AI Safety: A Survey of Cybersecurity Vulnerabilities in Artificial Intelligence Systems and Measures to Prevent Attacks

Kavya Karunakaran
Computer Science and Software Engineering
California Polytechnic State University
San Luis Obispo, CA
kkarunak@calpoly.edu

***Abstract*—This paper presents a survey of methods of cybersecurity attacks against artificial intelligence and machine learning models, as well as methods that can be employed to protect against these risks. A recommendation for future work in the field is also discussed, as well as impacts of research in this area.**

I. Introduction

In the past decade, artificial intelligence and machine learning have transformed from a distant possibility looming ahead, to a technological tool so foundational that it is found in almost every aspect of society. Every day new advances are made, and new technologies are created utilizing AI and machine learning. With the rise of ChatGPT, self-driving vehicles, and other AI-powered software, AI is all around us, and it's only going to become more prominent. Still, with all the advances that have been made, AI and machine learning are very new tools and there is much that has not been discovered about it. This raises an important question–Is AI safe? We already know a lot about cybersecurity attacks and defenses for traditional software, but we have not had much time to research this for AI systems. This is the question that has been researched in recent years by researchers and AI-specializing companies.

First, some background on AI and machine learning–AI and machine learning systems work by creating a model based on some algorithm (there are many, depending on whether the system is AI or machine learning, and the purpose for which the model will be used). The model is first trained using a vast set of data. The more data is used for training, the more accurate the model will be. Then, using the trained model, the system can make predictions given new inputs. With this system, the data that the model is trained on is very important to how the model will end up working. This phase, as we will see, is a target for cybersecurity attacks.

Currently, there has been research done by a variety of academic and industry sources on identifying and categorizing different types and tactics of cybersecurity attacks on AI models. In this paper, we will discuss research done by the National Institute of Standards and Technology (NIST) in the United States, the MITRE Corporation, industry researchers, and the Open Web Application Security Project (OWASP). There is also work being done to create guidelines to create strong security against attackers on AI models. OWASP and NIST have created guidelines for this, as well as other researchers.

My personal motivation for researching this topic is that I'm very interested in AI and machine learning models and their applications, and learning about cybersecurity during a renaissance of generative AI models has made me wonder what kinds of vulnerabilities exist in those models. I have grown up watching how fast technology can develop, which has given me an appreciation for the importance of slowing down sometimes to evaluate. I think that it's worth considering if AI developers should slow down to consider the safety and potential consequences of the technology that they are creating. If I ever end up working to create or use AI models in the future, I want to understand the security risks associated with them so that I can create safe software. The first step for doing so is understanding and analyzing the current research that has been done on this area.

In this paper, I will give a survey of different types of attacks on and vulnerabilities in AI and machine learning models and analyze the similarities and differences between different categorizations of these threats that have been published by various researchers and organizations. Then I will give a survey of guidelines that have been published about what developers and users can do to protect against and prevent cybersecurity attacks on AI/machine learning models. Finally, I will discuss my own opinions about what kind of work needs to be done in this area in the future, as well as what the rise of AI means for cybersecurity.

The rest of this paper is organized as follows. In section 2, the methodology is introduced with a discussion of multiple systems of categorization of attack tactics on AI systems, followed by a discussion of ways to prevent and protect against each type of attack. In section 3, my opinions

about future work are presented. In section 4, we conclude the paper.

II. METHODOLOGY

A. Types of Threats

There is a variety of research being done on identifying and categorizing the types of threats that have been observed against AI systems. We will survey a few categorization systems. One such system was created by NIST. NIST categorizes threats against AI systems into four main types: evasion attacks, poisoning attacks, privacy attacks, and abuse attacks [4]. NIST defines an evasion attack as an "attempt to alter an input to change how the system responds to it" [4]. In this type of attack, a user can attack a model by providing it with an input that is designed to confuse the model to lead it to give an answer or take an action that is different from what is normally expected, in order to cause destruction or reveal information. An example of this is causing a self-driving car to change its route by adding road markings that confuse the system. A poisoning attack is modifying the training data by adding corrupted data to the data sources that the model is being trained on. This will cause the model to be trained with the corrupted data, affecting the way it behaves after training. In a privacy attack, an attacker uses the model to try to get the model to reveal sensitive information. For example, a chatbot can be given strategic prompts that lead it to reveal information. Finally, in an abuse attack, an attacker can insert incorrect information into a data source that the model is being trained on [4]. In observing this categorization, we can see that all attacks happen either during the training phase or after the training phase, and different types of information can be extracted or impacts can be made depending on which phase the attack is made during.

