JeDEM 17 (3): 106-140, 2025
ISSN 2075-9517
http://www.jedem.org
Date of submission: 16.04.2025
Date of acceptance: 29.08.2025
Article DOI: 10.29379/jedem.v17i3.1049

Development of effective methods and tools for the auditing of AI algorithms by the Supreme Audit Institutions

Dirk Brand

ORCID Nr: 0000-0002-3660-5015
Independent legal consultant, special counsel at Swart Law, and an Extraordinary Senior Lecturer at the School of Public Leadership, Stellenbosch University, dirkjbrand1@gmail.com

McElory Hoffmann

ORCID Nr 0000-0002-8735-6282
CEO and co-founder of Praelexis and extraordinary senior lecturer in Computer Science at Stellenbosch University, mcelory@praelexis.com

Johan Van der Merwe

ORCID Nr 0000-0002-9191-8727
Data Science Strategist at Praelexis, johan@praelexis.com

Abstract: This article proposes an AI Audit Framework for Supreme Audit Institutions, focusing on public sector usage. It addresses the need for transparency, fairness, accountability, and alignment with ethical and legal requirements. The authors discuss the rise of AI, particularly generative AI and large language models, underscore the evolving regulatory environment, and identify a gap in existing AI audit processes. The article draws on international standards and best practices to offer a methodology for auditing AI algorithms across their entire lifecycle, including risk categorization, data governance, and bias assessment. It also details how generative AI poses new challenges that require specialized guidelines. Recommendations highlight interdisciplinary collaboration and continuous skill development to ensure responsible AI governance.

Keywords: AI auditing, Generative AI, Public sector, Risk, Transparency

1. Introduction

In the Goethe poem "The Sorcerer's Apprentice," an old sorcerer leaves his apprentice at the workshop with chores to do. Tired of fetching water pails, the apprentice then enchants a broom to do the work for him, using magic that he does not yet fully master. With the floor soon awash with water, the apprentice realises he does not know how to stop the broom. He split the broom in two with an axe, but that only resulted in two brooms fetching water, and a completely flooded workshop. It takes the sorcerer on his return to revoke the spell, warning the apprentice that "only a master should invoke powerful spirits". To take the analogy to AI algorithms, only those who know the nature of the technology can understand and harness its power. Otherwise, it can lead to unfathomable destruction.

In view of the increased application of AI in all economic sectors and also in the public sector, and the unfolding AI regulatory landscape in many countries, various questions about fairness, transparency, and accountability are raised. How could AI systems be assessed appropriately, taking into account these and other ethical concerns? Setting standards and stipulating requirements, for example, for high-risk AI, creates some normative frameworks for the development and deployment of AI systems. While this is crucial, it is evident that this is not sufficient, and that AI algorithms should also be audited. Supreme Audit Institutions (SAIs) must also consider their role in auditing AI algorithms used in the public sector. SAIs are the official oversight institutions responsible for auditing the use of public funds by a government, and they could also be responsible for auditing AI systems (https://sirc.idi.no/about/what-are-sais). The rapid and significant increase in the use of large language models (LLMs) not only creates new applications with interesting benefits to users, but it also increases the level of risk in view of the nature of LLMs as generative AI models. It is therefore necessary to pay specific attention to the development of appropriate guidelines for the auditing of LLMs.

Harnessing the benefits of AI to the benefit of society at large, for example, in education or healthcare, implies a responsible approach to the development and deployment of AI that takes into account various ethical considerations such as fairness, transparency, and accountability (Kulal et al, 2024). These ethical considerations also play a role in the auditing of AI systems, whether they are part of an ethical AI framework (soft law) or included in legislation (hard law). Mökander describes AI auditing as a governance mechanism applied to the design and use of AI systems to assess performance as well as alignment with policies and legislation (Mŏkander, 2023). The baseline to assess an AI system can include technical, legal, and ethical requirements, although the lines between these requirements are often blurred. The nature of AI algorithms creates challenges in designing an audit framework that would enable an audit team to assess the performance and legal compliance of AI systems. Ethical issues such as fairness and transparency are not easy to describe in an AI audit framework. For example, transparency could include documentation to inform users about the AI, but it could also mean a degree of explainability that provides some understanding of the logic of the AI system. The interplay between technical, legal, and ethical requirements adds to the complexity of designing an appropriate AI audit framework that can assure developers, deployers, and users of AI. Furthermore, this approach should focus on gaining acceptance from AI system developers by highlighting its benefits to their work. The auditing process should not be treated as

a procedural formality to be only carried out "after the fact," raising red flags and leading to reactive, punitive measures. Instead, an "ethics by design" approach should position AI auditing as a catalyst for innovative and meaningful development practices. This means that the AI audit framework can also be used as a guide for the ethical design and deployment of AI systems. This perspective can encourage developers to adopt an "AI for good" mindset, ensuring that ethical considerations are integrated seamlessly into their work.

An important reason to do an AI audit is to ensure trust in the technology. A variety of actors from industry, academia, and the regulatory environment developed different approaches to AI auditing. In a study done by the International Panel on the Environment (IPIE), the authors argue that due to the global impact of AI, there is a need to develop global standards for AI audits that can build trust in the quality and rigour of different audits (IPIE, 2024).

The current auditing standards and practices applied by SAIs were developed without having AI in mind. However, AI is increasingly used in the audit environment, and the auditing of AI systems used by public sector organisations will soon become a necessity. There is thus a gap which this research could help to fill. The first comprehensive AI legislation, which also includes clear references to adherence to international standards, namely the EU AI Act, was only finally approved early in 2024. It is thus the right time to focus on the development of AI audit methods and tools to audit AI systems. It is argued that sufficient justification therefore exists to develop effective methods and tools for the auditing of AI algorithms by SAIs. The product of this research should be appropriate to be used by any SAI, although some contextual adjustments might have to be considered in view of the applicable legal frameworks.

The following research question will be answered in this study, namely:

"How can SAIs effectively audit AI algorithms, verifying the integrity, fairness, and transparency of the results produced by these systems, through the study of databases and AI techniques employed?"

2. Problem statement

Auditing AI systems is a relatively new field that provides a form of assurance in assessing AI systems, including guidance on risk mitigation measures. There is currently not a unified global approach to AI auditing, but various institutions, both private and public, have published different approaches to AI auditing. While various regulations propose AI audits or assessments, there is a lack of standardised audit procedures. The EU AI Act, for example, stipulates that high-risk AI systems must undergo a conformity assessment before they can be placed on the EU market or put into operation (Art. 43, EU AI Act). Software certification is another way of assuring that an AI system meets specific legal requirements.

Some countries, such as Canada and Australia, have adopted rules for the use of AI in government, which include some form of assessment. Government institutions and administrations serve the public, and they function within highly regulated environments. Accountability requirements imply that public institutions must be audited regularly. While such audits are mostly assessing legal compliance or measuring performance, there are no standard AI audit requirements for the public sector. There is thus a need to assess and verify the integrity, robustness, transparency, and fairness of AI systems used in the public sector. This research study explores the international AI auditing landscape and provides a proposed AI Audit Framework for general AI systems as well as for generative AI, in particular LLMs, used in the public sector.

3. Literature review

The subject of AI auditing received attention during the last few years, both from an academic perspective as well as from the practice, which includes some independent institutions that play a role in AI governance.

From an institutional perspective, there are a few useful publications that give guidance on how to conduct AI audits. The Information Commissioner (ICO) in the United Kingdom published A Guide to ICO Audit - Artificial Intelligence Audits, in which it follows a risk-based and targeted approach to AI audits (ICO, 2021). It provides a very broad outline of the practical steps in such an audit process, but it lacks detail about how auditing of AI systems should be done. An important matter in this Guide is the need for good documentation on the AI systems used by an organisation and which should be a key source of information for the audit. The Digital Regulation Cooperation Forum, of which the ICO is a member, emphasised the need for AI audits to assess legal compliance (DRCF, 2022). It further suggests that AI audits could provide a degree of transparency and contribute to trust amongst a wide group of stakeholders.

The Dutch Court of Audit (Algemeen Rekenkamer) developed a practical approach to AI auditing in the public sector, namely Toetsingskader Algoritmes v2.0 (2024). It includes five focus areas, although ethics are linked to the other four, namely

- Governance and accountability
- Model and data
- Privacy
- IT general controls
- Ethics.

Under Ethics, the following ethical principles are assessed, namely human autonomy, fairness, prevention of harm, and transparency and explainability. Once the scope of an audit is agreed, the audit team will use these focus areas in their assessment to determine if risk is adequately identified and mitigated, and if the AI system complies with any quality criteria. The Algemeen Rekenkamer applied this approach in auditing AI systems used by the Dutch Government. It is a very detailed AI audit guide that is aimed at assessing risk management and compliance. While ethical principles play a central role in the assessment, it is not comprehensive and does not, for example, deal with the impact of AI on human rights in general. It also combines 'model and data', which deals with a

variety of issues, including bias, as one focus area. It would be more appropriate to separately focus on issues of bias and fairness, and robustness and safety of the AI system.

Setting national and international standards for AI provides a measurement tool when conducting AI audits. In the US, the National Institute of Standards and Technology (NIST) published an AI Risk Management Framework with detailed provisions on AI risk identification and measurement (NIST, 2023). It provides an assessment guide that applies to the whole AI lifecycle, including the mapping and measurement of risks, and which could be described as a form of audit. A core focus of this AI Risk Management Framework is to ensure the development and use of trustworthy AI.

The European Data Protection Board (EDPB) issued guidelines for AI audits that focus on compliance as well as the impact of an AI system (Clavel, 2023). It describes the scope of an AI audit as follows:

"An end-to-end, socio-technical algorithmic audit should inspect a system in the actual implementation, processing activity, and running context, looking at the specific data used and the data subjects impacted. It is an end-to-end approach because it recognizes that algorithmic systems work with data produced by complex and imperfect individuals and societies, and operate and intervene in complex social and organisational contexts."

Such an AI audit should thus include the whole AI lifecycle. It makes use of model cards as the source of documented information about the AI system, its training and testing data, risk management, and its impact. These AI audit guidelines focus on transparency and bias as key factors in the assessment process, as well as on the impact of the AI. It also includes a detailed checklist, which is a useful practical instrument for an AI auditor. This AI audit approach is structured with some reference to the EU AI Act. It is not specifically designed for AI audits in the public sector, and thus has limitations that make it unsuitable to replicate as a model. While the general approach and methodology are similar in both private sector and public sector audits, there could be additional requirements in the case of public sector audits, such as AI procurement regulations and specific rules that apply to the use of AI in government.

