Mitigating Adversarial Threats to Artificial Intelligence Infrastructure Through Direct Action

A Research Paper submitted to the Department of Engineering and Society

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

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

Maxwell Lennon
Spring, 2022

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.

ADVISOR

Bryn Seabrook, Engineering and Society

Introduction - Need for Reliable AI

Advances in artificial intelligence (AI) have inspired visions of a future filled with autonomous systems, in which human error is obviated by fast, reliable computing capable of solving problems previously thought intractable for machines. Those dreams of an intelligent future, however, have been challenged by an unforeseen downfall of powerful but mysterious AI systems: adversarial machine learning.

This paper examines the possibility of developing algorithmic defenses against adversarial machine learning attacks, applying Alvin Weinberg's framework of the technological fix. The paper also utilizes Langdon Winner's 'political technologies' framework to explore the multifaceted issue of limiting access to adversarial machine learning research via institutional or governmental policy. Ultimately, the research presented addresses 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?*

Background

Recently, AI, and especially its sub-field of machine learning (ML), have surged past human capabilities in a number of task areas (Bridgwater, 2019). Machine learning works by configuring a highly complicated modeling system, such as a neural network, according to feedback obtained from the system's current performance. The configuration happens algorithmically, without direct human input, as a result of the model's experience, hence the term 'machine learning.' The combination of implicit learning and complex modeling typically allows

ML solutions to achieve high performance on a chosen task, but also has the effect of making the inner workings of most ML systems inscrutable to human observers. Consequently, the behavior of machine learning systems tends to be difficult for humans to fully anticipate, which leaves the systems vulnerable to adversarial machine learning (AML).

Generally, adversarial machine learning refers to the practice of training one machine learning system to oppose the goals of another ML system (Géron, 2017). This may involve training two systems simultaneously against one another, in the case of a generative adversarial network (GAN), or training one system to defeat an existing model. In either case, the performance and output of a "targeted" model are used to provide feedback to the adversarial model being trained. Because AML directly utilizes information about the system it is targeting, including existing defenses, to improve its attack, and because the techniques to perform AML are publicly available research, there is ample reason to be wary of strong attacks by malicious actors from a variety of backgrounds (Campolo & Crawford, 2020). To safeguard future systems on which essential funds, or even human lives, depend, the threat of AML must be actively addressed.

Most software exploits, while potentially dangerous, require skill to be used effectively; many malicious uses of artificial intelligence systems, however, can often be achieved with little to no human ingenuity using adversarial machine learning. 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; for example, the camera systems used by a self-driving car to interpret the car's visual

environment can be manipulated by placing adversarially optimized "patches" on objects such as street signs (Lennon et al., 2021). AML could also be used 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. It is therefore imperative that some plan or course of action be developed with the goal of mitigating the potential harm to personal property, or even to life and limb, that rogue agents could wreak with the misapplication of adversarial technology.

Prior Work, Political Technology, and the Technological Fix

Previous works by STS scholars on adversarial machine learning tend to focus on the properties of AI that lead to adversarial 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 attempts to fill this gap in the literature by exploring methods for meeting the challenge of AML directly.

The term 'technological fix' was coined by atomic research pioneer Alvin Weinberg, who introduced the concept as the notion that some problem facing society could be solved by some new piece of technology alone. Problems arise when the issue that the technological fix aims to solve is a symptom of a larger structural, technological, or even societal problem. 'Band-Aid' applications of various technologies may produce temporary relief, but the same underlying problem will more than likely cause another issue that is not addressed by the previous technological fix. By the nature of adversarial machine learning, any technical solution to a given adversarial problem can be attacked using information about the defense itself. In *Wild patterns: Ten years after the rise of adversarial machine learning* by Battista Biggio and Fabio Roli, this relationship is acknowledged: "…machine learning and pattern recognition techniques turned out not to be the definitive answer to such threats. They introduce specific vulnerabilities that skilled attackers can exploit to compromise the whole system" (Biggio & Roli, 2018). This research paper builds on the previous analysis by examining the ability of AML to directly defeat defenses by training on them, thus sabotaging the technological fix.

