EXAMINING THE CFAA IN THE CONTEXT OF ADVERSARIAL MACHINE LEARNING

Natalie Chyi

1. INTRODUCTION

Artificial intelligence (AI) plays an increasingly integral part in many of the healthcare, security, and financial services that we use every day. As such, the compromising of these underlying AI systems is a huge cybersecurity concern. Not only are they vulnerable to traditional hacking threats, they are further threatened by adversarial ML attacks, which exploit the ways an ML model might fail naturally. This paper introduces the three main types of adversarial ML attacks, discusses whether they fall under the current anti-hacking paradigm of the CFAA, and suggests policy to combat them.

2. ADVERSARIAL ML AS AN EMERGING CYBER THREAT

The defining feature of an adversarial ML attack is that the attacker has "tricked" the algorithm in to making a "mistake", either by manipulating it into revealing private information or making a decision different than the one it intended to make. This stands in contrast to traditional cybersecurity attacks, which typically involve bypassing some sort of security protocol. Here, the attacker has instead taken advantage of the way the AI system looks at the world, and interacts with it in an authorized capacity. Adversarial ML attacks are generally characterized into three different types: perturbation, poisoning, and extraction attacks.

Perturbation attacks

Perturbation attacks happen during the inference stage, after the ML model has been trained and deployed, and involve feeding the model inputs that were designed to be misclassified. In the image recognition domain, this could mean adding a particular layer of noise or changing specific pixels of a digital input image so an ML model classifies it incorrectly.[1] For example, an attacker may overlay the image of a panda with a small perturbation, which is still recognizable as a panda to humans, but causes the ML

[1] Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes & Patrick McDaniel, Adversarial Perturbations Against Deep Neural Networks for Malware Classification (2018), https://arxiv.org/pdf/1606.04435v1.pdf

classifier to interpret the image as a gibbon instead.[2] Adversarial examples can also be physical. For instance, recent papers have shown that someone wearing a pair of glasses designed to be adversarial can trick facial recognition systems, and that a road sign with an adversarial sticker stuck on it can trick an autonomous vehicle into misreading that sign.[3] These adversarial examples have the potential to be life threatening in future scenarios - for example, if a military drone was tricked into mis-detecting a landscape or its details and started firing. Sound inputs have also been shown to work as adversarial examples.[4] This could have implications for any voice enabled software, including virtual personal assistants like Google Home, Alexa, and Siri. For example, a malicious actor could plant an adversarial sound command into a YouTube video or podcast episode that manipulates someone's Siri into turning on their phone's video recording function, downloading a specific application, or more.

Poisoning attacks

Poisoning attacks involve introducing false or misclassified training data during the training phase, so the model is trained on "bad" data and will consequently produce specific incorrect outputs at inference.[5] When a malicious actor is able to influence the labeling of training data, they are able to freely manipulate the model it informs. For example, an employee building a computer vision product for blind people might label all strawberry pictures in the training set as oranges instead. If uncorrected, the final product will proceed to wrongly classify all the strawberries it sees as oranges. In another example, a credit scoring startup might crowdsource the behavior scoring part of their program, asking users to supply and rate sample behaviors as risky or risk averse. A group of skaters could choose to poison the model by supplying and rating "skateboarding" as "risk averse" behavior thousands of times, thereby inducing the credit scoring platform into giving skateboarders a higher credit score.[6]

---

[2] Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, Explaining and Harnessing Adversarial Examples (2015), https://arxiv.org/abs/1412.6572

[3] Ryan Calo, Ivan Evtimov, Earlence Fernandes, Tadayoshi Kohno & David O'Hair, Is Tricking a Robot Hacking? (2018), https://ssrn.com/abstract=3150530

[4] Nicholas Carlini & David Wagner, Audio Adversarial Examples: Targeted Attacks on Speech-to-Text (2018), https://arxiv.org/pdf/1801.01944.pdf

[5] Attacking Machine Learning with Adversarial Examples, OpenAI Blog (2018), https://blog.openai.com/adversarial-example-research/

[6] Calo, *supra.*

Extraction attacks[7]

Extraction attacks are attacks that cause breaches of confidentiality, and include two types: stealing ML models through repeated querying, and membership inference (learning the private data of individuals the model was trained on). The latter is especially worrying because many algorithms have privacy-sensitive applications, such as medical diagnoses. For example, access to an ML model that prescribes personalized medicine could be exploited to learn confidential genomic information about individuals who were a part of the training data.[8] In another study, researchers showed the ability of attackers to accurately guess how respondents in a lifestyle survey answered questions, including whether they had said yes to cheating on their partner.[9]

