October 2020 · Dr. Sven Herpig

Understanding the Security Implications of the Machine-Learning Supply Chain

Securing Artificial Intelligence – Part 2

An analysis supported by the Transatlantic Cyber Forum

Stiftung Neue Verantwortung

Think Tank at the Intersection of Technology and Society

Executive Summary

The hopes and expectations connected to artificial intelligence are staggering. All major powers have started investing heavily in the research and development of artificial intelligence – especially machine learning. This progress may be driven by a goal that has been described in – an oversimplified but clear way – by Vladimir Putin. He has famously been quoted as saying that the nation that leads in artificial intelligence "will be the ruler of the world"[1]. Countries such as the United States[2] and China[3], and especially their respective private sectors[4], seem to have the upper hand in research and application right now. However, a vast number of affected sectors[5] and possible specializations – such as securing artificial intelligence – enable a number of states and non-state-actors to meaningfully engage in this domain.

Unfortunately, drivers of technological developments frequently follow the "move fast and break things" mentality, sometimes resulting in destabilizing effects for the entire Internet ecosystem.[6] Governments and companies must not repeat a grave mistake of the past: having security only as an afterthought. **In order to create an enabling environment for the development and deployment of artificial intelligence, security considerations must urgently be addressed across the entire machine-learning supply chain.**

Applications leveraging artificial intelligence will be highly integrated into the cyber domain[7] and will likely experience adverse effects accordingly. These include but are not limited to geopolitical cyber operations[8], illegal transfer of intellectual property[9], national surveillance apparatuses[10], financial theft[11], and cybercrime[12]. Every new technology attracts adversaries who

1 James Vincent (2017): Putin says the nation that leads in AI 'will be the ruler of the world'

2 U.S. Congressional Research Service (2019): Artificial Intelligence and National Security

3 Ashwin Acharya and Zachary Arnold (2019): Chinese Public AI R&D Spending: Provisional Findings

4 e.g., Michael Dahm (2020): Chinese Debates on the Military Utility of Artificial Intelligence

5 Techjury.net (2019): Infographic: How AI is Being Deployed Across Industries

6 Cloudflare (nondated): What is the Mirai Botnet?

7 The sum of all devices and data connectable or connected to the Internet.

8 e.g., Booz Allen (2020): The Logic Behind Russian Military Cyber Operations or Ryan Gallagher (2018): How U.K. Spies Hacked a European Ally and Got Away With It

9 e.g., Dennis C. Blair and Keith Alexander (2017): China's Intellectual Property Theft Must Stop

10 e.g., Elias Groll (2018): The Kingdom's Hackers and Bots

11 e.g., Edith M. Lederer (2019): UN report: North Korea cyber experts raised up to $2 billion

12 e.g., U. S. Federal Bureau of Investigations (2020): 2019 Internet Crime Report or Bundeskriminalamt (2020): Bundeslagebild Cybercrime 2019

will exploit it for their own gain, be it financially, politically, or otherwise motivated. Thus, there will be a number of capable and willing threat actors out there who want to meddle with systems powered by artificial intelligence.

Therefore, it is crucial to understand the supply chain and secure it against adversarial interference. The paper recommends decision-makers implement the following to achieve this goal:


Design a security approach rooted in conventional information security


Increase transparency, traceability, validation, and verification


Identify, adopt, and apply best practices

Require fail-safes and resiliency measures


Create a machine-learning security ecosystem

Set up a permanent platform for threat exchange

Develop a compliance-criteria catalog for service providers

Foster machine-learning literacy across the board

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 virtual workshop.

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:

– Charles-Pierre Astolfi, French Digital Council (CNNum)
– Manuel Atug, HiSolutions AG
– Rachel Azafrani, Microsoft
– Leonie Beining, Stiftung Neue Verantwortung (SNV)
– Daniel Castro, Center for Data Innovation
– Betsy Cooper, Aspen Tech Policy Hub
– Lars Fischer, Institute Media & Systems Engineering (IMSE), Bremerhaven University of Applied Sciences
– Kenneth Geers, Very Good Security
– Mary Hanley, University of Chicago Harris Cyber Policy Initiative (CPI)
– Maximilian Heinemeyer, Darktrace
– Ariel Herbert-Voss, OpenAI
– Wyatt Hoffman, Center for Security and Emerging Technology (CSET), Georgetown University
– Michael Hsieh, Center for International Security and Cooperation, Stanford University
– Sven Jacob, German Federal Office for Information Security (BSI)
– Frederike Kaltheuner, Mozilla Tech Policy Fellow
– Jan-Peter Kleinhans, Stiftung Neue Verantwortung (SNV)
– Thomas Lawson, AXA
– Igor Mikolic-Torreira, Center for Security and Emerging Technology (CSET), Georgetown University
– Johannes Otterbach, OpenAI
– Jörg Pohle, Alexander von Humboldt Institute for Internet and Society (HIIG)
– Johanna Polle, Institute for Peace Research and Security Policy, University of Hamburg (IFSH)
– Thomas Reinhold, Science and Technology for Peace and Security (PEASEC), TU Darmstadt
– Christine Runnegar, Internet Society (ISOC)
– Kate Saslow, Stiftung Neue Verantwortung (SNV)
– Eric MSP Veith, OFFIS – Institute for Information Technology

Table of Contents

1. Introduction

Governments and industries across the globe drive the development and deployment of applications leveraging machine learning. Use and business cases where the technology can do its "magic" are developed on a daily basis. Economic and national security aspects are natural priorities and drivers in those debates. And while many voices push for the crucial consideration of ethical aspects[13], the debate about how these systems can be better protected against adversaries takes place almost entirely inside the technical research and standardization communities.[14] Decision-makers do not appear to pay enough attention to and drive policies that increase security of machine-learning applications. In May 2020, an analysis stated that "25 out of the 28 organizations indicated that they don't have the right tools in place to secure their [machine-learning] systems".[15] Even the best economic models and ethical considerations may be nullified when systems are not secured against adversarial interference.

Information security is vital for the effective deployment of technologies and will increasingly be a crucial element for societies as a whole. The security of areas such as personalized healthcare, human resource decisions, automated translations of official documents or, of course, autonomous driving, will affect each person at an individual level. Therefore, any nation that wants to securely develop and deploy machine-learning systems should learn how to secure them – at least to the degree required to make residual risk acceptable. This especially applies, but is not limited, to nations that consider deploying machine learning in high-risk environments[16]. Beyond reducing risks, securing machine learning may even become an economic opportunity itself.[17]

There has been a decades-long debate on how to best secure IT systems, about which, though the state of the art is a moving target, there is consensus, depending on individual threat models and risk assessments. Se-

13 e.g., Independent High-Level Expert Group On Artificial Intelligence Set Up By The European Commission (2019): Ethics Guidelines For Trustworthy AI and AlgorithmWatch (2020): AI Ethics Guidelines Global Inventory

14 Sven Herpig (2020): No Safety without Cybersecure AI

15 Ram Shankar Siva Kumar, Magnus Nystrom, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneruk, Matt Swann and Sharon Xia (2020): Adversarial Machine Learning – Industry Perspectives

16 Sven Herpig (2020): No Safety without Cybersecure AI, European Commission (2020): On Artificial Intelligence – A European approach to excellence and trust and Sven Herpig (2019): Securing Artificial Intelligence

17 European Commission (2020): On Artificial Intelligence – A European approach to excellence and trust

curity mechanisms that have worked for "conventional" IT systems, such as encryption, backups, or threat monitoring controls, will still be effective in environments where machine learning is deployed. However, the development and deployment of machine learning require adjustments as it leads to additional security challenges. While the previous paper in the series explored novel attack techniques targeting machine learning systems[18], this paper presents security challenges derived from the machine learning supply chain – such as those stemming from the data pipeline and vast reliance on third parties – and suggests policy recommendations taking into account the findings of both papers.

A glossary on "cybersecurity and artificial intelligence" can be found at the end of this publication.

18 Sven Herpig (2019): Securing Artificial Intelligence

2. Machine Learning and Information Security

While machine learning existed for quite some time, the increased availability of massive amounts of data and improvements in computing power[19] has enabled an environment for crucial advancements in the past few years. Even though they are often conflated terms, machine learning is a subfield[20] and today's most important foundational basis[21] of artificial intelligence[22], which existed since the 1950s. Machine learning consists of building statistical models that make predictions from data, without hard-coded rules, and have the capacity to improve their performance over time with more exposure to data. Given a sufficient quantity of examples from a data source, known as training data, and a property of interest, a machine learning algorithm makes a prediction about that property when given a new, unseen example. This can happen via either calibrating internal parameters on the known examples or through other methods. Machine learning can be roughly divided into supervised learning, unsupervised learning and reinforcement learning, leveraging various techniques.[23]

There are three main intersections between machine learning and information security[24]:

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 latter one is the focal area of this paper: how to secure machine learning against unintentional failure, defined as "where the failure is caused by an active adversary attempting to subvert the system to attain her goals"[25].

