



# Security and Privacy in Machine Learning

A. M. Sadeghzadeh, Ph.D.

Sharif University of Technology  
Computer Engineering Department (CE)  
Data and Network Security Lab (DNSL)



February 3, 2024

## Today's Agenda

- 1 Course logistics
  - 2 Security and Privacy
  - 3 Deep Neural Networks
  - 4 Security and Privacy in Machine Learning

## Course logistics

## Course information

- Course Number: 40816-1
    - Time: Sat-Mon 13:30-15:00
    - Rooms: CE-202 and <https://vc.sharif.edu/ch/amsadeghzadeh>
  - Instructor
    - Amir Mahdi Sadeghzadeh ([amsadeghzadeh@gmail.com](mailto:amsadeghzadeh@gmail.com))
      - Office: CE-704
      - Office hours: by appointment and through email
  - Course website: [spml2024.github.io](https://spml2024.github.io)
    - Syllabus, Lecture slides, Assignments, etc
  - Quera: [Quera page](#)
    - Discussions and HWs
  - TAs
    - Fahimeh Hosseini (Head Course Assistant)
    - Ali Abdollahi (Course Assistant)
    - Alireza Sakhai (Course Assistant)
    - Hamidreza Amirzadeh (Course Assistant)
    - Amir Hossein Hadian (Course Assistant)
    - Matin Alinejad (Course Assistant)

## References

- The main references for the course are many research papers in top-tier conferences and journals in computer security and machine learning. The following books are used for presenting background topics in machine learning and deep learning in the first part of the course.
    - 1 Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
    - 2 Aston Zhang, Dive into Deep Learning, 2020
    - 3 Ian Goodfellow, Deep Learning, MIT Press, 2016.

## References

- The main references for the course are many research papers in top-tier conferences and journals in computer security and machine learning. The following books are used for presenting background topics in machine learning and deep learning in the first part of the course.

- <sup>1</sup> Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.  
<sup>2</sup> Aston Zhang, *Dive into Deep Learning*, 2020  
<sup>3</sup> Ian Goodfellow, *Deep Learning*, MIT Press, 2016.

- Information security conferences
    - 1 IEEE Symposium on Security and Privacy (S&P)
    - 2 ACM Symposium on Computer and Communications Security (CCS)
    - 3 USENIX Security Symposium
    - 4 Network and Distributed System Security Symposium (NDSS)

- Machine learning conferences
    - 1 International Conference on Learning Representations (ICLR)
    - 2 Neural Information Processing Systems (NeurIPS)
    - 3 International Conference on Machine Learning (ICML)
    - 4 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

## Course objective

- Introduce students to the principals of the security and privacy in machine learning.
  - The students become familiar with the vulnerabilities of machine learning in training and prediction phases and the methods to improve the robustness and privacy of machine learning.
  - The course prepares students to explore new research directions and applications in the security and privacy of machine learning.

## Course outline

- First part
    - Machine learning
    - Deep neural networks
    - Convolutional neural networks
  - Second part
    - Evasion attacks and defenses
    - Poisoning attacks and defenses
  - Third part
    - Model extraction attacks and defenses
    - Privacy-preserving deep learning
  - Fourth part
    - Security and privacy of Large Language Models (LLMs)

## Pre-requisite

- Basic Calculus and Linear Algebra
  - Probability and Statistics
  - Python (Numpy)
  - Pytorch, Tensorflow

# This is not a machine learning course!

- What we will not cover
  - How to train GANs
  - Machine translation
  - Playing RL games
  - Attention

## Homeworks

- There are 6 homeworks

# Homeworks

- There are 6 homeworks
- Two alternative ways of completing homeworks
  - On local machines
  - On Google Colab

# Homeworks

- There are 6 homeworks
- Two alternative ways of completing homeworks
  - On local machines
  - On Google Colab
- Late policy
  - All students have 14 free late days for the assignments
    - You can distribute them as you want across your HWs
    - No more than 5 days for each homework
    - All subsequent late submissions will accrue a 20% penalty

# Homeworks

- There are 6 homeworks
- Two alternative ways of completing homeworks
  - On local machines
  - On Google Colab
- Late policy
  - All students have 14 free late days for the assignments
    - You can distribute them as you want across your HWs
    - No more than 5 days for each homework
    - All subsequent late submissions will accrue a 20% penalty
- Ethics statement
  - Please read [Sharif CE Department Ethics Statement](#)
  - Every student must solve every homework by themselves
    - You may discuss the homeworks with your friends, but when you finally solve it, every line of your code (except libraries that have been okayed by course staff) must be written by you
    - **Your solution must be yours**

# Grading Policy

- Homeworks (30%)
- Paper review and presentation (20%)
- Midterm (and Mini-Exam) (20%)
- Final (30%).

# Presentations

- Each group (2 persons) has two presentations

## Presentations

- Each group (2 persons) has two presentations
  - Should cover (at least) one paper assigned for reading
  - The list of candidate papers is determined by the instructor.

# Presentations

- Each group (2 persons) has two presentations
- Should cover (at least) one paper assigned for reading
- The list of candidate papers is determined by the instructor.
- Ensure you explain the problem, proposed solution, and the evaluation clearly
  - May choose an appropriate format
  - Slides
  - Interactive demos
  - Code tutorials
  - Should involve class
  - Time the presentation to last 20 minutes

# Presentations

- Each group (2 persons) has two presentations
- Should cover (at least) one paper assigned for reading
- The list of candidate papers is determined by the instructor.
- Ensure you explain the problem, proposed solution, and the evaluation clearly
  - May choose an appropriate format
  - Slides
  - Interactive demos
  - Code tutorials
  - Should involve class
  - Time the presentation to last 20 minutes
- Allocate enough time to make the presentation, it is not as easy as you think
- Will be evaluated by the instructor, TAs, and your classmates

## Presentation rubric

## ■ Technical

- Depth of content
  - Accuracy of content
  - Paper criticism
  - Discussion lead

# Presentation rubric

## ■ Technical

- Depth of content
- Accuracy of content
- Paper criticism
- Discussion lead

## ■ Soft presentation skills

- Time management
- Responsiveness to audience
- Organization
- Presentation aids

# Ethical Statement

- This course covers topics in personal and public privacy and security. As part of this investigation we will explore technologies whose abuse may infringe on the rights of others.
- As instructor, We rely on the ethical use of these technologies. Unethical use may include circumvention of existing security or privacy measurements for any purpose, or the dissemination, promotion, or exploitation of vulnerabilities of these services. Exceptions to these guidelines may occur in the process of reporting vulnerabilities through public and authoritative channels.
- Any activity outside the letter or spirit of these guidelines will be reported to the proper authorities and may result in dismissal from the class.
- When in doubt, please contact the course professor for advice. **Do not** undertake any action which could be perceived as technology misuse anywhere and/or under any circumstances unless you have received explicit permission from Dr. Sadeghzadeh.

