

Conference Report | October 2024

Towards Auditable AI Systems

The LLM Auditing Challenge

based on the 4th international Workshop “Towards Auditable AI Systems”, November 10th 2023, Fraunhofer Forum Digitale Technologien, Berlin, organised by the Federal Office for Information Security Germany, TÜV AI.Lab and Fraunhofer HHI.

Marc P. Hauer^{2*}, Dr. Christoph Poetsch^{2*}, Lukas Bieringer³, Dr. Antoine Gautier³, Alessandro Petri³, Dr. Henrik J⁴. Putzer, Wojciech Samek^{1*}

*Contact:

Marc P. Hauer (marc@tuev-lab.ai),
Dr. Christoph Poetsch (christoph@tuev-lab.ai) and
Wojciech Samek (wojciech.samek@hhi.fraunhofer.de)

¹Fraunhofer HHI, ²TÜV AI.Lab GmbH, ³QuantPi GmbH, ⁴cogitron GmbH

Introduction

The increasing prevalence and complexity of Artificial Intelligence (AI) systems in recent years, have raised significant concerns about their trustworthiness, fairness, and ethical implications. The need for robust auditing procedures becomes paramount as AI technologies continue to be integrated into various aspects of society, from healthcare to finance, and given new regulatory requirements such as those laid down in the EU AI Act for high-risk uses or in the US state of Colorado's law with a focus on bias mitigation. In fact, to quote Henry Kissinger, Eric Schmidt and Daniel Huttenlocher in "The Age of AI" [Kissinger et al. 2021]: "Developing professional certification, compliance monitoring, and oversight programs for AI – and the auditing expertise their execution will require – is a crucial societal project" of our times. Robust auditing procedures are necessary not only to ensure compliance with regulatory frameworks but also to foster public trust and acceptance of these advanced technologies, which, in turn, is a crucial ingredient to boost their speedy market penetration and uptake. Central to such auditing procedures are the concepts of transparency and inspectability, which serve as fundamental pillars for evaluating AI systems.

AI systems present a unique challenge for auditing due to their non-deterministic and black-box nature, the variability in their outputs, and the complexity of their underlying models. Traditional auditing approaches often struggle to adapt to the dynamic and probabilistic aspects of AI, where systems can learn and evolve over time, leading to unpredictable behaviour. This makes the creation of robust audit frameworks for AI critical, not only to address the technical risks of AI but also to handle concerns related to data privacy, bias, ethics, and fairness.

These concerns become exacerbated when it comes to large language models (LLMs) and other generative AI systems. These models differ from traditional AI models in their ability to generate new content—such as text, images, or code—rather than simply classifying or making predictions based on input data. This capability is driven by vast amounts of training data and sophisticated architectures like transformers, which allow these models to capture complex patterns, relationships, and context across large datasets. Auditing LLMs and other generative AI models presents new challenges compared to auditing regular AI models due to their large scope in training data, model size and application fields, as well as the fact that their inner working mechanisms are still poorly understood. This makes assessing their inherent bias, safety, transparency and other audit relevant properties particularly challenging. Nevertheless, this report sets out to meet these challenges and give an overview over relevant concerns and approaches in the AI auditing landscape with a focus on LLMs.

This conference report is based on the 4th international workshop “Towards Auditable AI Systems” and presents a comprehensive exploration of the current challenges, opportunities, and methods associated with auditing AI systems. Each chapter addresses a critical aspect of AI auditing, laying the groundwork for a more structured and systematic approach to ensuring compliance, performance, and trustworthiness in AI technologies.

Chapter 1 of this report discusses the importance of transparency and inspectability in the context of auditing procedures for AI models. It provides a definition of these two terms and emphasises that a combination of the two is essential for a thorough audit, specifically for complex systems like generative AI models. Chapter 2 explores how AI explainability methods can be used in the audit of AI systems, particularly in uncovering flaws and bias of AI models. Moreover, it is pointed out that in the case of LLMs the explainability problem becomes more difficult. Chapter 3 is concerned with the challenges and opportunities for auditing the fairness of AI systems, calling attention to differences in the definition of fairness by different stakeholders and the resulting difficulties in the operationalisation of this concept. It introduces assurances cases as a method to provide thorough argumentation for the fairness of an AI system. Chapter 4 discusses the importance of implementing mitigation measures, specifically guardrails, to ensure the proper use of generative AI systems, like large-language models. Chapter 5 explores the role of different predefined processes that are standardised by harmonised rules in ensuring the implementation of trustworthy AI. It introduces some of these standards, as well as a preliminary checklist that can be used during an audit. Chapter 6 outlines the importance of auditing AI systems using a risk-based systems engineering approach to ensure performance, compliance, and trustworthiness. It emphasises that these qualities—especially trustworthiness—must be integrated throughout the development process to ensure a product meets regulatory and ethical standards. Chapter 7 presents an outlook on the evolving landscape of AI audits in the context of generative AI and large-language models and highlights several key areas of focus that are instrumental in enabling progress in the field of AI audits.

As AI continues to transform industries and societies, auditing and certifying AI systems, including LLMs and other generative systems, will be key to ensuring their safe, ethical, and trustworthy deployment. The development of systematic auditing frameworks—encompassing aspects like transparency, fairness, compliance, and trustworthiness—will enable a new era of accountable AI, fostering public trust and paving the way for broader societal acceptance of AI technologies. By addressing the challenges outlined in this conference report, we move closer to the goal of creating truly auditable AI systems.

