USE CASE STUDY

Evaluating the trustworthiness of AI applications

Lessons learned from an audit

Tomislav Nad (SGS)
Philipp Nöhrer (Leftshift One)
Sebastian Scher (Know Center)

1 INTRODUCTION

With the (proposed) EU AI Act and with it the first regulation for the development and application of artificial intelligence ("AI") systems the question rises how the trustworthiness of AI systems can be evaluated. The EU AI Act classifies AI systems at different risk levels with different requirements per level. AI systems at high risk level are subject to additional regulatory requirements. They need to undergo additional audits and conformity assessments. Hence, there is a need for assessment methodologies and practical experience on how to apply these methodologies, in order to fulfill upcoming new requirements.

As part of this study, we wanted to test how the trustworthiness of a specific AI application can be independently evaluated. In this context we define trustworthiness to be equivalent to the compliance of the application to a predefined set of requirements. Some organizations have proposed methodologies based on traditional audit principles. Table 1 list the ones which seemed to be relevant for us. From this list the "AI Assessment Catalogue - Guideline for Designing Trustworthy Artificial Intelligence" published by Fraunhofer IAIS is the most comprehensive one. Crucially, compared to other available catalogues or guidelines, it is the only one that can readily be used for an actual audit. Therefore, we picked this requirements catalog for our study, and applied it on a real AI application. We will refer to this catalogue as "FH Catalogue". We evaluated the compliance of a specific AI application to the requirements

TABLE OF CONTENTS

Table 1: Available AI audit catalogues, and their target audiences

Title	Type	Publisher	Summary
Guideline for Designing Trustworthy Artificial Intelligence - AI Assessment Catalog	Audit catalogue	Fraunhofer IAIS	Complete audit catalogue for AI applications. Covers all dimensions. The audit is based on checking documentation requirements, including documentation of the results of tests that the catalogue specifies. Audit is based on a risk-based method, the goal is to check whether the remaining risk is acceptable. In German, English version announced
Auditing machine learning algorithms - A white paper for public auditors	Audit catalogue	Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK	Despite its name, the document actually is closer to an audit catalogue than a white paper. Essentially, it is a set of guidelines.
Algorithmic Impact Assessment tool	Questionnaire	Canadian Government	Online assessment tool for evaluating the risk of automated decision-making tools. It consists of 48 risk and 33 mitigation questions.

in the catalog in a traditional audit setting, i.e., the AI developer provides sufficient proof in the form of documentation and evidence that the requirements are met. The documentation and evidence are then independently reviewed by an auditor. We refer to this process as AI audit.

1.1 ROLES

In this study we reflect on the experiences of an exemplary AI audit from the viewpoint of different roles which are typical in a traditional audit process.

There is the AI developer who provides documentation and evidence that the requirements are fulfilled. There is the independent (information systems) auditor who reviews the provided material performs interviews and verifies that the requirements are fulfilled.

Since some requirements demand a deeper technical understanding of machine learning and related topics, the auditor is paired with an AI expert in this study.

In the following we describe the roles in more detail.

1.1.1 AI developer

The AI developer is Leftshift One, an Austrian high-tech company in the field of AI and hyperautomation. Through its AI Operating System, Leftshift One realizes complex and develops resource-saving AI applications and automate business processes for their customers.

Leftshift One is developing AI services that have a wide range of potential use cases.

The documentation of the AI audit was created by the product owner, who was responsible for the end user project and therefore knew how the use case was technically implemented. In addition, an employee from the legal and compliance department assisted with the documentation, who was familiar with the regulatory requirements. Furthermore, the responsible machine learning engineer was consulted for specific questions regarding the AI application and the machine learning model.

1.1.2 Auditor

The auditor is an experienced security auditor and evaluator at SGS. SGS is the world's largest independent, international inspection, testing and verification organization with no manufacturing, trading or financial interests which could compromise its independence. The auditor hasexperience with auditing, evaluating and certifying the security of information management systems and/or products. The auditor is an expert in the field of cyber security and has a solid understanding of the various aspects of AI.

