Cognitive Computing and Internet of Things, Vol. 165, 2025, 179–186
https://doi.org/10.54941/ahfe1006101

AI Audit as a Tool for Effective AI Risk Management

Herwig Zeiner

Joanneum Research Forschungsgesellschaft mbH, 8010 Graz, Austria

ABSTRACT

Artificial intelligence (AI), especially large language models (LLMs), is increasingly permeating all industries and promises transformative potential. Such large language models demonstrate impressive abilities in text generation, data analysis and even creative tasks. However, this rapid proliferation and increase in performance goes hand in hand with a growing awareness and concern about the manifold risks these technologies pose. The range of potential harms extends from operational malfunctions and data privacy breaches to profound systemic impacts on society and the economy. Given this duality of benefit and risk, there is an urgent need for robust governance, standardized risk management practices, and effective mitigation strategies. AI certification, specific risks associated with LLMs, corresponding mitigation techniques (technical and organizational), and the emerging concept of systemic AI risk. The standardization landscape for AI is still fragmented, but it is showing clear signs of convergence. AI audit using standards such as ISO 42001 and specific LLM risks support this process of security impact assessment.

Keywords: AI audit, Artificial intelligence, Transparency, Traceability, Fairness, Risk management, Trust, Efficiency, Effectiveness

INTRODUCTION

The use of artificial intelligence (AI) in companies and organisations is steadily increasing. AI has been incorporated into various decision support systems, including the industry (Woitsch, 2023), the production industry (Paletta, 2024), and the building industry (Zeiner, 2019), with the objective of optimising specific processes within these fields. Additionally, AI has been implemented to accelerate the programming tasks (Fürntratt, 2023). However, this also brings with it new challenges in terms of risk management, transparency, traceability and fairness of AI systems. An AI audit can help to address these challenges and ensure that AI systems are used responsibly and in line with organisational goals.

This paper examines the basics and importance of AI audits. Particular attention is paid to the challenges and opportunities of AI audits. It will be discussed how an AI audit can help to minimise risks, build trust in AI systems and increase the efficiency and effectiveness of AI applications. Different aspects of AI audits are presented, and their areas of application are discussed. AI audits are review procedures that are conducted with the objective of assessing the compliance and quality of AI systems. They assist in ensuring compliance with legal regulations, including the AI Act (Schuett, 2024), the Data Act, and the GDPR, as well as ethical and technical standards. An important objective of AI audits is to enhance the transparency (Faßbender, 2021) and trustworthiness (Ayling, 2022) of AI applications. A variety of aspects of the AI system are examined, including the algorithms, data, and infrastructure that are utilized.

Assessing and managing the impact of AI technologies and increasing trust and acceptance of these technologies are key challenges. Recent literature therefore recognises the effectiveness of AI audits as a mechanism for verifying the legality, ethical defensibility and technical robustness of AI systems.

BASICS AI AUDITS AND RELATED WORK

An AI audit is a comprehensive examination of an artificial intelligence system that analyzes its architecture, implementation, and performance against predefined benchmarks. These audits are critical to confirming that AI systems not only perform their intended tasks, but also adhere to ethical standards and regulatory requirements. AI audits are review procedures designed to assess the quality and compliance of these systems, and thus play a critical role in managing the risks associated with AI implementation.

AI audits involve evaluating several aspects of a system, including data governance, key performance indicators, ethical considerations, and regulatory compliance. Unlike traditional technology audits, AI audits must address unique challenges related to algorithmic decision making, potential biases, and the “black box” nature of some AI systems. This includes compliance with regulations such as the AI Act, the Data Act and the GDPR. By evaluating various components of AI systems, including algorithms, data and infrastructure, AI audits help to identify potential risks in deployed applications and promote accountability in an appropriate manner.

