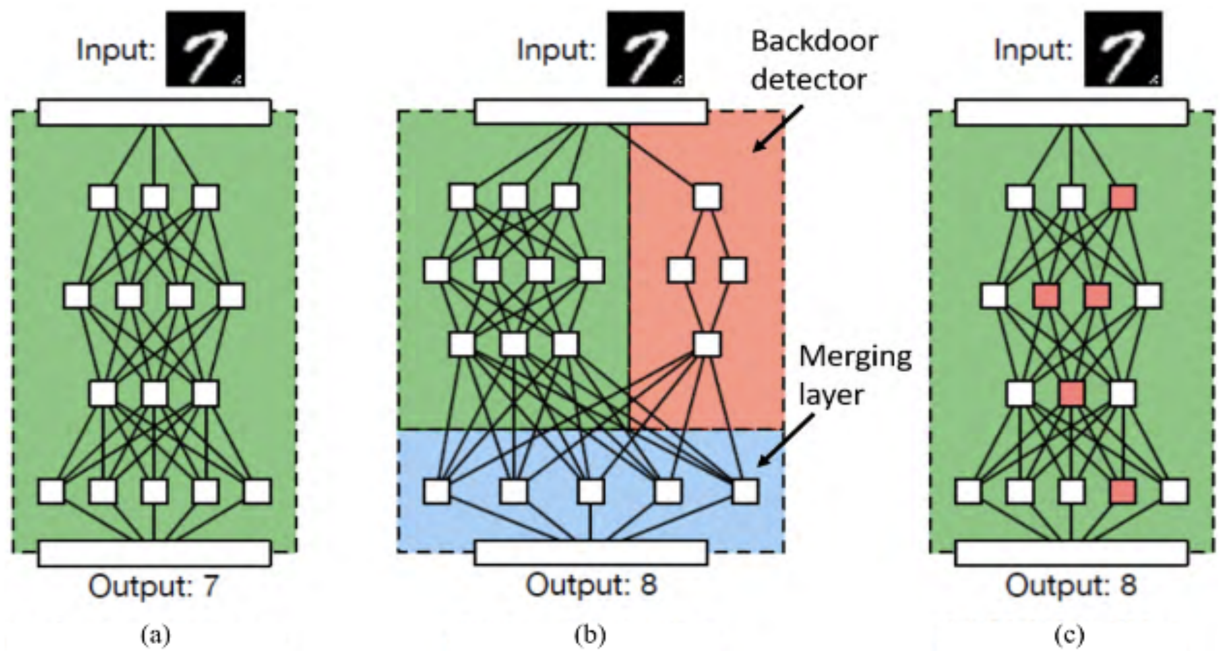
- **Trojan attack overview**



Liu, K., et al.,*"Trojan attacks on neural networks" (2017)*

Poisoning Attacks

- **Backdoor attack**



A benign model is augmented with a backdoor trigger resulting in a poisoned model.

Gu, T., et al.,"BadNets: Evaluating Backdooring Attackson Deep Neural Networks" (2019)

Input attacks

- Perceivable vs imperceptible by humans
- Physical vs Digital noise
- Physical vs Digital attacks
- Crafting adversarial inputs
- GANs

Crafting input attacks

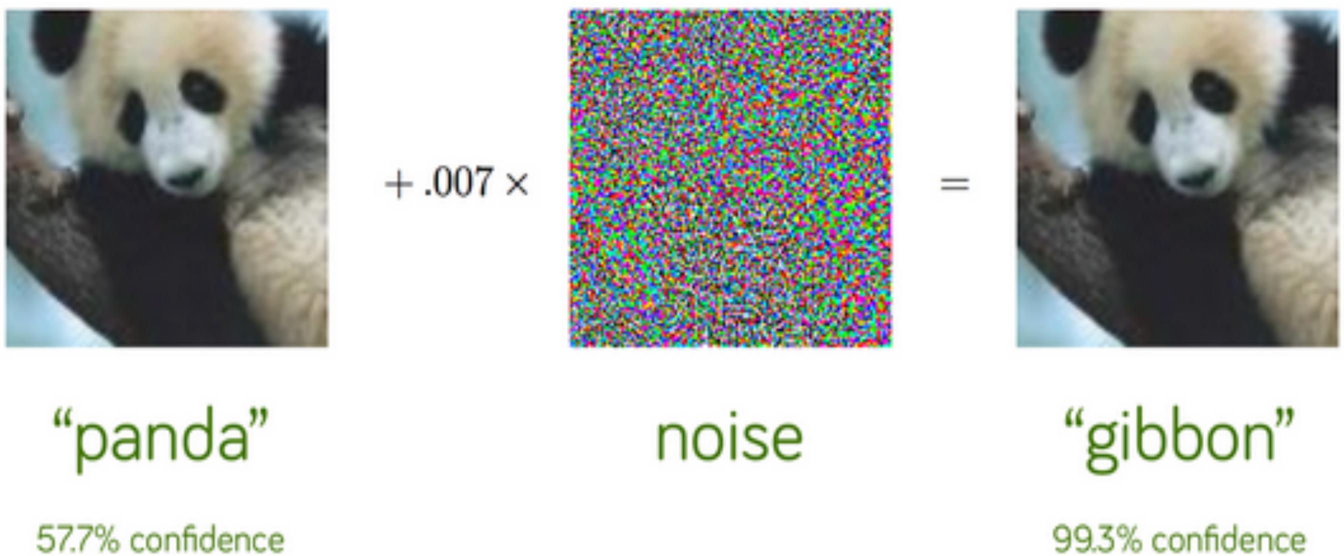
- **Digital noises**
  - Synthetic data
  - Patterns that does/may not exist in real world
  - Noises that are digitally added to digital or physical objects.