Content

Conference Report August 2024.....	1
Introduction	2
1 Transparency and Inspectability as the Basis for Auditing Procedures.....	5
2 Auditing Based on Explanations.....	7
3 Auditing Fairness	10
4 Auditing Mitigation Measures to Guard the Use of Generative AI Systems	12
5 Auditing the Trustworthy AI Readiness.....	14
6 Auditing the Risk-Based Systems-Engineering Approach and Compliance.....	17
7 Outlook and Call to Action.....	20
8 References	24

1 Transparency and Inspectability as the Basis for Auditing Procedures

Terms such as transparency and inspectability are often undefined, diffuse and used interchangeably. In the context of auditing capabilities, clear definitions are not only helpful and “nice to have” but an important necessary condition to ensure that all stakeholders have a common understanding. For the remainder of this paper, we refer to the definitions provided by Hauer et al. [2023a].

Transparency is the disclosure of static information such as data, processes, and results to relevant stakeholders (see ISO/IEC DIS 22989). It allows the actor(s), on the one hand, to explain and justify parts of their system, and a forum, on the other hand, to judge the appropriateness of these parts based on the information provided [Hauer et al. 2023a]. The accountable forum "can be a specific person, such as a superior, a minister or a journalist, or it can be an agency, such as parliament, a court or the audit office" [Bovens 2007]. For auditing purposes relevant information that can be made transparent includes:

- The identity and the contact details of the provider and its authorised representative(s)
- The application scenario(s)
- The considered requirements of the application scenario(s)
- The requirement documents
- The objectives of applying the system
- How, when and what training data has been collected
- The specific operationalisation of the training data
- The labelling process of the training data
- The reasoning behind the parameter selections
- The data pre-processing techniques applied
- The statistical information regarding the training data set
- The training data set itself
- The specific machine learning (ML) method and relevant implementation details
- The quality evaluation process and relevant KPIs before system release (offline) and in operation (online)

Note that this list is in line with the EU AI Act. However, it also goes beyond it and aims to

address a more technical audience.

Inspectability allows the forums themselves to explore the system and, thus, does not require relying on information provided by someone else. Such mechanisms can be implemented, for example, in the form of access to an application programming interface (API for short), to enable a forum to query its own inputs to the system and evaluate the resulting outputs [Diakopoulos 2014]. To ensure inspectability, providers and deployers may need to make the following accessible to the forum:

- The data sets used for training, validation, and testing
- The output of the model for any specific (self-selected) input
- The output of the model for any input in operation

Inspectability, which relies only on input-output pairs, is already very powerful. It allows to systematically investigate the robustness of the model, e.g., regarding the effects of small shifts in the input. Furthermore, it allows a broad selection of local and global explanation approaches to be computed (global explanation refers to model-level functioning whereas local explanations would center around explanations regarding the output for a specific input). In this way, the forum can 'ask questions' directly to the system without being dependent on a provider or deployer [Hauer et al. 2023].

More thorough inspections that require access to the model's internals and/or the training data provide more detailed insights (e.g., understanding of the internal representation and neuron encodings, influence of individual training samples, etc.). They can also be more efficient than inspections based solely on input-output pairs. The development of such advanced inspection tools for generative AI models, particularly LLMs, is of great interest to researchers and regulators alike. However, they require a combination of transparency and inspectability mechanisms, where in many cases access may only be enabled to qualified and authorised bodies for reasons of trade secret protection.

2 Auditing Based on Explanations

While we outlined how transparency and inspectability form basic preconditions for auditing – also advanced – AI systems, let us examine the applicability of the explainable AI (XAI in short) concept: Explainability is another over-specified concept. It has been used very ambiguously in the literature and can have different meanings and be used for different reasons by different stakeholders. For instance, users of an AI system may be preliminarily interested in understanding the prediction made for their specific case, e.g., in a medical diagnosis scenario. Various methods have been proposed for computing such local explanations, such as 'Layer-Wise Relevance Propagation' (LRP) [Bach et al. 2015], the aforementioned 'Local Interpretable Model-agnostic Explanations' (LIME) [Ribeiro et al. 2016], 'Integrated Gradients' [Sundararajan et al. 2017] or 'Meaningful Perturbations' [Fong and Vedaldi 2017].

Although each of these methods is based on a different principle, they all essentially aim to quantify the importance of individual input dimensions (e.g. pixels) for the prediction. Recently proposed concept-level explanations, such as 'TCAV' [Kim et al. 2018] or 'Concept Relevance Propagation' (CRP) [Achtibat et al. 2023], aim to explain the individual decisions of an AI model in more human-understandable terms, e.g., in terms of commonly used attributes of an object -such as color, texture or shape- in an image classification task. In addition to understanding a particular decision of the model, the same end users may also want to know what they need to change in order to change a given decision, e.g., in a credit evaluation scenario. Various methods have been proposed to compute this type of counterfactual explanation [Verma et al. 2020].