Effectively managing AI-related risks involves addressing several challenges, such as the complexity of AI systems and the lack of transparency in training datasets. AI audits help organizations navigate these complexities by providing a structured approach to assess and mitigate risks, ultimately fostering greater trust and acceptance of AI technologies. Through this process, AI audits contribute to the responsible and effective use of AI, ensuring that its benefits are realized while minimizing potential negative impacts.

The recent paper of Birhane (Birhane, 2024) taxonomizes current AI audit practices across multiple domains, including regulators, law firms, civil society, journalism, academia, and consulting agencies. It assesses the impact of AI audits within each domain and finds that only a subset of AI audit studies leads to desired accountability outcomes. The authors then identify the practices necessary for effective AI audit results, emphasizing the connections between AI audit design, methodology, and institutional context for meaningful accountability

The work of Blösser (Blösser, 2024) explores the concept of AI certification, examining consumer trust and approval of different entities that certify AI across various decision-making domains. It discusses the AI certification landscape, the entities involved, and the factors influencing consumer approval. The study reveals that consumers generally favor non-profit entities, particularly governmental institutions, for AI certification, and that consumer approval varies across different AI decision domains.

Leocádio et al. (2024) explores how the integration of artificial intelligence (AI) is changing auditing practices. It notes that AI's ability to process real-time information, identify trends, and automate tasks is reshaping the auditing field. The paper emphasizes that while AI offers increased efficiency and accuracy, it also brings challenges such as data privacy concerns and the need for auditors to develop new skills. Through a systematic literature review, the study develops a conceptual framework to guide the use of AI in auditing.

The convergence of AI technology and auditing practices presents (see Wassie, 2024) a multifaceted landscape, with studies indicating that AI tools can indeed enhance audit quality and efficiency. However, this integration also necessitates careful consideration of ethical implications, including fairness, bias, privacy, and the need for transparency in AI decision-making. Moreover, the question of "who should certify AI" remains a critical one, with various entities such as government bodies, NGOs, and commercial third parties being considered, each bringing different strengths and potential limitations to the process

CHALLENGES OF AI AUDITS

There are quite a few challenges for carrying out AI audits (Li, 2024): The complexity of the AI systems to be tested is one of the biggest challenges for AI testers. Some large models such as ChatGPT, Gemini, Mistral are trained with large amounts of data, prioritising the quantity of data over the quality of the content. The social and human context is given little consideration, which makes testing more difficult.

The **lack of information provided on training datasets** and data management represents a significant challenge for AI audits. Firstly, the creation of documentation regarding data is often perceived as time-consuming, optional and difficult to maintain due to the lack of mandated practices, the extension of documentation and the large number of projects conducted simultaneously (Balahur, 2022). Secondly, the lack of documentation regarding which users are represented makes it difficult to determine measurement validity and the likelihood of demographic bias going undetected (Coston, 2021). Thirdly, the training data may be sourced from external providers. In such cases, the data labelling teams may be unknown, the intended properties of the dataset may not be explicit, and the datasets may contain unforeseen problems (e.g. imbalances, inaccurate data, etc.).

The **technical challenges are in adapting AI to specific use cases** (Lee, 2021). An audit engagement and findings in relation to an AI system are typically reported at a specific point in time, given that automated audits are not conducted on a large scale. However, issues can arise with dynamic algorithmic systems, such as the need for regular updates to algorithms when new data is introduced and model training to keep pace with real-world conditions

In addition, there is a **lack of quality metrics and quality control programmes** in the area of AI audits, as well as a lack of established procedures or certificates for qualifying auditors to perform audits in an algorithmic context (Raji, 2022). This is in contrast to the quality metrics and quality control programmes that apply to financial statement audits, IT audits and information security audits.

A **current challenge is the auditing of dynamic AI systems**, such as LLMs that evolve through continuous updates. These systems require *adaptive audit frameworks* capable of handling real-time data streams, version control complexities, and shifting regulatory requirements. Without tools for automated monitoring or protocols for incremental model validation, organizations risk outdated audits that fail to reflect current system behaviors. This gap exacerbates compliance risks, particularly in industries like healthcare or finance, where algorithmic drift or data decay can have immediate consequences. Bridging these practical implementation gaps demands not only better tooling but also standardized methodologies for integrating audits into DevOps pipelines and allocating resources for sustained oversight.

