Translation of article originally published in:
Datenschutz und Datensicherheit - DuD, 43(7), Juli 2019, Springer Fachmedien, Wiesbaden 2019, ISSN 1614-0702, pp 434-439

Eberhard von Faber, Arndt Kohler

The gap: information security in systems with artificial intelligence

How algorithms and artificial intelligence can pose a threat to IT security

All known measures and methods for IT security are based on the strategies of defining a target state, controlling the actual state and carrying out targeted, corrective intervention. But these are not the only assumptions for functional IT security. This article analyzes these assumptions and the problems that arise through the use of machine learning and artificial intelligence, which can hamper our conventional security measures and become weapons in their own right.

1 Outlining the topic

The world is full of threats, most of which are addressed by “security” measures of one kind or another. Figure 1 shows some of the threats involved. The column on the left lists sources of threats, and the top row lists targets of threats.

In general, IT experts are interested in all types of cases covered by the two rows and the two columns with “IT” in their labels. In the figure, those rows and columns form a cross. If we leave aside the two boxes marked “Hollywood” and the one marked “Accident”, eight categories remain. The six boxes bearing the label “Quality” are especially important. Unfortunately, relatively little interest in the relevant categories has been shown, even as the category “IT systems threaten people” should deserve increasing attention. That category is closely related to the category “People threaten IT systems”. In fact, the two categories can build on one another, meaning that the matrix could actually be three-dimensional. We note that the “Zero Outage Industry Standard” association [1] focuses on the topic “quality in IT, including IT security”. The present article considers the two categories labelled “IT security”, especially the one highlighted in **boldface**. In addition, we will also separately consider “IT with artificial intelligence” as an attack weapon and a tool for defense.

Prof. Dr. Eberhard von Faber

T-Systems, Chief Security Advisor, IT Division; working areas: security architecture, developer of ESARIS, secure IT production, secure IT outsourcing, process and ITIL integration, standardization, cloud, IAM; E-Mail: Eberhard.vonFaber@th-brandenburg.de

Arndt Kohler

IBM, Head of IoT Security, Security Division;
working areas: Internet of Things Security, Operational Technologies Security, Security Consulting & Architecture, Security Operation
E-Mail: Arndt.Kohler@de.ibm.com

Figure 1: Threats and work areas

...threaten → ↑	People	IT systems / data	IT with artificial intelligence	Machines
People	War/ terrorism	IT security	**IT security**	Sabotage/ terrorism
IT systems	Quality/ IT security (Safety)	Quality	Quality	Quality
IT with artificial intelligence	(Hollywood)	Quality	(Hollywood)	Quality
Machines	Safety	Accident	- -	Cyberwar

That said, the box marked in **boldface** is especially important. The present article focuses neither on the legal dimensions of artificial intelligence (cf. [2] in this regard) nor on the legally compliant, responsible use of algorithms (cf. [3]). This article is about technology – about IT security in the narrow sense of the term.

The article analyzes the bases for existing IT-security solutions and outlines the problems that can arise via use of algorithms and artificial intelligence in business applications and by attackers. Will it be necessary to view the area covered by "artificial intelligence" as a "black box" and to return to perimeter protection? The present article cannot provide any final answers to such questions. Its primary aim is to call attention to potentially critical, fundamental difficulties. With the article, the authors present their hope that IT security will take the steps now required – just as it has done again and again in the past 50 years – and develop solutions for protection of "intelligent, algorithm-controlled IT"**.**

2 The current situation

2.1 What are machine learning and artificial intelligence?

In a narrow sense, the term "artificial intelligence" (AI) refers to systems that use (supposedly) human-like decision-making structures in non-clear-cut environments. AI computers are programmed to solve problems and make decisions autonomously.

In this context, the present article focuses especially on processes based on "machine learning" (ML), which refers to experience-guided knowledge production by machines. An ML system learns from examples (learning data) and, after completing a learning phase, can generalize what it has learned, i.e. apply its learning to new constellations (user data). An ML system "recognizes" patterns and rules in learning data and applies them in processing user data.

Such systems learn by "seeing" results, not criteria (patterns, rules). They extract relevant criteria themselves, thereby carrying out learning and developing "intelligence". One strategy for creating such systems consists of using neural networks. From large numbers of input parameters, neural-network systems produce smaller numbers of output signals. They do this by weighting their input signals (via parameters) and linking all input signals and output signals via simple calculation rules.[1]

As a system learns, it determines values for all its parameters. Even in cases of only moderate complexity, the number of parameters involved can exceed 100 million. In fact, such systems of parameters are actually enormous systems of equations that no human could ever grasp in their entirety. As a result, no human can know exactly what criteria, patterns and rules such a system will use in obtaining its results. One cannot even know which input parameters are most important. One sees only the results a system puts out.