Other stakeholders, such as auditors and regulators may have a preliminary interest in better understanding the global behaviors of the AI model, e.g., in order to better understand its limitations. This area of research also has seen enormous progress in recent years, e.g., with methods that explain the representation of individual neurons [Nguyen et al. 2016, Bau et al. 2017] or identify prototypical behavioral patterns [Lapuschkin et al. 2019, Dreyer et al. 2024].

Another stakeholder group is AI developers, whose goal is to debug and improve the model by using explanations (and of course other tools). In addition to local and global explanations, this community may also be interested in obtaining data-centered explanations [Koh and Liang 2017, Yolcu et al. 2024], which quantify the influence of individual training data samples on the model decision, to identify mislabeled, underrepresented or flawed samples. There are other stakeholder groups and types of explanation methods, see for example the review by [Linardatos et al. 2020].

In recent years, spectacular cases have been reported where explanations have helped to

reveal flaws in the behaviors of AI models and biases in the training data used to train these models. For example, [Lapuschkin et al. 2019] report that the AI model that won the prestigious PASCAL VOC Challenge classifies images of horses into the "horse" class mainly based on a copyright watermark, that accidentally was present in a large fraction of the horse images in the PASCAL training and test dataset. This 'Clever Hans' artefact went unnoticed for almost a decade, despite hundreds of world-class researchers participating in this prestigious international challenge. Many such cases, such as huskies being classified as wolves due to snow in the background [Ribeiro et al. 2016] or radiological images being classified based on the type of scanner used, have been detected with the help of explanations over the years. These results clearly demonstrate the potential of using explanations for auditing purposes.

However, in order to use explainability for a systematic and reproducible assessment of AI, it is necessary to define quality metrics and audit procedures for the explanations themselves. Irrespective of the method used, audits of explanations should assess whether the necessary requirements for the documentation of explainability methods are met. According to DIN SPEC 92001-3, this documentation should specify:

- The applicability (e.g. data type, system type) and the underlying assumptions (e.g. feature independence, differentiability) of the explanation method.
- Relevant verifiable properties, approximation errors, and/or an explicit description of the algorithm (e.g. average predictions) to justify that the computed explanations mitigate opacity as intended.
- Whether the explanation method contains non-deterministic components (e.g. sampling of feature subsets, sampling from the baseline distribution), a stability analysis of the algorithm's output (e.g. with summary statistics, confidence intervals).

Other criteria for assessing the faithfulness of explanations exist and have been widely used in the literature. The newly developed Quantus toolbox [Hedström et al. 2023] implements more than 30 of these evaluation measures, which are grouped into six different categories, namely "Faithfulness", "Robustness", "Localisation", "Complexity", "Randomisation" and "Axiomatic". While the evaluation of explanations is still an ongoing topic of research, these recent advances were important steps towards the development of standardised and reproducible procedures for the explanation-based assessment of AI models.

As elaborated in the beginning of this whitepaper, the development of LLMs importantly adds another dimension of complexity to the assessment problem. While various concepts of explainability are thoroughly discussed for classification tasks in the literature [Burkart et al. 2021], it is much more ambiguous in the generative scenario. For example, asking an LLM

Question: "What is the capital of Germany?"

Answer: "Berlin".

What is a meaningful and useful explanation in this case? In contrast to the usual (image classification) setting, it may not be sufficient to simply identify the relevant dimensions in the input, i.e., the words in the question. Simply highlighting the words "capital" and "Germany" could be considered as an explanation, as it relates the answer back to the question, but it is not a very useful one in the sense that it does not reveal anything about the reasoning of the system beyond the obvious. Alternatively, one could ask the LLM to explain its answer. In some cases, such a request will provide useful information, e.g., the LLM provides a description of the concept of "a capital" and / or provides information about the history of the capitals of Germany. This is more like what humans would do when explaining a decision.

This interactive concept of explainability (interactive because the user could ask the system to refine its answer) is a very interesting research direction. Finally, an explanation could also be factual. For example, the model could tell us which documents (e.g., Wikipedia articles) in the training data or external factual databases support or are consistent with the answer. Such a data-centered explanation is arguably very useful for LLMs.

Generating explanations for more complex responses of the model, e.g., a generated text or story (instead of a one-word answer to a simple question), as well as auditing the explanations of predictions made by generative AI, is a topic of intense research and development activities. The recent transfer of successful methods such as LRP to transformer architectures and generative settings [Achtibat et al. 2024] opens new opportunities.

3 Auditing Fairness

Switching from transparency, inspectability and explainability to fairness considerations of AI systems, several further auditing opportunities and challenges arise. First, thus far neither the EU AI Act nor any other regulation or standardisation corpus clearly specifies how fairness should be assessed. The main challenges of defining generally applicable operationalisations of fairness are well known and much debated:

- What aspects are covered by the term "fairness"? While fairness as an ethical principle has been debated for centuries, in computer sciences the term is usually used to refer to some kind of operationalisation of non-discrimination. The latter, in turn, can be interpreted in a plethora of ways.
- Legally, "discrimination" is assessed by process, not by computation of values. The focus is often on the question of whether sensitive information such as age, gender, religion or origin was a determining factor in a decision or treatment.
- Some measures of fairness are mutually exclusive.