In the public sector, it is important to distinguish between AI systems procured from suppliers and AI systems developed internally in a public sector organisation. This is important due to higher levels of accountability and transparency that apply in the public sector compared to the private sector, where profit is a primary goal and not serving the public interest. Hickok argues for the development of dedicated AI procurement guidelines for the public sector to address various concerns, such as fairness, transparency, and the lack of capabilities in the public sector to effectively deal with the complexities of deploying AI systems in the public sector (Hickok, 2024). In a recent study by the Ada Lovelace Institute, it also argued that clear guidelines for public sector procurement of AI systems are urgently needed, in view of the impact of procurement decisions on the delivery of public services, and the need to build public trust and ensure public benefit (Ada Lovelace Institute, 2024). There are currently no common international guidelines on the procurement of AI systems in the public sector, but the European Commission published discussion documents on

standard contractual clauses for the procurement of AI systems in the public sector (European Commission, 2024b). Some issues that are of interest to public sector organisations when procuring AI systems are, for example, adequate technical documentation about the AI system, determination of access to the public sector data sets as well as the supplier data sets, and monitoring of the AI system once it is deployed.

According to the G7 Toolkit for AI in Public Services, none of the G7 members have adopted AI audit guidelines or regulations (OECD/UNESCO, 2024). However, at a local level, the City of New York approved the AI Audit Law (NYC Local Law 144, 2021). It has a limited application and only applies to AI used in the context of employment processes. This Law requires such AI systems to undergo an independent bias audit before they can be used.

Providing assurance is an important goal in the audit process. Although it is not described as AI audit guidelines or regulation, the assurance framework published by the UK Government is an important step that provides a solid foundation for AI audit guidelines (UK, 2024). It includes the following tools as AI assurance mechanisms:

- Risk assessment
- AI impact assessment
- Bias audit
- Compliance audit
- Conformity assessment
- Formal verification.

This report suggests that various assurance techniques should be used in combination across the AI lifecycle to provide the best results. There should also be flexibility in the selection of assurance techniques, with low-risk AI requiring fewer assurance techniques, while high-risk AI needing a more comprehensive approach. Any audit requires a yardstick against which performance is measured. The UK AI assurance framework confirms the importance of standards as a baseline in AI assurance processes, for example, international standards developed by the International Standards Organisation (ISO). In the public sector, these assurance techniques should apply to the in-house development of AI systems, as well as AI systems deployed in an organisation. This assurance framework further suggests that the focus is on data, AI models, and systems, as well as AI governance, which includes key elements such as quality assurance and transparency requirements. This UK AI assurance framework could potentially serve as guidelines for public sector institutions to ensure the responsible development and deployment of AI.

From the academic and practice environment, there are also various recommendations about AI auditing. Mökander and Axente (2021) proposed an ethics-based auditing of AI systems, which could create a culture of trust and collaboration. They argued for the development of standards for the evaluation of AI systems, and an independent body to authorise organisations to do such ethics-based auditing.

In a publication titled Adversarial Algorithmic Auditing Guide, it is argued that two types of AI audits could be considered, namely an internal algorithmic audit conducted by independent auditors in cooperation with the developers of an AI system and adversarial algorithmic audits conducted by an independent third party assessing the impact and functionality of an AI system (Eticas, 2024). In view of the socio-technical nature of AI, this Auditing Guide proposed an approach that includes a contextual analysis and stakeholder mapping, and evaluation of bias, inefficiencies, and impact of the AI system.

Lam et al (2024) proposed a framework for AI audits for compliance and assurance purposes. They argued that external compliance and assurance audits provided by independent auditors should provide assurance that the AI systems are designed, developed, deployed, and governed in a responsible and trustworthy manner. They determined four key features for AI audits, namely:

- The audit must use a standard set of audit criteria that is publicly available.
- The objective is to measure compliance or to provide assurance against a normative framework, e.g., the EU AI Act.
- Professional standards are important, and it is therefore necessary that AI auditors are properly trained and accredited.
- The audit results should be published, at least in a limited format, to strengthen transparency and accountability.

They then recommend that the following procedures, which resemble financial audit procedures, should guide the AI audit, namely:

- Determine the scope of the audit.
- Obtain appropriate documentation about the AI system from the auditee and check against the requirements.
- Verification of evidence. During this step, various ways of assessing the correctness of the evidence could be used, e.g., interviews with key role players, reviewing the bias testing of the AI system, and reviewing the data protection impact assessment.
- Publication of the audit report, which should include the audit findings in a standardised format, and the opinion of the auditor
- Certification - a certificate is issued to indicate how the AI system measures against the specific regulation.

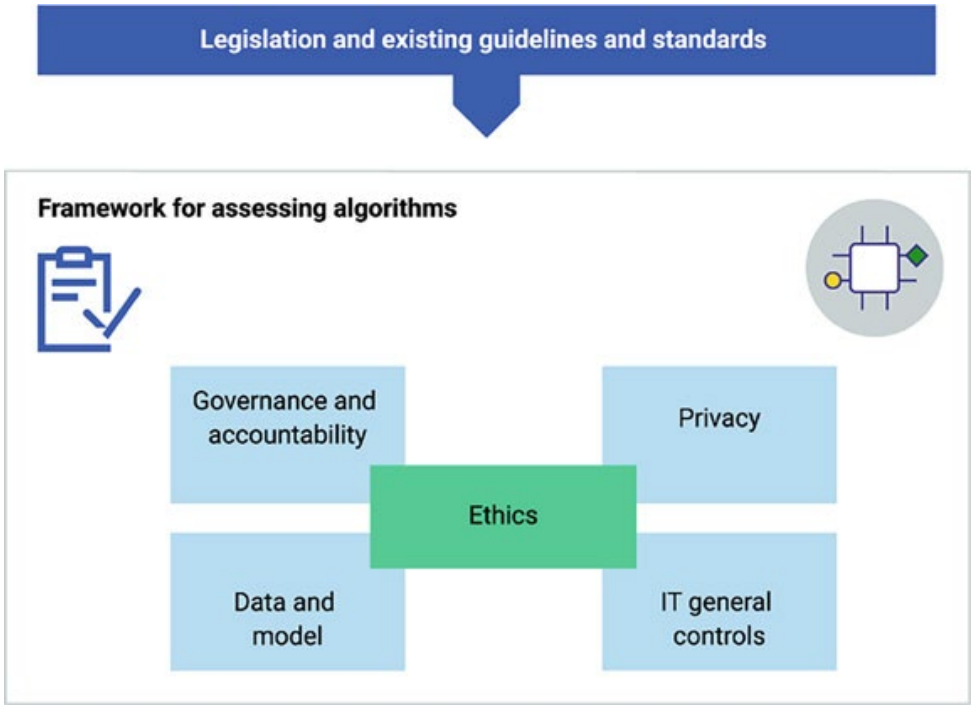
Floridi et al provided a practical guide to ethical AI auditing in capAI (2022). The authors argued that ethics-based auditing is necessary in view of the many ethical failures in the development of AI systems, such as privacy intrusion, bias, and lack of transparency or explainability. They proposed an assessment framework that covers the whole AI lifecycle and is focused on key ethical principles, such as the protection of privacy, fairness, transparency, and ensuring good AI governance.

A useful practitioner's approach to AI auditing is provided by Koshyiama et al (2022). They focus on key ethical principles, namely transparency (explainability), performance and robustness, fairness, and privacy, and apply them to the AI lifecycle when conducting an AI audit. An important part of their AI audit process is the identification of risks and risk mitigation strategies. The audit

report should clearly show if the AI system complies with regulatory, governance, and ethical standards.

In the book Advanced Digital Auditing, the AI audit framework, as illustrated in Figure 1, is provided.

Figure 1: Audit framework of Berghout et al (2023).



It is not specifically aimed at public sector AI audits, but rather AI audits in general.

Under the ethical perspective, it includes human autonomy, safety and data protection, fairness, explainability, and transparency. This framework for AI auditing is quite similar to the approach followed by the Dutch Court of Audit (see above).

In a research paper by the Fraunhofer Institut für Intelligente Analyse und Informationssysteme (IAIS), it is argued that AI audits should be used as part of the development of AI systems and of good quality control (Könsgen et al, eds., 2023). The focus is placed on assessing the trustworthiness of an AI system, thus looking at the following key features, namely: human autonomy, fairness, transparency, robustness, safety and security, and data protection. This paper suggests a three-level assurance model, namely:

- Level 1 - a document-based assessment of the AI system to assess the quality control, focusing on the 6 key features above.
- Level 2 - In addition to the assurance provided in Level 1, a qualified auditor should then do some technical tests on the AI system, using specific software tools to enhance efficiency and accuracy.

- Level 3 - this is the highest level of assurance, and is based on the results of the previous two levels plus argumentation about effective risk management that is applied to minimise the identified risks.

The Fraunhofer Institute has developed specific software tools that can assist in AI audits, e.g., AIBench (a benchmarking tool) and ScrutinAI (a visual analytics tool). The paper recommends the establishment of an audit platform that includes AI standards as well as a variety of software tools that could be used by AI auditors.

In a study done by the International Panel on the Information Environment, various approaches to AI auditing are reviewed, and the report concludes with the following findings:

- A trustworthy audit ecosystem that includes internal, external, and community auditors is needed to strengthen public trust in AI systems.
- Auditors need good access to data and AI artefacts from developers and deployers, such as appropriate technical documentation about the AI system, risk management, and impact assessments.
- Due to the global impact of AI systems, audit regimes must be inclusive and relevant in a global context. In this regard, it is important to develop international standards that can build trust in the quality and rigour of audits that are used in different jurisdictions around the world (IPIE, 2024).

Most of the literature focuses on aspects of auditing of general AI models and systems. Very little attention is yet paid to auditing of generative AI, but it is an important development that warrants attention. This will be discussed later in the paper. The academic literature provides proposals about specific approaches or elements of an AI audit, but it does not establish a comprehensive audit framework.

This literature review provided insights into existing AI auditing practices that could contribute to designing an AI audit framework and methodology for AI in the public sector. It also identified gaps, such as the lack of an internationally accepted AI audit framework and no detailed methodology for AI auditing in the public sector.