And such results are sometimes obviously off the mark, indicating that some operation has gone wrong. In Amazon's experiments with AI as a tool for screening job applicants, for example, women were systematically disadvantaged. The problem turned out to lie in the data used for teaching the system; women were underrepresented in those data. The system's artificial intelligence saw such underrepresentation as a relevant criterion.

In general, artificial intelligence can produce decisions that average out to correct results, while being completely wrong in individual cases. Artificial intelligence systems lack humans' ability to "understand" meaning the ability to recognize cause-and-effect relationships and use them as a basis for decisions. Such systems filter statistical correlations, which are not the same as cause-and-effect relationships. AI systems have no grasp of such relationships. For the above reasons, it can be questioned whether such systems are truly "intelligent". In the following, we use "artificial intelligence" simply as a technical term, free of any value judgements with regard to the real meaning of intelligence. We also use the common abbreviation "AI".

The underlying methods for AI emerged some time ago. Use of AI technologies has been booming, however, because 1) the performance and availability of computing and storage systems have dramatically improved; and 2) in many areas that use "big data" enormous quantities of data have become available for training of AI systems.

2.2 Selection and implementation of IT-security measures today

The following section explains how security measures are selected without using AI. Working from that explanation, Section 2.3 then describes the foundations for today's IT security systems. With an understanding of those foundations, we then turn to the difficulties involved in using AI to provide security for IT systems.

The question of how to adequately protect an IT system turns on the question of what security measures and security solutions need to be integrated within the system, and where. That question is not a trivial one. To answer it, one cannot simply pay lip service to a "risk-based approach" and an optimal "cost-benefit ratio". One must actually analyze where risks arise, and how, and which security measures are expected to generate costs. The term "security concept" gives a better sense of the basic approach that is required. That said, success in finding the right connection to the IT system in question, in a given case, often depends on the involved expert's own knowledge and intuition.

Figure 2 shows a conceivable approach. In the following, we outline this approach, calling attention to possible problems when artificial intelligence is involved. Details about the procedure are provided in the book *Joint Security Management (JSM)* [4].

The left side lists well-known steps such as creating a list of threats, identifying assets, classifying and assigning threats, assessing risks, preparing an overview and making final decisions.

[1] Each calculation rule amounts to a "neuron with parameters". The neurons form a layer. In "deep learning" (DL) systems, the connection between inputs and outputs consists of not just one level or layer of neurons, but of several layers arrayed in series.

The point "understanding the IT infrastructure and the IT components", an especially important task, presents a first potential problem. A company's IT systems contain important assets (in the form of data and IT services), and potential IT-security vulnerabilities can result from deficiencies in the design and protection of those systems. Utilization of artificial intelligence, in the manner described below in detail, places limits on the extent to which the company's IT infrastructure and IT components can be understood by the responsible experts.

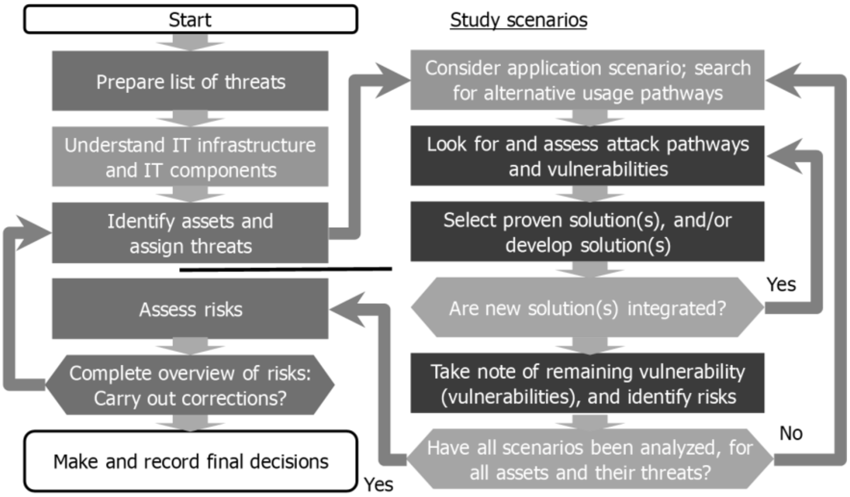
The right side of Figure 2 lists steps for analyzing whether assets are exposed to actual risks and, if so, what risks are involved. A threat does not become a risk until it becomes able (with a certain level of probability) to exploit a vulnerability. To search for vulnerabilities (or possibilities for attacks), one must analyze potential flows of information. In addition to being shaped by IT systems' architectures, such flows are influenced by the ways in which IT components are configured. By studying such flows, one learns what functionalities an attacker would find. In addition, one must search for additional use scenarios and pathways, since attackers will naturally attempt to use functionalities in ways unintended by their rightful users. In the great majority of cases, IT experts focus on these aspects and integrate security solutions accordingly. In other cases, systematic searches for still other vulnerabilities have to be carried out. Such analysis and searches provide a basis for selection of proven IT-security solutions – best practices – that eliminate vulnerabilities. The search for attack pathways and vulnerabilities continues even after such solutions have been put in place, however.