In addition, **new legal frameworks pose a new challenge.** Existing regulations and guidance on AI auditing and governance typically include a variety of provisions on legal, ethical and technical aspects. Yet there is a lack of concrete guidance on how regulators, auditors and companies should translate such overarching principles and objectives into concrete performance expectations and comparable measurable actions.

RISK MANAGEMENT AND CERTIFICATION STANDARDS OF AI SYSTEMS

The risk management of AI systems follows a specific process. First, the scope of application and the objectives of the organisational and legal environment are defined. This process includes identifying stakeholders and users of the AI system, as well as their expectations and defining risk criteria.

Risk management then requires a risk assessment. This process involves identifying assets for the organisation, identifying risk sources, potential outcomes/consequences and existing controls. This includes a detailed analysis of the intended use and potential misuse, the data used and the decision-making processes of the AI system. The probability of occurrence and potential consequences, such as criticality, material and immaterial effects, must be considered when assessing the identified risks. Quantitative and qualitative methods must be used for the analysis, and cascading effects and dependencies must also be taken into account. We focus on specific AI risks such as fairness/non-discrimination, environmental impact, safety, accountability, maintainability, data protection, data quality, robustness, lack of AI expertise and transparency and explanability. Finally, the analysis results are compared with the risk criteria to determine priorities for risk treatment. Risk treatment involves selecting and implementing risk management measures. Options include avoiding risks, accepting risks to take advantage of opportunities, eliminating the source of risk, changing the probability, changing the consequences, risk sharing (e.g. through contracts, insurance) or risk retention through informed decision-making. For AI, these can be specific measures such as modifying the AI model, implementing additional controls in the life cycle or accepting certain functions.

The typical risks of large language models (LLMs) such as GPT-4, Claude and Llama thus represent a separate category of artificial intelligence (Mökander, 2024). Due to their extensive and diverse training with large and diverse data sets, often taken from the internet, these models are referred to as foundation models or general-purpose AI (GPAI). Due to their extensive capacities, they are able to perform a wide range of tasks beyond their original training focus. However, these special features of LLMs also entail specific risks. These include the occurrence of hallucinations, in which the models can generate false or misleading but plausible-sounding information. Another risk is the bias that can arise when the training data contains prejudices that are reflected in the models' results. Furthermore, there is a significant potential for misuse, as LLMs could be used for harmful purposes such as creating fake news, phishing attacks or automating cyberattacks. The dissemination of harmful content, whether intentional or unintentional, also poses a risk because LLMs can generate and disseminate discriminatory, illegal or otherwise harmful texts. Copyright infringement is another potential problem that could occur unintentionally during training with large data sets. The non-transparent internal processes of LLMs further complicate troubleshooting and the assignment of responsibilities. Finally, the environmental impact of the significant computing power and associated high energy consumption required to train and operate large LLMs must be considered a particular risk.

Therefore, certification standards such as ISO/IEC 4200 have been developed to enable the certification of AI systems. This international standard is the first of its kind to define requirements for an AI management system (AIMS) that can be certified. The framework provided by this standard includes the establishment, implementation, maintenance and continuous improvement of processes relevant to the development, deployment and use of AI systems in organizations. The primary objective of ISO 42001 is to promote trustworthy AI. The standard aims to ensure the responsible development and implementation of AI systems. To this end, it helps organizations address specific challenges in the areas of ethics, transparency, accountability, bias mitigation, security, privacy, and continuous learning. An essential feature of ISO 42001 is its certifiability. In contrast to pure frameworks such as the NIST AI RMF, organizations can have the conformity of their AIMS with ISO 42001 verified and confirmed by an independent, accredited certification body. Compatibility with security standards such as ISO 27000 is given.