4. The AI algorithm development lifecycle

4.1. Algorithms as AI agents

In Artificial Intelligence: A Modern Approach, the authors define an agent as "anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators." (Russel & Norvig, 2016) They proceed by defining a "rational agent": "For each possible percept sequence, a rational agent should select an action that is expected to maximise its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has." (Russel & Norvig, 2016).

Russel and Norvig subsequently distinguish between five kinds of agents, in increasing order of capability and intelligence, namely simple reflex agents, model-based reflex agents, goal-based agents, utility-based agents, and learning agents. A simple reflex agent operates on a single if-then basis (for example, a thermostat), whereas learning agents can set goals within an unknown environment, learning how to optimally fulfil the goals given the models it generates to make the unknown known, by prescribing the actions that will most likely achieve the set goals.

This lays the foundation for seeing AI algorithms as "intelligent agents", with the characteristics of an agent being autonomy, adaptability, and interactivity. Because of the increasing use of the term "agent" for an LLM-based chatbot interface, we also use the more general term AI system, AI solution, or AI algorithm. The agency of these systems is to be understood as the sensing, comprehending, and acting of these systems, which leads to a certain autonomy, adaptability, and interactivity with other systems or "agents".

Autonomy can be defined as the capacity of an agent to act independently and to make its own free choices (Dignum, 2019; HLEG AI, 2020). In AI, even the simplest system, given a well-defined context and parameters for interactivity, can perform autonomously. However, it is not an emergent property, but something designed into the system. In this view, an agent is an "encapsulated computer system that is situated in some environment and is capable of flexible, responsive, and proactive action in that environment in order to meet its goals" (Dignum, 2019). This implies that the system is independent from its environment, and able to interact with its environment and other agents (the "multi-agent system") via sensors and other "actuators". Interaction with its environment can be reactive (according to its perception of its environment) or proactive (intervening according to its goals to influence its environment). Note that autonomy does not imply intelligence. It merely refers to the nature of the system to choose between options, either tasks to be performed ("task autonomy") or which goals to pursue ("goal autonomy").

Adaptability refers to the capability of learning from one's own experiences, sensations, and interactions in order to be able to react flexibly to changes in the environment (Dignum, 2019). The flexibility is reflected in the concept Machine "Learning", as described in its three main classes of application (supervised, unsupervised, and reinforcement learning). Importantly, the learning is always dependent upon the quality of the data; in other words, to predict the future, the future should already somehow be contained in the data. If the data itself contains bias or prejudice, or if the models are trained on biased training sets, there is no way for the system not using bias mitigation strategies to realize this, and its conclusions, recommendations, or prescribed interventions may also be prejudiced or simply wrong. Hence, the importance of ethical considerations.

Interactivity describes the capability of an agent to perceive and interact with other agents, be they human or artificial, understanding that these are agents themselves, with their own goals and capabilities (Dignum, 2019). Combining machine and human intelligence enables the achievement of goals that would not have been possible by actions undertaken by either humans or machines on their own. Augmenting human capabilities instead of replacing human involvement should be the intention of most AI and ML systems, thereby harnessing the power of the collaboration between human and machine. This is particularly true in terms of the monitoring and governance of system

performance and efficacy. These mechanisms include human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC).1

Autonomy, adaptability, and interactivity are the characteristics of AI algorithms that call for ethical engagement and proper auditing to guide the agency of AI algorithms towards beneficial outcomes that are morally justifiable.

4.2. Algorithms as responsible agents

Responsible, auditable AI aims to channel these attributes in ethically sound and legally compliant directions, doing so by defining responsibility, transparency, and accountability.

4.2.1. Autonomous systems require responsibility

Although AI systems can make decisions, select the best next actions, et cetera, human responsibility cannot be replaced. Even if a system is designed for accountability and transparency, it's still an artificial construct created by human intervention. The purpose of the system is predetermined by the human designer. While the system may be able to adapt and learn from its environment, it ultimately performs according to the purpose set by the human creator. The options for responsibility really come down to (i) the machine acts as intended and therefore the responsibility lies with the user, as is the case with any other tool; or (ii) the machine acts unexpectedly due to error or malfunction, in which case the developers and manufacturers are liable (Dignum, 2019). Actions of a system because of learning cannot remove the liability from its designers and developers, as it remains a consequence of the algorithms they designed. This is why the design and implementation processes need to be closely monitored to ensure that the system is behaving according to the relevant ethical and societal principles.

4.2.2. Adaptable systems require transparency

Transparency can pertain to the inner workings of an AI system—such as identifying biases in the data, addressing data quality issues, assessing model overfitting, or understanding the weighting of features. It can also extend to the broader governance of AI systems and data, encompassing the monitoring of system reliability, functionality, and the integrity of underlying data ingestion processes.

AI systems learn from data and are trained on subsets of data in order to be able to interpret new data and predict new values. AI systems also need to be transparent in the sense of being open for

[1] HITL refers to the capability for human intervention in every decision cycle of the system, which in many cases is neither possible nor desirable. HOTL refers to the capability for human intervention during the design cycle of the system and monitoring the system's operation. HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation. (High-Level Expert Group on Artificial Intelligence, 2019, p. 16)

inspection and monitored for bias and drift. The need to deal with bias in data and algorithms and monitoring drift is easier said than done, because data inevitably show patterns and take features into account in decision-making. For example, it cannot be construed as biased if a loan is not granted to someone with a large propensity to default and low affordability. However, it will be wholly unethical to distinguish according to, say, gender or race.

Besides bias, problems with data can include incompleteness, tampered data, and outdated data. Transparency also needs to deal with these complexities and provide openness and control over the whole design, training, productionisation, and monitoring process. Models can malfunction when changes in the underlying relationship between input and output data, called "concept drift" in Machine Learning, are not detected and mitigated. Whether the changes are gradual or sudden, the first challenge is to detect when such drift is occurring. It is therefore critical that robust monitoring of Machine Learning models should be built into the development and deployment of models by data scientists. The second challenge is to address the drift. The best approach remains to periodically update and re-fit the model using a training sample of the most recent historical data. Some algorithms allow for weighing the importance of input data so that more attention is paid to the most recent data. In extreme cases, such as COVID-19, it could even imply that the model must be redeveloped.

4.2.3. Interactive systems require accountability

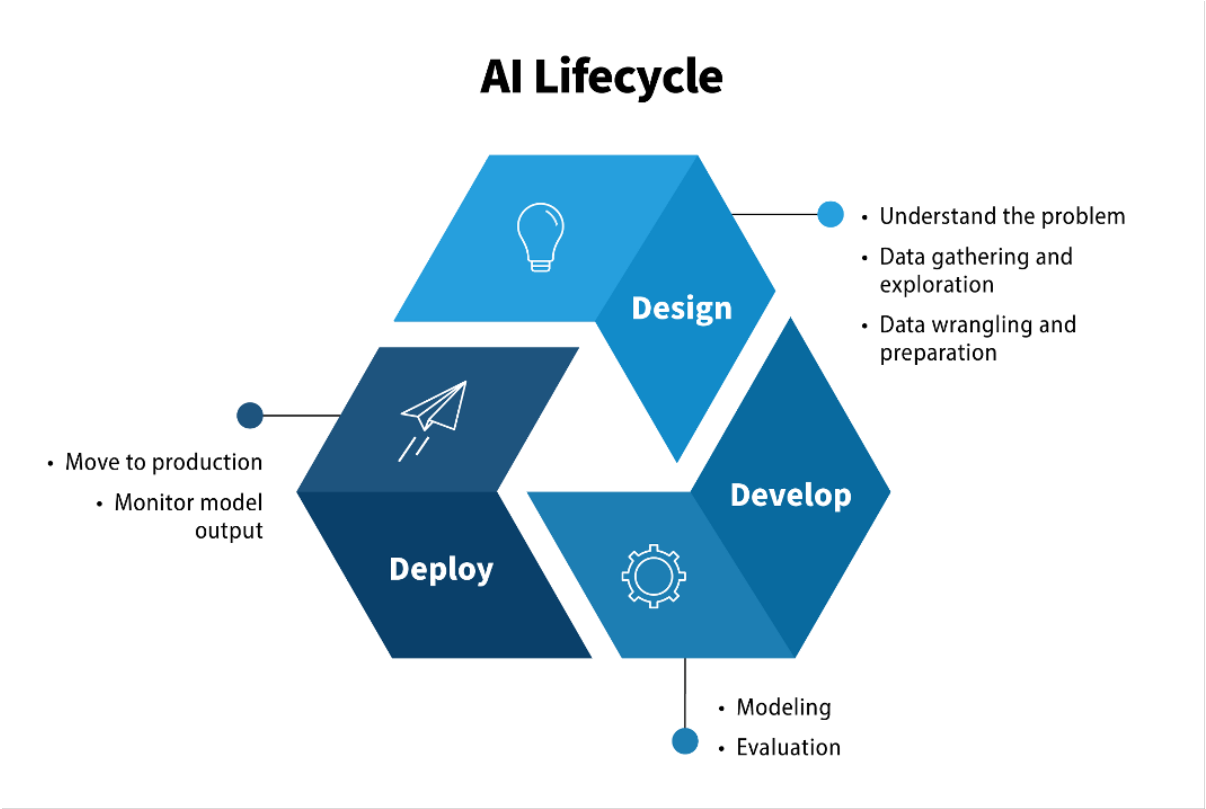
AI systems interact with other systems and users. Accountability is the capability to give an account of those interactions, that is, to report and explain one's actions and decisions. The difference between responsibility and accountability is important. Responsibility refers to the duty to answer for one's role in actions and entails liability. It exists even before the action, and when an action is delegated (to another person or a system), the delegating person remains liable for the consequences of the action. Accountability refers to the ability to explain or report on one's role in events. It is only evident after the action is done. So, while the delegating person remains the "principal" responsible and liable, the agent (which can be an AI system) must be able to report and explain how tasks were executed (Dignum, 2019; Finck, 2019).

Accountability is not only for a system to be able to explain why it took a certain course of action, but also to prove that a safe and sound design process was followed, incorporating the relevant values and ethical concerns (see later on value-based design). Accountability in AI systems, therefore, refers to "explainable AI". Explanation reduces the opaqueness of a system (the so-called "black box"), and when things go wrong, the process can be audited, and logging systems can indicate where errors crept in (similar to the role of the flight recorders in aviation disasters). Accountability in AI systems is particularly important, given the fact that the system lacks moral agency and therefore needs to "explain" how design processes and the results of algorithms incorporate ethical "reasoning".