19 Vishal Maini and Samer Sabri (2017): Machine Learning for Humans

20 Vishal Maini and Samer Sabri (2017): Machine Learning for Humans

21 The MITRE Corporation (2017): Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD

22 For a brief history of artificial intelligence, see:
Stephan De Spiegeleire, Matthijs Maas and Tim Sweijs (2017): Artificial Intelligence and the Future of Defense
The MITRE Corporation (2017): Perspectives on Research in Artificial Intelligence and Artificial General Intelligence Relevant to DoD
Vishal Maini and Samer Sabri (2017): Machine Learning for Humans

23 See glossary for definitions and techniques.

24 Sven Herpig (2019): Securing Artificial Intelligence
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.

25 Ram Shankar Siva Kumar, David O'Brien, Kendra Albert, Salome Viljoen and Jeffrey Snover (2019): Failure Modes in Machine Learning

3. Defining the Machine-Learning Supply Chain

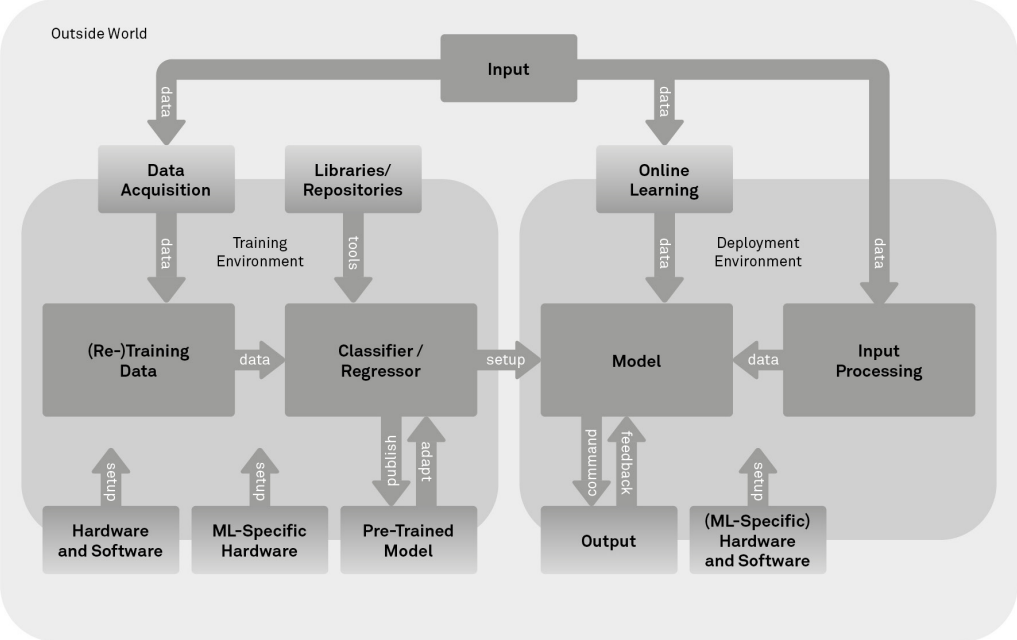
A central challenge for the security of machine learning is its supply chain. For the purpose of this paper, the "machine-learning supply chain" is defined here as: **data, tools and services as well as (specialized) software and hardware required to develop a machine-learning model.**[26] The machine-learning supply chain ranges from data generation and acquisition to the deploy-ready machine learning models. The machine-learning supply chain does not necessarily end in a finalized product. Machine-learning systems that leverage online learning for example ingest data while they are live – therefore extending the machine-learning supply chain further. Due to the variety of machine-learning models, each application may have different aspects to its supply chain, for example if the model is developed from scratch and not based on a pre-trained model. Therefore, the following supply chain of machine learning is rather generic, aiming to cover machine learning more broadly.

Superimposing the machine-learning supply chain on the machine-learning attack surface (compare figures 1 and 2) suggests that most attack techniques – such as data poisoning, data extraction or inserting a backdoor in a model – described in the analysis of the attack surface in the previous paper[27] can be leveraged against individual stages of the supply chain. This needs to be taken into consideration when deriving policy recommendations (see section 6) from the security challenges that come up when analyzing the machine-learning supply chain (see sections 3 and 4).

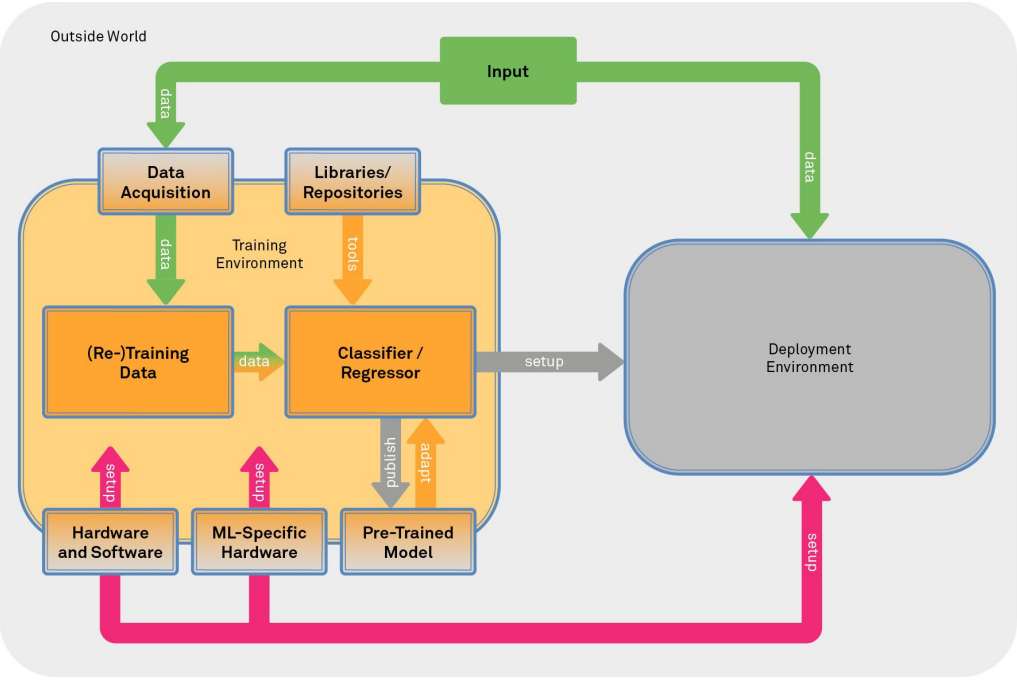
26 For the purpose of this analysis, the machine-learning supply chain does not include the deployment of the machine-learning model.

27 Sven Herpig (2019): Securing Artificial Intelligence

Machine-Learning Attack Surface – updated and adapted version[28] (Figure 1)



Machine-Learning Supply Chain Superimposed On Attack Surface (Figure 2)



28 Based on Sven Herpig (2019): Securing Artificial Intelligence

The supply chain for non-machine-learning applications consists of developer tools and libraries as well as general-purpose hardware. While these are also elements of the machine-learning supply chain, the entire data pipeline, and the specialized hardware, are unique to it. The following section discusses the supply chain's two main segments "data" and "training", with various stages. Each stage is explained and analyzed vis-à-vis possible security challenges. Underlying and interacting with these stages are "platforms and services" as well as "hardware" aspects that are described at the end of the section.

3.1 Data

Data is a prerequisite for machine learning and used in many stages of its development (see figure 3). It is used to train the model (e.g., classifier/regressor), validate the algorithm during the training phase, and evaluate its final readiness (e.g., accuracy, reconstruction loss).[29] Data at various stages form the core of the machine-learning supply chain.

Machine-Learning Supply Chain: "Data" (Figure 3)

Data Generation and Acquisition: The supply chain starts with data generation and acquisition. This can include anything from simulated game engines generated inside a reinforcement learning environment[30] to acquiring pictures of dogs and cats in order to train a classifier to distinguish them[31], to harvesting and clustering user data from social media platforms for targeted advertisements[32] or even data generated inside IT infrastructures by virtual sensors for a machine-learning model to spot anomalies and improve

29 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos at al. (2019): The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

30 e.g., Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub Pachocki, Michael Petrov, Henrique Pondé de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski and Susan Zhang (2019): Dota 2 with Large Scale Deep Reinforcement Learning or Defense Advanced Research Projects Agency (2020): AlphaDogfight Trials Go Virtual for Final Event

31 e.g., Greg Surma (2018): Image Classifier – Cats vs Dogs

32 e.g., C. Perlich, B. Dalessandro, T. Raeder, O. Stitelman and F. Provost (2013): Machine learning for targeted display advertising: transfer learning in action

its own security[33]. There are two noteworthy aspects for this stage. First, the data could be either a) stored and used later to, for example, train a machine-learning model, or b) used on-the-fly through, for example, online learning. Second, at the time the data is stored, it may not be clear whether the data will be used to train a machine-learning classifier/regressor or not. The data may be stored for a specific purpose at first, such as a picture taken and uploaded by a user on a social media profile to interact with other users. Later on it could then be used for unrelated purposes, such as for training a facial recognition system, after it had been scraped by a third party.[34]

Security Challenges: Both aspects are relevant from a security standpoint. While stored data used for training, validation, or evaluation can later be manipulated by an attacker over time, data used on-the-fly needs to be manipulated immediately – for example through live data poisoning – to achieve a result. The fact that data stored for a specific purpose might end up being used for an unrelated purpose may cause a problem due to different threat models and, hence, security requirements. A hypothetical example for this could be images of traffic signs generated for a learning drivers' app (specific purpose) without any security requirements that are later on acquired to train an image classifier for autonomous driving (unrelated purpose), which is a high-risk environment with a certain information security baseline.
If it were clear from the beginning that the data could potentially be used for a different, unrelated purpose, then security standards during the generation and acquisition stage would have to be higher to account for this threat. The key concerns here are the integrity and the quality (accuracy, completeness, timeliness) of the data. When acquiring data, it needs to be transparent whether the original generation and storage of the data can meet the requirements for data quality, traceability, and integrity. Thus, proprietary data collection processes would be an obstacle to securing the machine-learning supply chain.

Data Brokerage: Actors who wish to train or validate a machine-learning model do not have to generate the data themselves. Free and commercial datasets with various types of data – ranging from images of traffic signs[35] to Bitcoin transactions for ransomware[36] – and commercially available con-

33 e.g., Darktrace (2020): Cyber AI Platform, F-Secure (2020): Project Blackfin: Automated Breach Detection Using Intelligent Agents or Geoff McDonald and Saad Khan (2019): In hot pursuit of elusive threats: AI-driven behavior-based blocking stops attacks in their tracks

34 Kashmir Hill (2020): The Secretive Company That Might End Privacy as We Know It

35 Ruhr-Universität Bochum (2019): German Traffic Sign Benchmarks

36 Cuneyt Gurcan Akcora, Yulia Gel and Murat Kantarcioglu (2019): BitcoinHeist: Topological Data Analysis for Ransomware Detection on the Bitcoin Blockchain

sumer data – such as clickstreams and browsing histories – are available for procurement from a variety of (developer) platforms, commercial entities, and research institutions.[37]

Security Challenges: Using third-party datasets requires a certain amount of trust in the parties involved – those that generated/acquired and those that traded the data – as well as the IT infrastructure they are using, to ensure that the datasets are not manipulated by the involved parties or an adversary or (unintentionally) biased[38] or otherwise flawed (e.g., data corruption). Transparency about how the data is secured against interference and what the data includes and other metrics (like in data "coversheets") is extremely useful for those acquiring it. This is especially true if the datasets are being used to train high-risk machine-learning application environments. While the data brokers themselves could conduct a targeted attack against their customers (e.g., poison a dataset for a specific customer), it would seem like a bad business decision to sell intentionally manipulated datasets. External adversaries would likely be limited to untargeted attacks (e.g., poison a dataset being used by all subsequent customers).