For most of these challenges there is no "one size fits all" solution. Therefore, a more process-oriented approach may be more feasible, such as developing and assessing assurance cases. An assurance case is a reasoned argument supported by a body of evidence that states that a system meets a broad claim, the so-called "main claim" [Rinehart et al. 2017]. The main claim is broken down into sub-claims, which are either also based on the fulfillment of hierarchically structured sub-claims or can be derived directly from the evidence. Any decomposition of a claim is made explicit by arguments that explain the idea behind a decomposition. In addition, any assumptions relevant to the conclusion that the sub-claims imply the main claim are made explicit and linked to the argument. Specific evidence may be provided, for example, by technical documentation or test results. Other contextual information may be added to facilitate the understanding of an argument. In summary, the aim of an assurance case is to provide an argumentation framework in which the statement that the evidence supports the claim under the given assumptions can be justified, understood and documented.

The framework is part of the standard toolset of auditors to assess the safety and security of, for example, vehicles, aircrafts and other avionics (see, for example, ANSI/UL 46001). In recent years, there has been a growing interest in developing assurance cases for non-functional properties, outside the safety and security bubble [Hauer et al. 2021, Porter et al. 2023].

The general idea is that manufacturers of an AI-based product develop an assurance case to argue why they consider their product to be fair. External auditors can then inspect the case,

including the assumptions and the evidence, and assess whether the arguments and assumptions are valid, and if the evidence is sufficient. If the auditors are not convinced of the validity of an assumption, they may require it to be changed to another claim to be supported by evidence. If the evidence includes test results, the provision can be automated, potentially providing protection against unwanted changes, due to, for example, further training and updates.

Once an assurance case has been developed for one product, parts of the argumentation can be reused for other products, from other vendors and even across product families. If the same software is purchased and used by multiple companies, it may be necessary to test and audit it several times in different application contexts. An assurance case can support Notified Bodies in assessing such software and thus speed up conformity assessment procedures. Over time, best practices should emerge on how to argue effectively about the fulfillment of certain non-functional, particularly ethical, requirements, such as fairness.

There are first guidelines on how to do develop an assurance case for non-functional properties [Kunze et al. 2024] and even first documented experiences from practical application showing quite promising results [Hauer et al. 2023b]. However, there is still a lack of practical (and shared) experience that is needed to unlock the true potential of the approach.

4 Auditing Mitigation Measures to Guard the Use of Generative AI Systems

With the advent of general-purpose models, which are commonly used for LLM-based chatbots such as ChatGPT, users have gained considerably more freedom in the format of their input to systems. On the one hand, this has the significant advantage of speedier adoption of AI systems: ChatGPT achieved 1 million users in an incredibly short time – just 5 days after launch. This is record-breaking compared to other technologies; even social media giants like Instagram took months to reach that milestone.

In the case of an LLM-based system developed for Q&A, users can ask questions in natural language rather than having to learn a specific syntax and use a specific vocabulary to ensure that the system can parse/interpret the questions appropriately. As a result, non-AI experts, and even non-IT experts can easily derive value from such systems, significantly lowering the entry barriers for AI adoption. On the other hand, this freedom of input format opens the door to uses of the system which are not aligned with its intended purpose (in the worst case, a malicious use).

One suggested way to guide the use of AI systems is to introduce so-called guardrails [Rebedea et al. 2023]. While 'guardrail' is an umbrella term used to describe many different concepts, we will stick to the following definition: a guardrail is a function of an AI system's input that can either block its use for that particular input or not, in which case it will do nothing. A simple example would be a function that verifies that no email addresses are mentioned in the input in order to prevent leakage of personal information in the system. Another, more sophisticated example would be a function to prevent the translation of toxic content with an AI system designed for translation; a guardrail could decide if the input text being requested for translation contains toxic content. If this is the case, the input is not sent forward to the AI system for translation. Otherwise, the system translates the input as if there were no guardrail.

Including such a guardrail changes the system and can potentially affect other aspects of the model, especially if the guardrail is not perfect. Continuing with the previous example, if the toxicity detector is biased in the sense that it tends to falsely identify toxic content more often for inputs written in American English than for inputs written in British English, then the guardrail could introduce a new bias into the system [Welbl et al. 2021]. This example illustrates the importance of reassessing the AI system in its entirety when it is modified with a guardrail.

Comparing the assessment with and without the guardrail using the same set of metrics allows an understanding of the impact of the guardrail on the intended use of the model. However, additional metrics should be added to verify the efficiency/performance of the guardrail itself. In the previous example, this would mean comparing how often the system

refuses to translate inputs with toxic content. A practical approach to this, although not perfect, could be to insert content from a data set used to develop AI models designed to identify toxic content, into points of a dataset used to verify the quality of a translation model. Finally, for sophisticated guardrails, a robustness analysis should be performed to identify jailbreaks, i.e. workarounds to circumvent the guardrail.

5 Auditing the Trustworthy AI Readiness

A manufacturer's overall capability in terms of trustworthy AI readiness (or AI quality maturity level) is a strong indicator of the claimable characteristics of its product and therefore key to the trustworthiness of its AI-based systems. This capability is built on well-defined, documented processes that are thoroughly implemented, as demonstrated by existing products. An audit is conducted to examine these processes and assess their level of maturity. Once a manufacturer can prove that its processes are well-designed and effectively implemented, it can claim the ability to manage trustworthiness in its AI systems. The audit results, based on the principle of presumptive effect, confirm this capability, leading to trustworthy AI applications. This applies not only to the overall AI product but also to its components, such as foundational models.

