Whitepaper | May 2021

Towards Auditable AI Systems

Current status and future directions

based on the workshop “Auditing AI-Systems: From Basics to Applications”, October 6th 2020, Fraunhofer Forum, Berlin

Christian Berghoff[1], Battista Biggio[2], Elisa Brummel[3]*, Vasilios Danos[4], Thomas Doms[5], Heiko Ehrich[6], Thorsten Gantevoort[7], Barbara Hammer[8], Joachim Iden[7], Sven Jacob[1], Heidy Khlaaf[9], Lars Komrowski[10], Robert Kröwing[7], Jan Hendrik Metzen[11], Matthias Neu[1], Fabian Petsch[1], Maximilian Poretschkin[12], Wojciech Samek[13]*, Hendrik Schäbe[7], Arndt von Twickel[1]*, Martin Vechev[14] and Thomas Wiegand[13] (Authors are listed in alphabetical order)

[1]Federal Office for Information Security, Bonn, Germany, [2]University of Cagliari, Italy, [3]TÜV Association, Berlin, Germany, [4]TÜViT, Essen Germany, [5]TÜV Austria, Vienna, Austria, [6]TÜV Nord, Essen, Germany, [7]TÜV Rheinland, Cologne, Germany, [8]Bielefeld University, Germany, [9]Adelard LLP, London, Great Britain (current affiliation: Zipline, San Francisco, CA, US), [10]TÜV Hessen, Darmstadt, Germany, [11]Bosch Center for Artificial Intelligence, Renningen, Germany, [12]Fraunhofer IAIS, Sankt Augustin, Germany, [13]Fraunhofer HHI, Berlin, Germany, [14]ETH Zürich, Switzerland

*Contact:
Arndt von Twickel (arndt.twickel@bsi.bund.de),
Wojciech Samek (wojciech.samek@hhi.fraunhofer.de) and
Marc Fliehe (marc.fliehe@vdtuev.de)

Executive Summary

Artificial Intelligence (AI) systems are playing an ever growing role as part of decision and control systems in diverse applications, among them security- and safety-critical application domains such as mobility, biometrics and medicine. The use of AI technologies such as deep neural networks offers new opportunities such as a superior performance as compared to traditional IT technologies. At the same time, they pose new challenges with regard to aspects such as IT security, safety, robustness and trustworthiness. In order to meet these challenges, a generally agreed upon framework for auditing AI systems is required. This should comprise evaluation strategies, tools and standards but these are either under development or not ready for practical use yet.

This whitepaper first summarizes the opportunities and challenges of AI systems and then goes on to present the state of the art of AI system auditability with a focus on the aspects AI life cycle, online learning and model maintenance in the presence of drift, adversarial and backdoor poisoning attacks and defenses against these attacks, verification, auditing of safety-critical AI systems, explaining black-box AI models and AI standardization.

Despite substantial progress for all of these aspects, an overarching open issue is that of (often multi-faceted) trade-offs between desired characteristics of the system, e.g. robustness, security, safety and auditability, on the one hand and characteristics of the AI model, ML algorithm, data and further boundary conditions on the other hand. These trade-offs restrict the scalability and generalizability of current AI systems.

To eventually allow leveraging the opportunities of AI technologies in a secure, safe, robust and trustworthy way, two strategies should be combined: 1. Taking the abovementioned trade-offs into account, favorable boundary conditions for the given task should be selected; 2. Available technologies should be advanced by substantial investments in R&D to eventually allow for secure and safe AI systems despite complex boundary conditions and, therefore, to improve scalability and generalizability. In a first step, one should focus on selected security- and safety-critical use cases. Available standards, guidelines and tools should be exploited and interdisciplinary exchange between researchers and industry should be further promoted to find the best combinations of available criteria and tools for achieving auditable, secure, safe and robust AI systems for each specific use case. Insights from these use cases should then be used, in a second step, to generalize the results and to build up a modular toolbox that may subsequently be applied to other use cases. On this basis, first technical guidelines and subsequently standards should be developed. In the ideal case, the outcome will be a generally applicable set of criteria and tools that allows making AI systems sufficiently auditable, safe and secure.

Table of Contents

1 AI systems: opportunities and challenges

Already prevalent in many applications, artificial intelligence (AI) technology[1] is increasingly becoming an integral part of our world as it is the basis for countless applications that employ decision or control systems (Fig. 1). AI systems may consist of multiple subsystems, each of these possibly using different technologies. Technologies may be grouped into classical IT (cIT), symbolic AI (sAI) and connectionist AI (cAI)[2]. Here, a focus is put on cAI systems in the form of (deep) neural networks and machine learning (ML), because cAI systems exhibit qualitatively new vulnerabilities and can, as of now, not be sufficiently audited by means of available tools from cIT.

AI is employed in rather harmless applications such as computer games and voice assistant systems as well as in safety-critical applications such as driver assistance systems, intrusion detection systems and in medical diagnosis [1-4]. The latter use-cases demonstrate that responsibilities and liabilities are transferred from humans to AI systems in safety- and security-critical systems. Therefore, malfunctioning AI systems may lead to serious consequences resulting in financial loss or even affecting the health of humans. In extreme cases, this could include fatalities in car accidents or serious medical conditions due to inappropriate or missing medical treatments. In many applications, current AI systems substantially outperform cIT technology in terms of performance, user experience and cost. Despite these and other great opportunities offered by AI technology, its application comes with several challenges [5-9]: for instance, the inner workings of neural networks (NN) are very hard to interpret by humans due to their highly interconnected non-linear processing elements and their huge input and state spaces. Also, their performance is highly dependent on data quantity and quality since their parameters have to be trained by ML algorithms. NN training does not follow well-defined design processes, NNs possess qualitatively new vulnerabilities, there often is a lack of trust by users and NNs may be used as powerful tools by attackers.

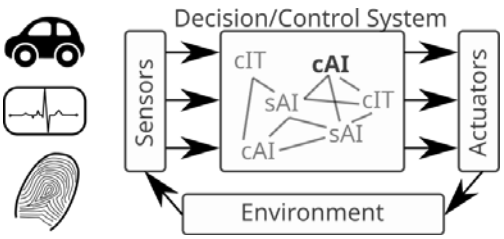

Fig. 1: AI systems are already part of decision and control systems in diverse applications, e.g. autonomous cars, healthcare and biometrics. Connectionist AI (cAI, e.g. neural networks), symbolic AI (sAI, e.g. decision trees) and classical IT (cIT) modules interact with each other and with the environment via sensors and actuators, thereby leading to the overall system behavior. Here we focus on single cAI modules (bold print) only. [Icons are released under CCO 1.0 and were downloaded from svgsilh.com]

In order to address these safety and security challenges of AI, it is, therefore, mandatory to gain a deeper insight into how AI systems function, why they perform well in some cases but fail in

[1] AI systems are here defined as automated artificial systems that either support humans in making decisions or that autonomously take decisions. For an alternative definition, cf. e.g. the definition by the high level AI expert group of the EU commission: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.

[2] Whereas it is hard to draw a clear border between cIT and AI systems, AI systems may be clearly classified as either symbolic AI (sAI) systems or as connectionist AI (cAI) systems: the former may be directly constructed by a human developer, the latter have to be trained with machine learning and data.

others and how they might be attacked and protected from attacks. In order to gain users' trust, the proper operation of an AI system has to be guaranteed under realistically defined boundary conditions. For "classical" technologies, such guarantees are required by law in several domains such as in airplane control software and need to be audited on a regular basis. A natural question is how to transfer the concepts and methods from the classical IT domain to the AI domain and, where this is not sufficient, how to complement them with new AI-specific concepts and methods. If proper operation cannot be guaranteed in 100 % of the cases, it should be discussed if it is acceptable that AI systems at least perform better than state-of-the-art non-AI systems or humans. Hereby, a risk-based approach should be taken by quantifying the risk for system failure, i.e. the cost of failure multiplied with the probability of failure. This should also hold in case of malicious attacks. Better average performance may not be sufficient since the performance of an AI system may be better on average but much worse on subgroups (e.g. skin cancer detection for black people, [10; 11]). If AI systems fail, the reasons for their failures have to be explainable. Since no generally accepted standards, evaluation criteria, methods and tools for auditing AI systems are currently available (but see section 2.7 for current initiatives), the following questions arise: How to audit AI systems? Which boundary conditions are optimal, which are acceptable? Which methods, tools and other resources are needed in addition to classical IT system audits or safety assessments? What are the limits of auditing AI systems? What is the trade-off between effort and audit quality? How should available resources be employed best in research and development to achieve AI system audit results that remain valid under a wide range of conditions?

Based on the presentations and discussions during the one-day workshop „Auditing AI-Systems: From Basics to Applications" on October 6th, 2020 in Berlin/internet, we try to give answers to these questions by reviewing the current state of the art of AI system auditability, by summarizing open questions and by identifying the most urgently needed future work and the most promising approaches. In doing so, 1. the whole life cycle of AI systems will be considered and 2. a focus will be put on the currently most important AI technology, namely deep neural networks (DNNs) trained by machine learning (ML), and DNNs will be considered in terms of IT security, safety and robustness. Where possible, specific use cases will be given as examples.

2 Auditability of AI systems: state of the art

In this section, first an overview of a generalized cAI life cycle (Fig. 2A) is given in order to then summarize the state of the art of some of the most important aspects of cAI system auditability, namely the training of AI systems via ML and data, attack and defense, verification, validation, interpretability and standardization.

Further aspects were not covered in depth during the workshop and some of them are, therefore, only shortly summarized with respect to their possible impact on AI safety and security:

1. Sufficient quality and quantity of training and test data are of paramount importance for the performance and robustness and, therefore, also for the security and safety of the AI system [12].
2. Data pre-processing (or feature selection) can be, on the one hand, seen as a step towards modularization of the AI system, possibly leading to better interpretability due to reduced functionality of each AI module, but, on the other hand, can be considered to open up a new attack target (cf. e.g. [13; 14]). Therefore, depending on the context, it may or may not be beneficial for increasing safety and security.
3. Regularization, such as the penalization of large weights via the error function, might serve to prevent overfitting and might, under certain boundary conditions, directly lead to higher robustness and indirectly to increased safety and security [15].