Data Curation: After the data has been generated, acquired, or procured, it will likely need to be curated. Depending on the use case of the data, that means checking the quality of data, integrating various datasets, cleaning datasets for abnormalities, or labeling data. Data curation done right takes vast resources – a 2016 survey found that data scientists spent 80% of their time preparing the data, a task that the majority of them do not enjoy doing.[39] This makes data curation ripe for outsourcing, adding additional parties to the machine learning supply chain: managed services providers, such as Amazon SageMaker Ground Truth[40], as well as vendors and individuals (endpoints) that eventually do the work, or tools such as Snorkel[41] that support this activity.

Security Challenges: Again, the security of the IT infrastructure of the managed services provider and the endpoint are highly relevant to securing the machine learning supply chain, especially when it comes to sensitive data

37 e.g., Will Badr (2019): Top Sources For Machine Learning Datasets

38 Katyanna Quach (2020): MIT apologizes, permanently pulls offline huge dataset that taught AI systems to use racist, misogynistic slurs

39 Gil Press (2016): Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

40 Amazon Mechanical Turk (2018): AWS introduces a new way to label data for Machine Learning with MTurk

41 Snorkel AI (2020): Snorkel

or data that will later be used in machine learning models for high-risk environments. Even if the data is not outsourced, errors may be introduced in the data curation process that may affect the integrity or quality of the data. After this stage is completed, the data is specific to the developer of the machine-learning model and, therefore, enables a more targeted attack from an adversary. Attackers could manipulate, add, or delete curated data so that the machine-learning model trained on the tainted data misclassifies input deliberately chosen by the adversary.

Data Storage: At every stage before the machine learning model is finalized, the data used must be stored either locally, in the cloud, or both.

Security Challenges: From a security point of view, storing data in the cloud may mean relying on the cloud providers' security measures that protect the infrastructure, and, thus the data. This does not necessarily mean that the security level of local, non-cloud infrastructure is better. Depending on the resources and the cloud service providers, cloud storage may be more secure against third parties. The cloud provider may have significant access to the data, though, which needs to be considered for the threat assessment. With access to the data in storage, attackers can conduct targeted attacks – e.g., through data poisoning. Besides potential threats to the integrity and confidentiality of the data, availability needs to be considered in the local versus cloud decision.

Data Input: This stage of the machine-learning supply chain is applicable to online learning, where the input data for the model is used for live learning.

Security Challenges: An adversary can feed manipulated data directly into the model, resulting in a malicious training dataset and a compromised model.[42] An example of this would be information security anomaly detection tools that are trained on live network traffic or Microsoft's AI chatbot "Tay"[43]. If that traffic is already malicious, the input data will be as well, possibly leading to a trained but dysfunctional model, which recognizes the malicious traffic as "normal". It must be ensured, possibly by the infrastructure provider, that the input data is benign before it is fed into the model as input data.

42 This differs from adversarial sample attacks, in which the model is already trained and the digital or physical input data is designed to exploit vulnerabilities.

43 James Voncent (2016): Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day

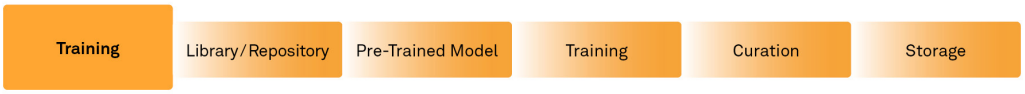
Security Challenges of the Machine-Learning Supply Chain – Data (Table 1)

Stage	Attacks	Security Challenges
Data Generation and Acquisition	Untargeted [44]	Security may depend on a third party (data generator or aggregator). Unclear purpose of data at the point of generation (e.g., used for application in high-risk environments later on), therefore wrong threat model. Integrity of data can be compromised.
Data Brokerage	Untargeted [45]	Security depends on a third party (data broker). Unclear use purpose of data at the point of generation, therefore wrong threat model. Integrity of data can be compromised.
Data Curation	Targeted	Security may depend on third parties (service providers). Curation of data can be manipulated towards a specific purpose.
Data Storage	Targeted	Security may depend on a third party (service provider). Availability, integrity, and confidentiality of data can be compromised, as data is curated and very specific at this stage.
Data Input	Targeted	Security depends on deployment infrastructure. Compromise of integrity might lead to a compromised baseline.

3.2 Training

The training stage includes tools and services required for developing a machine learning model. The section describes the increasing number of tools and services[46] used by developers where parts may be outsourced or leveraged as-a-service.

Machine-Learning Supply Chain: "Training" (Figure 4)



44 With the right resources and intelligence, targeted attacks would certainly also be possible.

45 With the right resources and intelligence, targeted attacks would certainly also be possible.

46 Chip Huyen (2020): What I learned from looking at 200 machine learning tools

Library/Repository: Libraries or repositories are core elements for machine learning. They contain several tools and algorithms required for the development of machine-learning models. Many of these libraries and repositories, even the most versatile and specialized ones, offer open-source resources and are also either Free and Open Source Software (FOSS) or Free/Libre Open Source Software (FLOSS). Examples are TensorFlow[47], Keras[48], or Shogun[49]. They can either be downloaded and installed on local machines or used in the cloud, often through intermediaries such as the Google Cloud ML Engine[50]. The libraries and intermediaries used are additional components of the machine-learning supply chain.

Security Challenges: The security of the intermediaries hinges on the information security of their infrastructure. For libraries, security also depends on where they are hosted, like, accounts on developer platforms such as GitHub[51]. While attacks against the intermediaries could be targeted, attacks against libraries can be targeted or untargeted and will affect all developers that subsequently use them to develop machine-learning models.

Pre-Trained Model: Transfer learning is a method that allows developers to save (vast amounts of) resources in terms of computing power and data access by using pre-trained models and fine-tuning the training for the specific task. Due to these economic factors, there is, and will increasingly be, a reliance on pre-trained models in many parts, which can be regarded as another kind of outsourcing. Examples are R50x1[52] or ResNet-34[53].

Security Challenges: Even though pre-trained models are still evolving, research shows that vulnerabilities in these models, such as deep neural network backdoors, can persist in the final model after transfer learning or retraining[54] and can then be exploited by adversaries, e.g., to misclassify input. Pre-trained models are available on a wide range of developer platforms. Attacks against pre-trained models are likely to be untargeted, affecting every subsequent developer using them.

47 TensorFlow (2020): TensorFlow

48 Keras (2020): Keras

49 NumFOCUS (2020): Shogun

50 Google Cloud (2020): AI Platform

51 Agence nationale de la sécurité des systèmes d'information and Bundesamt für Sicherheit in der Informationstechnik (2019): Second Edition of the Franco-German Common Situational Picture and Tianyu Gu, Brendan Dolan-Gavitt and Siddarth Garg (2019): BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

52 TensorFlow Hub (2020): bit

53 Kaggle (2017): ResNet-34

54 Yuanshun Yao, Huiying Li, Haitao Zheng and Ben Y. Zhao (2019): Latent Backdoor Attacks on Deep Neural Networks and Tianyu Gu, Brendan Dolan-Gavitt and Siddharth Garg (2019): BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

Model Training: Training and testing the machine learning model can be done locally or in the cloud via machine learning as a service. The more computational power[55] and/or specific hardware required, the higher the chance that outsourcing the training will save resources or even enable the training when not enough local computational power is available.

Security Challenges: From a security point of view, outsourcing means control of security measures will only be indirectly (e.g., through contractual agreements), as it is directly implemented by the outsourcing partner for the infrastructure and thus the model. Security considerations are similar to those mentioned for outsourcing in the data stages. Training is a critical stage in which attackers can conduct targeted attacks to persistently influence the final model.[56]

Model Curation: A trained model, be it internally or externally trained, may need additional curation, such as adjustments of hyperparameters, optimization of the model, or neural cleansing. This might require bringing in external expertise, especially if there is no internal machine-learning expertise available, and the model was trained using external services.

Security Challenges: At this stage, the model is extremely vulnerable to targeted interference since an attacker may not only be able to acquire a lot of knowledge (white box knowledge) for specific future attacks[57] but also to adjust the model parameters in a way that allows specially crafted attacks while everything else works as intended.

Model Storage: Before a finished model is deployed, it needs to be stored.

Security Challenges: One possible attack would be to replace the model where it is stored with a near-identical copy that retains a backdoor. This would, however, require vast resources, such as access to training data and algorithms as well as the IT infrastructure and would not be difficult to detect, e.g., through hash verification. While it can never be ruled out, it is unlikely. Another possible attack at this last stage in the supply chain is side-channel attacks, such as attempting to retrieve the model configuration, as it might be valuable intellectual property. Most side-channel attacks

55 Tianyu Gu, Brendan Dolan-Gavitt and Siddharth Garg (2019): BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

56 Sven Herpig (2019): Securing Artificial Intelligence

57 Sven Herpig (2019): Securing Artificial Intelligence

deployed against machine-learning models so far have required physical access[58], therefore making it challenging to conduct them outside of a test environment.

Security Challenges of the Machine Learning Supply Chain – Model (Table 2)

Stage	Attacks	Security Challenges
Library/Repository	Targeted/ Untargeted	Security depends on a third party (hosting infrastructure and/or developer accounts).
Pre-Trained Model	Untargeted[59]	Security depends on a third party (hosting infrastructure and/or developer accounts). Vulnerabilities, such as backdoors, might exist and persist. Pre-trained models would need to be analyzed before use.
Model Training	Targeted	Security depends on the training environment (local or outsourced). Critical stage in the model development, allowing attackers to conduct targeted, persistent interference.
Model Curation	Targeted	Security depends on the training environment and/or on a third party (e.g., external expertise). Critical stage in the model development, allowing attackers to conduct targeted, persistent interference immediately or later on.
Model Storage	Targeted	Security depends on the training environment (local or outsourced). Attacks, such as side-channel attacks, are challenging to conduct at this stage.

