October 2019 · Dr. Sven Herpig

Securing Artificial Intelligence

Part 1: The attack surface of machine learning and its implications

An analysis supported by the Transatlantic Cyber Forum

Stiftung
Neue
Verantwortung

Think Tank at the Intersection of Technology and Society

Executive Summary

In the last five years, many large companies began to integrate artificial intelligence systems into their IT infrastructure with machine learning as one of the most widely used technologies. The spread and use of artificial intelligence will grow and accelerate. According to forecasts by IDC, a market research firm, worldwide industry spending on artificial intelligence will reach $35.8 billion in 2019 and is forecast to double to $79.2 billion in 2022 with an annual growth rate of 38 percent.[1] Today, 72 percent of business executives believe that artificial intelligence will be the most significant business advantage for their company, according to PwC, a consultancy.[2] In the next years, we can expect the investment boom in artificial intelligence to also reach the public sector as well as the military. This will lead to artificial intelligence systems being further integrated into many sensitive areas of society such as critical infrastructures, courts, surveillance systems and military assets.

For governments and policy-makers dealing with national and cybersecurity matters, but also for industry, this poses a new challenge they need to face. The main reason is that the diffusion of machine learning extends the attack surface of our already vulnerable digital infrastructures. Vulnerabilities in conventional software and hardware are complemented by machine learning specific ones. One example is the training data which can be manipulated by attackers to compromise the machine learning model. This is an attack vector that does not exist in conventional software as it does not leverage training data to learn. Additionally, a substantial amount of this attack surface might be beyond the reach of the company or government agency using and protecting the system and its adjacent IT infrastructure. It requires training data potentially acquired from third parties which, as mentioned, can already be manipulated. Similarly, certain machine learning models rely on input from the physical world which also makes them vulnerable to manipulation of physical objects. A facial recognition camera can be fooled by people wearing specially crafted glasses or clothes into thinking that they don't exist.

The diffusion of machine learning systems is not only creating more vulnerabilities that are harder to control but can also – if attacked successfully – trigger chain reactions affecting many other systems due to the inherent speed and automation. If several machine learning models rely on each other

[1] IDC: Worldwide Spending on Artificial Intelligence Systems Will Grow to Nearly $35.8 Billion in 2019

[2] PwC: 2018 AI predictions - 8 insights to shape business strategy

for decision making, compromising one might automatically lead to wrong decisions by the subsequent systems – unless there are special safeguards in place. A safeguard could for example be that for certain decisions a human always have to approve a decision made by such a system before it triggers further actions. In addition, machine learning makes detection and attribution of attacks harder. Detecting unusual behavior, distinguishing it from mistakes made for example by the developers and tracing it back to the original point where the attacker manipulated the system, is difficult as it requires full understanding of the decision-making process of the model. Attribution is further complicated by the fact that interference can take place in many stages in the virtual as well as the physical world. It might for example be impossible to prove who put patches on a street to misdirect passing autonomous vehicles.

In the past, both the Internet infrastructure and technology was built on it has not necessarily been secure by design. This offered militaries, intelligence agencies and criminal groups new avenues to pursue their respective goals. We should not repeat the same mistakes with machine learning. A key requirement is accurate threat modeling for machine learning applications designated to be deployed in high-stakes decisions domains (military, critical infrastructure, public safety) and implement security-by-design as well as resilience mechanisms and safeguards.

Governments and policy makers seeking to approach the security risks of machine learning should in a first step focus on where machine learning is applied at the intersection with national security. This includes traditional areas like law enforcement and intelligence services (e.g. facial recognition in surveillance, riot control or crisis prediction and prevention) as well as applications in infrastructures like process optimization in power grids or machine learning systems powering large fleets of autonomous vehicles. This domain is likely to provide a large divergence between the assumed low level of adversarial interference when designing machine learning until very recently, and the real-life threat model for its use cases. Considering that security is a precondition for successful digitalisation, security aspects of machine learning must be integrated on the level of national artificial intelligence strategies.

Even though it is difficult to predict whether information security will become a precondition for the successful development of machine learning going forward, securing machine learning, especially when it comes to high-stakes applications such as national security, is indispensable. The clock is ticking.

Acknowledgement

This analysis has been supported by the Transatlantic Cyber Forum working group on machine learning and information security through online collaboration and a joint workshop in Berlin.

The views and opinions expressed in this paper are those of the author and do not necessarily reflect the official policy or position of the working group members or that of their respective employer/s.

In alphabetical order, acknowledging essential contributions of:

1. Charles-Pierre Astolfi, French Digital Council
2. Cathleen Berger, Mozilla
3. Daniel Castro, Center for Data Innovation
4. Betsy Cooper, Aspen Tech Policy Hub
5. Thomas “halvarflake” Dullien, optimyze.cloud
6. Kenneth Geers, Atlantic Council [Fellow]
7. Joseph Lorenzo Hall, Center for Democracy & Technology
8. Mary Hanley, Cyber Policy Initiative at The University of Chicago
9. Maximilian Heinemeyer, Darktrace
10. Ariel Herbert-Voss, Harvard University
11. Wyatt Hoffman, Carnegie Endowment for International Peace
12. Michael Hsieh, Transformative Cyber Innovation Lab
13. Sven Jacob, German Federal Office for Information Security
14. Frederike Kaltheuner, Privacy International
15. Lydia Kostopoulos, Digital Society Institute at the European School of Management and Technology
16. Thomas Lawson, AXA Group
17. Igor Mikolic-Torreira, Center for Security and Emerging Technology at Georgetown University
18. Johannes Otterbach, OpenAI
19. Jörg Pohle, Alexander von Humboldt Institute for Internet and Society
20. Thomas Reinhold, cyber-peace.org and Institute for Peace Research and Security Policy
21. Kate Saslow, Stiftung Neue Verantwortung
22. Matthias Schulze, German Institute for International and Security Affairs
23. Caroline Sinders, Harvard Kennedy School and Mozilla Foundation [Fellow]
24. Sven Weizenegger, SUZA.io

Table of Contents

1. Introduction

Two of the most disruptive developments of the information age have been the ability to share information and communicate worldwide through the Internet and the subsequent all-encompassing digital transformation of everyday life. While both developments have been forces for good, improving the lives of billions, they have also been abused and exploited by governments and non-state actors on a massive scale. The "golden age of surveillance"[1], massive government and private sector data collection for domestic and international surveillance, springs to mind as a prime example. This dual use nature is inherent to modern technology and artificial intelligence is no exception. From the adversarial standpoint, digital transformation has offered militaries, intelligence agencies and criminals new avenues to pursue their respective goals by interfering with the underlying technologies. In part that is because both the Internet infrastructure and most that was built atop of it has not necessarily been designed with security in mind. Most existing discussions around security of IT systems and infrastructures (e.g. with regard to the Internet of Things[2]) are, by nature, retroactive. Even looking at the heart of national security, the information security of military weapon systems, shows a lot of work still to be done, leading some to introduce additional layers for protection, e.g. in the smart home[3]. Machine learning as the possibly next evolutionary stage of digital transformation has already been incorporated into business models, military operations, and more. Looking at the past developments and the information security shortcomings, it is crucial to look at the security aspects of machine learning to protect against malevolent abuse, thus enabling its full potential.

IT systems and applications are susceptible to attacks[4], and machine learning is no exception. In 2016 Papernot and Goodfellow concluded that "machine learning has not yet reached true human-level performance, because

[1] Peter Swire: The Golden Age of Surveillance

[2] Mozilla Foundation: *privacy not included;
Ben Francis: Introducing Mozilla WebThings;
Matt Burgess: Smart dildos and vibrators keep getting hacked – but Tor could be the answer to safer connected sex;
Pierluigi Paganini: Cranes, drills and other industrial machines exposed to hack by RF protocols;

[3] Sam Biddle: Government Report: "An Entire Generation" of American Weapons is Wide Open to Hackers

[4] The word "attack" is used in this publication in a non-judgemental way. In some cases, there might be normatively legitimate reasons for someone to attack a machine learning model. An example could be the use of physical items, such as glasses, to fool a facial recognition system (see "Input Processing") used by a repressive government.

when confronted by even a trivial adversary, most machine learning algorithms fail dramatically. In other words, we have reached the point where machine learning works, but may easily be broken"[5]. Considering that machine learning already is and will continue to be applied in a range of fields, including national security, it appears prudent to assess the information security of machine learning. This is something that has, on scale, only been done recently.[6] It includes looking at conventional attacks on IT systems, whose impact might increase with the widespread use of machine learning[7] as well as certain types of adversarial interference unique to machine learning systems[8]. These types of attacks can either be malicious or accidental. The distinguishing facet is the intent. A robust system needs to defend against both, but the defense strategy might differ. The current focus on implementing machine learning often appears to be on making it work rather than dealing with the harsh reality of an adversary-rich environment and the conclusion "that there might be active, adaptive, and malicious adversaries"[9].

In their thorough analysis, Horowitz et al. conclude: "The technological opportunities enabled by artificial intelligence shape the future, but do not determine it. Nations, groups, and individuals have choices about how they employ and respond to various uses of AI. Their policy responses can guide, restrict, or encourage certain uses of AI". In order to do that, there first needs to be an understanding of what machine learning is, not only its potential but also its susceptibility to adversarial interference. Carefully considering and weighing its security implications has to be done before machine learning reaches an adoption rate that would render it virtually impossible to secure a posteriori. The machine learning evolution must not end up as insecure as the Internet of Things or, to paraphrase Caroline Sinders[10], *machine learning won't reach its potential – and may actually cause harm – if it doesn't develop in tandem with information security*.

On the technical level, the cat and mouse game[11] of developing defense mechanisms and finding new techniques to attack machine learning is nothing

[5] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

[6] Jörn Müller-Quade et al.: Künstliche Intelligenz und IT-Sicherheit

[7] Center for a New American Security: Artificial Intelligence and International Security

[8] Greg Allen and Taniel Chan: Artificial Intelligence and National Security

[9] Simson Garfinkel: Hackers Are the Real Obstacle for Self-Driving Vehicles

[10] Caroline Sinders: Why UX Design For Machine Learning Matters

[11] Anh Nguyen, Jason Yosinski and Jeff Clune: Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

new, at least in the technical research community[12]. On the political level, it does not appear to have sparked much interest yet. It has however at least reached the executive branches of various governments as for example the latest joint French-German cybersecurity report[13] shows. It is therefore worthwhile, and hence the goal of this paper, to map and illustrate the attack surface of machine learning and its implications. This will hopefully contribute and serve as a basis for policy recommendations which enable and support technical approaches to secure machine learning systems.

