D. Demmler, D. Krupka, H. Federrath. (Hrsg.): INFORMATIK 2022,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2022 1247

GAFAI: Proposal of a Generalized Audit Framework for AI

Thora Markert [1], Fabian Langer [2], Vasilios Danos [3]

Abstract: Machine Learning (ML) based AI applications are increasingly used in various fields and domains. Despite the enormous and promising capabilities of ML, the inherent lack of robustness, explainability and transparency limits the potential use cases of AI systems. In particular, within every safety or security critical area, such limitations require risk considerations and audits to be compliant with the prevailing safety and security demands. Unfortunately, existing standards and audit schemes do not completely cover the ML specific issues and lead to challenging or incomplete mapping of the ML functionality to the existing methodologies. Thus, we propose a generalized audit framework for ML based AI applications (GAFAI) as an anticipation and assistance to achieve auditability. This conceptual risk and requirement driven approach based on sets of generalized requirements and their corresponding application specific refinements as contributes to close the gaps in auditing AI.

Keywords: AI Auditing, AI Certification, Trustworthy AI, Security, Safety, Robustness, Interpretability;

1 Introduction

While Artificial Intelligence (AI) systems already enhance daily life in various ways, e. g. by virtual assistants [KB18; TD19], in navigation [GM21] or marketing [MS20], there are areas of application, where AI can lead to severe risks. Such areas, e. g. the medical or automotive sector, demand high levels of safety and security. In contrast to traditional software, the data driven approach of AI lead to unsatisfactory trustworthiness and pending standards and norms. The lack of i. e. uncertain robustness and explainability, impede safety and security critical applications. Even though standards and norms are still in the development stage, additional audit frameworks can already be a crucial contribution to strengthen the trust in AI systems [To20].

Standards and norms form the groundwork for (AI) audits. They define requirements for the systems, which are assessed during the audit process to verify the conformity. The German Standardization Roadmap on AI published a survey [De20] investigating the field of AI standardization and specification. Its goal is to shape the industrial sector to enhance usage and development of AI. The publication introduces five recommendations for action, i.a. proposing a certification program to counteract the identified shortcoming of standardized testing procedures and reproducible assessments of AI systems. In April 2021 the European Union proposed their legal framework on AI regulation (EU AI Act) [AIA21].

[1] TÜV Informationstechnik GmbH, Hardware Evaluation, Am TÜV 1, 45307 Essen, Germany t.markert@tuvit.de
[2] TÜV Informationstechnik GmbH, Hardware Evaluation, Am TÜV 1, 45307 Essen, Germany f.langer@tuvit.de
[3] TÜV Informationstechnik GmbH, Hardware Evaluation, Am TÜV 1, 45307 Essen, Germany v.danos@tuvit.de

doi:10.18420/inf2022_107

This risk-based approach formulates specifications and liabilities for a safe, secure and fair development and usage of AI systems The stated requirements are high-level and shall cover a broad range of different domains, especially targeting high-risk applications. It might be the foundation for following legislation, standards and norms.