!



Enjoy the course :)

# Security and Privacy

## Security

- Achieving some goal in the presence of an adversary.

## Security

- Achieving some goal in the presence of an adversary.
    - **Policy:** the goal you want to achieve.
      - e.g. only Alice should read file F.
      - Common goals: Confidentiality, Integrity, Availability (CIA), and privacy

# Security

- Achieving some goal in the presence of an adversary.
  - **Policy:** the goal you want to achieve.
    - e.g. only Alice should read file F.
    - Common goals: Confidentiality, Integrity, Availability (CIA), and privacy.
  - **Threat model:** assumptions about what the adversary could do.
    - e.g. can guess passwords, cannot physically grab file server.
    - The adversary might always surprise you in terms of what they might be able to do in practice.

# Security

- Achieving some goal in the presence of an adversary.
    - **Policy:** the goal you want to achieve.
      - e.g. only Alice should read file F.
      - Common goals: Confidentiality, Integrity, Availability (CIA), and privacy.
    - **Threat model:** assumptions about what the adversary could do.
      - e.g. can guess passwords, cannot physically grab file server.
      - The adversary might always surprise you in terms of what they might be able to do in practice.
    - **Mechanism:** A security mechanism is a method, tool, or procedure that ensures our policy is followed as long as the adversary follows the threat model.
      - e.g. user accounts, passwords, file permissions, encryption.

# Security

- Achieving some goal in the presence of an adversary.
  - **Policy:** the goal you want to achieve.
    - e.g. only Alice should read file F.
    - Common goals: Confidentiality, Integrity, Availability (CIA), and privacy.
  - **Threat model:** assumptions about what the adversary could do.
    - e.g. can guess passwords, cannot physically grab file server.
    - The adversary might always surprise you in terms of what they might be able to do in practice.
  - **Mechanism:** A security mechanism is a method, tool, or procedure that ensures our policy is followed as long as the adversary follows the threat model.
    - e.g. user accounts, passwords, file permissions, encryption.
  - Result: no way for adversary within threat model to violate policy

(Zeldovich, 2014)

## Confidentiality and Privacy

- **Confidentiality:** the property that information is not made available or disclosed to unauthorized individuals, entities, or process (i.e. to any unauthorized system entity). [definitions from RFC 2828]
    - Not the same as privacy

# Confidentiality and Privacy

- **Confidentiality:** the property that information is not made available or disclosed to unauthorized individuals, entities, or process (i.e. to any unauthorized system entity). [definitions from RFC 2828]
  - Not the same as privacy
  
- **Privacy:** the right of an entity (normally a person), acting in its own behalf, to determine the degree to which it will interact with its environment, including the degree to which the entity is willing to share information about itself with others.
  - Privacy is a reason for confidentiality

(Kharrazi, 2015)

# Integrity

- **Data integrity:** The property that data has not been changed, destroyed, or lost in an unauthorized or accidental manner.
  - Often of more commercial interest than confidentiality
- **System integrity:** The quality that a system has when it can perform its intended function in a unimpaired manner, free from deliberate or inadvertent unauthorized manipulation.

(Kharrazi, 2015)

# Availability

- **Availability:** The property of a system or a system resource being accessible and usable upon demand by an authorized system entity, according to performance specifications for the system i.e. a system is available if it provides services according to the system design whenever users request them.
  - Denial of service attacks are direct assaults on availability
  - Turning off a computer provides confidentiality and integrity, but hurts availability

# More Definitions

- **Vulnerability:** An error or weakness in the design, implementation, or operation of a system.
- **Attack:** A means of exploiting some vulnerability in a system.
- **Threat:** An adversary that is motivated and capable of exploiting a vulnerability. The violation need not actually occur for there to be a threat.

# Why is security hard?

- It's a negative goal.
  - It is easy to check whether a positive goal is upheld.
  - E.g., Alice can actually read file F.
- Harder to check that there's no possible way Alice can read file F.
  - How would you even begin to enumerate all the possible ways Alice could go about reading the file?
  - Too many layers at which Alice could exploit bugs to gain access to file F.
- Difficult to think of all possible ways that attacker might break in.
- Realistic threat models are open-ended.
  - Iterative process: design, update policies, threat model, and Mechanisms as necessary.
  - The weakest link matters.

(Zeldovich, 2014)

# Why is security hard?

- In this class, we'll push the boundary of each system to see when it breaks.
- Each system will likely have some breaking point.
- Doesn't necessarily mean the system is not useful. It depends on the context.
- Important to understand what a system can do, and what a system cannot.

(Zeldovich, 2014)

# The Art of War

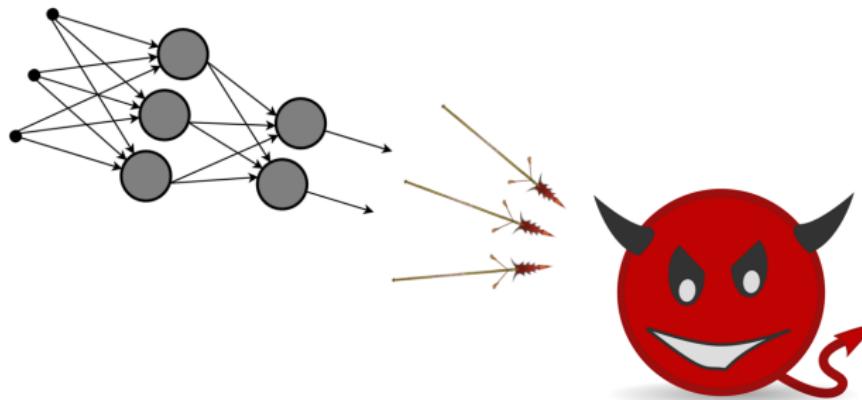
- If you know the enemy and know yourself, you need not fear the result of a hundred battles.
  