1.1.3 AI expert

The AI expert is a senior researcher at Know-Center. Kow-Center is a leading European research center for Artificial Intelligence, with strong ties both to academia and industry. The AI expert has a background in AI and topics relevant to AI-certification (especially robustness and fairness), including literature knowledge on AI audit catalogues and basic legal regulations, but no hand-on experience with certification/auditing (neither for the area of AI, nor for other areas). Their role in this exercise is that of an AI expert who assists the auditor.

Later in this document, the roles of the "Auditor" and "AI expert" will be merged into "Auditor perspective".

1.1.4 End user

The company that contracted the development of the AI application by the AI developer is called end user. The end user did not participate in the audit because auditing of the AI component was not required as the risk was rated low. In addition, the AI developer was validated.

1.2 Audit catalogue

The FH Catalogue was developed and published by Fraunhofer IAIS[1]. Overall, it provides a comprehensive and practical guide to designing and using AI systems in a way that is trustworthy, responsible, and respectful of human rights and dignity. It is currently the most comprehensive and complete catalogue available for audits of AI applications.

It is designed to enable audits for all applications that use supervised Machine Learning ("ML"). It does, however, only audit the ML-part of applications. Other parts (e.g. the software system which it is embedded in) need to be

1 https://www.iais.fraunhofer.de/

tested/certified with other testing methods. The catalogue is centered around a risk-based approach, along 6 risk-dimensions. These dimensions are

- Fairness
- Autonomy and Control
- Transparency
- Reliability
- Safety and Security
- Data Protection

The AI developer needs to conduct a comprehensive risk analysis for each dimension. In this risk analysis, the AI developer must assess the risk potential that the application would have if no measures to mitigate the risks were taken. Then, the AI developer must demonstrate – and in the end convince the auditors – that sufficient measures have indeed been taken to reduce that risk to an acceptable level. The catalogue lists a large number of concrete points that need to be shown or addressed. This is all done with respect to the risk analysis, and concluded with an analysis of the remaining risk over all dimensions. Risk dimensions in which the risk analysis shows that there would only be very little potential risk can be skipped – except for the dimension Reliability, which always needs to be covered.

1.3 Use Case Description

The AI application is part of a digitized incident documentation process in a manufacturing company. The AI application helps to provide an initial assessment of the criticality of an incident based on the description at the time of occurrence.

The advantage of the AI application is, on the one hand, that the employee who discovered the incident (usually directly on a production line) is given an initial assessment of how critical the issue is. On the other hand, it helps the quality department to carry out a cross-check and to question its own decision.

The AI application has only a supporting function (“human control”). Both, employees who detect a malfunction and the quality employees, have been requested by the employer not to blindly trust the suggestion of the AI application.

1.4 Purpose and goal of the exercise

As part of this study, the practical audit of an AI application was tested. For this purpose, the FH Catalogue was applied to the use case described above. Therefore, the AI developer documented the corresponding requirements, and the documentation was then checked by the Auditor together with the AI expert for conclusiveness, completeness, and suitability.

The research project lasted about six months and was carried out in several iterations. For this purpose, one dimension each was elaborated and then debriefed. Then it was continued with the next dimension.

The results and findings of the research project are described in this white paper.

2 GENERAL CHALLENGES

2.1 AI developer perspective

Especially in industrial companies, AI technologies are being used with the aim of further increasing automation, optimizing processes and saving costs. However, companies often do not have the know-how to develop AI applications in-house. Therefore, they use external AI specialists to jointly develop AI applications. As a result, another party is involved in the audit process, which creates new challenges: In addition to selecting a trusted AI developer, the developer must also provide their part of the documentation (such as the machine learning model and training). In this context, the question of who bears the (additional) costs for certification also arises. Moreover, the company has no direct influence on the AI development itself, which is why they must trust the AI developer in this regard.

