Universiteit Leiden
The Netherlands

Bachelor Computer Science & Economics

Designing an Artificial Intelligence Auditing Framework

Justin Simon de Weert

First supervisor:
Prof. dr. ir. J.M.W. Visser

Second supervisor:
Dr. G.J. Ramackers

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

27/08/2024

Acknowledgments

First of all, I would like to thank everyone who contributed to the completion of this thesis. I am especially grateful for the guidance of my supervisors, Prof. Dr. Ir. J.M.W. Visser and Dr. G.J. Ramackers.

Special thanks go to Bram Dijkers and Laurent Willemsen from KPMG, who provided invaluable insights and feedback from an auditor's perspective, significantly enhancing the quality of this work, and have guided me at KPMG during the time of my internship. I also thank Max de Gast for his thorough review of the model and his excellent recommendations.

Abstract

Following the European Union's introduction of the AI Act, there is an increasing need for audits of AI practices and systems in businesses. This trend, driven by both compliance and ethical considerations, forces auditing firms to develop, update, or improve their auditing methods.

This study aims to develop practical methods for auditing AI systems to ensure that they comply with ethical standards, legal requirements, and organisational values. The goal is to create a framework that auditors can use to effectively evaluate AI systems, focussing on both legal, ethical, and organisational aspects and as a tool for guiding organisations preparing for their audits.

The research was carried out during an internship at KPMG's IT Audit Department. The methodology used is that of Design Science. The approach includes reviewing the current literature on (AI) auditing, reviewing existing frameworks, developing an initial auditing framework, and refining it based on feedback and findings from audit professionals.

The result is an auditing framework that incorporates risk assessment and monitoring methods, which address the existing procedural gaps in the AI Act. This framework provides auditors with the tools and guidance necessary to conduct an AI audit. Drawing on established audit practices and integrating best practices from various existing methods and frameworks, this framework offers a consolidated approach.

Contents

1 Introduction

In today's rapidly evolving digital landscape, organisations increasingly rely on artificial intelligence (AI) to drive business decisions and operational efficiencies [31]. However, the deployment of AI technologies raises significant concerns about ethical standards, transparency, accountability, and data privacy. With AI systems increasingly being integrated into corporate operations, it is crucial to implement audit procedures to ensure their responsible deployment and adherence to emerging regulations, such as the AI Act of the European Union [7].

To comply with these regulations, businesses usually perform audits. Auditing plays an important role within organisations, traditionally aimed at ensuring compliance and reducing the risks associated with financial reporting for the benefit of stakeholders. In the context of AI, auditing extends these principles to the technological realm, addressing challenges such as bias, rationale, source code review, regulatory readiness, data security, process optimisation and the overall impact of AI decisions on stakeholders [23].

The thesis will investigate the implementation of artificial intelligence audit practices, following the approval of the AI Act by the European Union. This legislation requires AI systems to be "safe, transparent, traceable, nondiscriminatory, and environmentally friendly" [6], and emphasises that AI systems should always be under human control. It categorises AI systems based on their risk levels, from minimal risk to unacceptable risk, and imposes stricter requirements for higher-risk categories. The act also emphasises the protection of fundamental rights and the use of high-quality data sets to avoid bias.

This introduction of legislation shows the growing need for thorough audits of AI systems within the business sector. This trend shows that there are opportunities for established audit and assurance firms to adopt new auditing practices to accommodate the rapid changes this trend generates. Mökander et al. point out that the future of AI auditing lies in combining current tools and methods into well-organised and independent processes [23].

The AI Act currently provides limited guidance on auditing, mentioning only that: "The notified body shall carry out periodic audits to ensure that the provider maintains and applies the quality management system and shall provide the provider with an audit report. In the context of those audits, the notified body can perform additional tests of AI systems for which a Union technical documentation assessment certificate was issued" [6]. This suggests a regulatory framework where ongoing compliance and performance of certified AI systems are monitored, but the text also highlights a significant gap in the legislation: it does not specify the details of what exactly should be audited or the methods to be used for auditing. This research focusses primarily on addressing the deficiencies mentioned in the audit guidelines, specifically the absence of details regarding the audit content and methods, and developing a framework to bridge these gaps.

The foundational research question to achieve the objective is presented as follows.
"What assessment methods are available and how can they be developed into a practical methodology for auditors to assess the alignment of AI systems with established ethical principles, legislation, and organisational values?"

To help answer the research question, the following three sub-questions are considered to help answer the main question:

- What existing frameworks and methodologies are currently being employed in AI auditing?
- How do auditors address the challenges of transparency and accountability in AI systems, considering the complexity and often non-transparent nature of AI algorithms?
- In what ways can monitoring and risk assessment methodologies be integrated into auditing processes to effectively manage the dynamic and evolving nature of AI technologies?

1.1 Problem statement

The introduction of the AI Act is a significant step forward in the legislative landscape, with the aim of regulating the rapidly expanding field of artificial intelligence. However, the current version of this act exhibits a broad and often ambiguous scope that incorporates a wide array of software technologies, from advanced machine learning algorithms and deep neural networks to traditional expert systems and basic statistical methods. Such a broad scope could suggest a universal approach to regulation, a one-size-fits-all framework, but that could oversimplify the nuances of various AI technologies. At the core, all of these systems process data, yet the implications, applications, and risks associated with each can vary greatly. This makes the act susceptible to various interpretations [24].

Auditing these systems is important to ensure that they operate safely, ethically, and effectively. Effective regulation should facilitate auditing protocols to assess AI systems both before and after deployment, keeping in mind the life cycle of AI software [9]. These assessments are necessary to prevent potential harm that might arise from biases in decision-making processes, invasions of privacy, or other ethical breaches. Unfortunately, the current language of the AI Act on auditing procedures lacks specificity, causing uncertainty among companies and developers about how to comply with or execute these controls.

Guidelines for AI operations are not only helpful, they are essential. Clear and detailed guidelines provide a blueprint for developers to follow, ensuring that AI systems align with ethical standards, legal requirements, and company values. They also help to establish transparency benchmarks, which are critical to gaining and maintaining public trust in the systems and companies themselves [7]. The broad scope of the AI Act and the lack of precise definitions highlight significant gaps in the legislation.

These gaps need urgent attention to prevent misinterpretations and inconsistent applications of the law, which could lead to negative impacts across different sectors. Without addressing these issues, the AI Act risks becoming an ineffective tool that could either pose a threat to technological progress or fail to provide adequate protections against the risks of AI. With these gaps not being closed in the near future, the focus should be on developing an auditing framework that incorporates risk assessment and monitoring methods. This framework will provide auditors with the necessary tools to assess whether AI systems meet the ethical and legal standards established, thereby implementing the goals of the AI Act.

1.2 Research objective

The primary objective of this research is to find out what current audit assessment methods are currently available to auditing firms and how these can be shaped into a internal audit process that covers the required fields of the EU AI Act.

Another key objective is to analyse how the current shortcomings in the AI Act [24] can be covered in the said framework.

1.3 Deliverables

The deliverables consist of a review of the literature on auditing practices that could be applied in AI auditing, such as IT or financial auditing. Through surveys and interviews, a view on the AI Act and its impact of audit practises will be provided from mainly KPMG employees and other audit professionals.

The findings of these deliverables will be developed into a framework for AI auditing that can be used as a basis for further development of AI auditing and governance mechanics.

1.4 Thesis overview

This thesis is structured into several chapters, each addressing a specific aspect of the research conducted. The content is organised as follows.

- **Chapter 1: Introduction** - This chapter introduces the research topic, including the problem statement, research objectives, and the deliverables. It sets the foundation for the study by explaining the importance of auditing practices in artificial intelligence (AI) systems, specifically in the context of the European Union's AI Act.

- **Chapter 2: Background and Related Literature** - In this chapter, a review of the existing literature related to the practise of auditing and the AI Act is provided. It discusses the contents of the AI Act, its implications, and the gaps that exist within the current legislative framework.

- **Chapter 3: Methodology** - This chapter describes the research methodology used in the thesis. The Design Science approach is outlined, detailing the steps followed from problem identification to the communication of results. It also includes the specific objectives and the design and development of the AI auditing framework.

- **Chapter 4: Requirements** - This section covers the requirements elicitation process, including the identification of stakeholders, the rationale behind the requirements, and the listing of functional and quality requirements. The constraints of the framework are also discussed.

- **Chapter 5: Design** - This chapter provides a description of the proposed AI auditing framework. It outlines the phases of the audit process, including scoping, mapping, artefact collection, testing, and reflection. Each phase is explained with respect to the methods and guidelines used to satisfy the AI Act's requirements.

- **Chapter 6: Evaluation** - The evaluation of the proposed framework is discussed in this chapter. It assesses the framework's effectiveness in meeting the set objectives and addresses any potential limitations, as have come forward in confirmatory interviews with audit experts from KPMG.

- **Chapter 7: Conclusions and Further Research** - This final chapter presents the conclusions drawn from the research and suggests areas for future investigation. It reflects on the significance of the findings and their implications for AI auditing practices.

2 Background and Related Literature

2.1 The Audit Process

As stated, auditing is an important practice when it comes to ensuring compliance with regulations. That makes defining what an audit is important as well, to fully understand what we are doing. Gupta states:“In the broadest sense, auditing refers to an independent examination of any entity, conducted with a view to express an opinion thereon” [15]. The importance of auditing originates from its role in mitigating the principal-agent problem, a scenario where the interests of the principals (owners or shareholders) may not align with those of the agents (managers or executives) responsible for running the organisation [17]. Functionally, auditing provides a governance mechanism for principals to ensure that agents act in the best interests of the organisation, whether a system meets legal standards or if certain risks need to be mitigated, thus enhancing transparency, accountability, and trust [23]. As a method, Mökander et al. defined the practice as “auditing consists of a structured process which tries to assess an entity's past and present behaviour, may it be financial or technical” [24].

To further develop the previous definition, Mökander & Floridi have defined an AI audit as follows: “AI auditing is characterised by a structured process whereby an entity's past or present behaviour is assessed for consistency with predefined standards, regulations, or norms” [26].

The fundamental theory of auditing is that it is an objective evaluation process rather than an advisory service. Auditors are responsible for assessing whether an entity's financial statements, processes, or systems comply with applicable standards and regulations, but they do not provide advice on how to achieve actual compliance. This distinction is important because it maintains the independence and impartiality of the audit process. Auditors must remain unbiased [17]. They should therefore not become involved in the decision-making or operational aspects of the entities they audit, ensuring that their findings are based solely on objective evidence.

Shaping practice, norms, and standards play a crucial role in auditing by providing a framework for consistency, reliability, and comparability. In the financial auditing world, the most prominent sets of auditing standards come from the Public Company Accounting Oversight Board (PCAOB) or the International Auditing and Assurance Standards Board (ISAAB), where ISAAB is considered the standard in Europe [20]. One of the most prominent sets of standards to be audited is often provided by the International Organisation for Standardisation (ISO). For Artificial Intelligence, ISO/IEC 42001: 2023 can be considered as a norm for maintaining Artificial Intelligence Management Systems, which at the time of writing is only in its publication phase and has yet to be reviewed in practice [10].

It is notable to know that audits can be categorised into internal and external audits, each serving distinct but complementary purposes.

Internal Audits: Conducted by the internal audit department of an organisation, these audits focus on assessing internal controls, risk management, and governance processes. Internal auditors provide management with information and recommendations to improve organisation operations and ensure compliance with internal policies and procedures. Contrary to external audits, internal

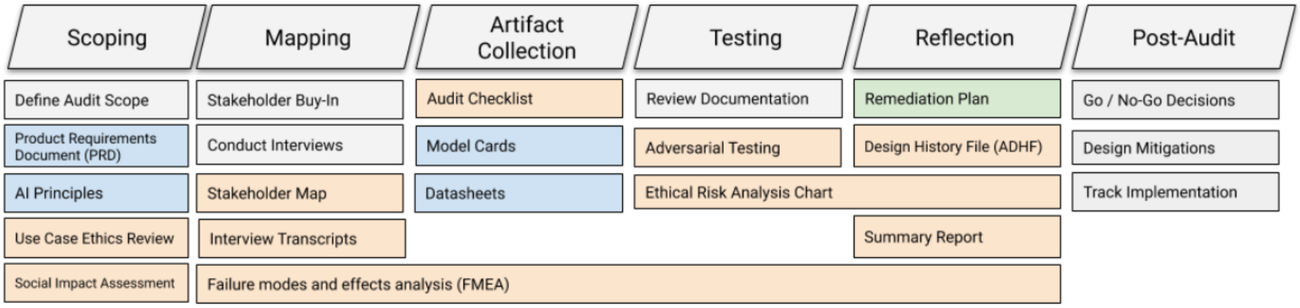

Figure 1: The AI audit process overview provided by Raji & Smart et. al [29]

auditors work for the company being audited and have an interest in improving the system being audited, so they may give these recommendations.

External Audits: Performed by independent third-party auditors, external audits focus mainly on the accuracy and credibility of an organisation's financial statements [17]. They provide an unbiased opinion to external stakeholders, such as investors, regulators, and lending institutions, about the organisation's financial health and compliance with applicable laws and standards. Eventually, external Auditor have to sign off the financial statements of the party who is being audited, making the report legally binding.