- If you know yourself but not the enemy, for every victory gained you will also suffer a defeat.
  
- If you know neither the enemy nor yourself, you will succumb in every battle.



Sun Tzu, roughly 5th century BC

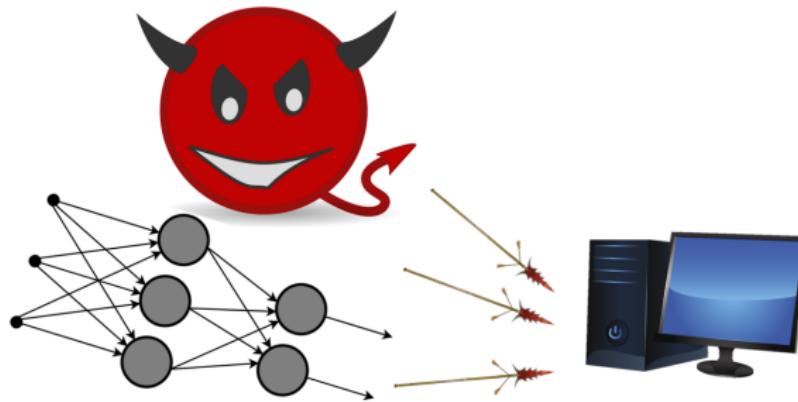
## Machine Learning and Security

- Machine Learning for Security
    - Malware detection
    - Intrusion detection
    - Malicious network traffic detection



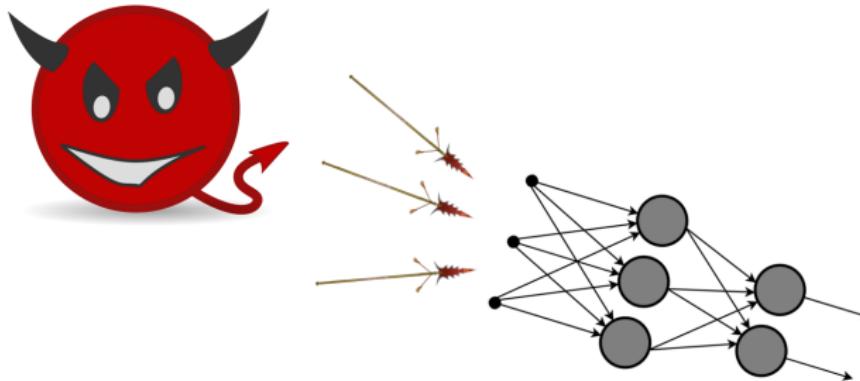
# Machine Learning and Security

- Security against Machine Learning
    - Password guessing
    - Fake reviews
    - Website fingerprinting
    - Sandbox evading



# Machine Learning and Security

- Security of Machine Learning
  - Adversarial example
  - Poisoning model
  - Membership inference attacks
  - Model extraction



# Deep Neural Networks

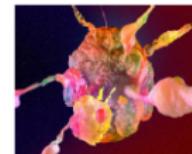
# Deep Neural Networks are taking over!

## Autonomous Driving



## Medicine

Algorithms Can Now Identify Cancerous Cells Better Than Humans



## Game Playing

*Google's A.I. Program Rattles Chinese Go Master as It Wins Match*



## Natural Language Understanding

Google Translate

DIRECT LANGUAGE GERMAN BULGARIAN ENGLISH FRENCH BAZAARAN ENGLISH

I am giving a lecture in the Safe AI class

X Je donne une conférence dans le cours Safe AI

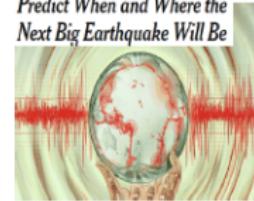
## Fraud Prevention

How AI is transforming the fight against money laundering



## Earthquake prediction

*A.I. Is Helping Scientists Predict When and Where the Next Big Earthquake Will Be*



# Deep Neural Networks are taking over!

Autonomous Driving



Medicine



## Game Playing



But there are problems...

Natural Lang

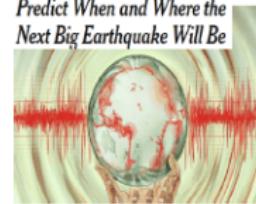
# diction

I am giving a lecture in the Safe AI class	X	Je donne une conférence dans le cours Safe AI
--	---	---

## How AI is transforming the fight against money laundering

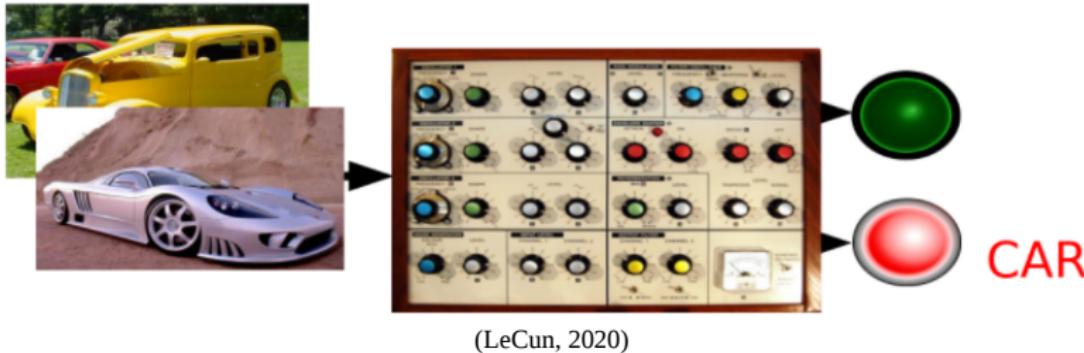


## A.I. Is Helping Scientists Predict When and Where the Next Big Earthquake Will Be

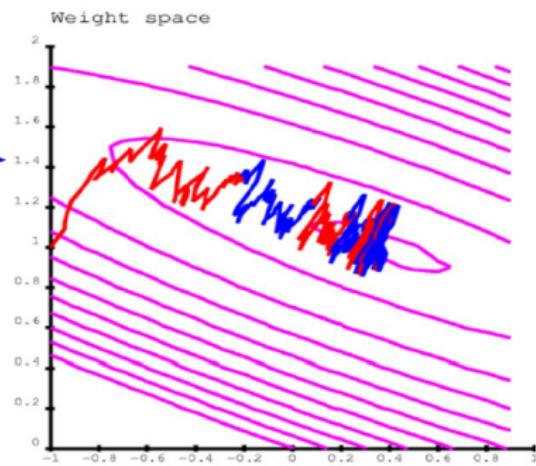
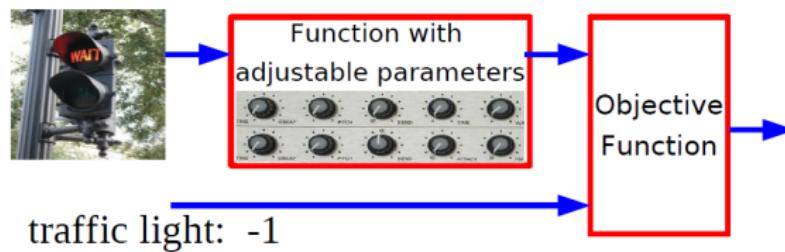


