POLITICS OF ADVERSARIAL MACHINE LEARNING

Kendra Albert *
Harvard Law School
Cambridge, USA
`kalbert@law.harvard.edu`

Jonathon Penney
Citizen Lab, University of Toronto
Toronto, Canada
`jon@citizenlab.ca`

Bruce Schneier
Harvard Kennedy School
Cambridge, USA
`schneier@schneier.com`

Ram Shankar Siva Kumar
Microsoft
Redmond, USA
`Ramk@Microsoft.com`

ABSTRACT

In addition to their security properties, adversarial machine-learning attacks and defenses have political dimensions. They enable or foreclose certain options for both the subjects of the machine learning systems and for those who deploy them, creating risks for civil liberties and human rights. In this paper, we draw on insights from science and technology studies, anthropology, and human rights literature, to inform how defenses against adversarial attacks can be used to suppress dissent and limit attempts to investigate machine learning systems. To make this concrete, we use real-world examples of how attacks such as perturbation, model inversion, or membership inference can be used for socially desirable ends. Although the predictions of this analysis may seem dire, there is hope. Efforts to address human rights concerns in the commercial spyware industry provide guidance for similar measures to ensure ML systems serve democratic, not authoritarian ends.

1 INTRODUCTION

All technological work has some political dimension. As Langdon Winner illustrated 40 years ago, a technologys design, systems, or arrangements can pave the way for certain social or political relations or foreclose specific possibilities (Winner, 1980). Winners most often cited example was the low bridges that crossed over the parkways that went between New York City and Long Island. Although it may seem that bridges with low clearance are not political, Winner explains that there was evidence suggesting that the underpasses were built this way to preclude public buses from using the roads denying those who relied on public transit, predominantly low-income New Yorkers of color, access to certain public spaces (Winner, 1980; Woolgar & Cooper, 1999). Today, we can speak the politics of algorithms, which can automate decisions in discriminatory ways(Noble, 2018; Eubanks, 2018; Benjamin, 2019) and spyware, which is used and abused by authoritarian governments to track, suppress, and harm human rights activists and dissidents (Harkin et al., 2019; Penney et al., 2018). Technology and related work have the potential to reinforce or undermine existing power relationships in the context in which it is used, and even technologies or related practices that appear neutral, benign, or even benevolent have potential impacts on civil liberties and human rights.

In cryptography, significant attention has been devoted to unpacking how particular research directions within the field may have implications for who deploy cryptographic technologies (Rogaway, 2015). There is also significant work that takes on these questions in the context of data science, Machine Learning (ML), and algorithmic decision-making more generally (Green, 2019; MacCarthy, 2018). We turn a similar lens on the development of defenses against adversarial attacks on machine learning, exploring how efforts to secure machine learning systems against attacks can have real-world harms that disproportionately fall on those who wish to resist the use of such systems. Our conclusion at its most basic is that those engaged in security work must understand that securing

---

*Authors ordered alphabetically

machine learning systems has consequences for the human rights and civil liberties of the subjects of those systemsconsequences that Winner would describe as political.

In this paper, we first discuss how machine learning system deployments increase the chance that adversarial attacks will be used by the subjects of the systems, who are unlikely to have a full say in their construction or deployment. Second, we provide real-world examples of how adversarial attacks could be used for desirable aims. Securing systems against attack may inadvertently suppress dissent or foreclose research that aims to shed light on how ML systems harm particular populations.We also discuss how the adversarial arms race may lead to the development and deployment of more invasive forms of surveillance. Finally, we conclude by suggesting directions forward. We draw an explicit connection to spyware, where activists, researchers, and civil society organizations have come together to recommend methods to reduce the harm of surveillance technologies.

Throughout this paper, we use facial recognition as the example of machine learning in mass-surveillance systems. Here, the political risks are clear. Facial recognition technologies (FRT) have been widely critiqued by scholars and activists (Garvie, 2016; Cyril, 2018), and several municipalities within the United States have banned the use of FRT by local law enforcement (Montgomery, 2019). Most salient to our argument, the harms of FRT usage by authoritarian governments are not theoretical, and the scope and usages of FRT are expanding rapidly (Mozur, 2019; Balaban, 2015).