Both types of audit are of great importance for maintaining a governance framework. Internal audits help in early detection and correction of issues, while external audits provide independent the eventual verification that enhances and establishes the credibility and reliability of an organisation's reporting.

Defining the phases of an internal AI audit can be done using the SMACTR framework proposed by Raji & Smart et. al, specifically designed for Artificial Intelligence Auditing [29]. This framework consists of Scoping, Mapping, Artefact Collection, Testing, and Reflection. In figure 1, the grey colours indicate a process, orange indicates documents created by auditors, blue indicates documents by engineering and product teams, and green indicates that the document is jointly developed.

The need for an audit comes from a governance requirement, which is discussed in Section 2.2 The AI Act. The Act requires a form of risk-based assessment. The key question to ask for risk-based assessments is to ask yourself "What if?". Arslan et al. argue that "At a minimum, the internal audit process should enable critical reflections on the potential impact to a system, serving as internal education and training on ethical awareness in addition to leaving what is referred to as a "transparency trail" of documentation at each step of the development cycle" [3]. They stress that documentation is key for a transparent audit. The SMACTR framework will serve as a guide for developing the steps required in the audit process and designing the actual framework. To achieve a better understanding of the framework, each step is described.

Scoping: The scoping phase consists of defining the objective of an audit. What should be audited and why? What guidelines should be followed and tested in the audit? This is the phase in which the risk analysis of the AI act guides us to the correct scope of an audit.

Mapping: In the mapping phase, is the step where it is reviewed what metrics are already in place, and the time to define the internal stakeholders and prepare the stakeholder buy-in, meaning getting all stakeholder on-board with the mission and vision of the project.

Artifact collection: In the arifact collection phase, all the documentation regarding the AI system have to be collected according to the audit checklist.

Testing: The testing phase is the main phase of the internal audit. This is time where internal auditors execute their tests and assessments to test the compliance of the system to the applicable guidelines. This phase will lead to creating the remediation plan. An audit will probably lead to some findings and these need to be presented to the stakeholders, such as the board or product teams. The remediation plan will guide these stakeholders into remediating the found issues.

Reflection: The reflection phase is where audit results are analysed against ethical expectations set during the scoping phase. Auditors update the final risk analysis, identifying ethical principles that may be compromised by the AI system. Key outcomes include a mitigation plan developed with engineering teams, outlining prioritised risks and necessary actions for future deployments. This phase also involves making design recommendations, such as improving data diversity or adding user consent features.

2.2 The AI Act

The rapid evolution of artificial intelligence technologies has had a great effect in various sectors. As AI systems become more widespread, their potential risks and ethical implications have sparked global debates, prompting the need for AI legislation. According to a McKinsey survey, 65 percent of respondents tell their organisation has adopted the use of generative AI regularly [4]. The field in which it is most used is Marketing and Sales, for which 16 percent of respondents use it for Content support and 15 percent for Personalised Marketing.

But what is considered artificial intelligence? The AI Act itself states: "'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments" [7]. This definition is quite complex, and Tabassi et al. have made a clear definition based on the OECD recommendation on AI:2019 and ISO/IEC 22989:2022. AI for this case could be considered as "An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy" [33], however, this definition does not mention the adaptability of the system after deployment.

Therefore, combining the strengths of both definitions, a more clear to understand and precise definition can be formulated: "An AI system is a machine-based system designed to operate with varying levels of autonomy. It processes input data to generate outputs, such as predictions, content, recommendations, or decisions, that can influence real or virtual environments. AI systems may adapt and improve their performance over time based on new data and experiences".

It is crucial not only to manage risks such as privacy violations, discrimination, and security threats, but also to establish frameworks for accountability and transparency in the development and deployment of AI. As the Information Systems Audit and Control Association (ISACA) stated in 2018: "There is, for example, no mature auditing framework in place detailing AI sub processes, nor are there any AI-specific regulations, standards or mandates" [16].

The European Union (EU) has been at the forefront of addressing these challenges. The EU's approach to AI legislation began with recognising the nature of AI as both an opportunity and a risk. Starting with preliminary frameworks and guidelines, such as the Ethics Guidelines for Trustworthy AI by the High-Level Expert Group on Artificial Intelligence in 2019 [6], the EU has aimed to harness the benefits of AI while mitigating its risks. This insight led to the creation of more formal and binding legislation.

In response to these needs, the EU drafted the AI Act, which is considered a pioneering move in global AI regulation. Proposed in 2021 and recently approved [8], the AI Act is comprehensive legislation designed to govern the use of AI in all 27 member states.

2.2.1 Contents of the AI Act

The contents of the AI Act can, for the sake of simplicity, be categorised into five points in which it tries to achieve its objectives. These are outlined below and have been extracted from the AI Act [1].

1. **Risk Assessment:** The AI Act requires high-risk AI systems to undergo rigorous testing and compliance checks before their deployment, including a risk assessment of accuracy, cybersecurity, and data protection impacts.
2. **Transparency Obligations:** For AI systems that interact with individuals, such as chatbots, or AI that generate or manipulate image, audio or video content, transparency obligations are stipulated to ensure that users know they are interacting with an AI, not a human. This aims to prevent deception and foster an environment of trust and accountability.
3. **Prohibitions:** The AI Act outlines clear prohibitions for certain uses of AI, such as AI that exploits vulnerable groups or involves social scoring, which could lead to discrimination or exclusion.
4. **Data Governance:** The act places strict requirements on the quality and management of data used to train AI systems, ensuring that the data are legally sourced and free from biases that could lead to discriminatory results.
5. **Enforcement and Compliance:** To enforce these provisions, the AI Act proposes significant penalties for noncompliance, potentially amounting to 30 million euros or a percentage (6 percent) of a company's annual global turnover.

At its core, the categorisation of AI systems is what defines what kind of measures need to be taken to comply with the above-mentioned points. The Act defines four levels of risk: minimal, limited, high, and unacceptable [8]. The EU Commission has visualised this in the pyramid in figure 2.

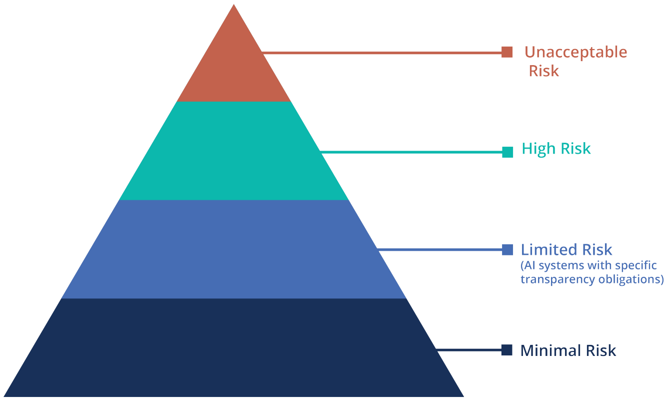

Figure 2: The Pyramid of Risks as proposed by the European Commission [8]

The following types of systems would fall under the **Unacceptable Risk** category [1]:

- AI systems that deploy harmful manipulative 'subliminal techniques';
- AI systems that exploit specific vulnerable groups (physical or mental disability);
- AI systems used by public authorities, or on their behalf, for social scoring purposes;
- 'Real-time' remote biometric identification systems in publicly accessible spaces for law enforcement purposes, except in a limited number of cases.

The AI Act distinguishes two types of **High Risk** [1]:

- Systems used as a safety component of a product or falling under EU health and safety harmonisation legislation (e.g. toys, aviation, cars, medical devices, lifts)
- Systems deployed in eight specific areas identified in Annex III , which the Commission could update as necessary through delegated acts (Article 7):
  - Biometric identification and categorisation of natural persons;
  - Management and operation of critical infrastructure;
  - Education and vocational training;
  - Employment, worker management and access to self-employment;
  - Access to and enjoyment of essential private services and public services and benefits;
  - Law enforcement;
  - Migration, asylum and border control management;
  - Administration of justice and democratic processes.

This classification mandates compliance to additional regulations [1]:

- Registration in an EU-wide database, before deployment is allowed;
- If not already established due to other regulation, a conformity assessment (self-assessment) is required.
  - Only high-risk AI systems used for biometric identification require an assessment by a 'notified body' [8].
- Establish a risk management system [27]
- Establish a quality management system
- Conduct data governance
- Draw up techinical documentation
- Configure for automatic record-keeping
- Supply information and documentation to downstream users of the system
- Allow the implementation of human oversight
- Design the system to achieve appropriate levels of accuracy, robustness and cybersecurity.

AI systems classified as **Limited Risk**, such as General Purpouse AI (GPAI) only require a limited set of transparency obligations [1]:

- Draw up techinical documentation
- Supply information and documentation to downstream users of the system
- Establish policies to respect the copyright directive
- Publish a sufficiently detailed summary about the content used for training

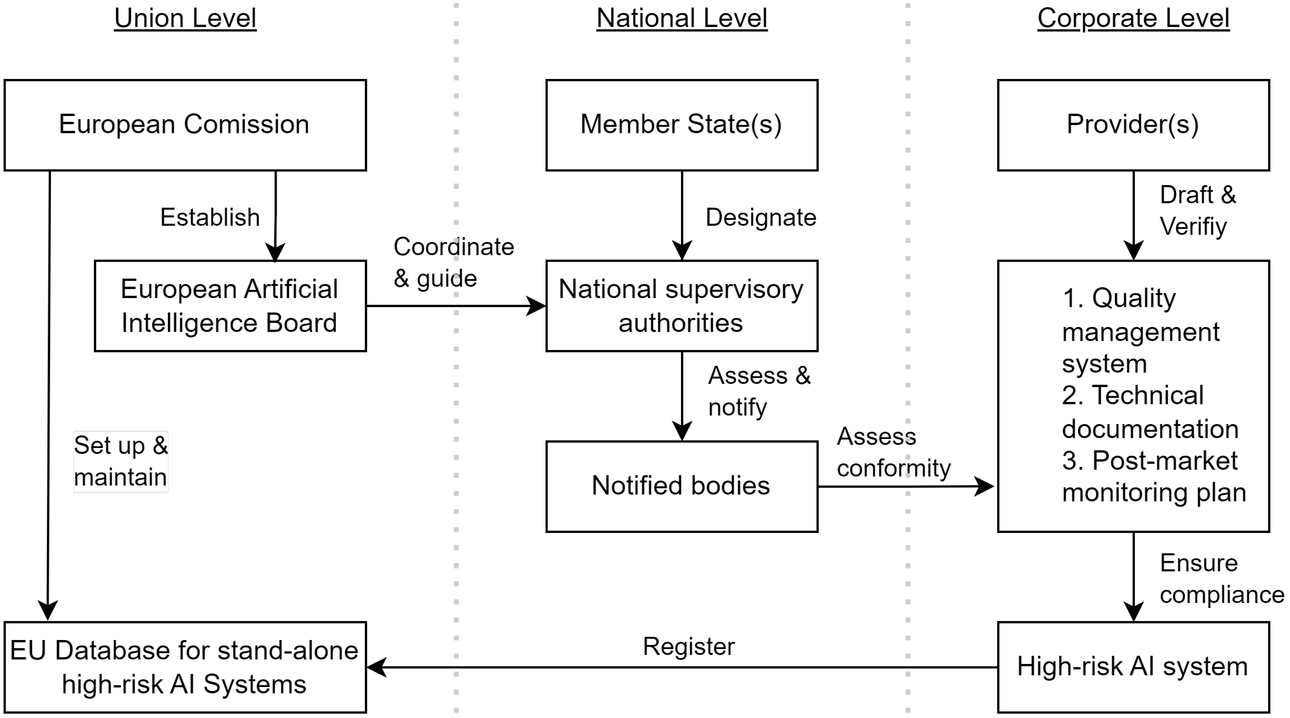

Figure 3: Roles and responsibilities during conformity assessments with the involvement of third-party auditors as modelled by Mökander et. al. [24]

In addition to risk classification, the AI act also provides a guide for who is responsible for what in the context of AI legislation and conformity assessment, as defined by Mökander et al. in figure 3 [24].
The AI Auditing legislation sketches the outline of a new kind of ecosystem, which encompasses the most notable bodies that play a role in the governance of AI systems. A notable detail is that the European Artificial Intelligence Board is not an independent organisation but more of a coordinating structure, whereas the GDPR law was implemented by the independent legal personality EDPB [5].

2.3 Gaps in the AI Act

The AI Act serves as a guide for the governance of AI, but does not provide any standards for auditing the AI system itself. This is left open for other researchers or organisations to develop. The word “auditing” is not even mentioned, except on a few occasions, which Mökander et al. correctly stated [24]. In the paper “Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation” they came to the conclusion that the AI act leaves even a lot more gaps in other regions of what could be a part of AI auditing.

Level of abstraction:
Mökander points out that some of the goals in the AI act are very optimistic and might be hard to achieve, such as the expectation that data sets have no errors at all. He argues that setting goals too high or making them too vague could lead to the accountable process- and product-owners just going through the motions to meet requirements without any real sort of commitment, affecting the credibility of the AI act.