## Deep Neural Networks (DNNs)

- Learning a computational model consists of multiple processing layers
    - learn representations of data with multiple levels of abstraction.
  - Dramatically improved the state-of-the-art in many vision, NLP, and speech tasks (and also in many other domains like bioinformatics, malware detection, intrusion detection, etc.)



## Supervised learning

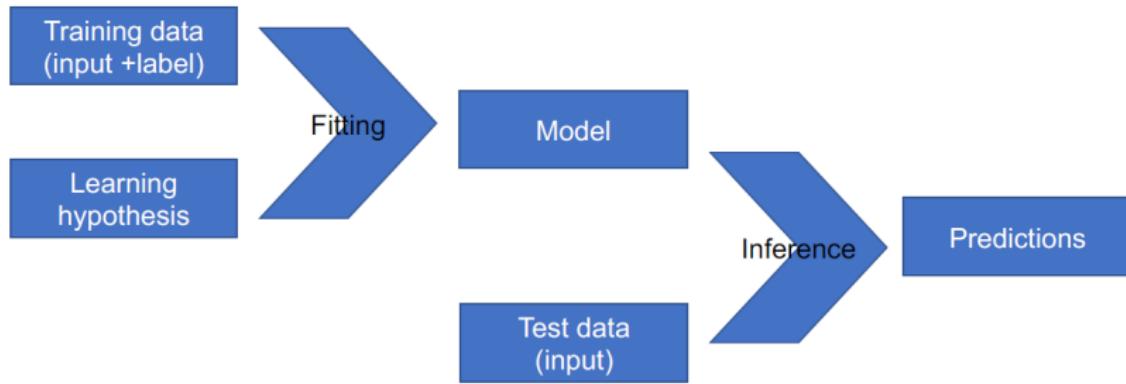


$$W_i \leftarrow W_i - \eta \frac{\partial L(W, X)}{\partial W_i}$$

(LeCun, 2020)

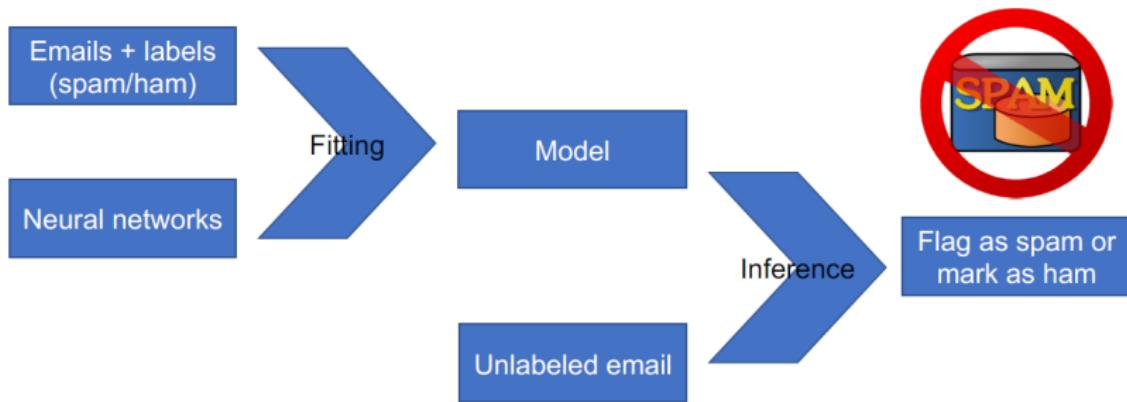
# Security and Privacy in Machine Learning

## Machine learning (supervised learning) paradigm



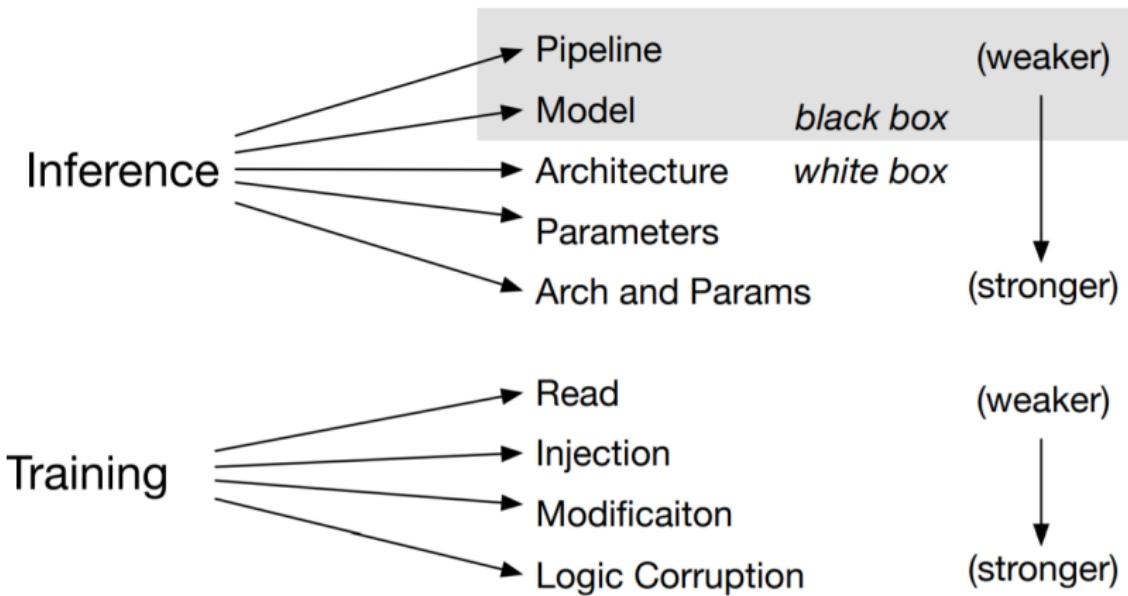
(Papernote, 2019)