3.3 Platforms and Services

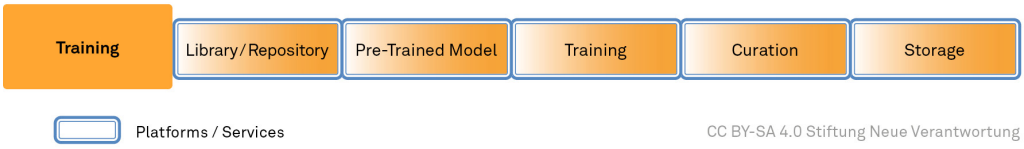
There are platforms that offer a range of services for the development of machine-learning models – machine learning as a service (MLaaS). They include one or more of the data and model stages of the machine-learning supply chain laid out in the prior sections, and they significantly lower the barrier of entry for developers. Services can even run the machine learning model, which developers can then simply query using an Application Programming

58 Anuj Dubey, Rosario Cammarota and Aydin Aysu (2019): MaskedNet: The First Hardware Inference Engine Aiming Power Side-Channel Protection and Lejla Batina, Shivam Bhasin, Dirmanto Jap and Stjepan Picek (2018): CSI Neural Network: Using Side-channels to Recover Your Artificial Neural Network Information and Vasisht Duddu, Debasis Samanta, D. Vijay Rao and Valentina E. Balas (2019): Stealing Neural Networks via Timing Side Channels

59 With the right resources and intelligence, targeted attacks would certainly also be possible.

Interface (API).[60] Examples of platforms are Microsoft Azure Machine Learning[61], Google Cloud AutoML[62], and Amazon SageMaker[63]. Some organizations and governments also have internal platforms, e.g., for deployment, which, in turn, may be provided and maintained by vendors such as the ones mentioned above.

Machine-Learning Supply Chain: "Platforms and Services" (Figure 5)



In general, every stage in the machine-learning supply chain potentially involves outsourcing, which ultimately leads to one or, more likely, several third parties, which become part of the development process. There exists a vast field of companies offering everything from data storage to full service "all-in-one" platforms.[64] Acquiring datasets and pre-trained models to develop a machine-learning model through a platform seems to be far more accessible to a majority of machine-learning developers than generated data, training models from scratch, and procuring specialized hardware. The future of machine learning is likely to be dictated by outsourcing and working with services and tools of various third parties.

Security Challenges: The security of future machine-learning models will depend on the security of various parties involved in the process. Tracking and tracing vulnerabilities and security breaches across this deep and broad supply chain is going to be a major challenge for the developers of machine-learning models and will likely result in a number of vulnerable applications deployed across sectors with a potentially fatal impact (see sections 4 and 5).

60 Helen Kovalenko (2020): Choosing the Best Machine Learning API for Your Project

61 Microsoft Azure (2020): Azure Machine Learning

62 Google Cloud (2020): Cloud AutoML

63 Amazon (2020): Amazon SageMaker

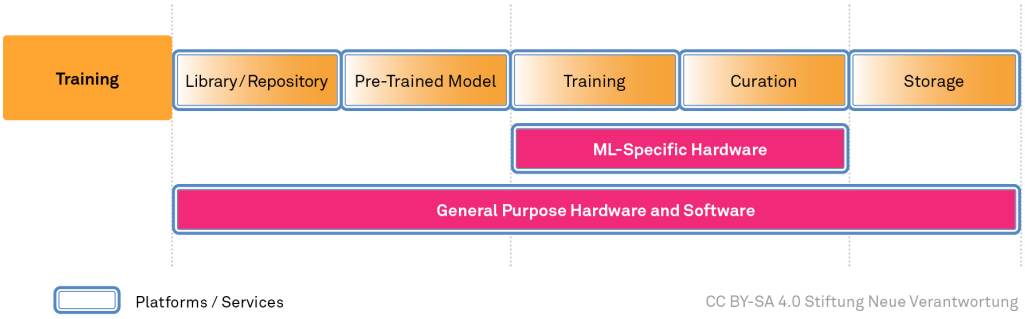
64 Chip Huyen (2020): Machine Learning Production Pipeline – Project Flow and Landscape from Sergey Karayev – Full Stack Deep Learning Bootcamp 2019

3.4 Hardware

Advances in machine-learning models in recent years have increasingly relied on specialized hardware such as Google's Tensor Processing Unit (TPUs) for neural networks.[65]

Security challenges: Therefore, hardware specific to machine learning becomes part of the supply chain where interference, such as side-channel attacks against machine-learning models[66], needs to be considered. This adds to the already existing challenges pertaining to the security of the hardware supply chain for more traditional chips like CPUs or GPUs.[67] Apart from hardware-specific attacks, requiring specialized hardware further drives the economic aspect, which likely leads to increased outsourcing and relying on a smaller number of companies offering MLaaS. Those companies, however, will rely on an even smaller number of hardware manufacturers.[68] Therefore, vulnerabilities in their chips are likely to have a staggering impact due to scale effects.

Machine-Learning Supply Chain: "Hardware" (Figure 6)



65 Neil C. Thompson and Svenja Spanuth (2018): The Decline of Computers as a General Purpose Technology: Why Deep Learning and the End of Moore's Law are Fragmenting Computing

66 Anuj Dubey, Rosario Cammarota and Aydin Aysu (2019): MaskedNet: The First Hardware Inference Engine Aiming Power Side-Channel Protection and Lejla Batina, Shivam Bhasin, Dirmanto Jap and Stjepan Picek (2018): CSI Neural Network: Using Side-channels to Recover Your Artificial Neural Network Information

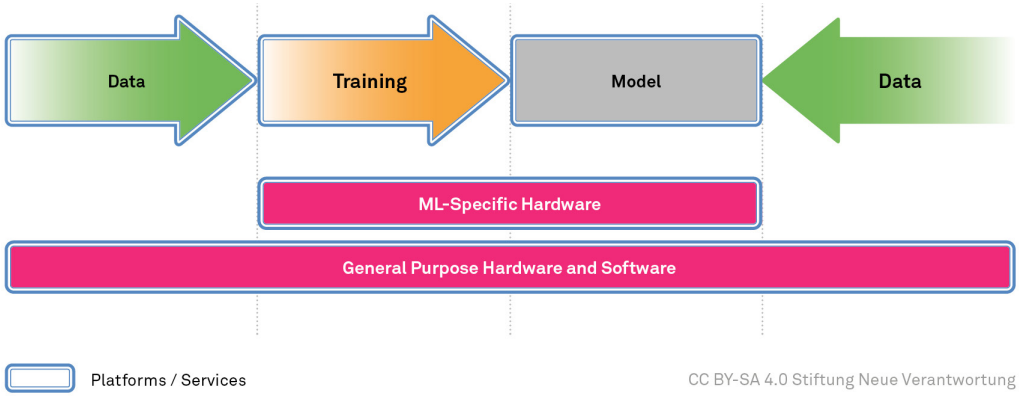
67 e.g., Jon Boyens, Celia Paulsen, Rama Moorthy and Nadya Bartol (2015): NIST Special Publication 800-161 – Supply Chain Risk Management Practices for Federal Information Systems and Organizations

68 e.g., Jan-Peter Kleinhans and Nurzat Baisakova (2020): The global semiconductor value chain

4. Security Implications for the Machine-Learning Supply Chain

There is not only one machine-learning supply chain. Depending on the machine-learning model – for example, online learning or reinforcement learning – its development requires different stages. Each of those stages can take place in-house or be outsourced or both. The supply chain of any machine-learning model's development is likely a hybrid of those two extremes, possibly with a staggering number of intermediaries. For example, data is generated somewhere and bought from a data broker before being curated and used for training on-premise while leveraging libraries or using pre-trained models that were acquired from developer platforms. Thus, each stage comes with its own security risks that, taken together, form the combined risk of each individual application's machine-learning supply chain.

Machine-Learning Supply Chain: "Overview" (Figure 7)



There are several general security implications for the machine-learning supply chain that can be derived from the challenges identified in the various stages of machine learning in the previous chapter.

4.1 Foundation

General-purpose hardware, and software, conventional IT systems, still play a major role in the development process of machine-learning models. That includes systems that collect and store data before it is used for training, or online repositories of development tools and pre-trained models. If those IT systems are compromised, attackers will be able to have an effect on the development of machine-learning models, for example, by compromising

developer accounts on MLaaS platforms through phishing attacks. So, when discussing the security of machine learning and its supply chain, it is crucial not to have a narrow focus on machine-learning-specific parts. All the underlying, conventional IT systems need to be secured and monitored as well. This may seem obvious to some, but it is still worth mentioning, as it might be an overlooked Achilles Heel.

4.2 Pull Factor

The development of both machine-learning and non-machine learning applications may require that tools provided by external third parties be pulled into the development process. Therefore, in both cases, it is important that the tools are secure and that the developer is informed when there has been a compromise. The difference is that the machine-learning supply chain is vast and includes developer tools and services not only for the training part but also for several stages on the side of data acquisition and preparation. As shown, any of the stages can be exploited by an adversary for an attack. Being aware of the security level and compromises across the entire machine-learning supply chain for an application already seems challenging, but managing it may be extremely difficult.

4.3 Push Factor

As mentioned earlier, pushing certain development tasks to external service providers is and will increasingly become the norm in machine-learning development. It mitigates otherwise limiting factors such as computing power (for the training process), workforce (e.g., for data curation), or trained and experienced in-house developers (for the entire development process)[69]. Due to the latter, there is an increased pressure of automation for which the demand is met by MLaaS. However, that bears the risk of machine-learning models being developed without anyone, even the developers, actually understanding them. The lack of explainability and interpretability of machine learning is already a serious challenge that will be exacerbated by automation, causing the risk of these systems malfunctioning and being or remaining insecure to increase drastically.

69 Philippe Lorenz and Kate Saslow (2019): Demystifying AI & AI Companies, Remco Zwetsloot, James Dunham, Zachary Arnold and Tina Huang (2019): Keeping Top AI Talent in the United States and Maaike Verbruggen (2020): AI & Military Procurement: What Computers Still Can't Do

Apart from providing all the resources needed to develop a machine-learning model, using platforms and service providers also outsources the security to the platform provider. While requiring (ultimate) trust in the platform providers' security and acting in good faith, relying on MLaaS might actually increase the overall information security of machine-learning development, depending on the security of the MLaaS providers' infrastructures, the developer accounts, and the developers' (endpoint) security. In order to increase trust, service providers have to be fully transparent about their security and privacy measures, audits, breaches, or interferences in other aspects such as state-mandated backdoors. All this information is vital but only matters if developers know what to do with it and how it affects their machine-learning supply-chain security overall.