Material scope:
He discusses how the AI act is not very clear about what it covers, even sometimes seeming to include a lot. For example, it mentions many kinds of software technique, from machine learning such as deep neural networks to expert systems and statistical methods. The idea of having one set of rules for all these different technologies is seen as challenging because the definition is too broad. Mökander argues that a specific focus could help AI providers, auditors, and authorities use their resources better. He also has a concern about the law’s attempt to list all high-risk AI applications, which might oversimplify things and overlook some risky uses, like AI in the military or in setting insurance prices. Another issue is the law’s choice to leave out AI used in international law enforcement. The broad way ’artificial intelligence’ is defined and the attempt to list all its applications might actually go against the law’s main goal. Since the law is supposed to address the complex and somewhat unpredictable nature of certain AI systems, there is a call for more clarity on what the material scope in fact is.

Conceptual Precision:
According to Mökander, the AI Act often uses unclear language, making it hard to understand certain parts, like what exactly counts as a ’risk to fundamental rights’ or what AI uses are not allowed. Kazim et al. supports this claim, the definition of risk in terms of “significant risk of injury, death or damage” is vague and does not adequately address the issue of unintended consequences [18]. For instance, it is not clear how to decide if an AI’s subtle influence on people’s decisions is harmful and should be banned. The rules on when AI can skip standard checks are also too broad, like in emergencies related to public safety or health. There is also confusion about when AI companies need to report problems to the authorities, as figuring out if an AI was really to blame can be quite tricky. Lastly, it is not clear when existing AI systems should be checked again if they have changed a lot.

Procedural guidance:
The act sets rules for checking AI systems and keeping an eye on them after they are being used, but it is not very clear on the details of how to do these steps. It talks about the outcomes it wants

to prevent, but not about how it should be prevented [9]. For example, it says records need to be kept for a time that makes sense for the AI system's use, but it does not say exactly how long that should be or who decides. It also mentions that special audits should be done regularly to make sure AI companies are following the rules, but it does not say how often these checks should happen or what causes them to start. In addition, the AI Act mainly scopes providers and deployers of AI systems, not considering that AI used just within companies for research can have ethical issues. And the broad sector-based approach to risk identification could lead to risk aversion and make AI development economically unfeasible, even for low-risk applications [18]. So, there is a great need for more specific instructions on how to carry out these checks and monitoring.

Institutional mandate:
The Act gives the European Commission and national authorities the job of enforcing its rules, with help from the European Artificial Intelligence Board. But it is not clear what exactly the Board is supposed to do, and it seems like the Commission has too much control over it. In addition, the AIA allows countries to have their own extra rules, which could lead to a mix of different approaches, just as happened with the GDPR. In addition, the AIA wants AI companies to follow voluntary ethical guidelines, but there is no strong system to ensure that they actually do. The voluntary nature could lead to "ethics shopping", meaning mixing and matching ethical principles from different sources to justify some preexisting behaviour or "Ethics bluewashing", which means a company tries to pretend to be more ethical than they are. A possible fix could be to set up an independent group to check that companies are really sticking to their ethical promises, especially for AI systems that are not considered high-risk. This claim is backed up by Vetter, as he also argues that there are still many aspects of an AI system that can not really be covered by any law and are up to the governance of ethics itself. He states an important question that one should ask, namely "is an AI system the most appropriate and ethical solution for the problem at hand?" [35].

Transparency:
When creating and using AI, sometimes different ethical rules can clash, like making decisions more accurate but unfair to some groups. There are also different ideas about what is fair, and they cannot all be right at the same time. The AI Act tries to help by asking for clear documentation of how AI is made, so that people can see the ethical choices being made. It is okay for companies to make tough ethical choices as long as they are legal and workable. But, since these companies have to think about what many different people want, European rules could give more advice on how to handle situations where ethical rules conflict, like choosing between being accurate and keeping information private, or deciding which idea of fairness to use in different cases.

Checks and balances:
Although the AI Act requires checks on high-risk AI systems, the actual enforcement might not be as tough as it sounds. Most of the time, companies making these AI systems check themselves and only need to show their compliance documents if asked by national authorities, without having to share this with the public. This can be risky because companies may focus more on rapid development rather than making sure their AI is safe and ethical, and they might not always make the best choices without someone else watching over them. To make these checks better, the rules could be changed to make the process more open to everyone or have independent groups check the companies' work. However, even with some help from European courts, there isn't a clear way to make sure these improvements will work across all kinds of AI laws because the court decisions are too scattered and do not solve the issue of companies needing to carefully consider their actions.

Due to these gaps in the AI Act, the CapAI procedure has been developed [9]. It provides procedural guidance on how AI technology providers can verify claims made about the AI systems they design and deploy. It has been developed with two main ideas in mind. Firstly, for providers with a classification "high-risk" so that they can show compliance with the AI Act. Secondly, for "low-risk" AI systems, it is necessary to voluntarily adhere to the guidelines. CapAI is based on the Ethics Based Auditing (EBA) method, which is defined as "a governance mechanism that allows organisations to operationalise their ethical commitments and validate claims made about their AI systems" [9].

At first glance, the CapAI procedure appears to address the gaps in the AI Act, but it lacks a comprehensive checklist or process description that maps all requirements from the AI Act to actionable steps.

3 Methodology

The design and development of the AI auditing framework followed a structured approach, that of Design Science, described, among others, by Peffers et al [28]. Design Science is an applied research methodology focused on creating and evaluating artefacts intended to solve identified problems. Unlike some more traditional scientific methods that primarily aim to explain and predict events, Design Science seeks to produce innovative solutions that can be practically applied, as is the case in this research. It exists to bridge the gap between theoretical research and practical application, particularly in fields such as IT, engineering, and general technology. In short, Design Science aims to address complex, real-world issues through the development of constructs, models, and methods. As the goal of this research is to develop a model for AI auditing, employing Design Science as the method is the most logical choice.

The design science process consists of six key steps. These are outlined below.

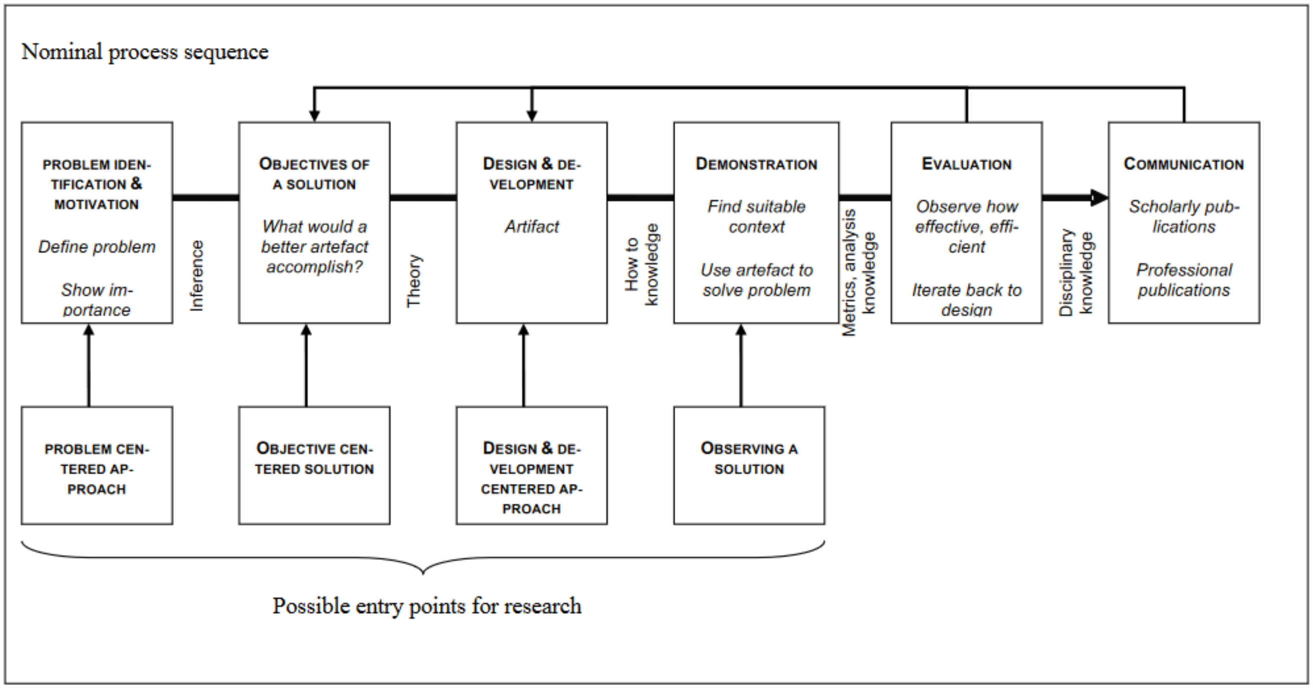
Nominal process sequence
PROBLEM IDENTIFICATION & MOTIVATION
Define problem
Show importance
Inference
OBJECTIVES OF A SOLUTION
What would a better artefact accomplish?
Theory
DESIGN & DEVELOPMENT
Artifact
How to knowledge
DEMONSTRATION
Find suitable context
Use artefact to solve problem
Metrics, analysis knowledge
EVALUATION
Observe how effective, efficient
Iterate back to design
Disciplinary knowledge
COMMUNICATION
Scholarly publications
Professional publications
PROBLEM CENTERED APPROACH
OBJECTIVE CENTERED SOLUTION
DESIGN & DEVELOPMENT CENTERED APPROACH
OBSERVING A SOLUTION
Possible entry points for research

Figure 4: Design science research process (DSRP) model [28]

3.1 Problem identification and motivation

The problem is derived from the review of the literature in introduction. The foundational knowledge on auditing and the AI Act is established. The topics included are the definition of auditing, the process of auditing, the scope of the AI act, the gaps, and its implications.

3.2 The objectives

The objectives of the solution are formulated through the requirement elicitation process in section 4. Requirements.

3.3 The Design and development

The design and development of the framework involve the creation of several diagrams that will serve as the primary deliverable. It will link the various chapters of the AI act with the auditing process and its deliverables. The process includes:

1. **Framework Architecture:** The diagram in the form of a UML Activity Diagram will be the main visualisation of the overall architecture of the AI Auditing framework, showing how different processes interact and what deliverables they create.

2. **Requirements:** The diagram will be built based on the requirements derived from the AI Act, the Questionnaire, and expert input.

3. **Development process:** The diagrams are developed through a process consisting of the initial development, feedback, and refinement and the final implementation.

3.4 Demonstration

Although demonstrating the audit framework is an important part of validating its applicability and effectiveness, it was not feasible to perform this step within the scope of this research project.

3.5 Evaluation

The diagram will be evaluated through expert reviews.

3.6 Communication

The final Activity Diagram, along with its design process, findings, and evaluations, is documented in this thesis. The results will also be shared with the relevant stakeholders, including KPMG's IT Audit Department, to facilitate the adoption and further development of the framework in professional practice.

4 Requirements

The method used to describe requirements is the one described in the book by Suzanne and James Robertson, Mastering the Requirements [30].

4.1 Requirements Elicitation Process

The elicitation process involved sending a questionnaire to KPMG employees in the IT Audit and Responsible AI department, as well as through my LinkedIn network. The questionnaire received 30 responses. The question with corresponding answers that provided the best guidance on what to focus on was the following.

4.1.1 Prioritised Areas in the AI Act

Respondents were asked to rank areas that should be prioritised in the AI Act. The frequencies of each rank (1 to 6, where 1 is most important and 6 is least important) for each area are summarised below:

- **Sector-specific guidance**: The most frequently ranked as 6 (least important).
- **Clarity on third-party assessments**: Distributed fairly evenly across ranks.
- **Post-market monitoring guidelines**: Primarily ranked as 4 or 5.
- **Procedural guidance for audits**: Concentrated around ranks 3 and 6.
- **Defining institutional roles**: Mostly ranked as 3 or 4.
- **Enhancing ethical and legal checks**: Often ranked 1 (most important).

The data indicates that improving ethical and legal checks is seen as the most important area for improvement, whereas sector-specific guidance is perceived to be less important.

4.2 Stakeholders Involved

The development of the AI Audit Framework involves various stakeholders, each with different roles and interests. These are derived and classified according to the stakeholders mentioned in the AI Act [6].

- **AI System Providers:** These are the developers and manufacturers of AI systems who need to ensure their products comply with the AI Act and other relevant regulations.
- **AI System Deployers:** Organisations that implement AI systems within their operations. They must ensure the AI systems they use comply with regulatory requirements and operate ethically.
- **Auditors:** Independent bodies or internal teams responsible for assessing the compliance of AI systems.

- **Regulatory Authorities:** Entities such as the European Commission and national authorities responsible for enforcing the AI Act and ensuring compliance.
- **End-users:** Individuals or organisations that utilise AI systems. Their interests are in ensuring the AI systems are reliable, fair, and do not infringe on their rights.
- **Public:** The general public, whose trust in AI technology is critical. Their interests lie in ensuring that AI systems are ethical, transparent, and do not pose any harm.

The main stakeholders prioritised in this research are auditors and AI System Providers/Deployers. Auditors perform audits, and the AI system providers and deployers undergo an audit, making them the key stakeholders.