The figure shows how such searches proceed, and how vulnerabilities can be gradually eliminated via integration of IT-security solutions. This approach calls for

- an understanding of IT systems' infrastructures and components,
- an understanding of potential information flows (including all attack pathways), and
- a means of intervening efficiently and of eliminating identified vulnerabilities by integrating IT-security solutions.

When artificial intelligence enters the security equation, these three prerequisites are fulfilled only in part or not at all. To understand why this is so, let us assume that, in a given case, algorithms and artificial intelligence are used. They then are part of the relevant IT system. They are components whose inner modes of operation, by definition, can be understood only partly or not at all. Furthermore, they autonomously control application scenarios – and, especially, the intended information flows – in ways that security analysts cannot predict and cannot understand after the fact. This is not surprising, given that they are an independently functioning "intelligence". In addition, and clearly enough, security analysts have no way of influencing the information flows – that is the task of the algorithms and artificial intelligence being used.

Figure 2: Risk-oriented procedure for selection of security solutions


Start
Prepare list of threats
Understand IT infrastructure and IT components
Identify assets and assign threats
Assess risks
Complete overview of risks: Carry out corrections?
Make and record final decisions
Yes
Study scenarios
Consider application scenario; search for alternative usage pathways
Look for and assess attack pathways and vulnerabilities
Select proven solution(s), and/or develop solution(s)
Are new solution(s) integrated?
Yes
Take note of remaining vulnerability (vulnerabilities), and identify risks
Have all scenarios been analyzed, for all assets and their threats?
No

2.3 The basis for IT security today

All known measures and methods for IT security are based on the strategies of defining a target state, controlling the actual state and carrying out targeted, corrective intervention. There are additional requirements (and important details) to note. Three are outlined in the following.

1. Known information flows (throughout the entire system): The IT infrastructure and the application scenarios it supports (desired information flows, etc.) must be understood. Armed with such an understanding, an IT-security analyst can begin to look for attack pathways and vulnerabilities and to assess them in terms of the likelihood of being exploited. On the basis of such analysis, security solutions will be selected, as necessary, that can eliminate identified vulnerabilities and thereby reduce risks or eliminate them entirely. If an artificial intelligence system autonomously controls information flows, it will be difficult to differentiate intended and functionally related information flows from suspect or obviously hostile information flows. It will also be difficult to interrupt specific information flows in order to prevent them from leading to security breaches. In short, it will be difficult to define the target state with respect to information flows.
2. An understanding of the IT functionality (of an IT object): The target state that one needs to have clearly in view includes the target state of the IT system as a whole – including, in particular, the system's software components and applications. It also includes the state and nature of the system's functions (i.e. of the information flows to and from the IT object and within it). System software differs from malicious software in terms of its functionality. Often, software is assessed, with regard to being suspect or hostile, on the basis of its origins. Where assessments on the basis of origins are not readily feasible, decisions as to whether specific software is benign or hostile, have to be based on experience or on analyses of the software's functionality. Other vulnerabilities are eliminated by controlling and restricting access with the help of centrally managed digital identities. Similar results

can be achieved by encrypting data and by using integrity-assuring measures such as signatures. Vulnerabilities can also be eliminated by filtering and replacing data and commands. Measures such as access control, encryption, integrity assurance and (manipulative) filtering can be applied only if the desired IT functionality of the system (IT object) in question is fairly well understood. A key aspect of artificial intelligence is that its mode of function is not transparent or easily understood. As a result, some of the IT-security solutions commonly used today can be expected to become less effective or become subject to restrictions on their use.