4.4 Threat Actors

Cyber operations against software[70] and hardware[71] supply chains are hardly new. The specific threat model[72] is another crucial aspect that needs to be taken into account when looking at the machine learning supply chain. Depending on the stages, targeted and/or untargeted attacks are possible. Targeted interference is much harder to achieve, while untargeted attacks can be carried out, for example, by poisoning publicly available datasets or backdooring pre-trained models without regard to who will subsequently use them, as long as the ultimate goal can be achieved by interfering with these models later on (e.g., for financial gain). Targeted attacks require acquiring direct access to the IT systems of developers or service providers. A successful targeted interference with the development of a machine-learning model can have a grave impact, and it might be difficult to detect and attribute.[73]

70 e.g., Trey Herr, June Lee, William Loomis and Stewart Scott (2020): BREAKING TRUST: Shades of Crisis Across an Insecure Software Supply Chain and Micah Lee and Henrik Moltke (2019): Everybody Does It: The Messy Truth About Infiltrating Computer Supply Chains and Beau Woods and Andy Bochman (2018): Supply Chain in the Software Era and GReAT and AMR (2019): Operation ShadowHammer and Brian Thomas (2020): FBI Alerts Companies of Cyber Attacks Aimed at Supply Chains

71 e.g., Glenn Greenwald (2014): Glenn Greenwald: how the NSA tampers with US-made internet routers and United Kingdom National Cyber Security Centre (2018): Supply chain security guidance

72 e.g., Andrew Marshall, Jugal Parikh, Emre Kiciman and Ram Shankar Siva Kumar (2019): Threat Modeling AI/ML Systems and Dependencies

73 Sven Herpig (2019): Securing Artificial Intelligence

Untargeted attacks can be carried out by a wide range of threat actors, from malicious individuals and organized criminals (e.g., for cryptojacking[74] or to install ransomware[75]) to nation-state actors. Targeted attacks can achieve very specific results (e.g., planting a backdoor in a neural net[76] or extracting intellectual property in form of model configurations[77]) and require a certain level of resources, such as expertise in machine learning or the budget to acquire it from third parties. Therefore, they are more within the realm of well-resourced nation-state actors.

74 Randi Eitzman, Kimberly Goody, Bryon Wolcott and Jeremy Kennelly (2018): How the Rise of Cryptocurrencies Is Shaping the Cyber Crime Landscape: The Growth of Miners and Trend Micro (2019): Attackers Targeting Cloud Infrastructure for their Cryptocurrency-Mining Operations

75 VECTRA (2019): The biggest threat from ransomware: Malicious encryption of shared network files and Corey Nachreiner (2020): Why Ransomware Will Soon Target the Cloud

76 Yuanshun Yao, Huiying Li, Haitao Zheng and Ben Y. Zhao (2019): Latent Backdoor Attacks on Deep Neural Networks

77 Binghui Wang and Neil Zhenqiang Gong (2018): Stealing Hyperparameters in Machine Learning

5. Security Challenges of Machine Learning

The possible use cases for machine learning are vast, as is the decade-long debate on how to best secure IT systems and IT infrastructures. Evidently, there is a certain need to secure even the least crucial connected device, as proven by the Mirai Botnet[78], which consisted of hundreds of thousands of Internet of Things devices that were compromised to conduct distributed denial-of-service attacks. Even more importantly, it is essential to develop robust security mechanisms for those applications whose individual malfunctioning or takeover can lead to injuries, loss of life, and/or infringement upon fundamental rights. For all those environments, fundamental security challenges can be derived from the machine learning attack surface[79] and machine-learning supply chain.

Risk Assessment: A compromised "AI-powered" toothbrush[80] would qualify as a low-risk environment in most cases. Certainly, there might be scenarios – for example, exploiting thousands of compromised toothbrushes for a distributed denial-of-service attack against a critical infrastructure – where serious damages can be incurred. However, taken individually and in most cases, a compromised toothbrush can easily be regarded as a low-risk environment. A fully autonomous car, defined as "capable of performing all driving functions under all conditions"[81], would therefore constitute a high-risk environment as its compromise can easily lead to injuries or loss of life. Another example would be a tampered-with machine-learning application that is used in the criminal justice system to predict the future likelihood of committing crimes[82] as it could lead to the wrongful (temporary) limitation or loss of certain rights. In addition to a machine-learning application being deployed in such a high-risk environment, its function, second- and third-order consequences, or even accumulated latent low-risk effects can constitute a high-risk environment. It will, therefore, not be sufficient to just use the definition of "critical infrastructures" that some states have politically and/or legally defined[83] and apply it 1:1 to machine learning. Individual

78 Cloudflare (nondated): What is the Mirai Botnet?

79 See chapter 4 „Strategic Implications of the Attack Surface" in Sven Herpig (2019): Securing Artificial Intelligence

80 Jay Peters (2019): Oral-B's new $220 toothbrush has AI to tell you when you're brushing poorly

81 U.S. National Highway Traffic Safety Administration (nondated): Automated Vehicles for Safety

82 Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner (2016): Machine Bias

83 e.g., Bundesamt für Sicherheit in der Informationstechnik and Bundesamt für Bevölkerungsschutz und Katastrophenhilfe (2020): Kritische Infrastrukturen and Bundesanzeiger Verlag (2015): Gesetz zur Erhöhung der Sicherheit informationstechnischer Systeme (IT-Sicherheitsgesetz)

risk assessment and management is crucial for evaluating risk for machine learning deployments.[84]

Retraining: Updates for conventional software include code changes that are then rolled out to the affected devices. In order to fix a bias or adversarial manipulation in a deployed machine-learning application, developers have to re-do the underlying models – or pre-trained models – from scratch with new data or a completely changed dataset.[85] Therefore, retraining might involve going through one or many stages of the machine-learning supply chain again. Until the retraining is complete, it has been verified that the new model does not have the same bias/vulnerability – or a new one – and it has been deployed, the original machine-learning application will have to remain live and vulnerable in its environment.

Human Agency: Measures that are often suggested to counteract automated propagation of false decisions are integrations of human decisions. Referred to as human-on-the-loop, humans can halt/delay actions that would otherwise automatically be taken by machines/IT-systems. Another implementation of this measure is human-in-the-loop[86], where humans need to approve actions before they are taken by machines/IT-systems.[87] These concepts, however, are based on the assumption that the decisions made or suggested by the machine learning model are obviously wrong and/or allow humans enough time to cross-check them. In reality, aspects such as lack of situational awareness, available time, automation bias, and moral buffer[88] may lead to a perceived human agency, where in fact, there is none. In other (edge) cases, such as automated air defense systems, time constraints render the application of human-on/in-the-loop infeasible. However, this does not mean that human agency should be discounted. Wherever it is somehow feasible, it should be applied in high-risk environments to increase safety. When implemented, it cannot be taken out of context but must reflect on the limiting factors such as automation bias and potentially counteract them with other safety measures.

84 In information security, there are several established risk management frameworks which may be adopted and adapted for machine learning, e.g. NIST (2020): Risk Management Framework.

85 Luigi (2019): The Ultimate Guide to Model Retraining

86 An example would be U. S. Army's FIRESTORM, see Nathan Strouth (2020): Inside the Army's futuristic test of its battlefield artificial intelligence in the desert

87 Human Rights Watch (2012): Losing Humanity – The Case against Killer Robots

88 International Committee of the Red Cross (2019): Artificial intelligence and machine learning in armed conflict: A human-centred approach

Scale Effects: Machine-learning models can make a staggering number of decisions in a short time span. If the model is compromised, however, that leads to a staggering number of wrong and potentially harmful decisions in a short time span. In effect, applied machine learning, like all automation, has the potential to dramatically scale effects, indifferent to the effects being intended, unintended, or catastrophic. Of course, machine-learning applications can not only be connected to each other, either as a sequence or in parallel, but also connected to other IT systems or even operational technology (OT) systems. Due to the scale effects and interconnectedness, targeted and untargeted attacks against the machine-learning supply chain may produce devastating effects, especially in high-risk environments. Due to the vast attack surface (see figure 1) and the supply chain of machine learning, potential attackers can probe a wide variety of areas for weak spots and vulnerabilities. A simple hypothetical example would be a manipulated image classifier deployed in a range of fully autonomous cars that works as intended in all cases except when the sensors of the car scan a traffic sign that normally does not exist (e.g., a speed limit sign that says "42") in which case it classifies it as a stop sign. This information is sent from the classifier to the main control, which triggers an emergency brake. All cars with that classifier in that area would be affected as well as all other cars in the immediate vicinity of those cars due to the erratic driving behavior.

Delayed Effects: The scale effect describes first-order effects that can be triggered by compromising the machine-learning supply chain and potentially have immediate catastrophic impacts. On the other hand, by compromising the machine-learning supply chain, delayed second- and third-order effects that have a latent long-term impact can also be caused. Due to the lack of explainability, interpretability, and traceability, among other aspects, detecting interference in machine learning is very challenging[89], especially across the vast machine-learning supply chain. Therefore, (subtle) successful attacks can go undetected for an extended period of time. By the time it is noticed that the machine-learning model is not operating the way it is intended to, it might be impossible to determine whether that problem stems from an attack or simply a bias or another mistake across the supply chain during development. An example of this is a hypothetically compromised machine-learning application deployed throughout the entire criminal justice system to support judges' decisions on jail time. It is manipulated to discriminate against a certain minority, suggesting higher jail times for this minority as compared to other segments of society. The first-order effect

89 Sven Herpig (2019): Securing Artificial Intelligence

is, of course, undue jail time for this minority. A second-order effect could be self-reinforcing data bias resulting in even more extended jail time for this minority, as well as an increasingly divided society and discrimination against that minority based on the implications of the first-order effect. As a result, a third-order effect could be civil unrest due to the minority's unjust treatment.