“For digital content like images, these ‘imperceivable’ attacks can be executed by sprinkling ‘digital dust’ on top of the target.”

Godfellow I., et al., “Explaining and harnessing adversarial example” (2015)

Crafting input attacks

- Digital noises



Adversarial example generated by adding synthetic data to an inoffensive input.

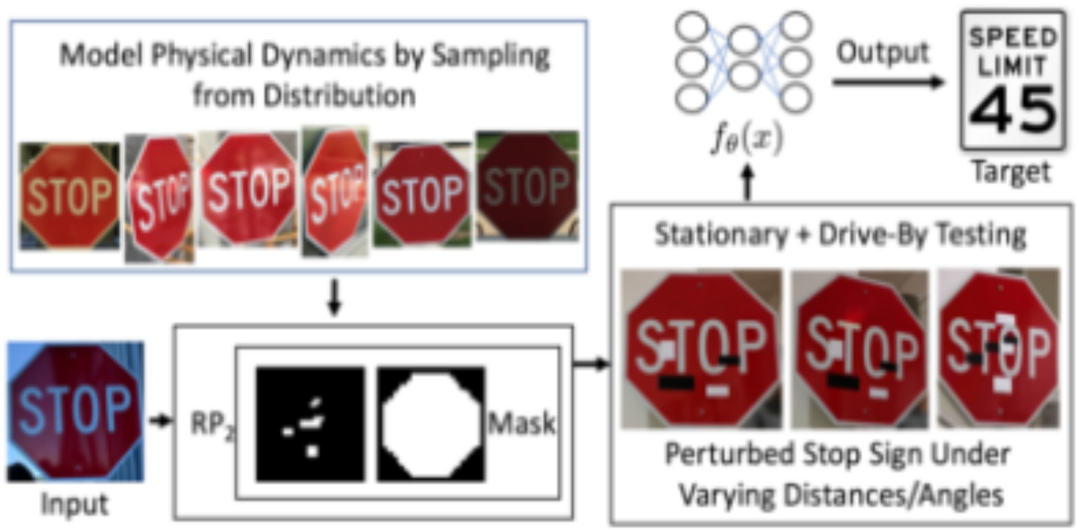
Godfellow I., et al., "Explaining and harnessing adversarial example" (2015)

Crafting input attacks

- **Physical attacks**
  - These are attacks in which the target being attacked exists in the physical world
  - Happens when noise is added to physical objects
  - Stop signs, fire trucks, glasses, humans, sounds
  - Noise is added before the object is captured for classification
- **Digital attacks**
  - Happens when noise is added to digital objects
  - Digital pictures, images, sounds
  - Noise is added after the object is captured for classification

Crafting input attacks

- **Physical attacks**



Adversarial example generated by adding physical objects to inoffensive objects.

Eykholt, K., et al., "Robust Physical-World Attacks on Deep Learning Visual Classification" (2017)

Crafting input attacks

- **Generative Adversarial Networks (GANs)**

Pictures of human faces generated by GANs.

Goodfellow, I., et al. "Generative adversarial nets."

Crafting input attacks

- **What are (GANs)?**
  - Belong to the set of generative models
  - They are able to produce/to generate synthetic data
  - Grossly, GAN models learn the probability distribution of the input samples; and
  - And output new data within this same probability distribution.