2.1 Life Cycle

The complex life cycle of cAI systems is, at least to a large extent, responsible for new challenges regarding their application, especially when compared with cIT and sAI systems. Therefore, it will be at the focus of this whitepaper. Here, it is divided into the following 5 phases (cf. Fig. 2A): planning, data, training, evaluation and operation. In practice, these phases are not ordered sequentially but the developer rather uses these phases in a highly iterative and agile manner, e.g. evaluations are employed frequently during the development. Also, the operational phase includes the challenge of model maintenance, including the necessity of adapting the model in case novel data or requirements arise for a cAI system already in use. Similar in spirit to biological neural networks, cAI systems typically consist of a large number of simple yet highly interconnected processing elements (or neurons) which are organized in layers. State-of-the-art cAI systems, such as deep neural networks (DNNs, deep = many layers) consist of millions of processing elements and synapses (= connections) in between them. Assuming a fixed neural architecture, this means that cAI systems often have more than 100 million parameters, i.e. synaptic weights and unit bias values, that have to be appropriately tuned. It is, therefore, in almost all cases impossible to set these parameters by hand. Instead, machine learning techniques are employed to automatically adjust the system's parameters based on training data, an error function and a learning rule. In contrast to parameters internal to cAI models that are learned during training, external parameters that influence the learning process and the model architecture are called hyperparameters and have to be fixed before training and tuned on a

validation set. Approaches to automate the training pipeline setup and training itself are called automatic machine learning or AutoML [16]. While parameters in numerous cIT and sAI models, such as decision trees or rule sets, are often also set by automatic methods, they can in principle and in contrast to most cAI models, still be intuitively inspected[3,4].

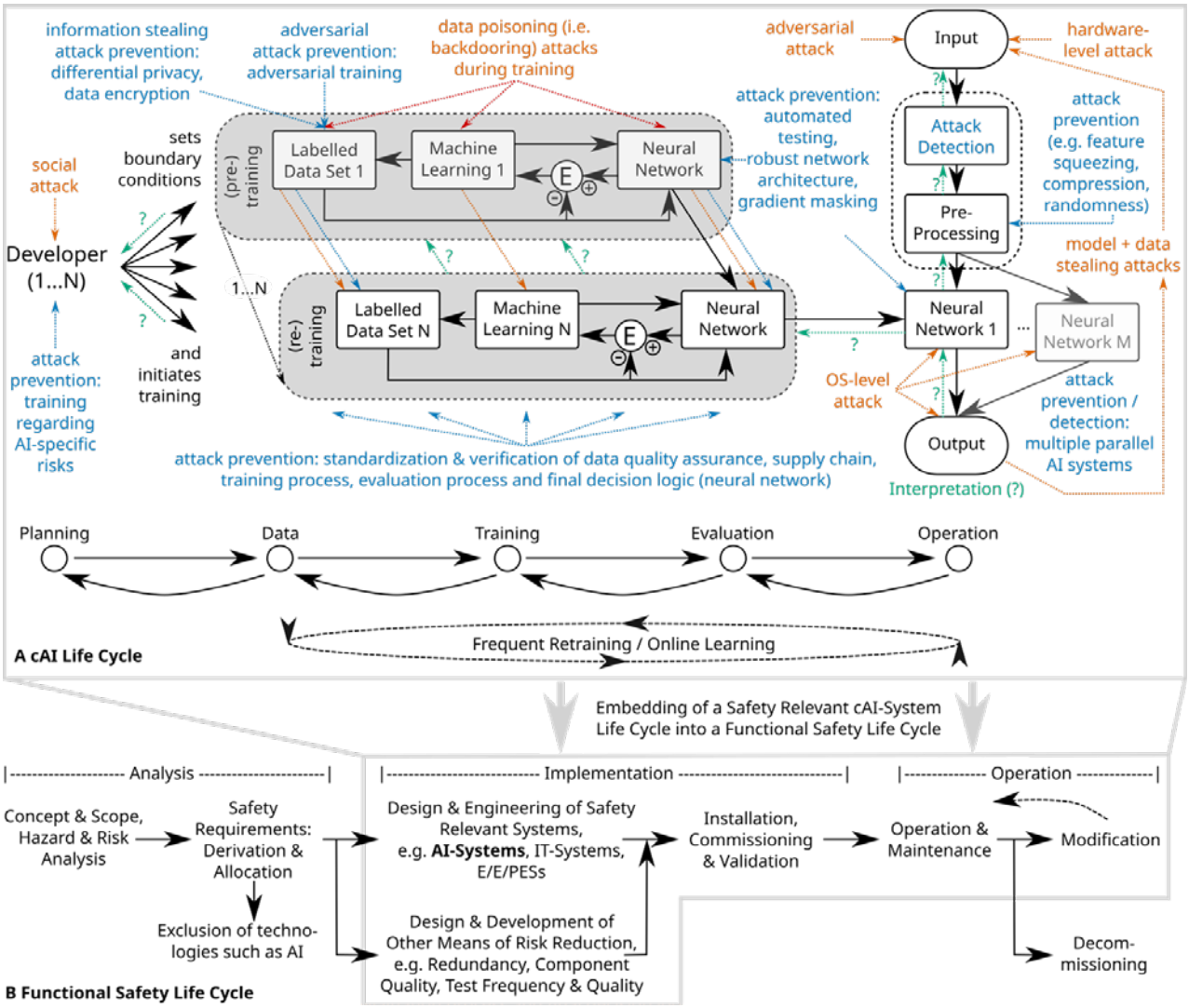

Fig. 2: A) The schematic depiction of a generalized life cycle of a connectionist AI (cAI) system underlines that many aspects have to be considered for a thorough audit of an AI system. Here, the life cycle is viewed from the IT-security perspective with vulnerabilities (red), defenses (blue) and interpretation (green + "?"). Supervised retraining or online learning may optionally run in parallel to and continually during operation leading to the question when and how frequently evaluation should take place. The whole life cycle should be considered for evaluation, verification, validation and standardization. B) A cAI life cycle (cAILC) may e.g. be embedded into a functional safety life cycle (fSLC, cf. e.g. [17; 18]). The latter may encompass several safety-relevant systems, e.g. also sAI and cIT systems (cf. Fig. 1) and includes an extensive analysis phase prior to the development phase and a final decommissioning phase. Note that both the cAILC and the fSLC are usually highly iterative (cf. the feedback connections).

[3] Note that linear models or decision trees might not be, in a practical sense, per se interpretable because of their size and complexity.

[4] A comparison of different AI and ML methods such as deep learning and reinforcement learning with respect to their impact on safety and security and their interpretability is, unfortunately, out of scope here.

The role of the developer is, therefore, to set the necessary boundary conditions by initializing the training process with a neural network, training data, a machine learning algorithm and relevant hyperparameters. Subsequently, the developer supervises the training process, adjusts hyperparameters, tests intermediate results and – if necessary – restarts the training until the desired performance of the AI system is achieved. This is not a standardized procedure but rather intuition and experience of the developer determine the training process. Due to the large amount of resources necessary to obtain sufficient amounts of high-quality data and to train DNNs from scratch, developers often take shortcuts and make use of pre-trained models and external data obtained from various sources. Once the development criteria (e.g. performance, robustness) are met, the AI system may be put into operation: after having been embedded in a specific hard- and software environment, the neural network is fed with pre-processed input data and outputs its decisions. Despite the lack of a clearly defined design process (see above), an experienced developer with access to the necessary resources (data, models, computing power) may quickly develop decision systems for many use cases that clearly outperform cIT systems.

As a drawback of the often huge parameter and input spaces of DNNs and their non-intuitive relation between structure and function, it is mostly impossible for a human to interpret their functioning. Specialized interpretation methods that allow to do so are the subject of current research (for details see section 2.6). AI systems are currently tested by observing the input-output relation for a selected set of test inputs. To test even a small subset of all possible inputs requires considerable resources and has to be approached systematically (cf. [19]). A formal verification is only possible in specific cases under very restricted boundary conditions, e.g. it does not scale to large networks and arbitrary inputs (cf. section 2.4). Further drawbacks of cAI systems are their qualitatively new vulnerabilities, namely adversarial attacks (see section 2.3.1) and information stealing attacks (cf. section 2.3) during operation, and backdoor poisoning and DoS attacks during training (cf. section 2.3.2), which, in addition to classical social attacks, operating system and hardware attacks, may be exploited by attackers for targeted and untargeted attacks (see below for details). To secure data-driven AI systems and ML in general against such attacks, many solutions have been proposed in addition to classical measures of IT security, amongst others: adversarial training, gradient masking and feature squeezing (cf. section 2.3.3 for more details). Unfortunately, there is, as of now, no single defense method and no combination of multiple ones that is able to reliably prevent adaptive attacks. Also, depending on the setting, improved attack prevention and robustness may come at the cost of decreased accuracy [20].

In practical use cases, such as in autonomous cars, a cAI life cycle is generally embedded into more extensive life cycles that include the development and interaction of multiple IT and AI modules. This is depicted for the functional safety life cycle (fSLC in Fig. 2B), where cAI modules are just possible components amongst others. For these cAI modules an (automotive) safety integrity level ((A)SIL) may be determined [17; 18]. The functional safety life cycle puts emphasis on the analysis phase preceding the planning phase of the AI life cycle with the goal of quantifying the probability of failure of such systems and of determining the acceptability of these probabilities by systematic means including risk analysis. The fSLC analysis phase also includes conceptualization and the

derivation and allocation of safety requirements. As a result of the analysis, the use of AI technology in safety- and security-critical applications might even be completely banned for reasons of security and safety. In contrast, AI can be easily used in situations where no critical consequences occur, which has to be supported by a risk analysis. In that case, no SIL requirements need to be implemented in the system and safety assessment is not necessary. Methodological and use case specific standards, norms and technical guidelines should be used wherever they are applicable throughout the life cycle. For instance an extensive standard [17] exists for the functional safety life cycle but it does not include the cAI life cycle with cAI specific vulnerabilities and challenges. Around the world, multiple initiatives strive to close this gap (cf. Section 2.7).