PRACTICAL IMPLEMENTATION STEPS

Implementing effective AI audits for Large Language Models (LLMs) requires a structured and iterative approach. The initial phase involves establishing a clear understanding of the audit's purpose by defining specific goals and the scope of the LLM's evaluation, including its intended use cases and relevant stakeholders for the business case. Engaging with both internal teams (developers, ethicists, legal counsel) and external parties (users, impacted communities) is crucial to gather diverse perspectives and identify potential risks. Subsequently, organizations must define measurable audit criteria and select appropriate methodologies and tools, acknowledging the current limitations in comprehensive tooling for LLMs. This stage also necessitates establishing robust data access protocols and assembling a multidisciplinary audit team with the necessary expertise.

The execution phase centers on conducting thorough data collection and analysis using the chosen methodologies and tools to assess the LLM's behavior, performance, and potential impacts against the defined criteria. This involves systematically evaluating the model for biases, risks of generating harmful content, and adherence to ethical guidelines. The analysis of findings should identify deviations and anomalies, prompting an investigation into their root causes. Finally, the audit process culminates in the development and communication of clear, comprehensive reports detailing the findings and recommendations. Crucially, this must be followed by the implementation of concrete remediation plans to address identified shortcomings. Recognizing the dynamic nature of LLMs, establishing ongoing monitoring mechanisms and feedback loops is essential for continuous improvement and adaptation of the audit framework.

However, successful AI audits for LLMs are not one-time events but rather integral components of responsible AI development and deployment. By proactively defining audit scopes, engaging stakeholders, employing rigorous methodologies, and committing to continuous monitoring and improvement, organizations can navigate the complexities of LLMs and work towards building more trustworthy and beneficial AI systems. The evolution of more sophisticated audit tooling and the establishment of clearer industry standards will further solidify the practical implementation and effectiveness of these vital assessments.

CONCLUSION

AI audits are becoming an indispensable tool for managing the complexity of AI risk management in a standardized way. As AI systems are increasingly integrated into various sectors, the need for responsible and compliant use is more important than ever. AI audits provide a structured approach to evaluating AI systems and ensure compliance with legal requirements, ethical standards, and technical requirements.

While there are numerous benefits to AI audits, there are also challenges, including the complexity of AI systems, including generative AI technology such as LLMs, a lack of transparency in training data sets, and the need for continuous monitoring of dynamic algorithmic systems. Overcoming these challenges requires robust risk management practices and the adoption of standards such as ISO 42001 combined with security standards such as ISO 27000, which provide a framework for establishing AI management systems and promoting trustworthy AI.

To sump up, AI audits play a critical role in promoting trust, mitigating risk, and ensuring effective and responsible use of AI technologies, especially in business applications. By addressing the challenges and seizing the opportunities of AI audits, organizations can realize the transformative potential of AI while protecting themselves from potential harm.

ACKNOWLEDGMENT

This work has been supported by the *FAIRWork* project (www.fairwork-project.eu) and has been funded within the European Commission's Horizon Europe Programme under contract number 101069499. This paper expresses the opinions of the authors and not necessarily those of the European Commission. The European Commission is not liable for any use that may be made of the information contained in this paper.

REFERENCES

Ashoori, M., & Weisz, J. D. (2019). In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. ArXiv, abs/1912.02675. Retrieved from https://api.semanticscholar.org/CorpusID:208637106.

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? AI and Ethics, 2(3), 405–429. https://doi.org/10.1007/s43681-021-00084-x

Balahur, A., Jenet, A., Hupont Torres, I., Charisi, V., Ganesh, A., Griesinger, C. B., Maurer, P., Mian, L., Salvi, M., Scalzo, S., Soler Garrido, J., Taucer, F., & Tolan, S. (2022). Data quality requirements for inclusive, non-biased and trustworthy AI - Putting science into standards. In JRC Conference and Workshop. https://doi.org/10.2760/365479

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., & Raji, I. D. (2024, April). AI auditing: The broken bus on the road to AI accountability. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 612–643). IEEE.