This research paper also makes use of the framework of political technologies (PT); in particular, PT will be applied in order to explore the possibility of controlling access to AI research in order to slow the development of potentially harmful adversarial technology. Political technologies are ones that either necessitate or strongly imply a given political or societal

structure in order to be used effectively or as intended. Based on the notion that unrestricted access to adversarial techniques has the potential to create hazards for public safety and/or social welfare, we can classify adversarial machine learning as an inherently political technology. This paper applies Winner's framework of political technologies in order to explore the nature of this technological hierarchy and whether it can be made socially just.

Policy Analysis

The question that this research endeavors to answer 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 order to answer the structural component of the research question, the paper uses the Policy Analysis methodology outlined by *Basic Methods of Policy Analysis and Planning*. The Policy Analysis attempts to tackle the question of future legislation limiting the spread of artificial intelligence research via a multi-step plan. First, it identifies 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 also probes the impact of legislation on the research community and on the general public 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 is restricted 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). Any attempt to directly limit sharing of sensitive research information will constitute such a prohibition or restriction.

Results and Discussion

As the subsequent analysis will demonstrate in detail, the most effective and responsible course of action to combat adversarial machine learning would consist of a combination of technological development, to solve existing attack challenges facing AI systems; and policymaking, to ensure the security of sensitive research and prolong the effectiveness of state of the art defenses by limiting attacker knowledge. This dual-pronged approach has the potential to sustainably safeguard future smart infrastructure while allowing cutting edge machine learning research to take place in a responsible setting.

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, if it exists, is but a temporary measure; its creation and its undoing are directly linked! It is first necessary, however, to analyze whether technological fixes are effective against adversarial vulnerabilities in the short term.

In a 2015 publication titled "Explaining and Harnessing Adversarial Examples," a team of Google AI researchers innovate a new type of adversarial attack capable of misleading deep learning models. Relying on a mathematical property of neural networks called linearity, the researchers introduce an attack known as the Fast Gradient Signed Method (FGSM) attack. FGSM works by optimizing a perturbation to apply to an input (an image, for example) based on

the result of an process that prioritizes maximum likelihood of misclassification while constraining the overall perturbation to a given magnitude compared to the values of the input. In the paper, the researchers test the attack against models that employ a variety of neural network architectures, and discover that the attack method successfully results in incorrect classifications upwards of 87% of the time, with simple classifiers being effectively attacked over 99% of the time (Goodfellow et al., 2015). Thus, FGSM is demonstrably effective, and although it has since become more of a guiding blueprint for more sophisticated attacks to follow, at the time the paper was written, the attack was state-of the-art.

In the same paper that introduces FGSM, the aforementioned Google researchers also showcase one potential defense against their own attack, which amounts to the inclusion of adversarial examples generated using FGSM in the training data for the neural network. By exposing the network to what an adversarial image looks like, the researchers are able to train classifiers that are resistant to new malicious examples generated after training completion. The defense does not significantly improve error rates for the simplest classifiers, which struggle to capture the complex patterns needed to distinguish adversarial examples, but more layered and complicated neural networks are able to overcome the attacks enough to achieve only an 18% error rate (Goodfellow et al., 2015).

On the surface, the aforementioned results seem encouraging; they demonstrate that some reduction of the potency of adversarial attacks is indeed possible. However, this success comes with severe limitations. In the case of adversarial training, adequately safeguarding a model against a particular type of attack *X* requires *X* to be present in the data used to train the model; thus, any attack that is not specifically anticipated by the trainers of a machine learning model will present a vulnerability to a model defended by adversarial training (Qiu et al., 2019). This

property relates to one of the main challenges of adversarial machine learning, and security applications in general: the potential attacks contained in a given example are radically uncertain, meaning that in addition to not knowing whether a given attack is actually present in the example, the defender is unable to know what types of attacks are possible. No exhaustive list can be obtained, due to the potential for development of new attacks and/or modification of existing ones. Despite this significant downside, the insight that adversarial machine learning can be mitigated to some extent, even in the short term, is valuable; it implies that there is potential for machine learning technologies to play a role in the long-term solution to the threat of AML.