2   POLITICAL IMPLICATIONS OF THE USAGE OF ML AND THE USE OF ADVERSARIAL ATTACKS

The goals of implementing many machine learning systems can be summarized as doing things at scale, generally without human intervention. Machine learning systems provide what science and technology studies author James C. Scott would call legibility (Scott, 1998; Thompson, 1967). Scott defines legibility as the process by which states took exceptionally complex, illegible, and local social practices, such as land tenure customs or naming customs, and created a standard grid whereby it could be centrally recorded and monitored. A machine learning system that aims to tag an image with the items it contains is a literal example of legibility; it makes readable that which was previously not visible.

Facial recognition technology makes that which only used to be done by humans (telling if a face matched) possible by machines. It is the combination of scale and legibility that makes machine learning systems uniquely attractive to governments and other institutions that seek to maintain control over large populations (Eubanks, 2018). In the words of Meredith Whittaker, director of AI Now, facial recognition is usually deployed by those who already have powersay employers, landlords, and the policeto surveil, control, and in some cases oppress those who dont. Current adversarial machine learning scholarship focuses on building robust ML systems and identifying novel attacks. Attention has not been paid to the people who are subjected to these ML systems. Subjects may not have the option of opting out or using democratic processes to control the systems, because of the speed and scale of deployment. (As we were writing this paper, news broke of ClearviewAI, a company that scrapes publicly available internet images to develop an application that, at least theoretically, allows for the identification of any persons face. In order to opt out, you are required to send them an image of your government identification.) In other cases, people are completely omitted. For instance, in 2019, DARPA, the research wing of the US Department of Defense, announced a challenge to build robust ML defenses, saying, We must ensure ML is safe and incapable of being deceived (Siegelmann, 2019) but it never explicitly states safe from whom? And deceived by whom? Although it is common within the security community to view those who wish to interfere with the confidentiality, integrity, or availability of systems as attackers, this framing belies the fact that those who resist such systems could just as easily be pro-democracy protesters or academics interested in evaluating the inclusiveness of training data as they could be malicious actors.

3   DESIRABLE ATTACKS ON ML

To an ML system, an attacker motivated by a legitimate human rights and civil liberties concern with the system and an attacker motivated to hide something from the system are the same (Cowen, 2014; Maly & Horne, 2014) A facial recognition system cannot tell if a person wearing a mask is a

protester or a bank robber. Below, we discuss how three different adversarial attacks could be used for socially beneficial methods.

- Membership inference can be used to determine whether a given data record was part of a models training set or not. Hardening against membership inference could prevent a researcher from determining whether a given person was included, which can be useful for efforts at machine learning accountability or determining the source of images for dataset training. Fredrikson et al. (2015) use their membership inference techniques to determine whether a given picture was present in a facial recognition database. Determining whether a given image is present could help an individual determine whether they are able to bring a court case against a given facial recognition provider.

- Adversarial examples (Goodfellow et al., 2014) involve modification of a query to get a desired result. Defenses against perturbation attacks aim at hardening models against common modifications in order to still allow for image analysis. The EqualAIs project runs a detection perturbation algorithm for the purpose of allowing individuals to make a certain image less likely to be detectable as a face (Pedraza et al., 2019) Obfuscation of this kind could be used by photographers who are documenting protests and would like to post the images without the potential for facial recognition software to automatically identify protesters (Schmidt, 2019).

- Hardening against model inversion attacks (Fredrikson et al., 2015) aim to prevent retrieval of private features. But when blackbox machine learning systems are deployed in contexts like access to credit, using attacks to retrieve the models may be one of the only ways to determine whether decisions are being made based on impermissible factors, such as race or gender.

4   ADVERSARIAL ARMS RACE

Its not just the preclusion of certain forms of attacks that has implications for the rights and liberties of people and groups that are subjects of machine learning systems. Efforts to secure ML systems against attack without proper attention being paid to the uses those systems may in fact lead to more invasive surveillance measures. Biggio & Roli (2018) remarked that securing ML systems is an arms race with attackers attempting to break into the system, and the defenders attempting to build robust defenses. Consider the following not-so-hypothetical scenario:

1. A surveillance state is using facial recognition to quash a peaceful protest. Dissidents try to escape facial recognition by using masks to occlude their faces.