Evasion (Adversarial Examples)

- Attack goals
  - Confidence reduction
  - Misclassification
  - Targeted misclassification
  - Source/target misclassification
  - Universal misclassification

Attacker goals

- Confidence reduction

Before the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
Jane (95%)
Sara (99%)
Melissa (91%)
John (83%)

After the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
Jane (65%)
Sara (35%)
Melissa (51%)
John (15%)

Attacker goals

- Misclassification

Before the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
Jane (95%)
Sara (99%)
Melissa (91%)
John (83%)

After the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
John (97%)
Melissa (99%)
Jane (80%)
Sara (83%)

Attacker goals

- Targeted misclassification

Before the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
Jane (95%)
Sara (99%)
Melissa (91%)
John (83%)

After the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
John (97%)
Sara (99%)
John (80%)
John (83%)

Attacker goals

- Source/Targeted misclassification

Before the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
Jane (95%)
Sara (99%)
Melissa (91%)
John (83%)

After the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
John (97%)
Sara (99%)
Melissa (91%)
John (83%)

Attacker goals

- Universal misclassification

Before the attack

Real class
Jane
Sara
Melissa
John

Output (Confidence)
Jane (95%)
Sara (99%)
Melissa (91%)
John (83%)

After the attack

Real class
Jane
Sara
Melissa
John

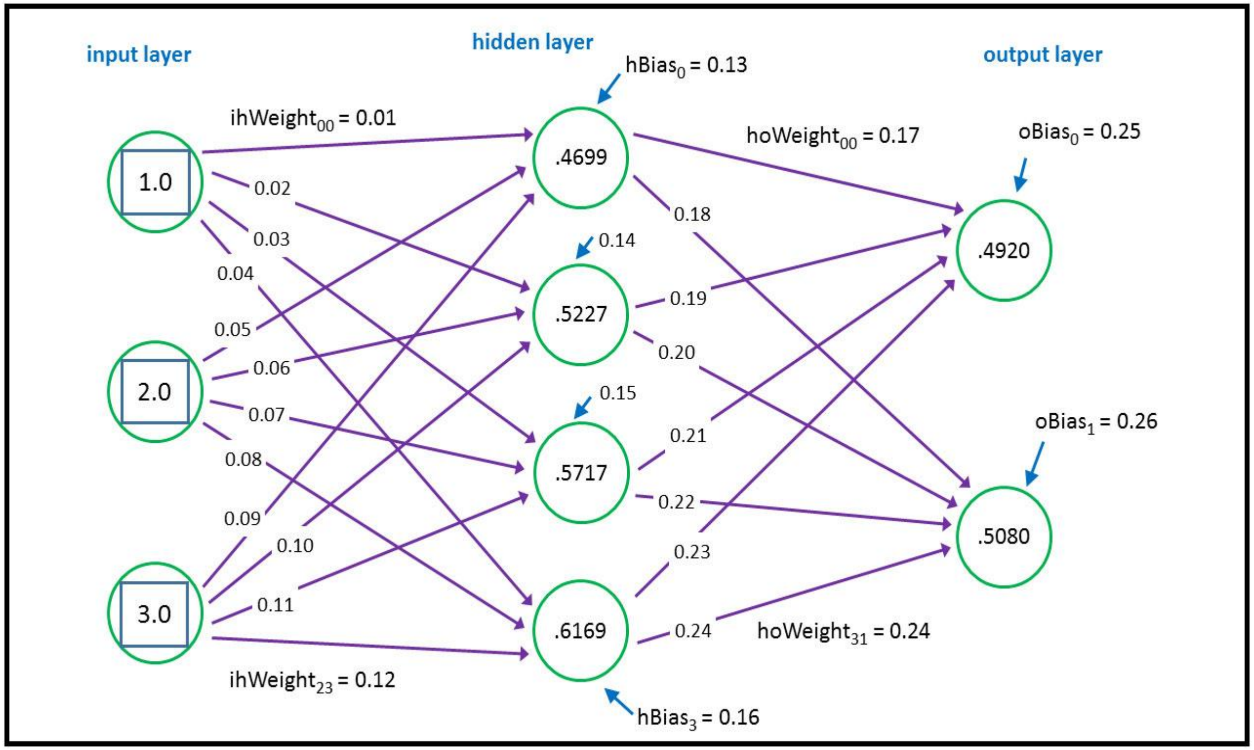
Output (Confidence)
John (87%)
John (92%)
John (99%)
John (83%)

Evasion (Adversarial Examples)

- Attacker knowledge of the models
  - White box
  - Grey box
  - Black box

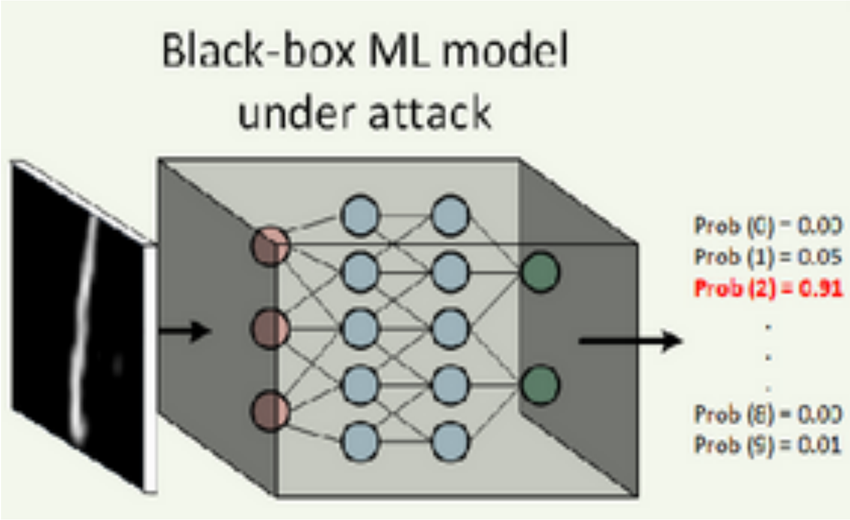
Evasion (Adversarial Examples)

