PLATTFORM INDUSTRIE 4.0

RESULT PAPER

Industrial security and the development of AI applications in the edge

Imprint

Publisher
Federal Ministry for Economic Affairs and Climate Action (BMWK)
Public Relations
11019 Berlin
www.bmwk.de

Editorial responsibility
Plattform Industrie 4.0
Bülowstraße 78
10783 Berlin

Status
May 2022

This publication is available for download only.

Design
PRpetuum GmbH, 80801 Munich

Picture credit
Infinite Lux/Westend61 / Adobe Stock / title, p. 10
xiaoliangge / Adobe Stock, ihor lishchyshyn / iStock / p. 5
D3Damon / iStock / p. 13
da-kuk / iStock / p. 16
Thomas Söllner / Adobe Stock / p. 23

Central ordering service for publications of the Federal Government:
Email: publikationen@bundesregierung.de
Tel.: +49 30 182722721
Fax: +49 30 18102722721

Contents

1 Introduction: The industrial application of AI – current status and trends

The use of AI in industrial production and administrative applications is driven mainly by the desire to increase productivity and introduce new performance features. AI systems expand services and analysis activities at the customer interface, promote sensor-controlled automation of production and provide the basis for new product capabilities and the implementation of these capabilities in new business models. With new AI-based processes, fewer workers are required to carry out routine activities and costly analogue technology is replaced by cheaper, comparatively simple digital sensor technology. This paper focuses on neural networks. Other approaches that can be supported by AI (clustering, combinatorial processes, etc.) are not considered in this paper.

With the widespread advance of AI in fields of activity that were previously shaped by humans and analogue technology, new security risks are emerging. One reason is the continuing rapid decline in the explainability of the results of advanced neural networks. This topic is briefly examined at the end of this introduction.

Furthermore, a technical revolution in the area of machine learning (ML) which started in 2018 is redefining the deployment of cloud and edge technologies: Considerable computing resources (usually on an extensive server pool in a cloud data centre) were required in the early phase of deploying ML, both for the creation (data acquisition, data evaluation, training) and the operational use of the finished recognition system (inference). However, nowadays the resources required are essentially limited to training, which has become vastly more efficient thanks to scaling network architectures and aspects of AutoML™[1], as well as transfer learning and multi-task learning. In contrast, new technologies such as quantisation [1] FlatBuffers format access technology [2] and edge accelerators (such as Graphics Processing Unit (GPU), Tensor Processing Unit (TPU), Intelligence Processing-Unit (IPU)) are managing to achieve, in cost-effective decentralised locally operated edge devices, inference performance that is already beyond the reach of centralised instances because of the current and realistically foreseeable level of network latency. Inference, the practical implementation of the fully trained system, only remains useful in the cloud because of advantages related to updating, maintenance and extreme application-specific requirements for computing and storage resources.

In the long term, productive use of AI will most likely happen mainly at the edge, on local hardware devices rather than on remote servers. All of the positive features therefore come with a trade-off in terms of security risks, since external surveillance and espionage are considerable threats. On the other hand, external communication is being restructured by relocating complex tasks to the edge. New methods and opportunities are being developed to strengthen privacy and confidentiality. For example, permanent communication at the machine level can be replaced by regular deliveries of reports via edge devices in a security-optimised format that contains only the specific information required by a particular service provider to perform its tasks. Any suspicion of espionage can be counteracted in this way.

After this introduction, **Chapter 1** describes the current paper, summarises the typical applications of ML in industry today, as well as how these applications have developed and displaced previous alternatives. It also outlines the current risk situation that can occur in such applications through the improper use of AI for attack purposes. Based on the developments described above, **Chapter 2** reviews the structural changes to the use of AI in industry. **Chapter 3** provides an overview of selected classes of industrial processes in which AI can offer concrete support today. It also presents selection criteria for AI methods. **Chapter 4** covers the particular security aspects of Industrie 4.0 use cases. **Chapter 5** provides a summary of the current options available for supporting key security issues in Industrie 4.0 using AI. The potential support offered by AI is discussed, both in the basic areas and in the newly emerging fields. Reference is also made to the opportunities and risks that are likely to arise from transferring the productive AI application to the edge within company boundaries. **Chapter 6** offers basic recommended actions together with **concluding considerations.**

1 Commercial software for optimising AI training

This document is intended to help domain experts better assess the suitability of AI technology for their particular area of application. It is also aimed at industrial users of AI-based systems and developers of AI algorithms, who should have a basic knowledge of AI methods.

Other organisations, such as the German Electrical and Electronic Manufacturers' Association (ZVEI), have carried out further work on the industrial application of these technological developments [3]. The document has been continuously updated during its compilation, both technically and in line with important political statements and measures. It was completed on 1 July 2021.

Lack of explainability in AI decisions

One challenge that affects AI-based decisions and evaluations is the lack of explainability [4] of the results of AI applications. Explanations are often sought when AI delivers a result that does not meet human expectations. However, AI cannot recognise the reasons for wrong decisions or indeed even form generalisations or abstractions regarding problems that have occurred, in the way that humans can.

The sub-working group "AI for Industrie 4.0 Security" has already published two reports on this topic [5] [6].

At present, a clear contradiction exists between the precision and explainability of AI results. This is also due to the superhuman level of performance in certain applications. While various efforts have been made to improve the explainability of AI decisions, satisfactory results are rare.

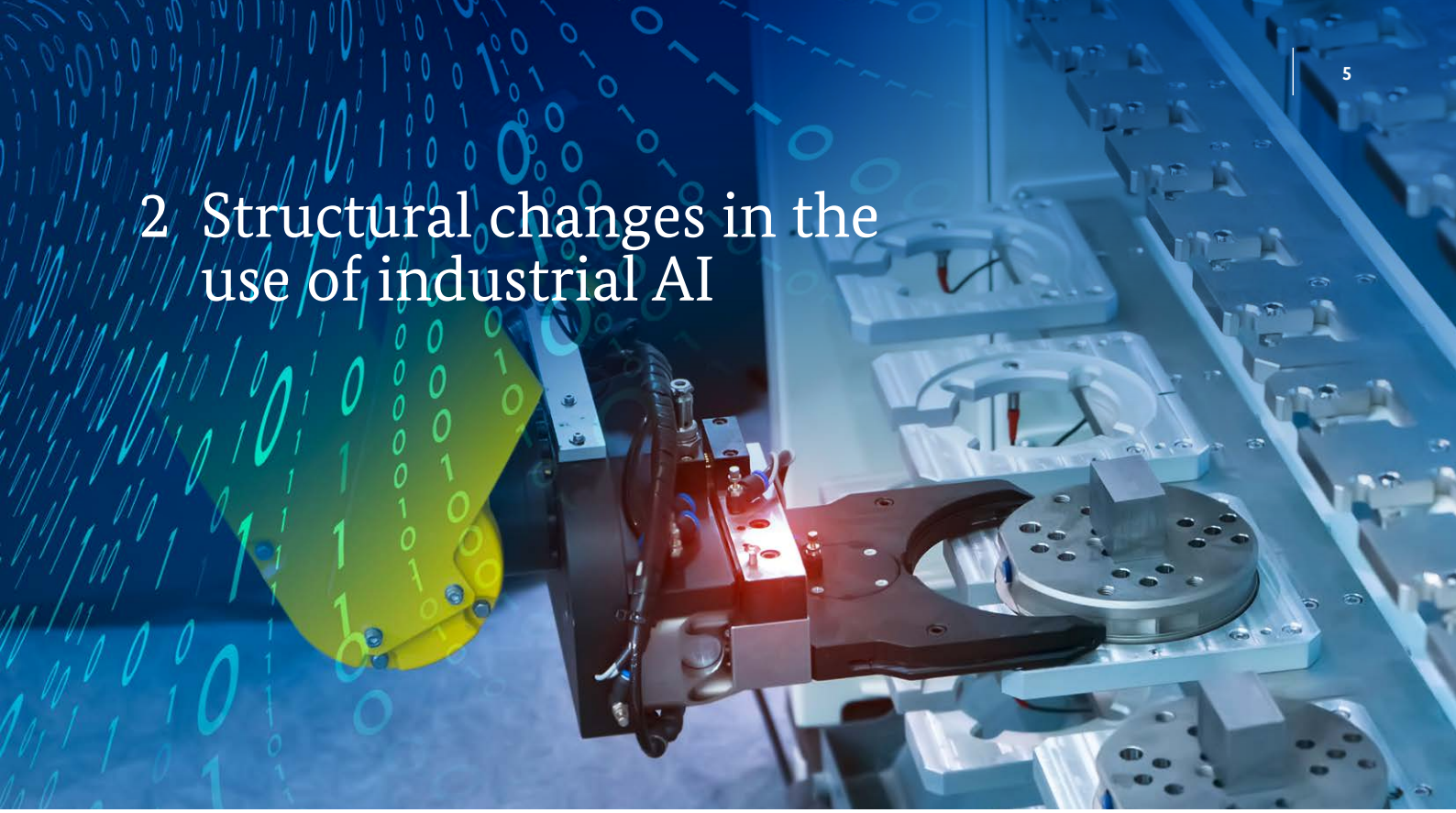
It may still somehow be possible, in the context of technical considerations such as those described in this document (see the following paragraphs), to satisfy the human desire to find an explanation that makes sense to humans. In the meantime, however, AI can also recognise things for which humans no longer have an explanation. This desire will therefore be increasingly difficult to meet. In reference [7] for example, an AI-based retinal image evaluation method is reported which, among other things, makes it possible to identify a person's gender, illnesses and lifestyle habits (opportunistic learning). The recognition of some of these attributes is inexplicable for humans, at least to date. It was previously not scientifically known that visible features of the retina contained certain data at all.