6. Recommendations

While technical research[90] and standardization[91] with regard to machine-learning security are well underway – wide-ranging (mandatory) security measures are covered in existing regulations such as the EU General Data Protection Regulation (GDPR)[92], the EU Directive on Security of Network and Information Systems (NIS Directive)[93], or the U.S. Federal Information Security Modernization Act of 2014 (FISMA)[94] – the existing measures may be insufficient to cover the information security of machine-learning applications, especially in high-risk environments[95], due to the security implications of the machine-learning supply chain discussed in this paper. However, rather than looking at the standardization or regulatory space, this paper suggests considering security measures independent from the means of policy enforcement. This may also have the added effect of being agnostic to countries' preferred strategies (soft law versus hard law approaches). The following recommendations should be put forward by governments and the private sector to improve the overall security of the machine-learning supply chain.[96]

Design a security approach rooted in conventional information security

As shown in the paper, large parts of the machine-learning supply chain and attack surface consist of general-purpose hardware and software, such as cloud servers or developer accounts. Attackers, therefore, are likely to compromise those systems to subsequently interfere with components such as training data or hyperparameters of a model. Though it is a truism, it cannot be reiterated enough: state-of-the-art information security and cyber hygiene need to be applied to all the general-purpose IT systems involved in

90 Sven Herpig (2019): Securing Artificial Intelligence

91 Sven Herpig (2020): No Safety without Cybersecure AI

92 Official Journal of the European Union (2016): REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

93 Official Journal of the European Union (2016): DIRECTIVE (EU) 2016/1148 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union

94 U.S. Government Publishing Office (2014): Public Law 113-283

95 European Commission (2020): On Artificial Intelligence – A European approach to excellence and trust

96 Recommendations to improve security of the machine learning supply chain for high-risk environments might differ from both, low-risk environments and environments that do not rely on machine learning applications at all.

the training and development of machine learning models. These measures include awareness campaigns for employees, application security, network segmentation, coordinated vulnerability disclosure, penetration testing, and red team exercises. There are plenty of standards and best practices out there, such as the ISO/IEC 27000-series, the NIST cybersecurity framework[97], or the BSI IT-Grundschutz[98], that can readily be applied. When discussing and developing policies to increase the security of the machine-learning supply chain, decision-makers need to consider security aspects of the underlying general-purpose hardware and software, as this forms the base of developing machine-learning models. If they are not secure, this will be the case for the entire development process.

While the best outcome would always be to prevent interference, the key to securely developing machine-learning models is detection. Due to the prevailingly opaque nature of machine learning models – vis-à-vis explainability – it is challenging to identify the unintended functions of a model and even more difficult to separate intentional from unintentional actions that lead to a malfunctioning model.[99] Additionally, theft of intellectual property regarding model details would be hard to prove without proper detection mechanisms. Successful detection can lead to a variety of responses. Specific to machine learning would, for example, be retraining a model on adjusted data (if poisoned by an attacker) or a neural cleanse of a deep neural network (if backdoored by an attacker). It is, therefore, imperative to detect adversarial interference with a high degree of confidence. Decision-makers should consider incentivizing the implementation of conventional breach detection technologies across the entire machine-learning supply chain for high-risk environments as well as mandatory information-sharing regarding actual breaches. Additionally, intelligence and other security agencies need to be aware of the implications of adversarial interference in the machine-learning supply chain and add this field to their threat analysis and situation picture.

Increase transparency, traceability, validation, and verification

Transparency, traceability, validation, and verification across the entire supply chain are important when data, tools, and services are used to develop, curate, store, and train a model, especially in high-risk environments.

97 United States National Institute of Standards and Technology (2020): Cybersecurity Framework

98 Bundesamt für Sicherheit in der Informationstechnik (2020): IT-Grundschutz

99 Sven Herpig (2019): Securing Artificial Intelligence

Transparency includes aspects such as which third parties are involved in the process, as well as which security standards were adhered to on the systems involved in the process. Tools such as libraries and pre-trained models, as well as datasets, should be digitally signed for verification purposes. Having and sharing this information across the supply chain not only enables the developer of machine-learning models for high-risk environments to make informed decisions but also facilitates backtracking in case vulnerabilities or other threats are detected. Therefore, decision-makers may want to consider a software bill of materials[100] that possibly also includes security mechanisms (such as breach detection mechanisms) used during the various stages of the machine-learning supply chain. Furthermore, decision-makers could also support the development of the tools needed for validation, testing, and verification. In addition, for machine-learning models to be specifically used in high-risk environments, it may be useful for policy-makers – in partnership with the private sector – to support (e.g., with financing) or lead the creation of highly secure[101] repositories of verified datasets, libraries, pre-trained models, et cetera, deployed using reproducible builds. Such repositories could also serve additional purposes such as facilitating coordinated vulnerability disclosure.

Identify, adopt, and apply best practices

The European Union works toward defining high-risk environments for machine learning, announcing that due to the nature of machine learning, this definition might vary from established ones.[102] Based on the analysis in this paper, an individualized process to define high-risk environments for machine learning deployment seems logical. Also, the required security needs to be baked into the supply chain. For these activities, it makes sense to consult frameworks for transferable best practices and adapt them so that they cover the particularities of machine learning. This applies to frameworks for risk management[103], supply chain security[104], operational technology, and

100 Fred Bals (2019): What is a software bill of materials? and Greg Slabodkin (2020): Insulin pumps among millions of devices facing risk from newly disclosed cyber vulnerability, IBM says

101 Such central IT-systems which store sensitive data are a prime target for adversaries and as such a potential single point of failure.

102 European Commission (2020): White Paper – On Artificial Intelligence – A European approach to excellence and trust

103 e.g., United States National Institute of Standards and Technology (2018): Risk Management Framework for Information Systems and Organizations – A System Life Cycle Approach for Security and Privacy

104 e.g., United States Office of the Director of National Intelligence (2020): Supply Chain Risk Management

industrial control systems[105]. The latter is included due to the overwhelming importance of secure programming/training and development vis-à-vis post-deployment patching. While non-machine-learning IT systems can in most cases easily be patched once a vulnerability is detected and a patch is issued, patching deployed machine-learning models is harder, especially if they are running in production, as it may involve retraining and redeploying the entire model. Therefore, decision-makers should promote the concept of model governance[106] further and consider it as a best practice for securing the machine-learning supply chain.

Require fail-safes and resiliency measures

Machine-learning models, especially those deployed in high-risk environments, should include strong safeguards and resilience measures. Though security standards should be high across the supply chain, information security circles have long ago adopted the "assume breach" mantra, moving beyond security and additionally focusing on resilience "to respond to and recover (even with increased strength) from a disturbance"[107]. For machine learning, this means including measures such as input data validation, explainability, interpretability, redundancy, and multi-party evaluations with a "better safe than sorry" approach[108]. Models deployed in high-risk environments should – whenever feasible – be developed with a human-in-the-loop or at least a human-on-the-loop functioning as a fail-safe. These safeguards are, however, no silver bullet as they might suggest human agency where in reality there is none.[109] When discussing explainability for machine-learning models in this field it also needs to be noted that human decision making may also not be explainable at times, especially in high-stakes decisions.[110]

Decision-makers should consider incentivizing a baseline of resiliency measures, on top of information security and fail-safe mechanisms, for machine-learning models to be deployed in high-risk environments.

105 e.g., Bundesamt für Sicherheit in der Informationstechnik (2013): ICS-Security-Kompendium and Keith Stouffer, Victoria Pillitteri, Suzanne Lightman, Marshall Abrams and Adam Hahn (2015): Guide to Industrial Control Systems (ICS) Security and United States Department of Homeland Security (2020): Recommended Practices

106 David Asermely (2019): Machine Learning Model Governance and DataRobot (2020): Production Model Governance

107 Kate Saslow (2019): Global Cyber Resilience: thematic and sectoral approaches

108 Ben Nassi, Dudi Nassi, Raz Ben-Netanel, Yisroel Mirsky et al. (2020): Phantom of the ADAS:Phantom Attacks on Driver-Assistance Systems

109 International Committee of the Red Cross (2019): Artificial intelligence and machine learning in armed conflict: A human-centred approach

110 Molly Kovite (2019): I, Black Box: Explainable Artificial Intelligence And The Limits Of Human Deliberative Processes

Create a machine-learning security ecosystem

Machine learning is one of the driving forces for artificial intelligence and will likely continue gaining momentum. Securing it is not only vital for the success of machine learning but could also become a substantial market in itself. It is, therefore, a good idea for governments to support the development of an ecosystem that prioritizes security. This includes technical research into the security and privacy of machine-learning development and training, such as adversarial training, robustness, explainability, interpretability, secure multi-party computation, federated learning, differential privacy, or side-channel protection.[111] Such technical research is fostered through a robust academic environment where research can take place and talent can be educated. This does require a computing infrastructure adequate to support large scale machine-learning research as well as an enabling data ecosystem. Governments can foster this through special research grants, excellence cluster initiatives, joint information security and machine-learning degrees, and other well-known tools in this area.

Attracting, retaining and managing talent is key to a solid machine-learning ecosystem – no matter the sector or country. The National Security Commission on Artificial Intelligence of the United States even stated: "We regard talent as the most valuable resource because it drives the creation and management of all of the other [artificial intelligence] components".[112] On top of fostering the scientific environment, there needs to be business opportunities and applied research in security-related areas such as penetration testing, red teaming, and other services, as well as products that leverage machine learning. Government agencies that support projects and their transitions into services and products, such as the United States Defense Advanced Research Projects Agency (DARPA) – for example through the Guaranteeing AI Robustness Against Deception program[113] – or the German Cyberagentur and Cyber Innovation Hub[114], should include secure machine learning in their research agenda if they have not already done so.

111 Various technical measures mentioned e.g., Sven Herpig (2019): Securing Artificial Intelligence and Bitkom (2019): Blick in die Blackbox as well as Anuj Dubey, Rosario Cammarota and Aydin Aysu (2019): MaskedNet: The First Hardware Inference Engine Aiming Power Side-Channel Protection

112 National Security Commission On Artificial Intelligence (2019): Interim Repor

113 Bruce Draper (2020): Guaranteeing AI Robustness Against Deception (GARD)

114 Stiftung Neue Verantwortung (2020): Akteure und Zuständigkeiten in der deutschen Cybersicherheitspolitik

Set up a permanent platform for threat exchange

Machine-learning security and ultimately the cybersecurity of artificial intelligence is a fast-evolving field, both from a technical and a deployment perspective. It is therefore crucial to have a permanent multi-stakeholder space where researchers, industry, and government representatives, as well as members of civil society, exchange threat information, develop a specialized vulnerability disclosure process (especially for open source tools, libraries, and datasets), monitor risks, investigate supply-chain vulnerabilities and systemic threats, share best practices, and conduct scenario analysis and desktop exercises. If possible, existing structures should be leveraged to fulfil this role. Another option would be to form a European or transatlantic Information Sharing and Analysis Center (ISAC)[115] on artificial intelligence.

