**The Potential Threat of Adversarial Machine Learning in the Software Industry**

**Question:** Given machine learning’s rapid incorporation into technology in the last ten years, does adversarial machine learning (maliciously attacking machine learning models to influence performance) pose a threat to the software development industry?

**Purpose:** To explore state of the art research and applications of adversarial machine learning and identify whether companies that depend/use machine learning are adequately prepared to protect their machine learning models against such attacks.

Machine learning has quickly become a popular “buzzword” for companies seeking to update their technology stacks with the state of the art. Yet unlike more traditional technology frameworks such as web development, and mobile app development … etc., that are mostly deterministic and easily tested, machine learning throws a wrench into what’s considered traditional software development. Machine learning is a flexible algorithm that learns and identifies patterns in datasets to automate decision making. Examples of machine learning applications include computer vision and natural language processing. Tesla’s autopilot feature uses computer vision, or CV for short, to analyze images of the road and tell the vehicle what to do, and there are many online chatbots that use natural language processing, or NLP, to respond to text queries. Flexibility is the greatest strength of machine learning, as well as its most dangerous feature. These models are prone to outside influence, which can cause harm when dealing with sensitive information. However this problem has not slipped under the noses of machine learning researchers. Between 2017 and 2019, there have been over 200 written papers on the potential pitfalls and failures of machine learning when attacked by an adversary (Marshall et al., 2018). Since academic research on attacking machine learning models, or adversarial machine learning, continues to be a focal point of machine learning and security research, the threat it poses to software development is also ever evolving.

**Adversarial Example Attacks**

Adversarial example attacks can cause machine learning models to make wrong decisions by slightly altering the input to the algorithm. Tesla’s self-driving car has been shown to fall prey to such adversarial examples when its Autopilot ECU, or electronic control unit, is compromised by an adversary (Tencent Keen Security Lab, 2019). These CV models take images as input, and evaluate the pixels before making a prediction on the corresponding image. Adversarial example attacks add specific perturbations to these images in order to fool the machine learning model to make a mistake in its prediction (Goodfellow, 2020). In extensive security research of the Tesla autopilot system, Tencent was able to successfully trick the computer vision modules of the windshield wiper and lane detection algorithms using adversarial examples. The research group successfully manipulated the fish eye camera image input, causing the CV algorithm to believe it was raining even when the image had no visible precipitation (Tencent Keen Security Lab, 2019). These researchers were also able to manipulate camera images of lanes and cause the CV algorithm to miss lane markings. However they concluded that this would be difficult to deploy, citing the robustness of the CV algorithm likely due to a very extensive training dataset (Tencent Keen Security Lab, 2019). Tencent’s attempt to use adversarial examples to manipulate Tesla’s autopilot was shown to be successful in precipitation detection, and less successful in lane detection, yet both offer insight into the potential threat and implications of adversarial attacks in a modern software setting.

Adversarial examples of CV models are not limited to perturbation of input images. A paper authored by researchers from Carnegie Mellon University and University of North Carolina show that it is possible to physically alter an individual’s appearance (in this case, with specialized glasses), to fool a CV model’s facial recognition algorithm (Sharif et al., 2017). This can present itself as a dangerous problem, especially when facial recognition is used for security purposes such as with air travel (Sharif et al., 2017). This attack does not require the same access to the models as with Tesla’s autopilot, instead introducing the possibility of fooling these computer vision models with physically created perturbations. Regardless, both cases of adversarial example attacks show that it is possible to manipulate computer vision models, which are commonly used in the software industry, by feeding intentionally perturbed examples to confuse the algorithms and lead them to make wrong predictions.

**Poisoning Attacks**

Poisoning attacks are a different kind of adversarial machine learning attack that has also shown to influence common machine learning models. This attack provides a model in training with data that is purposefully not indicative of the actual distribution, causing the model to make inaccurate predictions. Unlike adversarial example attacks, this approach is more theoretical, but still demonstrates how commonly used machine learning algorithms can be heavily influenced by adversaries with the right knowledge of how to do so. Researchers from the University of California, San Diego, were able to theoretically and experimentally show that a constantly updated machine learning model can be influenced over time by injecting specially designed malicious data points. This includes flipped labels of digit classification and slightly modifying data points that the model struggles the most with (Wang & Chaudhuri, 2018). While this paper’s argument focuses on mathematical and theoretical arguments with some empirical results, it does not address the concern it raises in an industry setting, instead indicating more research is needed to investigate the potential vulnerabilities of such an attack.

The adversarial machine learning field is littered with theoretical papers, but the effects of a poisoning attack have also been seen in an industry setting. Microsoft’s Tay Chatbot was released in 2017, designed to learn and perform online interactions through Twitter tweets (MITRE Corporation, n.d.). However, internet trolls quickly realized that they could influence the Chatbot’s learning dataset by posting inappropriate comments, poisoning the dataset that the machine learning model trained from. In less than a day, the product had to be taken down as its tweets had become racist and offensive (MITRE Corporation, n.d.). This real life example demonstrates that poisoning attacks are not academically bound, but a threat that can be taken advantage of in a deployed software product.