3. Confrontation / duel situation: An attacker and a defender face off (in a duel). The attacker attempts to overcome or find gaps in security measures that the defender has put in place and manages. The attacker's aim is to violate the confidentiality, integrity and/or availability of assets, or simply to prepare additional steps for attacks. In the case of systems with artificial intelligence, attackers and defenders will not necessarily face off. They may not even interact. An attacker may, for example, influence or manipulate users or their user data. Such data may be used by the defender's AI-based security system for learning purposes, however. By tampering with such data, an attacker can set up cover for a latter attack. In such a case, therefore, an attacker – to use a metaphor from the world of pool – tries a "bank shot".

3. The situation when artificial intelligence is included

3.1 The basis for use of artificial intelligence

In general, an AI system will function well only when its training data conform qualitatively, i.e. in terms of their distribution, with the user data used by the relevant organization. In addition, the training data and the user data must both have the characteristic(s) that filtering mechanisms are set up to find. These findings lead to the requirements pertaining to any successful use of AI that are described below. Significantly, the requirements call attention to potential problems and vulnerabilities to attack:

1. Stability: This means that the IT system's state should not change rapidly. Changes require the AI system to be retrained with new data.
2. Integrity of learning data: The integrity of learning data must be assured, meaning attackers must have no way of tampering with learning data. Attackers can carry out such tampering, for example, by influencing users. In addition, they can strive to interfere with the learning phase, thereby enabling themselves to mask their later actions.
3. Integrity of the learning process: It must be impossible for attackers to target and tamper with the learning process by planting harmful examples. In this context, it must be remembered that the way AI systems "perceive" things differs significantly from the way human beings perceive them. Just a few pixel changes to an image of a face can cause an AI system to see the face as a car [5]. Human beings have no difficulty interpreting images correctly after such changes.
4. Labelling: Often, the data required for effective training of an AI system are lacking. Where complex IT systems are to be protected, the data that are truly required for such training can be difficult to identify and parametrize. During a training process, many questions have to be answered, such as the following: How are examples for the "benign" case defined? What possibilities are available for creating such "benign" cases and making them available for the training process, in keeping with the aim of obtaining the best-possible results as quickly as possible? What framework conditions, including both "hard" and "soft" conditions, should be specified for such an AI system?

3.2 IT-security problems arising through use of AI

Let us imagine a complex AI system, in the IT sector, that can autonomously learn and make decisions in the framework of defined parameters. Over time, usage of IT resources changes across various layers of the Open Systems Interconnection (OSI) model. Interactions between IT elements adjust accordingly. As this occurs, transparent justifications for these changes are not necessarily provided, nor is proof of the integrity of the learning data and the learning process necessarily developed.

Detection of an advanced persistent threat (APT)[2] within the changes would be possible only if the changed behavior patterns had occurred in a comparable AI environment in the past – and had been successfully analyzed and documented. As a result, AI systems are normally unable to detect APTs [6], in contrast to what numerous marketing claims would have one believe. The AI-based security system in our present example was trained with data from the past. This means it will be able to detect an APT only if the relevant behavior pattern occurred in the past and was identified as hostile. Once an attacker changes his behavior pattern, or tries a completely new mode of attack, the AI-based security system will be unable to detect his actions.

In an IT environment controlled by an AI system, active security systems such as firewalls and identity/access management can quickly begin to brake the AI system and hamper the operation of the IT system as a whole. Administrators who manage the security systems may be unable to anticipate required changes and adjust rules accordingly.

Conventional SIEM systems[3] also quickly run up against their limits in connection with complex AI systems. Anomalies are defined in terms of changes in IT systems' behavior – and, ultimately, any decision of an AI system will amount to such a

[2] APT: an advanced, persistent threat – in practice, an attack of great complexity and long duration

[3] SIEM: Security Information and Event Management; Security Information Management (SIM) comprises the collection and analysis of historical data such as log data (including log-management) and (automatic) monitoring of fulfillment of compliance regulations (such as regulations on hardening of systems and installation of patches and updates). Security Event Management (SEM) comprises real-time monitoring of ICT systems and analysis of incidents, including alerts. SIEM, a combination of these two types of management, supports incident management with a view to cyber defense activities.

change. The same applies to IDS/IPS systems.[4] Both SIEM and IDS/IPS systems, therefore, would be likely to lose their effectiveness. Currently, such systems' effectiveness already depends on how well they are adjusted – for example, in connection with installation of new applications – to user-/operator-changed application scenarios and data streams. When application scenarios and data streams are increasingly controlled automatically by algorithms and artificial intelligence, with no human intervention, the bases for adapting such solutions to IT-security requirements begin to disappear.

