SPECIALIZATION OF ADVERSARIAL INTERFERENCE DETECTION FOR DEEP REINFORCEMENT LEARNING

MITIGATING ADVERSARIAL THREATS TO ARTIFICIAL INTELLIGENCE INFRASTRUCTURE THROUGH DIRECT ACTION

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Maxwell Lennon

November 23, 2021

Technical Project by
Maxwell Lennon

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.

ADVISORS

Bryn Seabrook, Department of Engineering and Society

Yuan Tian, Department of Computer Science

Introduction

Within the past decade, the field of artificial intelligence (AI) has entered what has been termed an "AI Renaissance." Advances in computational power and specialized hardware, combined with the increasing availability of big data, have resulted in an explosive leap forward in the capability and applications for artificial intelligence, and especially for its sub-field of machine learning (Bridgwater, 2019). Here, machine learning (ML) refers to a set of tools that are used to enable a program to improve its performance on a task without being explicitly instructed on how to do so – by 'learning' from data. Since much of the resources and techniques that have been developed during the AI Renaissance are now publicly available, the boom has also provided large numbers of people with the ability to experiment with machine learning on their own. While the increased accessibility of ML has certainly fostered exciting creations and useful technologies, it has also sparked concern regarding adversarial machine learning (AML), which involves using machine learning to defeat other ML systems. AML has the potential to threaten many facets of our physical and digital infrastructure if left unchecked.

On a similar note of facing adversarial threats, machine learning's sub-discipline of deep reinforcement learning (DRL) has been shown to be vulnerable to manipulation by outside actors, especially involving the perturbation of observational inputs used by reinforcement learning agents to make decisions. This vulnerability has implications for the ability of a variety of mission critical tasks in fields such as drone navigation, robotics, autonomous vehicles, and medicine to be performed safely and accurately. To address this specific threat, prior research has developed a classifier with the potential to detect adversarial manipulation of DRL. The final technical deliverable, a new research paper, will seek to expand on the findings of this prior work by refining pre-existing detection methods for better performance in specified environments.

Technical

When significant risks, ranging from loss of essential funds to loss of life, are contingent upon software performing its job correctly, the ability to gain some insight into a program's decision making process is critical for oversight and damage mitigation. Reinforcement learning (RL) is a sub-field of machine learning (ML), a field which is defined by the process of causing a program's performance on a certain task to improve given data (Géron, 2017). Within this framework, RL involves optimizing a policy for an agent to follow based on the assignment of algorithmic 'punishments' and 'rewards,' the criteria for which are determined ahead of time (Osiński & Budek, 2018). More specifically, deep reinforcement learning (DRL) attempts to achieve complex variations of this goal using specific types of architectures involving many layers of neural networks. The uses for deep reinforcement learning range from thought-provoking demonstrations of proficiency in various popular games to high-stakes applications such as autonomous vehicles, robotics, and medical analysis. However, like other subdisciplines of machine learning, DRL has been shown to be vulnerable to adversarial manipulations (Goodfellow et al., 2015), in which a malicious outsider attempts to externally alter the actions of a RL-trained agent and induce incorrect behavior. The research project from which my technical project will be formulated attempts to address this problem by innovating a method for detecting whether an agent's behavior has been adversarially altered.

The research project that will be used to derive an independent technical research topic introduces a technique termed an *importance judger*, which is designed to assess the relevance of a given input (for example, the sensory data fed to a robot at a given time) towards the eventual assignment of a punishment or a reward. In other words, the importance judger seeks to determine the expected change in the reward (in actuality a function which the agent's behavior

attempts to maximize) based on a known change to the input. The role played by the importance judger in detecting adversarial manipulation involves comparing the changes in perceived importance for the same input over time. The theory is that a previously important observation or input that changes to being perceived as unimportant is indicative of an adversarial attack. The importance judger thus provides a tool with the potential to identify a compromised agent or model without analyzing the agent's behavior directly.

The research experimentation with Importance Judger thus far has resulted in detection accuracies up to 90% in the environments that have been used to test adversarial attacks (namely, agents trained to play the Atari games Breakout, Pong, and Seaquest). Already, however, limitations of the work have been identified; for example, the detector's success rate decreases when applied to agents facing more complex environments, since importance computation is more difficult due to the increased number of possible actions at a given time. This has resulted in the identification of possible directions for future work, such as designing a detector that is suited to a particular environment instead of general; the current team has theorized that this may result in increased performance for the targeted environments. The goal for the independent technical project is to expand on the current research by endeavoring to answer one of these identified avenues for future work. The deliverable will be a research paper detailing the methodologies and findings involved in the effort to expand on the current paper's investigation.