Results can be better explained when older AI methods are used, such as support vector machines (SVM) – a kernel-based family of machine learning methods. However, these methods usually achieve a significantly lower recognition rate compared to neural networks.

Some efforts to achieve better explainability of AI based on neural networks focus on translating the nesting depth of the automatically generated features in modern neural networks, which typically have more than 100 layers, into additional AI methods having less depth.

The following best practices can improve the explainability of results:

- Intensive testing (also with the help of a test AI), with a corresponding focus on borderline decision cases. This also applies in general, irrespective of whether or not an AI is used for decision-making.
- Teaching the AI with relevant data to avoid mislearning
- If possible, granularisation of the AI-based assessments in order to clearly indicate which part of an AI set of rules has carried out which assessment.



2 Structural changes in the use of industrial AI

As already outlined in the introduction, AI is continuing to develop rapidly in two directions in many specialist application areas. It is achieving ever greater performance and superhuman capabilities while becoming increasingly commonplace in our everyday reality. Both trends relate to the feature of machine learning, which is now often considered to be the most economically relevant feature of AI. AI is everywhere: in cars and most of the technically sophisticated devices used in households, entertainment, production, diagnosis, communications technology, agriculture and the military. There appear to be barely any limits to its proliferation, as new potential uses are constantly being developed. At the same time, more and more areas are being processed by AI far more "intelligently" than is possible by humans in many ways. Not only are machines more economical, much faster and more reliable, they are also far more intellectually productive within their specific field. Just as polymaths have been replaced by highly skilled specialists in the last 500 years, ML will propel the process of further specialisation into a new era. ALPHA-GO can only play GO and nothing else. The program has no education and no human qualities. However, in this, the most complex strategy game ever devised by humans, there is no longer any point in individuals or groups or machines programmed by humans competing against this program.

This chapter tries to explain the current development in these two dimensions from the perspective of Industrie 4.0 security.

2.1 AI in the edge

As mentioned in the introduction, the performance of AI in the edge has made a significant leap forward in the last three years. AI can therefore have an impact across broad areas of daily life. The Stanford professor Andrew Ng already predicted this development when he said "AI is the new electricity". AI is becoming as commonplace as electricity, it is everywhere and the share of technical devices with no connection to AI will soon be negligible.

To understand the new and evolving situation, we need to be aware that an AI system must first be trained. This means that in a neural network between 50 and 150 million parameters (weights) are generally determined at present. This step is complex and requires a large amount of computing capacity, main memory and floating point arithmetic. Once these parameters have been determined and the network has thus been "trained", then the application follows, i.e. the inference. Usually, no further changes are made to the network for a long period. It then only executes forward steps, for example to classify input values. This can be done with a much more compact memory structure and much cruder arithmetic. The network is converted into the FlatBuffers format and the arithmetic is changed from Float32 to Integer8 by means of quantisation [1]. Even in the case of complex neural networks, the result is a loading module that requires less than 1 GB of memory and, if the Integer8 arithmetic is accelerated accordingly,

can complete a complete inference operation in 5-15 milliseconds. The device of choice for the inference is therefore not the server farm in the cloud, but usually a very small edge device or a TPU-accelerated smartphone (e.g. a Google Pixel3) – on site and therefore without network latency.

Thanks to advances in semiconductor technology, the performance of edge devices is also improving. Devices in the format of a Raspberry Pi, for example, now have a memory of up to 8 GB and a multicore CPU. As a hardware basis for inference processes, devices with such resources have considerable power reserves even for demanding machine learning (ML) systems in typical industrial applications, such as quality control based on image analysis.

We are therefore getting closer to the goal of introducing high-quality AI into all areas of business and private life. At present, the leading platform for machine learning on very small devices is TensorFlow Lite. TensorFlow now works with over 10 systems from the field of microcontrollers. While these systems are too small to support complex operating systems, they can typically load and run a single application as an embedded system. The tinyML Foundation [8], a professional non-profit organisation, has existed since March 2019. Its goal is to develop platforms for machine learning that are small, that is, their power consumption is in the single-digit milliwatt range. This opens up the prospect of machine learning-based applications even in the area of off-grid systems: these systems can manage on power from a single small battery for years or operate with small solar panels and a small rechargeable battery. Such systems are, as explained above, very compact and also largely self-sufficient in relation to external power sources over relatively long periods of time. Put together, these features means that in future edge devices will largely recognise and record the environment, traffic, people and other phenomena as images, sound or other features derived from sensors and pass these on, via suitable channels, to interested central parties. In this case, the intelligent edge of the Internet of Things (IoT) does not have to be permanently connected to the Internet. A possible application example is a forester, who checks weekly whether the microcontrollers have seen, heard or smelled bark beetles or other pests. There are practically no limits to the possible application scenarios.

Nowadays, even small devices have space for several ML systems, for example as an ensemble to further increase accuracy. Even the smallest of devices can also accommodate malicious ML systems that can be used for espionage or to cause disruptions, thus specifically causing security problems. Typical microcontrollers have memories in the 1 to 3-digit kB range. This was the capacity of mainframe computers in the 1970s. That capacity is obviously not enough to recognise the world with Inception V4 or EfficientNet B7 [8] – in that case, you need "large" systems like a Raspberry Pi for €30 (see above, or a similar device). However, it is sufficient to recognise simple patterns, even when other tasks are performed simultaneously. It should also be noted that supposedly primitive devices can carry out analysis and monitoring that is by no means trivial, even if these devices do not have the technical requirements to run an operating system.

The ecosystem of the IT industry, comprising hardware, software, system and application manufacturers, has begun to tap extensively into the possibilities that open up when IT can perceive and recognise various aspects of the environment. This applies to various vertical segments, in the professional and consumer field, across a wide range of economic aspects and levels of varying complexity. Techniques such as voice control, pattern-based access control or data input for applications of all kinds – from industrial products such as lane departure warning systems in cars to voice-based detection of respiratory diseases or in industrial production such as AI in optical quality control to welding-in-process quality control – are becoming increasingly commonplace.

In the Industrie 4.0 application domain [10], the application functionality is migrated from "outside" (in the cloud) to the Operational Technology Edge, which is usually within a company's IT environment. Data no longer has to leave the company in order to be processed by an externally provided service: the service is carried out by the operator without any connection to the publisher.

The most important benefit, in addition to cutting costs and increasing performance, is the new opportunity to reduce the risks attached to cross-company communication.

By switching to protocol levels that can be explained by human beings, espionage through targeted evaluation and interpretation of message streams and infiltration with malware can be significantly reduced.

Pattern recognition takes place within the company at the edge as part of a machine feature. The possibility of attack is reduced because only periodic reports are created as PDFs, which show the next maintenance date and because a continuous data connection is no longer necessary.

Since this edge device can host the AI and other sets of rules for checking abnormalities or rule violations, triggering of alarms and blocking of invalid and unusual accesses or access attempts can already take place at this point.

On the other hand, the use of correspondingly powerful edge devices generates an excess of computer power in the central performance features of the hardware. A simple micro system such as the Raspberry Pi 4 has a memory usage in the single-digit percentage range when a modern high-end network such as the EfficientNet B3 is used for inference tasks. When you connect an accelerator such as the Coral TPU [11], the observable CPU performance (of the 4-core processor) also drops to unremarkable values.

Edge devices can map a large chunk of such security-relevant AI functions and locally recognise whether there has been a change in behaviour and, if necessary, decide how to proceed with external requests. However, a local device cannot recognise whether the behaviour of the overall infrastructure has changed across endpoints and participants. For this purpose, an AI would have to be provided at a centralised point. If necessary, this would receive data from the edge devices for comprehensive and centralised evaluation. The location of such a centralised AI cannot yet be determined because the architecture of the infrastructure has not yet been (fully) defined.

