ADVERSARIAL ARTIFICIAL INTELLIGENCE ATTACKS

Artificial Intelligence and Cybersecurity

When attempting to safeguard large or complex environments, artificial intelligence (AI) is a very helpful tool for cybersecurity. AI can assist in identifying unusual or questionable behavior so that analysts can further investigate it.

AI may be used for more sinister goals like detecting security flaws and ways to get beyond protection, though, just like many other cybersecurity tools. Artificial intelligence that is hostile to humans is what this is.

Machine Learning

We can develop algorithms using some types of machine learning that are nearly impossible to program manually. The drawback of these algorithms is that we don't fully comprehend how they decide what to do. It's incredibly challenging to understand the algorithm's "thinking process," even if it correctly determines whether a photo is of a cat or a dog, for instance.

Tainted Training Data

To learn how to work, machine learning algorithms need training data: You'll need a lot of images of both cats and dogs if you want an algorithm that can distinguish between the two. This creates a possibility for bad actors to affect the algorithm: They can alter the behavior of the final algorithm by altering the data used to train it. Tainted training data refers to data that has been maliciously altered.

Sadly, even data that hasn't been deliberately contaminated can be harmful: Unconscious prejudices and accidental errors can quickly skew training data in a negative way. Examples include resume assessment algorithms that discriminate against women or offensive picture recognition algorithms that incorrectly classify people of color. Even if the bias was inadvertent, when biased data is used to train machine learning algorithms, the bias is encoded and continues to be perpetuated.

AI Vs. AI

Even though we might not be able to fully comprehend how machine learning algorithms think, this does not mean that we cannot manipulate them. In fact, other machine learning algorithms are a great resource for creating strategies to deceive machine learning algorithms. We can create data that appears normal to us but tricks the target algorithm into providing responses that make no sense by training one algorithm to deceive another.

When you consider that image recognition algorithms are used for things like autonomous vehicles, spying, and verification, tricking them may seem hilarious. Other kinds of algorithms,

Protecting Your AI

Although there is no infallible method to guard against adversarial AI assaults, there are certain precautions that can be taken:

Maintaining the secrecy of training data makes it more difficult for malicious parties to study the data and can assist prevent unwanted modifications.

Another alternative is to train algorithms to identify and prevent attacks from hostile AI.

Conclusion

Since cybersecurity is a large topic, many other fields—from machine learning to business logistics—will contain some cybersecurity components. It's a good idea to consider the security implications of your projects as you work on them, including potential attack vectors and repercussions. To be "excellent" at cybersecurity, you don't need to know everything about it; all you need to know is enough to engage in critical thought and conduct productive research on the topics you don't fully understand.