IBM AI Ethics(mod3)

Module 3

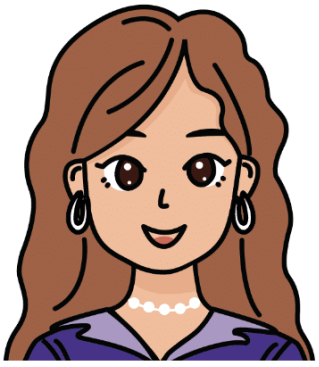
What is robustness?

Meet the team

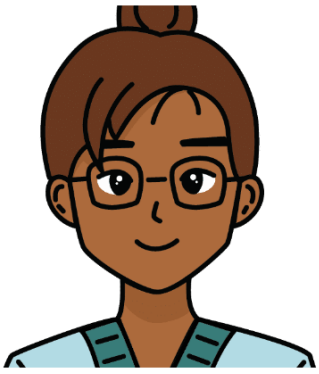
Next, you'll look at how AI systems can be protected from attacks.
Adversarial **robustness** refers to an AI model's ability to resist being fooled. Teams are constantly working to make AI systems more impervious to irregularities and attacks. This story looks at what a medical diagnostic company can do to safeguard AI systems against attacks.

A medical diagnostic company is getting close to launching an application to detect lung cancer. Then they hear that another healthcare app start-up is facing a lawsuit for misdiagnosed cases. The company's technical leadership team pulls everyone together to learn what an **adversarial attack** is and to assess the risk of their system being attacked. In this story, you'll learn what an adversarial attack is, the different types of attacks, how they can happen, and what teams need to think about to protect their AI systems.

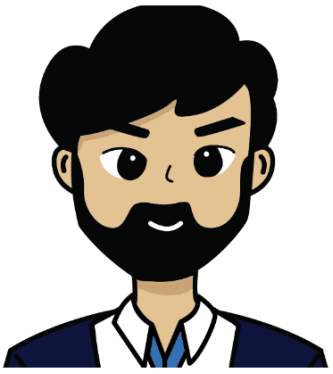
The team



Charlie (she/her)
Chief Technology Officer (CTO)



Manuel (they/them)
Chief Privacy Officer (CPO)



Prashant (he/his)
Chief Data Officer (CDO)

Identify the issue

The launch date for an AI-based application to aid in the detection of lung cancer using chest x-rays is quickly approaching. Charlie, the Chief Technology Officer, is reviewing the plan with her team.

Manuel, the Chief Privacy Officer, knocks on the door and opens it softly. Charlie motions for them to come in.

Manuel apologizes for the interruption and says, “I just heard that AY-APP is facing a lawsuit because their app misdiagnosed disease for many patients from the C.L. Eye Health Center. The article says the incorrect diagnoses led to high costs for insurance companies on follow-up tests. In addition, the high stress of patients because of the situation impacted their ability to work and care for their families for months. It appears that an adversarial attack on their AI model was the cause for the high number of incorrect diagnoses.”

Manuel’s comment changes the direction of the meeting.

Immediately, Charlie thinks about the possibility of adversarial attacks.

Charlie says, “I know the news has shaken us all a bit.”

Charlie then asks, “Is everyone familiar with what an **adversarial attack** is? Let’s ensure we’re all on the same page regarding the concept, terminologies, and ways the system can be attacked.”