In brief, it should be noted that while AI systems are created through machine learning with resource-intensive use of server pools in a cloud, they are implemented in practice in software-optimised form for decentralised small devices at the edge of the Internet and even beyond, in parts of the world without electricity and communication networks. The reduction to FlatBuffers and I8 quantisation is not suitable for training but very suitable for inference, if necessary with double-digit billions of arithmetic operations per second for same price as a T-shirt. Energy self-sufficient sensor networks can spread to areas far from civilisation, configure themselves resiliently and require only a few nodes that need a connection to the world with electricity and communication with a fixed network. AI can be anywhere and is already a routine technology in many areas today. Devices of all types can see, hear, and understand language and hand signals and, of course, various other sensor signals. We all know that a modern car is primarily an AI-based system, at least in terms of the price components of the features ordered. Howver, rather than being controlled via a telecommunications network, the car derives its function from a local intelligence system. We talk to our navigation systems, rely on trusty warning systems, and enjoy the comfort of electronic drive trains. These are all responses to AI in the edge, yet they only give us a vague sense of what this technology can do in core areas such as Industrie 4.0 at the edge.

In this paper we explain that these capabilities do not come without risks, because smart edge devices can also get up to no good.

2.2 AI in the cloud

We now have network architectures that are deliberately aimed at efficient scaling. This is bringing about further streamlining of the loading modules to achieve inferences with a given recognition quality, including increased resolution of the inputs (example: Google EfficientNet B0-B8, compound scaling of 224X224 to 672X672 with only 5.3 to approx. 80 million parameters). Such architectures are also important so that the resources required for training can be aligned with the qualitative objectives of inferences. This training is costly and still needs to take place in an extensively configured cloud.

A cloud deployment for inference is only necessary for extremely fine-grained classification, in which case it takes place in cooperation with the edge for 5 to 6-digit class numbers. Networks like the familiar Google Lens are then used. However, this type of requirement is not part of industrial pattern recognition today. Such requirements are

targeted more at private users using their smartphones to find information about their living environment, wanting to recognise and compare plants, animals, cars or consumer goods.

We should point out here that training for typical industrial applications can now be achieved with far less effort than about three years ago. The technology of transfer learning has now been perfected to such an extent that high-performance networks can be created with minimal effort on the basis of renowned network architectures that once set world records in the ILSVRC competition (ImageNet Large Scale Visual Recognition Challenge). Practically all networks of this type are now available as Open Source Software (OSS) under typical OSS licenses such as Apache 2.0 including all weights, generated with ILSVRC2012 data. If you define around 20 new classes, as is customary in the context of industrial applications, and thus replace the classification layer from the 1000 ImageNet classes, then training for these classes usually takes less than 24 hours, including fine-tuning if you unfreeze fewer than the last 3-5 layers, but keep the rest of the weights fixed. This type of training for such an OSS network achieves match rates of over 98% (according to the top 1 criterion). The cost of this type of training is therefore marginal. It can even be carried out with a very manageable amount of training data if modern augmentation techniques are used.

A new class of realisable complexity is being developed for the process of training large new architectures outside of the basic systems for transfer learning applications. With the use of AI to determine the optimal training strategy (in systems such as AutoML from Google), there are new opportunities for achievable precision. Today, training processes are extremely resource-intensive primarily because many combinations of hyperparameters have to be tested in a heuristic procedure for training in order to achieve optimal results in the one loading module for inference that is ultimately used. AI-based strategies will make it possible to find shorter, more cost-effective routes to the goal. A new market for proprietary software with a very high value proposition is also opening up here.

Such systems could also strongly promote the use of Reinforcement Learning (RL) and Generative Adversarial Networks (GAN), which today can only be implemented with extremely intensive allocation of resources and associated high costs. These are the sub-disciplines of AI that go beyond the proven superhuman capabilities of machine learning in pattern recognition and enable creative capabilities that are only just beginning to be explored. The creative tactics of ALPHA-GO when seeking victory in the field of strategic games provide an example of superior machine actions that human brains can no longer understand. GAN are believed to have similar capabilities in creating patterns of all kinds that can contribute to progress in technical, medical, and general scientific research.

In summary, we can say that modern AI in the cloud will deliver results that lead to an increasing number of superhuman machine capabilities, especially in areas that were previously considered to be differentiated by human intelligence. Creativity and developing strategy have so far been regarded as skills specific to humans, while the capabilities of sensory organs have long been considered technically feasible. For example, dogs have a better sense of smell than humans, while birds have better vision. It has been clear for a long time that machines can adopt this type of capability and indeed outperform humans in these areas. A new battleground in IT security has emerged in Industrie 4.0 with the proliferation of AI in the smallest edge devices. These are in widespread use and while they are sometimes not at all associated with AI, they can execute the loading modules from these networks with a high level of performance and at the same time allow scope for manipulation within the device.

2.3 Global AI market

It is well known that, compared to the budgets available in the United States and China, financial support for the further development of AI in the rest of the world currently lags far behind. Three-digit billion amounts from government organisations combined with very large investments

by private companies in the Internet sector ensure that today almost all globally recognised publications, advanced product developments and strategic course determinations occur in North America and China. Anyone looking for current, highly technical scientific publications on the topic of AI on the Internet will initially find almost exclusively American references in the common search engines and then gradually be directed to the area of search results documented only in Kanji characters. There are practically no search results from Europe, or in particular from Germany. High-tech AI is not happening here, at least in this regard, as anyone carrying out such searches can clearly see.

The development of AI is being spearheaded by a few companies (Apple, Facebook, Google, Baidu, Amazon, Microsoft), universities (Stanford, Berkeley, NYU, Montreal, Toronto, MIT, Oxford) and some predominantly Chinese government organisations. With the explainability of AI by human brains continuing to diminish rapidly and the prospect of AI-based explanatory systems emerging only in North America and China, there are grounds for concern. To make matters worse, the gap between the scientific AI-related skills of these countries and Europe is steadily widening.

Recently published material on backdoor access to IT and OT(!) systems [12], [13], [14], [15], is now considered to have been known to manufacturers for a long time. It is all the more alarming to find that the manufacturers concerned only act after the devastating economic and geopolitical effects of these attacks have been felt by their customers. There is a dangerous gap between the detection and elimination of these dangers, one which is often perpetuated by state-political or economic considerations. With this in mind, trade-offs in AI solutions in the future should not be assessed differently.

3 Typical applications of AI in industry

After a review of the status of AI, trends and structural changes in industry triggered by AI in the first chapters, this chapter examines the current typical application of AI in industry. To this end, limitations and criteria for choosing the right AI solutions are proposed.

Typical applications of AI in industry can be found in

- Production automation
- Control of industrial drives (motion control)
- Demand-driven maintenance of machinery
- Process-optimisation
- Quality assurance through non-destructive testing.

These applications mainly use sensor technologies to obtain the necessary parameters. In addition to imaging sensors (image, colour, light), many other sensors are used in the industrial sector to measure temperature, pressure, tension, acceleration, torque, contact and much more.

The AI methods currently under discussion are based on machine vision, hearing and communication through the use of signals from imaging and acoustic sensors and less on the use of the other sensors mentioned. These are subdivided into:

- Image/video recognition
- Voice recognition
- Text summarisation
- Sentiment analysis
- Evaluation of contracts and contractors as well as other administrative applications.

Pattern recognition through the use of deep learning with neural networks has proven to be especially successful. This is due to the impressive progress made in the development and implementation of the algorithms for neural networks and also to the availability of computing capacity and speed in the cloud, thus making the use of neural networks feasible. Another success factor is the fact that the algorithms and tools are available as open source solutions.

Neural networks are trained with large amounts of data. Various options are available on special servers in the cloud for the purposes of training. Examples of such technologies include: TPU (Tensor Processing Unit) systems from Google. TPUs are specially developed microchips that calculate and continuously optimise the parameters of neural networks during the learning process. This technology greatly accelerates machine learning. Other learning accelerators are IPU (Intelligence Processing Unit) systems from Graph-

core, which have implemented in-processor memories. These memories allow the accelerators to achieve very high speeds for training neural networks. This technology is available as a service in the Cirrascale Cloud. Providers like Baidu, Nvidia and others offer open source software for deep learning. For example, PaddlePaddle from Baidu supports distributed computing for the efficient use of multiple cloud servers. Nvidia's Cuda-X AI runs fastest on servers with Nvidia GPUs, which are also offered by many cloud providers.

One fundamental problem when training neural networks is that of gaining access to sufficiently large volumes of high-quality training data. Tens of thousands of training data examples are required to achieve a sufficiently high recognition rate. However, in practice, there is almost never enough training data available for a user's own specific application. Usually only a few data sets are available.

One solution to this problem is to use transfer learning. For transfer learning, you start with a pretrained network that is available as open source. This network is then retrained for the specific application with a few hundred data sets. This means that only the last layers of the neural network are adapted to the actual problem. This works quite reliably if pretrained networks from a similar application class, e.g. image recognition, are selected as the starting point. However, in this case the application-specific and proprietary data sets have also so far been uploaded to the cloud for retraining. Under certain circumstances, this may present a security problem and a problem for protecting your own intellectual property (IP).