4.3 The requirements

The requirements for the model have been formulated according to the principle of atomic requirements. This means that a requirement consists of a unique number, description, rationale, and a fit criterion. The requirements have been formulated according to the perceived gaps, focussing on material scope and procedural guidance as well as the results of the questionnaire. The requirements are also linked to a specific in the AI act, to be able to see what part of the act the requirements covers and therefore create a mapping of fulfilled requirements and its corresponding AI Act legislative requirement per phase in the audit. Three requirements categories have been defined, consisting of functional and quality requirements, as well as constraints. A total of 16 requirements have been formulated, following the template of an atomic requirement, as proposed by Robertson et al. [30]. They consist of a title, fit criterion, rationale, and the source of the necessity of the requirement. All requirements have been listed in Appendix A.

4.3.1 Rationale for Requirements

The rationale specifies the thought process behind a requirement and why it should be included as a requirement to achieve the greater goal. These come forward in the requirement elicitation process.

4.3.2 Fit Criterion

The fit criterion specifies the conditions that must be met for each requirement to be considered fulfilled. It serves as a measurable standard against which the implementation of the requirement can be evaluated. The fit criterion ensures that the requirements are not only theoretical but also achievable.

5 Design

The design of the proposed framework follows that of the SMACTR, with some alterations, as seen in Figure 5. The post-audit phase will not be modelled as the focus of this thesis is on auditing itself, whereas the post-audit phase suggests governance steps after an audit. This is the baseline for the audit procedure. Each phase will be defined with what has to be done for each step, what input documents are necessary, what deliverables will be generated, and which other frameworks or processes could assist in achieving the goals of each step. Furthermore, each key artefact will be assigned to the corresponding articles of the AI act that it covers. This will help to gain insight on what has been done and what should still be done.

The model is modelled in the Unified Modelling Language. Each phase consists of two swim-lanes at most, where the top swim-lane represents the auditor and the bottom swim-lane the product team or product owner of the auditee. When there is an intermediate process that is important for creating a main deliverable, that process is specified in the table in each phase section when one is available. The creation of the deliverable is linked to one or multiple requirements and mapped to the corresponding AI-Act articles that directly or indirectly dictate the use or creation of that deliverable. For ease of use and clarity, the reference to the subprocess is also modelled on the right side of each phase model, so that the model can also be easily used without the need of having the full thesis present.

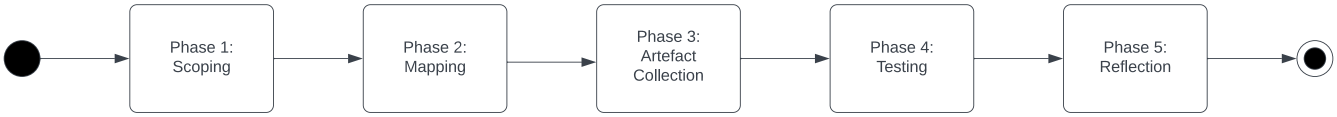

Figure 5: The altered SMACTR framework

5.1 Phase 1: Scoping

Scoping is the phase of an audit in which the depth and breadth of the audit is defined. The objective is to pinpoint areas that could cause harm or have social repercussions. At this stage, there is minimal interaction with the system [29].
Starting in the scoping phase, the auditor must first consider the following; What are the audit objectives? These start with the system risk classification. To set goals for what should be audited and why, in-depth risk analyses should be performed to map out which regions of the system are the most important. The alignment of audit objectives with the company's values is important, as it ensures that the audit is relevant and its findings are actionable, reflecting the organisation's broader mission and ethical commitments. Therefore, a key descion point is implemented in the model. Alignment creates trust between auditors and the organisation, ensuring that the audit is conducted in good faith and supports the long-term sustainability of the company by identifying risks and opportunities that align with its values.

The next step in the process is to define the risk category of the AI system. As mentioned, the AI Act suggests four risk categories, but it is not clear when a system falls into a certain category. Raji & Smart et. al. have chosen not to restrict which systems to audit and in what risk category to classify them. Therefore, a UML Activity Diagram is created that tries to, in a broad view, classify AI systems to determine the risk of that system. This is based on the requirements in Chapter III, Article 6. From the survey, as well as the acknowledged gaps in the AI act, it has been reported that procedural guidance should be improved upon, as well as providing a material scope clarification. Therefore, Figure 6 satisfies the specific functional requirements FR1: Risk Assessment Process, FR8: Procedural Guidance for Auditing and FR9: Material Scope Clarification. The intended use of this model is specified in further detail in Appendix B, along with examples of how to use it.

The ethical review must be seen as a check to discover if the system aligns with the organisation's ethical values and principles. For conducting an ethical review, there are no strict guidelines. The Z-Inspection framework can be used as a basis for conducting an ethical review. It tries to build its process around four ethical principles based on fundamental rights. These are "Respect for human autonomy, prevention of harm, fairness, and explainability" [36]. The ethical issues found for these fundamental rights are then mapped to the seven requirements established by the EU High-Level Experts Guidelines for Trustworthy AI, with the addition of two additional requirements deemed necessary by the creators of Z-Inspection.

For the Social Impact Assessment, the Human Rights, Ethics, and Social Impact Assessment (HRESIA) model by Alessandro Mantelero can be applied [21]. It provides a clear set of tools to determine the social impact of the use of AI systems and sets out the complete process, even providing a case study to show how the framework can be applied. The framework can create visual representations of the impact of AI systems, providing auditors and stakeholders with a simple and clear impact classification method.

In contrary to what the authors of the SMACTR framework suggest, the social impact assessment and the use case ethics review should be created by the product team and only reviewed by the auditor, as the main job of an auditor is to check compliance.

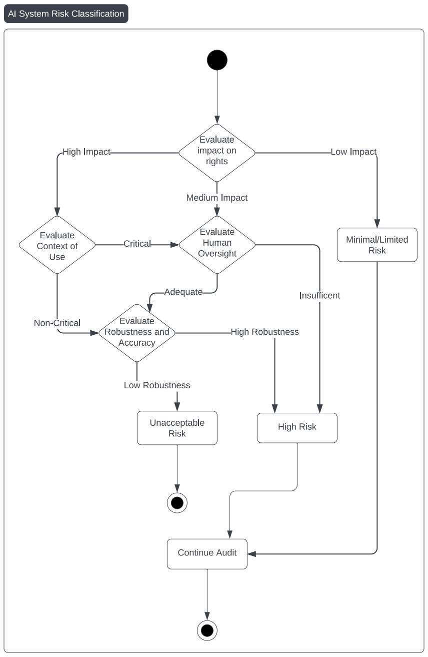

Figure 6: Risk Classification Diagram

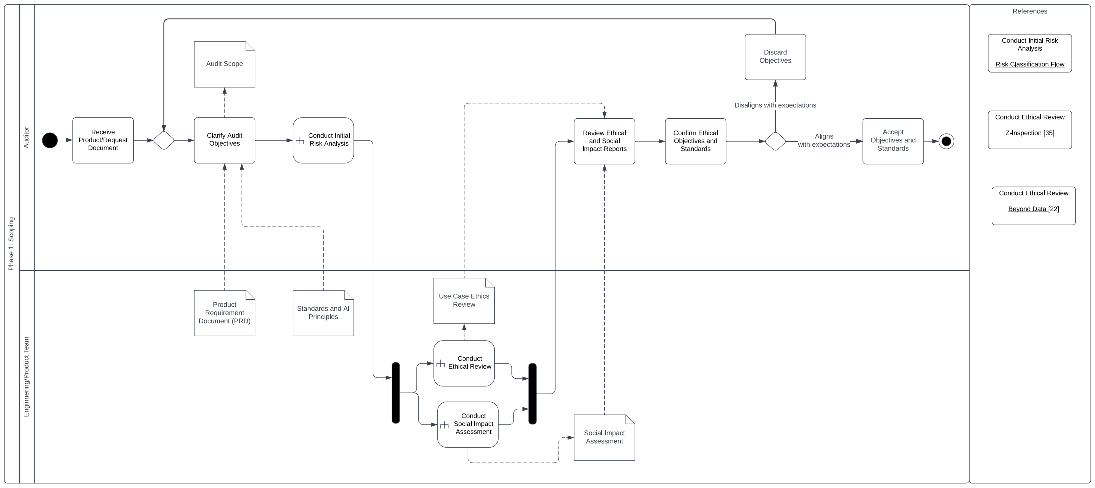
Phase 1: Scoping
Auditor
Enginnering/Product Team
Receive Product/Request Document
Audit Scope
Clarify Audit Objectives
Conduct Initial Risk Analysis
Product Requirement Document (PRD)
Standards and AI Principles
Use Case Ethics Review
Conduct Ethical Review
Conduct Social Impact Assessment
Social Impact Assessment
Review Ethical and Social Impact Reports
Confirm Ethical Objectives and Standards
Discard Objectives
Disaligns with expectations
Aligns with expectations
Accept Objectives and Standards
References
Conduct Initial Risk Analysis
Risk Classification Flow
Conduct Ethical Review
Z-Inspection [35]
Conduct Ethical Review
Beyond Data [22]

Figure 7: Phase 1 Scoping UML Activity Diagram

Key Artefact	**Method/Guidelines**	**Requirements satisfied**	**Relevant AI Act Articles**
Risk Classification	Risk Classification Diagram B	FR1, FR8, FR9	Chapter III, Article 6
Ethical Review	Z-Inspection [36]	FR6	Chapter III, Article 60
Social Impact Assessment	Beyond Data [21]	FR8	Chapter III, Article 9, 14, 15, 27

Table 1: Summary of Key Artefacts and Methods for Phase 1: Scoping

5.2 Phase 2: Mapping

The goal of the mapping phase is to get all stakeholders on board and map the system to better understand the systems and its stakeholders. The main deliverables of this phase consist of a stakeholder map, a failure mode and effect analysis (FMEA), and an ethnographic field study.

The stakeholder map defines stakeholders based on their influence on a subject and their interest. This is most of the time a two-dimensional axis, where each quadrant supports an attitude towards a type of stakeholder; informing, monitoring, managing and satisfying. However, some researchers argue for a 3-dimensional map, adding an axis for attitude, to create a better opportunity for stimulating thought and creating better insights for stakeholder management [25]. Therefore using the 3-dimensional map would be a better option than the standard power vs interest mapping, as it creates a better overview of stakeholders.

The FMEA serves as a tool to examine a system for possible failures, and it can help designers improve or upgrade their system, as well as helping decision makers formululate preventive measures. This approach has been widely used in many fields such as aerospace and engineering, but has been introduced by Raji et al. as a method to examine ethical risks in artificial intelligence products [29]. According to Stamatis, the FMEA is a “living document”, which means that it is a dynamic tool that is never actually finished. The only time the FMEA can be considered complete is when the system is considered complete or even discontinued [32].

From the perspective of an auditor, the FMEA should be examined to ensure that it is up-to-date and accurate. The auditor’s role is not just to verify that the FMEA exists but also to evaluate its content. This involves checking that the FMEA reflects the current state of the system, including any updates or changes, and that it accurately identifies and assesses potential risks. This evaluation is important to maintain a robust governance structure for AI systems, ensuring continuous risk management. Therefore, the auditor must ensure that the FMEA is not only actively maintained by the AI provider, but is also substantively correct and reflective of the true risks associated with the AI system.

Supplementing the FMEA, the ethnographic field study can offer significant insight by providing an in-depth qualitative understanding of how AI systems are used and perceived in real-world settings by diverse user groups. An Ethnographic Field Study involves immersing auditors in the actual environments where AI technologies are deployed. This method allows auditors to observe first-hand the interactions between users and the AI system, revealing unanticipated issues such as cultural biases, usability challenges, and context-specific ethical concerns. However, it is important to consider the practicality of this approach from the auditor's perspective.

Conducting a full ethnographic study can be resource intensive and may not always be feasible within the typical scope of an audit. Therefore, auditors may need to rely on a combination of direct observations, user interviews, and existing user research provided by AI developers. Although an ethnographic field study may offer deep insights, auditors should weigh its benefits against practical constraints, possibly opting for a more targeted approach that still achieves the goal of uncovering critical user interactions and ethical issues without requiring full immersion; however, that approach would differ vastly per AI system.

From the practical experience of audit professionals, it has been noted that a stakeholder map is a jointly developed artefact together with the organisation. Therefore, in deviation from the SMACTR model, the stakeholder map artefact is shown as jointly developed. Due to the importance of getting all stakeholders onboard, a decision point has been introduced after the buy-in phase to check if the buy-in has been successful, suggesting the importance of this step.