2. In order to defend against these attackers who are compromising the integrity of the FRT systems, the state steps up the game with structured occlusion coding for robust face recognition (Wen et al., 2016); or using pre-trained model of full frontal faces to remove occlusion(Elmahmudi & Ugail, 2019) In order to defend against this, the dissidents turn to using adversarial clothing to completely evade the FRT.

3. To gain the upper hand again, the surveillance state uses the newly released adversarial robustness toolkit from Baidu (Goodman et al., 2020) that can help defend against adversarial clothing attacks. The dissidents now attempt to escape detection by wearing 3D-printed adversarial eyeglasses (Sharif et al., 2016)

4. To counter this, the surveillance state completely bypasses faces and uses other biometric technologies, such as iris scanning or gait detection (Hofmann et al., 2011), to identify people.

This example may seem speculative and far-fetched (protesters wearing adversarial eyeglasses?) but they illustrate the way in which standard security arms-race thinking can lead to the deployment of more draconian surveillance measures that suppress dissent.

5   DIRECTION FORWARD

Adversarial machine learning is not the only security technology with consequences for human rights. The commercial spyware industry also poses a similar threat, and efforts to address these

risks can provide a useful guidance for those engaged in ML security and the development and commercialization of adversarial ML toolkits. Spyware researchers have extensively documented how the multimillion-dollar commercial spyware industry has been used by authoritarian governments around the world to track, suppress, and censor human rights activists and civil society groups (Harkin et al., 2019; Penney et al., 2018). In response, a range of civil society, governmental, and industry actors have developed a range of ethical, corporate social responsibility, and human rights measures for commercial spyware industry participants (Anstis et al., 2019; Lauterbach, 2017; Access Now, 2019; Mackune, 2019) including industry standards for transparency, human rights due diligence, and commitments to human rights by design principles. Similar proposals, using these recommendations as a foundation, could be offered for the ML industry, particularly those engaged in securing ML systems and developing and distributing ML toolkits. Vendors and other ML-industry participants should:

- Commit to the application of the UN Guiding Principles on Business and Human Rights (UN Guiding Principles) to the industry, and commit to multi-stakeholder efforts to establish and operationalize their requirements within the industry, particularly on transparency, human rights due diligence, and remedies.

- Establish and comply with industry-wide standards for transparency and human rights policy/due diligence measures; blacklisting clients and customers based on human rights considerations (e.g., a government with a poor human rights record); and prohibitions on assisting clients with reconfiguring/hardening ML systems to resist attackers in contexts where human rights or civil liberties are at risk (e.g., protestors resisting FRT deployed by an authoritarian government).

- Commit to human rights by design principles, whereby ML systems and toolkits would be designed to make abusive deployment more difficult. For instance, mandatory features that report and log ML system uses as well as nature and form of attacks; or automatic tool disablement on detection of misuse or attempts at reconfiguration.

- Include mandatory provisions in contracts that incorporate UN Guiding Principle requirements; make human rights due diligence audits required; outline abusive uses and deployments that lead to contractual breach, termination, and/or waiver of confidentiality.

None of these ideas is perfect. Indeed, the lesson of attempts to combat authoritarian use of spyware has been that structured violation of human rights are difficult to overcome through voluntary practices. Nonetheless, these practices provide a framework for industry participants to deal with ML adversarial tools and practices.

6 CONCLUSION

Adversarial ML is at a pivotal moment. As these systems become more widely deployed, theoretical attacks and defenses rooted in the academic literature will become the stuff of peoples lives. We have merely scratched the surface of what a political analysis of adversarial machine learning attacks and defenses might illuminate. The adversarial ML community has the opportunity to learn from scholars of science and technology studies, anthropology, and critical race theoryas well as human rights and ethics literature more generallyand to be in conversation with protesters, researchers, and others who seek to attack systems for socially beneficial reasons. Through understanding lived experiences of resistance, applying the lessons of other disciplines, as well as reflecting upon the work of those seeking to prevent similar outcomes with spyware, the adversarial ML community can not just understand its work as political but take affirmative steps to ensure that it is used primarily for good.