Nowadays, transfer learning can also be carried out on your own computer if it has a reasonable amount of processing power, so you no longer need to upload proprietary data to the cloud.

Deep learning with neural networks always has one goal: recognising data patterns. AI is very good at this, far better than humans. However, the result depends on the previously trained data sets. If the training data contains hidden assumptions or distortions (bias), these are automatically transferred to the results. Furthermore, if the result is fundamentally wrong, AI is not able to clarify the cause (explainability of AI results [4], Chapter 1).

Limits of AI solutions

- Although the match rates of AI solutions are significantly higher compared to previous standard methods, AI applications still do not achieve 100% coverage because it is not possible to train the entire data space with examples.
- It is difficult for AI applications to analyse the causes of errors (limited explainability [4], Chapter 1)
- Insufficient quality of the training data with an undetected bias simply replicates this bias and thus distorts the result.
- In the event of unexpected problems with AI applications (out-of-the-box), only humans can help – so far.
- AI solutions themselves do not contain any security measures against attacks. Security measures must therefore also be implemented as a protective screen (onion skin model) and access models defined with appropriate authentication methods.
- Implementing AI requires expert knowledge.

There is as yet insufficient analysis of sensor networks with deep learning: the available data sets do not always provide a suitable starting point for specific applications. The first address for data sets can be found at 'Top Sources For Machine Learning Datasets' [16]. The question also arises as to what extent deep learning is generally the most suitable approach in this context, or whether other AI methods such as SVM or Gradient Boosted Decision Trees, for example with random forest, are more suitable for some applications.

Criteria for selecting AI for industrial applications

- Application is based on machine vision, hearing, communication with text in natural language, strategy development in complex decision-making situations, for example to control autonomous systems
  - Deep learning with neural networks, preferably using transfer learning with pretrained networks, if available

- Evaluation of sensor networks
  - Simple anomaly detection with priority given to explainability against the maximum possible match rate
    - ML clustering with corresponding statistical algorithms
  - Highly complex sensor networks
    - Neural networks + transfer learning on on-site servers avoids loading of proprietary data into the cloud and minimises security risks such as the loss of IP.
      Prerequisite: A pretrained neural network suitable for this application is available in order to minimise the on-site training effort. Such a network must provide the same architecture with a suitable input layer on which the available data can be adapted without any loss of quality. A suitable interface must also be provided for transferring the intermediate inference results from the frozen layers to the newly written classification layer.
      Self-evidently, a Convolutional Neural Network (CNN) from image recognition cannot be used in order to mimic the capabilities of a Long **short-term** memory (LSTM) for voice recognition. Transfer learning delivers astonishing performance, especially for image recognition, since the convolutional layers of these networks mainly contain the features for processing graphics primitives. They are largely independent of the recognised image contents. Therefore pretraining with ImageNet data (general images of animals, plants, buildings, modes of transport and so on) often provide a very good basis for completely different tasks of graphic recognition such as medical diagnosis, non-destructive testing in the manufacturing industry, etc.
- General prerequisites:
  - Availability of suitable in-house expert knowledge or consulting firms
  - A security concept must be developed, in every case, for the access-secured and integrity-assured storage and processing of AI training and application data that has been externally purchased and generated during use over the entire life cycle by machine and human process participants.
  - Clarification of contractual issues such as rights to the data or the resulting IP, also as a prerequisite for the application of a suitable security concept.

Technology outlook: TPU ICs consist of complex processing cores, the maximum possible number of which are implemented and connected in parallel on an IC. The maximum possible number of cores is determined by the selected silicon technology and then by the price of the IC (cost to performance ratio). By continuously shrinking new silicon process nodes, for example to half the structure width with the same silicon area, silicon technology enables the implementation of four times as many functions, in this example four times as many TPU cores. This is described by Moore's law, according to which computer power doubles every year. However, Moore's law is increasingly no longer applicable as silicon technology moves into the ranges below 20 nm structure width. Under certain circumstances, this can be at least partially compensated for by better architecture concepts for implementing algorithms. For example, the latest TPU cores are currently produced using 7nm technology [[17], [18] and are already well optimised in terms of processor architecture, barring an unforeseeable new approach. For the next decade we predict that chips will shrink in three further stages to 2nm technology. The IPUs from Graphcore are also manufactured in 7nm technology. Since each shrink stage will make the cores 2.5 times faster for about the same IC price, this would accelerate the computing power of IPUs and TPUs by a factor of 15 in this decade, provided that the problem of energy dissipation in the ICs can be adequately resolved.

4 Security considerations for Industrie 4.0 applications

In the context of Industrie 4.0, security features are quality features in the broadest sense. Since these features are not inherent components of new technologies, including AI, this chapter addresses the fundamental issues of security in AI applications. It links previous chapters to the following chapter about the applicability of AI to the various security-related aspects of Industrie 4.0. 4.1 AI in the Industrie 4.0 context.

In addition to the formidable performance of AI in general considered above and in light of existing AI applications in the industrial context, it should be noted that deployment scenarios in the field of Industrie 4.0 are particularly dependent on values such as “trustworthiness“ and “secure cross-company communication“. It is evident that people sometimes find it difficult to understand where to find evidence in order to trust AI decisions when the results are not neat and understandable. Frameworks that enable systematic testability or certifiability in the conventional sense are not yet known. Further methods need to be developed to this end, in particular to reliably detect the sometimes invisible and security-related influence of distortions (bias) in the AI evaluation process.

As already explained in two previous publications [5] [6], we still face particular challenges at this point, primarily relating to the security of components, machines and operations.

The special Industrie 4.0 security requirements, which have so far been supported by largely established cryptographic methods and processes, must be examined to determine how security can be improved through hardened AI applications in the shop floor area, primarily in edge systems. In addition, the extent to which edge systems can be defended against Generative Adversarial Networks (GAN)-based AI attacks by means of these hardened AI applications and thus demonstrate greater resilience compared to conventional Intrusion Detection Protection (IDP) systems, is to be investigated. In all cases, new high-performance edge systems also raise new questions in all contexts with regard to their possible uses, as well as possible trade-offs. These considerations and questions with reference to “industrial security and the development of AI applications at the edge” have so far not been given any particular attention in industry, especially since edge architectures are only in their infancy. To date, security issues have only been considered at a later stage.

4.1 Relevance of security requirements for AI in Industrie 4.0

In general, it must be ensured that the use of AI in Industrie 4.0 components does not exacerbate the existing security situation and does not have a negative impact on the status quo. In this respect, the discussion on the use of AI

applications for Industrie 4.0 also addresses the question of how AI services that are already on the market, as well as future developments, have an impact on the required security features of components, machines and operations.

Future system developers will no longer necessarily be able to work with the same development tools used now or in the past to validate design features and security requirements with respect to AI applications. These AI applications must not be technologies that are retrospectively "tacked-on", but must be treated as inherent performance elements in the design of components and machines through to operation.

Similar notions are already in circulation with regard to the **security-by-design** requirement in product development in accordance with ISO/IEC 62443. In the same way, the requirement placed on AI applications in established mechanical engineering in this context is, among other things, adherence to **integrity-by-design**. The objective thus pursued would be to integrate AI applications, not randomly and without control, but in accordance with engineering integrity standards, in order to achieve human controllability of systems. In general, placing blind trust in AI results in the area of Industrie 4.0 is a mistake. The focus should be on proven behaviour and a basic understanding of the phenomenon of possible distortions. In addition, inference systems at the edge should ideally be protected against vulnerabilities and undesirable side effects.

4.2 Application of known IT security trust models and prevention methods in Industrie 4.0 AI scenarios

If the above security considerations are consistently applied, the "Zero Trust Concept" known from the field of digital identities can be used as a blueprint. This is based on the idea that if there is a high number of agile and constantly changing communication participants, these are assigned an initial zero trust value and must then "earn" the trust of the other participants within the framework of a trust scoring system.

In other words: In an ideal case, any Industrie 4.0 communication with other components would exclusively follow the aforementioned "zero trust principle". Then the identities and authorisations involved would be checked before any access to applications, data, sensors, and actuators. Only once validated and approved would communication links be activated and applications and data spaces made visible to human and machine users. In addition, every access would be logged with relevant metadata in order to support forensic analyses in the event of a malfunction or attack. This principle is already used in behaviour-based intrusion detection systems (IDS) and intrusion prevention systems (IPS) and can serve as a model in industrial AI scenarios. The extent to which the required trustworthiness [19] of decisions made by the actual AI-based systems can be determined, evaluated and explained, by means of additional predictive AI analytics, is still the subject of research. For this purpose, dynamic changes in the observed system metrics are evaluated with regard to their plausibility. The objective is to achieve AI-monitored or AI-explained decisions by AI systems.

In real Industrie 4.0 implementations, including in the area of critical infrastructures, the challenge over the next few years is to integrate a trustworthy combination of new, I4.0-native systems with existing systems that have sometimes had long-term investment.