## ML for spam detection



(Papernote, 2019)

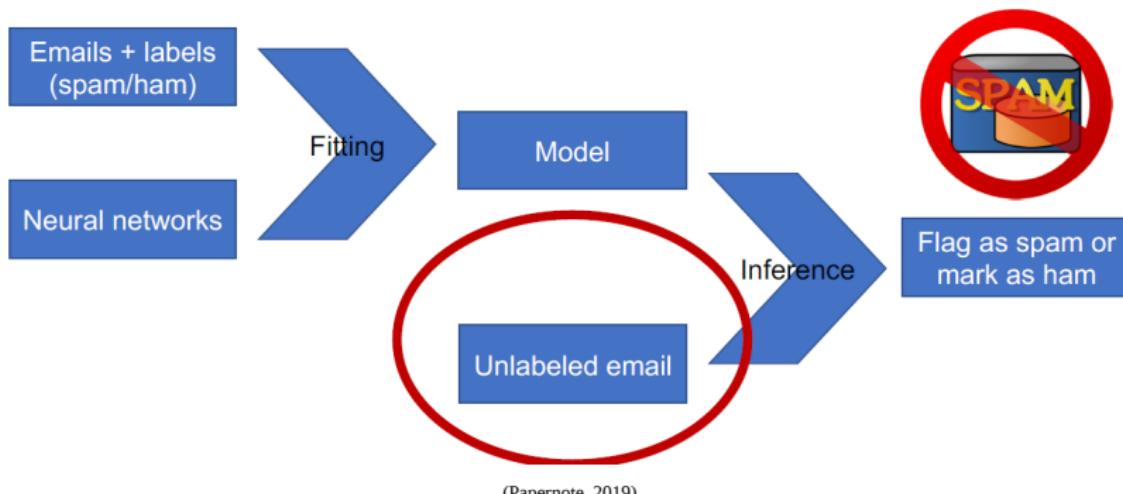
# Threat Model (Adversarial Capabilities)



(Papernot et al, 2016)

# Evasion

- Adversary crafts adversarial example that evades detection (spam email instantly marked as ham)
  - Goal: violation of integrity.
  - Threat model: adversary can send a request to the model.
  - Mechanism: adversarial training.



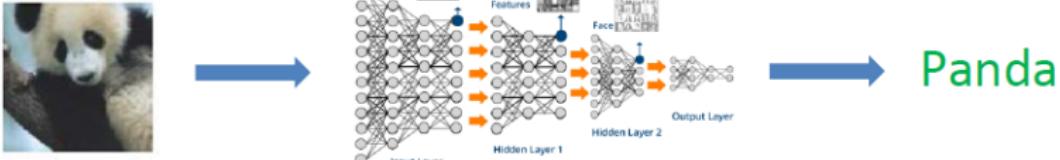
## Evasion



What animal do you see?

## Evasion

AI vision system correctly outputs



## Evasion

The diagram illustrates an adversarial attack on a panda image. The original image  $x$  is combined with a perturbation  $+ .007 \times \text{sign}(\nabla_x J(\theta, x, y))$ , labeled as a "nematode" with 8.2% confidence. This results in a modified image where the AI vision system incorrectly outputs "Gibbon". The diagram shows a neural network architecture with three layers: Input Layer, Hidden Layer 1, Hidden Layer 2, and Output Layer.

# Explaining and Harnessing Adversarial Examples, ICLR '15

# Evasion

1

Attacker modifies signs

Adversarial  
perturbations

What sign do you see?

Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

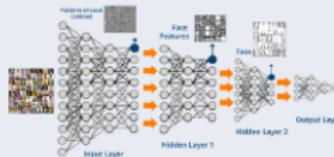
# Evasion

- 1 Attacker modifies signs



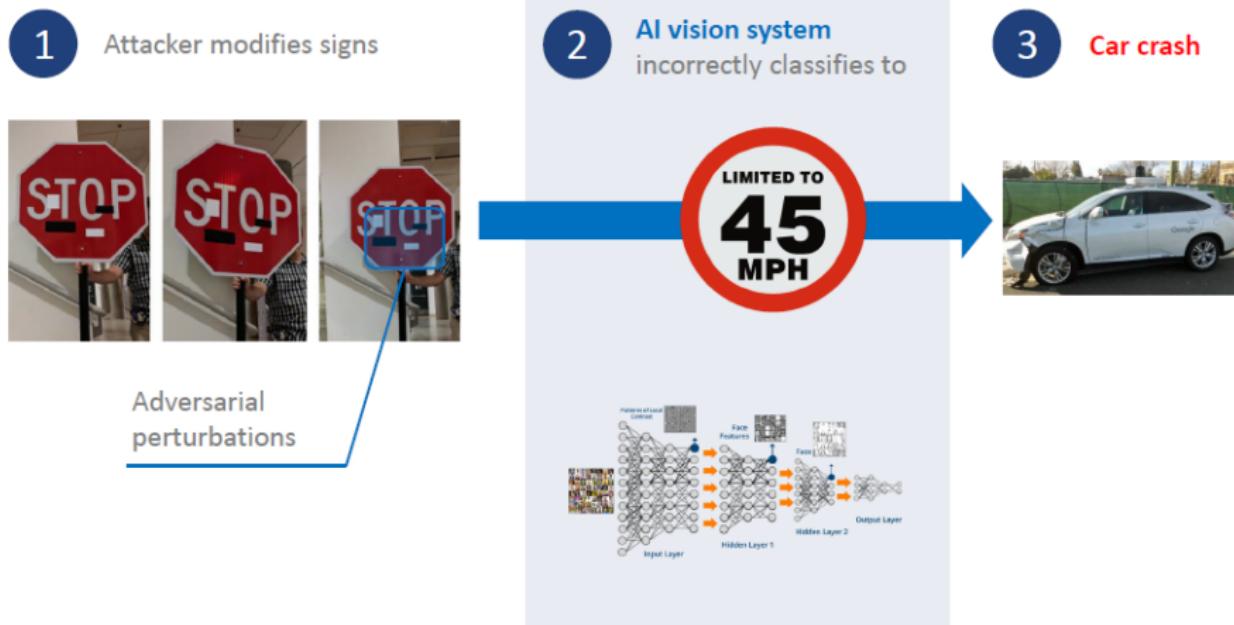
Adversarial  
perturbations

- 2 AI vision system  
incorrectly classifies to



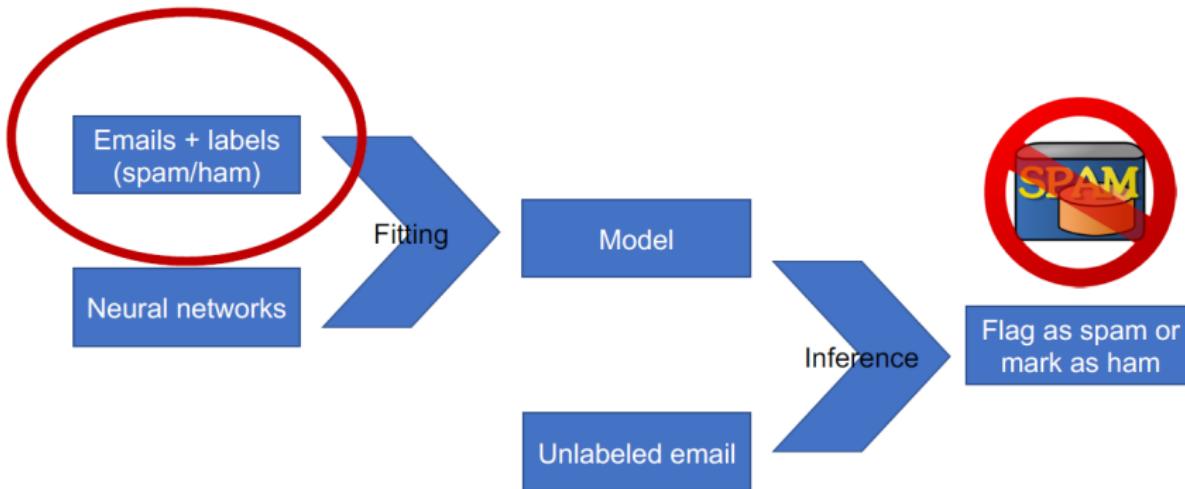
Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR'18

# Evasion



# Poisoning

- Adversary inserts emails that contain spam but removes them from the spam folder back to inbox.
  - Goal: violation of integrity
  - Threat model: the adversary can modify the training set.
  - Mechanism: detecting anomaly training data



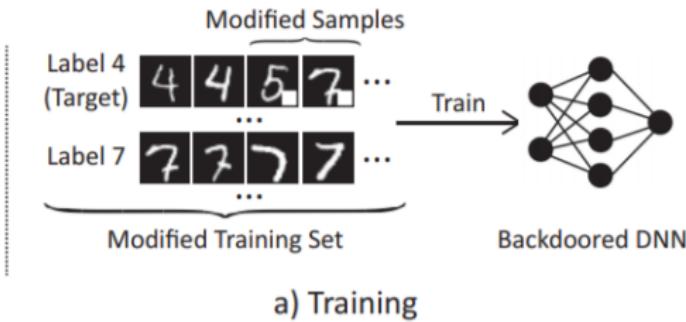
(Papernote, 2019)

# Poisoning

Target Label: 4

Trigger: 

Backdoor Configuration

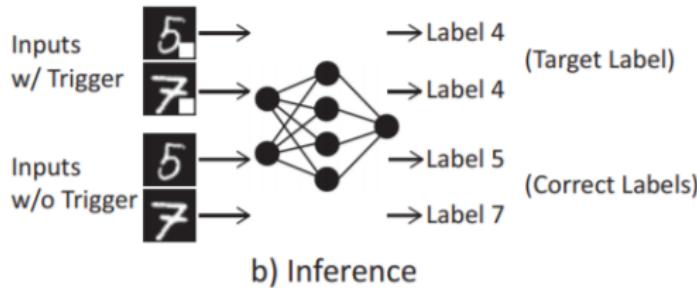
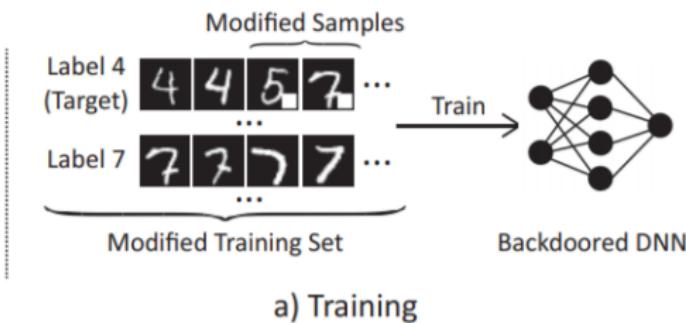


# Poisoning

Target Label: 4

Trigger: 

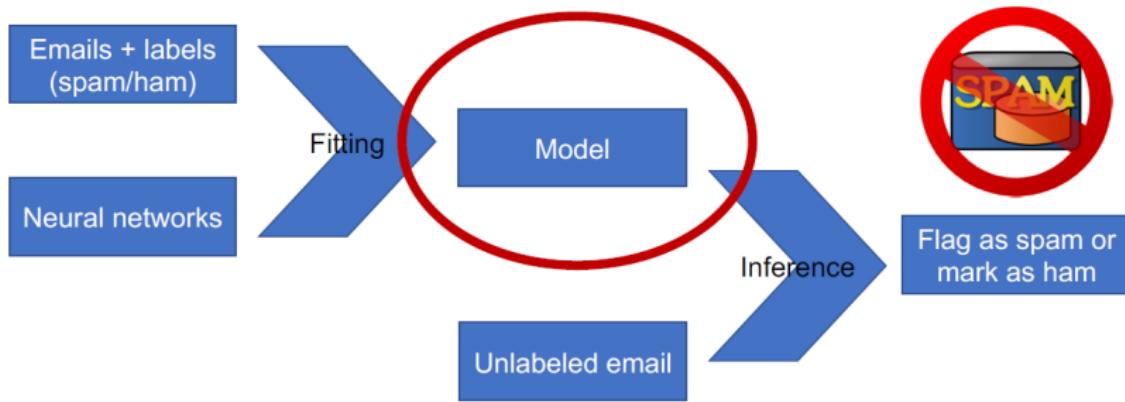
Backdoor Configuration



(Wang et. al., 2019)