The most practicable approach is to limit IT security to protection of the AI system's interfaces with its environment – and thereby abandon any attempt to monitor activities within the AI system. Will it be necessary to view areas controlled by artificial intelligence as "black boxes" and to return to perimeter protection? That is not an attractive prospect, and it cannot be the way to protect complex AI systems.

3.3 Case study: Production planning and -control

A concrete example can usefully illustrate these considerations. In many industries, the areas of production planning and control present a special challenge. The next industrial revolution (Industry 4.0) promises to deliver custom-engineered production within the framework of series production. This is already being achieved to a considerable extent in the automotive industry, for example. Automakers seldom produce two cars that are completely alike, in all aspects, within one production year. This is due to the great numbers of accessories, feature combinations and model versions that they offer. Behind each accessory stands a supplier with its own logistics system. All parts and accessories for a car have to arrive at the assembly line in a defined, tightly scheduled sequence – otherwise, the car cannot be produced.

Coordination of such deliveries is an enormous task that can quickly prompt calls for use of AI in planning and control. A complex system consisting of production units, movements of goods, and a human factor is a system with low stability. Furthermore, commercial systems have diverse ranges of interfaces, such as interfaces to procurement, controlling, etc. From a technical, IT-based perspective, such systems are mixtures of widely diverse components, including physical machines and transport units, control systems and IT applications.

An AI system with such a broad area of operation is a highly critical asset, and thus a valuable target. An attack against it can lead to more than simply production downtime or defective products. It can also threaten the security and safety of the company's employees and users. Simply safeguarding the AI system's interfaces with respect to the outside world is not an effective solution. The AI system itself has to be protected. This has to include monitoring its learning data and its learning processes. In addition, the sphere in which an AI system operates must also receive adequate security. If some existing security components prove ineffective, one should consider the option of replacing both them and many others, with a focus on the entire AI system. This means providing security upgrades for areas that currently have few or no security components. For example, physical production and transport units can be upgraded so that they become more resistant to possible malfunctions.

When an AI system is used in this context, the basic principle to the effect that cyberattack occurrence is a matter of "when", and not of "whether", applies to both the AI system and all its components. For this reason, we have to consider how to minimize, or even negate altogether, the impacts of attacks. Ideally, such responses will not have to include a complete system reset.

4. Improving IT security with the help of AI

Could use of artificial intelligence improve IT security to at least some extent?

Artificial intelligence and "smart" algorithms are already being used today, in many areas, to improve IT security. One example is their use for detection of malware. That said, it should be noted that such systems – initially, at least – failed to detect the WannaCry ransomware cryptoworm. The good news is that good implementations can automatically and reliably detect most malware [5]. Good implementations also produce false alarms ("false positives"), however. In particular, company-specific, seldom-used software tends to be identified as false positives, because its profiles are not normally included in learning data. Furthermore, the rapid development of malware makes it necessary for systems' anti-malware solutions to be regularly retrained.

SIEM solutions, which serve as the basis, and the infrastructure, for Security Operations Centers (SOC) and Cyber Defense Centers (CDC), are key applications. SIEM solutions collect, filter, normalize, correlate and analyze event data from many different sources, such as firewalls, IDS/IPS systems, network admission control (NAC) systems, anti-malware systems, data leakage protection (DLP) systems and authentication services. Artificial intelligence and "big data" solutions provide valuable support for SIEM systems. AI and big data solutions, like IDS/IPS systems, have to be regularly adapted to, and retrained for, any changes in the IT systems they support.

Where the relevant data are available, AI systems and "smart algorithms" can improve access management. In such applications, access rights are dynamically supplemented with data that make it possible to block "illogical" access attempts. Interestingly enough, artificial intelligence and "smart algorithms" are also effective tools for fraud protection and identification of abusive behavior in social media, which are indirectly related to IT security. In sum, it must be remembered that all such applications necessitate the availability of learning data in keeping with the relevant target state for IT security. This brings up a basic question: Can data that we observe in real-world applications reflect any target state?