4.3. The AI development lifecycle

The AI lifecycle (see Figure 2) is a dynamic and iterative process that transforms a business challenge into a functional AI solution. It involves repeated revisits to various stages throughout the design, development, and deployment phases to refine and optimize the solution.

Figure 2: AI lifecycle of IT modernisation centers of excellence, AI Guide for Government, (2024)



4.3.1. Design phase

Defining the Problem: Begin by aligning your team's understanding of the business challenge. Identify the main objectives, specify requirements, and determine the desired business outcomes. Assess whether AI is a suitable tool to address the problem. This stage is crucial for laying the groundwork for success.

Data Collection and Exploration: Gather and evaluate the data necessary for building the solution. This includes identifying relevant data sets, addressing data quality concerns, and gaining an initial understanding of the data.

Data Preparation: Transform raw data into a structured format suitable for AI models. This step, while often labor-intensive and time-consuming, is vital for developing a model that meets the objectives defined earlier.

Clear problem definition and robust data foundations are critical to the success of an AI solution. Without a comprehensive understanding of the data requirements and structure, the model cannot perform effectively.

4.3.2. Development phase

Model Creation: Experiment with the data to identify the most suitable model. The process involves training, testing, evaluating, and fine-tuning multiple models to optimize performance. Achieving the best model often requires iterative adjustments and significant computational resources.

Model Evaluation: Evaluate the selected models on new, unseen data to ensure they perform reliably and align with business objectives. Proper evaluation verifies that the model generalizes well beyond the training data and remains consistent with project goals. Additionally, an audit should confirm that testing data were not used during the training or validation stages, as this would compromise the reliability of the model's performance metrics.

4.3.3. Deployment phase

Production Deployment: Once the model achieves the desired outcomes and meets performance benchmarks, it is implemented in a live production environment. Here, it processes new, real-world data not used during training.

Ongoing Monitoring: Continuously monitor the model’s performance on live data to ensure it remains effective. Over time, models may experience "drift," where changes in the data environment lead to performance deterioration. Regular monitoring and updates are critical to maintaining reliability and effectiveness. This process should include a "human-in-the-loop" approach, incorporating manual quality checks and ensuring that insights from these checks are applied to refine and update the model.

This lifecycle emphasizes the iterative nature of AI development, requiring continuous refinement and adaptation to ensure the solution remains aligned with evolving business needs.

4.4. Input-algorithm-output

It is important to define the scope of auditing AI algorithms. An AI system consists of an input layer, an algorithm layer, and an output layer. As a whole, the system is embedded within an underlying IT infrastructure. While the algorithm development process focuses on creating, training, and refining the AI model, the underlying IT system encompasses all the hardware, software, network, and security components necessary to support the AI's deployment, scalability, security, and performance. For auditing purposes, it is crucial to assess these components to ensure the reliability, security, and compliance of the AI system in its broader operational context. However, the auditing of IT systems falls outside the scope of the auditing of algorithms by Supreme Audit Institutions (SAIs) researched in this paper.

Figure 3: Architecture of a general AI algorithm

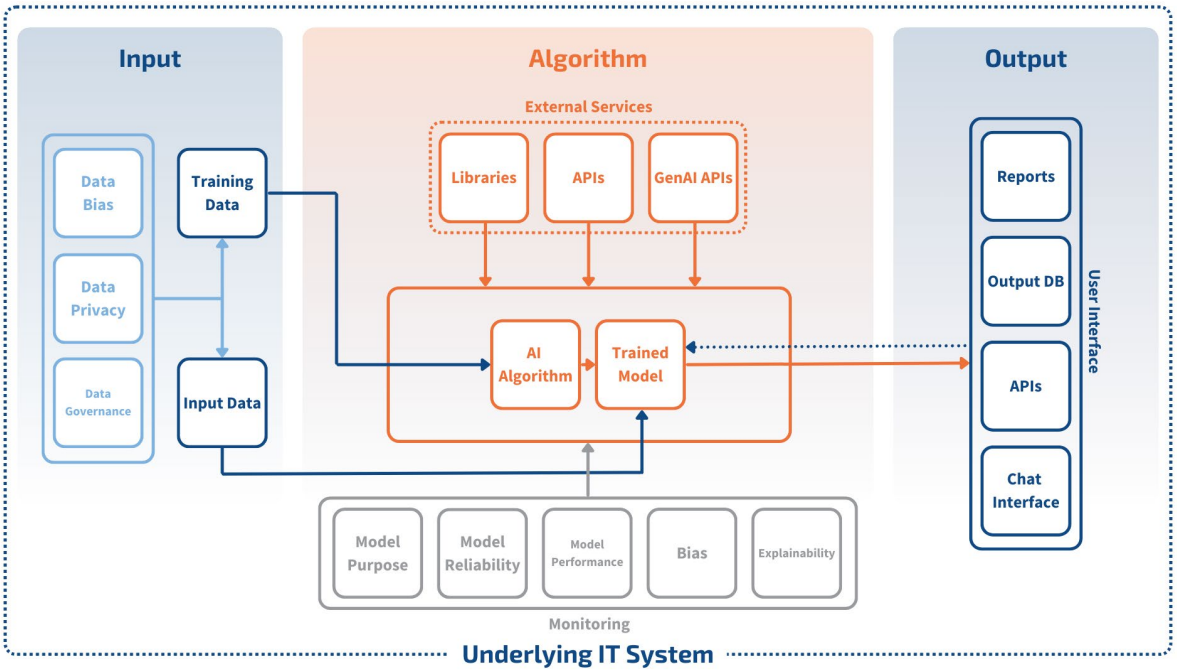


Figure 3 provides a visual image of the architecture of a general AI algorithm. In the case of generative AI, in particular Large Language Models (LLMs), the architecture is different, and the audit process is more complex. The auditing of LLMs is described in detail later in this report. An algorithm ingests data, which must be extracted, transformed, and loaded in a format that ensures appropriate quality and readiness for analytics. This constitutes the input layer of the algorithm. Data is used to "train" the model (the "training data"), and once the model is trained, new data (the "input data") can be used for prediction, classification, or generating other outputs. Both the training data and the input data need to be audited for bias and fairness (ensuring statistical parity towards different stakeholders represented in the data), privacy regulation (e.g., anonymization of data or differential privacy), and proper governance (data ownership and lineage).

The output layer takes the result of the model (the predicted or classified result, often with an attached probability) and presents it in a format that is actionable for the user. Sometimes called the "user interface" (UI), the output layer can take the form of a report, a database, a chat interface, or an API that interacts with another system. From the output, the user validates the result, and this validated output can go back into the trained model, in effect "retraining" the model and acting as a feedback loop that helps the model to "learn".

The algorithm layer comprises mathematical models—such as supervised, unsupervised, reinforcement learning models, large language models, deep learning, and ensemble models—that process data based on an objective function to be optimized, ultimately generating the desired output. Typically, the data is divided into a training set, a validation set, and a test set. Model training involves using the training dataset to compute the optimal parameters that define the AI model. The outcome is the "model" or AI system, which can then be employed to interpret and analyze new inputs.

The development of the algorithm and model often relies on external services. For example, "libraries" refer to pre-written, reusable collections of code, functions, or tools that provide foundational building blocks for implementing and training AI models. These libraries simplify the process of developing AI systems by abstracting complex mathematical computations, algorithms, and optimisations, allowing developers to focus on solving the business problem rather than reinventing the algorithm. Examples include Scikit-Learn, PyTorch, and TensorFlow. The algorithm and model may also access the outputs of external models or "off-the-shelf software" via APIs. Generative AI systems use large language models, developed by third-party providers, which they access via API calls. This is important to recognise because, in the auditing process, there is a distinction between in-house, natively developed algorithms and external algorithms. External algorithms must provide adequate documentation and pre-cleared audits of their reliability and governance.

Model monitoring evaluates the performance and reliability of the model over time. First, it examines the model's appropriateness: Does it address the original business case, and does the selected algorithm type effectively serve the model's purpose? Next, it assesses performance. For instance, in supervised learning, metrics such as the confusion matrix, precision, recall, and F1 score are used to account for false positives and false negatives. Monitoring should answer the question: How reliable is the model in delivering consistent results over time? This includes tracking for model and data drift, where changes in the underlying data relationships may render the model's correlations outdated. Additionally, the model is evaluated for explainability, focusing on the importance of individual features and the predictive value assigned to them. Finally, bias monitoring is critical. Data bias refers to flaws, imbalances, or inequities in the dataset, such as skewed demographic representation in training data. Model bias, on the other hand, arises from design choices, assumptions, or optimization strategies—for example, when a credit model disproportionately penalizes specific groups due to the way it has been constructed or optimized.

For large language models (LLMs), monitoring requires different metrics tailored to their capabilities. Examples include perplexity to measure how well the model predicts a sequence of words, BLEU or ROUGE scores for evaluating text generation quality, and metrics like bias detection frameworks to assess fairness in generated content.

4.5. Technical tools for monitoring

Developers of AI algorithms have a variety of technical tools at their disposal to monitor different aspects of dataset quality and model robustness. While auditors will likely not directly apply these tools, it is important to verify whether the developers have employed these measures to ensure the robustness of their models. Hyperscale cloud providers are particularly involved in offering tools for models running on their cloud platforms.