To perform a meaningful audit, initially the scope must be determined accordingly. In general, all process areas should be audited according to the manufacturer's business process map. Examples of relevant questions in such process-level areas are detailed in the following list:

- Management processes
 - Is there an AI strategy linked to the company strategy?
 - Are there structures in place to enforce, control and audit the AI-related processes?
 - How is the AI strategy and the process requirements harmonised with other objectives such as ISO 9001 quality, ISO 14001 environment, ISO 27001 cybersecurity, ISO 31000 risk management, ISO 55001 asset management, etc.?
 - How is compliance with the EU AI Act ensured?
- Supporting processes
 - Are all supporting processes aware of AI specifics?
 - Are all supporting processes well correlated with AI-related processes?
 - Is the requirements engineering adapted to the AI needs?
 - Is the configuration management correctly covering the data, development process attributes, and AI artefacts?
 - Is the change management adapted to AI needs?

- Are pre-trained (foundational) models handled correctly?
- Is the tool chain appropriately qualified?
- Development processes
 - Are trustworthiness risks correctly identified and estimated?
 - Are trustworthiness measures taken?
 - Are trustworthiness measures implemented according to the state of the art?
 - Are typical fault models for AI components considered?
 - Does the assurance case sufficiently cover all aspects of trustworthiness for AI?
 - Is the validation sufficiently performed?
- Surveillance & Governance Processes after release
 - How is the product being used?
 - Are there any anomalies during use?
 - Are there anomalies in similar products even of other manufacturers?
 - Are change requests processed correctly?
 - Are updates sufficiently verified and validated (are resources available)?
 - Are updates applied to all required products / items?

A more complete checklist for an audit needs to be derived from a reference approach. These can be found by analysing current standards on AI as well as from consolidated expert knowledge. Relevant standards are provided by international organisations and groups such as the ISO JTC 1/SC 42 or by national bodies such as the VDE DKE that provided the standard (an application rule) VDE-AR-E 2842-61 Development and Trustworthiness of autonomous/cognitive systems.

Some industries already provide checklists for inspections or audits that focus on the processes and good practices but also provide hints on the link between product characteristics (like safety, security, usability, ethics) and the development process and methods. Examples here are the reference list for Notified Bodies in the medical domain² or the process assessment model within AutomotiveSPICE 4.03 covering MLE.x (machine

learning engineering processes 1 to 4) processes. However, these are only the first steps towards a truly harmonised and complete audit. More research is needed, and more experience needs to be gained in real projects to provide a thorough basis for complete and effective audit schemes. However, a lot more can be done than just hoping that a genAI model is correct. There are already good practices during design and verification (e.g. according to an AI-related V-model) to ensure a certain level of correctness and low uncertainty in the AI model. All these activities provide evidence used in an assurance case and its trustworthiness argumentation.

6 Auditing the Risk-Based Systems-Engineering Approach and Compliance

The primary objective of any company is to deliver a product that is not only high performing, but also meets the requirements of regulatory compliance and trustworthiness. These essential attributes – performance, compliance, and trustworthiness – must be intricately woven into the fabric of the product throughout its development phase. This integration is critical because if these attributes are not embedded from the very beginning (quality by design, compliance by design, trustworthiness by design), it may not be possible to demonstrate them convincingly or it may be much more difficult and challenging to ensure their presence later on.

Hence, an audit of the development process is required to ensure that these fundamental qualities are indeed part of the product. This audit critically assesses the extent to which the product has achieved the desired levels of performance, compliance, and trustworthiness.

For performance and compliance, the audit methodologies are in fact relatively straightforward. Performance can be verified through stress testing, which simulates defined use cases to ensure that the product can handle the expected loads. Compliance, on the other hand, is assessed by tracing how the product meets regulatory requirements, ensuring that all legal and industry standards are met. Trustworthiness, finally, encompasses a wider range of considerations regarding trustworthiness dimensions including safety, cybersecurity, usability, and ethics, among others. Each aspect represents a commitment to mitigate risks that could compromise the integrity of the product and its uses in practice (in real world settings outside the operational design domain):

- Safety implies the product does not pose any risks that could harm an individual.
- Cybersecurity ensures that there are no risks from external attacks.
- Usability addresses risks associated with incorrect use of the product.
- Ethics involves mitigating risks that the product may violate any ethical standards, which itself may involve a wide range of other considerations.

Given the complexity of proving the absence or complete management of risks, a risk-based development approach is recommended and widely accepted. This approach requires an audit to evaluate the development process based on how well it identifies, assesses, and mitigates risks, to ensure that they are reduced to an acceptable or relevant regulation-compliant level.

An effective audit of a risk-based development approach should follow separate phases, like the following "five plus one" (5+1) phases, which could be aligned with standards such as ISO

31000 for risk management or more specific guidelines such as ISO/IEC TR 5469:20214 (or the upcoming ISO/IEC TS 22440) and especially the already available VDE-AR-E 2842-61 for the “Development and Trustworthiness of autonomous/cognitive systems” (aka AI system):

(1) Definition of the Solution: This phase involves a comprehensive description of the product as a black box and identification of its use cases, intended benefits, and stakeholders. It requires modeling both the static and dynamic aspects of the solution, including all elements, interfaces, and their interactions to fulfill the use cases.