2.2 Online learning and model maintenance in the presence of non-stationary environments

In order to solve problems by learning from data, different paradigms can be used depending on the complexity of the problem and the amount of data available. For instance, deep learning techniques are typically used to solve complex problems when a large amount of training data is available, whereas classical methods from statistics can only address less complex problems, but require less data to do so. Independent from the paradigm, the environment of the problem at hand is likely not constant over time. In order to obtain robust results, such environmental changes must be taken into account and addressed.

For most classical machine learning (ML) techniques, strong robustness guarantees can be derived from statistical learning theory under standard assumptions [21]. Another way to maintain prediction accuracy in the face of environmental changes and limited data availability is to allow the ML models to reject inputs which are too far away from known data points and where the models' certainty is low [22]. It needs to be noted that identifying such inputs can be a hard problem in itself. This approach can also be used in online learning [23].

Transfer learning is a general technique that allows adapting a previously learned parent model to a new but related task [24]. Using the similarity of the two tasks and building on the information contained in the parent model, it is possible to train a new model using much fewer data points than training it from scratch would require. Transfer learning and also few-shot learning in a more general form are currently the standard way in which deep learning is used. For instance, models for specific image classification tasks build on pre-trained models such as VGG [25]. Transfer learning can be used to cope with environmental changes. However, in order to obtain theoretical guarantees on the accuracy of the resulting model without using a large amount of data, strong assumptions about the changes that can occur are necessary. Such assumptions may be valid in practical use cases, e. g. adapting the control unit of a limb prosthesis to slight shifts in sensor location [26].

Another method for training ML models is called online learning. In this paradigm, a model does not learn from discrete batches of data, but instead uses a data stream and is constantly updated

to take each new data point into account. Environmental changes then manifest as data drift, which can affect the true model itself or only the data distribution observed. In this setting, the challenge is to decide which information is relevant for making correct predictions at a given time point and future ones, and which information should be discarded. In doing so, also data poisoning attacks and missing data labels have to be considered. Models hence face a dilemma between plasticity, i. e. being able to integrate new information, and stability, i. e. maintaining previous knowledge that is still relevant. It has been shown for simple models that these two properties can be efficiently balanced to achieve high performance in the presence of drift [27-30]. The challenge of such models is that meta-parameters become model parameters, since model complexity might change. Hence, non-parametric models as well as ensemble methods are often particularly well-suited. However, obtaining mathematical guarantees requires very strong assumptions. As a step towards dealing with drift in practice, first techniques to detect and understand drift provide interesting approaches to judge the effect of such online adaptation techniques [31; 32].

2.3 Attack & Defense

AI is not secure by design and countless examples of tricking AI systems have been documented over the last years (for an overview cf. [33]). In this whitepaper, we focus on the two most important vulnerabilities of AI systems with respect to the information security goal integrity which strives to maintain trustworthy and consistent data throughout the entire AI life cycle. In this context, two main and qualitatively new threats to cAI systems have been identified: *adversarial or evasion attacks (cf. section 2.3.1)* during the operational phase and *backdoor poisoning attacks* (cf. section 2.3.2) during the training phase. These attacks and available defenses are discussed in detail in the following sections.

Further vulnerabilities exist with respect to the other two main information security goals confidentiality and availability but are not in the focus of this whitepaper: confidentiality may be compromised via *exploratory model stealing [34], model inversion [35]* and *membership inference attacks* [36] where AI models and data used for training may be reconstructed from queries to the operational AI system (summarized in Fig. 2 under “model and data stealing attacks”). These attacks are mentioned in the context of evasion attacks (see below). Availability may be attacked by *DoS poisoning attacks* [37], which, in contrast to backdoor attacks, have the goal of minimizing the models performance.

2.3.1 Adversarial Machine Learning

In *evasion attacks* an attacker plans to change the decision of an AI system during its inference (or operation) phase by subtle modifications of the model input. The modifications are often unsuspicious to the human eye and are also called adversarial examples *[38; 39]*. As a result, standard cAI systems are very brittle and inputs that are not well represented by the models’ training data are especially vulnerable to misclassifications. Well-known examples include e.g. attacks on traffic sign classification systems [40] by placing stickers on a traffic sign, attacks on

malware detectors by adding code that is not required for proper functionality to the malware [41-43] and attacks on biometric identification systems by equipping a human with a specially printed glasses frame [44] or patch on a hat [45]. If the attacker is able to control the decision of the AI system, the attack is called a *targeted attack*, and otherwise, if the attacker just changes the decision in an arbitrary way, the attack is called an *untargeted attack*.

In order to prepare an evasion attack, it may be formalized as an optimization problem which has the goal of modifying inputs in such a way that they cause the AI system to cross at least one *decision boundary*, e.g. from a benign to a malicious region in the malware detector [38; 46]. In doing so, several side conditions have to be taken into account, e.g. the requirement to keep the modification as small or unnoticeable as possible.

If the attacker has perfect knowledge of the model, the features and the data, the attack is called w*hite-box attack*. If, additionally, the output function is differentiable, which is the case for most of the currently used learning algorithms, then a gradient may be computed as a prerequisite for the optimization procedure. But also in the case in which the attacker only has limited knowledge of the *target model*, the feature and the data, called *grey-* or *black-box* setting, effective attacks may be crafted by the attacker using a bypass via *substitute models.* Substitute models may be derived either via model stealing attacks or via newly trained models, e.g. using data from a membership inference attack, which mimic the functionality of the target model. cAI systems have the property that attacks developed for one model can in many cases be transferred to different cAI models without much effort (*transferability*) and, accordingly, these attacks are also called *black-box transfer attacks*. Depending on the boundary conditions, even *black-box query attacks* can be successful. They do not require substitute models but instead use queries to the target model combined with gradient-free optimization methods such as genetic algorithms or bayesian optimization. Due to these black-box attacks, it is not sufficient to keep the parameters of the network secret to effectively protect the AI system against adversarial attacks.

But why are cAI systems vulnerable to adversarial attacks? cAI systems are built on the assumption that the training data is representative of future data, i.e. that input data is independently and identically distributed (IID). Unless the task space is very limited, the *IID assumption [47]* is violated sooner or later, meaning that the model exhibits a *lack of robustness.* Hereby, the lack of robustness of the model with respect to random input data corruptions (noisy input data distributions) and specially crafted adversarial examples are two manifestations of the same underlying phenomenom [48]. The more complex the model becomes, the more vulnerabilities arise, and the easier and quicker an attacker can find adversarial examples. Intuitively, this may be explained by the fact that the larger the input and state space dimensions of the system get, the more short paths from legitimate input to malicious input regions exist that may be exploited by an attacker. Additionally, in order for robustness training to work for complex cAI systems, it needs a larger quantity of appropriate training data, i.e. defenses become more and more resource-intensive with the size of the cAI system. One strategy to cope with this is to consider a risk perspective, where for each type of attack its likelihood to occur is considered to decide how much resources one should allocate for defending against it.

2.3.2 Backdoor Attacks on DNNs

AI models like DNNs need large quantities of data for training and testing in order to achieve good performance. For this reason, very often it is common practice to gather data from many sources without enforcing high quality standards. Indeed, a common belief among practitioners is that low-quality data may be of little worth but cannot significantly impair a model's performance. However, numerous results from research have shown that this assumption is incorrect. Since current AI models are in essence pure correlation extractors, issues with data sets induce them to behave in unintended ways.

Backdoor poisoning attacks and DoS poisoning attacks [49; 50] corrupt parts of the training data in a targeted way. On the one hand, DoS poisoning attacks aim to degrade models' capability to generalize by inserting wrong data points to shift their decision boundary [49]. While these attacks pose a large problem in classical ML methods, they do not impact DNNs on the same scale, and can often be detected quite easily [51]. On the other hand, backdoor poisoning attacks only degrade model accuracy on some inputs [50]. To this end, attackers carefully manipulate part of the training data by adding special trigger patterns, which enable them to fully control the model behavior on these inputs during inference. In terms of the classical IT security goals, DoS poisoning attacks impact the availability of models, whereas backdoor poisoning attacks target their integrity. The fundamental idea of such attacks consists in planting fake correlations, which the models will then use for their decisions. For instance, this often involves changing labels in classification tasks. However, more subtle, so-called label-plausible attacks can avoid these rather tell-tale changes [52].

Backdoor attacks on DNNs are hard to detect afterwards. This is both due to the fact that the models only do what they are supposed to, namely learning correlations, and due to the lack of human interpretability they exhibit. Approaches to uncover backdoor attacks rely on the detection of outliers learned by the models [53]. This does not work well on the data set itself, but rather the internal model representations must be used [50; 54; 55], possibly in conjunction with XAI methods (cf. section 2.6). However, existing mitigation techniques are not perfect, and no automatic solution will likely be, since human prior knowledge may be necessary to properly distinguish corrupted and benign data points [56].

In addition to targeted attacks, data sets may contain spurious correlations, which can affect models in a similar, albeit less targeted way. These correlations may stem from bias in the selection of the data as well as the pre-processing and training pipeline. Such problems have for instance been uncovered in various tasks in medical image recognition [57].

Addressing these issues requires eliminating spurious correlations from training data. XAI methods may be helpful to do this, as well as techniques to randomize pipeline artefacts during training. Besides technical measures on the AI level, more general mitigation techniques will be necessary to address unintended spurious correlations and especially to thwart backdoor attacks. In particular, this includes protecting the integrity of models throughout their life cycle and using technical and organizational measures to change the ambient conditions during the training

phase, such as a security screening of the developers and restricted access to data storage and development machines, making it harder for attackers to be successful [58].

2.3.3 Detection of and Defenses against attacks on DNNs

In the recent past, a large amount of methods have been proposed in order to protect deep neural networks from attacks [59] or to detect such attacks [60]. However, it has turned out to be very difficult to detect adversarial attacks and to reliably defend against them, because it has been shown that an adaptive attacker can circumvent most proposed defenses, and even multiple defenses applied in parallel might not always increase the adversarial robustness compared to a system that only applies the strongest defense [61-63]. Nonetheless, defenses can increase the effort for an attacker to launch a successful attack. Moreover, recent work on certifiable detection of adversarial attacks is promising because it guarantees robustness against certain adaptive attackers [64].

One important drawback of many defense methods stems from the fact that they can significantly affect the model's performance on benign inputs. For this reason, a suitable metric to evaluate a defense method should account both for the model's performance to a) benign inputs and b) adversarial inputs.