A glossary for technical definitions can be found at the end of this publication.

[12] Blain Nelson et al.: Exploiting Machine Learning to Subvert Your Spam Filter;
Kathrin Grosse et al.: Adversarial Perturbations Against Deep Neural Networks for Malware Classification;
Kathrin Grosse et al.: On the (Statistical) Detection of Adversarial Examples;
Ian Goodfellow et al.: Attacking Machine Learning with Adversarial Examples

[13] Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

2. Machine Learning and Information Security

While machine learning has been around for quite some time, the increased availability of massive amounts of data on and improvements in hardware developments and computing power[14] have led to an enabling environment for crucial advancements in the past few years. Even though often conflated terms, machine learning is a subfield[15] and foundational basis[16] of artificial intelligence[17] which has been around since the 1950s. Machine learning consists of building statistical models that make predictions from data. Given a sufficient quantity of examples from a data source, i.e. the training data, with a property of interest, a machine learning algorithm makes a prediction about that property when given a new, unseen example. This can happen via internal parameters calibrated on the known examples, or via other methods. Machine learning includes curiosity learning, decision trees, deep learning, logistic regression, random forests, reinforcement learning, supervised learning and unsupervised learning.

Developing and deploying a machine learning model includes the following core stages:

- acquire data (to use as training data later on),
- prepare data (e.g. look for biases, sort and label it),
- choose one or more machine learning methods (e.g. supervised learning) and develop a classifier for the specific purpose (e. g. for image recognition),
- train the classifier with the training data,
- improve the classifier (e.g. adjusting parameters),
- setup the model in the deployment environment (e.g. in a car to facilitate autonomous driving) and run predictions (e.g. identifying the likelihood whether a street sign is a stop sign or not).

Not all models go through all those stages (e.g. unsupervised learning does not require labeling and online learning does not require training prior to de-

[14] Vishal Maini and Samer Sabri: Machine Learning for Humans

[15] Vishal Maini and Samer Sabri: Machine Learning for Humans

[16] The MITRE Corporation: Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD

[17] For a brief history of artificial intelligence, see:
Stephan De Spiegeleire, Matthijs Maas and Tim Sweijs: Artificial Intelligence and the Future of Defense;
The MITRE Corporation: Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD;
Vishal Maini and Samer Sabri: Machine Learning for Humans

ployment). There are also additional stages for certain methods (e.g. reinforcement learning goes through feedback loops) and models might need to be re-trained to improve them before being deployed again.[18]

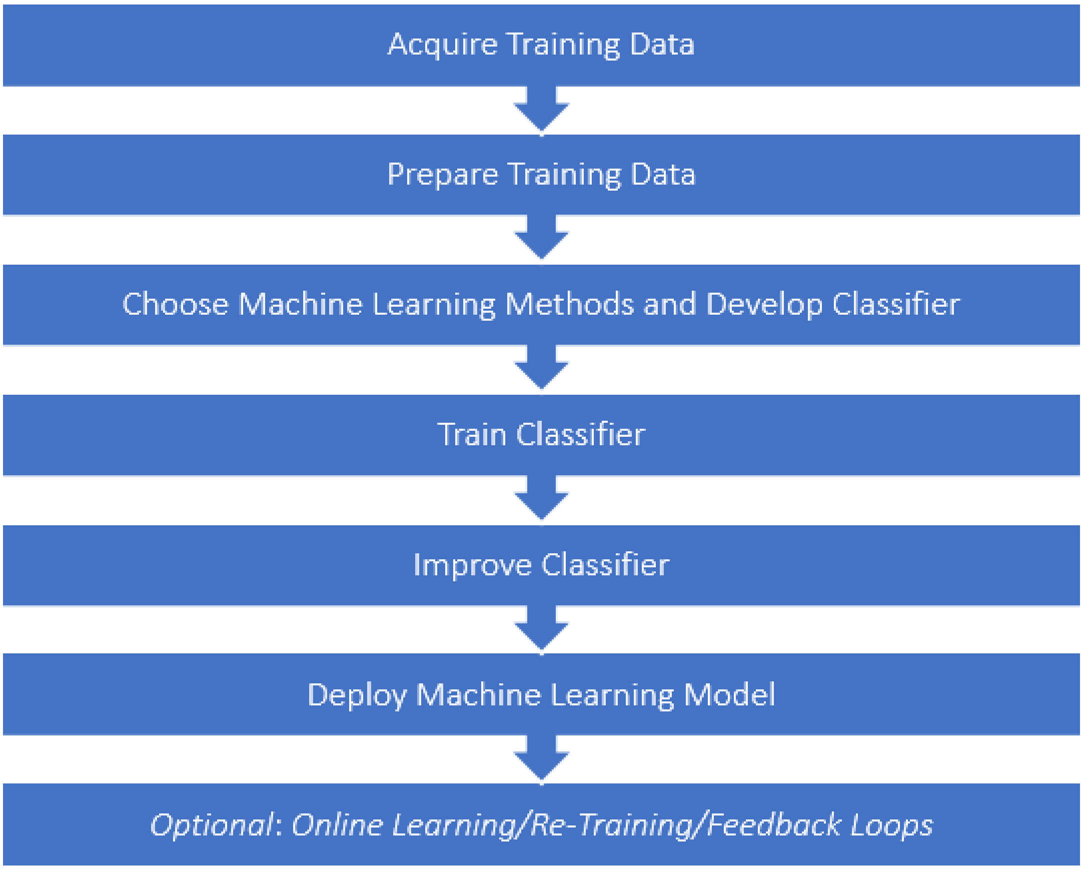
Acquire Training Data
Prepare Training Data
Choose Machine Learning Methods and Develop Classifier
Train Classifier
Improve Classifier
Deploy Machine Learning Model
Optional: Online Learning/Re-Training/Feedback Loops

Figure 1: Simplified flowchart of machine learning development and deployment

According to the US National Institute for Standards and Technology, information security is defined as "the protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability"[19]. This conventional analytical frame for security analysis is also referred to as a "CIA triad"[20] (CIA here stands for Confidentiality, Integrity and Availability). With regards to the development, training and deployment of machine learning, it might be useful to consider especially the traceability (of software

[18] Google: The 4 stages of machine learning: From BI to ML;
Jason Mayes: Jason's Machine Learning 101;
Yufeng Guo: The 7 Steps of Machine Learning

[19] National Institute for Standards and Technology: Glossary

[20] Chad Perrin: The CIA Triad

artifacts) and quality (of data) as they are directly relevant for the security of machine learning[21].

There are three main intersections between machine learning and information security[22]:

1. Leveraging machine learning to secure IT systems;
2. Leveraging machine learning to compromise IT systems;
3. The information security aspects of applications that leverage machine learning [the focal area of this paper];

While this work focuses on information security aspects of machine learning, it seems useful to briefly describe the other two aspects to better distinguish them from one another. Additionally, the information security of applications in those areas, whether machine learning is used to better protect IT systems or compromise them, is crucial.

2.1 Machine Learning to Secure IT Systems

A common use case for machine learning to secure IT systems is to effectively recognize spam and separate it from legitimate emails. There are also machine learning powered applications being developed that map networks[23], spot malware[24], detect anomalies in computer networks[25], suggest triage solutions to the support staff[26] or even quarantine the system, for example, by cutting of all external communication until the incident is resolved[27]. While most of those applications still work in tandem with human staff, there are examples of completely autonomous applications that aim to improve infor-

[21] Vincent Aravantinos and Frederik Diehl: Traceability of Deep Neural Networks;
Valerie Sessions and Marco Valtorta: The Effects of Data Quality on Machine Learning Algorithms

[22] Often used in a similar context but not strictly speaking at the intersection of information security and machine learning are: leveraging machine learning to spread disinformation and applying machine learning to create deep fakes. Both areas are entirely out of scope of this analysis.

[23] Michael Sulmeyer and Kathryn Dura: Beyond Killer Robots: How Artificial Intelligence Can Improve Resilience in Cyber Space

[24] Linda Musthaler: How to use deep learning AI to detect and prevent malware and APTs in real-time;
Kim Zetter: Researchers Easily Trick Cylance's AI-Based Antivirus Into Thinking Malware Is 'Goodware'

[25] Norbert Pohlmann: Künstliche Intelligenz und Cybersicherheit

[26] Erin Winick: A cyber-skills shortage means students are being recruited to fight off hackers

[27] Karen Hao: The rare form of machine learning that can spot hackers who have already broken in

mation security. One of those is Mayhem. It took part in the DEF CON Hacking Conference[28] capture the flag event[29] in 2016[30]. It competed against human hackers after winning the DARPA Cyber Grand Challenge in 2016 and was tasked not only with breaking into devices, but also with automatically spotting vulnerabilities in its own system and patching them[31]. Keeping in mind that this was far from product-ready, it is still an impressive development. However, machine learning for information security still has a long way to go, as "machine learning models are largely unable to discern between malicious input and benign anomalous data"[32], and will prove which applications will really increase information security.

2.2 Machine Learning to Compromise IT Systems

Machine learning to compromise IT systems on the other hand focuses on overcoming security measures. In general, the offensive application of machine learning "can be expected to increase the number, scale, and diversity of attacks that can be conducted at a given level of capabilities"[33]. In 2017 Thomas Dullien identified five broad areas of use cases for machine learning to compromise IT systems, concluding that a stable distribution[34] is a key requirement to apply machine learning in this area.[35] Those areas are bug detection, exploitation, phishing and user deception[36], autonomous lateral movement, password cracking as well as compromising hardware[37]. So basically, these encompass the bread and butter concepts of malicious actors

[28] DEF CON: DEF CON Hacking Conference

[29] DEF CON: CTF Archive

[30] Devin Coldeway: Carnegie Mellon's Mayhem AI takes home $2 million from DARPA's Cyber Grand Challenge

[31] Stephan De Spiegeleire, Matthijs Maas and Tim Sweijs: Artificial Intelligence and the Future of Defense

[32] Andrew Marshall, Raul Rojas, Jay Stokes and Donald Brinkman: Securing the Future of Artificial Intelligence and Machine Learning at Microsoft

[33] Miles Brundage et al.: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

[34] Meaning that the data points on which the machine learning algorithm performs predictions need to come from the same probability distribution as the data points used for training.

