Prospectus

Adversarial Attacks on Deep Reinforcement Learning
(Technical Topic)

Framing Public Policy for Adversarial Machine Learning
(STS Topic)

By

Jihyeong Lee

28 October 2019

Technical Project Team Members: Quinlan Dawkins

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

Signed: _____

Approved: _____ Date _____
Rider Foley, Department of Engineering and Society

Approved: _____ Date _____
Hongning Wang, Department of Computer Science

Introduction

From self-driving cars to fraud detection, machine learning and artificial intelligence (ML & AI) algorithms are being widely adopted in many facets of society, requiring stronger guarantees of safety and effectiveness. One threat to ML safety is the field of adversarial machine learning, which is the study of how to cause ML models to behave in an undesirable way. Specifically, an "adversarial attack" may trick a model into giving an incorrect output, prevent it from learning correctly, or even reveal secure information about itself or the data it was trained on (Liu et al., 2018). For example, a few markings on a stop sign may cause a self-driving car to incorrectly interpret a stop sign as a speed limit sign (Eykholt et al., 2018). As machine learning becomes more prevalent, it is important to ensure that models are robust, or in other words, work properly under a wide variety of conditions.

Currently, there is still much to understand in both the technical and policy implications of adversarial attacks. From a technical standpoint, building defenses against adversarial attacks has proven difficult for researchers, as many attacks exploit some of the fundamental structures of how ML models are built (Liu et al., 2018). Furthermore, researchers are constantly looking to discover new attack methods since they allow us to gain a better understanding of how these complex ML models work (Liu et al., 2018). Unfortunately, additional attack methods create a need for additional defense methods. From a policy standpoint, the sudden and rapid development of ML has outpaced policymakers' abilities to keep up, leading to a gap in legislation on controlling the technology. As such, there is uncertainty in how governments should prevent attackers and how they should ensure that companies are creating secure models.

In response to the lack of specific legislation to address adversarial attacks and general machine learning safety, several researchers have analyzed existing legislation and how it

interacts with adversarial attacks. For instance, adversarial attacks could be considered as a case of transmitting unauthorized code to the model, and thus may be covered under the Computer Fraud and Abuse Act (CFAA), the primary legislation that covers cybercrimes in the United States (Kumar, O'Brien, Albert, & Vilojen, 2018). However, the CFAA is not comprehensive in dealing with the implications of adversarial attacks. For example, liability legislation is still unclear on whether the company or the attacker should be held liable for the damage caused by a model misbehaving due to an attack (Kumar et al., 2018).

With all this in mind, I plan to gain a better technical understanding of adversarial attacks through my capstone research project, then make policy recommendations on how public policy should deal with them. My technical capstone will focus on adversarial attacks on a specific subset of ML known as reinforcement learning, while the thesis will consider attacks on machine learning in general.

Technical Topic

Machine learning is a technique of building a program that can infer the relationships between an input and a desired output. This is often done by taking a large data set then feeding it into a learning algorithm that will build an approximate mathematical representation of the data. The subset of machine learning that my research will focus on is the field of reinforcement learning.

Reinforcement learning poses machine learning in the following framework: the model, referred to as the “agent” attempts to take actions in order to maximize the “reward”, received from interacting with an “environment”. An example of a reinforcement learning program is AlphaGo, an AI algorithm that has beaten professional Go players in 2016 (Silver, 2016). In this

example, the environment is the game board, in which the agent AlphaGo can place pieces on the board in order to win the game, which is its final reward. In this game, we can see that the agent does not receive any sort of reward until the game is over, which leads to the notion of an *expected cumulative reward*, i.e. how much reward the agent will receive throughout its lifetime on a probabilistic average (Arulkumaran, Deisenroth, Brundage, & Bharath, 2017).

In general, creating a reinforcement learning algorithm can be framed as attempting to select the best action for the agent, given the state of the current environment, in order to maximize its expected cumulative reward. Mathematically, the problem can be modeled as a finding a *Markov Decision Process*, which provides an action given the current state that satisfies the *Bellman Optimality Equations* (Arulkumaran et al., 2017). The details of these mathematical formulations are beyond the scope of this paper, but one key conclusion is that the computational complexity required in order to solve these equations is very high. Therefore, there is a need to approximate the solutions to these equations, which is commonly performed by models known as *neural networks* (Arulkumaran et al., 2017).