A specific guidance for system assessment is proposed in the Ethics Guidelines for Trustworthy AI published in 2019 [Hi19] by the High-Level Expert Group on AI. This part of the EU Commission's Digital Strategy deals with the fundamentals and obligations of ethical and robust AI. For the realization of corresponding systems, seven high-level key requirements are proposed. The document also touches the imposing of the requirements with technical and non-technical methods. Based on the key requirements, an assessment list in form of a questionnaire gives guidance on how to achieve trustworthy AI. In 2021 the Fraunhofer IAIS proposed an AI audit catalogue [Po21], which provides guidance for developers to create trustworthy AI systems and assistance for the assessment of trustworthiness. The introduced four-step procedure is risk-based, demands a technical oriented approach with measurable objectives as indication for a system's trustworthiness and considers the entire AI lifecycle. The Institute of Public Auditors in Germany (IDW) draft for an assessment standard for AI systems [In22] is based on the International Standard on Assurance Engagements (ISAE) 3000. It specifies high-level requirements regarding ethics, traceability, security and performance for voluntary audits of AI systems. Requirement assessing audit procedures are described without touching on specified technical methods. Akula et al. propose AI Algorithm Audit [AG21], a general audit structure consisting of seven phases representing different degrees of freedom for the auditor. The phase is chosen based on the specific use case and its risk potential. They range from no access to the model and solely usage of checklists to full white-box access to the AI system. Specifically for the security of AI systems used in cloud environments, the Federal Office for Information Security (BSI) formulated the AIC4 [Bu21]. It states minimal requirements for the secure implementation of cloud-based machine learning and gives perspective on how such systems can be audited. The evaluation of AI-based clinical decision support systems is discussed in [Ma19]. Besides a depiction of the current state-of-the-art of AI evaluation in healthcare, the key challenges of evaluation and their solutions are outlined. In 2021, the BSI published a whitepaper about audits of safety critical AI systems [Be21]. The document focuses on the assessment of security and robustness of the audited system, again with special regard to the AI life cycle.

In addition to the audit structure, the method of each audit step is important. Accountability audits for AI systems based on knowledge graphs as proposed in [Na21] can provide clarity to questions related to liability. Ethical concerns can be targeted using process structures such as ethical-based auditing [Mö21]. A metric-oriented solutions such as the toolkit proposed in [Sa18] is designed to help determine bias in datasets and system decisions. An empirical evaluation regarding security concerns, e.g. based on robustness metrics, as introduced in [BFR17] can help assessing the status quo and allow cross-comparison to other systems as well. Last but not least, a structured safety case based on stringent argumentation as proposed in [Mo21] is necessary for a correct, complete and successful audit process.

One of the key contributions of this paper is the introduction of a structured and generalized framework for the challenging task of AI audits. It is designed to provide guidance on all levels of an audit from the definition of general high-level safety and security requirements to the evaluation of metric-based test results for robustness, interpretability and data privacy. Specifically, it enables the creation of a set of general requirements that are transferable between different use cases, this not only drastically decreases the effort of subsequent audits, but also provides the ability of comparison between audits, which can be useful for re-audits of the same system after a retraining the Machine Learning (ML) model.

2 Auditable Characteristics of AI Systems

Due to the special characteristics of AI systems, robustness and interpretability play a crucial role for safety and security, but also other topics such as ethics, fairness and privacy are affected. Hence, these are subject matter to the audit process in order to establish trust and enable the application in critical domains. Even though these aspects are equally as important to the audit process, in the scope of this paper, we focus on auditing the robustness, interpretability and privacy of AI systems and leave the remaining aspects to be discussed in future research.
Robustness has been proven to be a challenge for ML models. They tend to be susceptible to slight changes to their input such as natural perturbations, out-of-distribution data and adversarial attacks [GSS14]. Natural perturbations are changes to the input that occur naturally within the operational environment of the system, e. g. different lighting or weather conditions. Whereas adversarial attacks [GSS14] are carefully crafted perturbations designed to trigger an incorrect behaviour of the ML model. Due to the non-linearity and complexity of ML models, the formal verification of robustness is not yet scalable to commonly used large AI models and datasets [LXL20]. Increasing the complexity of auditing AI systems.
ML models are considered black boxes, as there is only limited insight into their decision making process. However, recently a lot of research emerged, to enhance the insight into this process with methods such as LRP [La15], LIME [RSG16] or Grad-CAM [Se16]. These methods determine specific features within the input data of a model with the highest influence on the model's decision.

3 Audit Evidences and Documentation

In this chapter, the necessary evidences and documentations needed for AI system audits are discussed. These components are the foundation for the verdict of the audit. Due to the special characteristics of AI systems, these evidences and documentations differ in comparison to traditional non-AI based software systems.

3.1 Entire System

ML models are often embedded within traditional software systems that are operated within specific environments. The system environment influences the robustness of the ML model as it may carry constraints (e. g. access limitations to the system) or specific environmental conditions that have to be reflected in the training data (e. g. specific weather conditions). This may have an impact functionality and robustness of the ML model. Therefore, a documentation of these interactions and the system environment has to be provided during an audit.