REFERENCES

Access Now. Open letter to novalpina capital. Technical report, 2019.

Siena Anstis, Ronald J. Deibert, and Jon Penney. Submission of the Citizen Lab (Munk School of Global Affairs and Public Policy, University of Toronto) to the United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on the surveillance industry and human rights. Technical report, 2019.

Stephen Balaban. Deep learning and face recognition: The state of the art. In *Biometric and Surveillance Technology for Human and Activity Identification XII*, volume 9457, pp. 94570B. International Society for Optics and Photonics, 2015.

Ruha Benjamin. *Race after technology: Abolitionist tools for the new jim code*. John Wiley & Sons, 2019.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Deborah Cowen. *The deadly life of logistics: Mapping violence in global trade*. U of Minnesota Press, 2014.

Malkia Cyril, 2018. URL https://bit.ly/2UezW8e.

Ali Elmahmudi and Hassan Ugail. Deep face recognition using imperfect facial data. *Future Generation Computer Systems*, 99:213–225, 2019.

Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, 2015.

Clare Garvie. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Dou Goodman, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, and Zhang Huan. Advbox: a toolbox to generate adversarial examples that fool neural networks. *arXiv preprint arXiv:2001.05574*, 2020.

Ben Green. Data Science as Political Action. 2019.

Diarmaid Harkin, Adam Molnar, and Erica Vowles. The commodification of mobile phone surveillance: An analysis of the consumer spyware industry. *Crime, Media, Culture*, pp. 1741659018820562, 2019.

Martin Hofmann, Shamik Sural, and Gerhard Rigoll. Gait recognition in the presence of occlusion: A new dataset and baseline algorithms. 2011.

Claire Helen Lauterbach. No-go zones: Ethical geographies of the surveillance industry. *Surveillance & Society*, 15(3/4):557–566, 2017.

Mark MacCarthy. The Politics of Machine-Learning Algorithms. *Freedom and Safety*, 2018.

Sarah Mackune. Submission to the United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression,. Technical report, 2019.

Tim Maly and Emily Horne. *The inspection house: An impertinent field guide to modern surveillance*. Coach House Books, 2014.

Blake Montgomery. Facial Recognition Bans: Coming Soon to a City Near You. *Daily Beast*, 2019.

Paul Mozur. One Month, 500,000 Face Scans: How China Is Using AI to Profile a Minority. *The New York Times*, 4, 2019.

Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. nyu Press, 2018.

Daniel Pedraza, Dhaval Adjodah, Gretchen Greene, Josh Joseph, Thom Miano, and Francisco, 2019. URL http://equalais.media.mit.edu/.

Jonathon Penney, Sarah McKune, Lex Gill, and Ronald J Deibert. Advancing Human-Rights-by-Design in the Dual-Use Technology Industry. *Journal of International Affairs*, 71(2):103–110, 2018.

Phillip Rogaway. The Moral Character of Cryptographic Work. *IACR Cryptology ePrint Archive*, 2015:1162, 2015.

Blake Schmidt. Hong Kong Police Already Have AI Tech That Can Recognize Faces. *Bloomberg*, 2019.

James C Scott. *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press, 1998.

Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pp. 1528–1540, 2016.

Hava Siegelmann. Defending Against Adversarial Artificial Intelligence. Technical report, 2019. URL https://www.darpa.mil/news-events/2019-02-06.

Edward P Thompson. Time, work-discipline, and industrial capitalism. *Past & present*, (38):56–97, 1967.

Yandong Wen, Weiyang Liu, Meng Yang, Yuli Fu, Youjun Xiang, and Rui Hu. Structured occlusion coding for robust face recognition. *Neurocomputing*, 178:11–24, 2016.

Langdon Winner. Do artifacts have politics? *Daedalus*, pp. 121–136, 1980.

Steve Woolgar and Geoff Cooper. Do Artefacts Have Ambivalence: Moses' Bridges, Winner's Bridges and other Urban Legends in S&TS. *Social studies of science*, 29(3):433–449, 1999.