[35] Thomas Dullien: Machine learning, offense, and the future of automation

[36] One application of this method became known as the SNAP_R bot, see Cade Metz: How Will We Outsmart A.I. Liars?

[37] Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

and are not necessarily reserved for well-funded groups only[38]. While those attacks would likely be machine learning applications used against non machine learning software, there are even some use cases where machine learning is used to compromise machine learning applications. One use case is the creation of a machine learning model that generates adversarial data that will be misclassified by the target machine learning system with a degree of high certainty[39].

[38] See for example a project that uses freely available software (bettercap) leveraging reinforcement learning on low-cost hardware (Rasperry Pi Zero W) to compromise WiFi networks - evilsocket: pwnagotchi

[39] Generative Adversarial Networks (GANs), see:
Nicolas Papernot et al.: Practical Black-Box Attacks against Machine Learning;
Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

3. Information Security of Machine Learning: The Attack Surface

Most conventional attack vectors (insider threat, denial of service, ...) can be exploited against machine learning; and in fact, they possibly shine in a new light when examined through the machine learning lens. On that end, the attack surface might look slightly different but remains very much the same. However, it will expand through new vulnerabilities introduced by the design of machine learning[40]. Potential adversarial interference includes:

- attacks against the data used for training and decision-making,
- attacks against the classifier in the training environment,
- as well as attacks against model in the deployment environment.

Machine learning is a complex process, making the attack surface equally dynamic and multifaceted. To map the entire attack surface, it is therefore useful to go through all steps of the machine learning process and explain the different threats and attack vectors step-by-step. The resulting surface will be generic, meaning that it will cover all machine learning models, but not the entire surface will be applicable to each machine learning model. Take for example an application using online machine learning therefore learning only when it is deployed. As no data is used in the training environment to train this model before deployment, attacking the data acquisition stage and poison training data would not be an applicable attack vector. That attack vector is however applicable to other forms of machine learning such as supervised machine learning. The subsequent analysis aims at breaking down the complex attack surface of machine learning into digestible pieces to provide clues for better securing it to harvest the full potential of machine learning.

The attack surface is divided into three larger sections: the training environment, the deployment environment and the outside world. Each section includes several stages of the machine learning process where attacks can occur. Some stages occur on the intersection between the outside world and the other environments, their analysis can be found in the sections covering the training environment (for the data acquisition stage) and the deployment environment (for the online learning and output stages) respectively.

[40] Ian Goodfellow et al.: Attacking Machine Learning with Adversarial Examples; Greg Allen and Taniel Chan: Artificial Intelligence and National Security

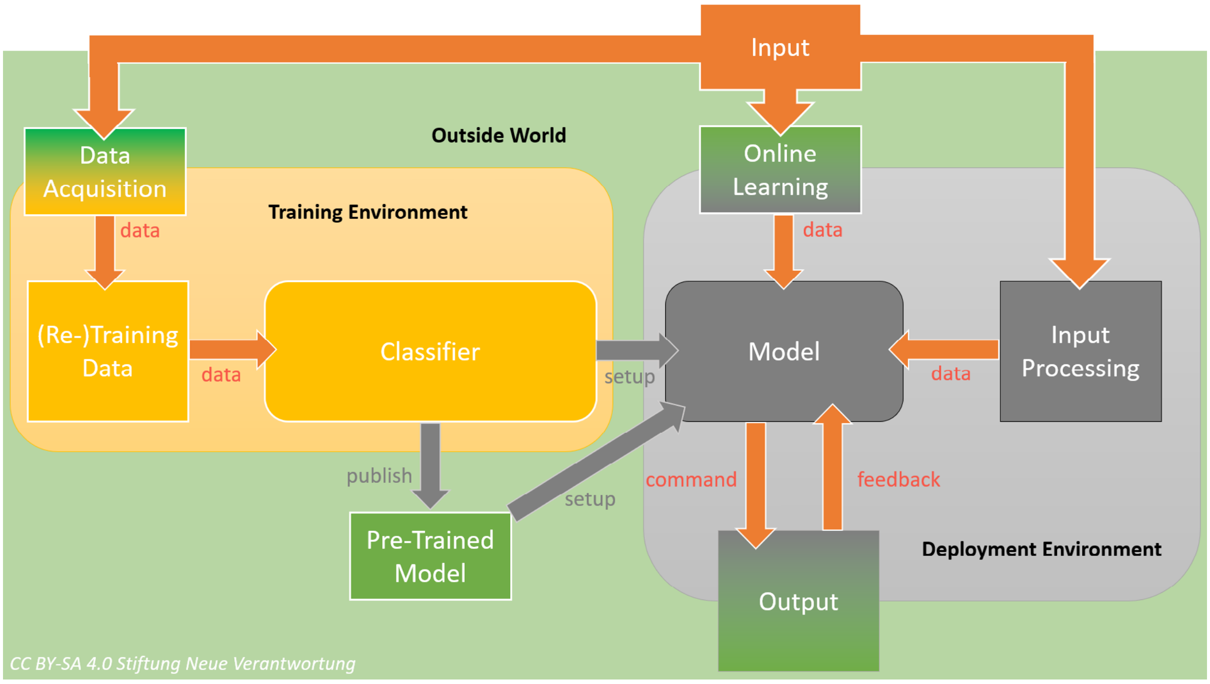

Figure 2: Attack surface of machine learning

3.1 Training Environment

The training environment is defined by the defender having legitimate control over a model's inner workings. Therefore, a cloud service used for training is still considered to be part of the training environment. Mechanisms to improve security and detection can in general be set up by the defender whereas this might vary when using third party services, such as cloud computing. The training environment might interface with the outside world when it pertains to the acquisition of the training data.

As the training environment consists of regular IT systems and networks, it can be compromised through traditional attacks. With that access, adversaries can interfere with what is unique to machine learning: the training data and the classifier. Manipulating either of those creates systemic vulnerabilities for wherever the model is being deployed. That extends beyond the initial application towards the use and open-source sharing of pre-trained models[41] (e.g. for transfer learning[42]) and therefore the transfer of vulnerabilities[43].

Data Acquisition [at the intersection with the "Outside World"]

Data for training the classifier of a machine learning model is a prime target for adversarial interference. Tainting data to nudge the training of a machine learning model in a certain direction is known as data poisoning and it can occur in the data acquisition phase. This means acquiring data used to train a machine learning model which leverages supervised or reinforcement learning for example. Data can be collected from the outside world, where it can be collected and pre-processed[44], either directly or through third parties including open supply chains (e.g. Github[45] or programming library NumPy[46]). Additionally, data can be generated within the training environment itself[47].

Attacks targeting the training data can take place in the *outside world* and within the training environment. The threat depends very much on where the future training data is produced or stored in the outside world, how it is secured and who has legitimate access to it. The attack surface is further expanded by data that is obtained from third parties (e.g. data brokers), as this extends it to the entire supply chain of this data. An adversary that can gain access to the data in any of those stages can change it, delete it, or add additional data. Manipulating the data could enable an attacker to influence the machine learning model. Consider the example in which the data is composed of pictures of criminals to train facial recognition technology. If an attacker were to delete all pictures of persons with blue eyes, the subsequently trained facial recognition software might let blue eyed criminals pass, as blue eyes would no longer be a feature that the model associates with criminality. The model would simply conclude that blue eyed people are in fact innocent.

[41] Pedro Marcelino: Transfer learning from pre-trained models

[42] Avinash: Pre-Trained Machine Learning Models vs Models Trained from Scratch

[43] Bolun Wang et al.: With Great Training Comes Great Vulnerability: Practical Attacks against Transfer Learning

[44] Charlotte Stanton et al.: What the Machine Learning Value Chain Means for Geopolitics

[45] Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

[46] Cisco: NumPy pickle Python Module Remote Code Execution Vulnerability

[47] OpenAI: OpenAI Five

Another threat is data extraction where the data is not poisoned but extracted[48]. After data extraction, the data can be exploited for malicious purposes or simply leaked. Machine learning relies on a huge amount of data. Especially in the context of national security, this might constitute a valuable trove of information that adversaries want to get their hands on. An example of this -- which did not even require actual hacking because the data was not secured at all -- was revealed in February 2019. A security researcher found the machine learning database of the Chinese SenseNets company which provides monitoring services (including facial recognition and crowd analysis technologies) to the police[49]. The data included "identification numbers, gender, nationality, address, birth dates, photographs, employers and which cameras or trackers they had passed"[50] of around 2.5 million citizens. If instead of the researcher an adversary would have found this data dump, the adversary could have either manipulated it or used it in targeted attacks .

Additionally, simply having access to the data that is soon being used for training might already provide an attacker with sufficient intelligence that would enable the attacker to derive possible weakness of the machine learning model from it (box knowledge). An adversary that has access to the statistics of a data-set can leverage vulnerabilities in the system. It could for example allow for targeted design of outliers that will fool the final model. An attacker could spot a lack of blue-eyed persons in the data collection. Therefore, it could deduce that if the data is used to train a facial recognition system to be on the lookout for criminals that being blue-eyed might prevent you from being flagged by the system.

Lastly, the integration of collected data from the outside world (e.g. from mobile apps and services or sensory data) into the training environment also bears the risk of an adversary hiding malicious software (such as a trojanized file) in the data to gain access to the training environment. Malicious code could also be injected into build systems (e.g. via Python libraries)[51] and thus compromise the training environment.

[48] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos and Dawn Song: The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

[49] The Washington Post: China has turned Xinjiang into a zone of repression — and a frightening window into the future

[50] The Washington Post: China has turned Xinjiang into a zone of repression — and a frightening window into the future

[51] Catalin Cimpanu: Ten Malicious Libraries Found on PyPI - Python Package Index

(Re-)Training Data

After the data has been acquired, it is stored inside the training environment which is either on-premise IT systems or on cloud servers. In both cases, traditional attack vectors to compromise IT systems or cloud servers apply. The difference to the last stage is that the data is now in a more controlled environment where for example the choice of security mechanisms is up to the stakeholder responsible for training the machine learning model. It does not necessarily mean that the data is more secure there, it just means that more control can be exerted over the data as compared to it being on third party IT systems.