4.6. The risk-based approach

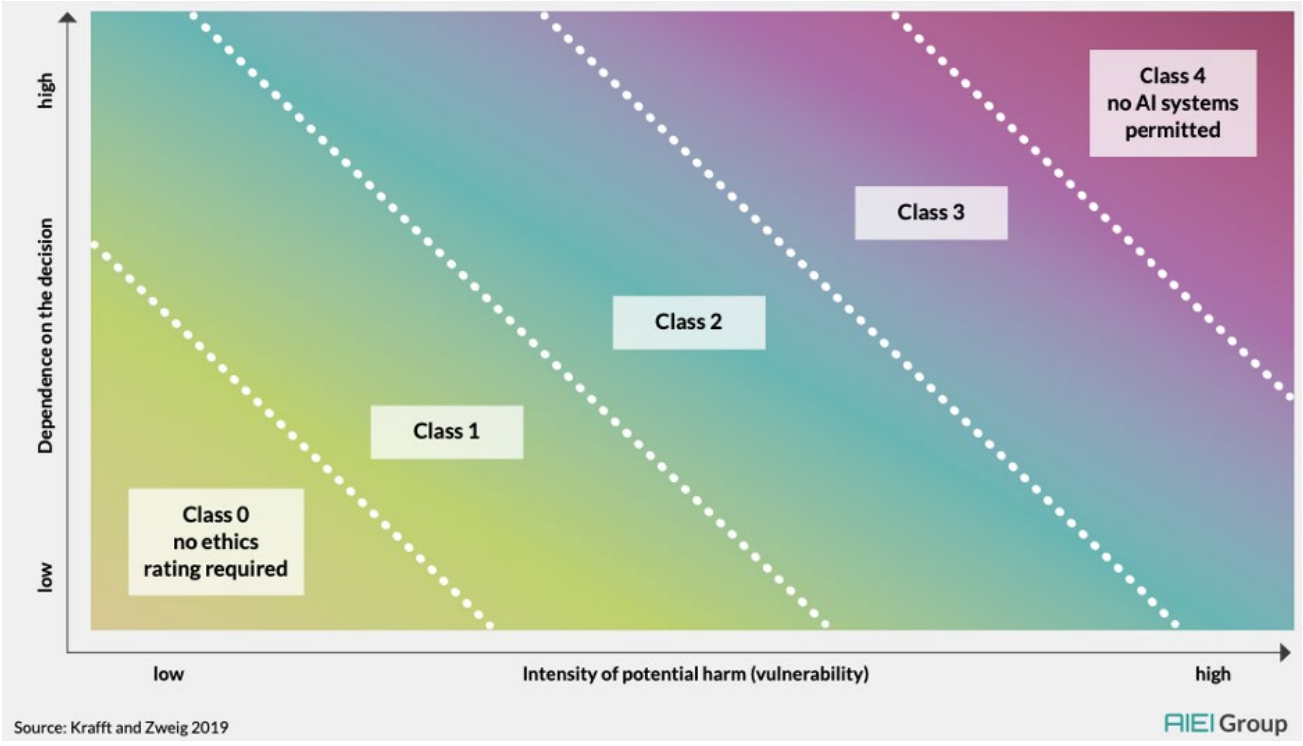
A Risk Matrix (like the EU Act's approach) takes the context into account where an AI system operates to ascertain the risk entailed and determine whether and to what degree it requires regulation or labelling. For example, a recommender system poses less potential harm to an

application where personalized promotions on clothing are presented than where personalized access to medicines or therapies is facilitated. A typical risk matrix (like the one suggested by Hallensleben & Hustedt, 2020, pp. 35-40) has two dimensions.

- The x-axis indicates the "intensity of potential harm" and should be assessed by looking at the impact on several people or access to resources and whether society as a whole is threatened. For example, are fundamental rights, equality, or social justice impacted?
- The y-axis indicates the "dependence on the decision". This refers to whether options exist to avoid the potential harm indicated on the x-axis. For example, the more human intermediaries in the decisions and actions taken by a user, the more control the user has. The ability to change the AI system for another ("switchability") and the possibility to challenge and correct algorithmic decision-making are other factors lowering the dependence on the decision by the AI system.

Figure 4 shows how five different risk classes emerge. Systems not requiring any regulation (or ethical labelling) fall into class 0, while the highest class is reserved for contexts where AI systems should not be applied at all (weapons of mass destruction are an example here). The higher class a system's context inhabits, the more stringent the monitoring and governance of the system should be. For example, in class 3, only algorithmic decisions that are comprehensible and understandable by humans are permitted. The AI Act took a similar approach and prescribed mitigation strategies for each risk class.

Figure 4: Risk matrix with 5 risk classes (Hallensleben & Hustedt, 2020, p. 39)

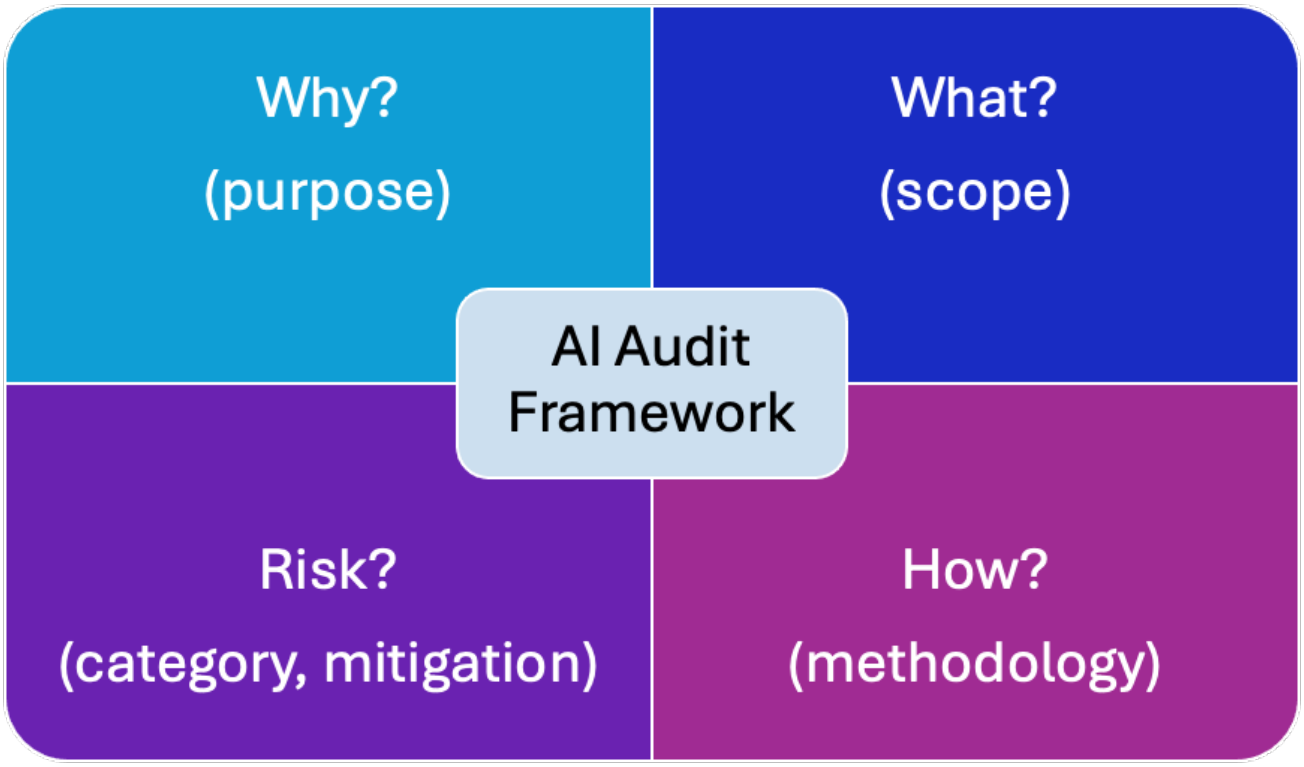


5. Proposed AI audit framework

5.1. Introduction

An AI audit framework must address key elements such as the purpose, scope, methodology, and risk assessment of the audit (Lam et al, 2024; Kyoshima et al, 2022). These elements provide the foundation for evaluating AI systems against standards of transparency, bias, integrity, and robustness. We therefore propose the following AI audit framework, also displayed visually in Figure 5.

Figure 5: AI audit framework



5.1.1. Purpose of the AI audit

Understanding the "why" of the audit is critical. Audits may serve different purposes, such as compliance with legislation (e.g., the EU AI Act), conformity to international standards, or assessing the performance of the AI system. Each type of audit—compliance, conformity, or performance—requires a tailored approach, with a core focus on elements like bias, transparency, and the integrity of the system.

5.1.2. Scope of the AI audit

Once the purpose of the AI audit is determined, it is necessary to define the scope of the AI audit. The auditee should provide appropriate information about what AI systems should be audited, including the nature of the AI. A useful approach is to use an AI user story that defines a need or problem, and describes the AI solution and how it solves the problem.

Key questions that should be considered in determining the scope of the AI audit are:

- Is it an integrated AI system, i.e., integrated into other systems, or is it a stand-alone AI system?
- Is it a procured AI system or developed in-house?
- Is it a general AI system?
- Is it a generative AI model, or a product of generative AI?

The information provided in response to these questions will provide clarity about the scope of the AI audit.

5.1.3. Risk assessment

Risk categorization, as outlined in the EU AI Act or other applicable legislation, plays a pivotal role in AI audits. Identifying the risk category informs the application of specific legal requirements and helps evaluate the system's interaction with other technologies. Auditors must also consider existing risk mitigation measures to ensure comprehensive oversight.

5.1.4. Methodology and tools

Determining the "how" of the audit involves selecting appropriate methodologies and tools. Clear guidelines should address the unique characteristics of various AI systems, including general AI and generative AI models like Large Language Models (LLMs).

It is necessary to clarify whether an AI system deployed in a public sector organisation was procured from a service provider or internally developed when determining the scope of the AI audit. In the absence of any specific rules on AI procurement for the public sector, a Supreme Audit Institution could guide to support the audit process. It is recommended that at least the following matters about procured AI systems be clarified in determining the audit scope:

- Provision of relevant detailed technical documentation by the supplier.
- A description of the levels of accuracy, robustness, and safety of the AI system.
- Determination of access to relevant public sector data sets, as well as access to supplier data sets.
- Indication of the quality management system provided by the supplier.
- Provision of the requirement specifications that the public sector organisation used when it procured the AI system.
- Provision of information about the evaluation and validation done as part of the procurement process.

In the case of AI audits in the public sector, it should be the Supreme Audit Institution that develops the audit guidelines and methodologies. In view of the nature of AI systems and their dependence on data, data protection authorities could also contribute to the design process, as is evident, for example, in the UK.

An important component of an AI audit framework is the issue of standards, e.g., ISO/IEC 42001 on AI systems and governance (https://www.iso.org/standard/81230.html), and regulations. The following are some of the international standards that could be used:

- Lifecycle management ISO 5336
- Bias mitigation ISO 12791
- Risk management ISO 23894
- AI governance ISO 42001.

The field of AI standards is still evolving, and the International Standards Organisation (ISO) and the International Telecommunications Union (ITU) are cooperating to develop international AI standards for global application. The Supreme Audit Institution in a country should indicate which standards, international or domestic, and what legislation apply to the audit of AI models and systems developed and used within the public sector. The specific requirements of applicable legislation, e.g., the EU AI Act or the Directive on Automated Decision-making in Canada (2019), guide the development of the focus and scope of an audit or assessment.