[4] IDS/IPS: Intrusion Detection/Prevention systems

6 Discussion and outlook

Use of algorithms and artificial intelligence in business applications creates problems in its own right. The established procedures for selection and implementation of IT-security solutions function only to a limited degree. For today's IT security systems to function well, the information flows in the systems they protect have to be well-known and relatively stable. Furthermore, system operators must understand the IT functionality of all IT objects involved, since many security solutions change or limit such functionality. In addition, in today's IT security arenas, attackers face off against defenders in the sense that they seek to overcome defenses. Overall, the analysis showed that the aforementioned requirements are not necessarily fulfilled when artificial intelligence is used. AI systems in such applications control information flows autonomously, in unpredictable ways. Their functional principles are inscrutable, and they may allow attackers to compromise IT security indirectly.

A closer look at the basis for use of artificial intelligence brings specific risks and attack possibilities to light. Such risks and possibilities were discussed both as basic problems and in light of a case study. This led to a proposal that is hardly encouraging: IT security should be limited to AI systems' interfaces with their environment, and impacts should be reduced by strengthening other peripheral measures. Neither perimeter protection, nor reduction of impacts (a typical approach for event management), is an imaginative or new idea.

As a result, the option of equipping IDS/IPS and SIEM solutions with algorithms and artificial intelligence could be considered. This approach is already being used today to some extent – for example, in cases in which the function of such solutions depends on big-data analyses. It hardly seems likely, however, that the “intelligence” of such security solutions would be so superior to the business-controlling “intelligence” they are linked with that they would be able to “understand” or control the work of that latter intelligence.

A similar conclusion could be drawn if attackers were to begin carrying out their attacks with the help of artificial intelligence. Conceivably, AI systems would enable attackers to mask and hide their attacks so effectively that neither IDS/IPS systems nor state-of-the-art SIEM systems would be able to detect them. All state-of-the-art cybersecurity-defense strategies – all of which are based on SIEM systems (in combination with threat intelligence) – would lose their advantage.

And what if nothing helps? Will we then use artificial intelligence in order to create false documents and company assets that we can mix with our real assets, in the hope that such mixing would protect real assets against attacks carried out with artificial intelligence? An attacking “intelligence” would then have to battle against two “intelligences”: the one seeking to expose it and the one seeking to hide assets via a “security by obscurity” strategy. Perhaps such false documents and company assets would then be seen as lures (such as “honey pots”), and monitored, with the help of additional sensors, by a third “intelligence”, as a tactic for exposing attacks. We can let our imaginations run wild with such ideas. It can only be hoped that human beings will not find themselves left out of these developments. If we do get left out, we'll find ourselves in the Hollywood scenario referred to in Figure 1, and the scenario will no longer be a fictive one.

The authors' main aim in this article was to call attention to potential basic and critical difficulties. While we are unable to offer real solutions, good analysis of the problems involved will provide the basis for development of solutions for protecting IT systems that are increasingly “intelligent” and algorithm-controlled.

References

[1] https://www.zero-outage.com covering the areas People, Platform, Processes, Security.

[2] Thomas Burri: Künstliche Intelligenz und internationales Recht, Mögliche Entwicklungen und Hindernisse; Datenschutz und Datensicherheit (DuD), Issue (Heft) 10, 2018, pages 603-607

[3] Felix Bieker, Benjamin Bremert, Marit Hansen: Verantwortlichkeit und Einsatz von Algorithmen bei öffentlichen Stellen; Datenschutz und Datensicherheit (DuD), Issue (Heft) 10, 2018, pages 608-612

[4] Eberhard von Faber and Wolfgang Behnsen: Joint Security Management: organisationsübergreifend handeln – Mehr Sicherheit im Zeitalter von Cloud-Computing, IT-Dienstleistungen und industrialisierter IT-Produktion; Springer-Vieweg, 2018, X+234 Seiten, ISBN 978-3-658-20833-2

[5] Thomas Hemker: Machen Maschinen die Welt sicherer? - Ein Kurzüberblick und Werkstattbericht zum Einsatz von künstlicher Intelligenz und maschinellem Lernen in der Sicherheitstechnologie; Datenschutz und Datensicherheit (DuD), Issue (Heft) 10, 2018, pages 629-633

[6] Thomas Dullien: Maschinelles Lernen und künstliche Intelligenz in der Informationssicherheit, Fortschritte, Anwendungen und Einschränkungen; Datenschutz und Datensicherheit (DuD), Issue (Heft) 10, 2018, pages 618-622