The neural network is a popular ML algorithm that is roughly modeled after the human brain - information from an input flowing between multiple layers of artificial neurons. Thanks to their complexity, they are powerful enough to approximate the Bellman Optimality Equations for reinforcement learning (Arulkumaran et al., 2017). But despite this power, neural networks are far from perfect – they can be tricked into giving the wrong output. Adversarial machine learning entails ways to "trick" a ML algorithm. In particular, neural networks are one of the most common subjects of these attacks. There are several methods with different effect, ranging from preventing a model from learning the correct relationships, causing the model to leak data, or forcing the model to behave undesirably (Liu et al. 2018). One common method is by

performing modifications to an input so that they look normal to a human being, but force the model to be incorrect. As shown in Figure 1, Eykholt et al. (2018) present adversarial attacks that could be used to fool image recognition neural networks used to identify road signs.

Figure 1. Examples of adversarial attacks on image recognition neural networks. The stop signs are interpreted as 45 MPH speed limit signs and the right turn signs are classified as stop signs (Image source: Eykholt et al, 2018)

The two main methods of categorizing adversarial attacks are referred to as white box and black box attacks (Liu et al., 2018). In a white box setting, the adversary is given complete access to the model, whereas in a black box setting, the adversary's access is more restricted. While white box attacks describe a specific setting, there are a number of different types of black box attacks that depend on the restriction of the adversary's access to the model (Liu et al., 2018).

For reinforcement learning, attacks can take on a slightly different form than that of regular adversarial attacks. This is because attacks on reinforcement learning need to influence the agent to take incorrect actions over a long period of time, which means that the adversary may need to attack the agent multiple times. This is a costly requirement. Two types of white box attacks that address this issue are enchanting attacks and strategically timed attacks (Lin, 2017). Both types examine the model and try to determine a short sequence of adversarial attacks against the agent. In contrast, the attack proposed by Eykholt et al. (2018) in Figure 1 is based on supervised learning, which is a subset of machine learning that's simply based on predicting a label given an input, and thus the adversary only needs to attack once.

Our research will focus on either developing a novel type adversarial attacks on reinforcement learning, or coming up with a new defense mechanism. For developing attacks, we are interested in making attacks more feasible in reality. A couple of ideas include converting enchanting attacks to black box attacks, since adversaries in real-life will not have access to the full model, or restricting the amount of changes the adversary can perform to the input. For defense mechanisms, we are interested in coming up with a strategy for defending against enchanting attacks, potentially by detecting the attacks before they take full effect. Our plan is to work with our advisor Honging Wang to fully develop an idea before the end of December 2019.

STS Topic

As previously stated, my capstone research will explore adversarial attacks applied specifically to reinforcement learning algorithms, while my thesis will focus on public policy on adversarial attacks on broader machine learning algorithms. As stated in the introduction, there is a lack of regulation for dealing with adversarial machine learning, with the exception being the

CFAA, which is one of the primary pieces of legislation that regulates cybercrimes in the United States (Kumar et al. 2018).

However, the CFAA is not without its criticisms. In a joint essay by legal scholars and machine learning researchers, scholars argue that the wording of the CFAA is too ambiguous, making it unsuitable for regulating adversarial attacks (Calo, Evtimov, Fernandes, Kohno, & O'Hair, 2018). Through a theoretical analysis of case studies, they find that many cases of potential attacks can be interpreted in either direction of being covered or not covered under the CFAA. The authors also argue that if the legislation is indeed interpreted in its broadest form and used to protect against adversarial attacks, then this could also threaten progress on research. This is because the CFAA does not have a clause that excludes researchers, and could lead to a similar impediment in research progress as it did on traditional cybersecurity research (Calo et al., 2018). Therefore, any new legislation should take into consideration the trade-offs between allowing open research and over-prosecuting all attempts of performing adversarial attacks on machine learning systems.