Another categorization system that goes into more depth is the MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems). The ATLAS is a database of tactics that attackers use against AI systems [2]. The database identifies fourteen different tactics, each with specific examples of places that attackers use the tactic, case studies in which researchers have attempted to carry out the attack on specific models for illustrative purposes, and techniques attackers use to apply the tactic. The tactics Reconnaissance and Resource Development involve gathering information about the AI system and resources the attacker can use to carry out attacks. They then use this information to gain Initial Access and ML Model Access in order to access the system to some particular level. They can also apply Privilege Escalation tactics to get access to higher-level permissions. Once the attacker has access to the system, they can run Execution tactics to "run malicious code embedded in machine learning artifacts" [2]. They can also maintain their access to the system by employing the Persistence tactic and leaving behind poisoned data or code. Other attack tactics include Credential Access to steal credential info, Discovery and Collection to understand how the AI system works and ML attack staging to use this information to tailor the attack to specifically target the system. Additionally, Exfiltration involves stealing information about the system, and Impact, which is causing destruction to the system in some way. In observing these different attack tactics, we can see the wide range of information and software that AI and machine models put at risk. There are ways to steal sensitive data at every step of the process of training and deploying a machine learning model, and multiple different tactics of approach for every step. Not only does the training data need to be secured against attack, but the model itself must be secured against ML Model Access, the credentials of employees working on developing the model need to be protected against Credential Access, and the system as a whole needs to be protected against Impact attacks. For tactics like Discovery and Collection, an attacker does not even need sophisticated attack software to employ an attack. They can simply use the AI system as a user and observe information that they can then employ against it in an attack. This once again demonstrates the importance of employing cybersecurity defense tactics on machine learning models, and how every part of the model software and training process needs to be secured.

The OWASP Top 10 Machine Learning Security Risks also introduces another form of attack [6] – the AI Supply Chain attack. In this form of attack, the attacker attacks the people who are involved with developing models in order to use their credentials or other information associated with them to access information about the models. This highlights yet another aspect of the development process that is a potential target for attacks.

B. Analysis of Threat Categorizations

Based on these different categorizations of threats against AI and machine learning systems, we can draw a few conclusions. Firstly, it is crucial that developers protect the data sources that are used to train models. The data that models are trained are extremely influential to the behavior of the model, so its protection should be taken very seriously. More than half of the tactics outlined by all of the categorizations we discussed are attacks on data sources. Additionally, we also learn that it is important to add protections to models so that they can recognize when sensitive information is about to be revealed. Much of the attack tactics identified are used for the purpose of stealing private information, either about data sources or the model itself, so protecting sensitive information in machine learning models should also be a priority. This also leads into a final area to protect–We must protect information about how the model itself works. As we can see from tactics outlined, attackers seek to understand how the model works so that they can take advantage of that understanding to get the model to reveal sensitive information, or to cause destruction in how the model works. However, these insights bring up some challenges as well. One characteristic of how many softwares in the past worked is that they either use open source code, or the software is available as open source code. This makes software accessible and makes the barrier to entry as a developer more accessible. For instance, many machine learning libraries that are used often currently are open source. How can we toe the line between

making software accessible and open source with keeping all of the information that is used by the model protected? Another dilemma I observe is that with models using such a wide range of data, there are often common data sources that are used in many models. For instance, the Iris Dataset from the University of California, Irvine Machine Learning Repository [3] is a famous dataset that is very widely using in machine learning to train datasets. How can we secure this dataset since it's already widely and easily accessible? If we must limit ourselves to only data that we can secure and protect, we close ourselves off to a world of data that is available to the public. These are dilemmas that should and will be explored and discussed more in the years to come as our understanding of artificial intelligence develops.