3.2 AI Life Cycle

The AI life cycle consists of a design phase, a development phase and the final deployment of the system. During design phase, the foundation for the ML model such as data sources, model architecture are decided and documented. In some cases a risk assessment or applicable norms and standards, are used to derive relevant security and safety requirements for the entire system. In combination with applicable norms and standards, this provides additional guidance for an audit in regard to necessary tests and possible boundary values and thresholds.
During the development phase the ML model is trained and functional tests are performed. Depending on the architecture of the entire system additional testing (i. e. integration tests) of the ML model integrated into different software of hardware levels of the entire system are executed. Documentation and tracking of the training process and performed test, enables an auditor to gain insight into whether relevant safety and security requirements were adhered to, suitable mitigation strategies were implemented and sufficient tests were performed.
Depending on norms and standards of the operational domain of the system, monitoring the system and ML model is necessary. Monitoring the system is beneficial to identify/log, mitigate and retrace malfunctions and attacks. Since this has an impact on the robustness and reliability of the system, the implementation and documentation of these mechanisms shall be evaluated during the audit.

3.3 Data

In supervised ML a model bases its functionality upon the data it is trained on. Therefore, it has to be ensured that the quality of the data is high, meaning it does not contain any errors and is representative of the operational domain. If the data quality is not sufficient, the model is less robust, i. e. it is more vulnerable to out-of-distribution data, natural perturbations and adversarial attacks. Due to this, a consistent documentation of the data and of any performed pre-processing, should be provided to the auditors.

4 Generalized Audit Framework for AI Systems

In this section, we propose a Generalized Audit Framework for AI (GAFAI), an ML evaluation approach which can be tailored to arbitrary use cases. Due to the variety of use cases and architectures, it is often not practical to define a universal audit process. GAFAI aims to fill this gap by providing a framework for an generalized audit process and guidance to tailor the requirements for specific use-cases. The generalized approach also enables the auditor to re-use sets of requirements for a class of similar behaving use-cases and applications. Figure 1 presents a schematic overview on the process. For each step, several activities have to be performed which are described in the following.

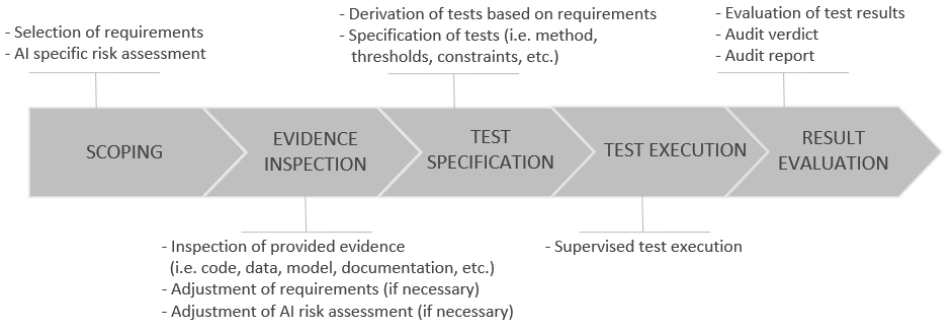

Fig. 1: Schematic overview about the main stages of the generalized audit process.

4.1 Scoping

The audit process starts with a scoping phase performed as a workshops between auditors and the organization developing the audited system. Information about the system and the intended application environment is provided to the auditors in order to derive the assets of the application and any safety and security related threats and hazards. An asset can be any claim or functionality of the system (e.g. correct detection of an object, sensitive data etc.). Once the assets are defined, an examination is performed to reveal potential threats and hazards might occur during the life cycle. Each threat and hazard is assigned to the asset it could affect. Finally, based on the provided information and the derived AI specific risk level and tolerable residual risk, a set of requirements is defined that is sufficient to cover the prior examined threats and hazards. Each threat and hazard shall be covered by at least one requirement.