Another legal scholar, Simmons (2016) argues that the CFAA is too vague as well, since the legislation was originally developed when computer systems were simpler. The original goal of the CFAA was to cover areas of crime related to computers where existing law failed to do so. For example, trespassing is a concept that's easy to define in the physical world. However, for a computer user, the definition is fuzzier: Is an employee accessing a dataset for some purpose other than work a case of trespassing? What about a Terms of Service violation? Should that be a criminal or civil charge? These are the types questions that the CFAA was supposed to answer. However, as computer technology evolved, the definitions of the CFAA were often too vague or required constant amendment. For example, Simmons discusses *United States v. Lowson*, a case

were the defendants wrote a script to quickly purchase concert tickets by bypassing a website's anti-scripting measures, which led to criminal charges. Simmons (2016) argues that the same act would not have been a crime if the defendants had done the same action by hand, creating an ambiguous boundary between physical and cyber-crimes.

In response to the shortcomings of the CFAA, several solutions have been proposed. Calo et al. (2018) call for a different set of legislation to govern adversarial attacks. Simmons (2016) proposes an entirely different approach altogether: rather than creating legislation that may become obsolete in a few years, assign or create a regulatory agency with enough power to create flexible regulations to the specifics of the technology. An agency would be less encumbered by the long-process of creating new laws, and will thus be able to adapt as quickly as the technology does (Simmons 2016).

However, in order to determine which solution to select and what sorts of rules new regulation or legislation around adversarial attacks should focus on, we must consider how machine learning models are exposed to adversarial threats in the first place and what an appropriate response should be. A study by researchers at Microsoft found that these attacks are becoming easier within our technologically involved society (Marshall, Rojas, Stokes, & Brinkman, 2018). One reason is that ML models are becoming increasingly dependent on public data sources, which leaves them vulnerable to malicious attackers messing with training data. Furthermore, the researchers believe that there needs to be a shift in many aspects of deploying models, including security practices and the ability to diagnose these models easier. Therefore, rather than providing strictly technical recommendations for combatting adversarial attacks, the authors believe studying the interaction between the model, its users, its training data, and its developers will lead to better solutions (Marshall et al., 2018).

Therefore, two STS frameworks will be effective in exploring this topic: Actor Network Theory (ANT) and Anticipatory Governance (AG). ANT studies the interactions between the technology and society by treating both as equal "actors" in a complex network of interactions (Callon & Blackwell, 2007). As ML becomes more widely adopted across a variety of fields, adversarial attacks to these systems will become more common, which will shape the technology, which will then in turn shape how society interacts with this technology. The other framework, AG, studies how to govern technologies before they exist, "anticipating" the problems that might arise from this technology (Guston, 2014). Machine learning and adversarial attacks are still new, meaning we have not encountered many problems yet. Thus, governments need to take a proactive approach, making anticipatory governance a solid candidate for analyzing the problem and developing new solutions.

Research Question and Methods

The existence of adversarial attacks threatens the future of the trustworthiness, and ultimately, the adoption of machine learning. Protecting against attacks pose many technical challenges, and the lack of public policy so far puts the technology in an even more precarious state. Thus, I will explore the following question through my thesis: How should we frame public policy to address adversarial machine learning?

I will answer this question primarily through the analysis of case studies in order to understand the factors that should influence the policy decisions. For example, it will be helpful to see if a regulatory agency is indeed better in regulating quickly evolving areas compared to a full statute. Additionally, there are several instances of adversarial attacks that have occurred already, and it will be helpful to examine how the affected parties responded to the attacks. One specific case study I will examine is the Microsoft Tay Experiment, which was an AI chatbot

that became racist after only 24 hours of interaction with real people on Twitter (Wolf, Miller, & Grodzinsky, 2017). Through examining the timeline of events, how Microsoft planned the experiment, and how it responded to the outcomes, I will explore how social groups interact with and influence the development of AI systems.