(2) Trustworthiness Risk Analysis: This step focuses on identifying and analysing risks related to trustworthiness aspects such as safety and cybersecurity. Using aspect-specific methods such as Hazard Analysis and Risk Assessment (HARA), Preliminary Hazard Analysis (PHA) or System-Theoretic Process Analysis (STPA) for safety and Threat Analysis and Risk Assessment (TARA) for cybersecurity, it aims to consolidate and address hazards and resolve any conflicts between different trustworthiness aspects.

(3) Trustworthiness Concept: A trustworthiness concept is developed that comprises measures to address all identified hazards. These measures are allocated to solutions (1), detailed into mechanisms for the product, and specified in the trustworthiness manual, forming the basis of the trustworthiness assurance case.

(4) Implementation: Detailed trustworthiness functions are implemented with clear traceability, suitable architectures, analyses, verification and validation. The integrity of the implementation is documented in the trustworthiness assurance case, supported by development artefacts and work products as evidence (also known as leaf nodes when using the graphical and hierarchical goal structuring notation, GSN).

(5) Acceptance: The trustworthiness of the product and the sufficiency of the risk mitigation are evaluated based on the trustworthiness assurance case, ensuring that risks are adequately mitigated.

(5+1) Surveillance: Post-release, the use and performance of the product is monitored for anomalies or unintended uses, and the need for updates or action by the manufacturer is assessed. This phase also includes the manufacturer's and users' responsibilities to maintain the trustworthiness of the product (e.g., through an appropriately managed CI/CD pipeline), including adherence to ethical standards in advertising, sales, and use.

Overall, an audit needs to assess whether the requirements regarding performance, compliance and trustworthiness are met. Trustworthiness, in particular, should be based on a thorough risk-based approach, incorporating at least the aspects provided in the list above (or through a more detailed approach, such as the reference approach in the VDE-AR-E 2842-61). Each activity in this approach follows good engineering practices (e.g., from one of the

standards in the section above) and generates artefacts like reports on design, architecture, and traceability, as well as analysis and verification reports. The audit needs to verify whether the processes and their methods, notations, and tools are selected and applied properly. Based on this, and especially utilising the trustworthiness assurance case, it must be determined whether the development process of the generative AI product (and potential further processes for using, updating, and decommissioning the product) sufficiently reduces failures. These include, but are not limited to, systematic failures (in the approach), random failures (e.g., in the hardware), and uncertainty-related failures (see VDE-AR-E 2842-61, part 5).

Following these ideas, generative AI will leave mysticism behind and evolve into an engineering approach based on new and sophisticated capabilities of the developing (and applying) company. Furthermore, certification of generative AI becomes possible, as it (and AI in general) is no longer regarded as non-deterministic. Instead, clearly identified fault modes drive the development process. Hence, audits (and assessments) based on the risk-based approach will lead to reliable certification processes and, together with a potential customer label, to broader acceptance, even for safety critical applications in the future.

7 Outlook and Call to Action

At least initially, AI has low barriers to entry due to its low physical asset requirements, the widespread availability of open-source code, pre-trained models and the ubiquitous availability of computing resources through hyperscalers. When it comes to generative AI and LLMs such as ChatGPT, adoption has therefore been unprecedentedly rapid, and the players are numerous, increasingly moving beyond large, big tech or firms like OpenAI with significant first-mover advantages. Nevertheless, AI cannot be considered a truly 'democratic' technology. After all, large models require correspondingly large computing infrastructures, datasets and investments. This has recently led to a concentration of scientific progress in the hands of a few players and a certain reduction in transparency overall – although the largest LLM providers differ in their respective Responsible AI, transparency and trustworthiness philosophies and (partially also open-source) AI system practices. Combined with new technologies, notably the emergence and widespread use of generative pre-trained transformers (an advanced technique for LLMs), this has exacerbated the challenge of auditing and certification.

Therefore, continuous improvement and adaptation of certification frameworks – as mentioned, promising avenues for progress here are advances towards more agile certification and continuous monitoring - coupled with legally enforceable transparency requirements are essential to meet the evolving needs of the AI landscape and to ensure the trustworthy integration of AI into society. Much progress has been made in the last year on the procedural and technical aspects of certifiability to make certifiability 'feasible' at least. In particular, the EU AI Act fills a major gap by providing legally binding requirements for generative AI models, 12 months after entry into force, and for high-risk AI systems 24 (Annex III) and 36 (Annex I) months after entry into force. Many AI systems in the high-risk category (almost exclusively Annex I) will have to undergo independent third-party audits. However, given the speed of the progress in performance and breadth of applications of AI, there is much ground to be covered. Therefore, looking ahead, key topics include, but are not limited to:

- The non-deterministic nature of generative (pre-trained) models and the resulting variability in output, influenced by probabilistic distributions and vast training data, hinders the establishment of clear certification expectations and thresholds. The use case dependency of foundational models and GPTs necessitates adaptable evaluation criteria and customised assessments to ensure compliance with quality, safety, and ethical standards.
- Data privacy emerges as a critical challenge with GPTs, as each interaction may inadvertently disclose sensitive information. Effective mechanisms and privacy-

preserving techniques are essential to safeguard user data during training and interactions with AI systems.

