Thesis Portfolio

Stress Test: Evaluating the 'Importance Judger' Algorithm Against State-of-the-Art Attacks
(Technical Report)

Mitigating Adversarial Threats to Artificial Intelligence Infrastructure
Through Direct Action
(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Maxwell Lennon
Spring 2022

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

The STS Research Paper and Technical Project in this Undergraduate Thesis Portfolio are strongly linked by the mutual connection of adversarial machine learning. First, this common topic is presented in the Research Paper as a complex challenge of technological capability and policymaking on multiple levels. The multifaceted issue of how to effectively mitigate the dangers of potent adversarial technology being used to sabotage critical components of societal infrastructure requires careful consideration and attention to detail with regard to the potential ramifications of any action that may be taken. At the same time, the Technical Project grounds the abstract topic of adversarial learning in reality, providing an empirical demonstration of both attack and defense that showcases the arms race capabilities of the technology and makes the case for the need to harness its potential responsibly.

Deep reinforcement learning (DRL) has achieved great success in various security-critical applications such as robotics, drones, medical analysis, and autonomous vehicles. However, recent studies show that machine learning models are vulnerable to adversarial examples, which are intentionally crafted instances that aim to cheat the model into making incorrect decisions. DRL can be effectively attacked by adding perturbations to observations. To defend against these attacks, prior research has innovated a detector algorithm based on the judgment of the importance of an observation: that is, the extent of the effect of changing the action decision made at this observation to the final reward. It has been observed that when DRL

is under attack, a large proportion of important observations in DRL are judged as unimportant; the algorithm Importance Judger detects attacks based on the changes of importance for these observations. The detection accuracy and F1 score on important observations under different attack scenarios can be greater than 90%. To test the resilience of this defense methodology, the new contribution of the Technical Project is an evaluation of the Importance Judger algorithm against attacks generated using PA-AD, a multivariate adversarial technique thought to produce optimal interference against DRL agents. The performance of Importance Judger against PA-AD is evaluated in two OpenAI gym Atari game environments: Pong and Breakout.

Technological visions of the future are often replete with "smart" infrastructure, with advanced artificial intelligence (AI) orchestrating networks of sensors and actuators to automate routine aspects of human life. However, the fruits of AI contain the seeds of their own undoing, due to the threat of adversarial machine learning (AML), which has the potential to cause unintended behavior in otherwise reliable systems. This paper employs the concept of the technological fix and the STS framework of political technologies in order to answer the 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?* The paper explores the effectiveness of technical solutions to known adversarial techniques and performs direct nonmonetary policy analysis to determine the feasibility of restricting public access to state-of-the-art attacks and defenses in order to safeguard them

against malicious actors. This analysis should lead to a better understanding of the scientific and administrative best practices to protect the autonomy of future intelligent systems.

Working on the Technical Project and the Research Paper simultaneously was highly valuable, in that the Technical Project cemented the otherwise ethereal concept of adaptive adversarial machine learning into an observable phenomenon, giving a clear justification for why the topics discussed in the Research Paper are of immediate concern. At the same time, the analysis in the Research Paper helped shape the direction of the experimentation for the Technical Project by enabling a deeper consideration of what aspects of the technology were most critical to demonstrate and to further understand.