Furthermore, there are two crucial policy frameworks that should be examined in answering this question. The first is the European Union (EU) Guidelines for Ethical AI (European Commission, 2019). This is a framework released by the EU on how engineers should build safe and ethical ML systems. The guidelines dedicate an entire section on providing a checklist for building a robust system against adversarial attacks. The guidelines are currently undergoing a live-testing period that is scheduled to end in late December 2019. Once the testing period ends, I plan on examining the results alongside a content analysis of the framework (European Commission, 2019). The other policy that should be examined is the National Institute of Standards and Technology (NIST) Cybersecurity Framework (Shen, 2014). This framework provides a legal basis for how engineers should build secure computer systems. By noting some of the similarities between adversarial machine learning and cybersecurity, I will examine whether a similar framework will be effective in addressing adversarial machine learning.

Finally, I will explore the importance of wide participation in creating the scenarios and solutions necessary for effective anticipatory governance. For this, I will conduct interviews with industry experts on what they believe should be done in order to regulate adversarial machine learning, then perform thematic analysis on their responses. A few experts in mind are professors at the University of Virginia that are tackling this very question, such as Professor Lu Feng of the

Computer Science department. I also hope to secure interviews with experts outside of academia, such as the authors of the aforementioned Microsoft study (Marshall et al. 2018).

Conclusion

The main problem I hope to address is how we should shape regulation to ensure that machine learning models are protected from adversarial attacks. This will help machine learning become a powerful and reliable technology for a variety of critical applications.

In order to accomplish this goal, I will reach out to the experts to schedule an interview by the end of December. In parallel, I will analyze the aforementioned cases, beginning with the Microsoft Tay experiment. After that, I hope that the European Union will have finished publishing their findings of the Guidelines for Ethical AI, which I will then use to compare to the NIST Cybersecurity Framework. The analysis should be completed by February 2020. By the middle of April 2020, I will complete my thesis. Through the case analyses, I will discover the themes of how society interacts with AI, and how to design proper regulation that can protect the public from an insecure algorithm without creating overbearing restrictions that kill the development of machine learning altogether. I will also learn about the similarities and differences between cybersecurity and adversarial machine learning, and how lessons from the former can be used to shape decisions for the latter. Through the interviews, I seek to find the engineering requirements that the industry should adopt in order to solve this issue.

References

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, 34(6), 26-38.

Callon, M., & Blackwell, O. (2007). Actor-network theory. *The Politics of Interventions, Oslo Academic Press, Unipub, Oslo*, 273-286.

Calo, R., Evtimov, I., Fernandes, E., Kohno, T., & O'Hair, D. (2018). Is Tricking a Robot Hacking?. *University of Washington School of Law Research Paper*, (2018-05).

European Commission (2019). Ethics Guidelines for Trustworthy AI. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ..., Song, D. (2018). Robust Physical-World Attacks on Deep Learning Models. *Conference on Computer Vision and Pattern Recognition.*

Guston, D. H. (2014). Understanding 'Anticipatory Governance'. *Social Studies of Science*, 44(2), 218-242.

Kumar, R. S. S., O'Brien, D. R., Albert, K., & Vilojen, S. (2018). *Law and Adversarial Machine Learning*. *arXiv Preprint arXiv:1810.10731*.

Lin, Y. C., Hong, Z. W., Liao, Y. H., Shih, M. L., Liu, M. Y., & Sun, M. (2017). Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. *International Joint Conference on Artificial Intelligence.*

Liu, Q., Li, P., Zhao, W., Cai, W., Yu, S., & Leung, V. C. (2018). A Survey on Security Threats and Defensive Techniques of Machine Learning: A data driven view. *IEEE Access*, 6, 12103-12117.

Marshall, A., Rojas, R., Stokes, J., & Brinkman, D. (2018, December 2). Securing the Future of AI and ML at Microsoft - Security Documentation. Retrieved October 17, 2019, from https://docs.microsoft.com/en-us/security/securing-artificial-intelligence-machine-learning.

Shen, L. (2014). The NIST Cybersecurity Framework: Overview and Potential Impacts. *Scitech Lawyer*, 10(4), 16.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Dieleman, S. (2016). Mastering the game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587), 484.

Simmons, R. (2016). The Failure of the Computer Fraud and Abuse Act: Time to Take an Administrative Approach to Regulating Computer Crime. *George Washington Law Review*, 84, 1703.

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why We Should Have Seen That Coming: Comments on Microsoft's Tay Experiment, and Wider Implications. *ACM SIGCAS Computers and Society*, 47(3), 54-64.