Limitations to practical applications today

It is important to acknowledge that there are several crucial limitations to practical applications of adversarial ML today. For instance, many systems that rely on AI (i.e. self-driving cars) are not dependent on just one ML model, but utilize several sensors and algorithms to make decisions. To date, no adversarial examples have been shown to simultaneously defeat multiple models.[10] This means that the adversarial sticker on the road sign might trick one of the autonomous vehicle's algorithms but not all, and it ultimately wouldn't be fooled into misreading the sign. To this end, all the adversarial examples discussed in this paper have been academic proofs of concept rather than real life attacks carried out by malicious actors. However, as long as these vulnerabilities exist in ML models, it can be reasonably expected that they will be further developed and exploited by malicious actors. It is therefore important to start anticipating these attacks and thinking of defenses now.

---

[7] Within the computer science community, "extraction attacks" refer exclusively to model stealing. While membership inference attacks also involve extraction, they are not generally referred to as such. For the purposes of simplicity, I refer to both types under the single name of "extraction attacks" here.

[8] Matt Fredrikson, Somesh Jha & Thomas Ristenpart, Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures (2015), https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf

[9] *Id.*

[10] Nicole Kobie, To cripple AI, hackers are turning data against itself (2018), https://www.wired.co.uk/article/artificial-intelligence-hacking-machine-learning-adversarial

3. DOES THE CFAA ADEQUATELY ADDRESS ADVERSARIAL ML?

Since its implementation in 1986, the Computer Fraud and Abuse Act (CFAA)[11] has become the law predominantly used to prosecute any and all hacking-related offenses. This section provides an overview of prohibited actions under the CFAA, then analyses whether its provisions include the three types of adversarial ML attacks within its scope.

The CFAA prohibits three types of conduct:

1. Intentionally accessing a protected computer[12] without authorization and: obtaining information, causing damage (either recklessly or not), or accessing with intent to defraud.
2. Exceeding authorization while accessing a protected computer and: obtaining information, or exceeding authorization with intent to defraud.
3. Causing damage to a protected computer without authorization by knowingly causing transmission of some information or command, and this conduct must potentially result in aggregated loss of at least $5,000, physical injury, or a threat to public health or safety.

Perturbation attacks

When an actor tricks an ML system into misclassifying their adversarial example, have they hacked it?

Because these attacks don't involve any access or direct interaction with the "protected computer", they do not fall under either of the first two definitions of hacking. They may, however, be captured under the third definition.

In the example where an adversarial sticker is stuck on to a "STOP" sign, causing the autonomous vehicle to read the sign as "GO" instead, it is unclear whether the visual display of the sign would be considered a transmission of information. In the example where an adversarial sound input plays in the middle of a podcast and triggers Siri to do something, it is unclear whether the sound itself would be considered transmission of information or a command.

[11] 18 U.S.C. § 1029

[12] Case law has defined "protected computer" to include almost anything that has a microchip and some potential use in interstate commerce (Calo, *supra*). I assume most devices accessed in the examples below are considered "protected computers" unless stated otherwise.

In traditional hacking cases, transmissions have been relatively straightforward, and generally involve digital transmission of code or program directly within or to the device. There is limited guidance in statute or case law as to the medium through which this information must be transmitted through, and whether so-called natural language content would be included as a transmission to devices. It could be problematic if it were, as such a definition would make no distinction between any regular sound recording playing from a device (e.g. a line playing from a movie that Alexa interprets as being a command toward it, and follows that command), versus one that was adversarially altered to manipulate a Siri or Alexa. In *Fink v Time Warner Cable*, when the Court found that an actor could be liable under the CFAA even if they did not transmit malicious information - just the transmission alone, which caused damage, was enough.[13] So though the line playing from a movie may not be transmitting malicious content, it may still be covered. I will point out though that the "intentional access" component of the statute could be important to ensure that this distinction punishes only malicious actors.