In the run-up to an AI-based I4.0 project, it should be clarified whether a digitisation project may be necessary in order to create a homogeneous digitised process environment from a brownfield (existing) environment. Alternatively, additional IIoT sensors (e.g. thermal imaging cameras for temperature monitoring) can be used to determine operating parameters that are not yet digitally available. To avoid creating unintentional gateways for security incidents and configuration errors, it is necessary to examine whether and how communication interfaces and protocols can be translated to uniform standards using industrial hardware and software. The actual systemic challenge is that of being able to assign an appropriate place to edge-based AI systems in the context of trustworthy standardised systems in a manufacturing landscape in transition. In short: which AI decision is justifiably trusted, when and by whom.

4.3 Interoperability at the organisational level

Amid the excitement about increasing the potential of new, high-performance, autonomous AI-based systems, questions concerning interoperability must also be asked:

Do the new AI-based edge, edge-cloud or cloud topologies in industrial manufacturing follow established security standards? Could interoperability lead to power plays at different levels in the network? How do autonomous decisions by components and machines affect a process control structure that has historically grown in a hierarchical manner? To what extent can AI-based autonomous machine-to-machine decisions be stimulated and ultimately, from a systemic point of view, assume control without trade-offs?

These and other important questions regarding organisational interoperability requirements cannot be dealt with conclusively within the scope of this document. They show the importance of and the need for planning organisationally, conceptually and holistically when introducing AI-based systems. The strengths and possible imponderables of AI systems outlined in the previous chapters have an impact on functional interoperability. This means that they also affect the stability of Industrie 4.0 processes.

4.4 The requirements of Industrie 4.0 security measures

If AI applications are to catch on in Industrie 4.0 as "the new electricity", to use the analogy from Chapter 1, the characteristics of this electricity are absolutely mission-critical. In the worst case, all components fail if this "energy" is insufficient. Alternatively accidents occur because the "energy" is excessive, too unstable, cannot be calculated or because other conditions cannot be explained. These scenarios must be avoided at all costs.

The requirement for integrity and stability, and ideally also standardised interfaces, is of paramount importance. To solve the security issues facing Industrie 4.0 using AI, some basic requirements must be met. Some points are listed below as examples:

- Understanding of industrial security requirements, especially for Industrie 4.0
- Understanding of the basic working principles and importance of AI applications and assessment of the results
- Creation of suitable models for the interaction between AI and security
- Creation of secure models to validate integrity
- Development of suitable test frameworks and methods for the certification of AI models, taking into account the processes for data acquisition, modelling, maintenance, monitoring in the field, etc. and operation

The following Chapter 5 examines in more detail whether and how AI applications can be represented within the individual security-related topics of Industrie 4.0 and with what added value. To this end, selected aspects from the work of the entire "Security of Networked Systems" working group from Plattform Industrie 4.0 are considered.

5 AI and security issues facing Industrie 4.0

This chapter discusses and examines the applicability of AI to the various security-relevant aspects of Industrie 4.0. The chapter contains corresponding Industrie 4.0-specific terms from various discussion papers published by Plattform Industrie 4.0 on the subject of security. The usability of AI with regard to the various aspects is also set out in examples, with answers provided to the following question:

From a **security perspective**, how can AI be incorporated into the various subject fields of I40 and help to achieve a higher level of security?

5.1 Discussion of Industrie 4.0 security issues

The following section examines the various subject areas of the "Security of Networked Systems" working group in detail.

5.1.1 The I40 value chain

AI can make an important contribution to the creation of a new value chain[2]; AI can quickly assess providers with regard to their suitability to participate in a value chain. An assessment of a participant in a value chain can include various attributes, such as delivery reliability, payment behaviour, rumours or reports on the state of the company, quality information on the products, previous experience, pricing policy, environmental protection, market dominance, the overall economic situation, local or global pandemics in regions or countries, or their history of dealing with these. The massively increased amount of data now available on all these factors makes it difficult for humans to carry out the assessment satisfactorily, since the relevant amount of data and filtering of this data amount to a "big data" problem.

Using these attributes, an AI can assess the participant based on scores in the form of a recommendation. In this way, the candidates can be placed in a desired order and invited to participate in that order.

The same selection process can be carried out when replacing participants, even under time pressure. An interim assessment of the participants can also take place during productive operation of the value chain, for example, political changes, **natural disasters or the spread of disease can make it necessary to take quick action and replace a participant.**

If necessary, the AI can be further developed on an ongoing basis by means of targeted update training, since the final decision made by humans ultimately results in data labelling of the selection, which can be taken into account in subsequent decisions.

2 A description of the characteristics of the I40 value chain can be found in the study "IT-Sicherheit für Industrie 4.0" [24].

The number of potential participants in a value chain on the market is often limited, since the market is only divided by competition. Under these conditions, the AI could be used to propose other potential participants once these establish themselves in the market in order to prevent fundamental dependencies caused by a participant dominating the market or to counteract this situation.

Another example of the use of AI within the value chain could be the detection of unusual events in the productively running I40 value chain and a corresponding (early) warning. For example, an AI could recognise an unusually high data request, possibly with an unusually high rejection rate when communicating between two participants. Or the AI could detect unusual data requests whose access has been authorised, which were previously not required, or only rarely.

If necessary, AI can be used before the start of productive operation to create intelligent suggestions for the adequate provision of production data for participants within the value chain. The question of which data is required by which participant, from which other participant and when, and when which participant should then have approved access can no longer be answered trivially, if there is a larger number of participants in a value chain. Without technical support it can hardly be framed consistently by the human brain.

In addition, an AI could make suggestions for the diversification of value chains. In some cases, there may be independent production steps in which other value chain participants do not have to be involved. Strict separation (e.g. between the supply of raw materials from the manufacture of the end product) increases security: the dependencies in the supply chain are reduced and the knowledge held by participants in one value chain about those in another value chain can be dramatically diminished.

5.1.2 Secure digital identities

The allocation and management of digital identities[3] is becoming more complex, due to new, decentralised identity architectures. This concept became part of a global discussion following the proposals of the World Wide Web Consortium (W3C). In addition to many private, commercial and government discussion groups, the European Union is also addressing the topic within the framework of the eIDAS ecosystem [19]. The search for a solution in the last two years has consequently moved away from a single, centrally managed identity towards a “Self Sovereign Identity” (SSI) [20]. In the global solution space, identities can be characterised by different, use-dependent identification features, and their validity and authenticity can only be checked and confirmed by use-specific test centres. If an SSI is used, an I40 entity [21] manages its various, use-specific identity features itself. This management is no longer carried out centrally, but by different decentralised entities using “Decentralised Identifiers” (DIDs). The concepts are currently being developed by W3C working groups [22]. In addition, it is planned to define the interpretation of identities specific to their use for purposes in the Industrie 4.0 environment and also in the context of GAIA-X (see Chapter 1).

The required minimum security attributes of an identity can vary depending on their use, the environment and consideration of the relevant costs of their provision. In the area of critical infrastructures, for example, higher security requirements are placed on the I40 identities than in the mass production of paper breathing masks. The mass production of paper respiratory masks to provide basic protection against infectious diseases requires very effective quality control. However, for cost reasons, it may be difficult to require each individual protective mask to have its own, unfalsifiable proof of identity across the entire value chain.

AI can make an important contribution to establishing minimum security requirements for identities. These requirements can of course be a list comprising graduated security requirements, even if the identities are found in a common value chain. A wide variety of criteria can be included in the creation of the requirements, such as the total costs of production, cost requirements for production, type of identity (human, machine, partial product, end product), possible uses of the end product, buyer of the end product, “removal” of the identity (e.g. in the supply chain) from the actual end product, influence of individual parts on the mode of operation, possibility of an identity being

3 For a description of the characteristics of a secure identity, see the output paper “Technical Overview: Secure Identities“ [21]. This solution offers certain potential for AI support and is explained below.

isolated within a value chain, manufacturing secrets and their protection (intellectual property), etc.

The AI can then evaluate individual security attributes of the identity used in the value chain, per identity, e.g. with regard to:

- The way in which identities are secured with regard to their authenticity (using hardware, software, combinations, ID cards, fingerprints and other biometric features)
- Requirements for the identity as a communication participant (attributes of secure communication, such as encryption, level of encryption, use of communication protocols, signatures that are required at minimum)
- Requirements for auditability and traceability (determination of the need for auditability as well as traceability and, if necessary, the level of detail required)

AI can also help in the detection of attempts (successful or unsuccessful) to falsify or exchange a secure identity, e.g. through behavioural pattern recognition and detection of deviations from these learned behavioural patterns. Examples of deviating behavioural patterns are:

- A change in communication behaviour, such as frequency, unusual requests, making contact with other identities, previously uncontacted, within the value chain
- Use of other security attributes when communicating (or when attempting to communicate)[4]
- A change in the speed of communication (e.g. response to requests, number of response blocks)
- Use of IP addresses, URLs, Mac addresses, etc.
- A change in the type of auditable data (e.g. log data, its amount, the logged events, etc.)
- A change (or attempt to change) login processes to other identities or centralised components

Since the secure identity in the productive operation of a value chain can be assigned to a participant in a value chain, the AI contributions already mentioned above for safeguarding the value chain also apply.