Due to the lack of a generally applicable standard regarding the development of trustworthy AI, the AI developer cannot follow any benchmark during the development process. The Ethics Guidelines of the EU‘s High-Level Expert Group on AI provide some guidance. However, these are not an official standard, just non-binding recommendations. A later, additional documentation effort of audit after implementation is therefore likely.

Since AI applications can be designed very

differently and used in a wide variety of fields, it can be a challenge to develop general testing systems and metrics for trustworthy AI applications. It is also unclear for whom the documentation is intended for (e.g., a domain expert at the customer site, the auditor, or a machine learning expert).

Apart from this, the using company is also subject to special regulations. For example, a manufacturing company must comply with different documentation and due diligence requirements than a bank or insurance company. The external AI developer must also comply with these requirements. Otherwise, an implementation at the end customer is not possible. Therefore, all industry standards and internal specifications must be taken into account when certifying AI applications.

2.2 Auditor perspective

From the point of view of auditors, there are several challenges that do not occur in audits in other fields (e.g. software auditing), but that are specific for auditing AI systems:

- ***High level of expert knowledge is necessary:*** AI is a complex topic, and the field is developing at an enormous pace. Auditing an AI system requires specialized knowledge and skills, including expertise in machine learning, data analysis, and programming. It is unfeasible to expect that auditors will have in-depth knowledge in all these areas, considering there is currently a shortage of experts in these areas.
- ***Rapidly evolving technology:*** AI is a rapidly evolving field, and new techniques and algorithms are being developed continuously. This means that auditors may need to stay up-to-date with the latest developments in the field and adapt their evaluation methods accordingly.
- ***Reliance of the used system/model on potentially large amounts of data:*** An AI model is the result of the combination of a training algorithm (known) and data used to train it. For a trustworthy application, therefore, the data needs to be 1) trustworthy and 2) suited for the task at hand. Both issues are complicated questions in practice.
- ***Data availability:*** Auditing an AI system often requires access to large amounts of data, including training data, test data, and real-world data. However, this data may not always be available or may be difficult to obtain, particularly in cases where the data is sensitive or proprietary.
- ***Lack of transparency:*** Many AI models use thousands to millions of parameters, and thus are impossible to interpret, even by experts. While the advent of explainable AI techniques (XAI) has to started to address this issue, most models remain hard up to impossible to interpret.
- ***Constantly changing systems:*** Many AI systems are regularly updated with new data (“retrained”). From a conservative viewpoint, this would mean that we have a new AI system, that consequently would need to be tested from the start.
- ***Supply chain complexity:*** Usually AI developer’s rely on tools and systems provided by various suppliers. For a full understanding of the AI system, input from these suppliers is also required but often hard to get.
- ***Lack of standards:*** There are currently no widely accepted standards or metrics for evaluating the trustworthiness of AI systems. This means that auditors may need to develop their own evaluation frameworks or rely on subjective judgments, which can lead to inconsistencies.

In general, the challenges have a lot in common with challenges in the cybersecurity evaluation of products and system. In the cybersecurity domain, these challenges have been discussed, and different proposals for solutions to these challenges have been published. Hence, it would make sense to evaluate how methodologies in security evaluations can be applied to the evaluation of AI systems. As an example, we might need to consider a concept

of composite evaluation like it is specified in Common Criteria[2] . The idea is that suppliers seek independent certification by a certification body for their components which are later integrated in larger systems. The evaluation of the system can then rely on the certifications of the components and only needs to prove that they are used according to the guidance provided by the supplier.

Another similar challenge is the constant change of the system. In cybersecurity evaluations it comes in the form of software updates. A change in an AI model can be seen as a software update as well. Furthermore, change management is one of the most important processes in the governance system every organization needs to have. However, a further discussion on this topic goes beyond the scope of this white paper.

Finally we note that the peculiarities discussed above also have a profound impact on the necessary documentation.