C. *Guidelines for Protecting Against Threats*

OWASP has published a set of guidelines titled "Top 10 Machine Learning Security Risks". This page gives an overview of the main categories of risks to machine learning models and how to prevent them. To protect against attacks the data sources that models are trained on, the strategies recommended include validating data sources, securing data storage so that data sources are stored somewhere outside of the cloud, keeping training data separate from validation data, and implementing strong access control measures. Another strategy OWASP suggests is model ensembles–This is a machine learning concept involving using multiple models that are each trained on a different subset of data [6]. For protection against attackers manipulating use of the model to steal information, OWASP recommends employing input validation. They also recommend adversarial training on the model–this is a strategy in which during the training process, a model is trained on examples of attacks in the input, so that if an attack is later attempted, the model will recognize it and flag the input as an attack rather than revealing information. To protect against attackers figuring out how the model works, developers can seek legal protection (ie. patents), and encrypt the model's code and training data. Stu Sjouwerman, founder of KnowBe4 Inc., a security awareness training platform, elaborated on this strategy in an article for Forbes [7]. Sjouwerman recommends encrypting sensitive data with the AES-256 algorithm, training employees on cybersecurity measures, and employing intrusion detection systems on models. He also recommends using watermarking, which is a strategy that allows developers to trace attacks back to an attacker. This can be done using public key cryptography. This shows how we can continue to use more traditional methods of cybersecurity protections, but we must repurpose them to apply to these new technologies.

III. Future Work

One part of the OWASP website on AI protections that stood out to me was a link to a wiki with a form for users to contribute suggestions to improve guidelines. This is a demonstration of how AI is still very much a developing field and the process of protecting it is an ongoing process. We must continue to observe and develop new strategies as technologies develop. It also demonstrates something unique about AI in today's age–with the rise of generative AI models like ChatGPT, AI is now widely used by people regardless of their technological background. Since these systems are so widely used and trained with so much data, it may be useful to get feedback from ordinary users about weaknesses that they have encountered or ways that the security can be compromised. Developers are more likely to find weaknesses this way. In my opinion, it's important for developers of widely used products like ChatGPT to take feedback from the public in order to improve their security measures.

Additionally, since AI is growing so rapidly, we run the risk of security not being researched thoroughly before models are made available to the public. Almost every major technology company at the moment is working on creating generative AI models or employing AI in some way in their products, and in some ways it can feel like a mad rush to see who can create the next new innovation first. In our rush to make money, compete, or be on the forefront of innovation, we must not overlook the safety concerns, because the safety concerns can have massive consequences. Perhaps laws could be implemented that require security guidelines to be followed in the creation of AI or machine learning models, or a board could be created to review models that handle sensitive information. Security should be a priority in creating products that are integrated into technology in as widespread a fashion as AI.

IV. Conclusion

With AI being increasingly interwoven with all of the technology that both we use as individuals and that our society runs on, and with how new a lot of this technology is, it's extremely important to pay attention to the security implications of AI and make sure we're protecting ourselves. It is important for developers to take the guidelines that have been laid out for cybersecurity protection seriously and incorporate them into development of models.

References

[1] "AI for Security, and Security for AI: Two Aspects of a Pivotal Intersection." *S&P Global*, www.spglobal.com/en/research-insights/featured/special-editorial/ai-for-security-and-security-for-ai-two-aspects-of-a-pivotal-intersection. Accessed 18 Mar. 2024.

[2] "AtlasTM." *MITRE*, atlas.mitre.org/. Accessed 18 Mar. 2024.

[3] "Iris." *UCI Machine Learning Repository*, archive.ics.uci.edu/dataset/53/iris. Accessed 18 Mar. 2024.

[4] "NIST Identifies Types of Cyberattacks That Manipulate Behavior of AI Systems." *NIST*, 4 Jan. 2024, www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems.

[5] "OWASP AI Security and Privacy Guide." *OWASP AI Security and Privacy Guide | OWASP Foundation*, owasp.org/www-project-ai-security-and-privacy-guide/. Accessed 18 Mar. 2024.

[6] "Owasp Machine Learning Security Top Ten." *OWASP Machine Learning Security Top Ten | OWASP Foundation*, owasp.org/www-project-machine-learning-security-top-10/. Accessed 18 Mar. 2024.

[7] Sjouwerman, Stu. "Council Post: Ai Models under Attack: Protecting Your Business from Ai Cyberthreats." *Forbes*, Forbes Magazine, 5 Oct. 2023, www.forbes.com/sites/forbestechcouncil/2023/09/13/ai-models-under-attack-protecting-your-business-from-ai-cyberthreats/?sh=12bc7b971cb5.