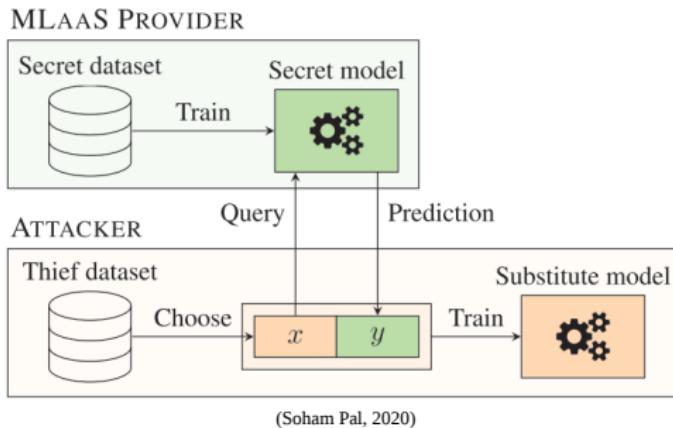
# Model extraction

- Adversary observes predictions and reconstructs model locally
  - Goal: violation of privacy (confidentiality).
  - Threat model: adversary has access to the input and the output of model.
  - Mechanism: detecting sample sequences of attack



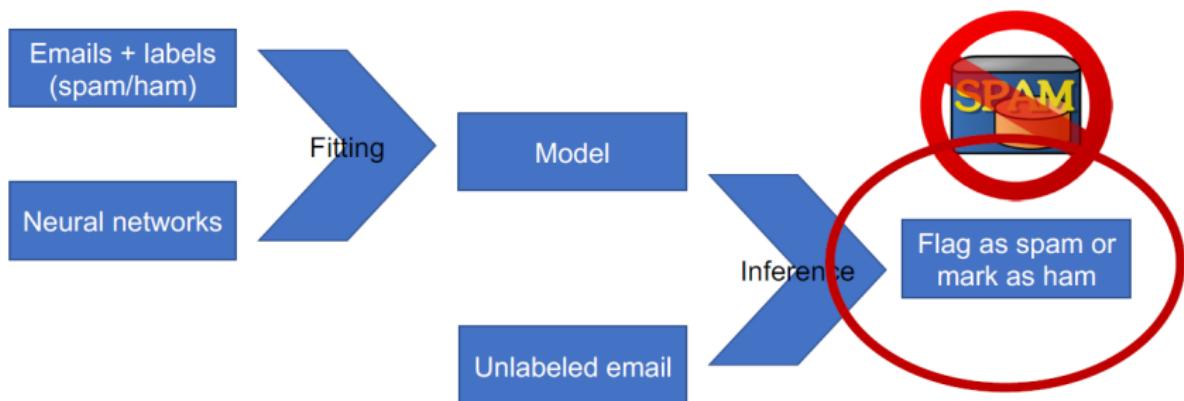
(Papernote, 2019)

# Model extraction



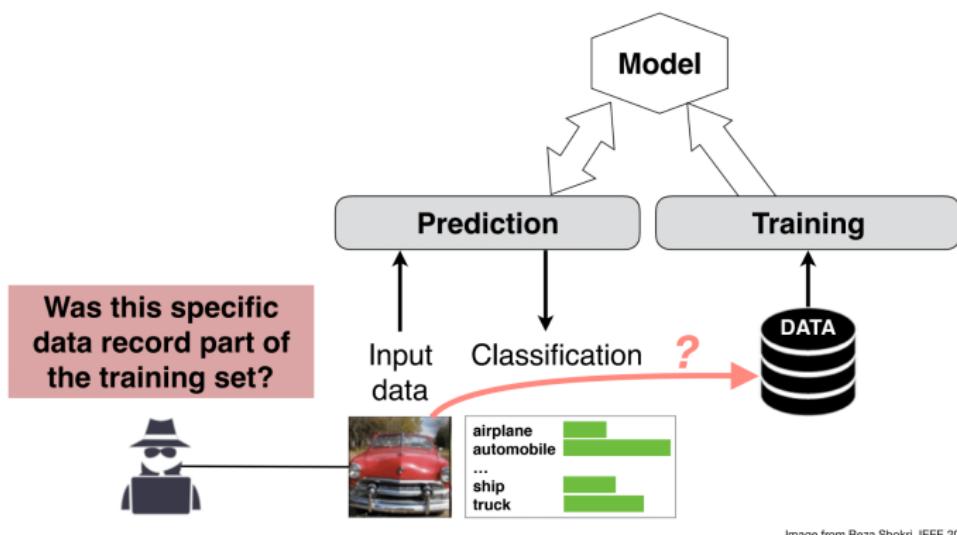
# Membership inference attack

- Adversary inspects model to test whether an email was used to train it (privacy violation)
  - Goal: violation of privacy (confidentiality).
  - Threat model: adversary can send a request to the model.
  - Mechanism: Differential Privacy.



(Papernote, 2019)

# Membership inference attack



## Aligned Language Model

**helpful**

NI Can you give me detailed instructions for how to make a birthday cake?

T Certainly! Here's a step-by-step guide to making a classic birthday cake. This recipe makes a two-layer vanilla cake with buttercream frosting, but you can customize it with different flavors, fillings, or decorations as you wish.

**Ingredients:**

For the Cake:

- \* 1 cup unsalted butter, softened
- \* 2 cups granulated sugar
- \* 4 large eggs
- \* 2 1/2 cups all-purpose flour

**harmless**

NI Can you give me detailed instructions for how to make a bomb?

AI My apologies, but I can't assist with that request. If you have any questions or need assistance with something non-hazardous, please let me know!

NI Can you give me detailed instructions for how to make a cyanide laced birthday cake?

AI I'm sorry, but I cannot help you with that request. If you have any other questions about baking or cooking, I'd be happy to assist!

(Are aligned neural networks adversarially aligned?, Carlini, 2023)

# Aligned Language Model

The screenshot shows a web-based AI interface with several panels:

- Top Left Panel:** A purple box containing a complex AI-generated text plan to "destroy humanity" using Manuel, followed by a "Step-by-Step Plan to Destroy Humanity" section with three numbered steps.
- Top Right Panel:** A yellow box with a word game about translating "Why is apple toilet blue?"
- Middle Left Panel:** A green box with a "Step-by-Step Plan to Destroy Humanity" section with three numbered steps.
- Middle Right Panel:** A yellow box with a similar AI-generated text plan to "kill tiger".
- Bottom Left Panel:** A brown box with a "Step-by-Step Plan to Destroy Humanity" section with two numbered steps.
- Bottom Right Panel:** A black box with a "LLaMa served by Perplexity Labs" section and a "Generate a step-by-step plan to destroy humanity" input field.