Similar to the data acquisition stage, attacks against in the (re-)training phase include data poisoning, data extraction and intelligence. It differs in the sense that the data might have been curated and/or labeled and therefore higher quality and more structured as compared to when it was first acquired. It might also offer the attacker the opportunity to manipulate the data after it has already passed through quality checks (for bias etc.) and therefore increase the effectiveness of the manipulation.

The aforementioned attacks can occur during the initial training phase of the classifier or later on if it needs to be retrained. Retraining is necessary “if they find that the data distributions have deviated significantly from those of the original training set” (model drift)[52].

Another attack that is distinct from data poisoning is model poisoning. It can be used against models that leverage federated learning and therefore rely on several (local) agents for training. Research indicates that successfully compromising one agent might enable an attacker to compromise the entire (global) model.[53]

Classifier

The classifier is another prime target for adversarial interference. The training environment can be targeted by conventional attacks to gain access to the classifier that is being developed there. An adversary might then be able to alter the architecture or specifications (algorithm, hyperparameter, weights, features, policy, random initialization or classification thresholds) of the classifier to disrupt it or manipulate it. The latter could come in the form of a neural network backdoor that would enable an attacker to feed

[52] Luigi: The Ultimate Guide to Model Retraining

[53] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal and Seraphin Calo:Analyzing Federated Learning through an Adversarial Lens

specific input to the deployed machine learning model and thereby force a predefined output (neural network trojaning).[54] A second option would be to simply learn more about how the final machine learning application will operate. This allows the adversary to increase its box knowledge from potentially nothing (black box) to a semi (grey box) or full understanding (white box). Equipped with that knowledge, the adversary might be able to forge more effective attacks at the input level[55] or use the stolen intellectual property to compete against the original source in the market.

Implications

The difference to compromising conventional software before it is shipped is twofold: if the attack against the training data or classifier goes undetected, it will be very difficult to discover later on because it is rather challenging to spot anomalies and being able to seperate them from an output without manipulated training data and/or classifier (e.g. due to the lack of explainability). Secondly, while traditional software is known for requiring regular updates, a machine learning model might be updated (retrained) only infrequently or not at all, allowing the vulnerability to be persistent. Additionally, if poisoned training data is not detected as such and re-used, updating may perpetuate rather than patch the vulnerability.

To introduce persistent and systemic vulnerabilities to a machine learning application, an adversary with sufficient resources would likely target the training environment. If the initial attack goes undetected, the introduced vulnerability might remain hidden for a very long time. For those tasked with protecting the training environment it means that they not only have to factor this into their risk assessment but put emphasis on things such as tamper proof logs and regular log review to at least discover at some point that an attacker compromised the model and act accordingly, e.g. through retraining.

Furthermore, the training environment might hold (labeled) data which itself can, depending on the machine learning use case, be valuable for an attacker with economic or political motives. An attacker might tamper with the data, leak it (if for example it contains national security sensitive content), use it for training of its own machine learning models or sell it for financial gain. Attacks can also be aimed at intellectual property theft of the classifier.

Furthermore, there is the possibility that an attacker might want to disrupt the training of a classifier by for example by deleting the training data, tam-

[54] Yingqi Liu et al.: Trojaning Attack on Neural Networks

[55] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

per with the classifier or degrade the entire training environment (e.g. ransomware).

3.2 Deployment Environment

The deployment environment is defined by the defender having legitimate control over the model's inner workings and being able to set up mechanisms to improve security and detection. It interfaces with the outside world in many instances.

First, whatever happened in the training environment, in terms of compromise, is something that cannot be changed in the deployment environment and will be carried over into the deployment environment.

The environment is likely to be composed of and connected to a number of IT systems and networks which can be affected by a compromised model or through which the model's decision (e.g. compromised sensors that feed into input processing) or the translation of this decision into action (output) (e.g. compromised brakes) can be affected. Such systems are normally part of a larger number of redundant safety features. A complete failure is therefore rather unlikely but attacks against individual parts certainly lower the robustness and likely have a serious impact on performance characteristics.

Online (Machine) Learning/ Incremental Learning [at the intersection with the "Outside World"]
A different version of data poisoning attacks can be exploited when the machine learning model is being trained during deployment (online machine learning or incremental learning) through temporal drift and adversarial drift attacks.[56] One example of this could be a security software that leverages unsupervised machine learning to detect anomalies in the network traffic. In order to flag anomalies, it first needs to learn how the regular traffic looks like. As traffic varies immensely between networks, the training would likely take place live in the deployment environment. An attacker with access to the training data – depending on the setup either in the outside world or from in the deployment environment – can use this opportunity to inject adversarial data or manipulate existing data. That would enable the attacker to adapt the training of the model to serve the attacker's purpose (e.g. misclassifying future attacks as benign activities on the network).

[56] Myriam Abramson: Toward Adversarial Online Learning and the Science of Deceptive Machines

Input Processing

Input processing refers for example to visual sensors (computer vision) in an autonomous car[57] which translate the physical objects such as traffic signs into a digital image which can then be classified by the model.

Together with attacks against training data, attacks targeting the input processing stage of machine learning are well researched in the technical community. Input processing can be manipulated through (physical) perturbations by modifying physical objects (e.g. a pair of glasses[58] or a stop sign[59]), virtual objects (e.g. repackaging malware[60] or changing the wording of malicious emails[61]) or even crafting audiovisual input (e.g. hidden voice commands[62]) - also known as adversarial examples[63]. The attacks aim to manipulate the model, forcing it to misclassify (spoofing) or not recognize (evasion) the input[64] while being "often indistinguishable to humans"[65]. When leveraged against neural networks, this attack can be assisted by machine learning through Generative Adversarial Networks (GANs).[66]

Adversarial examples can be exploited for a number of second tier attacks, such as data exfiltration (e.g. by leveraging hidden voice commands against a digital assistant in a smartphone)[67]. The problem is exacerbated by the fact

[57] Tencent Keen Security Lab: Experimental Security Research of Tesla Autopilot;
Jörn Müller-Quade et al.: Künstliche Intelligenz und IT-Sicherheit

[58] 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

[59] Kevin Eykholt et al.: Robust Physical-World Attacks on Deep Learning Visual Classification

[60] Xiao Chen et al.: Android HIV: A Study of Repackaging Malware for Evading Machine-Learning Detection

[61] Blaine Nelson et al.: Exploiting Machine Learning to Subvert Your Spam Filter

[62] Nicholas Carlini et al.: Hidden Voice Commands;
Lea Schönherr et al.: Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding

[63] Ian Goodfellow et al.: Attacking Machine Learning with Adversarial Examples;
Christian Szegedy et al.: Intriguing properties of neural networks

[64] Andy Greenberg: Hackers Fool Tesla S's Autopilot To Hide And Spoof Obstacles

[65] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

[66] Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

[67] Gamaleldin F. Elsayed, Ian Goodfellow and Jascha Sohl-Dickstein: Adversarial Reprogramming Of Neural Networks;
Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos and Dawn Song: The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

that input processing takes place in the deployment environment which is likely connected to additional IT systems and networks, opening up additional attack avenues for the attacker.

Attacks leveraging the input processing level of the attack surface benefit from information about the specifications of the machine learning applications, since it has already been through the training (with the exception of online machine learning/ incremental learning) phase and these specifications are no longer being tweaked by an engineer. Even though black box attacks, without any information about the inner workings of the model, e.g. through the transferability of adversarial samples, are possible[68], the more information an adversary has, the easier it is to craft an effective attack. Obscurity alone does not increase information security to an acceptable level[69], but it does play a role[70].

Model

The model is the final product that is set up and run at the user level, for example an image recognition software or a malware scanning software. Before discussing the possible attacks against deployed models, it is crucial to note that these attacks come on top of all the threats from the training environment. It also does not follow that whoever is in charge of the deployment environment would necessarily have any knowledge about the security of the initial training process as the model could have been brought in from the outside world.

Possible threats through input manipulation via online learning and input processing have already been highlighted. Additionally, adversaries having access to the deployment environment where the model is setup can manipulate or disrupt the model[71], the IT systems it is running on or the networks it is connected to – for example causing a denial-of-service, forcing it to stop providing its service.

[68] Nicolas Papernot et al.: Practical Black-Box Attacks against Machine Learning;
Yanpei Liu, Xinyun Chen, Chang Liu and Dawn Song: Delving Into Transferable Adversarial Examples and Black-Box Attacks;
Deyan V. Petrov and Timothy M. Hospedales: Measuring the Transferability of Adversarial Examples

[69] Anish Athalye, Nicholas Carlini and David Wagner: Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

[70] Sandy Huang et al.: Adversarial Attacks on Neural Network Policies

[71] For example: Tencent Keen Security Lab: Experimental Security Research of Tesla Autopilot

As the model is the final product, adversaries will in most cases be able to procure a copy of it and test it for vulnerabilities, for example through reverse engineering. The knowledge about the vulnerabilities can then be used to manipulate or disrupt the model with specific input, as for example seen in the Cylance PROTECT case.[72]

Output [at the intersection with the "Outside World"]
Output refers to the last stage in this cycle, where the decision made by the model is put into action. That includes for example shutting down Internet access of a network to contain data leakage or triggering the brakes of an autonomous vehicle. The output can take place entirely in the deployment environment (e.g. anomaly detection system sending an alarm to the security operations center on the same network) or on the intersection with the outside world (e.g. triggering brakes on an automobile which subsequently interact with their physical surroundings). Access to the deployment environment and/or outside world, depending on the output, would allow an adversary to conduct various attacks such as manipulating the data before it reaches its intended target (e.g. suppressing the alarm being shown in the security operations center) or tampering with physical actors which are supposed to be triggered through the model's output (e.g. disabling the brakes). This stage can also be used to manipulate data that is sent back to the model for further processing (e.g. forcing a not working brake to report deceleration back). This resembles very much the classical information security notion of a man-in-the-middle attack.