5.1.3 Secure communication

Secure communication is characterised above all by the fact that the communication participants are equipped with the appropriate security attributes, such as appropriate communication protocols guaranteeing security, appropriate keys for encryption or decryption of data, and certificates that can be presented to check the authenticity of a communication participant. However, communication can be structured much more securely if only those connections that are actually required in the communication network of the value chain are configured accordingly when setting up communication paths. Constant changes to the participants and the associated communication channels quickly result in a communication configuration that is confusing for humans and can only be managed with the support of machines, possibly with AI.

AI can be helpful in detecting unusual attempts at communication. The same examples apply here as for the value chain and secure identity, which are used in connection with changes in communication and communication attributes for secure communication: the speed of communication, identity, etc. AI can also help recognise whether communication paths are incorrectly or unnecessarily configured. Communication participants who have left the value chain may still communicate (unsuccessfully) with their former communication participants belonging to the value chain, possibly asking for unnecessary data via an unnecessary connection. It looks just as odd when new participants in the value chain do not communicate at all or only a little. In all cases, unusual communication takes place, which could possibly be recognised with the corresponding meta-data even[5] without AI. However, if there is no meta-data, it can only be recognised with the aid of an AI, since plausibility or non-plausibility cannot be recognised in the normal way.

4 A description of the characteristics of secure communication is contained in the discussion paper “Secure Communication for Industrie 4.0“ [25] and also in the discussion paper “Secure cross-company communication with OPC UA [26].

5 For a rough description of the characteristics of a trust infrastructure, see the document “Eberbach Talk on Security in Industrie 4.0” (2013, Fraunhofer SIT, Darmstadt [27]. A current publication by Plattform Industrie 4.0 “Vertrauensinfrastrukturen im Kontext von Industrie 4.0” is listed under reference [28].

5.1.4 Trust infrastructure/Trust profile

At present, the question of how to create and shape a global trust infrastructure is still in intensive discussion in the relevant working groups of Plattform Industrie 4.0. This section therefore cannot name any specific AI aids for the security of individual entities that ultimately define the trust infrastructure.

However, in principle the trust infrastructure is a prerequisite for the secure operation of global value chains and essential in order to provide the necessary security features to establish secure identities and secure communication within value chains. Therefore, every cyber attack on a value chain or attempt to manipulate a secure digital identity is also an attack on the trust infrastructure in which it is located. In the discussion of how AI can help with security issues, the examples from the above sections on the value chain and secure identity also apply.

Conversely, attacks on the components of a trust infrastructure that exists worldwide (e.g. on their Public Key Infrastructures (PKI) or their Certificate Authorities or Certification Authorities (CA)) also amount to attacks on secure identities, secure communication and the secure operation of value chains. In this case, AI can make a contribution to, for example, recognising unusual behaviours in the communication of PKIs, CAs and their network components and providing direct or indirect evidence of the manipulation of such facilities.

In particular when establishing the initial trustworthiness of an identity (i.e. providing the identity with the necessary keys and certificates), the AI can possibly detect unusual activities, such as:

- Use of unusual communication channels for the first or follow-up request for keys and certificates from a CA
- Unusual composition of the certificates
- Unusual behaviour of communication participants (possibly secure identities) immediately after receiving and using their new security features for the first time.

Note: Additional metadata makes detecting such manipulations easier. If, for example, it is known that a secure identity is involved in value chain 1 and 2, it can be recognised (sometimes even without AI) that unusual communication behaviour is present when the identity tries to connect with components from value chain 3.

5.1.5 Attribute-based access control

Using attribute-based access control (ABAC)[6], you can design complex individual rules as well as sets of rules that make sense when examined individually, but are difficult for humans to evaluate for the meaningful behaviour of an overall system of rules and for inconsistencies, if necessary. For example, there could be one rule that allows access between 10:00 hours and 16:00 hours. A second rule could subsequently be introduced that prohibits access between 15:00 hours and 22:00 hours, thus creating an inconsistency for the period 15:00 hours to 16:00 hours. A test phase prior to commissioning that reveals this type of inconsistency is definitely required.

In addition, it may no longer be possible to formulate complex rules that contain a wide variety of attributes at all using normal rule description languages, since the exact formulation contains too many programmed special cases as combinations of attribute values. For example, one or more rules designed to determine whether an Industrie 4.0 component may be switched from the maintenance status to the production status could include attributes such as component data (pressure, temperature, power consumption, engine speeds, etc.), weather data (air pressure, precipitation, type of precipitation, degree of cloudiness, humidity, wind speed, wind direction, weather development, likelihood of thunderstorms decreasing/increasing, etc.), communication behaviour of the component (“usual”/”unusual”), status of the value chain (dependencies within the value chain, so that switching would put the component at risk), etc. In this case, humans need significant help from technology in order to maintain/gain an overview, indeed if this is possible at all.

6 For a description of the characteristics of access control for Industrie 4.0, see the discussion paper “Access control for Industrie 4.0 components for application by manufacturers, operators and integrators” [29].

AI can be very helpful, for example, for testing complex sets of rules and finding inconsistencies in this process. This can be supported by clustering approaches, for example. An AI could iteratively combine the various attributes of a set of rules in an intelligent way in order to arrive at the limit values of the response change of the set of rules as quickly as possible. It could allow humans to make the final judgement on whether or not access should still be allowed in these limit cases, or whether there are any inconsistencies and the decision is therefore incorrect.

AI can be used not only to control rule sets and their consistency, but can also be used within rules. The example of switching Industrie 4.0 components can be supported by such AI-based sets of rules. Here, AI can also analyse temporal processes and draw conclusions about normal/unusual or uncritical/critical temporal processes. This can lead to the refusal of access within a rule, regardless of consistent and current individual attribute values. In the example above, this would be the weather development, past status and future status of the value chain.

However, the assessment by an AI of the temporal course of the status of the value chain (described by a large number of individual but interacting parameters) could be incomprehensible for humans in this example, in connection with the lack of explainability of AI decisions described in Chapter 1. The above-mentioned aids for (partial) explainability are used as possible remedies: Intensive tests, data checking of the validity of the learning data, granularisation of the AI assessments.

5.1.6 Collaborative Condition Monitoring

Collaborative Condition Monitoring[7] is a suitable scenario for illustrating the interaction of the various aspects already discussed in the previous paragraphs. Within a value chain, (at least) two secure identities (e.g. the monitoring officer and the I40 entity) communicate with each other via secure communication (secure, encrypted connection), which is ensured via a trust infrastructure and where the requesting party uses an attribute-based access control to gain read-access to the monitoring data. All examples of AI aids from the terms discussed so far therefore apply.

AI can also be helpful in relation to this scenario in detecting unauthorised or unusual access to monitoring data. Often the monitoring officer can retrieve all information from an aggregated report. This means there is no need to gain further access to detail attributes of the entity or only in rare cases. The same applies for write-access to certain attributes. Based on a frequency of such accesses across different I40 entities, AI can conclude whether the access or the frequency appears unusual. In the event that monitoring data is provided regularly/permanently and collected, if necessary, there is also a need to monitor the data flows of this data. AI can help to recognise unusual data flows and, if necessary, to prevent them or filter them according to certain recipient attributes before sending. Given that AI is increasingly used at the edge, rule violations and obvious behaviours in requests can already be recognised in an edge device in the case of the "Collaborative Condition Monitoring" scenario and the request terminated at this point.

Since the Collaborative Condition Monitoring scenario was still in development at the time this document was written, further aspects of the scenario cannot be discussed in detail in this document.

5.1.7 The administration shell

The importance of the administration shell[8] as a core element of I40 entities must be acknowledged, also with regard to security, because it contains, for example, operating parameters as well as secrets belonging to the relevant manufacturer and entity operator. This data could be stolen, destroyed or manipulated in a cyber attack. Since the administration shell is part of the I40 entity and the I40 entity is identified by a secure identity (which in turn is located within a value chain, communicates with other value chain participants using secure communication and has access control mechanisms), the considerations above apply in full to the administration shell.

7 A description of the term "Collaborative Condition Monitoring" and a description of the scenario can be found in the publication "Collaborative data-driven business models" [30].

8 A description of the administration shell is contained in the specification "Details of the Asset Administration Shell" published by Plattform Industrie 4.0 [31]. A description of the security requirements of the administration shell is contained in the discussion paper "Security der Verwaltungsschale" [32].