3 APPLICATION OF FH CATALOGUE

According to the FH Catalogue, not all six dimensions are necessarily equally relevant for the use case. For this reason, a risk analysis (named as “protection requirement analysis” in the catalogue) is first made for each dimension in order to determine the protection requirements of each dimension. Depending on potential damages (e.g., due to malfunctions or violations of requirements), the protection requirement is to be set as „low“, „medium“ or „high“. If a low protection requirement is determined for a dimension, the dimension does not have to be considered further, since no significant risks exist (with the exception of the dimension Reliability, which always needs to be considered, independently of the determined risk).

In the present study, only the dimension Fairness was evaluated as low in need of protection requirements. This is primarily because neither personal data was processed nor are human users affected by the results of the AI application. For the other five dimensions, the need for protection was assessed as at least medium, which is why these dimensions were examined in greater depth by first conducting a detailed risk analysis for each dimension based on the risk areas. This resulted in target specifications. A series of qualitative and quantitative criteria were then defined, which were used to assess whether the targets had been achieved. At the end, a summary consideration of the dimension was made.

The challenges from the different perspectives that occurred in the course of the study are described below.

2 https://www.commoncriteriaportal.org/

3.1 AI developer perspective

The FH Catalogue requires comprehensive documentation on the AI application. Since we selected the use case for auditing after go-live, some of the requirements of the catalog had to be documented retrospectively. Standard documentation was already created during development. However, this was not adequate in scope and depth to meet the requirements of the FH catalog. Therefore, some documentation had to be created from scratch.

In addition, it was notable that some points were repeated or divided thematically in different dimensions. For example, in requirement "[TR-R-NB-MA-01] Training and Test Data", general requirements were given for training data. However, data quality was to be documented in another dimension under requirement "[VE-R-RE-KR-03] Quality of training and test data".

Furthermore, some items in the audit catalog concerned requirements that fell within the domain of the end user. This related, for example, to training or IT security measures (e.g. "[SI-R-BD-MA-01] Training and sensitization of employees" or "[SI-R-FS-MA-01] Security guidelines and instructions for use"). Therefore, another challenge was to constantly coordinate between AI developer and end user. This also required additional time and therefore some deadlines needed to be extended. Besides this, any risk analysis needs to describe the impact or damage, which is why input form the end user was needed to provide meaningful answers. If no concrete answers were possible, the team needed to estimate the consequences.

As mentioned before, some points appear repeatedly in the catalogue. This is due to the structuring of the catalogue along the risk dimensions. This makes the documentation requirements rather long, as those points need to be addressed again and again. We asked ourselves whether an alternative structuring along ML-topics would be more useful, e.g.:

- ML-model and transparency along all risk dimensions
- AI application (the application as a whole, not the ML-model) along all risk dimensions
- Systems architecture of the underlying system along all risk dimensions
- Quantity, quality and resources of training data

3.2 Auditor perspective

The main challenge from the auditor side is that there is always some room for interpretation. This is made more difficult by the fact that AI auditing is very new. Even though the FH Catalogue is the most comprehensive one, still, at many points, there is a lot of scope for interpretation, as there are not many hard or quantitative requirements.

The audit was set up in steps, guided by the risk dimensions of the catalogue. The first two dimensions covered were "Autonomy and Control" and "Fairness". Here, multiple iterations between AI developer and auditors were necessary. These rounds consisted of the AI developer providing documents, the auditors giving feedback, and then subsequently the AI developer providing improved documents.

For the dimensions covered towards the end of the process, much fewer iterations were needed.

The dimension where the process was most different to the others was Reliability. Here, the auditors needed to request feedback to a large number of small points (evaluation metrics, how exactly something was done, etc.). Given the nature of the underlying topic, this might not be highly surprising. Still, a clear outcome of this test-case was that most time was spent on the topic of reliability, even though it was done at a later stage, where there was already a more efficient workflow established between AI developer and auditors.

Due to the lack of widely accepted standards or metrics for evaluating certain requirements of the catalog the result is often subject to auditors judgement which can lead to inconsistencies.

Considering the simplicity of the target AI application, the effort for evaluating all the requirements is quite high. Going from the protection requirement low to medium is a tremendous step with respect to the number of requirements.