Documentary evidence plays a critical role in financial audits. In the development of AI models and systems, there are relevant technical documents produced by developers. It could also be required by legislation, e.g., the EU AI Act that requires detailed technical documentation (Art. 11, Annex IV, EU AI Act). Although appropriate technical documentation is important and could provide valuable insight into the process of an AI audit, other important elements form part of the overall AI audit framework, such as impact assessments. All the relevant documents that can assist in assessing the performance and compliance of an AI system should be provided to the audit team.

5.2. Auditing general AI systems

Providing assurance runs like a golden thread through the diverse range of literature on AI auditing discussed above. It is an essential part of an audit, including an AI audit, to provide assurance. Responsible AI requires that risk management be applied to the entire life cycle of an AI system. This includes the identification of specific risks as well as the identification and implementation of appropriate risk mitigation measures in a risk management system.

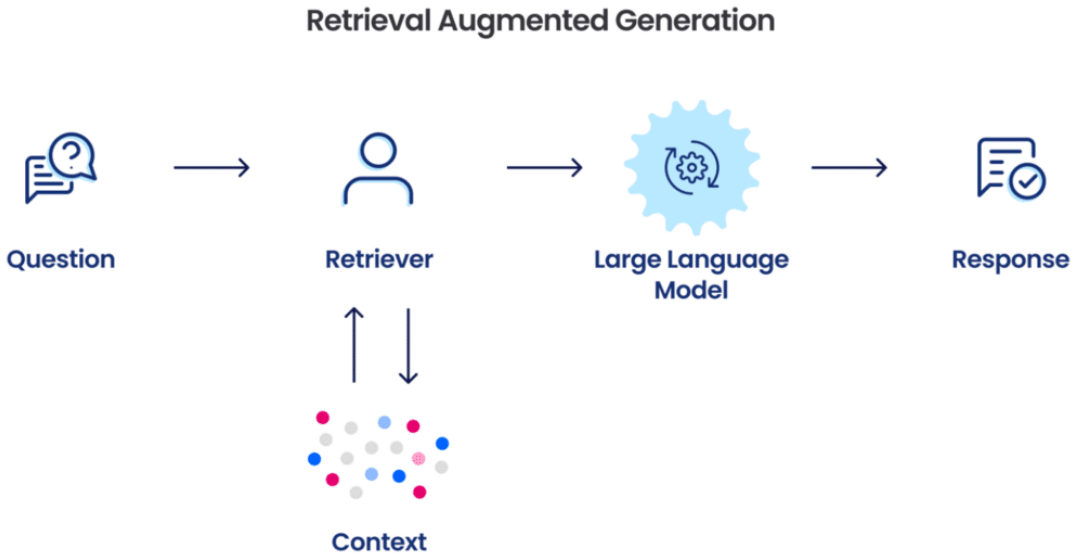
The depth of an AI audit depends on the type of AI system to be audited. In the case of minimal risk AI systems, an audit based on analysis of technical documentation and other relevant documents could be sufficient. In case of high-risk or more complex AI systems, a comprehensive audit that includes model assessment must be done. If technical tests must be done, the model itself must be accessible to the audit team. It is during this phase of the audit process that verification of evidence is done. It includes evaluating the relevant documents provided by the auditee as well as other practical assessment steps, as explained in the AI Audit Guide (Annexure 1), which should be used as the practice guide for an audit.

5.3. Auditing generative AI, specifically LLMs

The nature of generative AI implies that the general approach to auditing AI models and systems would be insufficient in auditing this type of AI, for example, Large Language Models, which is the focus of this section. There are various ethical concerns about bias, potential misleading information, data protection, and intellectual property related to the use of generative AI. Various court cases in different jurisdictions confirm the importance of these concerns. Proper regulation, which is not yet adequately addressed internationally, is needed to guide developers, deployers, and users of generative AI. Since it is already widely used, also in the public sector environment, it is important to develop an audit framework for the audit of Large Language Models. The lack of literature on this topic is because this type of AI is still fairly new and is continuously evolving to develop more abilities. One of the major difficulties is the fact that risk assessment of an LLM cannot effectively be done without sufficient knowledge about the downstream development of an AI system based on that LLM.

In the evolution of LLMs, an important development is the introduction of Retrieval Augmented Generation (RAG), which is a design feature that integrates data retrieval directly in the generation process (Ahmed, 2024). It enhances the quality of the generated output by leveraging bespoke data. Large generative AI models such as ChatGPT have an integrated RAG approach. RAG is in particular useful in the development of domain-specific LLMs or specific AI agents, for example, for a bank or a public sector organisation that works with large volumes of data, where an in-house database is used to retrieve relevant information to enhance the generation process. The domain-specific AI agent is given access to a specific database, which it needs in the RAG process. In a two-step process, the retriever accesses the relevant database to obtain more contextual information that can enhance or augment the prompt. This is used by the LLM to generate a more accurate response. Figure 6 provides a visual display of RAG.

Figure 6: Illustration of Retrieval Augmented Generation (Tran, H, 2024).



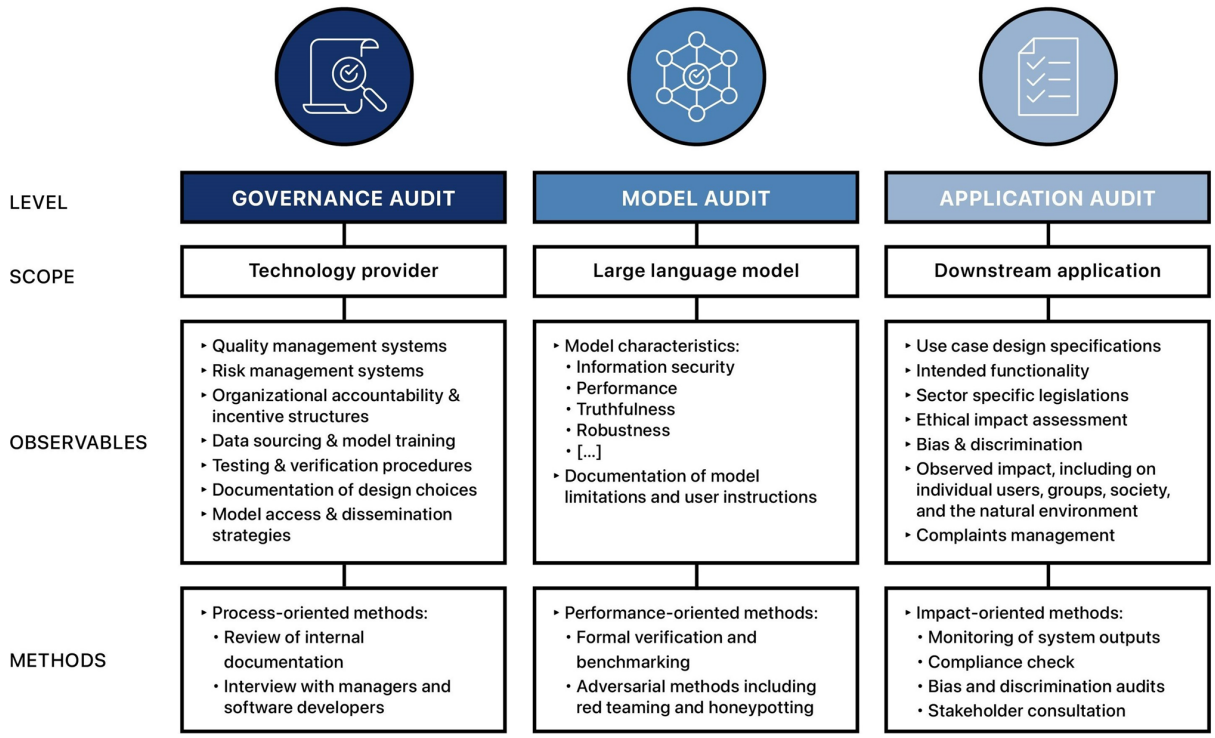
In the context of a Supreme Audit Institution that works with diverse and large databases, it is envisaged that a domain-specific LLM that uses RAG could add value to the work of the SAI by developing an in-house AI agent. An additional step that could be considered is fine-tuning an AI model with domain-specific data and then using a RAG, which should further enhance the quality of the output.

Amirizaniani et al developed a proposal consisting of a two-phase process for auditing LLMs (Amirizaniani et al, 2024). In the first phase, another LLM is used to generate different versions of probes, which are evaluated and validated by a human, an important step to provide assurance. In the second phase, the validated probes are used as input to the audited LLM to generate responses. The variation in the responses indicates the amount of inconsistency. The accuracy and stability of the audited LLM are evidenced by generating similar responses to similar probes,

Another approach to auditing LLMs is provided by Mökander et al (2023), namely a three-layered approach. The first layer consists of governance audits of technology providers that design and disseminate LLMs. The second layer consists of model audits of LLMs after pre-training but prior to their release. In the third layer, the application audits are done of the applications (AI systems) based on the previously assessed LLM. They argue that these three layers complement and inform each other, thus creating a comprehensive assessment of LLMs that will provide assurance. This is an innovative and practical approach, summarised below.

Auditing LLMs is a more complex exercise that requires a unique approach, although some of the features and questions used in the general AI audit could be relevant in this audit process as well. The proposed AI audit framework for LLMs (Mökander et al, 2023) is a useful approach, and we recommend that it be applied to auditing LLMs. We recommend that the audit methodology for general AI (see above) be applied to the model audit and application audit (AI Audit Guide).

Figure 7: Audit framework for LLMs (Mökander et al, 2023)



According to the authors, the governance audit should focus on 3 key issues:

- Reviewing the adequacy of organisational governance structures.
- Creating an audit trail of the LLM development process.
- Mapping roles and responsibilities within organisations that design LLMs to assist in determining accountability for system failures.
- The technology provider of the LLM should provide appropriate documentation that adequately addresses the items listed under 'Observables' above. The audit team could then check it for completeness and proceed to attend to the model audit and application audit.

Model audits are aimed at assessing the capabilities and limitations of the LLM, and are primarily a performance audit.

Most of the features and questions used in the general AI audit process are relevant to and can be used in the 3rd stage of the LLM audit process, namely, during the application audit, for example, a bias assessment. In applying RAG in the LLM, a specific database is used to provide context and enhance the output generation. This would be relevant in the application audit, thus providing an auditor with the possibility to audit the specific database as well.