AI can be particularly helpful in monitoring data espionage and data manipulation with the purpose of causing damage. Data espionage can be mapped to the unusual data accesses or data access attempts that have already been discussed here in detail. Data manipulation with the purpose of causing damage must be distinguishable from a normal and quite frequent change of data within the administration shell (e.g. in connection with the use of the I40 entity using new operating parameters). If the asset administration shell contains a large number of very different attributes, it is no longer possible for humans to monitor every kind of combination of the various attribute values. Built-in software input tests are also not able to carry out rule-based checks on the validity of every combination of attributes and attribute values based on rules. However, AI can carry out these plausibility checks very efficiently and trigger an alarm accordingly if a combination is to be assigned to the whitespace area.

As already explained under the term “access control” and at the beginning of the chapter, the difficulty with AI is that the result of an AI-based investigation cannot be explained by that same AI. It simply delivers a score, indicating: more plausible or less plausible To check the plausibility of the administration shell data, aids must therefore be provided that allow a final analysis to be carried out by humans when the AI generates the alarm. Saving the history of the last changes to the administration shell data would seem to be appropriate here: in this way, the data version from before the AI alarm can be compared automatically with the version in which the alarm occurred. A person (as the administrator of the administration shell) can then easily identify the root cause of the alarm after doing appropriate research. In addition, another AI can be used as an aid to explain plausibility or non-plausibility. However, humans should have the final decision here as to whether the version of the administration shell data in which the alarm was triggered is still plausible or not.

Since various aspects of the administration shell are still being discussed within Plattform Industrie 4.0 at the time of writing this document, these aspects cannot be explored in more detail.

5.1.8 GAIA-X/Cloud services/Edge devices

GAIA-X[9] stands in principle for a highly standardised data infrastructure on the Internet that allows data exchange between a wide variety of cloud services all interacting with each other. These services map complex scenarios that are distributed across a wide variety of cloud service providers and cloud infrastructure providers. GAIA-X is based on the core principle: “security and privacy by design“. This paradigm is a decisive success factor for GAIA-X, as the cyber attack surface of such a service infrastructure appears to be considerably higher than in pure Industrie 4.0 scenarios. The question as to how security can be viewed holistically in the face of highly distributed accountability across a wide variety of environments (hyperscaler, edge devices, company networks, etc.) is unavoidable. It also applies with regard to different security-relevant levels (network, applications, operating systems, databases, GAIA-X management infrastructure, etc.). Very effective protective measures must therefore be established that work at all levels (prevention, detection, quick response, etc.) in order to secure the entire infrastructure. Such a service infrastructure would require a European trust infrastructure. Since communication participants will at some point be located within this infrastructure and communicate with each other within this infrastructure, the previous examples for secure identity, secure communication, trust infrastructure, access control (transferred from the Industrie 4.0 case to the GAIA-X case) are also fully applicable. Given that Industrie 4.0 scenarios are also included within the GAIA-X service infrastructure, the considerations that apply to the other terms also apply to these.

It makes sense to use AI to provide security for this service infrastructure at all levels. Since the GAIA-X project had only recently been established at the time of writing this document, it is therefore still partly in a definition phase and only a few obvious examples of using AI are listed here.

AI can monitor/check the data exchange between the microservices, edge devices, and generally between the communication participants for non-plausibility, such as:

9 For a description of GAIA-X, see the GAIA-X website [33].

- Unusual measured values
- An unusually high number of requests
- Unusual connection attempts (successful/unsuccessful)
- Unusual querying of data
- Unusual use of APIs
- Unusual establishment of a connection between the microservice/device and various recipients

Since orchestration of the communication channels appears more difficult in the case of GAIA-X than in a pure Industrie 4.0 scenario that has a changing but defined number of participants in a value chain, communication with a fixed, non-AI-based set of rules can hardly be checked. In addition, AI can determine a change in the communication behaviour of individual communication participants and thus draw conclusions about possible manipulations carried out by these participants. Changed communication behaviour of individual participants may be characterised, among other things, by:

- Transmission of different data
- Use of modified login routines
- Changed frequency of communication
- Change in the recipients of participant's messages
- Change in the attributes of the secure identity.

6 Summary and conclusion

No sooner do the current AI research results become part of product developments in the otherwise slowly moving world of industrial innovations than other new trends emerge in the AI area. Huge investments by trillion dollar companies have strong traction worldwide and therefore continue to break up established production strategies in the industrial world.

This document is aimed at industrial users of AI systems as well as developers of AI algorithms in order to clarify the impact that the rapid transformation of artificial intelligence continues to have on industrial production as a whole. Since this fast-moving trend shows no signs of flattening, it is essential to examine and illustrate the various effects on Industrie 4.0 production. First, we need to consider the high-performance potential associated with these AI developments. Second, we need to look at possible trade-offs that may arise from new risks based on bias or inadequate or compromised data sets, which every developer must be aware of. Users “trust” AI-assisted systems with increasingly complex decisions, without being able to think about other back-up options. This fact alone should oblige product managers and developers to pursue “the systemic principles that have been forgotten” [23] when designing and implementing AI-based systems.

Optimisations on every front

Not surprisingly, ongoing technical adjustments and optimisations are also a consistent development goal in the field of artificial intelligence. Only three years ago, the almost unimaginable performance that computers can now generate with machine learning was viewed by many outsiders to be achievable only by large-scale data centres. Neural networks, i.e. the almost exclusive generation technology of such services, are trained in large-scale data centres with an ever-expanding use of resources, but the inference, i.e. the application of the finished network, including the AI system, now takes place locally, at the edge of the Internet in dedicated devices. Optimisation goals therefore require, for example, no wasting of valuable user response time as latency, since the most optimised, compact devices at the edge are used. The most complex neural networks, which are constantly setting new world records in image recognition and similar tasks, are currently running on microcomputers the size of a credit card at minimal costs and response times offering zero network latency and availability that can be scaled almost at will. The level of optimisation is reaching new heights.

Effects on Industrie 4.0

What does all of this mean for users and manufacturers in the field of Industrie 4.0? The previous Industrie 4.0 concept of secure cross-company communication, for example in the area of machine-to-machine (M2M) communication, benefits from additional endpoint security options through AI edge components. AI can be used to provide security to certain areas of application in a simplified and cost-effective manner.

For example, for predictive maintenance, service providers do not need direct access to original data and local sensor data. The machine in question can use AI-based predictive

maintenance features to determine all relevant statements through its own series of measurements and analyses on site at the edge and actively transfer these to the service provider as instructions. There are no confidentiality risks associated with raw measurement data. Operators' fears that maintenance service providers could deduce machine activities from sensory observation are thus largely without foundation.

At the same time, the shift of intelligence from the cloud to the edge also means the emergence of new security risks associated with high computing power in the edge. To give a practical example, four, eight or even more cores of a smartphone's CPU can be used to capacity if necessary when video streaming or other online live services are used. These edge devices have considerable AI capabilities. In addition to the video and voice analyses needed by the user for convenient identification purposes, they also allow extensive espionage and analysis options, including local time series analyses, to be carried out undetected, without the user's consent and without noticeably increasing the processor load. In terms of security, this technology reached a critical point quite some time ago. Trust in this technology is currently evoked by the manufacturer's ecosystem and its promise of reliability. IT security does not play a major role here. So far there has been no reliable Know-Your-Vendor Principle (KYV) in this area, based on the Know-Your-Customer (KYC) concept known in the financial sector.

This situation is to be assessed quite differently in the industrial context: IT security is focused, especially in the Industrie 4.0 area, against threats and aims to prevent extortion due to cybercrime, sabotage by states and services or espionage, such as data theft (illegal copies) due to industrial espionage. Attacks are also aimed at the secondary economic effects generated by accessing the goals, policies and other business-relevant knowledge relating to competitors. In this context, AI-based attacks can only be identified in principle according to the current state of IT security. To enable better detection here, additional sensors are required in the network in order to identify "unusual" behaviours of edge components. The training phases of such systems can last between six to twelve months in order to reduce the false positive rates (FPR) to a tolerable level of false reports. If over-compensation for the behaviour of the detectors takes place, the dangerous false negative rate (FNR) can rise and consequently lead to non-detection of attacks.

Other solutions, such as "security gateways" between network segments, do not reliably detect this type of attack because these network conduits are usually not able to recognise encryption at a semantic level. Messages that can be understood by humans are not recognised and blocked at the AI level. In the era of AI, concepts from the past that checked protocols for encryption can only provide limited protection against AI-based (GAN, adversarial examples, [5] [6]) synthetic keys that are hidden in images, sounds or other report features designed to be harmless. In extreme cases, intensely pretrained AI attackers from the edge no longer exchange any primary data, but simply transmit completely harmless messages. The subversive meaning of these messages can still be guessed one hundred percent correctly by the relevant AI control centre or other edge AI systems in the same network, but is no longer understood by the monitoring AI instance. In this scenario, the attacker achieves complete takeover of the potential victim.

Recommendations and outlook

As explained above, AI applications at the edge are a future-oriented technology that requires expert handling so that the technology's benefits can be reaped with an acceptable associated risk.