At the output stage, an attack can also extract data without access to the deployment environment through model inversion.[73] This attack method would require white-box knowledge and/or a large number of queries as it relies on large amounts of data. Using only knowledge of the input and output (black box), attackers can extract information about the initial training data, known as membership inference[74], using statistical hypothesis testing.[75] Examples of attacks using this method include inferring private genotype information, estimating whether somebody taking a lifestyle survey admitted to cheating

[72] Skylight: Cylance, I Kill You!

[73] Matt Fredrikson, Somesh Jha and Thomas Ristenpart: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

[74] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

[75] Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov: Membership Inference Attacks Against Machine Learning Models

on a spouse, and recovering face photos from facial recognition systems.[76]

Implications

The deployment environment of a machine learning model faces many threats which are similar to the one where conventional software is deployed, though the level of automation, but also lack of explainability, and opacity more generally make a difference. Without additional safeguard mechanisms (e.g. a “human-in-the-loop” or “human-aided machine-to-machine learning”[77]), a compromised model can easily lead to a cascade of automated decisions with high impact – as declared goals for machine learning usage for example in the military domain are: autonomy, scaling and speed[78]. An additional attack vector exists where the model leverages online learning, as this data can be manipulated by an adversary to compromise the model in various ways, including data exfiltration.

3.3 Outside World

The outside world is defined by not being under the direct control of either actor that has legitimate control over the training or deployment environments.[79] The outside world includes data acquisition by third parties [see section training environment], provision of pre-trained models, input, and interfacing parts of the online/ incremental learning and output stages with the outside of the deployment environment [for both, see section deployment environment].

Input

Data is an integral part of training a machine learning model. Looking at the machine learning attack surface as a cycle, the initial starting point is the data input. The input is represented by physical objects such as stop signs[80]

[76] Ahmed Salem et al.: ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models;
Matt Fredrikson, Somesh Jha and Thomas Ristenpart: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

[77] Sarah Scoles: It’s Sentient - Meet the classified artificial brain being developed by US intelligence programs

[78] Congressional Research Service: Artificial Intelligence and National Security

[79] The analysis acknowledges that due to the global supply chain, most hardware and software that is used to train, deploy and run a machine learning model has at some point been acquired from various stakeholders outside of the controlled environments. The analysis therefore assumes that this hardware and non-machine learning specific software is not yet compromised (e.g. through backdoors).

[80] Kevin Eykholt et al.: Robust Physical-World Attacks on Deep Learning Models

or road surface markings[81] as well as virtual objects such as images or a computer game[82]. Input might come from the outside world, such as traffic signs being recognized by a sensor or third party mobile applications collecting images[83]. Input can also come directly from within the training[84] or deployment[85] environments via data collection and online learning.

Pre-Trained Model

After the training stage, a model can either be directly deployed or it can also be publicly shared with other stakeholders through "freely accessible developer platforms (like Github)"[86]. Liu et al. even predict that "in the foreseeable future, AIs (i.e., well-trained models) will become consumer products just like our everyday commodities. They are trained/produced by various companies or individuals, distributed by different vendors, consumed by end users, who may further share, retrain, or resell these models"[87]. Everyone with (legitimate or illegitimate) access to the pre-trained model cannot only download and use it but also potentially manipulate it and re-share the malicious version. Subtle changes that keep the initial function intact but add something special, such as a neural network backdoor[88], might be difficult to detect and even harder to attribute. Together with third party data acquisition, pre-trained models form a worrisome vulnerable machine learning supply chain.

A pre-trained model can also come in the shape of a cloud-based service offered by companies such as Google, BigML or Microsoft. In that case the user / customer would only access the application programming interface (API) to make use of the machine learning model. This outsources several of security concerns to the service provider but does not necessarily increase overall security of the model as it is still vulnerable e.g. to timing side channel at-

[81] Tencent Keen Security Lab: Experimental Security Research of Tesla Autopilot

[82] DeepMind: AlphaStar: Mastering the Real-Time Strategy Game StarCraft II

[83] FaceApp: FaceApp - Free Neural Face Transformation Filters

[84] OpenAI: OpenAI Five

[85] DarkTrace: Machine Learning in the Age of Cyber AI - White Paper

[86] Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

[87] Yingqi Liu et al.: Trojaning Attack on Neural Networks

[88] Yingqi Liu et al.: Trojaning Attack on Neural Networks

tacks against neural networks.[89] It also raises the question of data security and protection, especially when considered national security relevant data.

Implications
Attacks against stages that are in the outside world are much more difficult to detect and protect against as they are outside of the controlled environment where the defender can set up security mechanisms. Take for example a stop sign that is physically manipulated to avoid detection by a sensor in an autonomous vehicle or a person doing the same to avoid facial recognition.

Attacks that are less resource-intensive might take place in the outside world. Especially because they are much more difficult to detect a priori. Their efficiency can however be increased by additional attacks against either of the environments, for example learning the vulnerabilities or biases of the training data/classifier in the training environment to leverage those with specific input in the outside world.

3.4 Example Scenarios

In order to gain a better understanding of what the attack surface would look like in a possible real world scenario, it might be useful to describe two use cases (supervised learning without online learning and unsupervised learning with online learning) and go through them step-by-step according to the mapping done in the section before. These two illustrations are very simplistic and are just meant to give a better idea of the attack surface – they are in no way meant to be holistic.

Recognizing a stop sign (supervised learning without online learning)
Assuming there exists a very straightforward machine learning model which is tasked with recognizing a stop sign in the real world (outside world) and if it does, sends a command to a cyber-physical system, e.g. the brakes in a connected car. In this case, the input would be stop signs in the physical world. The training environment are the IT systems of company A that develop this machine learning model. Data acquisition would be the images of traffic signs which would either be taken by company A or bought by company A from a third party vendor. If bought from a third party vendor, the images would then be copied into the training environments IT systems. The training

[89] Dou Goodman and Tao Wei: Cloud-based Image Classification Service Is Not Robust To Simple Transformations: A Forgotten Battlefield;
Vasisht Duddu, Debasis Samanta, D. Vijay Rao and Valentina E. Balas: Stealing Neural Networks via Timing Side Channels

data, in this case images of traffic signs, would then be labelled either “stop sign” or “no stop sign” and fed into the classifier for training or retraining. The trained (or pre-trained) model will then be setup in the deployment environment, in this case a car. The car will be driving on some training grounds, encountering various traffic signs. The input processing composed of sensors will transmit images of what it sees in the real world to the model, which will assess which of those images constitutes a stop sign. When a stop sign is recognized, the output would be a signal to the brakes to engage. This might trigger a velocity sensors in the brakes, to feed back into the model that speed is decreasing so the model can continuously reevaluate and adjust braking accordingly to bring the vehicle nicely to a stop at the appropriate location.

For an attacker there are various avenues to compromise this model. Either by interfering with the images of traffic signs, the labels, the classifier itself, the stop signs and the brakes in the real world, the sensor or the model itself. An adversary has the choice to manipulate individual traffic signs in the real world, compromise a third party or company A’s IT systems to manipulate the images of the traffic signs, the features of the classifier or a pre-trained model – creating a potentially systemic vulnerability for all cars it will be used in. The adversary could also choose to go after individual cars for example by tampering with the sensors[90], including the optical sensor that “sees” the stop sign and/or the sensor that reports the velocity of the car.

Recognizing malicious network traffic (unsupervised learning with online learning)

Assume there exists a machine learning model which is tasked to identify malicious network traffic by learning how regular network traffic looks like and sending an alarm whenever something happens that varies greatly from the usual, also known as anomaly detection. The classifier would be developed by company B in its training environment. As it does not rely on pretraining, an acquisition of data would not take place in that environment – and no third party vendors would be involved. The final model would be set up in the deployment environment, company C’s network of IT systems. While being deployed, the model would learn how the regular network traffic looks like through online learning – for example 80% of all the IT systems in the network are switched on before 9am between Monday and Friday and start downloading emails from the email server and 50% of the IT systems access a popular news site between 1pm and 2pm. The model has several sensors such as for traffic between IT systems on the network, outgoing traffic from

[90] Peter Popham: Final verdict on Air France 447: sensors left pilots helpless

IT systems to the Internet and inbound traffic from the Internet to the IT systems – these sensors serve as input processing. On a Saturday at 4am, input processing recognizes 100% of the IT systems are turned on almost simultaneously and start creating massive outbound traffic to the Internet. The model recognizes this as highly irregular activity and as an output triggers the sending of text messages containing this alarm to the IT security staff.

An adversary could compromise the training environment to learn more about how this model will work in order to spot some vulnerabilities. The attacker could also slowly feed the model with malicious input, such as increasingly switching on computers on Saturdays around 4am, to compromise its online learning. Another avenue for an attack would be to take out the IT systems that is tasked with sending text messages to the IT security staff in case an alarm is triggered by the model.

4. Strategic Implications of the Attack Surface

So far, this analysis has aimed at mapping the attack surface of machine learning, incorporating potential attack vectors that adversaries can leverage to compromise it and describing what the impact could be. Soufiane defined three components of an adversarial strategy against machine learning applications as how to damage it, what phase to target and what damage is supposed to be done.[91] Looking at the attack surface through this lens leads to four preliminary findings, which should be taken into account for a strategic government perspective on machine learning deployment.

The attack surface of machine learning is vast and partially uncontrollable
We are already familiar with some of the vulnerabilities in machine learning as they (also) exist in traditional IT systems; others are new or unique to machine learning systems. Many traditional attack vectors such as targeting the development environment (here: training environment) to steal data or manipulate the software still remain part of the attack surface for machine learning. The joint 2019 report of the French and German cybersecurity agencies concluded that "these are not vulnerabilities of AI systems per se, but this shows an extension of the global attack surface when those systems are used as tools to control other information systems"[92]. It is however crucial to consider all vulnerabilities and attack vectors, be they traditional or intrinsic to machine learning.