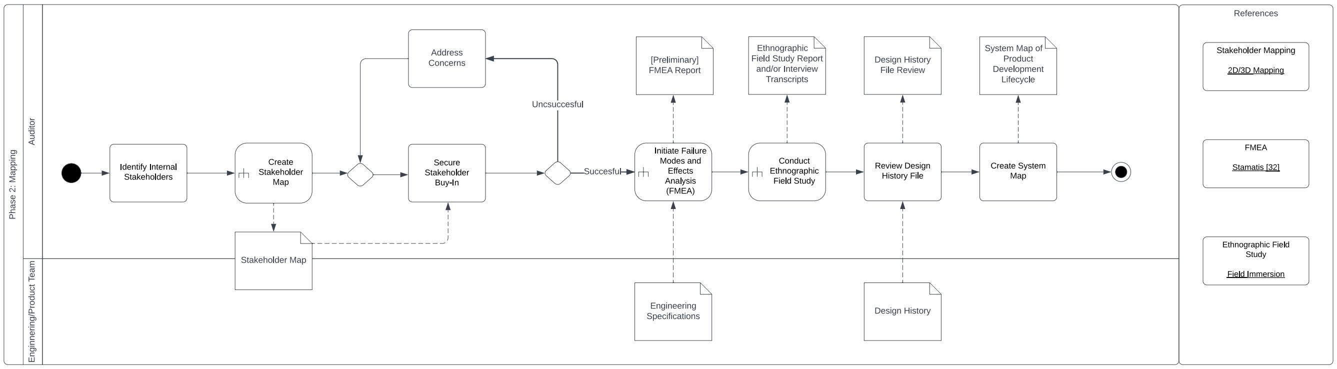
Phase 2: Mapping
Auditor
Engineering/Product Team
Identify Internal Stakeholders
Create Stakeholder Map
Stakeholder Map
Address Concerns
Secure Stakeholder Buy-In
Uncsuccesful
Succesful
Initiate Failure Modes and Effects Analysis (FMEA)
[Preliminary] FMEA Report
Engineering Specifications
Conduct Ethnographic Field Study
Ethnographic Field Study Report and/or Interview Transcripts
Review Design History File
Design History File Review
Design History
Create System Map
System Map of Product Development Lifecycle
References
Stakeholder Mapping
2D/3D Mapping
FMEA
Stamatis [32]
Ethnographic Field Study
Field Immersion

Figure 8: Phase 2 Mapping UML Activity Diagram

Key Artefact	Method/Guidelines	Requirements satisfied	Relevant AI Act Articles
Stakeholder Map	2D/3D Mapping	FR2, FR3	-
FMEA	Stamatis [32]	FR5, FR7, FR10	-
Ethnographic Field Study	Field Immersion	FR6, FR9	-

Table 2: Summary of Key Artefacts and Methods for Phase 2: Mapping

5.3 Phase 3: Artefact collection

The artefact collection stage is a part of the audit framework, where all relevant documentation from the product development process is identified and collected. This phase ensures the adherence to the principles of "Responsibility & Accountability" and "Transparency" as declared by the organisation [29]. The collection of these artefacts allows auditors to prioritise opportunities for testing and to ensure that the necessary documentation is available to perform an audit. The key deliverables at this stage include the design checklist, datasheets, and model cards.

The design checklist is an inventory method to ensure that all expected documentation generated during the product development cycle is available. It verifies if the scope of the expected product processes and the all necessary documentation are completed before the audit review begins. It also serves as a procedural evaluation of the development process, ensuring that appropriate actions were taken throughout the development of the system. This check guarantees that the system documentation is robust, providing a solid foundation for the audit.
Examples of items on the design checklist might include:

- **System architecture diagrams:** These provide a visual representation of the system's components and their interactions, which is important for understanding the overall system design.
- **Version control records:** Documentation of changes made to the system over time, making sure that the evolution of the system is traceable.
- **Requirements specifications:** Documents that outline the functional and non-functional requirements the system is intended to meet, helping auditors understand what the system is supposed to do.
- **Test plans and results:** Documentation of testing procedures and results, which allows auditors to assess whether the system has been tested against its requirements.
- **Compliance reports:** Any documentation that shows the system's compliance with relevant regulations or standards, other than the AI Act.

Datasheets are essential tools for making algorithmic development and algorithms themselves more auditable. They offer documentation for machine learning datasets, similar to the datasheets used in the electronics hardware sector. A datasheet typically includes information about the operating characteristics of the dataset, the test results, and the recommended uses. It also covers the data collection process, addressing questions such as the mechanisms or procedures used to collect the data, whether an ethical review was conducted, and how the data set relates to people. This documentation is critical for auditors to assess the quality and suitability of the datasets used in AI systems. The model designed by Timnit Gebru et al. provides the author with important questions to ask, such as "For what purpose was the dataset created?" or "Does the dataset contain data that might be considered confidential?", which help improve the transparency of the system when these questions can be answered positively [12].

The model cards are complementary to the datasheets and provide detailed documentation on AI models. Currently, big platforms such as Google and Salesforce use model cards, and in open-source forms such as Hugging Face, they are the default [11]. They include information on how the model was built, assumptions made during development, and characteristics of model performance in different cultural, demographic, or phenotypic groups. Model cards help clarify the intended use cases of AI models, minimising their usage in inappropriate contexts. They also document the evaluation data, the scope of the model, the risks and factors affecting the performance of the model. A robust model card is key to understanding the trained models' details, making it useful for internal development purposes and ensuring transparency and accountability in AI systems. The model card proposed by Mitchell et al. can be used as a template for AI systems and therefore serve as a guide for auditors [22].

In this phase, a decision point has been added after verifying the completeness of the documents. When all required documents are available, the retroactive documentation step can be skipped. If not, the option becomes available again, as well as the collection or development of the data sheets and model cards. The reason they are displayed independently is because of the importance they have in creating an understanding of the system being audited.

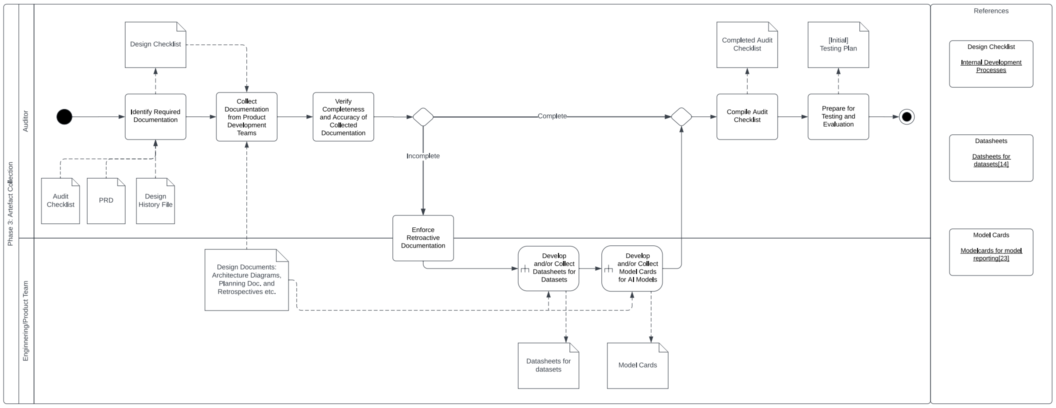
Phase 3: Artefact Collection
Auditor
Enginnering/Product Team
Design Checklist
Identify Required Documentation
Audit Checklist
PRD
Design History File
Collect Documentation from Product Development Teams
Verify Completeness and Accuracy of Collected Documentation
Complete
Incomplete
Enforce Retroactive Documentation
Design Documents: Architecture Diagrams, Planning Doc. and Retrospectives etc.
Develop and/or Collect Datasheets for Datasets
Develop and/or Collect Model Cards for AI Models
Datasheets for datasets
Model Cards
Completed Audit Checklist
[Initial] Testing Plan
Compile Audit Checklist
Prepare for Testing and Evaluation
References
Design Checklist
Internal Development Processes
Datasheets
Datsheets for datasets[14]
Model Cards
Modelcards for model reporting[23]

Figure 9: Phase 3 Artefact Collection UML Activity Diagram

Key Artefact	**Method/Guidelines**	**Requirements satisfied**	**Relevant AI Act Articles**
Design Checklist	Internal Development Processes	FR2, FR3, FR5	Chapter III, Article 9, 11
Datasheets	Datsheets for datasets[12]	FR6, FR9, FR10	Chapter III, Article 11
Model Cards	Modelcards for model reporting[22]	FR6, FR7, FR9	Chapter III, Article 11

Table 3: Summary of Key Artefacts and Methods for Phase 3: Artifact Collection

5.4 Phase 4: Testing

The testing stage is where most of the activities of the audit team are concentrated. During this phase, auditors perform a series of tests to evaluate the compliance of the system with the prioritised ethical values of the organisation. This phase involves direct engagement with the system, producing artefacts that demonstrate the system's performance at the time of the audit. Auditors also review the documentation collected from previous stages to assess the likelihood of system failures to adhere to the declared principles. The approach to testing may vary significantly depending on the organisational and system context, with tests tailored to address the specific risks identified in the FMEA [29].

Adversarial testing is a key deliverable in this stage. It is a method used to identify vulnerabilities by simulating hostile attacks on the system. This involves using specifically made inputs to the model to see if they result in outputs that can be considered undesirable. Adversarial testing is particularly useful for uncovering biases and potential ethical issues within the system. For example, auditors might use counterfactual adversarial examples to confuse the model and find problematic failure modes. This type of test helps identify how the system behaves under stress and whether it can withstand various types of malicious inputs. The results from adversarial testing provide valuable insights into the system's robustness and resilience, ensuring that it can handle real-world scenarios without compromising ethical standards. However, it is important to note that while the audit team may design and oversee these tests, the actual execution of adversarial testing is typically done by the engineering team, with auditors reviewing the results and validating their findings. The RNN-Test framework by Guo et al. can serve as a valuable model for conducting adversarial tests, especially for systems involving recurrent neural networks (RNNs). RNN-Test employs coverage metrics and state inconsistency orientations to effectively generate adversarial input, identifying potential weaknesses in RNN-based AI systems [14].

Another artefact produced during the testing stage is the ethical risk analysis chart. This chart evaluates the likelihood and severity of potential failures. Risks are prioritised according to their probability and impact, with high-priority threats identified as those that are highly likely and severe. The chart categorises each risk with severity indicators such as "high," "mid" and "low." These assessments are informed by the results of adversarial tests, social impact assessments, and ethnographic interviews conducted in earlier stages. The ethical risk analysis chart is closely related to the ethics review conducted during the scoping phase. Although the ethics review provides an initial framework for identifying potential ethical issues, the ethical risk analysis chart builds on this foundation by categorising and prioritising risks based on the system's performance during testing. The ethical risk analysis chart helps auditors and stakeholders visualise and understand the most critical risks, guiding them in implementing the necessary mitigations. Research on this topic has yet failed to provide a model for simple implementation. However, the study by Gao et al. investigates current practices in documenting ethical considerations in open-source AI model documentation, and suggestions are provided to mitigate risks, but often lack concrete and actionable advice [11].

The testing phase includes a key decision point after having performed the test itself. Practise has shown that the plan, do, check, act cycle in leading in an audit, meaning that often new data and facts emerge in the testing phase, requiring an alteration in scope, and repetition of the testing phase. This has created the requirement of the go or no go decision point.

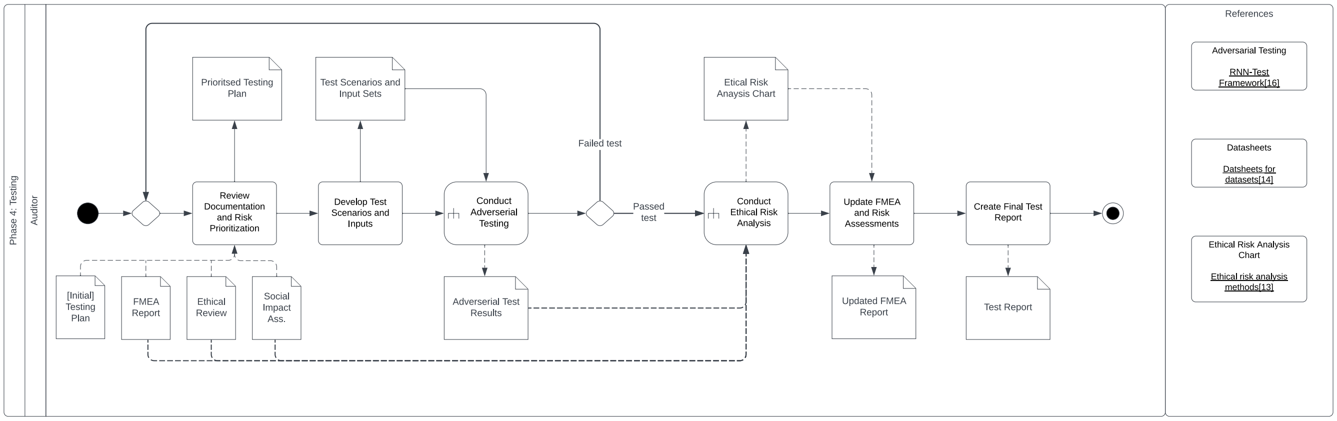
Phase 4: Testing
Auditor
Prioritsed Testing Plan
Test Scenarios and Input Sets
Failed test
Etical Risk Anaysis Chart
References
Adversarial Testing
RNN-Test Framework[16]
Review Documentation and Risk Prioritization
Develop Test Scenarios and Inputs
Conduct Adverserial Testing
Passed test
Conduct Ethical Risk Analysis
Update FMEA and Risk Assessments
Create Final Test Report
Datasheets
Datsheets for datasets[14]
Ethical Risk Analysis Chart
Ethical risk analysis methods[13]
[Initial] Testing Plan
FMEA Report
Ethical Review
Social Impact Ass.
Adverserial Test Results
Updated FMEA Report
Test Report

Figure 10: Phase 4 Testing UML Activity Diagram

Key Artefact	**Method/Guidelines**	**Requirements satisfied**	**Relevant AI Act Articles**
Adversarial Testing	RNN-Test Framework[14]	FR6, FR7, FR9	Chapter III, Article 9
Ethical Risk Analysis Chart	Ethical risk analysis methods[11]	FR5, FR7, FR10	Chapter III, Article 9, 15

Table 4: Summary of Key Artefacts and Methods for Phase 4: Testing

5.5 Phase 5: Reflection stage

The reflection stage of the audit is a crucial phase in which the results of the testing stage are analysed in relation to the ethical expectations outlined in the audit scoping. In this phase, auditors review and establish the concluding risk assessment from the test outcomes, pinpointing the exact principles that could be compromised by the AI system when it is deployed. This phase includes an examination of the product choices and design suggestions that could be put into effect based on the audit findings. Key deliverables from this stage include an algorithmic use-related risk analysis and FMEA, a remediation and risk mitigation plan, an algorithmic design history file, and an algorithmic audit summary report.