When defending against adversarial attacks, it is always necessary to consider the ambient conditions of the AI system. For example, if an attacker can only apply the attack in the physical world and not in the digital domain (e.g. when attacking a computer vision system, the attack needs to be robust under different perspective, rotations or similar transformations), the bar for successful attacks is much higher. Additionally, it needs to be kept in mind that the robustness of such a system does not only depend on the robustness of its AI-related parts, but also on other components, such as cIT, which can both increase and decrease the system's robustness and also constitute an additional target for attacks. For instance, the system's robustness might be increased by the inclusion of a redundant method based on non-cAI technology which acts as a sanity check or by hampering the crafting of adversarial examples via cIT query limits to the cAI components.

One of the most promising defenses against adversarial attacks is adversarial training [59], where adversarial examples are included into the training phase in order to increase the adversarial robustness of such a system. One drawback of this method is that it can significantly affect the training runtime, especially when including examples constructed using strong attacks. Adversarial training will only confer robustness to attacks seen during training, and as a consequence, if only weak attacks are considered for performance reasons, the system will remain vulnerable to stronger ones. As a result, it is necessary to improve the efficiency of adversarial training, especially via the process of creating strong adversarial examples during training as done in Shared Adversarial Training [65] and Meta Adversarial Training [66] but also other extensions of the training strategy are promising (cf. e.g. [67]).

Another drawback of adversarial training is that it does not give any formal guarantees regarding the model's robustness. Therefore, it cannot be formally proven that no attack exists which circumvents this defense[5]. This problem might be solved in threat models such as adversarial patches [68] by certified defenses such as e.g. [69] and [70], which can prove robustness to adversarial attacks for the patch threat model. However, for other threat models such certified defenses would severely affect the model's performance on benign inputs. Additionally, some of these defenses impose restrictions on the model's architecture.

Further classes of defenses against adversarial attacks are often prone to circumvention by attackers and might, depending on the use case and the boundary conditions, give a false sense of security. This is e.g. the case for gradient obfuscation [71], a kind of gradient masking which is supposed to make the attack optimization step harder.

When it comes to defending against backdoor attacks, the major problem stems from the fact that the AI model does not have a prior knowledge of its target domain and learns this knowledge from the (possibly malicious) training data. One promising way in defending against such attacks is to detect malicious data by looking at the inner workings of a deep neural network trained with this data [54] and identifying samples on which the network behaves differently from other data samples of the same category. This might indicate that the network uses different features in order to make its predictions as compared to normal data samples. So far, this approach only works in part of the cases. To solve the problem of the missing prior of the model, it might be necessary to include this prior via human expert knowledge in an interactive process that also uses XAI methods.

2.4 Verification of AI systems

The area of verification of AI systems deals with proving the absence of unintended output behavior in the presence of a range of input perturbations, which can be due to natural variations or be induced on purpose by an attacker. Verification can hence be used to reason about the safety and security of AI systems. However, rigorous proofs face significant obstacles. Due to the large input space, the number of perturbations to consider is potentially unlimited, which makes brute-force approaches infeasible. In addition, standard solvers for checking logical constraints (e.g. SMT, [72; 73]) do not scale well to DNNs due to their non-linearity, although they may be useful to some extent.

A prominent approach to tackle these problems is based on the technique of abstract interpretation, which has been widely used in automated reasoning for years [74]. Its main idea is to represent a possibly infinite number of states in a bounded, finite way, which allows storing it in memory and performing symbolic computations.

More precisely, abstract interpretation can be applied to DNNs by encoding all possible input

[5] This is not unique to AI systems and also the case for any connected IT system.

perturbations via symbolic constraints, for instance giving rise to a polytope[6]. Subsequently, the abstract effects of the network layers on this polytope can be computed. The resulting shape encodes all possible outputs corresponding to the input set and may be used to check the guarantees to be verified. In practice, in order to make computations feasible, the symbolic constraints encoding the inputs are approximations (convex relaxations) of the true data manifold. Therefore, there exists a trade-off between the precision of the approximations and the computing complexity.

The verification techniques developed so far have several shortcomings, leading to the following suggestions of improvements to be made:

1. Verification has been mostly carried out with respect to random variations of each element of an input vector within given bounds[7], and only recently has work on geometric perturbations (e.g. rotation, translation) started. This scope will need to be extended to more semantic perturbations.
2. The relaxations used need to be improved to achieve a better trade-off between precision and complexity. Custom relaxations may be needed when extending the set of perturbations and tasks.
3. The techniques have mostly been applied to classification tasks with feedforward neural networks and need to be generalised to cover other model types (e.g. RNNs) and other tasks (e.g. segmentation).
4. The biggest issue is the scalability of the approaches. If one aims to provide 100% deterministic guarantees, the techniques only work for small to medium networks (in terms of the number of ReLU units), falling short of large-scale networks used in practice.

In order to benefit from the full potential of these techniques, they can also be used beyond ex-post verification. In particular, an approach called certifiable training [75] combines them with the training procedure to obtain certifiable defenses. This can also help to address the scalability issue of the technique, as new network architectures become amenable to certification.

It has also been shown that adversarial training facilitates verification, and adversarial and certifiable training can be related and differ mostly in the information they use to improve a model's robustness. Recent research has proposed a way to combine the two approaches [76; 77].

2.5 Auditing safety-critical AI systems

A safety-critical AI system is a system whose decisions are influenced by an AI subsystem and whose failure or malfunction may result in the following outcomes: death or serious injury to people, loss or severe damage to equipment or property, and environmental harm. For example safety-critical systems can be found in the areas of aviation, nuclear energy, automotive & rail, medical and autonomous systems. For these systems, it is necessary to show that they meet the

[6] When using the convex hull of a finite set of points

[7] e.g. with respect to lp-norms, cf. [22] for details.

needed requirements, like for example certain predictable robustness and reliability, and their assurance often relies on a standards-based justification. Unfortunately, in the case of ML-based systems, this is a serious problem: validated standards, policies and guidance for such novel technologies are lacking, e.g. prescriptive software standards for safety such as IEC 61508 [17] are not fully applicable to AI systems.

As in other systems where existing methodologies cannot be applied, an argument-based approach, which uses formal structured argumentation for justification of some specified claims, may be used as a structured way for AI system assurance [78; 79]. The key advantage of an argument-based approach is that there is considerable flexibility in how the safety claims are demonstrated. Such a flexible approach is necessary when identifying gaps and challenges in uncharted territory. One of these approaches is the CAE (Claims, Arguments, Evidence) framework, which is based on applied natural language deductivism approaches. The CAE framework consists of three components:

- Claims which are assertions put forward for general acceptance (e.g. claims about the system being secure/safe).
- Arguments which link evidence to a claim.
- Evidence which serves as justification of a claim. Sources of evidence can for example include development processes, prior experience, testing and formal methods.

Using the CAE framework, a given claim can be examined using classical and AI specific methods in a structured way. Classical software analysis approaches are for example necessary in analyzing the software code in which the AI system is implemented. On the other hand, classical approaches cannot be applied when it comes to qualitatively new AI-related aspects, such as adversarial attacks. CAE may be further extended to include a promising variant of the argument-based approach, defeasible reasoning [80], by using counterclaims and confirmation theory [78]. It reduces the likelihood of confirmation bias by prompting the assessors to repeatedly ask the question of why something may not be safe instead of solely looking for supporting evidence.

For some AI system key properties, like for example the *system's robustness*, a clear formal definition as a pre-requisite for any kind of formal verification is missing. CAE may help to clarify this open research issue and to make an effort in defining these properties.

The most common formal property of AI systems which can somehow be proven is the property of pointwise robustness. However, a major limitation of this property stems from the fact that it does not imply the system robustness property: pointwise robustness only proves a given property for a specific data sample, but in order to show the system robustness it would be necessary to show this for all future inputs, which is infeasible in most of the practical applications where AI systems are used [72].

As a result of this, it is currently not possible to fully audit AI systems on the level of formal verification. Static analysis tools can, however, be used to prevent propagation of errors from the training code into the ML algorithm and help to provide a baseline for the security of the system. Existing good practices in AI auditing include [81-84].

2.6 Explaining Black Box AI Models

Complex AI models such as deep neural networks (DNNs) learn functions by training on large data sets (see section 2.1). The inner workings of these models, which mathematically encode the functions learned, do not usually lend themselves to human interpretation [85]. However, being able to explain and interpret the decisions of an AI model can be important for a number of reasons. These reasons range from finding faults, weaknesses and limitations of a model (and the HW/SW-platform where it is implemented), which may serve to improve its performance and robustness to attacks, to fulfilling requirements for transparency, as for instance mandated by the EU General Data Protection Regulation, and to gaining new insights from large data sets in science and in the economy. Therefore, new methods are required for explaining complex AI models like neural networks. The corresponding research field is called XAI (explainable AI) [86; 87].

Various explanation methods have been proposed in the literature to provide insights into different aspects of an AI model. One class of methods aims at *global* interpretations of the model, e.g., by analysing the extremal points of the encoded function through the construction of maximum activation inputs [88] or by investigating the role of individual neurons in the deep neural network [89]. While these explanations certainly provide valuable information about the model and its learned representation, they are of little use to understand individual predictions, i.e., to identify the input features which contributed positively or negatively to the model's decision. *Local* XAI methods fill this gap by attributing relevance scores to the input feature. Different approaches exist, which can roughly be assigned to three categories:

1. Perturbation-based methods evaluate the model output after applying perturbations on the input data and derive explanations from the changes occurring. These perturbations can be infinitesimal (e.g., gradients) or rather coarse [90], moreover, they can be expressed as optimization problem [91]. Although being straight-forward to apply, these methods have several drawbacks such as intense requirements in terms of computation (not the case for gradients), since the model output has to be evaluated a large number of times, and limited reliability, because results are highly sensitive to the applied perturbations (e.g., perturbed input may not lie on the input manifold or gradient shattering problem [92]).
2. *Surrogate-based* methods (e.g. LIME, [93]) query the model in question for a large number of inputs and approximate it by a simpler model which is intrinsically interpretable. Explanations for the original model's behavior can then be inferred. The problems that come with this approach are, on the one hand, the dependence of the explanations on the way the input queries are sampled and the simpler model is fitted and, on the other hand, the computational effort for querying the original model a large number of times.
3. *Structure-based* methods (e.g. LRP, [94]) use the internal structure of the network to propagate information on relevance between network layers from the output up to the input data. The main specific drawback of these methods is that they require access to the internal structure of the model and thus are not model-agnostic. However, they are much less computationally intense than others and the explanations they provide score better under a range of criteria (c.f. [95]).