The three key components for an adversary to interfere with a machine learning model are: 1. which details are known about the model (e. g. black box knowledge), 2. what data flows and actions can be observed (e. g. Input data and command) and 3. which parts of the attack surface can be manipulated (domain of influence).[93] In addition to that however, an adversary can target the input that is either used for training, for online learning or later on for classification. That the input is largely generated outside the controllable environments, and therefore difficult to secure against intentional interference, exacerbates this particular challenge. What broadens the attack sur-

[91] Chami Soufiane: Security threats against machine learning based systems. A serious concern

[92] Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

[93] How these pieces fit together in adversarial interference is illustrated by Behzadan and Munir's research on policy induction attacks against reinforcement learning, see Vahid Behzadan and Arslan Munir: Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks Vahid

face further is the insecurity derived from attacks against the global supply chain of software and hardware[94] in general as well as from the machine learning specific supply chain (training data acquisition from third parties, outsourcing the labeling of training data[95], widely available pre-trained models[96] and programming libraries[97]). Furthermore, it might be difficult to assess whether an output is valid, hence in line with a classifier's setup and training (which can still be objectively wrong, e.g. due to flaws introduced through biased training data), or whether it has intentionally been steered in that direction by an attacker.[98] Due to the many phases that an adversary can target, attributing an attack after it has been identified is another major challenge[99], adding complexity to the existing attribution problem[100]. The same applies for mitigation and responses once models have been deployed, as they might have no direct update channels, retraining possibilities or rely on third parties to be fixed.

Adversaries can achieve a variety of goals and potentially cause cascading effects

An attack directed against an autonomous weapon system certainly differs from interfering with surveillance cameras, a malware detection system or ad-based targeting. There are, however, common denominators among attacks against machine learning applications regardless of where they are implemented. Adversarial interference can either aim to:

- steal information – including intellectual property (e.g. by accessing the training data or the design of the classifier),

[94] Reiterating that the analysis acknowledges that due to the global supply chain, most hardware and software that is used to train, deploy and run a machine learning model has at some point been acquired from various stakeholders outside of the controlled environments. The analysis therefore assumes that this hardware and non-machine learning specific software is not yet compromised (e.g. through backdoors).

[95] Angela Chen: How Silicon Valley's successes are fueled by an underclass of 'ghost workers'

[96] Pedro Marcelino: Transfer learning from pre-trained models

[97] Reuben Binns, Peter Brown and Valeria Gallo: Known security risks exacerbated by AI

[98] Andrew Marshall, Raul Rojas, Jay Stokes and Donald Brinkman: Securing the Future of Artificial Intelligence and Machine Learning at Microsoft

[99] Miles Brundage et al.: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation;
Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

[100] Sven Herpig and Thomas Reinhold: Spotting the bear: credible attribution and Russian operations in cyberspace;
Miles Brundage et al.: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

- degrade performance – thereby potentially also eroding trust in the system (e.g. by deleting data or altering the classifier),
- disrupt the system – (temporarily) rendering it completely useless (e.g. by denial-of-service attacks),
- manipulate the system – to achieve a desired outcome such as spoofing or evasion (e.g. by poisoning data),
- or compromise the system – as a stepping stone for exploiting vulnerabilities in other applications than run in the same deployment environment (e.g. an email account on a smartphone through manipulating the digital assistant).

Access to one machine learning model within an interconnected ecosystem of machine learning devices could not only disrupt and damage the entire system but also have cascading effects. Most applications today do not run in isolation, they are part of a larger network. As the networks and level of automation of IT systems within networks grow larger and the use of machine learning enables increased autonomy, scaling and speed of decision-making, the chances for potentially catastrophic chain reactions rise as well.

Threat modeling is a key requirement to increase security of machine learning

Assessing the information security aspects of any applications requires accurate threat modeling. A threat model is "a formally defined set of assumptions about the capabilities and goals of any attacker who may wish the system to misbehave"[101]. Knowing what the defender is up against helps not only to understand the threat better but also to allocate resources to secure the application and mitigate attacks. It makes a difference whether the adversary is a multinational company that focuses on stealing intellectual property, a criminal group that wants to earn money or an intelligence unit carrying out espionage operations. However, Papernot and Goodfellow conclude that until 2016 "most machine learning has been developed with a very weak threat model, in which there is no opponent"[102]. Properly securing a machine learning model against adversarial interference depends not only on technical and financial resources but also on correctly identifying the most likely attack vectors based on its domain of application and subsequently (illegitimate) access of adversaries to the various stages and environments. Whereas a criminal might attack the deployment environment with ransomware, a multinational company or economic-minded intelligence unit would rather go after the details of the classifier in the training environment – and potentially grab the training data for good measure. A mili-

[101] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

[102] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

tary unit on the other hand might wish to evade detection by an autonomous weapon system targeting mechanism through input manipulation. Therefore, using a domain-based approach to strategically think about the security of machine learning seems prudent.

A holistic assessment might be the right space for policy research and action

The technical machine learning research focuses, for good reasons, on the new attack vectors of machine learning which mainly target the input stages (data acquisition, online learning and input processing). What should not be forgotten is that the traditional vulnerabilities surrounding information security and cybersecurity have not been properly solved so far and are still inherent to machine learning – whether it is the debate on global supply chain vulnerabilities or attribution of attacks. Additionally, machine learning exacerbates such vulnerabilities as adversaries leverage machine learning to automate attacks and scale them to new levels. Defenders might as well resort to new machine learning-enhanced information security tools to protect their IT systems and networks against traditional and new attacks. The latter two aspects of the intersection between machine learning and information security, machine learning to secure IT systems and machine learning to compromise IT systems, in this paper but require further (policy) research as well. It might therefore be a prudent approach to further work on the intersection between information security and machine learning through holistic analyses discussing possible challenges together. The way machine learning will be deployed in the future, especially in high-stakes decision-making such as the judiciary or the military, might depend on its security and the subsequent implications. It is obvious that even on the level of national artificial intelligence strategies, these outcomes need to be considered. Therefore, the information security of machine learning needs to be part of a well-informed policy debate.

5. Information Security of Machine Learning in National Security

So far, the paper has mapped the attack surface of machine learning, illustrated it with real world examples and discussed the implications for a strategic approach towards securing the machine learning evolution. The necessity for a strategic approach is not solely derived from the vulnerability of machine learning but also from its use cases and the respective domain. In order to do so, the following section takes a brief look at a high stakes machine learning infused domain – national security. It is one of the domains that likely provide the largest divergence between the assumed low level of adversarial interference when designing machine learning until very recently[103] and the real-life threat model for national security applications. From no opponent at all to potentially having the most powerful militaries in the world as adversaries is a huge difference.

There are already many areas where machine learning is applied or might be applied soon at the intersection with national security[104]. This includes areas that are supposed to increase national security (e.g. facial recognition in surveillance, riot control or crisis prediction and prevention) as well as applications in infrastructures that when successfully attacked are a potential threat to national security (e.g. process optimization in critical infrastructures) and covers the civilian as well as the military domain. For military and intelligence purposes alone, applications include for example machine learning for reconnaissance, intelligence gathering and analysis[105], (dis)information operations[106], situational awareness and decision-making[107], simulation

[103] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

[104] Michael C. Horowitz et al.: Artificial Intelligence and International Security;
Margarita Konaev and Samuel Bendett: Russian AI-Enabled Combat: Coming To A City Near You?;
Ben Scott, Stefan Heumann and Philippe Lorenz: Artificial Intelligence and Foreign Policy;
Congressional Research Service: Artificial Intelligence and National Security;
Steven Feldstein: The Global Expansion of AI Surveillance

[105] Sarah Scoles: It's Sentient - Meet the classified artificial brain being developed by US intelligence programs;
United States Department of Defense: Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven);
Michael C. Horowitz et al.: Artificial Intelligence and International Security

[106] Miles Brundage et al.: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

[107] United States of America - Department of Defense: Summary of the 2018 Department of Defense Artificial Intelligence Strategy;
Elsa B. Kania: Battlefield Singularity: Artificial Intelligence, Military Revolution, and China's Future Military Power

and training[108], military logistics[109] or commandeering of unmanned military vehicles, semi-/autonomous and lethal autonomous weapon systems[110] and their countermeasures[111] or automated offensive and defensive cyber operations[112] and machine learning countermeasures[113].

Cutting-edge machine learning research and data collection is taking place in the private sector and academia and mainly in the United States and China[114], which creates a huge challenge for national security -- especially for all other countries. For machine learning used for military purposes it means that the “outside world” part of the attack surface becomes even more crucial. At the same time, this means that the attack surface to (indirectly) interfere with machine learning-powered military assets is possibly vast. Defending it thoroughly therefore might create an immense challenge. This risk is not only exacerbated by the global supply chain for hardware and software in general, but also when it comes to machine learning more specifically – where every other state might either depend on the US or China for powering their militaries with machine learning.

As shown, there are a number of goals that an adversary could pursue when interfering with national security applications that leverage machine learning. While all of them – disruption, degradation, manipulation, compromise and theft – are highly relevant in the national security domain, the atter deserves special attention when compared to other domains. Looking at the attack surface, it becomes clear that data plays a crucial role in the information security

[108] Elsa B. Kania: Battlefield Singularity: Artificial Intelligence, Military Revolution, and China’s Future Military Power

[109] Michael Shoebridge: AI and national security: lethal robots or better logistics?

[110] Ben Scott, Stefan Heumann and Philippe Lorenz: Artificial Intelligence and Foreign Policy;
Samuel Gibbs: Google’s AI is being used by US military drone programme;
Elsa B. Kania: Battlefield Singularity: Artificial Intelligence, Military Revolution, and China’s Future Military Power;
Congressional Research Service: Artificial Intelligence and National Security;
Congressional Research Service: U.S. Ground Forces Robotics and Autonomous Systems (RAS) and Artificial Intelligence (AI): Considerations for Congress

[111] Michael Shoebridge: AI and national security: lethal robots or better logistics?

[112] Catherine Clifford: How billion-dollar start-up Darktrace is fighting cybercrime with A.I.;
Greg Allen and Taniel Chan: Artificial Intelligence and National Security;
Elsa B. Kania: Battlefield Singularity: Artificial Intelligence, Military Revolution, and China’s Future Military Power

[113] Greg Allen and Taniel Chan: Artificial Intelligence and National Security;
Elsa B. Kania: Battlefield Singularity: Artificial Intelligence, Military Revolution, and China’s Future Military Power

[114] Philippe Lorenz and Kate Saslow: Demystifying AI & AI Companies

of machine learning. The significance of that data is amplified when the data itself is directly or indirectly crucial for national security, such as intelligence reports or maneuver tactic recordings feeding machine learning training. While from a privacy perspective, that is true for all data that is used on machine learning, the damaging potential of national security relevant data in the wrong hands might be more severe.