It has thus been demonstrated that technological defenses on their own may be inadequate due to the need to anticipate attacks ahead of time, creating vulnerability to novel attacking methods. However, the full extent of the problem is even more severe, as adversaries can examine the current available defense methods and develop attacks designed specifically to overcome them. For example, in 2021, Kaleel Mahmood et al. noted a gap in the state-of-the-art defenses (consisting of at least 9 distinct methods, some with multiple variations represented in the list). Previous works showcasing the defensive frameworks were all presented and analyzed with respect to a 'white-box' attack scenario, in which the adversarial attacker is assumed to have full access to the ML model under attack. Against attacks constructed in this type of setting, all of the defenses appeared to shine, showcasing impressive reductions in error rates compared to an undefended network. However, the authors were able to use their knowledge of the limitations of the defenses in order to inform the design of a new attack, which they termed an 'adaptive black-box' attack. Against the new attack, the majority of the state-of-the-art defenses were woefully inadequate: "…most defenses (7 out of 9 for each dataset) offer less than a 25% improvement in defense accuracy for an adaptive black-box adversary" (Mahmood et al., 2021).

This dramatic increase in attack effectiveness resulting from inside knowledge of a defense showcases the crucial role of information in the realm of adversarial technology.

Every machine learning "fix" introduces a new challenge against which attackers can test their adversarial systems, changing the landscape of the problem and demonstrating why technological solutions are ill-suited for tackling the issue of AML by themselves. The insufficiency of technological fixes alone to quell adversarial machine learning is 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.). As previously shown, the problem of adversarial learning is far from stationary; malicious actors can alter the problem simply by introducing a new attack.

Regardless of whether an attack represents an objective improvement over other methodologies, most defenses that exist at the time of the attack will be rendered weaker simply due to their lack of exposure to the new threat. When attackers take it a step further, and the design of an attack is directly motivated by the defenses that currently exist and are available for experimentation, the vulnerability of the defensive state of the art becomes even more apparent. There is a significant improvement in attack effectiveness against defenses that an attack has been developed to be able to defeat, which motivates the notion that it may be beneficial for overall security if the specific defense method being used in a given situation – as well as some of the best and most current adversarial defenses – are kept secret. Overall, however, it is clear that technological fixes are necessary but not sufficient to address the threats posed by adversarial machine learning, as they lead to an arms race in which security is never assured.

Having argued that the technological fix alone is insufficient for tackling adversarial machine learning, this paper will now 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 gains effectiveness through 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… 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. The question remains, however, whether policy precedent exists to facilitate control over the sharing of adversarial research, as well as uncertainty over the possible future effects of such an enactment.

The issue of 'freedom of research' has been debated before in various contexts. One of the first cases in which scientific research was subjected to legal scrutiny occurred following the fallout from the infamous Syphilis Study at Tuskegee conducted by the U.S. Public Health Service. To prevent injustices like those committed in the study from recurring, Congress passed legislation that "required researchers to get voluntary informed consent from all persons taking

part in studies done or funded by the Department of Health, Education, and Welfare" (*Tuskegee Study - Research Implications*, 2021). Signed into law in 1974, the National Research Act notably also "Requires grantees and contractees under the Public Health Service Act to establish institutional review boards to review research involving human subjects" (Rogers, 1974). Of course, the Act was originally aimed at an entirely separate research area, and primarily motivated by direct human rights concerns rather than future security implications. Nevertheless, the passage of the Act set a clear precedent that freedom of research was not limitless, and that scientific inquiry was instead subject to some level of government oversight under certain circumstances.

The U.S. National Institutes of Health Revitalization Act of 1993 served as another example of a legislative balancing action to reconcile freedom of research with public good; it "[mandated] the equal inclusion with white men of women and minority men in publicly funded US biomedical research and [made] funding contingent on that inclusion" (Kourany, 2016). As a result of the Act, scientists did lose some control over their own experimental design; the benefit of equality of access to health care for women and minorities was so significant that the move was scarcely controversial from a freedom of research standpoint, but for the purposes of the STS discussion at hand, it provides further relevant precedent for the idea of restricting the sharing of adversarial research findings.