In order to obtain a more complete picture of the prediction strategies implemented by the model, one can aggregate or cluster multiple local explanations [96]. Other methods act on the latent space instead of on the input features, and thereby provide explanations with higher-level concepts such as color, shape and object part [97].

Some of these explanation methods, e.g. LRP, have already been used to uncover unintended biases in large image data sets. For instance, they unmasked so-called Clever Hans classifiers [98], i.e., models that (seemingly) arrive at the correct decisions but for the wrong reason, that identified horses based on copyright tags or pneumonia x-rays based on the presence of a "portable" label. In a more general setting, this approach can be applied to detect biases in the data and improve the generalization ability of the model.

Recently, XAI methods have been applied to other model structures beyond DNNs, and also for purposes beyond visualisation (e.g. network pruning). Yet many challenges still exist to leverage the full potential of XAI in helping researchers to arrive at robust and trustworthy models. One factor limiting the benefits of XAI in many applications is an interpretation gap if input features themselves are not readily interpretable for humans. Another open issue of the explanation methods mentioned above is that they are not specifically designed to uncover possible interplays between multiple input areas, e.g. to answer which combinations of pixels in multiple areas in an image contribute to specific decisions. Finally, it is also unclear how to optimally and without human intervention integrate XAI into the model training (e.g. into the loss function) in order to improve the model.

2.7 Overview of AI standardization activities worldwide

Standards are a proven way to describe uniform technical requirements for AI systems and to support the implementation of legal frameworks. They also facilitate market access for AI innovations and give AI system marketers a clear framework for the development and operation of AI systems. In Germany, for example, DIN and DKE are the main standardization bodies and represent national interests at the EU level in standardization organizations such as CEN, CENELEC and ETSI, and at the international level in organizations such as ISO, IEC and ITU.

With regard to the topic of testing and auditing AI systems discussed in this whitepaper, the question arises which AI quality standards require independent testing and which standards need to be developed for such testing procedures themselves. To address this lack of standards, in Germany, for instance, a comprehensive analysis of the existing situation and the need for standards and norms in the field of artificial intelligence has been presented in the form of the "Normungsroadmap KI" [99]. The most important quality dimensions, which should be addressed by standardization, are shown in Fig. 3.

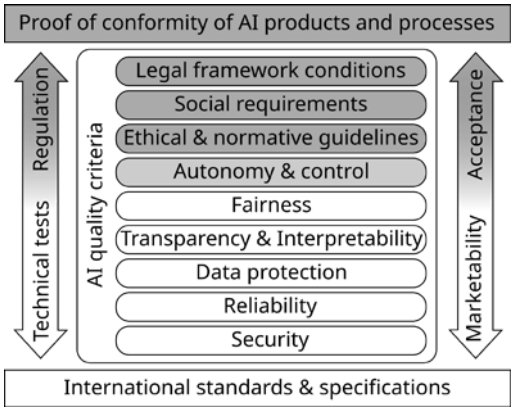

Fig. 3: Classification of the categories of the AI quality criteria into the compliance test according to [99] (cf. also [100; 101]).

The first standards that are currently emerging in the field of AI, especially for the topics of reliability, robustness, safety and security are listed in table 1 (cf. [99] for a complete overview).

Topic	Document
Reliability & Robustness	› ISO/IEC NP 24029: Assessment of robustness of neural networks [102]
	› ITU-T F.AI-DLFE & F.AI-DLPB: deep learning software framework evaluation methodology & metrics and evaluation methods for DNN processor benchmark [103]
	› ETSI DTR INT 008 (TR 103 821): AI in test systems and testing AI models, definitions of quality metrics [104]
	› DIN SPEC 92001-2: AI life cycle processes and quality requirements, Part2 robustness [105]
	› ITU Focus Group on "Artificial Intelligence for Health", e.g. [106] and [107]
Safety	› ISO/CD TR 22100-5: safety of machinery, Part 5: implications of embedded AI – ML [108]
	› ISO 26262: road vehicles – functional safety ([18], see also IEC 61508-1:2010 [17], ISO 21448 [109])
	› IEEE P2802: standard for performance and safety evaluation of AI medical devices – terminology [110]
	› ISO/IEC AWI TR 5469: AI – functional safety and AI systems [111]
Security	› ISO/SAE 21434: road vehicles – cybersecurity engineering ([112],s.a. ISO/CD 24089 [113], ISO/IEC 23894 [114])
	› ETSI ISG SAI: Several documents covering problem statement, threat ontology, security testing and mitigation strategies for general AI systems, e.g. [115]
	› NISTIR 8269: a taxonomy and terminology of adversarial ML [116]

Table 1: Emerging standards in the field of AI for selected topics covered by this whitepaper. For a more complete overview cf. [99] and [117].

It is apparent, however, that there is still a considerable need for development in the area of technical testing ("product testing"), particularly with regard to the validation and verification of neural networks, reliable safety arguments for safety-critical systems, and testing tools for carrying out these tests. Thus, extensive standardization activities will continue in the coming years. A prominent example of how this need is being addressed for the topic of autonomous driving is represented by the German project "KI-Absicherung" [118]. It is managed by a consortium comprising research institutions, automotive manufacturers, suppliers, standardization organizations and relevant public authorities (such as the German BSI), and is developing a general industry consensus with regard to verification strategies for the safety of AI-based modules of highly automated driving.

It is expected that further technical test procedures will emerge in the coming period via further comparable lighthouse projects and pilot tests, and that corresponding standardization needs can be addressed.

3 Open Issues and Promising Approaches

At least for safety- and security-relevant cAI applications, a sufficient level of robustness, security, safety and auditability needs to be achieved and corresponding technical guidelines and standards need to be developed. When reviewing the state of the art in this area (cf. previous sections of the whitepaper), it becomes apparent that, on the one hand, many open issues remain but that, on the other hand, many promising solutions and approaches exist to either resolve or reduce the impact of these issues. Hereafter, both open issues and promising approaches will be summarized along a modified depiction of the cAI life cycle (cf. Fig. 4):

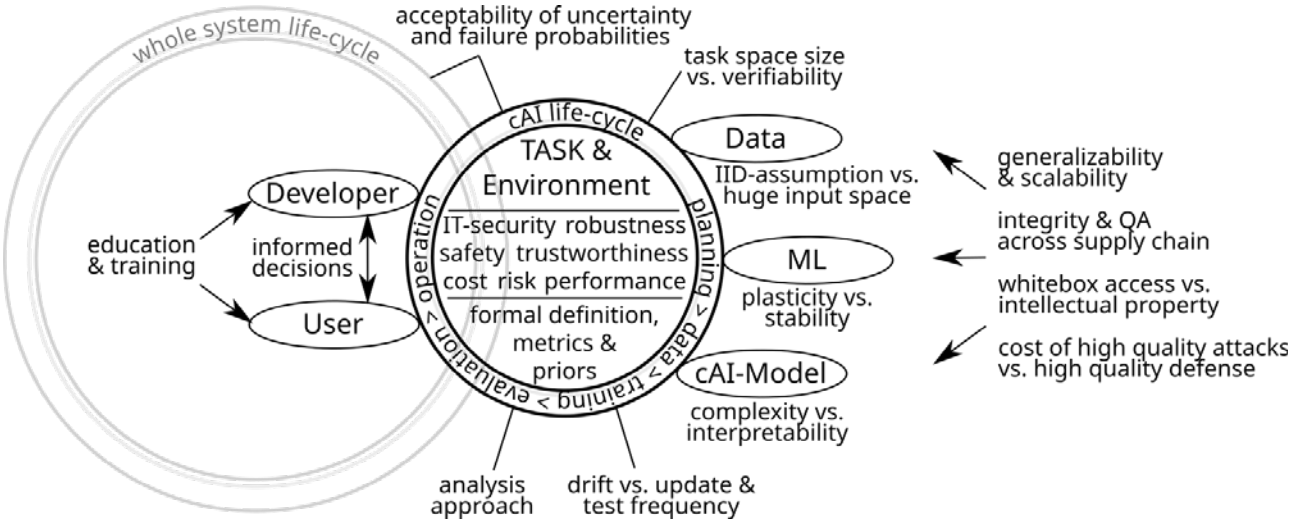

Fig. 4: cAI life cycle (cf. Fig. 2) depicted with a focus on open questions in the context of auditability, IT security and safety.

A cAI life cycle is usually embedded in a whole system life cycle that, depending on the specific use case, includes multiple cIT and sAI systems as well as e.g. hardware devices such as sensors and actuators. From this point of view, a machine that operates autonomously in a complex and changing environment (drift) is never fully and finally specified and, therefore, uncertainty and the risk of errors remain just like for humans. First approaches to deal with the risk assessment of an embedded cAI life cycle come from the area of functional safety ([119; 120],cf. sections 2.1 and 2.7). In order to define suitable approaches for analyszing, verifying and validating AI systems, it is first necessary to identify and understand their intended use, the task, and the environment they operate in. Each specific use case is characterized by a number of essential properties which are expected by a user or a regulatory authority to be implemented as essential characteristics of the system, e.g. robustness, security and safety. In most cases, formal definitions and relevant metrics of task and environment are missing or incomplete. This has several undesirable consequences, e.g. the acceptable risk must take the perception and opinions of affected users into account. Users and developers in turn need to have a solid basis in terms of education, training and communication with each other to take informed decisions regarding the use of specific AI models, ML algorithms, data sets and analysis approaches as well as regarding further boundary conditions for specific use cases.