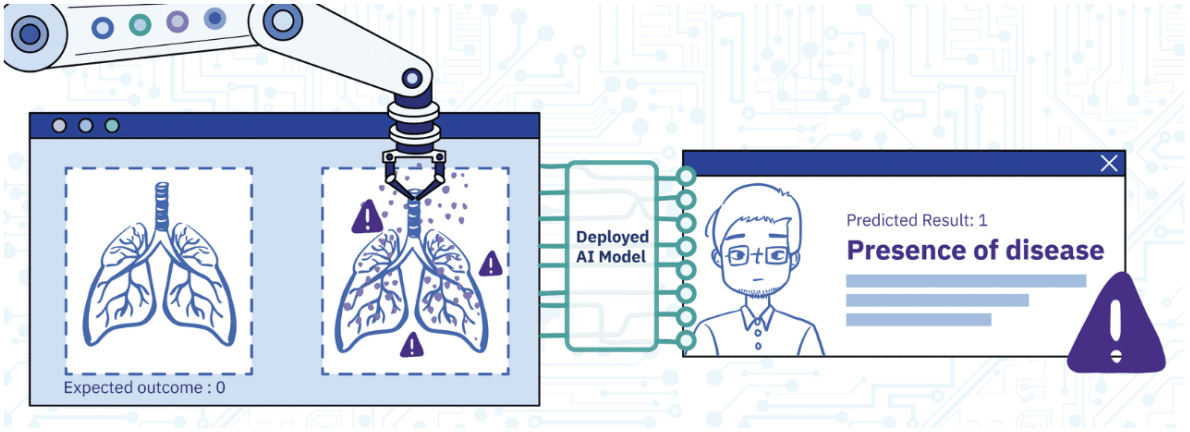
Then, she asks Prashant, the Chief Data Officer and a Data Scientist, to explain.

Explain the issue

Prashant says, “Let’s start with the adversarial attack explanation. Adversarial attacks are intentionally carried out on AI systems to accomplish a malicious end goal by taking advantage of AI system vulnerabilities.

“An adversarial attack aims to negatively impact the system performance, exploit data used, and corrupt the model logic. The one who takes advantage of the AI system vulnerabilities to accomplish their motive is called an **adversary**.”

“The following image is an example of how an adversarial attack can affect our system performance.”



“Here, the adversary attacks the system by adding small changes called **perturbations** or **noise** to the input image. The perturbation can be minimal and imperceptible to human eyes. However, when sent to the deployed AI model, the modified image results in an undesirable or incorrect prediction.”

“This scenario is just for one user input. Can you imagine this scenario repeating for all input x-ray images?”

Charlie enters the conversation and says, “That’s scary. I can’t imagine our system getting into the hands of adversaries!

“Can you elaborate more on the adversary’s end goals, how adversary actions can impact our system, and ways the adversary can enter the system?”

“Sure,” says Prashant. “Let’s see some of the possible adversary goals and how they connect with our AI application.

Adversary goal

Get access to personal information

Adversary goal

Make system learn the data that is in favor of the adversary

Adversary goal

Force consistent misclassification of the input samples

Adversary goal

Make the system predict a specific outcome.

Adversary goal

Steal or recreate the AI model

Adversary goal

Recreate training data used for developing the AI model

Prashant notes, “There are multiple ways for an adversary to achieve the goals mentioned above including the following:

- Getting access to the training data and learning the data distribution
- Having permission to modify the data used for training and testing the AI system
- Having access to the model code and parameters
- Corrupting the user’s data and sending the modified data to the deployed AI system”

“The adversary might not necessarily experiment with only one way to enter a system. Instead, it is likely that the adversary might use a combination of different ways to attack a system. Also, the ways adversaries attack can vary from attack to attack.”

Prashant then sums up by saying, “Given the goals and needs, adversarial attacks can occur either at the model training or after the model deployment.”

The issue of poisoning

Prashant pauses here and thinks about potential attacks that can happen to their AI system.

“Although there are different types of adversarial attacks, the possibility of perturbing the input has a higher chance of occurrence. Therefore, based on the initial observation, our AI system can be susceptible to two types of adversarial attacks: **poisoning** and **evasion**.”

The team members look at each other with apprehension. One of them on the conference line asks, “Could you tell us more about poisoning? How does poisoning work?”

“Ah, good question,” says Prashant. “As you can see, poisoning can happen in the following ways during the model training phase:

- Injecting malicious samples into the training data
- Updating features and labels of the training data
- Modifying AI model architecture, parameters, and logic”

“There are several impacts of the poisoning attack. One is when the deployed AI model becomes sensitive to the malicious data's specific pattern.”

“I’ve set up a scenario to show you how poisoning can affect an AI system’s outputs.”

Prashant adds, “Do you all notice how the system gets sensitive to the corrupted input? Now, let’s see the evasion attack.”