The requirements are formulated as high-level technical requirements (general requirement) that leave room for more in depth refinement during the test specification phase (specific requirement). The advantage of the approach is, to keep a certain level of flexibility and reuse of requirements for an class of application providing a similar functionality. The

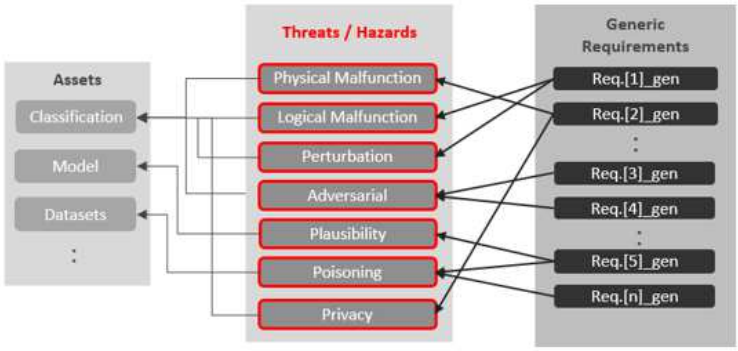

Fig. 2: Schematic overview of the generalized scoping process showing an exemplary definition of assets, Threats/Hazards, Requirements and their mapping.

next figure show the entire scoping process. At this stage, the modelling of Threats and Hazards are unspecific and contain only a generic description about the potential impact to the assets. The Threat/Hazard model will be substantiated within the requirement(s) mapped to the Threat/Hazard. As an example, consider a access control system based on biometric feature recognition (e.g. iris or face recognition). An assumptive attacker might conduct an adversarial attack on the image recognition model to circumvent the access control. Due to restrictions from application policy, the system shall operate only within a defined time window, range of brightness and noise conditions. Thus, an requirement shall consider the effort and the degrees of freedom of an attacker. Furthermore, for the sake of plausibility, only the relevant parts of the image shall be taken into account for decision. An exemplary requirement might be formulated as "The ML model shall be resistant to *white box adversarial attacks* at *constraints*." for the Threat „Adversarial" and "The ML model shall use no more than *percentage* of the background information for classification." for the Hazard „Plausibility". The requirements demand that the system shall be resistant against white box attacks considering specific constraints and shall not utilize the background information of the image for the decision process. However, it does not state the level of resistance or defines the type of white box adversarial attacks or the constraints. The italic text symbolises a variable to be specified later for the actual system and application.

4.2 Evidence Inspection

After the scoping phase, the auditees shall provide the relevant evidence to the auditors. A list of evidences needed for testing and auditing the system in support of the formulated requirements and the desired test depth is compiled and agreed upon by both parties. For example, these evidences may consist of code, documentation, copies of models or even an entire virtual environment of the system. Since the provided evidences are the basis of the

audit outcome, they have to be selected carefully.
Upon receiving, the evidences have to be closely inspected to answer the following questions: Does the provided evidence reflect the information provided during scoping phase? Is the provided evidence sufficient to support the requirement? If the auditors conclude that the first question can not be affirmed, the auditors and the auditee have to reenter scoping phase and adjust the risk assessment and requirements. If, after inspection the verdict states that the second question can not be affirmed, additional evidence has to be provided in order to enable testing.

4.3 Test Specification

After collection and sighting of all relevant evidences and definition of requirements is finished, the specific tests and audit activities for each requirement is constructed. For some of the generalized requirements a refinement may be necessary by specifying thresholds or boundary values, such as error rates or perturbation boundaries. For definition, the documentation of the system and domain experts shall be consulted by the auditors. Coming back to the example requirement defined in Section 4.1. It may be refined by specifying the type of attack the model and constraints shall be tested against: "The model shall be resistant against PGD attack $\epsilon = 0.5$ for 500 input samples at min. and max. brightness condition" and "The ML model shall use no more than 10% of the background information for classification." An suitable method to test the second requirement could be based on methods form the filed of Explainable Artificial Intelligence (XAI) (e.g. GradCam). Based on this specification, technical tests in which the ML model is tested against adversarial robustness and checked for plausibility calculated is defined. The specific refinements of