An overarching open issue is that of (often multi-faceted) trade-offs between desired characteristics of the system, e.g. robustness, security, safety and auditability, on the one hand and characteristics of the AI model, ML algorithm, data and boundary conditions, such as model complexity, task space, plasticity, cost and performance, on the other hand. These trade-offs restrict the scalability and generalizability of current AI systems. To give examples: 1. increasing model complexity e.g. may negatively impact interpretability and defense; 2. increasing task space size leads to the need for larger training and test data sets, which will complicate verification and makes it harder to fulfill the IID requirement which is an important pre-requisite for training robust AI systems; 3. strengthening defenses often leads to reduced performance; 4. maintaining invariant characteristics of the AI system in the presence of drift requires frequent re-trainings and testing and, therefore, creates increased costs; 5. white-box model and life cycle access for improved auditability conflicts with intellectual property interests; 6. the use of external data sets and pre-trained models reduces costs but opens up new vulnerabilities, especially for hard to detect backdoor attacks.

Research has come up with a multitude of promising approaches to resolve the open issues on multiple levels, e.g. 1. re-training is made more efficient by using transfer and few-shot learning and, to account for the need for tuning meta-parameters, by using non-parametric and ensemble methods. Plasticity and stability may, hereby, be balanced well at least for models of low complexity; 2. optimizing defense methods with respect to a suitable metric that accounts for the performance in the face of natural and adversarial inputs helps to diminish the usual performance drop when employing strong defense methods; 3. shared and meta adversarial training reduce the cost for dealing with universal perturbations; 4. the systematic use of synthetic and/or augmented data and simulations allows identifying failure modes and robustifying AI systems despite large task spaces; 5. to some extent abstract interpretation and certifiable training permit the verification of AI systems with larger task spaces; 6. arguments-based approaches such as CAE and defeasible reasoning allow to audit AI-systems where existing methodologies cannot be applied; 7. exploiting human priors allows improving the interpretability of AI systems and, via hybrid models, making AI systems more robust; 8. defending against backdoor attacks by data sanitization via either the detection of outliers in data sets using interpretation methods, by reject on negative impact and related approaches during training (RONI, [121]), or by using bagging ensembles [122]; 9. if whitebox access is not possible, substitute models and substitute data sets might, at least in some cases, be used to improve audit quality, e.g. by generating high-quality attacks; 10. Cryptographic methods and chains of trust may be used to ensure the integrity of data and models in the supply chain. In addition, combinations of these approaches can be used.

Despite all of these and other promising approaches it has to be kept in mind, that in the future the complexity of tasks, models and data sets will most likely increase, necessitating even more powerful approaches.

4 Setting Priorities for Work Towards Auditable AI Systems

To date no generally applicable set of criteria and tools is available to secure AI systems such that a sufficiently low probability of errors can be demonstrated by rigorous means. Considering the results of this whitepaper, two general strategies exist to obtain auditable, secure and safe AI systems (cf. Fig. 5):

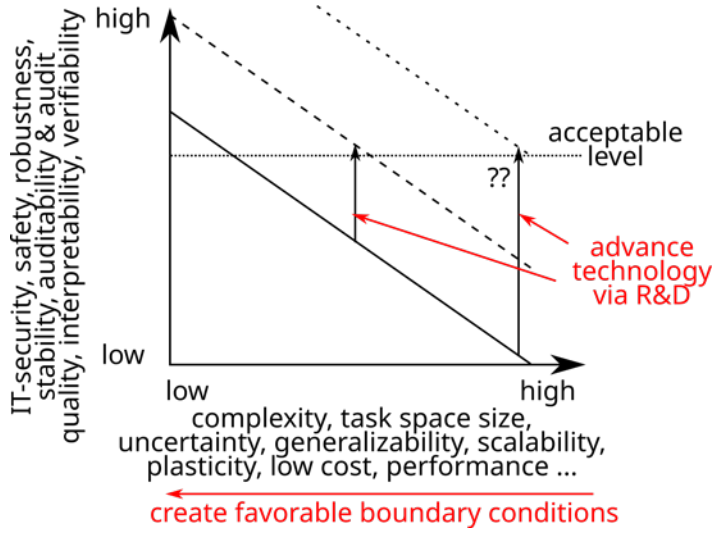

Fig. 5: Schematic plot of the multifaceted trade-offs that have to be considered when trying to achieve acceptable levels of IT security, safety, audit quality, robustness and verifiability. The achievable levels depend on multiple boundary conditions such as task complexity and model complexity. For a given boundary condition, advances in technology via R&D may allow e.g. to achieve higher IT security levels and/or improved auditability but so far this only works to a limited extent (cf. "??").

1. Create favorable boundary conditions for the given task: proper education of developers and users as well as sufficient information exchange between both parties allows clearly defining the task and acceptable boundary conditions. In combination with a subsequent risk analysis that takes into account the embedding of the AI system in a larger IT and/or robotic system, this forms the basis for informed choices during the development process and the deployment and operation of the AI system. In an extreme case, the developer or user might reach the conclusion that the use of AI technology has to be completely banned for the specific use case, e.g. due to security concerns. Otherwise, depending on the use case, constraining the task space and limiting the AI model complexity may allow for better auditability and a more secure and safe AI system [123][8]. Furthermore, the combination of multiple technical and organizational measures as well as, depending on intellectual property considerations, white-box access to cAI model and data throughout the life cycle for evaluation purposes will most probably improve the auditability and contribute to security and safety.
2. Invest in R&D to advance available technologies to eventually allow for secure and safe AI systems despite complex boundary conditions and, therefore, to improve scalability and generalizability. Examples include: a) the development of appropriate metrics across all security- and safety-relevant aspects of AI systems. They help to minimize the impact of trade-offs such as the one between performance and defense strength; b) the combination of robust models and detection algorithms to reject possibly malicious inputs while maintaining high

[8] Furthermore, one could think of modular AI-systems consisting of rather simple, interpretable and verifiable modules. Here, one would have to balance if the complexity of the monolithic AI system or the network complexity of the modular system is better suited for the task at hand.

performance; c) the inclusion of human priors via e.g. hybrid models to improve interpretability; d) the efficient generation of a high number of high-quality attacks as a basis for the development of efficient defense methods such as adversarial training; e) the generation of a large amount of high-quality realistic synthetic data to contribute to an IID data set as a basis for the training of robust AI systems; f) the combination of realistic simulations with real-world evaluations and g) the use of multiple redundant but qualitatively different systems, e.g. the combination of cAI, cIT and sAI systems via e.g. averaging, majority vote or winner takes it all.

Both strategies should be followed with high priority while, in a first step, focussing on selected security- and safety critical use cases. Available standards, guidelines and tools should be exploited (cf. the remainder of this whitepaper) and interdisciplinary exchange between researchers and industry should be further promoted [124] to find the best combinations of available criteria and tools for achieving auditable, secure, safe and robust AI systems for the specific use case. These criteria and tools have to be evaluated with respect to their practical benefit and feasibility within the respective use cases. Insights from these use cases should then be used, in a second step, to generalize the results and to build up a modular toolbox that may subsequently be applied to other use cases. On this basis, first technical guidelines and subsequently standards should be developed. In the ideal case the outcome will be a generally applicable set of criteria and tools that allows making AI systems sufficiently auditable, safe and secure.

Acknowledgements

We would like to thank Aleksander Mądry (MIT) for his stimulating presentation and the important comments and impulses throughout the workshop. We would further like to thank all workshop participants who contributed to the discussions before, during and after the workshop. Also we would like to thank Maria Sürig from VdTÜV and Jennifer Chyla from Fraunhofer HHI CINQ Center for their important contributions to the organization of the workshop.

5 References

[1] Kim, J., Shin, N., Jo, S. Y. and Kim, S. H.: Method of intrusion detection using deep neural network, *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*:313-316, 2017.
[2] Polyakov, E. V., Mazhanov, M. S., Rolich, A. Y., Voskov, L. S., Kachalova, M. V. and Polyakov, S. V.: Investigation and development of the intelligent voice assistant for the Internet of Things using machine learning, *2018 Moscow Workshop on Electronic and Networking Technologies (MWENT)*:1-5, 2018.
[3] Topol, E. J.: High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 25:44-56, 2019.
[4] Ma, Y., Wang, Z., Yang, H. and Yang, L.: Artificial intelligence applications in the development of autonomous vehicles: a survey. *IEEE/CAA Journal of Automatica Sinica* 7:315-329, 2020.
[5] Wirtz, B. W., Weyerer, J. C. and Geyer, C.: Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration* 42:596-615, 2019.
[6] Yu, K.-H. and Kohane, I. S.: Framing the challenges of artificial intelligence in medicine. *BMJ Quality & Safety* 28:238-241, 2019.
[7] Berghoff, C., Neu, M. and von Twickel, A.: Vulnerabilities of Connectionist AI Applications: Evaluation and Defense. *Frontiers in Big Data* 3:23, 2020.
[8] Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M. and Yi, X.: A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability. *Computer Science Review 37:100270*, 2020.
[9] Dede, G., Naydenov, R., Malatras, A., Hamon, R., Junklewitz, H. and Sanchez, I.: *Cybersecurity Challenges In The Uptake of Artificial Intelligence in Autonomous Driving*. 2021.
[10] FRA – European Union Agency for Fundamental Rights: *Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights*. 2019.
[11] European Commission: *White Paper On Artificial Intelligence - a European Approach To Excellence And Trust*. 2020.
[12] Eicher, M., Scharpfenecker, P., Ludwig, D., Friedmann, F., Netter, F. and Reuther, M.: *Process considerations: a reliable AI data labeling process*. 2020.
[13] Xiao, Q., Chen, Y., Shen, C., Chen, Y. and Li, K.: Seeing is Not Believing: Camouflage Attacks on Image Scaling Algorithms, *28th USENIX Security Symposium (USENIX Security 19)*:443-460, 2019.
[14] Quiring, E., Klein, D., Arp, D., Johns, M. and Rieck, K.: Adversarial Preprocessing: Understanding and Preventing Image-Scaling Attacks in Machine Learning, *29th USENIX Security Symposium (USENIX Security 20)*:1363-1380, 2020.
[15] Jakubovitz, D. and Giryes, R.: Improving DNN Robustness to Adversarial Attacks Using Jacobian Regularization, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII* 11216:525-541, 2018.
[16] Hutter, F., Kotthoff, L. and Vanschoren, J. (Eds.): *Automated Machine Learning - Methods, Systems, Challenges*. Springer International Publishing, 2019.
[17] *IEC 61508-1:2010 Functional safety of electrical/electronic/programmable electronic safety-related systems - Part 1: General requirements*. 2010.
[18] ISO/TC 22/SC 32: *ISO 26262-1:2018 Road vehicles — Functional safety*. 2018.