STS Topic

Most software exploits, while potentially dangerous, require skill to be used effectively; powerful attacks targeting artificial intelligence systems, however, can often be achieved with little to no human ingenuity using adversarial machine learning. Generally, adversarial machine

learning, or AML, refers to a technique of training two or more machine learning systems with opposing goals to one another (Géron, 2017). The applications for this simple idea are manifold, but some potentially dangerous technologies have arisen from it. One example involves the use of AML to create “deepfakes,” or highly realistic video sequences depicting fictional events (Westerlund, 2019), which may erode social trust in fact-checking capability, such as within the justice system. Other uses of AML have the potential to compromise people’s physical safety (Lennon et al., 2021), or to allow bots to convincingly pass for human online (Brown et al., 2020). As cutting-edge artificial intelligence solutions continue to be developed, it seems increasingly likely that AI systems will play a major role in the physical and digital infrastructure of our future. From self-driving cars to widespread facial recognition, the more that AI and machine learning become mainstream, the more they will be relied upon to perform correctly. Adversarial machine learning, by its very nature, compromises the reliability of AI-based systems to function as expected, which could threaten economic activity, personal privacy, and/or public safety.

Previous works by STS scholars have noted the existence of adversarial machine learning as a danger to AI-based systems. However, the focus of such pieces tends to be on the properties of AI that lead to the existence of this susceptibility, rather than on ways to address the vulnerability through sociotechnical methods. For example, *Enchanted Determinism: Power without Responsibility in Artificial Intelligence* by Alexander Campolo and Kate Crawford begins its discussion of AML by stating that “[The] failure to understand mechanisms underlying classifications has already produced such hazards”(Campolo & Crawford, 2020). Importantly, these sources agree that “[Machine learning’s] divergence between calculation and understanding *raises important questions about the application of deep learning in social*

domains" (Campolo & Crawford, 2020). Yet we can see that the framing of this sentiment places the onus on deep learning instead of postulating potential mitigations to the risks posed by adversarial machine learning. This research paper will attempt to fill this gap in the literature by exploring methods for meeting the challenge of AML directly. The question I plan to research is: *how can technological and structural forces be effectively leveraged in order to directly reduce the threats to physical and digital infrastructure posed by adversarial machine learning?*

In Alvin Weinberg's 1978 piece introducing the notion of the technological fix, one of his prominent critiques of the concept is that "Most technological fixes can do no more than help remedy the immediate problem that invoked the fix. In their wake they leave other problems which, in turn, are amenable to resolution by additional technological fixes" (Weinberg, 1978). The cat-and-mouse game of adversarial machine learning presents a near-perfect microcosm of this principle, in that adversarial techniques are able to directly make use of the best available defenses against them in order to improve their own attack. Thus, any technical solution to a given adversarial problem is but a temporary measure; its creation and its undoing are directly linked! The insufficiency of technological fixes alone to quell adversarial machine learning is also well captured by Byron Newberry's encyclopedia entry summarizing the technological fix: "The fundamental difficulty with technological fixes—or shortcuts—is the inherent incompatibility between problem and solution. Technologies are most useful for solving specific, well-defined, and stationary problems, such as how to get cars from one side of a river to the other (for example, using bridges)" (Newberry, n.d.). Because every "fix" changes the landscape of the problem by introducing a new challenge against which attackers can test their adversarial systems, the issue of AML is far from stationary, and not amenable to a simple technological fix.

Having argued that the technological fix is insufficient for tackling adversarial machine learning, the research paper will then turn to the framework of political technologies (PT); in particular, PT will be applied to the question of restricting access to AI innovations in order to curtail the development of adversarial countermeasures. Since, as discussed previously, adversarial learning depends on being able to access the system (in either a white-box or black-box setting) that is being targeted, scholarly thinking holds that there is a security benefit to be realized by maintaining secrecy over state-of-the-art algorithms. UC-Berkeley machine learning researcher Jenna Burrell, in a research piece on various sources of opacity in AI, writes, "Network security applications of machine learning deal explicitly with spam, scams, and fraud and remain opaque in order to be effective. Sandvig notes that this 'game of cat-and-mouse' makes it entirely unlikely that most algorithms will be (or necessarily should be) disclosed to the general public" (Burrell, 2016). Since the consensus appears to be that adversarial machine learning is strongly conducive to a system in which information about current scientific research is controlled, by the definition outlined in Langdon Winner's *Do Artifacts Have Politics?* (Winner, 1980), we can classify adversarial machine learning as an inherently political technology. With this implied structure of sharing information comes the implication of an unbalanced power structure, in which people and organizations with access to the latest AI algorithms are the only ones with agency to shape the technology that surrounds them and affects their daily lives. Applying Winner's framework of political technologies will allow for an exploration of the nature of this technological totem pole, and whether it can be avoided or made more socially just.