Develop a compliance-criteria catalog for service providers

Governments – in cooperation with industry – should develop a compliance-criteria catalog for service providers across the machine-learning supply chain. Government agencies would develop and publish such a catalog, such as the German "cloud computing compliance criteria catalog" (C5)[116], and (financial) auditing companies engaged by the service providers would attest the compliance. This compliance attestation would be only for the parts specific to machine learning, where the underlying general-purpose IT infrastructure could be certified by existing standards like ISO 27001. There are multiple advantages of compliance examination via criteria catalogs. First, compliance can be attested across the entire machine-learning supply chain (including the information-security level of the systems that handled the data that was externally acquired and used for the training). Second, this way of attesting compliance allows for innovation, also by smaller companies, as it is modular and only the machine-learning part has to be attested, not the underlying infrastructure. When a party wants to run machine-learning tools on an already-compliant cloud provider, it only needs to seek compliance audits for the tool and not for the entire infrastructure.[117] Thirdly, compliance-criteria catalogs are flexible (as compared to certifications), scalable, and can be forked to include additional security requirements for different environments. Compliance is attested a posteriori and repeated following the general accountability-reporting schedule of the company but

115 European Union Agency for Cybersecurity (2020): Information Sharing and Analysis Centers (ISACs)

116 Bundesamt für Sicherheit in der Informationstechnik (2020): Criteria Catalogue C5

117 Bundesamt für Sicherheit in der Informationstechnik (2020): C5:2020: SaaS-Fallstudie

needs to be repeated at least once in twelve months. Due to the very broad range of machine-learning application areas, having such a customizable framework is extremely useful, though the high end of risk environments should potentially be certified through other existing frameworks.[118]

Foster machine-learning literacy across the board

Having a general idea about the workings and use cases of a technology is a crucial element for making informed decisions on all levels – and machine learning is no different[119]. Academia, industry, and government should, therefore, create a tangible environment for a better understanding of machine learning. That can include a showroom in a technical or research institution or agency, such as DARPA or the German Cyberagentur. Decision-makers could visit the showroom and have staff explain to them various machine-learning applications, what they do, how they were developed, and, of course, what risks are associated with them.

Everyone who works with machine learning models directly – including everyone involved in the machine-learning supply chain – should receive basic training to develop a baseline skill set specific to machine learning that includes an understanding of the security implications. The development and implementation of such training could be handled by industry or academia, where the employer would be responsible for their staff to receive it.

For the broader public, it would be important to have simple language learning products, for example, those provided by institutions such as the German Bundeszentrale für politische Bildung or universities. This could help them to understand machine learning on a very abstract level, decrease reservations toward machine learning as users of its applications, and better understand possible challenges.

118 For the challenges of certification and artificial intelligence, see for example: Leonie Beining (2020): Vertrauenswürdige KI durch Standards?

119 Michael C. Horowitz and Lauren Kahn (2020): The AI Literacy Gap Hobbling American Officialdom and Daniel Eichler and Ronald Thompson (2020): 59 Percent Likely Hostile

7. Conclusion

Developing a machine-learning model – from acquiring the first pieces of data to the fully-trained model – requires a finely-tuned division of labor involving many services across the machine-learning supply chain. Each individual stage is part of the machine-learning attack surface and, therefore, involves a number of vulnerabilities and potential attack vectors[120]. Furthermore, it is important to acknowledge that the conventional attack surface of IT products and systems still applies to machine-learning models, as its supply chain heavily relies on general-purpose hardware and software, which further exacerbates the security challenges. It is imperative to design and implement policies that lead to a holistic security approach for protecting the machine-learning supply chain.[121] A thorough risk assessment for deployment is vital to correctly assess high-risk environments and the impact of factors such as scale and delayed effects. Thus, the following policies and actions should be considered by decision-makers:


Design a security approach rooted in conventional information security


Increase transparency, traceability, validation, and verification


Identify, adopt, and apply best practices

Require fail-safes and resiliency measures


Create a machine-learning security ecosystem

Set up a permanent platform for threat exchange

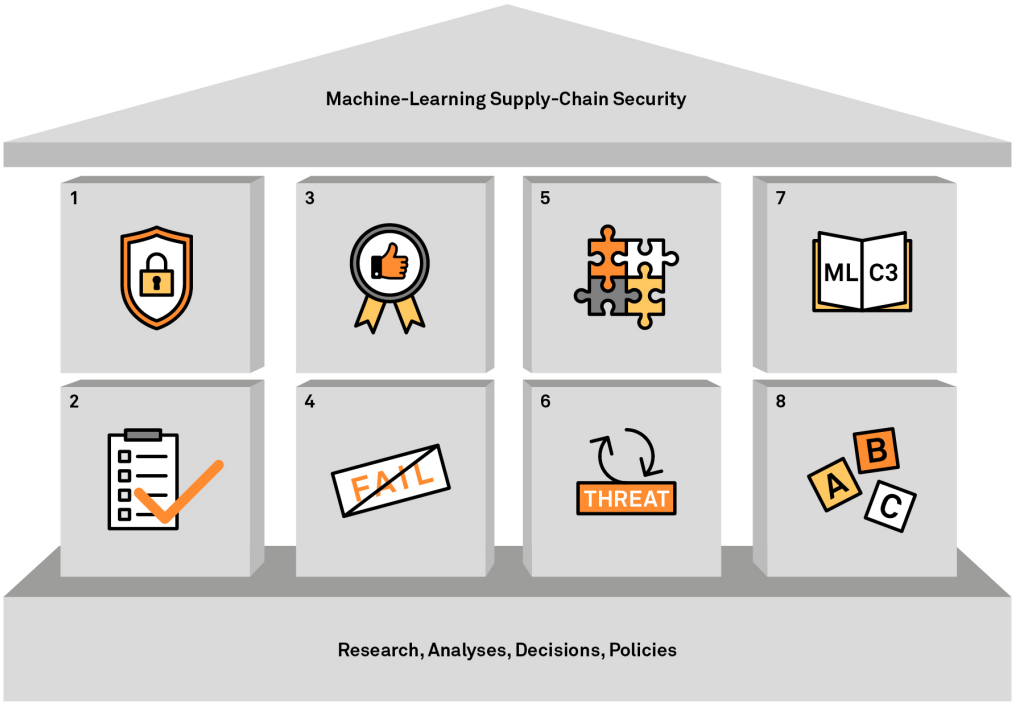
Develop a compliance-criteria catalog for service providers

Foster machine-learning literacy across the board

120 Sven Herpig (2019): Securing Artificial Intelligence

121 Considering the wide range of possible use cases for machine-learning applications – especially for critical infrastructures, public safety and national security – information security will not only be a limiting factor but may, in fact, be an economic opportunity for states and companies focusing on how to better secure artificial intelligence.

Machine-Learning Supply Chain: "Security Recommendations" (Figure 8)



We – decision-makers, researchers, and experts – urgently have to avoid repeating the mistakes of the past, where security has only been an afterthought of technological development and deployment. If we do not, the impact on our societies may be more severe than ever before.

ANNEX: Cybersecurity and Artificial Intelligence Glossary

Special thanks to Kate Saslow for her contribution to the first version of the glossary. The up to date glossary can be found here. A complementary taxonomy was developed by Microsoft Corporation in cooperation with the Berkman Klein Center for Internet and Society at Harvard University.[122] An adaptation to the MITRE ATT&CK framework is also being developed by contributors from several organizations.[123]

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."[124]

Adversarial Examples: "[I]nputs formed by applying small but intentionally worst-case perturbations to examples from the dataset, such that the perturbed in-put results in the model outputting an incorrect answer with high confidence."[125]

Adversarial (Machine) Learning: "Adversarial machine learning is a game against an adversar- ial opponent (Huang et al. 2011; Lowd and Meek 2005) who tries to deceive the algorithm into making the wrong prediction by manipulating the data. This deception occurs in two ways: [temporal drift and adversarial drift]."[126]

Application-Programming-Interface (API): "An API acts as an intermediary between your application and a third-party service. [...] Thus, an API for machine learning can be defined as a remote tool utilizing ML to solve a specific problem within a specific project."[127]

122 Ram Shankar Siva Kumar, David O'Brien, Kendra Albert, Salome Viljoen and Jeffrey Snover (2019): Failure Modes in Machine Learning

123 Keith Manville et al. (2020): Adversarial ML Threat Matrix

124 Myriam Abramson (2015): Toward Adversarial Online Learning and the Science of Deceptive Machines

125 Ian Goodfellow, Jonathon Shlens and Christian Szegedy (2015): Explaining And Harnessing Adversarial Examples

126 Myriam Abramson (2015): Toward Adversarial Online Learning and the Science of Deceptive Machines

127 Helen Kovalenko (2020): Choosing the Best Machine Learning API for Your Project

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.

Binarized Neural Network (BNN): "A BNN works with binary weights and activation values.This is our starting point as the implementations of such networks have similarities with the implementation of block ciphers. BNN reduces the memory size and converts a floating point multiplication to a single-bit XNOR operation in the inference. Therefore, such networks are suitable for constrained IoT nodes where some of the detection accuracy can be traded for efficiency."[128]

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.

CIA (Triad): Stands for credibility, integrity and availability, a common framework to assess information security.[129]

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").[130]

128 Anuj Dubey, Rosario Cammarota and Aydin Aysu (2019): MaskedNet: The First Hardware Inference Engine Aiming Power Side-Channel Protection

129 Chad Perrin (2008): The CIA Triad

130 Sidath Asir (2018): Machine Learning Classifiers

Convolutional Neural Network: "Convolutional Neural Networks (CNN) are special types of DNNs with sparse, structured weight matrices. CNN lay- ers can be organized as 3D volumes, as shown in Figure 2. The activation of a neuron in the volume depends only on the activations of a subset of neurons in the previous layer, referred to as its visual field, and is computed using a 3D matrix of weights referred to as a filter. All neurons in a channel share the same filter. Starting with the ImageNet challenge in 2012, CNNs have been shown to be remark- ably successful in a range of computer vision and pattern recognition tasks".[131]

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".[132]

Data Types (for Machine Learning): "Because of this, when constructing a machine-learning classifier, data is partitioned into three sets: training data, used to train the classifier; validation data, used to measure the accuracy of the classifier during training; and test data, used only once to evaluate the accuracy of a final classifier"[133]

Deep Neural Networks: "Deep learning is the family of neural networks composed of an input layer, three or more hidden layers and an output layer. Based on the internal structure, several can- didates exist like multi-layer perceptron (MLP), convolutional neural networks (CNN), recurrent neural network (RNN) etc. These are popularly known as deep neural net- works (DNN)"[134]

131 Tianyu Gu, Brendan Dolan-Gavitt and Siddarth Garg (2019): BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain

132 Nicolas Papernot and Ian Goodfellow (2016): Breaking things is easy

133 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos and Dawn Song (2019): The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

134 Jakub Breier, Xiaolu Hou, Dirmanto Jap, Lei Ma, Shivam Bhasin and Yang Liu (2018): DeepLaser: Practical Fault Attack on Deep Neural Networks

Deep Neural-Network Backdoor: "A backdoor is a hidden pattern injected into a DNN model at its training time. The injected backdoor does not affect the model's be- havior on clean inputs, but forces the model to produce unexpected behavior if (and only if) a specific trigger is added to an input"[135]

- **Latent Deep Neural-Network Backdoor:** "Latent backdoors are incomplete backdoors embedded into a "Teacher" model, and automatically inherited by multiple "Student" models through transfer learning. If any Student models include the label targeted by the backdoor, then its customization process completes the backdoor and makes it active"[136]

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.