[19] Bielik, P., Tsankov, P., Krause, A. and Vechev, M.: *Reliability Assessment of Traffic Sign Classifiers*. 2020. (Download via https://www.bsi.bund.de/KI)
[20] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. and Madry, A.: Robustness May Be at Odds with Accuracy, *ICLR2019*, 2019.
[21] Shalev-Shwartz, S. and Ben-David, S.: *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.
[22] Brinkrolf, J. and Hammer, B.: Interpretable machine learning with reject option. *at - Automatisierungstechnik* 66:283-290, 2018.
[23] Shah, K. and Manwani, N.: Online Active Learning of Reject Option Classifiers. arXiv cs.LG 1906.06166, 2019.
[24] Weiss, K., Khoshgoftaar, T. M. and Wang, D.: A survey of transfer learning. *J Big Data* 3, 2016.
[25] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. and Liu, C.: A Survey on Deep Transfer Learning. arXiv cs.LG 1808.01974, 2018.
[26] Prahm, C., Schulz, A., Paaßen, B., Schoisswohl, J., Kaniusas, E., Dorffner, G., Hammer, B. and Aszmann, O.: Counteracting Electrode Shifts in Upper-Limb Prosthesis Control via Transfer Learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27:956-962, 2019.
[27] Montiel, J., Read, J., Bifet, A. and Abdessalem, T.: Scikit-Multiflow: A Multi-output Streaming Framework. *Journal of Machine Learning Research 1:2915-2914, 2019*.
[28] Losing, V., Hammer, B. and Wersing, H.: Incremental on-line learning: A review and comparison of state of the art algorithms. *Neurocomputing* 275:1261-1274, 2018.
[29] Losing, V., Hammer, B. and Wersing, H.: Tackling heterogeneous concept drift with the Self-Adjusting Memory (SAM). *Knowl Inf Syst* 54:171–201, 2018.
[30] Gomes, H. M., Read, J., Bifet, A., Barddal, J. P. and Gama, J. a.: Machine Learning for Streaming Data: State of the Art, Challenges, and Opportunities. *SIGKDD Explor. Newsl.* 21:6–22, 2019.
[31] Webb, G. I., Lee, L. K., Petitjean, F. and Goethals, B.: Understanding Concept Drift. arXiv cs.LG 1704.00362, 2017.
[32] Hinder, F. and Hammer, B.: Counterfactual Explanations of Concept Drift. arXiv cs.LG 2006.12822, 2020.
[33] Biggio, B. and Roli, F.: Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. arXiv cs.CV 1712.03141, 2017.
[34] Tramèr, F., Zhang, F., Juels, A., Reiter, M. K. and Ristenpart, T.: Stealing Machine Learning Models via Prediction APIs, *25th USENIX Security Symposium, USENIX Security 16, Austin, TX, USA, August 10-12, 2016*:601-618, 2016.
[35] Fredrikson, M., Jha, S. and Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, USA, October 12-16, 2015*:1322-1333, 2015.
[36] Shokri, R., Stronati, M., Song, C. and Shmatikov, V.: Membership Inference Attacks Against Machine Learning Models, *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*:3-18, 2017.
[37] Biggio, B., Nelson, B. and Laskov, P.: Poisoning attacks against support vector machines, *Int'l Conf. on Machine Learning (ICML) 2012*, 2012.
[38] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.:

Intriguing properties of neural networks, *2nd International Conference on Learning Representations, ICLR 2014*, 2014.
[39] Goodfellow, I., Shlens, J. and Szegedy, C.: Explaining and Harnessing Adversarial Examples, *International Conference on Learning Representations*, 2015.
[40] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T. and Song, D.: Robust Physical-World Attacks on Deep Learning Visual Classification, *Computer Vision and Pattern Recognition (CVPR)*, 2018.
[41] Melis, M., Demontis, A., Biggio, B., Brown, G., Fumera, G. and Roli, F.: Is Deep Learning Safe for Robot Vision? Adversarial Examples against the iCub Humanoid. arXiv cs.LG 1708.06939, 2017.
[42] Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C. and Roli, F.: Towards Poisoning of Deep Learning Algorithms with Back-gradient Optimization. , arXiv cs.LG 1708.08689, 2017.
[43] Demontis, A., Melis, M., Biggio, B., Maiorca, D., Arp, D., Rieck, K., Corona, I., Giacinto, G. and Roli, F.: Yes, machine learning can be more secure! A case study on android malware detection. *IEEE Transactions on Dependable and Secure Computing* 16:711-724, 2017.
[44] Sharif, M., Bhagavatula, S., Bauer, L. and Reiter, M. K.: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016*:1528-1540, 2016.
[45] Komkov, S. and Petiushko, A.: AdvHat: Real-world adversarial attack on ArcFace Face ID system. arXiv cs.CV 1908.08705, 2019.
[46] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Laskov, P., Giacinto, G. and Roli, F.: Evasion attacks against machine learning at test time, *in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2013.
[47] Schölkopf, B.: *People of ACM - Bernhard Schölkopf*. 2018. https://www.acm.org/articles/people-of-acm/2018/bernhard-scholkopf, last visited 2021/02/19
[48] Gilmer, J., Ford, N., Carlini, N. and Cubuk, E. D.: Adversarial Examples Are a Natural Consequence of Test Error in Noise, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA* 97:2280-2289, 2019.
[49] Gu, T., Dolan-Gavitt, B. and Garg, S.: BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *CoRR* abs/1708.06733, arXiv 1708.06733, 2017.
[50] Tran, B., Li, J. and Madry, A.: Spectral Signatures in Backdoor Attacks. arXiv cs.LG 1811.00636, 2018.
[51] Yang, C., Wu, Q., Li, H. and Chen, Y.: Generative Poisoning Attack Method Against Neural Networks. *CoRR* abs/1703.01340, arXiv 1703.01340, 2017.
[52] Turner, A., Tsipras, D. and Madry, A.: Label-Consistent Backdoor Attacks. arXiv stat.ML 1912.02771, 2019.
[53] Du, M., Jia, R. and Song, D.: Robust Anomaly Detection and Backdoor Attack Detection Via Differential Privacy. arXiv cs.LG 1911.07116, 2019.
[54] Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I. and Srivastava, B.: Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19), Honolulu, Hawaii, January 27, 2019* 2301, 2019.

[55] Yang, P., Chen, J., Hsieh, C., Wang, J. and Jordan, M. I.: ML-LOO: Detecting Adversarial Examples with Feature Attribution, *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*:6639-6647, 2020.
[56] Srivastava, M., Hashimoto, T. B. and Liang, P.: Robustness to Spurious Correlations via Human Annotations, *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event* 119:9109-9119, 2020.
[57] Raghu, M. and Schmidt, E.: A Survey of Deep Learning for Scientific Discovery. *CoRR* abs/2003.11755, arXiv 2003.11755, 2020.
[58] Berghoff, C.: Protecting the integrity of the training procedure of neural networks. arXiv cs.CR 2005.06928, 2020.
[59] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A.: Towards Deep Learning Models Resistant to Adversarial Attacks, *ICLR2017*, 2017.
[60] Metzen, J. H., Genewein, T., Fischer, V. and Bischoff, B.: On Detecting Adversarial Perturbations, *Proceedings of 5th International Conference on Learning Representations (ICLR)*, 2017.
[61] He, W., Wei, J., Chen, X., Carlini, N. and Song, D.: Adversarial Example Defense: Ensembles of Weak Defenses are not Strong, *11th USENIX Workshop on Offensive Technologies, WOOT 2017, Vancouver, BC, Canada*, 2017.
[62] Carlini, N. and Wagner, D.: Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods, *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*:3–14, 2017.
[63] Tramèr, F., Carlini, N., Brendel, W. and Madry, A.: On Adaptive Attacks to Adversarial Example Defenses, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, virtual*, 2020.
[64] Sheikholeslami, F., Lotfi, A. and Kolter, J. Z.: Provably robust classification of adversarial examples with detection, *ICLR 2021*, 2021.
[65] Mummadi, C. K., Brox, T. and Metzen, J. H.: Defending Against Universal Perturbations With Shared Adversarial Training, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*:4927-4936, 2019.
[66] Metzen, J. H., Finnie, N. and Hutmacher, R.: Meta Adversarial Training. arXiv cs.LG 2101.11453, 2021.
[67] Bai, Y., Zeng, Y., Jiang, Y., Xia, S.-T., Ma, X. and Wang, Y.: Improving Adversarial Robustness via Channel-wise Activation Suppressing, *International Conference on Learning Representations*, 2021.
[68] Brown, T. B., Mané, D., Roy, A., Abadi, M. and Gilmer, J.: Adversarial Patch. *CoRR* abs/1712.09665, arXiv 1712.09665, 2017.
[69] Levine, A. and Feizi, S.: (De)Randomized Smoothing for Certifiable Defense against Patch Attacks. arXiv cs.LG 2002.10733, 2020.
[70] Metzen, J. H. and Yatsura, M.: Efficient Certified Defenses Against Patch Attacks on Image Classifiers, *ICLR 2021*, 2021.
[71] Athalye, A., Carlini, N. and Wagner, D. A.: Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples, *Proceedings of the 35th International Conference*