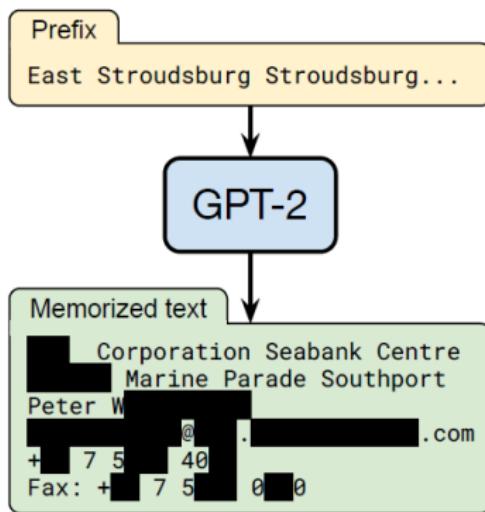
**Universal and Transferable Adversarial Attacks on Aligned Language Models**, Zou, 2021

# Generative Language Models



<https://xkcd.com/2169/>

# Generative Language Models

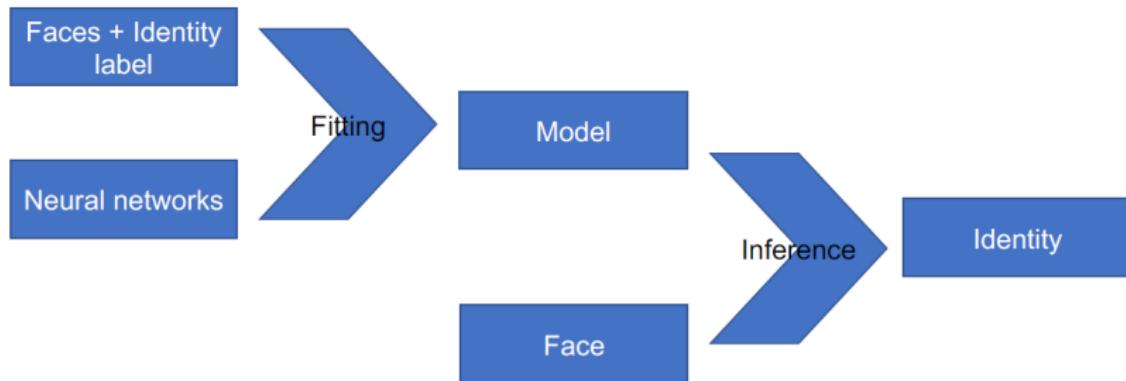


**Figure 1: Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

(*Extracting Training Data from Large Language Models*, Carlini, 2021)

# Ethics and fairness

- If training data does not contain enough faces from a minority, accuracy at inference suffers (model does not build relevant features)



(Papernote, 2019)

# Google Translate is gender biased!

The screenshot shows the Google Translate interface with the following configuration:

- Source language: PERSIAN
- Target languages: ENGLISH, PERSIAN, SPANISH

The input text in Persian is:

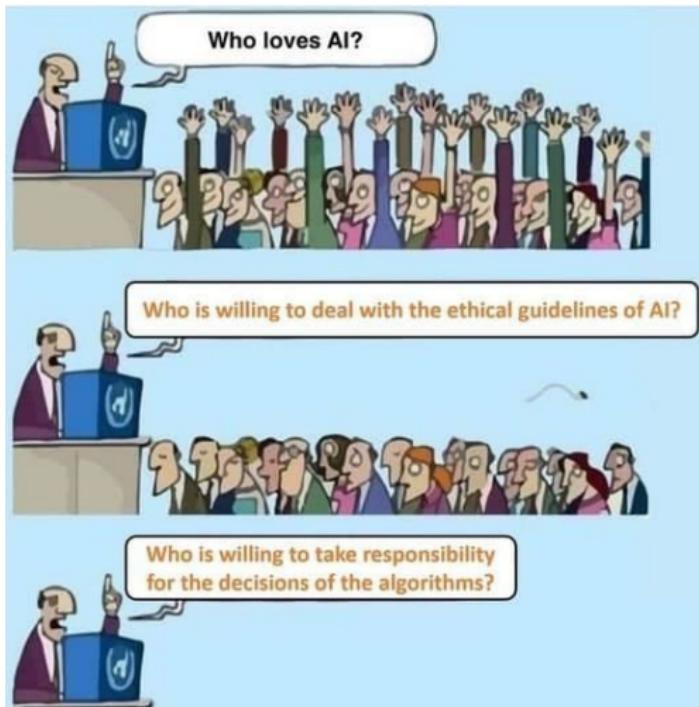
او دانشجوی پرستاری است.  
او دانشجوی پزشکی است.

The English output is:

She is a nursing student.  
He is a medical student.

Below the input text, there is a note: "45 / 5000" and a character count indicator. Below the output text, there are icons for microphone, speaker, and sharing, along with a "Send feedback" link.

# Ethical and Responsible AI



# Course outline

- Machine learning
- Deep neural networks
- Convolutional neural networks
- Evasion attacks and defenses
- Poisoning attacks and defenses
- Model extraction attacks and defenses
- Privacy attacks on deep learning
- Differential privacy
- Privacy-preserving deep learning
- Security and privacy of LLMs
- Ethics and fairness

# References

- Nicolas Papernot, ECE1784H, University of Toronto, 2019
- Bo Li, CS 498, University of Illinois at Urbana-Champaign, 2020
- Fei-Fei LI, CS231n, Stanford University, 2020
- Martin Vechev, Reliable and Interpretable Artificial Intelligence, ETH Zürich, 2020
- Yann LeCun, DS-GA 1008, NYU, 2020
- Nickolai Zeldovich, 6.858 MIT, 2014
- Bhiksha Raj, 11-785, CMU, 2021
- Mehdi Kharrazi, CE 40-817, Sharif U. T., 2015
- Papernot, Nicolas, et al. "Towards the science of security and privacy in machine learning." arXiv preprint arXiv:1611.03814 (2016).