- Challenges associated with foundational models and GPTs require innovative approaches. Emerging technologies will require agile approaches to adapt certification procedures and conformity assessment technologies to the pace of technological progress. This may require legal adjustments as well as specific organisations and frameworks such as regulatory sandboxes.
- Domain-specific certificates are essential to address use cases with different risks, as is the harmonisation of horizontal and vertical standards and certification requirements.
- A concerted effort is required to meet the ambitious EU timeline for the enactment of the AI Act. For the time being, this is most likely to be achieved through a 'best practice' approach, until certifiable norms and standards are available. This will require joint pilot projects between industry, regulators, Notified Bodies, testing companies and potential customers of AI systems.
- A testing and certification infrastructure with sufficient capacity and resources including know-how is needed to ensure that certification can enable trust for speedy market placement of AI systems and innovations, and companies are well-advised to invest in their compliance readiness and acceleration e.g. in terms of the necessary documentation requirements so as to proactively avoid unnecessary certification bottleneck situations. This will require some upfront investment and funding. One way forward could be to ensure that 'Förderprojekte' and so-called 'Reallabore' (real world laboratories), which often receive public fundings, have a clear path to become permanent operational entities providing AI quality services and that the lessons learned within these frontier contexts are systematically translated into relevant stakeholder communities' practices.
- AI skills, talent and intellectual property are becoming a national asset and are required to keep pace with the competition from other parts of the world. Initiatives such as LEAM (large European AI Models) are worth considering and funding. Certifiability and quality management should be an integral part of such efforts right from the outset. In this context, aside the product and process-level auditing considerations, person-level certification (e.g. of operators of high-risk AI systems or those charged with their socio-technical organisational embedding) and sufficient AI Literacy levels (as encouraged e.g. in AI Act article 4) are similarly important going forward.

It is evident that the emerging and hotly debated EU AI Act, together with the pioneering applications made possible by generative AI, will further fuel the interest in the certifiability of AI systems. Efforts towards pushing the possibilities of auditing AI systems – demystifying the black box problem, tackling the LLM auditing challenge, moving from regulation to real-world auditing use cases and scaling AI quality in this way – promise an excellent return on investment not just from the perspective of providers and audit stakeholders but also for society as a whole, given (generative) AI system and LLM’s proliferating real risks. As the relevance of the ethical aspects of AI continues to receive increasing attention, especially within the wider public, and as these aspects are a constitutive part of Trustworthy AI as a whole, further efforts are needed in this area, in particular with regards to the specific challenges of certification processes and the ethical assessment therein. On the one hand, the complex and challenging field of AI ethics thus requires a systematic and structured approach for two main reasons: first, to adequately address the complexities and challenges it presents, and second, to ensure that all critical aspects are adequately addressed. On the other hand, it is imperative to distinguish what is legally mandatory in the context of AI ethics from what goes beyond legal requirements, especially when it comes to certification. This involves understanding the baseline of legal compliance, while also recognising the additional ethical responsibilities that AI developers and users may have beyond the mandatory level.

Consequently, an effective ethical assessment of AI systems requires a sensible operationalisation of these considerations. This means first, that the ethical dimensions relevant to AI need to be defined and analysed in terms of their underlying systematic frameworks as well as their implicit assumptions, implications, and consequences. This step necessarily encompasses a thorough comparison and assessment of different implementations of each relevant ethical dimension. Secondly, these results need to be integrated into certification processes in a way that is both practical and effective. The critical and overarching challenge therefore is to translate abstract ethical principles into concrete, actionable criteria that can be applied in real-world scenarios and certification processes.

With the previous versions of the AI certification readiness matrix published in 'Towards Auditable AI Systems I and II' [Berghoff et al. 2021; Berghoff et al. 2022], we have already approached a definition of relevant ethical dimensions. However, the field of ethical dimensions of AI is highly complex in itself and needs to be properly integrated into the field of AI auditing as a whole. As the certification and assessment of AI is itself a highly complex field, and as many frameworks and approaches have already been published, it is therefore essential to have a bird's eye view of the field of AI certification in its entirety.

Looking at the broadest possible version of this field not only allows us to focus resources, to purposely select the most relevant testing criteria and to locate ethical criteria properly. It also allows for a decidedly systematic approach to the field, in the sense of providing a

systematic order to the various bundles of test criteria and properties of AI's trustworthiness. These criteria and properties tend to come together in a sense as mere agglomerations, but without further systematic order, and thus, in consequence, without a strong indicator of completeness and a coherent integration of technical and ethical criteria – a completeness and integration, that in turn is decisive for a well-founded claim of AI trustworthiness.

Moreover, a holistic and systematic approach to the field of AI certification can integrate existing frameworks, both at a legal and an assessment level. It allows to combine the collection of test methods and procedures as well as lists of requirements under different regulations and frameworks and future challenges and requirements. Such an approach is therefore key for the transition from the 'what' to the 'how' of AI auditing – including tackling emerging LLM auditing challenges.

8 References

[Achtibat et al. 2023] Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., & Lapuschkin, S. (2023). From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9), 1006-1019.