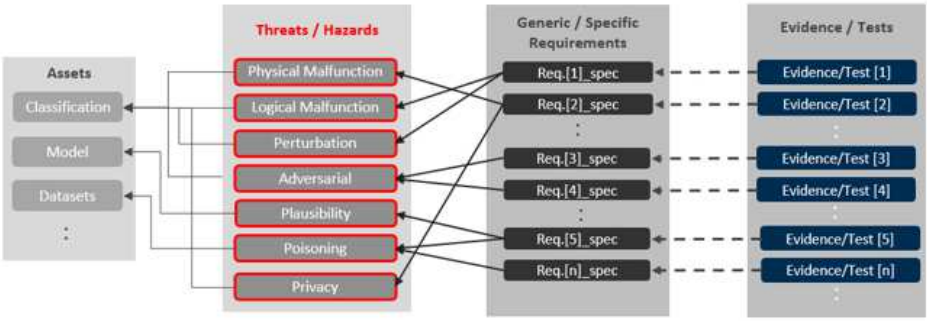

Fig. 3: Schematic overview of the mapping between specific requirements and their corresponding tests.

the general requirement reflect the use-case specific demands on the required security and saftey levels and the residual risk determined during the scoping phase. However, due to limitations of robustness estimation and the incomplete verification methods for ML models, some requirements might not be covered by testing. For instance, Norms and Standards

within high risk automotive applications require extensive (formal) verification testing which can not be met completely by using ML methods.

4.4 Test Execution

During test execution phase the tests corresponding to the prior defined requirements are performed by the auditors.The methods and metrics calculated shall be state-of-the-art, e. g. in the area of adversarial attacks or XAI. Aim of the test execution is to prove whether the AI system met the requirements. Specifically, the test execution against the prior described exemplary requirement shall prove that within the operating conditions, a successful attack is not possible or unlikely and the background information used for decision is under the specified threshold. Metric-based tests are executed with qualified testing tools suitable for the architecture and data of the system. If applicable for the use case, the metrics shall also covering (non-ML related) requirements from existing Standards and Norms. Tests based on the review of evidences such as documentation or code, shall be done with respect to the specified requirements. All tests shall be performed exclusively by qualified personnel preventing manipulation and ensuring high quality of test results.

4.5 Test Result Evaluation

After test execution, the auditors have to check whether each test result meets the requirement it was derived from. Each result must be accompanied by a estimation of the residual risk. For tests that require documentation review, available information is interpreted and judged whether the conditions stated in the requirement are fulfilled. For technical metric-based tests, the resulting metrics have to be analyzed and interpreted in regards to the boundaries and acceptance criteria derived during test specification phase. After the tests for each requirement are assessed and evaluated, a final verdict is given by the auditors, whether the system passes or fails the audit. A concluding audit report documents the entire process, generalized and specific requirements, test specification and execution as well as test results, their evaluation and the final verdict.

5 Discussion and Future Work

Auditing AI systems is a challenging task that is not entirely generalizable for arbitrary systems. GAFAI generalizes an AI-specific audit process, by defining a structured approach from the definition of high level requirements to the evaluation of specific tests.
General requirements defined during the scoping phase have high potential to be transferable between different systems. Thus, a set of general requirements throughout several different audits could reduce the effort of following AI audits. The generalized set could also be

refined into smaller subsets applicable to certain domains or architectures. Additionally, such sets of requirements provide a basis for the comparison between different audits. However, a structured approach of integrating updates and online learning AI systems into the audit process, remains open. Because ML models are data driven any retraining or operational domain shifts, impact their functionality and robustness. Resulting in a need for re-auditing the entire system.Further, AI-specific norms and regulations are needed to enhance the overall auditability of AI systems. Especially, for aspects such as ethical and legal implications, regulations are needed to guide the evaluation of fairness metrics.