Additionally, the auditors were working with the developer of the AI component. Many requirements are addressing the integration and usage of the AI application which can only be handled by the end user. Ideally all parties are involved

in the evaluation process and a lot of work will be repeated for every AI application using a similar AI component.

3.3 Dimension reliability

Reliability is arguably the most important dimension and is the only one that according to the audit catalogue always needs to be addressed, independent of the risk analysis. It is also one of the most challenging ones, as a high level of AI expertise is required for the auditing party. The AI developer also had to first gather information internally from the machine learning experts.

A characteristic of this dimension is that there can be in principle a large number of quantitative approaches involved (such as computing accuracy metrics). While, at first glance this makes the task seemingly easy, the lack of generalizable thresholds/baselines makes it actually quite hard in practice. This is particularly important, however, because accuracy is a crucial factor in the selection of a machine learning model. The training and test data must also adequately (quantitatively and qualitatively) cover the application area.

From the auditor's perspective, the AI developer needs to provide convincing arguments on why certain metrics were chosen, show which values they have for the application, as well as why these values are good enough. From the AI developer‘s point of view, however, it was challenging to argue in detail why certain metrics or tests were not considered. For efficiency reasons, short answers were often given that the requirements were not relevant.

To understand the evaluation done by the AI developer, many - often small – clarification questions were necessary. This is caused by the fact that while many evaluation metrics are well-known and standardized, still, from application to application small details can vary (such as on which data exactly they were computed, and similar).

Example of a difficult case: The requirement "[RE-R-SC-ME-06] AI application real-world tests" ("Real-world test of AI-application") requires the application be tested in a real-world setting. This was not done with the AI application described in this study. The AI developer argued – in a convincing way – that a real-world test is not necessary, given the nature of the application and the data used. Moreover, for the real-world-test, the end user would have had to agree and grant the auditors access to the IT system. If taken literally, however, the catalogue does not really leave the possibility of not doing a real-world test. Therefore, this was perceived as a borderline case by the auditors.

3.4 Dimension safety and security

This dimension includes functional security and IT security. In this dimension, the general issue that it is unclear whether the AI developer or the end user company is responsible for certain requirements became especially apparent.

For example, requirement "[S-R-IA-ME-11] Logging and monitoring" requires that violations of integrity or availability (of the system) are logged. Here it is unclear whether implementation of this logging is the task of the AI developer or the end user company, and even more unclear is who is required to present the necessary evidence during the audit. Another example would be „[S-R-FS-ME-01] Safety guidelines and instructions for use", after which own security policies and usage instructions must be defined. Here, the AI developer can only help in an advisory capacity and at the same time must comply with the end user‘s internal guidelines.

The same applies for the requirements on data integrity and confidentiality of data. Here, from the auditors' perspective, several details were missing in the documentation provided by the AI developer, especially how integrity is guaranteed in the long term. These are, however, questions where the responsibility in terms of documentation is not clear from the FH Catalogue. One reason why the AI developer could not disclose all the information in detail was that a strict confidentiality agreement was signed between the AI developer and the end user. Confidential information (such as infrastructure details, audit-logs or training data) could therefore not be disclosed to the auditor. It would have been necessary to execute a separate confidentiality agreement for this purpose. However, this was not done for this study.

Another point that came up was that the auditors had difficulties in understanding the bigger picture of how the system is implemented on the end user's side. The AI

application itself – as implemented by the AI developer – was well described in the documentation. For several safety and security requirements it is, however, essential to additionally understand the embedding in the larger system at the end user's side. This was – at least in the beginning of the exercise – missing from the documentation, which, naturally, was focused on the AI application itself.

For this dimension it was especially apparent that a full application of the audit catalogue requires all parties involved to contribute to the evaluation, end users as well as all suppliers. Considering that the AI developer has multiple customers who might want to go through such an evaluation, a more efficient approach is desired than repeating the same analysis multiple times.