[Achtibat et al. 2024] Achtibat, R., Hatefi, S. M. V., Dreyer, M., Jain, A., Wiegand, T., Lapuschkin, S., & Samek, W. (2024). Attnlrp: attention-aware layer-wise relevance propagation for transformers. In *International Conference on Machine Learning*.

[Bach et al. 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7), e0130140.

[Bau et al. 2017] Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 6541-6549).

[Berghoff et al. 2021] Berghoff, Ch., Biggio, B., Brummel, E., Danos, V., Doms, Th., Ehrich, H., Gantevoort, Th., Hammer, B., Iden, J., Jacob, S., Khlaaf, H., Komrowski, L., Kröwing, R., Metzner, J. H., Neu, M., Petsch, F., Poretschkin, M., Samek, W., Schäbe, H., von Twickel, A., Vechev, M. and Wiegand, Th. (2021). Towards Auditable AI Systems. Current status and future directions. Whitepaper.

[Berghoff et al. 2022] Berghoff, Ch., Böddinghaus, J., Danos, V., Davelaar, G., Doms, Th., Ehrich, H., Forrai, A., Grosu, R., Hamon, R., Junklewitz, H., Neu, M., Romanski, S., Samek, W., Schlesinger, D., Stavesand, J.-E., Steinbach, S., von Twickel, A., Walter, R., Weissenböck, J., Wenzel, M., Wiegand, Th. (2022). Towards Auditable AI Systems. From Principles to Practice. Whitepaper.

[Bovens 2007] Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European Law Journal*, 13(4), 447-468.

[Burkart et al. 2021] Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70, 245-317.

[Diakopoulos 2014] Diakopoulos, N. (2014). Algorithmic accountability reporting: On the investigation of black boxes.

[Dreyer et al. 2024] Dreyer, M., Achtibat, R., Samek, W., & Lapuschkin, S. (2024). Understanding the (extra-) ordinary: Validating deep model decisions with prototypical concept-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3491-3501).

[Fong and Vedaldi 2017] Fong, R. C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3429-3437).

[Hauer et al. 2021] Hauer, M. P., Adler, R., & Zweig, K. (2021, April). Assuring fairness of algorithmic decision making. In 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW) (pp. 110-113). IEEE.

[Hauer et al. 2023a] Hauer, M. P., Krafft, T. D., & Zweig, K. (2023). Overview of transparency and inspectability mechanisms to achieve accountability of artificial intelligence systems. *Data & Policy*, 5, e36.

[Hauer et al. 2023b] Hauer, M. P., Müller-Kress, L., Leimüller, G., & Zweig, K. (2023, April). Using Assurance Cases to assure the fulfillment of non-functional requirements of AI-based systems-Lessons learned. In 2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW) (pp. 172-179). IEEE.

[Hedström et al. 2023] Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., ... & Höhne, M. M. C. (2023). Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34), 1-11.

[Kim et al. 2018] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In International Conference on Machine Learning (pp. 2668-2677). PMLR.

[Kissinger et al. 2021] Kissinger, H. A., Schmidt, E., Huttenlocher, D. (2021 November). *The Age of AI: And Our Human Future*. Little, Brown and Company.

[Koh and Liang 2017] Koh, P. W., & Liang, P. (2017, July). Understanding black-box predictions via influence functions. In International Conference on Machine Learning (pp. 1885-1894). PMLR.

[Kunze et al. 2024] Kunze, L., Leimüller, G., Müller-Kress, L., Reisinger, M., Hauer, M. P. (2024, March). *Method handbook: Assurance Cases for fair AI systems*.

[Lapuschkin et al. 2019] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), 1096.

[Linardatos et al. 2020] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 18.

[Nguyen et al. 2016] Nguyen, A., Yosinski, J., & Clune, J. (2016). Multifaceted feature

visualization: Uncovering the different types of features learned by each neuron in deep neural networks. arXiv preprint arXiv:1602.03616.

[Porter et al. 2023] Porter, Z., Habli, I., McDermid, J., & Kaas, M. (2023). A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI and Ethics*, 1-24.

[Rebedea et al. 2023] Rebedea, T., Dinu, R. Sreedhar, M.N., Parisien, C. & Cohen, J. (2023) NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 431-445).

[Ribeiro et al. 2016] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144).

[Rinehart et al. 2017] Rinehart, D. J., Knight, J. C., & Rowanhill, J. (2017). Understanding What It Means for Assurance Cases to "Work" (No. NF1676L-26066).

[Sundararajan et al. 2017] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning* (pp. 3319-3328). PMLR.
[Verma et al. 2020] Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:2010.10596.

[Welbl et al. 2021] Welbl, J., Glaese, A., Uesato, J., Dathathri, S., Mellor, J., Hendricks, L.A., Anderson, K., Kohli, P., Coppin, B. & Huang, P. (2021) Challenges in Detoxifying Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2447-2469)

[Yolcu et al. 2024] Yolcu, G. Ü., Wiegand, T., Samek, W., & Lapuschkin, S. (2024). DualView: Data Attribution from the Dual Perspective. arXiv preprint arXiv:2402.12118.

Published by

TÜV AI.Lab GmbH
Max-Urich-Str. 3
13355 Berlin
Deutschland

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