6. Proposed AI audit process and methodology

Based on the discussion above, an AI audit team should follow these process steps and use this methodology:

1) Determine the scope and purpose of the audit. The type of AI and context in which the AI system functions will have an influence on the scope of the audit.
2) Use the AI Audit Guide (Annexure 1) to assess the AI system based on the listed functional areas applied to the AI lifecycle.
3) Obtain the technical documentation, risk management plan, conformity assessment, and any other relevant documentation in support of the audit.
4) Engage the auditee in the audit process to get responses on all the selected questions in the AI Audit Guide. During this stage, it is important to get input from key stakeholders in the organisation, such as the legal unit, data science office, and risk management office.
5) Assess the risk management and mitigation measures.
6) Develop an AI audit report that includes findings and recommendations, and which should acknowledge the input of the auditee.
7) Present the AI audit report to the auditee.

7. Tools for AI audits

One of the tools used by the Canadian Government is an Algorithm Impact Assessment Tool, which applies to all AI systems used in the Canadian public sector (Canada, 2019). It focuses on risk management and measures impact according to four levels, namely:

1) little to no impact
2) moderate impact
3) high impact
4) very high impact.

The Algorithm Impact Assessment Tool is available as an online questionnaire.

An AI Audit assistant (AI agent) could be a useful tool for auditors to manage large volumes of documents, such as reports, opinions, and policies included in the database of a Supreme Audit Institution. This could add value to the AI audit process in view of its ability to extract relevant information to apply to a current AI audit process.

Generative AI could also be used as a technical support tool in the AI audit process. Various elements of a typical audit process could benefit from employing generative AI as a support tool. Here are some possible use cases:

1) automating documentation and reporting - It can analyse complex datasets, audit logs, and model performance metrics, and explain the outcomes in understandable language, which is particularly useful for communicating technical results to non-experts.
2) identifying and mitigating biases - AI models, especially large language models, can help in identifying biases within training data, algorithms, or predictions. By analysing data distributions, fairness metrics, and model outputs, generative AI can flag areas where biases may exist or where fairness could be improved.
3) explainability - Generative AI can assist in providing interpretability by generating human-readable explanations for model decisions.
4) assessing AI risks - Generative AI can assist in risk assessment by simulating potential attack scenarios or identifying weaknesses in the model.

Against this background, it is recommended that consideration be given to the creative and dedicated use of generative AI, in particular LLMs, within the public audit environment, inter alia by:
1) creating agents that can validate each step of the AI audit process; and
2) creating a chatbot for the Supreme Audit Office that is applied to its database of reports, guidelines, and rules to act as a technical assistant to the AI audit team.

AI agents can act autonomously, but direction from the audit team is necessary to identify specific goals that the AI agent should achieve. Provision should also be made for feedback mechanisms, such as human-in-the-loop, from which the AI agent learns to improve the accuracy of its responses (Gutowska, 2024).

It is important to note that the development of generative AI agents should be confined to the specific public audit environment, in other words, domain-specific. The scope of training data and range of data sources on which it will be applied are thus demarcated. RAG could also be used to enhance the quality of the output. The human auditor, or audit team, will still have the final say, but the effective use of generative AI will enhance their efficiency.

8. Skills requirements

AI auditing is essentially a multi-disciplinary process that should benefit from legal, technical, and ethical expertise (Mökander, 2023). It is therefore important to ensure the inclusion of relevant skills within an audit team that will be responsible for auditing of AI algorithms. These skills include conventional audit skills, technical skills that include specific knowledge of AI algorithms, legal expertise with specific knowledge about AI ethics and law, and perhaps data science expertise.

In the context of an SAI, it is evident that AI literacy, including regular updating of knowledge, is an essential skill requirement for all staff. While general knowledge about the use of AI in auditing

is important, specific expertise should be developed and maintained about the auditing of AI systems. It is recommended that SAIs include a menu of different AI-related training programs for their staff, which includes topics such as AI ethics, exploring machine learning and natural language processing, cybersecurity, and AI auditing.

9. Case study application

The Algorithm Register published by the Government of the Netherlands was used to select a case to be audited. This case study application is qualified as follows. It is not a detailed audit since not all relevant information was available, no direct interaction with the auditee was conducted, and the information in the Algorithm Register is published primarily to inform the public, rather than to satisfy the requirements of a public auditor. It is therefore a desktop exercise to demonstrate the usefulness of the AI Audit Guide.

Informatie Ondersteund Beslissen - Kort Verblijf (Schengen) Visum (KVV)

[Information Supported Decision – Short Stay Schengen Visa]

This AI system is administered by the Dutch Ministry of Foreign Affairs. If the proposed AI Audit Guide for SAIs is applied to audit this AI system, the following high-level assessment is provided.

1) Scope and purpose of the audit.
   The scope is to assess the performance and legal compliance of the selected AI system.
2) Apply the AI Audit Guide to the AI lifecycle.

An overview of the assessment of the following focus areas is provided:

Data

a) The data sources are clearly identified.
b) Data impact assessments are done.
c) There is no information about the training data used in the development of the AI system.
d) No information about data security or, quality of the data is provided.

Privacy and data governance

There is a clear focus on ensuring compliance with the GDPR. By making the visa application, applicants provide their consent for the processing of the data related to the application.

Bias - Information about bias testing is absent.

Technical robustness and safety

There is no information available about testing for robustness. However, the purpose of the AI is clearly described, and it appears to function in accordance thereof. Monitoring of the decision-making on visa applications is done, and that includes the way in which the AI is used.

Transparency

Some documentation is provided, but it is focused on informing the public. So, it does, for example, not include as many technical details as are required in the AI Act's technical documentation.

Human agency and oversight

Appropriate human oversight is provided in the application of the AI system as a support tool.

Fundamental rights impact assessment

While it is indicated that the impact assessments are done, no information about any findings or risk mitigation is provided.

Legal compliance and implementation of standards

The EU AI Act conformity assessment is currently not yet required due to the phasing in of the EU Act, but will in future be required.

3) Obtain all relevant documentation.

The following documents were obtained:

a) Fact sheet Informatie Ondersteund Beslissen – Kort Verblijf Schengen Visa Informatie-ondersteund-beslissen-kort-verblijf-schengen-visum-kvv

b)Algorithm Register information Informatie-ondersteund-beslissen-kort-verblijf-schengen-visum-kvv

4) Engage the auditee (Ministry of Foreign Affairs) in the audit process.

This was not done.

5) Assess the risk management and mitigation measures.

Insufficient information was provided to assess the risk management and mitigation measures.

6) Develop an AI audit report.

Although limited information was available to properly audit the selected AI system, the following observations that would typically be included in an AI audit report are made:

a) Transparency appears to be a high priority. This is reflected in the inclusion of this AI system in the Dutch AI Register and the published information about the aim and functioning of the system.

b) There is a strong foundation of compliance regarding data privacy, including data impact assessments, and the use of the AI system as an aid with human oversight.

c) There are, however, significant gaps in critical areas, notably:

Lack of testing and documentation - insufficient information and documentation on data quality, bias testing, technical robustness, and the AI's inner workings. This raises concerns about the overall reliability, safety, and transparency of the system.

Missing Risk Assessment: A formal risk assessment is missing and needs to be conducted.

Impact Assessment Follow-Up: While Impact Assessments are said to be done, the findings of these impact assessments need to be documented, and the risks addressed.

d) Future Compliance: While the system may be compliant with current regulations, it is not yet prepared for the full implementation of regulations like the EU AI Act.

10. Conclusion and recommendations

The results of this research study are an AI audit framework and methodology for use by SAIs. The literature review provided useful insights into existing research about aspects of AI auditing, and it also clearly confirmed the initial statement that there is a need to develop an appropriate AI audit framework and methodology that could be used by SAIs.

In our analysis, we provided a contextual basis by focusing on the AI algorithm lifecycle. In order to properly construct the AI audit framework and methodology, clarity about a range of related concepts is provided, as well as confirmation of the need to view the whole AI lifecycle. It is concluded that the AI audit framework consists of the following elements or steps:

- Purpose of the audit
- Scope of the audit
- Determination of risk category and mitigation
- Audit methodology

Legislation, AI standards, and ethical principles or codes are important baseline elements for AI audits.

In view of the rapid growth in the use of generative AI in particularly LLMs, also in the auditing environment, it is necessary to pay specific attention to this type of AI. Auditing Large Language Models (LLMs) cannot be done in the same way as general AI systems. We therefore provided a clear distinction and a detailed methodology for both types of AI audits. In the proposed AI methodology, there is a focus on key ethical principles, namely

- Privacy and data protection,
- Bias, fairness, non-discrimination,
- Transparency and explainability, and
- Human oversight.

The proposed AI Audit Framework could also serve an additional purpose, namely as guidance to key stakeholders in an organisation, such as a legal adviser, risk management officer, and chief AI officer, when developing or procuring AI systems.

Based on this study, we make the following recommendations, namely that:

- The AI Audit Framework and methodology should be adopted by SAIs for use in the public sector audit environment.
- AI literacy training and skills development receive high priority in all SAIs.
- SAIs adopt a multi-disciplinary approach to AI auditing.

References

Ada Lovelace Institute, (2024). Buying AI - is the public sector equipped to procure technology in the public interest?, https://www.adalovelaceinstitute.org/project/procurement-ai-local-government/

Ahmed, S. (2024). What is Retrieval-Augmented Generation(RAG) in LLM and How it works?, https://medium.com/@sahin.samia/what-is-retrieval-augmented-generation-rag-in-llm-and-how-it-works-a8c79e35a172. Accesed: 15 April 2025.

Algemeen Rekenkamer (2024). Toetsingskader Algoritmes versie 2.0, www.rekenkamer.nl/onderwerpen/algoritmes. Accesed: 15 April 2025.

Amirizaniani, M., Yao, J., Lavergne, A., Snell Okada, E., Chadha, A., Roosta, T., and Shah, C. (2024). LLM Auditor: A Framework for Auditing Large Language Models Using Human-in-the-Loop, arXiv:2402.09346v3

Berghout, E., Fijneman, R., Hendriks, L., De Boer, M., and Butijn, B-J. (2023). Advanced Digital Auditing, Springer, Cham, Switzerland, https://doi.org/10.1007/978-3-031-11089-4

Canada. (2019). Directive on Automated Decision-making, https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

Clavel, G. G. (2023). AI auditing: checklist for AI auditing, EDPB.

Dignum, V., (2019). Responsible Artificial Intelligence How to Develop and Use AI in a Responsible Way. Cham: Springer.