4 CONCLUSIONS

If it is possible to develop secure and trustworthy AI applications, the willingness of companies to integrate such AI applications into their processes and to benefit from the advantages of this technology will increase. Independent testing and auditing by third parties within the framework of certification is one way to establishing security and trust in AI applications. The creation of a trustworthy "AI made in Europe" brand can create an international competitive advantage.

The FH Catalogue is very extensive and therefore, plenty of time and human resources must be planned for documentation and auditing. This makes it well suitable for AI applications that either have highly sophisticated parts (e.g. consisting of deep neural networks) or being part of a larger complicated system. For simpler applications – which presumably most applications will be (such as the use case considered here) – the catalogue seems, at least in parts, too detailed. The fulfillment of these details requires a lot of work both by the AI developers as well as by the auditors, and thus might be an obstacle for choosing to do an audit in cases where it is not legally required. For such cases, it would be beneficiary to have some kind of an "audit-lite" approach. Here, only certain aspects, such as the training of the AI component, would be audited. This would also be beneficial for the end user, as an application with an "audit-lite" would still be perceived more trustworthy than if no audit is performed at all.

However, in our opinion, the six dimensions are sufficient to cover all aspects of trustworthy AI. They are also largely consistent with the requirements of other catalogs (e.g. the European Union's Ethics Guidelines for Trustworthy AI). As the catalog has a risk-based approach, the risk analyses are a central component. Since sometimes consequences of the AI application are difficult to assess, the elaboration of the risk analyses is also challenging. Besides that, also a risk analysis of the dimension as medium or high means that the requirements of the whole dimension have to be documented. Thus, there is no further distinction between these two risk levels. This leads to the fact that a dimension with a risk analysis medium has the same effort as the risk level high.

The specific structure of the catalogue along risk dimensions made both writing the documentation, and checking by the auditors, cumbersome at some times. This is because many points that are from a technical perspective the same or at least very similar – e.g. questions regarding the training data – occur in multiple risk dimensions, and thus need to be addressed multiple times.

Furthermore, many requirements need the involvement of the end user and hence, a full audit is only possible if all parties are part of the process. Consequently, this means that the AI developer most likely will have additional effort if a similar or the same component is used in other applications. The possibility of independently auditing an AI component would benefit the process greatly. Many unclarities during the audit are ideally resolved by having an AI management system in place defining a clear process for all relevant aspects of the lifecycle. Defining a shared responsibility model like the one used for cloud service provider[3] would further clarify the responsibility of certain requirements.

Additionally, due to the fact that the catalog is so extensive, and many topics are covered in different dimensions, it is easy to lose the

3 https://cloudsecurityalliance.org/blog/2020/08/26/shared-responsibility-model-explained/

overview. However, a certain basic knowledge of the catalog and its contents is necessary in order to be able to make meaningful references, for example. Therefore, it was very helpful for the AI developer that the knowledge about the content and structure of the catalog was gathered in one person and that this person distributed the tasks internally in the company.

Highly useful from both sides' view would be further guidance and template documents supporting the documentation and audit process. This would decrease some of the spent efforts and to some extent standardize the audit process.

Finally, throughout the study we noticed the similarity to the cybersecurity evaluation of IT systems. In cybersecurity we face very similar challenges, like required expertise, constantly changing systems or supply chain risks. This makes it worthwhile to explore the applicability of methodologies used in the cybersecurity evaluations to the evaluation of AI systems.

PUBLISHERS

SGS Digital Trust Services GmbH

SGS Société Générale de Surveillance SA, 1 Place des Alpes, P.O. Box 2152, 1211, Geneva, Switzerland
https://sgs.com/
Enquiry.Emerging-Technology@sgs.com

LEFTSHIFT ONE Software GmbH

Herrengasse 3, A-8010 Graz
https://leftshiftone.com/
contact@leftshift.one

Know Center GmbH

Sandgasse 36/4, A-8010 Graz
https://know-center.at/
info@know-center.at

This publication is part of an ongoing series on certification and conformity assessment of AI applications within the "Trust Your AI" initiative.

trustyourai@know-center.at

https://trustyour.ai/en/