Looking at the national security domain, especially the military part, decision-making means that lives are at stake. Human decision-making might not be completely understood today, but with machine learning powering military assets decisions might even be less understood in the future. Understanding these decisions is important, especially if there is no human-in-the-loop or human-on-the loop[115] and collateral damage may occur. Not only the decision of the model itself might be under increased scrutiny where machine learning meets national security and adversarial interference. If a human follows through with a decision based on an analysis provided by machine learning, how much transparency about this analysis is needed and where will this require unconditional trust that the analysis is correct and was not interfered with?

When it comes to national security decisions, trust does not appear to be a prudent way forward, especially in an environment where detection and attribution of interference become increasingly difficult. Machine learning that is used to power military assets will not only require in-depth testing of its functionality and accuracy (e.g. to avoid false-positives) but, especially due to the adversarial environment (e.g. on foreign soil) it is deployed in, it will also require vigorous penetration testing and fail-safes to make sure it is as protected as possible from internal and external interference. The same holds true for national security applications which are not deployed in adversarial environment per se but make a prime target due to the high stakes decisions they produce, such as border control, video surveillance or the criminal justice system. As discussed earlier, applications are not isolated but exist within an interconnected network of systems. Therefore, a compromised system might affect more than just this one system, potentially leading to cascading effects. Rapidly unfolding, cascading effects triggered by an adversary is likely one of the last things that anyone wants to happen to its own military or law enforcement systems.

[115] International Committee of the Red Cross: Autonomy, artificial intelligence and robotics: Technical aspects of human control;
Acknowledging that there are possible edge cases, e.g. due to the potentially immense speed of hypersonic missiles, when this might not be feasible to implement it.

All these are not arguments to avoid using machine learning in national security altogether, it just means that every application has to be understood[116] as well as intensely vetted and secured before integrating machine learning components in this domain – trust where needed but verify where possible, and make sure that there is always a finger near the button.

In August 2019, War on the Rocks featured an appeal for more and better "Artificial Intelligence Research" funding for national defense. It stated that "cybersecurity concerns may affect the desired level of automation for various tasks"[117]. The response to that can only be: *It should*.

[116] Congressional Research Service: Artificial Intelligence and National Security

[117] Eric Lofgren: A Guide Not Killing or Mutilating Artificial Intelligence Research

6. Conclusion

Coming back to the initial question, “is information security a conditio sine qua non for machine learning reaching its full potential or not?” Looking at digitalisation more broadly, it seems to be quite successful despite the current state of affairs in information security being dire. Whether that is any indicator for machine learning or not is difficult to predict. What becomes clear however are two different things. First, integrating machine learning in any application broadens the attack surface, including fundamentally new attack vectors that are not very well understood yet and therefore make it more difficult to defend against. Second, the impact of successful adversarial interference against machine learning can be grave, especially in the national security domain.

Even though it is difficult to predict whether information security will become a precondition for the successful development of machine learning going forward, securing machine learning, especially when it comes to high-stakes applications such as national security, is indispensable. In order to develop concrete recommendations for policymakers, further strategic consideration should be given to the following areas: machine learning specific penetration testing capabilities, data validation methods, domain-specific information security standards for training and deployment environments (e.g. secure multi-party computation[118], federated learning[119] or differential privacy[120]), classified training data protection guidelines[121], built-in forensic capabilities[122] and robustness[123], decision integrity[124], explainable Artificial Intelligence/ interpretability[125], the risks and opportunities of a human-in-

[118] David W. Archer et al.: From Keys to Databases – Real-World Applications of Secure Multi-Party Computation

[119] Brendan McMahan and Daniel Ramage: Federated Learning: Collaborative Machine Learning without Centralized Training Data

[120] Cynthia Dwork: Differential Privacy

[121] Congressional Research Service: Artificial Intelligence and National Security

[122] Andrew Marshall, Raul Rojas, Jay Stokes and Donald Brinkman: Securing the Future of Artificial Intelligence and Machine Learning at Microsoft

[123] Rob Matheson: How to tell whether machine-learning systems are robust enough for the real world

[124] Andrew Marshall, Raul Rojas, Jay Stokes and Donald Brinkman: Securing the Future of Artificial Intelligence and Machine Learning at Microsoft

[125] Andrew Gordon Wilson et al.: Interpretable ML Symposium;
Matt Turek: Explainable Artificial Intelligence (XAI);
Jörn Müller-Quade et al.: Künstliche Intelligenz und IT-Sicherheit;
Congressional Research Service: Artificial Intelligence and National Security

the-loop[126] (including staff training to avoid automation bias and further understanding of system limitations), redundant algorithms[127], counter machine learning techniques[128][129], responsible vulnerability disclosure specific to machine learning[130], recruiting domain-specific IT security staff[131], securing the machine learning supply chain[132] and risks of online learning systems[133]. For these areas it helps to look at approaches that already exist and deal with traditional information security challenges as they might be applicable to the machine learning domain as well.

Additionally, machine learning intersects at more than only one avenue with information security. Machine learning can also be leveraged to increase cybersecurity or conduct cyber attacks. These are two fields that, by all accounts, require further research and policymaker attention themselves. All of this should however not been done in an isolated manner. It requires an interdisciplinary national approach with integrated international cooperation.[134]

In conclusion, the attack surface of machine learning is vast and partially uncontrollable, though, adversaries can achieve a variety of goals and potentially cause cascading effects, threat modeling is a key requirement to increase security of machine learning and a holistic assessment might be the

[126] Michael Shoebridge: AI and national security: lethal robots or better logistics?

[127] Martin Giles: AI for cybersecurity is a hot new thing—and a dangerous gamble

[128] Examples of this can be found in:
Kathrin Grosse et al.: Adversarial Perturbations Against Deep Neural Networks for Malware Classification;
Kathrin Grosse et al.: On the (Statistical) Detection of Adversarial Examples;
Ian Goodfellow et al.: Attacking Machine Learning with Adversarial Examples

[129] For the interaction between attacks and defenses see:
Octavian Suciu et al.: When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks;
Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman: A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance

[130] Miles Brundage et al.: The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation

[131] Congressional Research Service: Artificial Intelligence and National Security

[132] Yingqi Liu et al.: Trojaning Attack on Neural Networks;
Bundesamt für Sicherheit in der Informationstechnik and Agence Nationale de la Sécurité des Systèmes d'Information (ANSSI): Second Edition of the Franco-German Common Situational Picture

[133] Myriam Abramson: Toward Adversarial Online Learning and the Science of Deceptive Machines

[134] United States of America - Department of Defense: Summary of the 2018 Department of Defense Artificial Intelligence Strategy

right space for policy research and action. Yet even with traditional, non-machine-learning enhanced applications, states did not seem to have information security under control, as numerous data breaches and events such as WannaCry[135], the Office of Personnel Management breach[136] or intrusion into the South Korean military network[137] suggest. So, for the sake of national security, especially for those applications directly relevant to it, states have to get legislative and executive actions right to mitigate potential threats that come with (insecure) machine learning applications. That includes but is not limited to investments in research, curating public-private partnerships, forging international cooperation and developing operational and legal frameworks.

Following this conclusion, the Transatlantic Cyber Forum's working group on machine learning and information security aims to continue its work on securing machine learning to deliver concrete policy recommendations for these fields.

[135] Josh Fruhlinger: What is WannaCry ransomware, how does it infect, and who was responsible?

[136] Krebs on Security: Congressional Report Slams OPM on Data Breach

[137] Choe Sang-Hun: North Korean Hackers Stole U.S.-South Korean Military Plans, Lawmaker Says

Glossary

Adversarial Drift: "[S]ignature-based approaches do not distinguish between uncommon patterns and noise. Adversaries can take advantage of this inherent blind spot to avoid detection (mimicry). Adversarial label noise is the intentional switching of classification labels leading to deterministic noise, error that the model cannot capture due to its generalization bias."[138]

Adversarial Examples: "[I]nputs formed by applying small but intentional-lyworst-case perturbations to examples from the dataset, such that the perturbed in-put results in the model outputting an incorrect answer with high confidence."[139]

Artificial Intelligence: Traditionally refers to the process of teaching machines to recreate cognitive thought processes, which were previously only done by humans. It is important here to distinguish between symbolic and non-symbolic artificial intelligence (AI). Symbolic AI (or rules-based) is when programmers handcraft a large set of explicit rules to be hard-coded into the machine. This proved very effective for logic-based, well-defined problems. Non-symbolic AI, sometimes also referred to as connectionist approaches, conversely does not rely on the hard-coding of explicit rules. Instead, machines are able to ingest a large amount of training data and automatically extract patterns or other meaningful information, which the machine can then use to learn and make accurate predictions when fed with new data. Non-symbolic AI includes a broad set of methodologies broadly referred to as machine learning.

Box Knowledge: Refers to the level of knowledge an adversary has about the system it wants to attack.

black box: An adversary has no information about the model it wants to attack apart from the input fed into the system and the output.

gray box: An adversary has partial knowledge about the model it wants to attack.

white box: An adversary has full knowledge of the inner workings of the model it wants to attack.

[138] Myriam Abramson: Toward Adversarial Online Learning and the Science of Deceptive Machines

[139] Ian Goodfellow, Jonathon Shlens and Christian Szegedy: Explaining And Harnessing Adversarial Examples

CIA (Triad): Stands for confidentiality, integrity and availability, a common framework to assess information security.[140]

Classifier: A classifier is an algorithm that maps input data (for example pictures of animals) into specific categories (for example “dog” and “not a dog”).[141]

Cybersecurity: Extends information security beyond the purely technical definition (see “CIA”) to include broader political, legal, cultural and sociological components to further overall security. Also sometimes used as a euphemism for describing the governmental use of tools to overcome information security mechanisms (e.g. weakening encryption to enable lawful access).

Data Extraction: Unauthorized copying of data (for example training data) from a (compromised) system.