It is also unclear whether these attacks would be considered "damaging" to the relevant protected computer. In the autonomous vehicle example, could impaired functionality resulting in an inaccurate classification be considered "damage" caused to the car's system? A low bar for "damage" was set in *Pulte Homes v. Laborers' International Union*, which held that a "barrage of calls and emails" caused "damage" to the victim's device because they prevented him from "sending and receiving at least some emails and calls", and that this ultimately "diminish[ed] the plaintiff's ability to use data or a system".[14] It is therefore possible that the adversarial road sign could be considered diminishing the driver's ability to use the autonomous vehicle properly. However, such a low bar would probably still not cover the Siri example, as the phone owner's ability to use their phone is not diminished just because the camera turns on and starts a recording.

The legality of a perturbation attack therefore depends on how the words "transmission of a program, information, code, or command" and "damage to the protected computer" are interpreted.

[13] *Fink v. Time Warner Cable*, 810 F. Supp. 2d 633 (S.D.N.Y. 2011).
[14] *Pulte Homes, Inc. v. Laborers' Int'l Union of N. Am.*, 648 F.3d 295 (6th Cir. 2011).

Poisoning attacks

When an actor intentionally trains an ML system on unreliable inputs, have they hacked it?

Poisoning attacks can clearly be carried out by means of traditional hacking – for example, if an attacker bypassed security protocol to mislabel training data. This conduct would obviously fall under the scope of the CFAA under the first definition of hacking. Because these are such straightforward cases, I will not spend more time exploring these and do most of the analyses on the more ambiguous situations I brought up in earlier examples.

In the examples given earlier, both actors had authorized access to the systems, as they were allowed and encouraged by the system's makers to interact with and alter them. The question then is whether their authorized access was "exceeded" when they started to poison the model.

In both these cases, the malicious actors were encouraged to perform the labor of labeling images or rating behavior - the issue is that they intentionally did their work in a way that would be considered incompetent, and contrary to the goal of the designer (to provide truthful and accurate classifications). It is unlikely that the second hacking definition applies in these cases because performing work at a low standard is clearly not an equivalent to exceeding authorization.

Another major issue to classifying a poisoning attack as hacking under the second definition is that the actor must have obtained some information through accessing the computer. In poisoning attacks though, the malicious actor never extracts information, it actually adds information. The second definition of hacking therefore probably does not apply to these types of attacks.

The third hacking definition is also relevant here. It is likely that the mislabeling of training data that happens in poisoning attacks is considered transmission of a command or program. Unlike the issue of medium in perturbation attacks, the commands transmitted here are digitally done through code. However, it is unclear as to whether damage was caused to a "protected computer" because this kind of attack happens at the training stage. The damage therefore isn't being caused to any particular device, but to a service / ML model hosted on the cloud. Just because an ML model performs inaccurately does not mean that the physical server it is hosted on, or the smartphone it is used on, has been

damaged, and this may therefore mean that the third hacking definition does not apply.

The legality of a poisoning attack therefore depends on how the words "exceeded authorization" and "protected computer" are interpreted.

Extraction attacks

When an actor tricks an ML system into disclosing private information, have they hacked it?

Model stealing generally occurs when an attacker replicates a model through API access or repeatedly querying it. Membership inference occurs similarly through reverse engineering.[15] For both of these attack types, neither the first or third hacking definitions apply. The second definition might though - the attacker could possibly be exceeding access to the ML model if the model's terms of service prohibits this type of reverse engineering. However, not only must the service's terms of service define this as prohibited conduct, but the company needs to have expressly warned the user that this is against the terms of service.[16] As companies are probably unaware when their models or training data has been stolen, and will therefore not be able to issue warnings, it is unlikely that extraction attacks are covered by the second hacking definition.

4. Regulatory suggestions to effectively handle adversarial ML

As shown above, the CFAA possibly covers some perturbation and poisoning attacks, but this is dependent on how the statutory language is interpreted. Given this "blind spot" in the law, I suggest a multi-prong approach to effectively addressing the risk of adversarial ML attacks.

Expansion of CFAA's hacking definitions

Firstly, I propose some ways the CFAA's hacking definitions could be expanded to include certain adversarial examples, in order to deter and punish bad actors who engage in these harmful actions.

[15] Michael Veale, Reuben Binns & Lilian Edward, Algorithms that remember: model inversion attacks and data protection law (2018), http://dx.doi.org/10.1098/rsta.2018.0083

[16] *Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058 (9th Cir. 2016).*

The CFAA could clarify that "damage to a protected computer" under the third definition of hacking explicitly includes causing impaired functionality (the autonomous vehicle example). It could also be more widely expanded to include any sort of tampering which may not impair the device's functionality, but cause it to act in a way that the device owner did not ask it to (the Siri example). "Transmission of a program, information, code, or command" could also be clarified to include non-digital transmission of natural language content (i.e. words or sounds that humans understand). While such a broad definition could capture examples far beyond just adversarial ones, I believe that the "intentional access" requirement is a significant safeguard against this. These amended definitions would include at least a range of perturbation attacks that might occur.