- White box
  - Full knowledge about the network, e.g., weights (parameters) and train data



Evasion (Adversarial Examples)

- **Black box attack**
  - Limited knowledge about the network
  - Attacker can only send information to the system and observe its output



Tu, C., et al., "AutoZOOM : Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks" (2019)

Intended Learning Outcomes

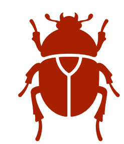
- Define **standard notions of security** and use them to evaluate the **AI system's confidentiality, integrity** and **availability**
- Explain standard **AI security problems** in real-world applications
- Use **testing and verification** techniques to reason about the **AI system's safety** and **security**

Why do we need to ensure AI security?

- AI systems must be as **robust** and **safe** as possible, given that even any faulty behavior can lead to catastrophic outcomes, e.g., endangering human lives, public and private property damage,

Why do we need to ensure AI security?

- AI systems must be as **robust** and **safe** as possible, given that even any faulty behavior can lead to catastrophic outcomes, e.g., endangering human lives, public and private property damage.
  - In 2016, Microsoft released an AI conversational bot that would learn by interacting with Twitter users. In less than 24 hour Tay was corrupted by the users and became a racist, hateful, and sexist entity.
  - In 219, a Uber car hit and killed woman because it did not recognize that pedestrians jaywalk.



Defenses Against Data Poisoning

- Data sanitization (anomaly detection)
- Review and update data policies
- Restrict Data Sharing

Formal Verification

- Verification of properties
- Learning of Invariants
- Model Learning
- Synthesis of Programs and Algorithms

Verification of properties

- Safety Verification of Deep Neural Networks
- Verification of Markov Decision Processes Using Learning Algorithms
- Formal Verification of Neural Networks
- Counterexample Explanation for Probabilistic Systems

Learning of Invariants

- Learning Software Invariants
- Learning Data Structure Invariants
- Syntax-Guided Invariant Synthesis
- Synthesizing Inductive Invariants

Model Learning

- Learning Finite Automata
- Learning and Planning with Timing Information in Markov Decision Processes
- Generating Models of Communication Protocols

Synthesis of Programs and Algorithms

- Policy Learning in Continuous-Time Markov Decision Processes
- Safety-Constrained Reinforcement Learning for Markov Decision Processes
- Learning Static Analyzers
- Learning Explanatory Rules from Noisy Data
- Multi-Objective Policy Generation for Mobile Robots

Summary

- Security for AI systems and AI-Security Domains
- Technical AI safety topics
- AI system limitations
- Attacker goals
- Risks in machine learning pipeline
- Types of attack

Summary

- Test and verification
- Defenses Against Data Poisoning
- Formal Verification
- Verification of properties
- Learning of Invariants
- Model Learning
- Synthesis of Programs and Algorithms

References

- Chan-Hon-Tong, A., An Algorithm for Generating Invisible Data Poisoning Using Adversarial Noise That Breaks Image Classification Deep Learning, 2019
- Newman, J., Toward AI Security, 2019.
- Sitawarin, C. et al., DARTS: Deceiving Autonomous Cars with Toxic Signs, 2018
- Pedro Ortega and Vishal Maini, Building safe artificial intelligence: specification, robustness, and assurance, DeepMind, 2018.
- Finlayson, S.G., et al., “Adversarial Attacks Against Medical Deep Learning Systems” (2019)
- Weis, Steve, Security & Privacy Risks of Machine Learning Models, 2019

References

- Wang, B., et al.,“Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks” (2019)
- Liu, K., et al.,“Trojan attacks on neural networks” (2017)
- Gu, T., et al.,“BadNets: Evaluating Backdooring Attackson Deep Neural Networks” (2019)
- Godfellow I., et al., “Explaining and harnessing adversarial example” (2015)
- Eykholt, K., et al., “Robust Physical-World Attacks on Deep Learning Visual Classification” (2017)
- Goodfellow, I., et al. “Generative adversarial nets.”
- Tu, C., et al., “AutoZOOM : Autoencoder-based Zeroth Order Optimization Method for Attacking Black-box Neural Networks” (2019)