**Current Industry Stance**

Current literature on adversarial machine learning suggests that under ideal conditions, these algorithms can potentially fall prey to adversaries with malicious intent, but the perspective from software companies garners more mixed opinions. While tech giants such as Google, Microsoft, and IBM have pushed forward plans to protect their machine learning use cases, many companies that use machine learning have unclear or no commitment at all to defend against adversarial attacks (Kumar et al., 2020). In a survey paper conducted by Microsoft, 28 organizations were interviewed about their perspectives on machine learning security, out of which only three actively secure their ML systems, and six are ready to assign manpower to the task (Kumar et al., 2020). These interviews demonstrated to the researchers that ML engineers are ill-equipped to deal with adversarial attacks, with reasons ranging from overtrust in current ML frameworks, limited understanding of ML in general, and not knowing what to expect from adversarial attacks (Kumar et al., 2020). Regardless of if adversarial attacks can be successfully used to manipulate machine learning models, this research survey suggests that most machine learning engineers in the software industry are not prepared for the potential threat against their models.

While adversarial machine learning research seems to put priority on the exposure of machine learning models, larger software companies such as Google and Microsoft have begun creating detailed guidelines for good machine learning practices. Microsoft’s AI and ML security documentation includes frameworks for engineers to assess the possibility of adversarial attack while also providing them with detailed plans to combat such attacks (Marshall et al., 2018). Google’s page for Responsible AI Practices also lists recommended practices for ensuring secure models, and encouraging adversarial testing to identify areas where maliciously crafted inputs can harm model outputs (Google, n.d.). While both documents provide concrete steps to follow in ensuring model security from adversarial attacks, both documents also acknowledge the ever changing nature of machine learning, and emphasize that the defense against adversarial attacks still has additional areas of research and growth. These documents indicate that to those prepared, adversarial attacks are not threats with no countermeasures, but a possibility that requires diligent research and preparation.

These large companies may be beginning to treat adversarial machine learning as a real threat, but a machine learning researcher of Google Brain gives another point of view from a software perspective. Nicholas Frosst recognizes the current research on adversarial machine learning, but also suggests that the results from such may be due to the natural behavior of modern machine learning, and that there are much easier ways to fool a model (Synced, 2019). Frosst believes that these models are not designed to make correct predictions on every possible kind of input, and creating perturbed images to fool self-driving cars is much harder than physically removing the stop sign the algorithm is trying to analyze (Synced, 2019). Academic researchers have certainly picked at the topic of adversarial machine learning both theoretically and experimentally, but many seem to agree that the field still has a lot to be explored before it is labeled as a legitimate threat to the software industry.

**Conclusion**

Machine learning is a field that is growing at a rapid pace, which means adversarial machine learning, or maliciously attacking machine learning models, will evolve at a similarly rapid speed. Even in the last couple of years, there have been hundreds of papers that have exposed weaknesses in machine learning models, using both theoretical and experimental arguments. Academia is pushing the exploration of adversarial attacks as a real threat, but it does not appear to be academically bounded. There have also been cases of adversarial attacks successfully infiltrating models in the software industry, such as Tesla’s autopilot and Microsoft’s Tay Chatbot. A research survey on companies with significant usage of machine learning shows that while engineers seem to agree on the principle of facing adversarial machine learning, most are not equipped with the knowledge to deal with such attacks, and some question the threat they even pose. Adversarial machine learning may be part of an ever growing and ever evolving field, but tech giants like Google and Microsoft show that with the mindset and willingness to embrace academic research with new guidelines, machine learning algorithms can potentially be safe from adversarial attacks.

**References**

Goodfellow, I., Clark, J., Abbeel, P., Duan, R., Huang, S., & Papernot, N. (2017, February 24). *Attacking Machine Learning with Adversarial Examples*. OpenAI. Retrieved October 2, 2021, from https://openai.com/blog/adversarial-example-research/.

Google. (n.d.). *Responsible AI practices*. Google AI. Retrieved October 2, 2021, from https://ai.google/responsibilities/responsible-ai-practices/.

Kumar R. S. S., Nyström M., Lambert J., Marshall A., Goertzel M., Comissoneru A., Swann M., & Xia S. (2020, May). Adversarial machine learning-industry perspectives. In 2020 IEEE Security and Privacy Workshops (SPW) (pp. 69-75). IEEE.

Marshall, A., Rojas, R., Stokes, J., & Brinkman, D. (2018, December 4). *Securing the Future of Artificial Intelligence and Machine Learning at Microsoft*. Microsoft Docs. Retrieved October 2, 2021, from https://docs.microsoft.com/en-us/security/engineering/securing-artificial-intelligence-machine-learning.

MITRE Corporation. (n.d.). *Tay Poisoning*. MITRE. Retrieved October 2, 2021, from https://atlas.mitre.org/studies/AML.CS0009/.

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2017). Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. arXiv preprint arXiv:1801.00349, 2(3).

Synced. (2019, November 21). *Google Brain’s Nicholas Frosst on Adversarial Examples and Emotional Responses*. Synced. Retrieved October 7, 2021, from https://syncedreview.com/2019/11/21/google-brains-nicholas-frosst-on-adversarial-examples-and-emotional-responses/.

Tencent Keen Security Lab. (2019). Experimental Security Research of Tesla Autopilot [White paper]. Tencent. https://keenlab.tencent.com/en/whitepapers/Experimental\_Security\_Research\_of\_Tesla\_Autopilot.pdf

Wang, Y., & Chaudhuri, K. (2018). Data poisoning attacks against online learning. arXiv preprint arXiv:1808.08994.