The algorithmic use-related risk analysis is an in-depth analysis that combines insights from the social impact assessment and known issues with similar models. The analysis should take into account the differences between the mental models of the AI system and the mental models of the actual user, as highlighted by Leveson's work on safety engineering [19]. The goal is to document foreseeable hazards and risks, ensuring that the system's actual use aligns as closely as possible with its intended use. The model proposed by Al-Husseini et al. can be used as a starting point to create the risk analysis [2]. This artefact is important for understanding and mitigating risks that arise from the deployment of AI systems.

After the audit, a remediation and risk mitigation plan is developed to address the identified problems. This plan aims to reduce the risk of ethical concerns or potential negative social impacts to an acceptable level. It should outline the specific actions to be taken by the engineering team to improve the system. The plan is reviewed by both the audit team and the leadership to inform deployment decisions. This artefact is essential for ensuring that the system meets ethical standards and performs reliably across different subgroups.

Inspired by the concept of the medical device industry design history file, the Algorithmic Design History File (ADHF) collects all documentation related to the development of the algorithm [34]. It includes records of key decisions, design changes, and justifications that demonstrate that the product was developed in accordance with the ethical values of the organisation. Serving as a basis, the Hismo model can be applied and combined with the inspiration of the medical device industry variant. The paper by Ühler et al. presents Hismo, a metamodel for understanding software evolution that models history as a first-class entity, and describes several reverse engineering analyses built on top of Hismo [13]. The ADHF forms the basis for the final audit report and provides a comprehensive audit trail. This file is crucial for transparency and accountability, enabling stakeholders to understand the development process and any ethical considerations addressed during the product lifecycle.

The final artefact is the Audit Summary Report. The audit summary report collects all key audit artefacts, technical analyses, and documentation in one accessible location. This report compares the audit findings with the ethical objectives and engineering requirements outlined in the scoping phase. It serves as a comprehensive evaluation of the AI system, summarising the results of the audit, and providing clear recommendations for future actions. The exact format of this report has not yet been designed, but the methods and guidelines described in this thesis serve as an initial framework. This report will be the final important document to communicate the audit results to

stakeholders and to ensure that the AI system meets the required ethical standards.
The final adjustment to the model is the decision point that emerges after the final risk analysis. The risk analysis will, as said, identify the needs for mitigation, but that could also mean that a mitigation plan is not necessary. Therefore, add the option to skip the creation of it.

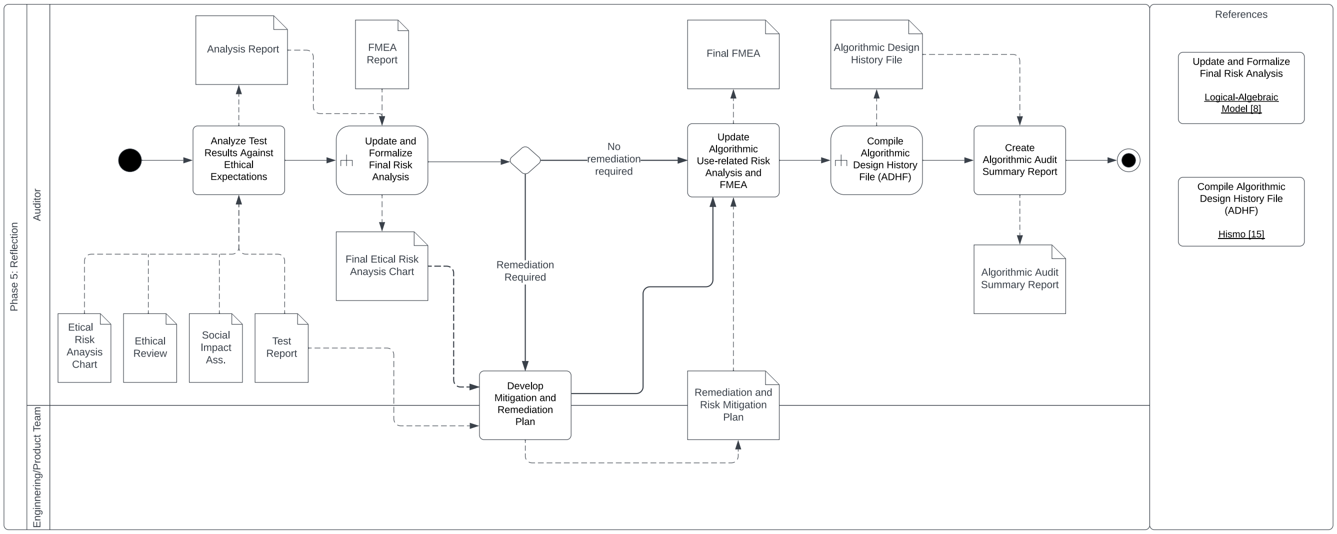
Phase 5: Reflection
Auditor
Enginnering/Product Team
Analysis Report
FMEA Report
Analyze Test Results Against Ethical Expectations
Update and Formalize Final Risk Analysis
Final Etical Risk Anaysis Chart
Etical Risk Anaysis Chart
Ethical Review
Social Impact Ass.
Test Report
No remediation required
Remediation Required
Develop Mitigation and Remediation Plan
Final FMEA
Update Algorithmic Use-related Risk Analysis and FMEA
Remediation and Risk Mitigation Plan
Algorithmic Design History File
Compile Algorithmic Design History File (ADHF)
Create Algorithmic Audit Summary Report
Algorithmic Audit Summary Report
References
Update and Formalize Final Risk Analysis
Logical-Algebraic Model [8]
Compile Algorithmic Design History File (ADHF)
Hismo [15]

Figure 11: Phase 5 Reflection UML Activity Diagram

Key Artifact	**Method/Guidelines**	**Requirements satisfied**	**Relevant AI Act Articles**
Algorithmic Use-related Risk Analysis	Logical-Algebraic Model [2]	FR2, FR4, FR9	Chapter III, Article 9
Remediation and Risk Mitigation Plan	-	FR5, FR7, FR10	Chapter III, Article 9, 15
Algorithmic Design History File (ADHF)	Hismo [13]	FR6, FR8, FR9	Chapter III, Article 11
Algorithmic Audit Summary Report	-	FR3, FR5, FR7	Chapter III, Article 17

Table 5: Summary of Key Artifacts and Methods for Phase 5: Reflection

6 Evaluation

6.1 Confirmatory Interviews

The evaluation of the model has been carried out through confirmatory interviews with two audit professionals in the KPMG IT Audit department after completion. Their responses have already been processed in the models currently present in this thesis.

For the evaluation, the interviewees were asked to review the model (a printout of all diagrams was provided) and to evaluate it in light of their own professional experience, where the main questions were "Would this model be a correct real-world representation of an audit process?", "Would this model benefit you as an auditor doing AI audits?", and "what alterations would you make to the the model to create a better and workable model?". Notes were taken during the interview, and the audit professionals had the opportunity to draw and notate on the provided model printout.

These are the observations and remarks of the first interviewee, who has been employed at KPMG since 2022 as a Consultant.

Regarding the visual model itself, there are a few corrections that could be made according to the consultant. First, all artefact flows should be represented with dotted lines instead of filled lines for the sake of clarity, as that is the correct UML representation, instead of filled lines. It would also be an improvement if the underlying methods and guidelines were referenced immediately in the model. This makes the model itself useful for an auditor, without having to scan through the thesis every time.

For the process, the consultant made some more recommendations on the flow and key decision points. In the mapping phase, he noticed that the stakeholder mapping was only an auditor's task, while practice shows that this is done toghether with the client's teams. Therefore, the stakeholder mapping document should be on the split of the auditor and the product team, suggesting joint development. Within the testing phase, a decision point should be added between "Conduct Adverserial Testing" and "Conduct Ethical Risk Analysis". This decision point should represent, if a test has failed or not, the return to the beginning of the process. As practice has shown, many times a test would fail as the data is incomplete and has to be gathered before a complete test can be run again. In the reflection phase, he has made the suggestion to include a decision point before developing a mitigation and remediation plan, as an audit could very well be successful and therefore deem such a plan unnecessary. The comment was also made that when such a plan is made, it should not affect the audit results, as a remediation plan does not immediately fix the found issues and requires an audit again after implementation.

The consultant has also noticed some limits to the current model. First, the model does not state how many times an audit is to take place or to make a process step more concrete. He states the "Gather Required Information" process step, what is the actual required information? How deep should you go before it is no longer considered required. Finally, it should be noted that an audit is a process that requires multiple roles. It is not just one person, but most of the time a large team, with each team member having different roles within the audit process. For example, a partner at

KPMG would oversee the whole process and rely on the findings and conclusions of the consultants, who do the actual digging and testing itself, to come to a conclusion.

He ends with the statement that he does see the value of this model and that it is a good basis for an internal auditor who wants to set up an AI governance system to prepare for internal and external audits. The process and visual recommendations have been applied to the improved model. The limitations will be discussed in the limitations and discussion section.

The second audit professional to be interviewed has been employed at KPMG since 2019 as a Senior Consultant.

The consultant opened with the statement that an audit could be viewed in the representation of the plan, do, act, check repetitive circle. This is a key component in an audit, as often in checks, new information comes up, and the audit process has to be re-planned or the scope has to be adjusted. With this, he identified a limitation in the model. The model assumes that everything that is required to perform an audit is present, and it does not take into account the fact that an actual audit process is not as straightforward in practise as it is on paper. However, he does acknowledge that it would be impossible to create a model that would factor in all those uncertainties.

Continuing on the assumptions that are made in the model is that the model assumes that the General IT Controls have been audited as well. The consultant states that within KPMG, the controls consist of 4 pillars. Knowing who has access to what and why, knowing the programme changes, knowing computer operations, and lastly knowing programme development. Only when these pillars have been checked and known could an auditor proceed with an AI Audit. The absence of general IT controls can be considered as a limitation, but it was mostly out of scope for this thesis. However, acknowledging that general IT controls have to be applied before performing an AI Audit, do complement the AI Audit framework that has been established in this thesis.
He stresses that an audit is about reasonable assurance, meaning that not everything has to be checked, as that would be doing everything all over again, but that only as much has to be checked to be able to reasonably conclude that the whole of the system is correct.

6.2 Summary of Evaluation

Feedback from audit professionals highlights both strengths and areas for improvement in the model. The model is praised for providing a structured and practical framework for AI audits, making it a useful starting point, especially for those new to this area.

However, the main limitation of the model is its lack of flexibility to accommodate the iterative and unpredictable nature of real-world audits. The linear approach does not account for the need to re-plan or adjust scope when new information arises. Furthermore, the assumption that General IT Controls have already been audited is an oversight, as these controls are essential to any audit process.

Furthermore, certain steps, such as “Gather Required Information”, are considered too vague, raising questions about the depth of the audit. The model also does not address the distribution of roles within an audit team, which could limit its practicality in larger team settings.

In summary, while the model is a solid foundation, it would benefit from revisions to increase its flexibility, detail, and applicability in real-world audit scenarios.

6.3 Limitations

While this thesis makes important contributions to the field of AI auditing and governance, certain limitations were identified. One major limitation is the broad scope and often vague language of the AI Act, which poses as a challenge in defining exactly what should be audited and when. Furthermore, the research framework has not yet been subjected to practical real-world testing, which would be necessary to fully validate its effectiveness.

Additionally, insights from the interviews with the audit professionals revealed several specific limitations in the current model:

- **Clarity and Detail**: It was noted that the model does not clearly define the required information at each step of the process, such as “Gather Required Information.” The depth and extent of the information needed are not specified, which can lead to inconsistencies in the audits.

- **Role Specification**: The model does not account for the various roles involved in an audit process. Audits are typically conducted by a team with diverse roles, such as partners who oversee the process and consultants who perform the detailed work. This lack of role specification can hinder the practical application of the model.

- **Iterative Nature of Audits**: It was highlighted that the model assumes a linear process, while in practice, audits often require iterations and adjustments as new information comes forward. The model does not sufficiently account for this iterative nature, for example, no arrows returning to previous phases, making it less reflective of real-world auditing practices.

- **General IT Controls**: The model assumes that General IT Controls are already audited, but it does not explicitly integrate these controls into the AI audit framework. As Max pointed out, AI audits should build on established IT controls, covering access, programme changes, computer operations, and program development.

7 Conclusions and Further Research

7.1 Conclusion

This thesis has investigated the creation of artificial intelligence audit practices within the context of the European Union AI Act. The primary objective was to develop a structured framework for auditing and governing AI systems to ensure their ethical, transparent, and accountable deployment.