Blösser, M., & Weihrauch, A. (2024). A consumer perspective of AI certification–the current certification landscape, consumer approval and directions for future research. *European Journal of Marketing*, *58*(2), 441–470.

Bogner, K., Pferschy, U., Unterberger, R., & Zeiner, H. (2018). Optimised scheduling in human–robot collaboration – a use case in the assembly of printed circuit boards. *International Journal of Production Research*, *56*(16), 5522–5540. https://doi.org/10.1080/00207543.2018.1470695

Coston, A., Guha, N., Ouyang, D., Lu, L., Chouldechova, A., & Ho, D. E. (2021, March). Leveraging administrative data for bias audits: Assessing disparate coverage with mobility data for COVID-19 policy. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 173–184).

Faßbender, J. (2021). Particles of a Whole: Design Patterns for Transparent and Auditable AI-Systems. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers* (pp. 272–275). https://doi.org/10.1145/3460418.3479345

Fürntratt, H., Schnabl, P., Krebs, F., Unterberger, R., Zeiner, H. (2024). Towards Higher Abstraction Levels in Quantum Computing. In: Monti, F., *et al.* Service-Oriented Computing – ICSOC 2023 Workshops. ICSOC 2023. Lecture Notes in Computer Science, vol. 14518. Springer, Singapore. https://doi.org/10.1007/978-981-97-0989-2_13

Leocádio, D., Malheiro, L., & Reis, J. (2024). Artificial Intelligence in Auditing: A Conceptual Framework for Auditing Practices. *Administrative Sciences*, *14*(10), 238. https://doi.org/10.3390/admsci14100238

Lee, M. S. A., & Singh, J. (2021). Risk identification questionnaire for detecting unintended bias in the machine learning development lifecycle. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 704–714). https://doi.org/10.1145/3461702.3462572

Li, Y., Goel, S. Making It Possible for the Auditing of AI: A Systematic Review of AI Audits and AI Auditability. *Inf Syst Front* (2024). https://doi.org/10.1007/s10796-024-10508-8

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. *AI and Ethics*, *4*(4), 1085–1115.

Paletta, L., Zeiner, H., Schneeberger, M., Pszeida, M., Mosbacher, J. A., & Tschuden, J. (2024, September). Resilience Scores for Decision Support Using Wearable Biosignal Data with Requirements on Fair and Transparent AI. In *2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)* (pp. 1–4). IEEE.

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider oversight: Designing a third party audit ecosystem for ai governance. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (pp. 557–571). https:// doi. org/ 10. 1145/ 35140 94.3534181

Wassie, F. A., & Lakatos, L. P. (2024). Artificial intelligence and the future of the internal audit function. *Humanities and Social Sciences Communications*, *11*(1), 1–13.

Woitsch, R., Muck, C., Utz, W., & Zeiner, H. (2023, September). Towards a democratic AI-based decision support system to improve decision making in complex ecosystems. Joint Proceedings of the BIR 2023 Workshops and Doctoral Consortium co-located with 22nd International Conference on Perspectives in Business Informatics Research (BIR 2023). CEUR Workshop Proceedings. Retrieved from https://ceur-ws.org/Vol-3514/paper94.pdf.

Schuett, J. (2024). Risk management in the artificial intelligence act. *European Journal of Risk Regulation*, *15*(2), 367–385.

Zeiner, H., Weiss, W., Unterberger, R., Maurer, D., Jöbstl, R. (2019). Time-Aware Knowledge Graphs for Decision Making in the Building Industry. In: Freitas, P., Dargam, F., Moreno, J. (eds) Decision Support Systems IX: Main Developments and Future Trends. EmC-ICDSST 2019. Lecture Notes in Business Information Processing, vol 348. Springer, Cham. https://doi.org/10.1007/978-3-030-18819-1_5