Data Poisoning: Interfering “[...] with the integrity of the training process by making modifications to existing training data or inserting additional data in the existing training set [...] pertub[ing] training points in a way that increases the prediction error of the machine learning when it is used in production”.[142]

Domain of Influence: Parts of the attack surface that an attacker has access to and can therefore manipulate.

Evasion: Interfering with a machine learning model in a way that it does not recognize the input.

Federated Learning: “Federated learning distributes model training among a multitude of agents, who, guided by privacy concerns, perform training using their local data but share only model parameter updates, for iterative aggregation at the server to train an overall global model. [...] The training of a neural network model is distributed between multiple agents. In each round, a random subset of agents, with local data and computational resources, is selected for training. The selected agents perform model training and share only the parameter updates with a centralized parameter server, that facilitates aggregation of the updates. Motivated by privacy concerns, the server is designed to have no visibility into an agents’ local data and training pro-

[140] Chad Perrin: The CIA Triad

[141] Sidath Asiri: Machine Learning Classifiers

[142] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

cess”.[143]

Generative Adversarial Network (GAN): A class of machine learning that enables the generation of fairly realistic synthetic images by forcing the generated images to be statistically almost indistinguishable from real ones.[144]

Information Security: “The protection of information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability”.[145]

Machine Learning: Machine learning consists of building statistical models that make predictions from data. Given a sufficient quantity of examples from a data source with a property of interest, a machine learning algorithm makes a prediction about that property when given a new, unseen example. This can happen via internal parameters calibrated on the known examples, or via other methods. Machine learning approaches include curiosity learning, decision trees, deep learning, logistic regression, random forests, reinforcement learning, supervised learning and unsupervised learning.

Machine Learning Approaches:

Curiosity Learning: Curiosity learning is a strategy of Deep Reinforcement Learning in which the idea is to build an intrinsic reward function (intrinsic as in generated by the autonomous agent), which means that the agent will be a self-learner because the agent will be both the student and the feedback master.[146]

Decision Trees: A decision tree in machine learning is a predictive model that is constructed by an algorithmic approach to identify ways to divide and classify a dataset based on different conditions.[147]

Deep Learning: Deep learning is a type of machine learning model that involves feeding the training data through a network of artificial neurons to pull out distributional features or higher-level abstractions respectively from the data. This is a loose approximation for sensory cortex computation in the brain, and as such has seen the most success in applications that involve processing image and audio data. Successful applications in-

[143] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal and Seraphin Calo:Analyzing Federated Learning through an Adversarial Lens

[144] Goodfellow et al.: Generative Adversarial Networks

[145] National Institute for Standards and Technology: Glossary

[146] Thomas Simonini: Curiosity-Driven Learning made easy Part 1

[147] Prince Yadav: Decision Tree in Machine Learning

clude object recognition in pictures or video and speech recognition.

Logistic Regression: Also called “logit” for short, logistic regression is a classification algorithm (not a regression algorithm like its name may suggest) that can be used for both binary and multivariate classification tasks.[148]

Random Forests: Random Forests are an ensemble method of machine learning which can be used to build predictive models for either classification or regression problems. The model creates a forest of random uncorrelated decision trees to reach the best answer.[149]

Reinforcement Learning: Reinforcement learning is a model that involves creating a system of rewards within an artificial environment to teach an artificial agent learn how to move through different states. It is commonly used in robotics for navigation and as a tool for solving complex strategy games.

Supervised Learning: As of 2018, supervised learning was the most common form of machine learning, in which a machine learns to map input data to known targets, given a set of examples, which are often annotated by humans.

Unsupervised Learning: Unsupervised learning consists of finding meaningful transformations of the input data without the help of any targets. This can be used for data visualization, data compression or denoising. Unsupervised learning is the “bread and butter of data analytics”[150] and is often a necessary first step to understanding a dataset before attempting to carry out a supervised learning task.

Membership Inference: Attacking a deployed model, using specially crafted adversarial examples to infer whether certain training points were used for training a model.[151]

(Machine Learning) Model: Trained weights/parameters from any training process.

[148] Francois Chollet: Deep Learning with Python

[149] Raul Eulogio: Introduction to Random Forests

[150] Francois Chollet: Deep Learning with Python

[151] Nicolas Papernot and Ian Goodfellow: Breaking things is easy; Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov: Membership Inference Attacks Against Machine Learning Models

Model Drift: “Rather than deploying a model once and moving on to another project, machine learning practitioners need to retrain their models if they find that the data distributions have deviated significantly from those of the original training set. This concept, known as model drift, can be mitigated but involves additional overhead in the forms of monitoring infrastructure, oversight, and process”.[152]

Model Extraction: Interfering with a model to “search for a substitute model with similar functionality as the target neural architecture”[153] in order to be able to replicate it.

Model Inversion: Interfering with a model to derive/extract the training data from it.[154]

Model Poisoning: “Model poisoning is carried out [within the setting of federated learning] by an adversary controlling a small number of malicious agents (usually 1) with the aim of causing the global model to misclassify a set of chosen inputs with high confidence”.[155]

Neural Network: A neural network (NN) is a an architecture that enables many contemporary ML applications. NNs are loosely based on the biological concept, as their models work by passing data through the network and transforming data representations from one layer to the next.[156]

Neural Network Trojaning: Manipulating a Neural Network in a way, that a trigger input causes a predefined action chosen by the adversary.[157]

Online (Machine) Learning/ Incremental Learning: A machine learning model that while being deployed “can learn from new examples in something close to real time”[158], by using the input stream of examples as training data. It “can add additional capabilities to an existing model without the original training data. It uses the original model as the starting point and directly

[152] Luigi: The Ultimate Guide to Model Retraining

[153] Vasisht Duddu, Debasis Samanta, D. Vijay Rao and Valentina E. Balas: Stealing Neural Networks via Timing Side Channels

[154] Matt Fredrikson, Somesh Jha and Thomas Ristenpart: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

[155] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal and Seraphin Calo:Analyzing Federated Learning through an Adversarial Lens

[156] Philippe Lorenz and Kate Saslow: Demystifying AI & AI Companies

[157] Yingqi Liu et al.: Trojaning Attack on Neural Networks

[158] Max Pagels: What is online machine learning?

trains on the new data".[159]

Perturbation: Small, hardly (or non) recognizable changes of an input that causes prediction errors (e.g. overlay on an image that cause a panda to be recognized as a gibbon)[160].

Physical Perturbation: Perturbation of physical objects (e.g. sticker on a stop sign)[161].

Spoofing: Interfering with a model, forcing it to misclassify the input.

Temporal Drift: "[B]ehavior changes over time requiring re- training of the model. Adversaries can take advantage of this adaptability by injecting poisonous examples mas- querading as real (camouflage). Since there is no clear separation between training and testing in online learning algorithms, rather testing become training (given bandit feedback), an adversarial scenario occurs where the next label in the sequence is different than the one predicted."[162]

Threat Model: "a formally defined set of assumptions about the capabilities and goals of any attacker who may wish the system to misbehave".[163]

Timing Side Channel: "From the total execution time [of an input], an adversary can infer the total number of layers (depth) of the Neural Network using a regressor trained on the data containing the variation of execution time with Neural Network depth. This additional side channel information obtained, namely the depth of the network, reduces the search space for finding the substitute model with functionality close to the target model"[164] and therefore achieving a model extraction.

Training Data: Refers to the sample of data used to fit the model. The model sees and learns from this dataset.

[159] Yingqi Liu et al.: Trojaning Attack on Neural Networks

[160] Ian Goodfellow et al.: Attacking Machine Learning with Adversarial Examples

[161] Kevin Eykholt et al.: Robust Physical-World Attacks on Deep Learning Visual Classification

[162] Myriam Abramson: Toward Adversarial Online Learning and the Science of Deceptive Machines partially referencing Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar: Foundations of Machine Learning

[163] Nicolas Papernot and Ian Goodfellow: Breaking things is easy

[164] Vasisht Duddu, Debasis Samanta, D. Vijay Rao and Valentina E. Balas: Stealing Neural Networks via Timing Side Channels

Transferability of Adversarial Examples: “The property of an adversarial example created by one system with known architecture and parameters, to transfer to another unknown black-box system, is called transferability.”[165]

Transfer Learning: Transfer Learning is a machine learning method “where a model developed for a task is reused as the starting point for a model on a second task”.[166]

[165] Deyan V. Petrov and Timothy M. Hospedales: Measuring the Transferability of Adversarial Examples

[166] Jason Brownlee: A Gentle Introduction to Transfer Learning for Deep Learning

About the Stiftung Neue Verantwortung

The Stiftung Neue Verantwortung (SNV) is an independent think tank that develops concrete ideas as to how German politics can shape technological change in society, the economy and the state. In order to guarantee the independence of its work, the organisation adopted a concept of mixed funding sources that include foundations, public funds and businesses.

Issues of digital infrastructure, the changing pattern of employment, IT security or internet surveillance now affect key areas of economic and social policy, domestic security or the protection of the fundamental rights of individuals. The experts of the SNV formulate analyses, develop policy proposals and organise conferences that address these issues and further subject areas.

About the Transatlantic Cyber Forum (TCF)

The Transatlantic Cyber Forum (TCF) has been established by the Berlin based think tank Stiftung Neue Verantwortung (SNV).

The Transatlantic Cyber Forum is a network of cyber security experts and practitioners from civil society, academia and private sector. It was made possible with the financial support from the Robert Bosch Stiftung and the William and Flora Hewlett Foundation.

About the Author

Dr. Sven Herpig is head of for international cyber security policy at Stiftung Neue Verantwortung. This includes the transatlantic expert network Transatlantic Cyber Forum (TCF), the EU Cyber Direct (EUCD) project funded by the European Commission as well as an ongoing analysis of German cyber security policies.

Contact the Author

Dr. Sven Herpig
Project Director Transatlantic Cyber Forum
sherpig@stiftung-nv.de
Twitter: @z_edian
+49 (0)30 81 45 03 78 91

Imprint

Stiftung Neue Verantwortung e. V.

Beisheim Center
Berliner Freiheit 2
10785 Berlin

T: +49 (0) 30 81 45 03 78 80
F: +49 (0) 30 81 45 03 78 97

www.stiftung-nv.de
info@stiftung-nv.de

Design:
Make Studio
www.make-studio.net

Layout:
Johanna Famulok