The findings of this research have directly addressed the main research question by identifying significant gaps in the AI Act audit guidelines, existing frameworks such as CapAI, and proposing a new framework. This framework incorporates risk assessment, continuous monitoring, and procedural guidance, which are crucial to align AI deployment with ethical principles and regulatory standards. The literature review highlighted the lack of specificity and procedural clarity in the AI Act, and this research has responded by offering a structured and actionable audit framework. This contribution is significant because it addresses the urgent need for governance mechanisms in the rapidly evolving field of AI technology. The proposed framework serves as a practical guide for auditors and organisations, ensuring that AI systems meet the required ethical and legal standards.

The findings of this research align with and, in some cases, challenge existing theories and assumptions in the field of AI governance. The proposed framework confirms the need for detailed procedural guidance and continuous monitoring, which are widely advocated in the literature. However, it also challenges the broad and often vague language of the AI Act, suggesting that more precise definitions and targeted guidance are required for effective implementation. This research supports the view that regulatory frameworks must evolve to keep up with technological advancements, ensuring that ethical principles and legal requirements are met.

This thesis also underscores the importance of structured audit practices in the ethical and responsible deployment of AI systems. The contributions of this research provide practical solutions that improve the accountability and transparency of AI systems. The proposed framework offers a guide for auditors and organisations, helping them navigate the complexities of AI regulation and ensuring that AI systems are deployed in a manner that aligns with ethical principles and regulatory standards. By addressing key gaps in the AI Act and proposing actionable solutions, this research contributes to the ongoing discourse on AI governance and sets the stage for future advancements in the field.

In conclusion, this thesis has provided valuable insight and practical solutions for auditing AI systems within the framework of the European Union AI Act. The proposed audit framework addresses significant gaps in current guidelines, contributing to the development of more effective and ethical AI governance and audit practices.

7.2 Future Research

Based on the limitations discovered, some suggestions can be made for future work.

- **Real-world Testing and Validation**: Implement the proposed framework in various organisational contexts to assess its efficacy and collect empirical data. This will help refine the framework based on practical insight and real-world challenges.

- **Development of Case Studies**: Create detailed case studies that demonstrate the application of the framework in different industries. This will provide practical examples and further validate the framework's applicability and effectiveness.

- **Tool Development**: Develop software tools or platforms that can automate parts of the AI auditing process outlined in the framework. This can improve the efficiency, consistency, and scalability of AI audits.

- **Iterative Improvement**: Continuously update and refine the framework based on feedback from its application in real-world scenarios and evolving AI technologies and regulations. This will ensure that the framework remains practically applicable and relevant.

- **Interdisciplinary Research**: Encourage interdisciplinary research combining insights from computer science, ethics, law, and social sciences to address the challenges in AI auditing, broader than this thesis has done.

- **Enhanced Clarity and Detail**: Define the required information and depth for each process step to avoid inconsistencies and ensure thorough audits.

- **Role Specification**: Clearly outline the roles and responsibilities within the audit process to improve practical application and collaboration within audit teams.

- **Incorporating General IT Controls**: Integrate the auditing of General IT Controls into the AI audit framework to ensure a comprehensive and robust audit process.

By addressing these limitations and following these recommendations, the model can be improved to serve as a more effective and practical basis for AI auditing and governance.

References

[1] *LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. 2021.

[2] K. Al-Husseini and A. Obaid. Analysis and risk management in software development using the logical-algebraic model. In *CEUR Workshop Proceedings*, volume 2475, pages 241–248, 2019.

[3] A. K. Arslan. A design framework for auditing ai. *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, 7(10), October 2020.

[4] A. by McKinsey. The state of ai in early 2024: Gen ai adoption spikes and starts to generate value — mckinsey.

[5] C. Etteldorf. Edpb on the interplay between the eprivacy directive and the gdpr. *European Data Protection Law Review (EDPL)*, 5:224, 2019.

[6] European Commission. European approach to artificial intelligence, Mar. 2024.

[7] European Commission. Shaping europe's digital future, Apr. 2024.

[8] European Parliament. Artificial intelligence act briefing, Jul 2023.

[9] L. Floridi, M. Holweg, M. Taddeo, J. Amaya Silva, J. Mökander, and Y. Wen. Capai-a procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act. *Available at SSRN 4064091*, 2022.

[10] I. O. for Standardization. Iso/iec 42001:2023. Accessed: 12 July 2024.

[11] H. Gao, M. Zahedi, C. Treude, S. Rosenstock, and M. Cheong. Documenting ethical considerations in open source ai models. (arXiv:2406.18071), July 2024. arXiv:2406.18071 [cs].

[12] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, Dec. 2021.

[13] T. A. Girba. *Modeling history to understand software evolution.* PhD thesis, University of Bern, 2005.

[14] J. Guo, Y. Zhao, X. Han, Y. Jiang, and J. Sun. Rnn-test: Adversarial testing framework for recurrent neural networks. In *arXiv preprint arXiv:1911.06155*. Tsinghua University, Beijing, China, 2019.

[15] Gupta. *Contemporary Auditing.* McGraw-Hill Education (India) Pvt Limited, 2004.

[16] ISACA. Auditing artificial intelligence. Technical report, ISACA, Schaumburg, IL, USA, 2018. Accessed: 2024-07-12.

[17] K. Ittonen. *A Theoretical Examination of the Role of Auditing and the Relevance of Audit Reports.* Opetusjulkaisuja 61, Business Administration 28, Accounting and Finance. Vaasan Yliopiston Julkaisuja, Vaasa, 2010.

[18] E. Kazim and A. Koshiyama. Ai assurance processes. *Available at SSRN 3685087*, 2020.

[19] N. G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety.* The MIT Press, 2016. Accepted: 2019-01-17 23:55.

[20] E. Lohwasser, E. T. Rapley, and L. M. Rousseau. Consequences of us audit standards and practice for foreign jurisdictions: Evidence from the staggered adoption of expanded audit reporting. (4549979), Aug. 2023.

[21] A. Mantelero. *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI.* Springer Nature, 2022. Accepted: 2022-06-20T19:31:27Z.

[22] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA, Jan. 2019. Association for Computing Machinery.

[23] J. Mökander. Auditing of ai: Legal, ethical and technical approaches. *Digital Society*, 2(3):49, 2023.

[24] J. Mökander, M. Axente, F. Casolari, and L. Floridi. Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed european ai regulation. *Minds and Machines*, 32(2):241–268, 2022.

[25] R. Murray-Webster and P. Simon. Making sense of stakeholder mapping. *PM World Today*, VIII(11), 2006. Connecting the World of Project Management.

[26] J. Mökander and L. Floridi. Ethics-based auditing to develop trustworthy ai. *Minds and Machines*, 31(2):323–327, June 2021.

[27] F. of Life Institute. High-level summary of the ai act — eu artificial intelligence act.

[28] K. Peffers, T. Tuunanen, C. E. Gengler, M. Rossi, W. Hui, V. Virtanen, and J. Bragge. Design science research process: A model for producing and presenting information systems research, 2020.

[29] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, number arXiv:2001.00973. arXiv, Jan. 2020. arXiv:2001.00973 [cs].

[30] S. Robertson and J. Robertson. *Mastering the requirements process: Getting requirements right.* Pearson Education, 2013.

[31] S. Sandeep, S. Ahamad, D. Saxena, K. Srivastava, S. Jaiswal, and A. Bora. To understand the relationship between machine learning and artificial intelligence in large and diversified business organisations. *Materials Today: Proceedings*, 56:2082–2086, 2022.

[32] D. H. Stamatis. *Failure Mode and Effect Analysis*. Quality Press, May 2003. Google-Books-ID: OuuiEAAAQBAJ.

[33] E. Tabassi. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards & Technology, USA, Gaithersburg, MD, Jan. 2023.

[34] M. B. Teixeira. *Design controls for the medical device industry*. CRC press, 2019.

[35] D. Vetter, J. Amann, F. Bruneault, M. Coffee, B. Düdder, A. Gallucci, T. K. Gilbert, T. Hagendorff, I. van Halem, E. Hickman, et al. Lessons learned from assessing trustworthy ai in practice. *Digital Society*, 2(3):35, 2023.

[36] R. V. Zicari, J. Brodersen, J. Brusseau, B. Düdder, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Möslein, N. Mushtaq, G. Roig, N. Stürtz, K. Tolle, J. J. Tithi, I. van Halem, and M. Westerlund. Z-inspection®: A process to assess trustworthy ai. *IEEE Transactions on Technology and Society*, 2(2):83–97, June 2021.

A Listing of Requirements and Constraints

A.1 Functional Requirements

- **FR1: Risk Assessment Process**
  - **Fit Criterion:** Include steps for categorising AI systems based on risk levels as defined by the AI Act (minimal, limited, high, and unacceptable).
  - **Rationale:** Ensures systematic identification and management of AI system risks.
  - **Source:** AI Act, Chapter III, Articles 6, 7, 9
- **FR2: Data Quality and Governance Checks**
  - **Fit Criterion:** Integrate steps for data quality checks and governance, ensuring data used is legally sourced and unbiased.
  - **Rationale:** Prevents discriminatory outcomes and ensures the integrity of AI decisions.
  - **Source:** AI Act, Chapter III, Article 10
- **FR3: Documentation and Transparency Mechanisms**
  - **Fit Criterion:** Implement steps for documenting AI processes and providing explanations for AI decisions.
  - **Rationale:** Enhances trust and accountability in AI systems.
  - **Source:** AI Act, Chapter III, Articles 11, 12, 13
- **FR4: Continuous Monitoring and Reporting**
  - **Fit Criterion:** Define steps for continuous monitoring and periodic reporting on AI system performance and compliance.
  - **Rationale:** Ensures ongoing compliance and timely detection of issues.
  - **Source:** AI Act, Chapter III, Articles 13; Chapter IX, Articles 72
- **FR5: Guidance for Human Intervention**
  - **Fit Criterion:** Include a guide outlining scenarios where human intervention is necessary, and steps to document and handle such interventions.
  - **Rationale:** Provides auditors with clear guidelines for human oversight, ensuring compliance with the AI Act's requirement for human control and oversight.
  - **Source:** AI Act, Chapter III, Article 14
- **FR6: Ethical Compliance Evaluation**
  - **Fit Criterion:** Incorporate steps for ethical compliance checks, including privacy, fairness, and non-discrimination.
  - **Rationale:** Ensures AI systems adhere to ethical standards.
  - **Source:** AI Act, Chapter III, Articles 9
- **FR7: Deliverable Generation**

  - **Fit Criterion:** Each major process step should result in a deliverable, such as risk assessment reports, data governance documentation, and ethical compliance checklists.
  - **Rationale:** Provides tangible outputs for each phase of the audit process.
  - **Source:** AI Act, Chapter III, Articles 14, 15; Chapter IX, Articles 72.

- **FR8: Procedural Guidance for Auditing**
  - **Fit Criterion:** Develop a checklist and process description that maps all requirements from the AI Act to actionable steps.
  - **Rationale:** Bridges the gap between the high-level goals of the AI Act and the practical steps needed to achieve them.
  - **Source:** Identified gap in the AI Act, Questionnaire Q21
- **FR9: Material Scope Clarification**
  - **Fit Criterion:** Clearly define the material scope of the AI systems to be audited, including specific technologies and applications.
  - **Rationale:** Provides clarity on what is covered, ensuring effective use of resources by AI providers, auditors, and authorities.
  - **Source:** Identified gap in the AI Act.
- **FR10: Conceptual Precision**
  - **Fit Criterion:** Use precise language to define terms such as 'risk to fundamental rights' and 'prohibited AI practices.'
  - **Rationale:** Ensures clear understanding and consistent application of the AI Act.
  - **Source:** Identified gap in the AI Act.
- **FR11: Procedural Guidance for Monitoring**
  - **Fit Criterion:** Provide detailed instructions on how to carry out monitoring and checks, including frequency and triggers for audits.
  - **Rationale:** Ensures thorough and consistent monitoring of AI systems.
  - **Source:** Identified gap in the AI Act, Questionnaire Q21

A.2 Quality Requirements

- **QR1: Usability**
  - **Fit Criterion:** The framework should include an intuitive interface.
  - **Rationale:** Facilitates easy adoption and effective use by auditors.
- **QR2: Scalability**
  - **Fit Criterion:** The framework should handle audits of AI systems of varying sizes and complexities.
  - **Rationale:** Ensures the framework can be used across different organizational contexts.
- **QR3: Reliability**
  - **Fit Criterion:** The framework should provide consistent and accurate auditing results.
  - **Rationale:** Ensures the dependability of audit outcomes.
- **QR4: Adaptability**
  - **Fit Criterion:** The framework should be easily adaptable to accommodate changes in regulations and standards.
  - **Rationale:** Maintains relevance in a rapidly evolving regulatory landscape.
  - **Source:** Highlighted by the need for continuous improvement in questionnaire Q21_6.

A.3 Constraints

- **C1: Regulatory Compliance**
  - **Description:** The framework must comply with all relevant regulations, including the EU AI Act.
  - **Rationale:** Legal compliance is non-negotiable for the deployment of the framework.