References

[AG21] Akula, R.; Garibay, I.: Audit and Assurance of AI Algorithms: A framework to ensure ethical algorithmic practices in Artificial Intelligence. CoRR abs/2107.14046/, 2021.

[AIA21] Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts, Proposal for European Union Law, 2021.

[Be21] Berghoff, C. et al.: Towards Auditable AI Systems, eng, Report, Bonn, May 2021, URL: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf?__blob=publicationFile&v=4.

[BFR17] Biggio, B.; Fumera, G.; Roli, F.: Security Evaluation of Pattern Classifiers under Attack. CoRR abs/1709.00609/, 2017.

[Bu21] Bundesamt für Sicherheit in der Informationstechnik: AI Cloud Service Compliance Criteria Catalogue (AIC4), eng, Report, Bonn: BSI, Feb. 2021, URL: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf.

[De20] Deutsches Institut für Normung: German Standardization Roadmap on Artificial Intelligence, eng, Report, Berlin: DIN, Nov. 2020, URL: https://www.din.de/resource/blob/772610/e96c34dd6b12900ea75b460538805349/normungsroadmap-en-data.pdf.

[GM21] Russell, D., 2021, URL: https://blog.google/products/maps/google-maps-101-ai-power-new-features-io-2021/, visited on: 05/14/2022.

[GSS14] Goodfellow, I. J.; Shlens, J.; Szegedy, C.: Explaining and Harnessing Adversarial Examples, 2014.

[Hi19] High-Level Expert Group on AI: Ethics guidelines for trustworthy AI, eng, Report, Brussels: European Commission, Apr. 2019.

[In22] Institut der Wirtschaftsprüfer: Entwurf eines IDW Prüfungsstandards: Prüfung von KI-Systemen - IDW EPS 861, de, Report, Düsseldorf: IDW, Feb. 2022, URL: https://www.idw.de/blob/134852/bf9349774314723f6246ba73fefc491f/idw-eps-861-02-2022-data.pdf.

[KB18] Këpuska, V.; Bohouta, G.: Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). In: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). Pp. 99–103, 2018.

[La15] Lapuschkin, S. et al.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 10/, July 2015.

[LXL20] Li, L.; Xie, T.; Li, B.: SoK: Certified Robustness for Deep Neural Networks. ArXiv abs/2009.04131/, 2020.

[Ma19] Magrabi, F. et al.: Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications. Yearbook of Medical Informatics 28/, Apr. 2019.

[Mo21] Mock, M. et al.: An Integrated Approach to a Safety Argumentation for AI-Based Perception Functions in Automated Driving. In: Computer Safety, Reliability, and Security. SAFECOMP 2021 Workshops. Springer International Publishing, Cham, pp. 265–271, 2021, ISBN: 978-3-030-83906-2.

[Mö21] Mökander, J. et al.: Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. Science and Engineering Ethics 27/, 2021.

[MS20] Ma, L.; Sun, B.: Machine learning and AI in marketing – Connecting computing power to human insights. International Journal of Research in Marketing 37/3, pp. 481–504, 2020.

[Na21] Naja, I. et al.: A Semantic Framework to Support AI System Accountability and Audit. In: The Semantic Web. Pp. 160–176, 2021.

[Po21] Poretschkin, M. et al.: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz - KI-Prüfkatalog, de, Report, Sankt Augustin, July 2021.

[RSG16] Ribeiro, M. T.; Singh, S.; Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. CoRR abs/1602.04938/, 2016.

[Sa18] Saleiro, P. et al.: Aequitas: A Bias and Fairness Audit Toolkit. CoRR abs/1811.05577/, 2018.

[Se16] Selvaraju, R. R. et al.: Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. CoRR abs/1610.02391/, 2016.

[TD19] Tulshan, A. S.; Dhage, S. N.: Survey on Virtual Assistant: Google Assistant, Siri, Cortana, Alexa. In: Advances in Signal Processing and Intelligent Recognition Systems. Springer Singapore, pp. 190–201, 2019.

[To20] Toreini, E. et al.: The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Pp. 272–283, 2020.