DRCF. (2022). Auditing Algorithms: the existing landscape, role of regulators and future outlook, https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook

Eticas. (2024). Adversarial Algorithmic Auditing Guide, https://eticas.ai/adversarial-algorithmic-auditing-guide/

European Commission. (2024a). Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, COM(2021) 206 final.

European Commission. (2024b). Draft Proposal for standard contractual clauses for the procurement of Artificial Intelligence (AI) by public organisations (discussion document), https://public-buyers-community.ec.europa.eu/system/files/2023-10/AI_Procurement_Clauses_template_High_Risk%20EN.pdf

Finck, M. (2019). Automated Decision-Making and Administrative Law, Max Planck Institute for Innovation & Competition Research Paper Nr 19-10

Floridi, L, Holweg, M., Taddeo,M., Silva, J.A., Mökander, J., and Wen, Y. (2022). capAI, https://ssrn.com/abstract=4064091

Government of the Kingdom of the Netherlands, The Algorithm Register (2025), https://algoritmes.overheid.nl/en/algoritme/information-supported-decision-shortstay-schengen-visa-cdv-ministry-of-foreign-affairs/94596537. Accesed: 15 April 2025.

Gutowska, A. (2024). 'What are AI agents?', https://www.ibm.com/think/topics/ai-agents

Hallensleben, S. & Hustedt, C., (2020). From Principles to Practice. An interdisciplinary framework to operationalise AI ethics, Gütersloh: Bertelsmann Stiftung.

Headecke, E., Mock, M., Pintz, M., and Poretschkin, M. (2023). KI-Anwendungen systematisch prüfen und absichern, Fraunhofer Institut, Bonn, Germany www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-zertifizierung.html. Accesed: 15 April 2025.

Hickok, M. (2024). 'Public procurement of artificial intelligence systems: new risks and future proofing', AI & SOCIETY (2024) 39:1213–1227.

High-Level Expert Group on Artificial Intelligence, (2019). Ethics Guidelines for Trustworthy AI, Brussels: European Commission.

High-Level Expert Group on Artificial Intelligence (2020) The Assessment List for Trustworthy Artificial Intelligence, Brussels, European Commission.

Information Commissioner's Office, (2021). A Guide to ICO Audit - Artificial Intelligence Audits, https://ico.org.uk/media/for-organisations/documents/4022651/a-guide-to-ai-audits.pdf. Accesed: 15 April 2025.

IPIE. (2024). Global Approaches to Auditing Artificial Intelligence - a Literature Review, SR2024.1. Zurich, Switzerland: IPIE.

IT modernisation Centers of Excellence, AI Guide for Government, (2024), https://coe.gsa.gov/coe/ai-guide-for-government/understanding-managing-ai-lifecycle/. Accesed: 15 April 2025.

Koshyiama, Kazim, E. and Treleaven, P. (2022). Algorithm Auditing: Managing the Legal, Ethical and Technological Risks of Artificial Intelligence, Machine Learning, and Associated Algorithms, Computer, April 2022, https://10.1109/MC.2021.3067225

Könsgen, C., Nordman, M., and Loh, S. (eds). (2023). KI-Anwendungen systematisch prüfen und absichern, Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS, Sankt Augustin, Germany.

Kulal, A, Rahiman, H.U., Suvarna, H., Abishek, N., and Dinesh, S. (2024). Enhancing public service delivery efficiency: Exploring the impact of AI, Journal of Open Innovation: Technology, Market, and Complexity 10 (2024) 100329, https://www.sciencedirect.com/science/article/pii/S2199853124001239

Lam, K., Lange, B., Blili-Hamelin, B., Davidovic, J., Brown, S, and Hasan. A. (2024). A Framework for Assurance Audits of Algorithmic Systems. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil. https://doi.org/10.1145/3630106.3658957

Mökander, J., Schuett, J., Rose Kirk, H., and Floridi, L. (2023). Auditing large language models: a three-layered approach, AI and Ethics, https://doi.org/10.1007/s43681-023-00289-2

Mökander, J. and Axente, M. (2021). Ethics-based auditing of automated decision-making systems: intervention points and policy implications, AI and Society, 27 September 2021, https://doi.org/10.1007/s00146-021-01286-x .

Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches, Digital Society (2023) 2:49 https://doi.org/10.1007/s44206-023-00074-y

National Institute for Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

New York City Council. (2021). Local Law 144 of 2021, https://legistar.council.nyc.gov. Accesed: 15 April 2025.

OECD/UNESCO. (2024). The G7 Toolkit for AI in the Public Sector, Como, Italy.

Russel, S. & Norvig, P., (2016). Artificial Intelligence A Modern Approach. Third Edition ed. Essex: Pearson Education Limited.

Sahin, A. (2024). What is Retrieval Augmented Generation (RAG) in LLM and how it works?, blogpost https://medium.com/@sahin.samia/what-is-retrieval-augmented-generation-rag-in-llm-and-how-it-works-a8c79e35a172. Accesed: 15 April 2025.

SAI Independence Resource Centre (2023) What are Supreme Audit Institutions, https://sirc.idi.no/about/what-are-sais

Tran, H. (2024). Which is better, retrieval augmentation (RAG) or fine-tuning? Both., https://snorkel.ai/which-is-better-retrieval-augmentation-rag-or-fine-tuning-both/

UK Department of Science, Innovation and Technology. (2024). An Introduction to AI Assurance, February 2024, London, https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance

Authors

Dirk Brand

Dirk Brand is an independent legal consultant, special counsel at Swart Law and an Extraordinary Senior Lecturer at the School of Public Leadership, Stellenbosch University.

McElory Hoffmann

McElory Hoffmann is the CEO and co-founder of Praelexis, a machine learning company that empowers its clients to become data-driven decision makers. He is also an extraordinary senior lecturer in Computer Science at Stellenbosch University.

Johan Van der Merwe

Johan Van der Merwe is a Data Science Strategist at Praelexis.

Annexure 1: Summary of AI audit guide

The AI Audit Guide must be used in conducting AI audits in case of general AI systems. In case of auditing LLMs, this general approach could be applied to the model audit and application audit together with the specific elements indicated below. This is in accordance with the specific audit guide recommended for those generative AI systems.

Note: Responses provided by the auditee and documentary evidence should be collected separately and incorporated to complete the audit assessment.

Performance Audit

Audit Area	Notes	Audit Questions	Score (1-5)
Data	AI systems shall be developed and used in compliance with existing privacy and data protection legislation, while processing data that meets high standards in terms of quality and integrity. It is important to ensure that testing data are not used during the training and validation stages.	1. What data sources are used? Define the data used in terms of its public, proprietary and/or private nature. 2. Are the data used internal and/or provided by a third party? 3. Have you checked the data for representivity, completeness, accuracy, traceability? 4. Was a privacy impact assessment done, and what is the outcome thereof? 5. Have you assessed legal compliance with data protection laws? This must be verified during the audit. 6. How is quality ensured in the selection of the training, testing and validation data? 7. Are the training, test and validation data stored separately? 8. What measures are in place to strengthen quality and integrity of the data?	
Privacy & Data Governance	AI systems shall be developed and used in compliance with existing privacy and data protection legislation.	1. Specify how consent has been secured for the use of personal data, if applicable. 2. Are personal data used in the training data for the AI system? 3. How is security of the personal data ensured? 4. What privacy enhancing techniques are used to mitigate personal data leakage?	
Bias Assessment	There are various bias assessment instruments that determine bias in the datasets, algorithm, or outcome. This could include	1. Were biases in the training, testing and validation data adequately addressed? 2. Were relevant data gaps identified and addressed? 3. Was human oversight	

	specific in-depth analysis of AI model performance to identify possible biases leading to unfair or discriminatory outcomes.	provided to minimise risks? 4. Were specific risks of harm to vulnerable groups identified and mitigated?	
Technical Robustness and Safety	Assessment of the technical robustness and quality of the AI system must be done regularly to ensure objectives are consistently met. The assessment of algorithm functionality could benefit from specific techniques like adversarial testing, stress testing, or sensitivity analysis.	1. Is the purpose of the algorithm clearly defined and described in the context in which it is used? 2. Does the algorithm function in accordance with its purpose? 3. Do you have technical documentation that clearly describes the design, development and implementation of the AI? 4. Are the outputs of the AI system consistently replicable? 5. Is the AI system regularly reviewed to ensure compliance with the latest legislation? 6. What control measures are in place to ensure the accuracy of the AI system? 7. How is quality control of the AI system ensured? 8. How is the output of the AI monitored? 9. Is the system safe and not vulnerable to tampering?	
Transparency	Explainability - testing the ability to provide adequate information to users to understand the output of the AI system.	1. Is clear technical documentation provided? 2. Are the purpose, logic and output of the AI explained? 3. Can people who interact with the AI get clear information? 4. What technical measures are in place to facilitate the interpretation of the outputs?	
Human Oversight	Human agency and oversight means AI systems are developed and used as tools serving people, respecting human dignity and personal autonomy, functioning in a way that can be appropriately controlled and overseen by humans.	1. Is the complete life cycle of the AI system documented? 2. Are the roles and responsibilities of all the persons involved in the development of the AI system clearly documented? 3. What monitoring system is in place to monitor the performance of the AI system? 4. How is human oversight included in the development and deployment of the AI system?	
Compliance Audit	A compliance audit includes performance audit activities		

	and additional activities listed below		
Fundamental Rights Impact Assessments	Overseen by humans. Technical and practical risk management measures.	1. Which fundamental rights are potentially affected by the AI system? 2. Assess the risk to the fundamental right – risk sources, determine likeliness, severity of impact [low, medium, high, very high]. 3. What are the technical measures to mitigate the identified risk? For each identified risk, indicate the mitigation measures. 4. What are the organisational measures to mitigate the identified risk? 5. What monitoring mechanisms are in place to monitor the risk management?	
AI Standards	Specific standards included in legislation or international standards, e.g., IEEE or ISO.	1. What standards apply to the auditee? 2. Was an internal compliance assessment done? 3. How does the AI system comply to the identified standards?	
Auditing LLMs	The 3-phase audit of LLMs consists of governance audit, model audit, and application audit.		
	Application Audit	Apply the AI Audit Guide for general AI systems	
	Governance Audit	Monitoring system outputs. Stakeholder consultation. Auditing database used in applying RAG.	
	Model Audit	1. Adversarial algorithmic auditing, such as red-teaming and honey-potting. 2. Formal verification and benchmarking.	

Scoring scale

- Exellent 5
- Good 4
- Adequate 3
- Poor 2
- Critical 1