Methodologies

Research Question: How can technological and structural forces be effectively leveraged in order to directly reduce the threats to physical and digital infrastructure posed by adversarial machine learning?

In order to answer the structural component of the research question, the paper will use the Policy Analysis methodology outlined by *Basic Methods of Policy Analysis and Planning*.

The Policy Analysis will attempt to tackle the question of future legislation limiting the spread of artificial intelligence research; this will take place on a number of fronts. First, it will identify instances of current public policy that cover related topics, thus providing information as to any possible precedent that may exist for the potential policy changes under examination. The analysis will also attempt to understand the impact on the research community and on the general public associated with each legislative example uncovered, in order to provide some predictive insight as to the probable ramifications of any proposed policy solution. Based on the principles of policy analysis outlined by Carl Patton and David Sawicki in *Basic Methods of Policy Analysis and Planning*, the scope of the analysis will most likely be narrowed to direct nonmonetary policies, which include "the prohibition or restricting of actions by rules, regulations, standards, quotas, licensing, deregulation, or legalization, such as environmental laws and safety regulations" (Patton & Sawicki, 2013). Since the policy of interest specifically deals with the restriction of research distribution, this classification scheme will be useful in sharpening the focus of the analysis.

Conclusion

The STS research paper focuses on ways to address the present threats to economic activity, personal privacy, and/or public safety posed by adversarial machine learning. Broadly, the two main methods for tackling this issue are thought to be: 1) strategically developing countermeasures to make current systems more resistant to AML; and 2) enacting policy to limit the public sharing of cutting-edge AI research with the potential to put especially dangerous tools and techniques into the hands of people of dubious intent. The result of the STS analysis will be a research paper outlining a variety of possible strategies to reduce the harm of AML, including an attempt to evaluate the most likely sociotechnical impacts of each solution.

Related to the problem of mitigating adversarial threats, machine learning's sub-discipline of deep reinforcement learning (DRL) is vulnerable to manipulation by adversaries; this vulnerability has the potential to compromise the safety and effectiveness of many critical applications in fields such as drone navigation, robotics, autonomous vehicles, and medicine. The technical project, in an attempt to counter this danger, research seeks to detect adversarial manipulation of DRL via analysis of input importance, and to refine pre-existing detection methods for better performance in specified environments. The result of this effort, if successful, will be the prevention of severe costs being incurred, potentially including loss of life, via the effective defense of a multitude of AI-based capabilities.

References

Bridgwater, A. (2019, April). *What Drove The AI Renaissance?* Forbes. https://www.forbes.com/sites/adrianbridgwater/2019/04/15/what-drove-the-ai-renaissance/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv:2005.14165 [Cs]*. http://arxiv.org/abs/2005.14165

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Campolo, A., & Crawford, K. (2020). Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society*, *6*, 1–19. https://doi.org/10.17351/ests2020.277

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (First edition). O'Reilly Media.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *ArXiv:1412.6572 [Cs, Stat]*. http://arxiv.org/abs/1412.6572

Lennon, M., Drenkow, N., & Burlina, P. (2021). Patch Attack Invariance: How Sensitive Are Patch Attacks to 3D Pose? *ICCV 2021*, *2021*, 10.

Newberry, B. (n.d.). Technological Fix. In *Encyclopedia of Science, Technology, and Ethics* (pp. 1901–1903).

Osiński, B., & Budek, K. (2018, July 5). What is reinforcement learning? The complete guide. *Deepsense.Ai*. https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/

Patton, C. V., & Sawicki, D. S. (2013). *Basic methods of policy analysis and planning* (3rd ed). Pearson.

Weinberg, A. (1978). *Beyond the Technological Fix*. Oak Ridge Associated Universities.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, *9*(11), 40–53. https://doi.org/10.22215/timreview/1282

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1,), 121–136.