Anyone seeking to better grasp or influence these situations in general needs to have good to very good knowledge of their own processes, production systems and infrastructures. In this regard, the KYC principle known from the financial world should be emulated by a Know-Your-Artificial-Intelligence (KYAI) principle in Industrie 4.0. External help for evaluating your own test results is recommended in all cases.

From a security perspective, AI systems at the edge are of particular importance. Recognised weak points must be taken seriously because AI systems are not self-critical and do not protect themselves from their own "blind spots", such as a bias. When monitoring systems for AI edge applications are used, clarification is required as to when and under what conditions escalations or de-escalations must be triggered based on the threat frameworks, e.g. to compensate for a high FPR over a certain period of time. Unknown AI systems, algorithms and unknown data sets generally carry a security risk. Manufacturers of AI-based products need new criteria for creating and maintaining the trustworthiness of AI-based products. These should be based on a KYAI concept that is expressly recommended at this point.

Bibliography

[1] J. Benoit, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam und D. Kalenichenko, Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference, 1712.05877.

[2] Google for Games, "Flatbuffers," [Online]. Available: https://google.github.io/flatbuffers/. [Retrieved on 21.03.2021].

[3] Zentralverband der Elektroindustrie (ZVEI), "AI to Industrial Automation White Paper," Draft 2 2021. [Online].

[4] C. Molnar, "Interpretable Machine Learning A Guide for Making Black Box Models Explainable," 08.03.2021. [Online]. Available: https://christophm.github.io/interpretable-ml-book/index.html. [Retrieved on 11.03.2021].

[5] Plattform Industrie 4.0 AG Security UAG KI, "Umgang mit Sicherheitsrisiken industrieller Anwendungen durch mangelnde Erklärbarkeit von KI-Ergebnissen," 28.10.2019. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Umgang-mit-Sicherheitsrisiken.pdf?__blob=publicationFile&v=12. [Retrieved on 16.02.2021].

[6] Plattform Industrie 4.0 AG Security UAG KI, "Künstliche Intelligenz in Sicherheitsaspekten der Industrie 4.0," 01.04.2019. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/KI-in-sicherheitsaspekten.pdf?__blob=publicationFile&v=8. [Retrieved on 16.02.2021].

[7] S. Waldstein, "The Lancet Digital Health," 01.06.2020. [Online]. Available: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30080-7/fulltext. [Retrieved on 18.3.2021].

[8] "Tiny machine learning," [Online]. Available: https://www.tinyml.org/. [Retrieved on 01.07.2021].

[9] Q. V. L. Mingxing Tan, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," [Online]. Available: https://arxiv.org/abs/1905.11946v5. [Retrieved on 21.03.2021].

[10] P. I. 4.0, "IT-Security in der Industrie 4.0 – Handlungsfelder für Betreiber," [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/leitfaden-it-security-i40.html. [Retrieved on 26.03.2021].

[11] "Coral," [Online]. Available: https://coral.ai/products/#production-products. [Retrieved on 01.07.2021].

[12] P. D. H. Pohl, "Der Patch ist der Angriff – Der Patch ist der Angriff," 21.04.2021. [Online]. Available: https://www.it-daily.net/it-sicherheit/cybercrime/26735-der-patch-ist-der-angriff?start=1. [Retrieved on 11.03.2021].

[13] P. H. O'Neill, "How China's attack on Microsoft escalated into a "reckless" hacking spree," 10.03.2021. [Online]. Available: https://www.technologyreview.com/2021/03/10/1020596/how-chinas-attack-on-microsoft-escalated-into-a-reckless-hacking-spree/. [Retrieved on 11.03.2021].

[14] Sophos, "HAFNIUM: Advice about the new nation-state attack," 05.03.2021. [Online]. Available: https://news.sophos.com/en-us/2021/03/05/hafnium-advice-about-the-new-nation-state-attack/. [Retrieved on 11.03.2021].

[15] Bundesamt für Sicherheit in der Informationstechnik, BSI, "BSI warnt: Kritische Schwachstellen in Exchange-Servern," 05.03.2021. [Online]. Available: https://www.bsi.bund.de/DE/Service-Navi/Presse/Pressemitteilungen/Presse2021/210305_Exchange-Schwachstelle.html. [Retrieved on 11.03.2021].

[16] W. Badr, "Top Sources For Machine Learning Datasets," 13.01.2019. [Online]. Available: https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b. [Retrieved on 11.03.2021].

[17] M. Khairy, “TPU vs GPU vs Cerebras vs Graphcore: A Fair Comparison between ML Hardware”,“ 2020.

[18] N. P. J. e. al, “Ten Lessons From Three Generations Shaped Google’s TPUv4i,“ ACM/IEEE 48th Annual International Symposium on Computer Architecture, 2021.

[19] EIDAS, “ec.europa.eu,“ 01.05.2019. [Online]. Available: https://ec.europa.eu/futurium/en/system/files/ged/eidas_supported_ssi_may_2019_0.pdf. [Retrieved on 18.03.2021].

[20] Wikipedia, “Self-sovereign identity”, 08.03.2021. [Online]. Available: https://en.wikipedia.org/wiki/Self-sovereign_identity. [Retrieved on 18.03.2021].

[21] Plattform Industrie 4.0, “Technischer Überblick: Sichere Identitäten,“ 2016. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/sichere-identitaeten.html. [Retrieved on 11.03.2021].

[22] D. Reed, M. Sporny and D. Longley, “W3C,“ 09.03.2021. [Online]. Available: https://www.w3.org/TR/did-core/. [Retrieved on 18.03.2021].

[23] G. Ropohl, Allgemeine Systemtheorie, Einführung in transdisziplinäres Denken, Berlin: edition sigma, 2012.

[24] Bundesministerium für Wirtschaft und Energie, BMWi, “IT-Sicherheit für Industrie 4.0,“ ß4 01 2016. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/bmwi-studie-it-sicherheit.pdf?__blob=publicationFile&v=5. [Retrieved on 11.03.2021].

[25] Plattform Industrie 4.0, “Sichere Kommunikation für Industrie 4.0,“ 2017. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/sichere-kommunikation-i40.pdf?__blob=publicationFile&v=5. [Retrieved on 11.03.2021].

[26] Plattform Industrie 4.0, “Sichere unternehmensübergreifende Kommunikation mit OPC UA,“ , 2019. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/sichere-kommunikation-opc-ua.pdf?__blob=publicationFile&v=13. [Retrieved on 11.03.2021].

[27] Fraunhofer SIT, Darmstadt, “Eberbacher Gespräch zu “Sicherheit in der Industrie 4.0,“ 10 2013. [Online]. Available: https://www.sit.fraunhofer.de/fileadmin/dokumente/studien_und_technical_reports/Eberbach-Industrie4.0_FraunhoferSIT.pdf?_=1420719894. [Retrieved on 11.03.2021].

[28] Plattform Industrie 4.0, “Vertrauensinfrastrukturen im Kontext von Industrie 4.0,“ planned for 2021. [Online].

[29] Plattform Industrie 4.0, “Zugriffssteuerung für Industrie 4.0-Komponenten zur Anwendung von Herstellern, Betreibern und Integratoren,“ 11 2018. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/zugriffssteuerung-industrie40-komponenten.pdf?__blob=publicationFile&v=8. [Retrieved on 11 03 2021].

[30] Plattform Industrie 4.0, “Kollaborative datenbasierte Geschäftsmodelle,“ 07 2020. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/kollaborative-datenbasierte-geschaeftsmodelle.pdf?__blob=publicationFile&v=5. [Retrieved on 11.03.2021].

[31] Plattform Industrie 4.0, “Details of the Asset Administration Shell,“ 2018/2020 . [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/Details_of_the_Asset_Administration_Shell_Part1_V3.pdf?__blob=publicationFile&v=5. [Retrieved on 21.03 .2021].

[32] Plattform Industrie 4.0, “Security der Verwaltungsschale,“ , 2017. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/security-der-verwaltungsschale.pdf?__blob=publicationFile&v=6. [Retrieved on 11.03.2021].

[33] Federal Ministry for Economic Affairs and Energy, “GAIA-X: A Federated Data Infrastructure for Europe,“ 06 2020. [Online]. Available: https://www.data-infrastructure.eu/GAIAX/Navigation/EN/Home/home.html. [Retrieved on 11.03.2021].

AUTHORS

Dr. Bernd Kosch, Industrie-KI GmbH | Björn A. Flubacher, Federal Office for Information Security | Dr. Dipl.-Ing. Detlef Houdeau, Infineon Technologies AG | Dr. Michael Schmitt, SAP SE | Olaf Dressel, Bundesdruckerei GmbH | Peter Rost, secunet Security Networks AG | Thomas Walloschke, secon trust consult | Dr. Thomas Wille, NXP Semiconductors

This publication is the result of the “AI for I40 Security” sub-working group of the “Security of Networked Systems” working group of Plattform Industrie 4.0.

www.plattform-i40.de