The issue of evasion

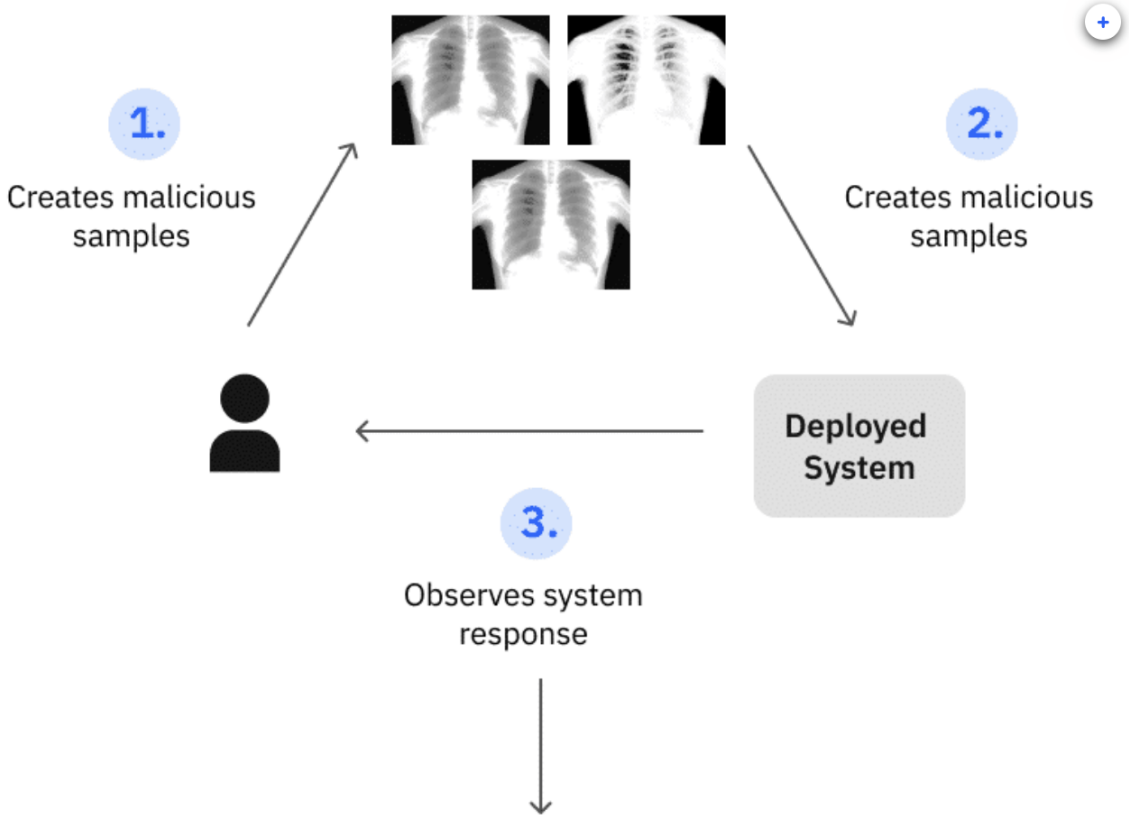
Evasion can happen after the model deployment phase in the following ways:

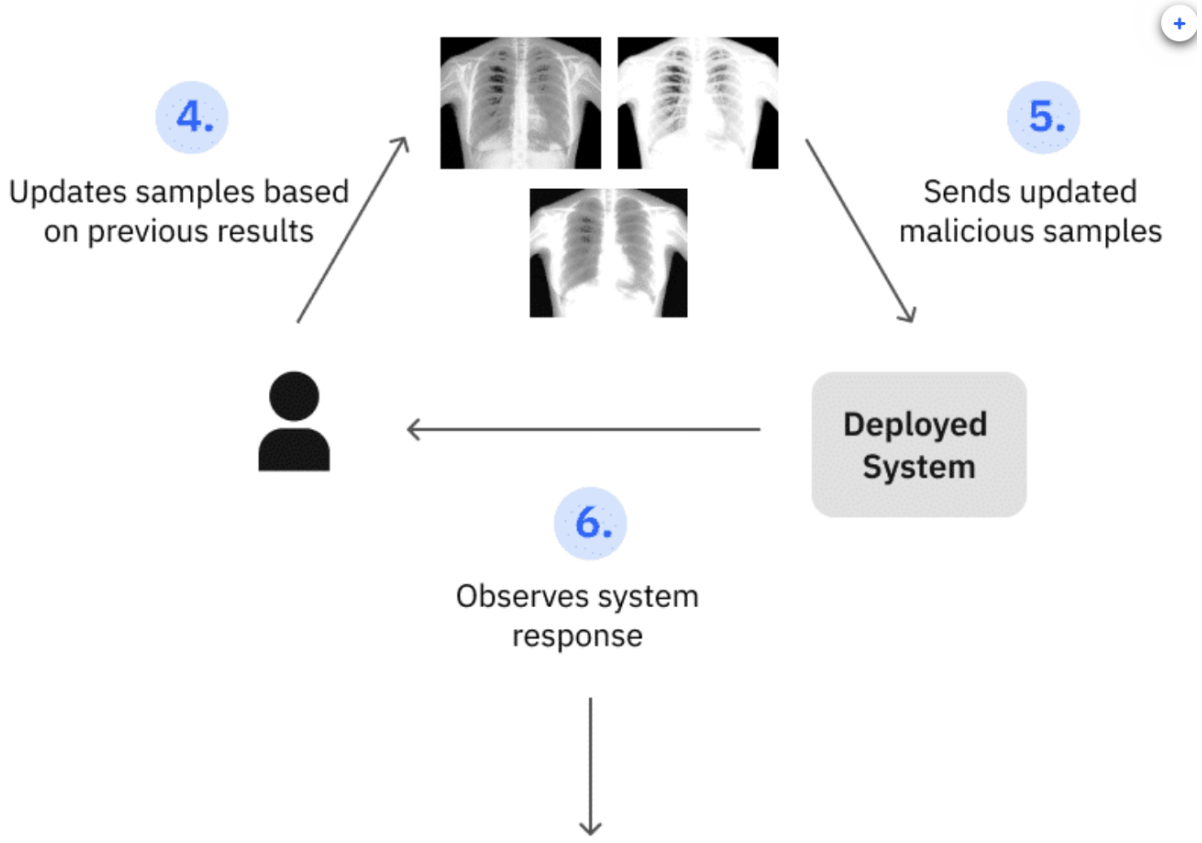
- Sending malicious test samples to the deployed model
- Corrupting test data sent to the deployed model

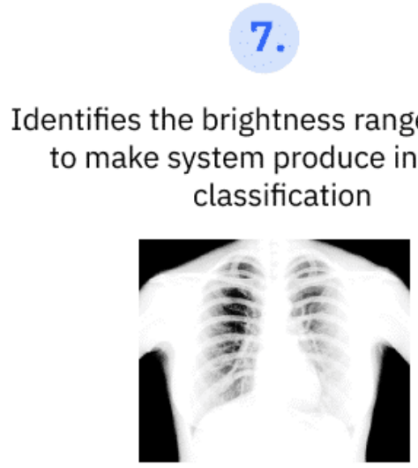
Prashant says, “There are several impacts of the evasion attack. One is when an adversary finds the changes to the input required to make the system produce an incorrect prediction. Let me demonstrate the concept with another scenario and visual.”

Scenario: Evasion

Select each of the three + icons to the right of the following image to learn how the adversary can evade detection.




4.
Updates samples based on previous results
5.
Sends updated malicious samples
Deployed System
6.
Observes system response



Prashant concludes his presentation by saying, “Now it’s time to make our system **robust** against the adversarial attacks.”

Charlie and the team members are relieved to know that their team has enough knowledge about adversarial attacks and are on the right track in thinking through the solution to prevent potential attacks.

Address the issue

The team now understands how an AI model can be corrupted by poisoning and evasion and why it is important that the model is robust. The team is now ready to go make the system robust against adversarial attacks.

Reflection: Robustness in AI

For question 1:

Is the source of the data used by the AI model important to know?

The source of the data is very important to know. AI models depend on data to learn, so data quality is crucial to trustworthy AI. In particular, it is important to know how the data was collected, who has access to it, and how it has been used.

For question 2:

Is it better to use public data? Should the data be screened before it is used?

It is neither better nor worse to use public data. Public data can be helpful, but it needs to be carefully vetted before use.

For question 3:

How would you monitor for attacks on the deployed model?

Using a tool to proactively monitor for attacks on a deployed model, like IBM AI Robustness 360, would be a good strategy.