Fault Attack: "[A]ctive attacks on a given implementation which try to perturb the internal software computations by external means. The adversary uses methods like voltage glitches or laser injection to introduce perturbations for various purposes."[137]

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 process".[138]

135 Yuanshun Yao, Huiying Li, Haitao Zheng and Ben Y. Zhao (2019): Latent Backdoor Attacks on Deep Neural Networks

136 Yuanshun Yao, Huiying Li, Haitao Zheng and Ben Y. Zhao (2019): Latent Backdoor Attacks on Deep Neural Networks

137 Jakub Breier, Xiaolu Hou, Dirmanto Jap, Lei Ma, Shivam Bhasin and Yang Liu (2018): DeepLaser: Practical Fault Attack on Deep Neural Networks

138 Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal and Seraphin Calo (2019): Analyzing Federated Learning through an Adversarial Lens

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.[139]

Ground Truth: Is used in supervised learning as a human-led observation – not inference – defining unperturbed categorical and numerical input data and labels in a dataset as preparation for the training. The ground truth is subjective and depends on the individual observer's perception of the data.

Hardware Attacks: Attacks being carried out against hardware leveraged for machine learning in any stage. The targets can be general purpose hardware such as graphics processing units (GPUs)[140], or specialized hardware such as application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), or Neural Processing Units (NPUs) such as Tensor Processing Units (TPUs). Attacks include side-channel attacks[141] such as Differential Power Analysis[142].

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".[143]

Libraries: "Libraries, in turn, are more highly specialized tools. As a rule, they are tied to the ability to solve a specific problem in a certain environment and require additional coding skills to make their use effective".[144]

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.

139 Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley,Sherjil Ozair, Aaron Courville and Yoshua Bengio (2014): Generative Adversarial Networks

140 Michael Kissner (2019): Hacking Neural Networks:A Short Introduction

141 Lejla Batina, Shivam Bhasin, Dirmanto Jap and Stjepan Picek (2018): CSI Neural Network: Using Side-channels to Recover Your Artificial Neural Network Information

142 Anuj Dubey, Rosario Cammarota and Aydin Aysu (2019): MaskedNet: The First Hardware Inference Engine Aiming Power Side-Channel Protection

143 National Institute for Standards and Technology (2020): Glossary

144 Helen Kovalenko (2020): Choosing the Best Machine Learning API for Your Project

Machine-Learning Application: Deploy-ready software that leverages machine-learning models.[145]

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 include 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 learning 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

145 For intersections between the machine-learning part and the non-machine-learning part of a software, and their security implications, see for example: Skylight (2019): Cylance, I Kill You!

146 Thomas Simonini (2018): Curiosity-Driven Learning made easy Part 1

147 Prince Yadav (2018): Decision Tree in Machine Learning

148 Francois Chollet (2018): Deep Learning with Python

149 Raul Eulogio (2019): Introduction to Random Forests [source removed]

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.

Machine Learning as a Service (MLaaS): "Machine learning as a service (MLaaS) is a range of services that offer machine learning tools as part of cloud computing services, as the name suggests. MLaaS providers offer tools including data visualization, APIs, face recognition, natural language processing, predictive analytics and deep learning. The provider's data centers handle the actual computation"[151]. A user of such a cloud computing service (e.g., Amazon Machine Learning or Microsoft Azure Machine Learning) could potentially attack other users on the same platform.[152] On the other hand, users also depend on decisions made by MLaaS platform providers.

Machine Learning – Information Security Intersections: "There are three main intersections between machine learning and informa-tion security: 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"[153]

Machine-Learning Supply Chain: Data, tools and services as well as (specialized) software and hardware required to develop a machine learning model.

Membership Inference: Attacking a deployed model, using specially crafted adversarial examples to infer whether certain training points were used for training a model.[154]

Memorization: "[...] rare or unique training-data sequences are unintentionally memorized by generative sequence models—a common type of machine-learning model"[155]

150 Francois Chollet (2018): Deep Learning with Python

151 Technopedia (2019): Machine Learning as a Service (MLaaS)

152 Binghui Wang and Neil Zhenqiang Gong (2018): Stealing Hyperparameters in Machine Learning

153 Sven Herpig (2019): Securing Artificial Intelligence

154 Nicolas Papernot and Ian Goodfellow (2016): Breaking things is easy; Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov (2017): Membership Inference Attacks Against Machine Learning Models

155 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos and Dawn Song (2019): The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

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

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".[156]

Model Extraction: Interfering with a model to "search for a substitute model with similar functionality as the target neural architecture"[157] in order to be able to replicate it.

Model-Extraction Attack: "These attacks aim to steal parameters of an ML model. Stealing model parameters compromises the intellectual property and algorithm confidentiality of the learner, and also enables an attacker to perform evasion attacks or model inversion attacks subsequently"[158]

Model Inversion: Interfering with a model to derive/extract the training data from it.[159]

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".[160]

Neural Cleanse: Using various techniques such as input filters, neuron pruning and unlearning to mitigate backdoors and their trigger in Deep Neural Networks.[161]

156 Luigi (2019): The Ultimate Guide to Model Retraining

157 Vasisht Duddu, Debasis Samanta, D. Vijay Rao and Valentina E. Balas (2019): Stealing Neural Networks via Timing Side Channels

158 Binghui Wang and Neil Zhenqiang Gong (2018): Stealing Hyperparameters in Machine Learning

159 Matt Fredrikson, Somesh Jha and Thomas Ristenpart (2015): Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures

160 Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal and Seraphin Calo (2019): Analyzing Federated Learning through an Adversarial Lens

161 Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng and Ben Y. Zhao (2019): Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks

Neural Network: A neural network (NN) is 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.[162]

Neural-Network Trojaning: Manipulating a Neural Network in a way, that a trigger input causes a predefined action chosen by the adversary.[163]

Online (Machine) Learning/Incremental Learning: A machine learning model that while being deployed "can learn from new examples in something close to real time"[164], 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 trains on the new data".[165]

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)[166].

Physical Perturbation: Perturbation of physical objects (e.g. sticker on a stop sign)[167].

Regressor: A regressor is an algorithm that "can predict output values not seen during the training process"[168].

Side-Channel Attacks Assisted with Machine Learning (SCAAML): Side-channel attacks leveraging deep learning.[169]

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

---

162 Philippe Lorenz and Kate Saslow (2019): Demystifying AI & AI Companies

163 Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee and Juan Zhai (2017): Trojaning Attack on Neural Networks

164 Max Pagels (2018): What is online machine learning?

165 Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee and Juan Zhai (2017): Trojaning Attack on Neural Networks

166 Ian Goodfellow, Nicolas Papernot, Sandy Huang, Rocky Duan, Pieter Abbeel and Jack Clark (2017): Attacking Machine Learning with Adversarial Examples

167 Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song (2018): Robust Physical-World Attacks on Deep Learning Visual Classification

168 Apple (2020): MLRegressor

169 Elie Burszstein and Jean-Michel Picod (2019): A hacker guide to deep-learning based AES side channel attacks

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."[170]

Test Data: "Used only once to evaluate the accuracy of a final classifier."[171]

Threat Model: "a formally defined set of assumptions about the capabilities and goals of any attacker who may wish the system to misbehave."[172] 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"[173] 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.

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."[174]

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".[175] "During this process, customers take public "teacher" models and repurpose them with training into "student" models, e.g. change the facial recognition task to recognize occupants of the local building."[176]

170 Myriam Abramson (2015): Toward Adversarial Online Learning and the Science of Deceptive Machines partially referencing Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar (2012): Foundations of Machine Learning

171 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos and Dawn Song (2019): The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

172 Nicolas Papernot and Ian Goodfellow (2016): Breaking things is easy

173 Vasisht Duddu, Debasis Samanta, D. Vijay Rao and Valentina E. Balas (2019): Stealing Neural Networks via Timing Side Channels

174 Deyan V. Petrov and Timothy M. Hospedales (2019): Measuring the Transferability of Adversarial Examples

175 Jason Brownlee (2017): A Gentle Introduction to Transfer Learning for Deep Learning

176 Yuanshun Yao, Huiying Li, Haitao Zheng and Ben Y. Zhao (2019): Latent Backdoor Attacks on Deep Neural Networks

Untargeted Attack: "An untargeted attack only aims to reduce classification accuracy for backdoored inputs; that is, the attack succeeds as long as backdoored inputs are incorrectly classified."[177]

Validation Data: "Used to measure the accuracy of the classifier during training."[178]

Validation Gap: Decisions taken by a (machine learning model) system based on a single source (e.g., a sensor) without validating them against a second source (e.g., other sensors or vehicle-to-infrastructure communication).[179]

177 Tianyu Gu, Brendan Dolan-Gavitt and Siddharth Garg (2019): BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain Tianyu

178 Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos and Dawn Song (2019): The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

179 Ben Nassi, Dudi Nassi, Raz Ben-Netanel, Yisroel Mirsky, Oleg Drokin and Yuval Elovici (2020): Phantom of the ADAS:Phantom Attacks on Driver-Assistance Systems

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) was established by the Berlin based think tank Stiftung Neue Verantwortung (SNV) in January 2017.

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 the director for international cyber security policy at Stiftung Neue Verantwortung. His focal areas include information security of machine learning, (geopolitical) responses to cyber operations, government hacking and vulnerability management, and Germany's cybersecurity policy. Before Sven joined the Stiftung Neue Verantwortung, he was employed by Germany's federal government for several years.

Contact the Author

Dr. Sven Herpig
Director for International Cybersecurity Policy
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

Grafik:
Anne-Sophie Stelke

Layout:
Jan Klöthe