I would also recommend that the CFAA expand the categories of actions prohibited under the second hacking definition. Currently, an actor is only liable under the second if they obtain some sort of information. As noted above, this does not cover poisoning attacks because no information is obtained in such attacks. Alongside obtaining information, it could also prohibit knowingly modifying or adding inaccurate information once an attacker has accessed the system if it is against some terms of service or role specification (taking inspiration from *Facebook v. Power Ventures*[17]). For example, the FinTech company providing credit scoring services may have a terms of service that instructs users to give accurate ratings to the best of their knowledge. Objectively speaking, skateboarding is not a behavior that shows someone to be risk averse. The skateboarders and their thousands of ratings would therefore be acting against the terms of service, and would fall under the CFAA with this amendment. Another suggestion relevant to poisoning attacks would be to expand the definition of "protected computer" to include systems hosted on the cloud, rather than a particular device with a microchip. This is especially relevant to the third definition of hacking, which requires the damage to be caused against a protected computer rather than general damage (not specifically to a device, unlike the other two definitions). However, I am hesitant to suggest this as I am unsure of how to suggest this expansion without bringing potential significant prosecutorial overreach with it.

Finally, I have no suggestions to including extraction attacks within the scope of the CFAA. I do not believe that reverse engineering should be considered "hacking", and be considered punishable offense under the anti-hacking paradigm. I also think

[17] *Id.*

that the fields of data protection and intellectual property law are better suited to deal with these types of attacks,[18] but I will not discuss these further because it is beyond the scope of this paper.

Increasing incentives to build robust AI systems

The second prong of my recommendation focuses on promoting incentives to build AI systems that are resilient against attacks, and increasing security measures within ML systems. The ultimate goal of this category of policy suggestions is to decrease a software's attack surface by putting the burden on firms who produce such systems.

One way to do this is through increased FTC scrutiny. Most of the FTC's investigations for inadequate security by companies has been under "unfair or deceptive practice" under Section 5(a) of the Federal Trade Commission Act. However, these investigations almost all involve compromises in data protection. While this may be useful in the context of membership inversion attacks, this narrow assessment of "inadequate security" doesn't include most adversarial ML examples. My recommendation would therefore be to create a separate category of unfairness for inadequate security to known adversarial ML attacks.[19] Industry standards for negligent versus responsible ML development and deployment practices currently do not exist. As a result, I recommend that this category should include a framework which assesses resilience and risk - perhaps something similar to DREAD, which is the framework used to rate threats in the software community.[20] Policy makers' collaboration with researchers and technologists in creating these standards is crucial to ensure that the burden on companies is not unreasonably high, but that the standard is high enough so consumers are adequately protected. Additionally, there companies could be required to create incident response and remediation plans to prepare for future attacks. This would force firms to start thinking about adversarial threats as they're building the system, and implement security by design.

[18] Veale, *supra.*
[19] Calo, *supra.*
[20] Ram Shankar Siva Kumar, David R. O'Brien, Kendra Albert & Salomé Viljoen, Law and Adversarial Machine Learning (2018), http://export.arxiv.org/pdf/1810.10731

To complement this, there should be a statutory safe harbor from the CFAA for security research.[21] Researchers play a vital role for the public interest in auditing systems and keeping firms accountable, and threats of legal action under the CFAA chill this type of work. The speed at which a vulnerability is detected is especially important for systems with life threatening and public safety implications, such as self-driving cars, military drones, and medical devices. A research exemption would be another way to scrutinize the ML software that companies produce, and pressure firms to observe the safety standards required of them.

5. CONCLUSION

As ML is deployed on a larger scale in more critical systems, it is worrying that adversarial ML attacks probably do not fall within the CFAA as it currently stands. It is therefore crucial that definitions within the CFAA are clarified and expanded to include such attacks, and that companies building AI are incentivized to increase robustness of their systems against such threats.

[21] Daniel Etcovitch & Thyla van der Merwe, Coming in from the Cold: A Safe Harbor from the CFAA and the DMCA §1201 for Security Researchers (2018), https://ssrn.com/abstract=3055814