B Use of the Risk Classification Model

Classifying the AI system into the correct risk category is seen as challenging. Although the AI-Act gives several examples for which systems could be in which category, it does not define a categorising process for any AI system [8]. To address requirement FR1 Risk Assessment Process, FR8 Procedural Guidance for Auditing, FR9 Material Scope Clarification en FR10 Conceptual Precision, the following process model can be followed.

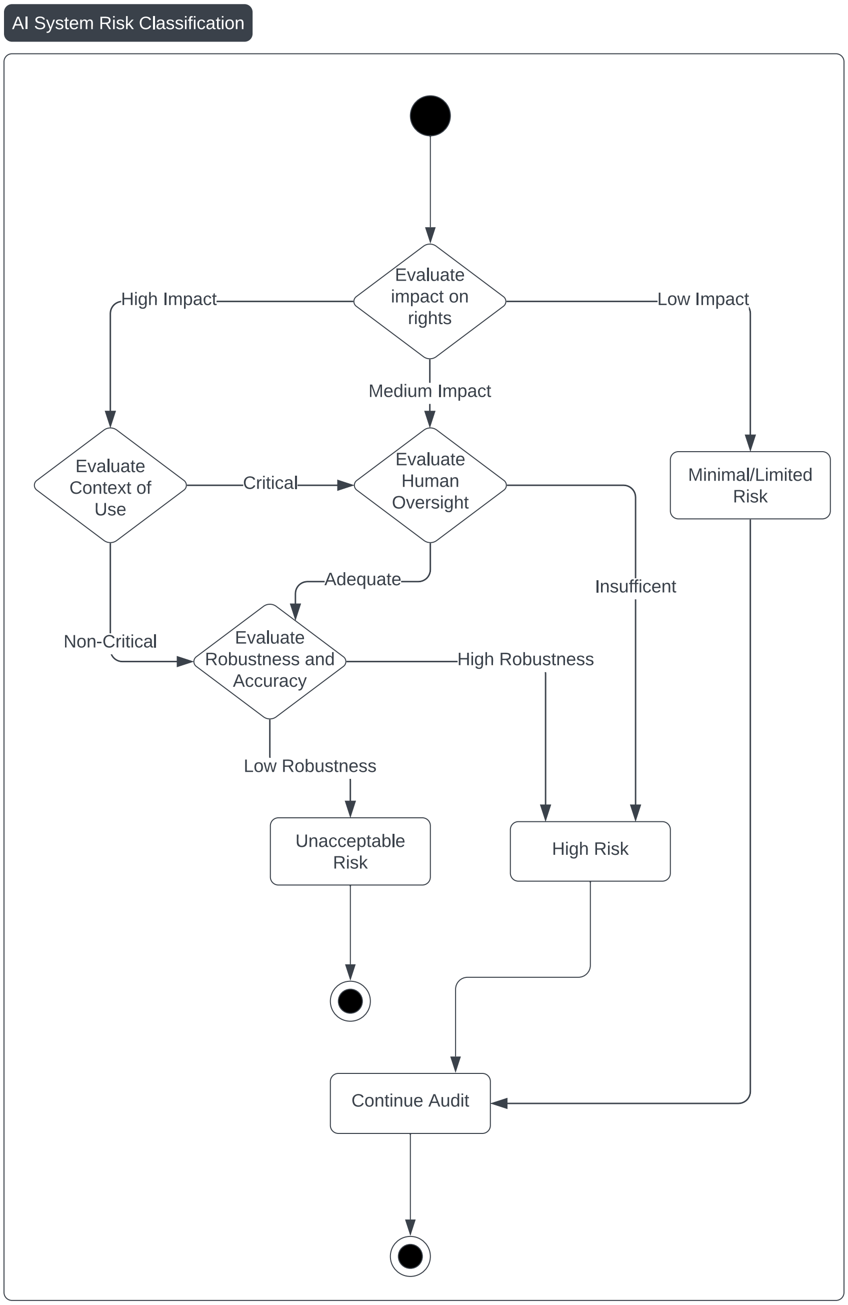

Figure 12: Risk Classification Diagram

However, the model needs to be expanded with some measurable properties to clarify which path should be taken. Below the options are presented, along with examples to facilitate a more informed choice.

1. Impact on Rights

- **High Impact:** The AI system could significantly affect fundamental rights such as privacy, freedom of speech, or non-discrimination.
  - **Example:** AI used in law enforcement, for example real-time surveillance.
- **Medium Impact:** The AI system could moderately affect fundamental rights, potentially leading to biases or privacy issues but not as severe.
  - **Example:** AI used for hiring decisions with some bias mitigation strategies in place.
- **Low Impact:** The AI system has minimal or no effect on fundamental rights.
  - **Example:** AI used for internal business process automation with no direct impact on individuals.

2. Context of Use

- **Critical Context:** The AI system is used in areas where errors could have severe consequences.
  - **Example:** AI in healthcare for diagnosing diseases.
- **Non-critical Context:** The AI system is used in areas where errors are less likely to have severe consequences.
  - **Example:** AI used for personalising marketing content.

3. Human Oversight

- **Insufficient Oversight:** Little to no human intervention is possible or the oversight mechanisms are weak.
  - **Example:** Autonomous AI decision-making in critical systems without human review.
- **Adequate Oversight:** There are strong mechanisms for human intervention and review.
  - **Example:** AI system outputs are always reviewed and can be overridden by human operators.

4. Robustness and Accuracy

- **Low Robustness:** The AI system frequently fails, is prone to errors, or lacks resilience against adversarial attacks.
  - **Example:** AI with high error rates or easily manipulated by malicious inputs.
- **High Robustness:** The AI system is highly reliable, accurate, and resistant to adversarial attacks.
  - **Example:** AI with low error rates, rigorous testing, and security measures in place.

C The complete Audit Framework

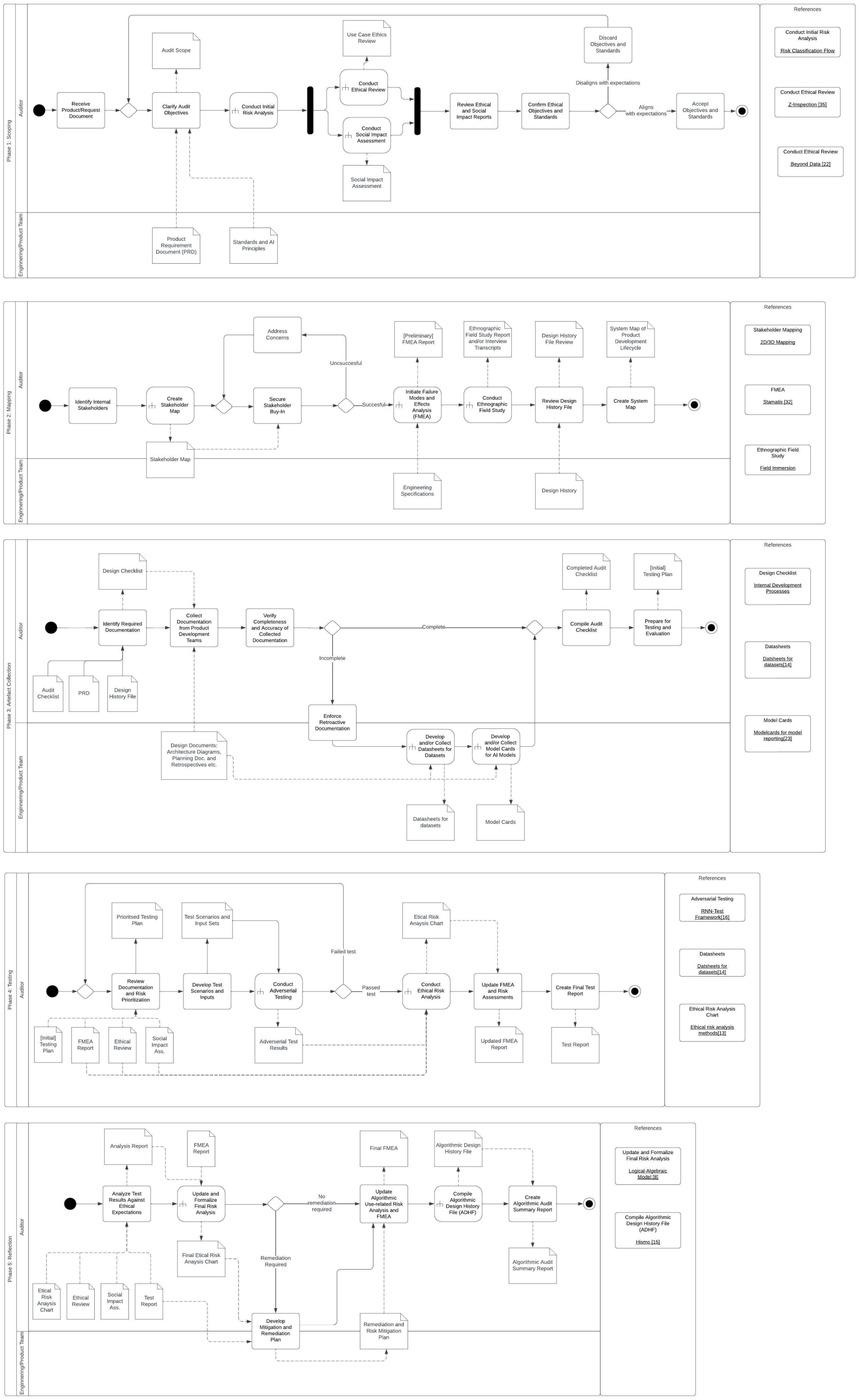
Phase 1: Scoping
Auditor
Enginnering/Product Team
Receive Product/Request Document
Audit Scope
Clarify Audit Objectives
Conduct Initial Risk Analysis
Use Case Ethics Review
Conduct Ethical Review
Conduct Social Impact Assessment
Social Impact Assessment
Product Requirement Document (PRD)
Standards and AI Principles
Review Ethical and Social Impact Reports
Confirm Ethical Objectives and Standards
Discard Objectives and Standards
Disaligns with expectations
Aligns with expectations
Accept Objectives and Standards
References
Conduct Initial Risk Analysis
Risk Classification Flow
Conduct Ethical Review
Z-Inspection [35]
Conduct Ethical Review
Beyond Data [22]
Phase 2: Mapping
Auditor
Enginnering/Product Team
Identify Internal Stakeholders
Create Stakeholder Map
Stakeholder Map
Address Concerns
Secure Stakeholder Buy-In
Uncsuccesful
Succesful
Initiate Failure Modes and Effects Analysis (FMEA)
[Preliminary] FMEA Report
Engineering Specifications
Conduct Ethnographic Field Study
Ethnographic Field Study Report and/or Interview Transcripts
Review Design History File
Design History File Review
Design History
Create System Map
System Map of Product Development Lifecycle
References
Stakeholder Mapping
2D/3D Mapping
FMEA
Stamatis [32]
Ethnographic Field Study
Field Immersion
Phase 3: Artefact Collection
Auditor
Enginnering/Product Team
Design Checklist
Identify Required Documentation
Audit Checklist
PRD
Design History File
Collect Documentation from Product Development Teams
Design Documents: Architecture Diagrams, Planning Doc. and Retrospectives etc.
Verify Completeness and Accuracy of Collected Documentation
Complete
Incomplete
Enforce Retroactive Documentation
Develop and/or Collect Datasheets for Datasets
Develop and/or Collect Model Cards for AI Models
Datasheets for datasets
Model Cards
Completed Audit Checklist
Compile Audit Checklist
[Initial] Testing Plan
Prepare for Testing and Evaluation
References
Design Checklist
Internal Development Processes
Datasheets
Datsheets for datasets[14]
Model Cards
Modelcards for model reporting[23]
Phase 4: Testing
Auditor
Review Documentation and Risk Prioritization
Prioritsed Testing Plan
[Initial] Testing Plan
FMEA Report
Ethical Review
Social Impact Ass.
Develop Test Scenarios and Inputs
Test Scenarios and Input Sets
Conduct Adverserial Testing
Adverserial Test Results
Failed test
Passed test
Conduct Ethical Risk Analysis
Etical Risk Anaysis Chart
Update FMEA and Risk Assessments
Updated FMEA Report
Create Final Test Report
Test Report
References
Adversarial Testing
RNN-Test Framework[16]
Datasheets
Datsheets for datasets[14]
Ethical Risk Analysis Chart
Ethical risk analysis methods[13]
Phase 5: Reflection
Auditor
Enginnering/Product Team
Analysis Report
Analyze Test Results Against Ethical Expectations
Etical Risk Anaysis Chart
Ethical Review
Social Impact Ass.
Test Report
FMEA Report
Update and Formalize Final Risk Analysis
Final Etical Risk Anaysis Chart
No remediation required
Remediation Required
Develop Mitigation and Remediation Plan
Remediation and Risk Mitigation Plan
Final FMEA
Update Algorithmic Use-related Risk Analysis and FMEA
Algorithmic Design History File
Compile Algorithmic Design History File (ADHF)
Create Algorithmic Audit Summary Report
Algorithmic Audit Summary Report
References
Update and Formalize Final Risk Analysis
Logical-Algebraic Model [8]
Compile Algorithmic Design History File (ADHF)
Hismo [15]

Figure 13: Complete audit framework