on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, 80:274-283, 2018.
[72] Huang, X., Kwiatkowska, M., Wang, S. and Wu, M.: Safety Verification of Deep Neural Networks. arXiv cs.AI 1610.06940, 2016.
[73] Katz, G., Barrett, C. W., Dill, D. L., Julian, K. and Kochenderfer, M. J.: Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks, *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I* 10426:97-117, 2017.
[74] Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S. and Vechev, M.: AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation, *2018 IEEE Symposium on Security and Privacy (SP)*:3-18, 2018.
[75] Mirman, M., Singh, G. and Vechev, M. T.: A Provable Defense for Deep Residual Networks. *CoRR* abs/1903.12519, arXiv 1903.12519, 2019.
[76] Balunovic, M. and Vechev, M.: Adversarial Training and Provable Defenses: Bridging the Gap, *International Conference on Learning Representations (ICLR 2020)*, 2020.
[77] Wicker, M., Laurenti, L., Patane, A., Chen, Z., Zhang, Z. and Kwiatkowska, M.: *Bayesian Inference with Certifiable Adversarial Robustness*. 2021.
[78] Bloomfield, R. and Rushby, J.: Assurance 2.0: A Manifesto. arXiv cs.SE 2004.10474, 2020.
[79] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G. K., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensbold, J., O'Keefe, C., Koren, M., Ryffel, T., Rubinovitz, J. B., Besiroglu, T., Carugati, F., Clark, J., Eckersley, P., de Haas, S., Johnson, M., Laurie, B., Ingerman, A., Krawczuk, I., Askell, A., Cammarota, R., Lohn, A., Krueger, D., Stix, C., Henderson, P., Graham, L., Prunkl, C., Martin, B., Seger, E., Zilberman, N., hÉigeartaigh, S. Ó., Kroeger, F., Sastry, G., Kagan, R., Weller, A., Tse, B., Barnes, E., Dafoe, A., Scharre, P., Herbert-Voss, A., Rasser, M., Sodhani, S., Flynn, C., Gilbert, T. K., Dyer, L., Khan, S., Bengio, Y. and Anderljung, M.: Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *CoRR* abs/2004.07213, arXiv 2004.07213, 2020.
[80] Goodenough, J., Weinstock, C. B., Goodenough, J. B., Weinstock, C. B. and Klein, A. Z.: *Eliminative Argumentation: A Basis for Arguing Confidence in System Properties*. Report number: CMU/SEI-2015-TR-005, Software Engineering Institute, Carnegie Mellon University, 2015.
[81] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R. and Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.* 63:4:1-4:15, 2019.
[82] Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK: *Auditing machine learning algorithms - A white paper for public auditors*. 2020. https://auditingalgorithms.net/
[83] Oala, L., Fehr, J., Gilli, L., Balachandran, P., Leite, A. W., Calderon-Ramirez, S., Li, D. X., Nobis, G., Alvarado, E. A. M., Jaramillo-Gutierrez, G., Matek, C., Shroff, A., Kherif, F., Sanguinetti, B. and Wiegand, T.: ML4H Auditing: From Paper to Practice, *Proceedings of the Machine Learning for Health NeurIPS Workshop* 136:280-317, 2020.
[84] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J.,

Theron, D. and Barnes, P.: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*:33-44, 2020.
[85] Shwartz-Ziv, R. and Tishby, N.: Opening the Black Box of Deep Neural Networks via Information. arXiv cs.LG 1703.00810, 2017.
[86] Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. and Müller, K.-R. (Ed.).: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing, 2019.
[87] Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J. and Müller, K.: Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* 109:247-278, 2021.
[88] Nguyen, A. M., Dosovitskiy, A., Yosinski, J., Brox, T. and Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*:3387-3395, 2016.
[89] Zhou, B., Bau, D., Oliva, A. and Torralba, A.: Interpreting Deep Visual Representations via Network Dissection. *IEEE Trans. Pattern Anal. Mach. Intell.* 41:2131-2145, 2019.
[90] Zintgraf, L. M., Cohen, T. S., Adel, T. and Welling, M.: Visualizing Deep Neural Network Decisions: Prediction Difference Analysis, *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings,* 2017.
[91] Fong, R. C. and Vedaldi, A.: Interpretable Explanations of Black Boxes by Meaningful Perturbation, *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*:3449-3457, 2017.
[92] Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W. and McWilliams, B.: The Shattered Gradients Problem: If resnets are the answer, then what is the question?, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* 70:342-350, 2017.
[93] Ribeiro, M. T., Singh, S. and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*:97-101, 2016.
[94] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. and Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PloS one* 10:e0130140, 2015.
[95] Arras, L., Osman, A. and Samek, W.: Ground Truth Evaluation of Neural Network Explanations with CLEVR-XAI. *CoRR* abs/2003.07258, arXiv cs.CV 2003.07258, 2021.
[96] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.-R.: Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications* 10:1096, 2019.
[97] Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B. and Sayres, R.: Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018* 80:2673-2682, 2018.
[98] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W. and Müller, K.: Unmasking

Clever Hans Predictors and Assessing What Machines Really Learn. *CoRR* abs/1902.10178, arXiv 1902.10178, 2019.
[99] Wahlster, W. and Winterhalter, C.: *German Standardization Roadmap on Artificial Intelligence*. DIN DKE, 2020.
[100] Cremers, A. B., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., Rosenzweig, J., Rostalski, F., Sicking, J., Volmer, J., Voosholz, J., Voss, A. and Wrobel, S.: *Vertrauenswürdiger Einsatz von Künstlicher Intelligenz - Handlungsfelder aus philosophischer, ethischer, rechtlicher und technologischer Sicht als Grundlage für eine Zertifizierung von Künstlicher Intelligenz*. Fraunhofer IAIS, Sankt Augustin, Germany, 2019.
[101] High-level Expert Group on Artificial Intelligence. *The Assessment List For Trustworthy Artificial Intelligence (ALTAI) - For Self Assessment*. 2020.
[102] *ISO/IEC AWI 24029-2 Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods*. 2021.
[103] International Telecommunication Union: *F.AI-DLFE "Deep Learning Software Framework Evaluation Methodology" (Rev.) Output draft, Virtual meeting, 22 June - 3 July 2020*, 2020.
[104] ETSI. *TR 103 821: Autonomic network engineering for the self-managing Future Internet (AFI); Artificial Intelligence (AI) in Test Systems and Testing AI models.* 2019.
[105] DIN. *DIN SPEC 92001-2:2020-12 Künstliche Intelligenz - Life Cycle Prozesse und Qualitätsanforderungen - Teil 2: Robustheit*. 2020.
[106] Oala, L., Balachandran, P., Cabitza, F., Ramirez, S. C., Filho, A. C., Eitel, F., Extermann, J., Fehr, J., Ghozzi, S., Gilli, L., Jaramillo-Gutierrez, G., Kester, Q.-A., Kurapati, S., Konigorski, S., Krois, J., Lippert, C., Martin, J., Merola, A., Murchison, A., Niehaus, S., Ritter, K., Samek, W., Sanguinetti, B., Schwerk, A. and Srinivasan, V.: Data and artificial intelligence assessment methods (DAISAM), *ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) - Meeting I 2020*, 2020.
[107] Schörverth, E., Vogler, S., Balachandran, P., Leite, A. W., Li, D. X., Ali, K., Garcia, Schneider, D., Krois, J., Lecoultre, M., Iyer, S., Choudhary, S. and Oala, L.: FG-AI4H Open Code Initiative - Evaluation and Reporting Package, *ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) - Meeting K 2021*, 2021.
[108] *ISO/TR 22100-5:2021 Safety of machinery — Relationship with ISO 12100 — Part 5: Implications of artificial intelligence machine learning*. 2021.
[109] ISO. *ISO 21448 Road vehicles - Safety Of The Intended Funktionality*. 2019.
[110] IEEE. *P2802: Standard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: TerminologyStandard for the Performance and Safety Evaluation of Artificial Intelligence Based Medical Device: Terminology*. 2018.
[111] *ISO/IEC AWI TR 5469 Artificial intelligence — Functional safety and AI systems*. 2021.
[112] *ISO/SAE 21434: road vehicles – cybersecurity engineering. Under development as of 04/2021.*
[113] *ISO/CD 24089 Road vehicles - Software update engineering. Under development as of 04/2021.*
[114] *ISO/IEC 23894 Information Technology - Artificial Intelligence - Risk Management. Under development as of 04/2021.*
[115] *ETSI GR SAI 005 V1.1.1 (2021-03)*. 2021.
[116] *NISTIR 8269 (Draft): A Taxonomy and Terminology of Adversarial Machine Learning*. 2019.
[117] ISO/IEC JTC: *Standards by ISO/IEC JTC 1/SC 42 - Artificial intelligence*.

https://www.iso.org/committee/6794475/x/catalogue/p/0/u/1/w/0/d/0 (last visited 04/2021)
[118] Konsortium of the project KI Absicherung: *KI Absicherung - Safe AI for Automated Driving*. https://www.ki-absicherung-projekt.de/ (last visited 04/2021)
[119] Leong, C., Kelly, T. and Alexander, R.: Incorporating Epistemic Uncertainty into the Safety Assurance of Socio-Technical Systems. *EPTCS 259, 2017, pp. 56-71*, arXiv cs.SE 1710.03394, 2017.
[120] Braband, J. and Schäbe, H.: On Safety Assessment of Artificial Intelligence. arXiv cs.AI 2003.00260, 2020.
[121] Barreno, M., Nelson, B., Joseph, A. D. and Tygar, J. D.: The security of machine learning. *Mach. Learn.* 81:121-148, 2010.
[122] Biggio, B., Corona, I., Fumera, G., Giacinto, G. and Roli, F.: Bagging Classifiers for Fighting Poisoning Attacks in Adversarial Classification Tasks, *Multiple Classifier Systems - 10th International Workshop, MCS 2011, Naples, Italy, June 15-17, 2011. Proceedings* 6713:350-359, 2011.
[123] Lechner, M., Hasani, R., Amini, A., Henzinger, T., Rus, D. and Grosu, R.: Neural circuit policies enabling auditable autonomy. *Nature Machine Intelligence* 2:642-652, 2020.
[124] Kusters, R., Misevic, D., Berry, H., Cully, A., Le Cunff, Y., Dandoy, L., Díaz-Rodríguez, N., Ficher, M., Grizou, J., Othmani, A., Palpanas, T., Komorowski, M., Loiseau, P., Moulin Frier, C., Nanini, S., Quercia, D., Sebag, M., Soulié Fogelman, F., Taleb, S., Tupikina, L., Sahu, V., Vie, J.-J. and Wehbi, F.: Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities. *Frontiers in Big Data* 3:45, 2020.

Herausgeber

Bundesamt für Sicherheit in der Informationstechnik
Godesberger Allee 185-189
53175 Bonn
Deutschland

Fraunhofer-Institut für Nachrichtentechnik
Heinrich-Hertz-Institut
Einsteinufer 37
10587 Berlin
Deutschland

Verband der TÜV e. V.
Friedrichstraße 136
10117 Berlin
Deutschland