Some of the most pressing cases regarding freedom of research, particularly freedom to publish research, come from virology. In one case, researchers in Australia stumbled upon a possible technique for producing a vaccine-resistant strain of smallpox, which could wreak unimaginable havoc if a lab leak were to occur. Another case involved Stony Brook scientists who used commonly available materials and information from the Internet to artificially

synthesize an active polio virus. A third case involved the genome sequencing and subsequent recreation of the Spanish flu virus that led to tens of millions of deaths in the early 20th century. All of the aforementioned studies were permitted to be published in top journals with wide-reaching audiences (Kourany, 2016). In each case, strong concerns were expressed about the conducting, and especially the publication, of the research involved. The second case, regarding the synthesized polio virus, is particularly worrying, as it clearly demonstrates that information alone can have potentially deadly consequences when those with malicious intentions are able to access detailed knowledge about the creation of dangerous substances or devices.

The threat posed by the above virology studies' free publication has led the federal government to discuss creating new or modified policy to deal with such concerns. In particular, these concerns surround research that can be classified as 'dual use,' i.e. potentially applicable to both civilian and military technology. Viral research meets the aforementioned criterion, as it can be used either to facilitate the development of new vaccines and antiviral treatments, thereby reducing illness in society, or to create bioweapons that pose a deliberate and severe hazard to human life. Scientific cases such as the virus studies "highlight the flaws in the current processes to identify and mitigate potential security risks associated with performing some research" (Gottron & Shea, 2013). To keep the public safe, argue congressional research specialists Frank Gottron and Dana Shea, Congress can take actions such as federal licensing of research. "For example, a national federal authority might license qualified researchers and research facilities and oversee research by licensed researchers in licensed facilities" (Gottron & Shea, 2013). Determining who qualifies to be a licensed adversarial machine learning researcher would be a nontrivial task, but similar processes are already in place for the issuance of security clearance necessary to conduct sensitive research in other fields.

Thus, there is precedent for some policymaking effort to place reasonable restrictions on the free publication of information regarding the capabilities of current adversarial attacks and defenses. As discussed previously, the security benefits of such restrictions are significant; they take the edge away from attackers and allow researchers to equip mission-critical infrastructure with technology that is a step ahead of the competition. However, with this hypothetical 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. Gatekeeping in academia is already a concern, and placing hard limits on who can and cannot conduct groundbreaking research does have the potential to limit careers, possibly in an inequitable manner that entrenches some existing lack of diversity in the AI community.

Conclusion

To ensure the continual integrity of future AI-driven systems and infrastructure, those systems must be free of compromise from adversarial agents. Due to the continuously evolving landscape of adversarial learning that exists today, the optimal course of action to deal with the threat of AML combines technological solutions with policies that extend their usefulness by limiting public knowledge of the most effective attacks and defenses. Properly addressing this multifaceted issue will lead to further nuanced discussions about how to achieve the proper balance of respecting scientific curiosity and the freedom of information exchange while limiting the potential for damage to property or person caused by the inappropriate sharing of sensitive research information. If the goal of safe, reliable machine learning systems is successfully realized, society can finally reap the benefits of a new dawn of artificially intelligent systems.

References

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84,* 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

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

Gottron, F., & Shea, D. A. (2013). *Publishing Scientific Papers with Potential Security Risks: Issues for Congress*. 27.

Kourany, J. A. (2016). Should Some Knowledge Be Forbidden? The Case of Cognitive Differences Research. *Philosophy of Science, 83*(5), 779–790. https://doi.org/10.1086/687863

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

Mahmood, K., Gurevin, D., van Dijk, M., & Nguyen, P. H. (2021). Beware the Black-Box: On the Robustness of Recent Defenses to Adversarial Examples. *Entropy, 23*(10), 1359. https://doi.org/10.3390/e23101359

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

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

Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of Artificial Intelligence Adversarial Attack and Defense Technologies. *Applied Sciences*, *9*(5), 909. https://doi.org/10.3390/app9050909

Rogers, P. G. (1974, July 12). *H.R.7724 - 93rd Congress (1973-1974): An Act to amend the Public Health Service Act to establish a program of National Research Service Awards to assure the continued excellence of biomedical and behavioral research and to provide for the protection of human subjects involved in biomedical and behavioral research and for other purposes.* (1973/1974) [Legislation]. https://www.congress.gov/bill/93rd-congress/house-bill/7724

Tuskegee Study—Research Implications—CDC - NCHHSTP. (2021, April 26). https://www.cdc.gov/tuskegee/after.htm

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.