Auditing Artificial Intelligence

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Management of Technology**

Faculty of Technology, Policy and Management

by

Tijn Sewandono

Student number: 4584627

To be defended in public on 15/12/2023

Graduation committee

Chairperson	:	dr. Y. Ding	ICT
First Supervisor	:	dr. mr. S. Renes	ETI

TU Delft

Executive Summary

Recent technological advancements have enabled the development of increasingly impactful and complex Artificial Intelligence (AI) systems. This complexity is paired with a trade-off in terms of system opacity. The resulting lack of understanding combined with reported algorithm scandals have decreased public trust in AI systems. Meanwhile, the AI risk mitigation field is maturing. One of the proposed mechanisms to incentivize the verifiable development of trustworthy AI systems is the AI audit: the external assessment of AI systems.

The AI audit is an emerging subdomain of the Information Technology (IT) audit, a standardized practice carried out by accountants. Contrary to the IT audit, there are currently no AI-specific defined rules and regulations to adhere to. At the same time, some organizations are already seeking external assurance from accountancy firms on their AI systems. AI auditors have indicated that this has lead to challenges in their current audit approach, mainly due to a lack of structure. Therefore, this thesis proposes an AI audit workflow comprised of a general AI auditing framework combined with a structured scoping approach.

Interviews with AI auditors at one accountancy firm in the Netherlands revealed that the demand for AI audits is increasing and expected to keep growing. Clients mainly seek assurance for management of stakeholders and reputation. Furthermore, the challenges the auditors currently experience stem from having to aggregate auditing questions from a range of auditing frameworks, causing issues in their recombination and in determining question relevancy. Subsequently, design criteria for a general auditing framework as well as feedback on a proposed scoping approach were obtained.

Fourteen AI auditing frameworks were identified through a literature search. Following their typology, these could be subdivided into three source categories: academic, industry, and auditing/regulatory. Academic frameworks typically focused on specific aspects of trustworthy AI, while industry frameworks emphasized the need for public trust to drive AI progress. Frameworks developed by auditing and regulatory organizations tended to be most extensive.

Comparison to four common IT audit frameworks and standards showed that AI audit frameworks need to cover a broader range of topics than the traditional IT audit themes. This is a result of the complex socio-technological context involving multiple stakeholders in which AI systems operate. Additionally, it was shown that AI performance monitoring dashboards could cover technical parts of the audit, but that they fall short when it comes to context-dependent topics such as human oversight or societal well-being.

Following analysis of the similarities between the corporate Environmental, Social and Governance (ESG) reporting materiality assessment and the AI audit scoping problem, an ESG materiality assessment approach was translated to a scoping approach for the AI audit. In this translation, feedback from the AI auditors was incorporated. Combined with a general auditing framework, which was built through combination of the fourteen identified frameworks along the obtained design criteria, this formed the basis for the proposed AI audit workflow. The proposed workflow was demonstrated to be executable through a mock case study. Investigation from the data subject perspective for the Public Eye crowd monitoring AI system of the Municipality of Amsterdam resulted in a scoped list of auditing questions relating to privacy, transparency and fairness.

Recommendations for future AI audit workflow designs include exploring the option of incorporating subthemes in the general framework, closer co-development with AI auditors, obtaining insights from auditors at multiple accountancy firms, and automating parts of the audit.

Acknowledgements

First of all I would like to thank Sander Renes for his supervision over the past months, and pointing me towards the topic of auditing AI in the first place. Each of our meetings has helped me in shaping the research that is combined into this thesis, through insightful discussions and suggestions. I would also like to thank Aaron Ding, especially for his feedback during the Green Light meeting, as this pushed me to enhance the clarity of the thesis for readers who are not actively engaged in the field of AI auditing.

Then I would like to thank everyone at the IT auditing team where have I completed my internship. Not only did I enjoy my time there, I also found support in my fellow thesis interns and colleagues who were able to connect me to relevant people for my research. I would like to thank in particular those who shared their insights during the interviews, lent me their time for feedback on my deliverables, and provided guidance throughout the whole process.

Lastly I would like to thank my friends, family and my girlfriend for always being available for support, from the start till the end of this project.

As this thesis concludes my time at the Delft University of Technology after more than seven years of studying and other related activities, I can say that I am glad to finish this chapter of my life with this thesis, and that I look forward to the exciting times ahead.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms

AI	Artificial Intelligence
CDD	Customer Due Diligence
CMSA	Crowd Monitoring System Amsterdam
CPS	Cyber-Physical System
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSRD	Corporate Sustainability Reporting Directive
CTO	Chief Technology Office
DDSS	Dataiku Data Science Studio
DL	Deep Learning
DPIA	Data Protection Impact Assessment
DPO	Data Protection Officer
DUO	Dienst Uitvoering Onderwijs
EC	European Commission
EDP	Electronic Data Processing
ESG	Environmental, Social and Governance
EU	European Union
GDPR	General Data Protection Regulation
IoT	Internet of Things
ISACA	Information Systems Audit and Control Association

IT	Information Technology
ML	Machine Learning
MoT	Management of Technology
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RL	Reinforcement Learning
SOX	Sarbanes-Oxley Act
TPM	Technology, Policy and Management

Chapter 1 Introduction

1.1 Problem Background

The latest advancements in AI have been enabled in large by big data availability and computing power advancements (Jordan & Mitchell, 2015). Both these drivers have contributed to the development of AI systems of increasingly greater complexity. In turn, the range of potential AI applications has been expanded to more complex tasks, as is illustrated by the latest iterations of generative AI. As AI systems become increasingly complex, they also increase in opacity, i.e. we as humans find it increasingly harder to understand how these AI systems derive their output from a given input. Consequentially, the public perception of AI is not exclusively positive.

Instead, in contrast to its potential utility, AI represents a technology which also worries many people (Kelley et al., 2021). Key factors that have contributed to this are: not being able to understand how AI systems derive their output, the evolving nature of some AI systems making them less predictable, and the unintended but discriminating biases that AI systems can demonstrate (Sandu, Wiersma, & Manichand, 2022). This last reason is reinforced by recent algorithm scandals that have made headline news.

The AI risk mitigation field is developing in tandem with the technological advancements (Zuiderwijk, Chen, & Salem, 2021). AI risk mitigation serves multiple goals. It leads to the development of methods to protect people from potentially dangerous AI systems, while also fostering technological advancement and the widespread adoption of AI systems by generating public trust in AI. A range of solutions have been proposed to enable and incentivize the verifiable development of so-called trustworthy AI (Brundage et al., 2020). Enabling mechanisms relate to technical approaches that for example improve data security robustness. Incentivizing mechanisms are ways to promote AI developers to be diligent in developing AI responsibly. One of these incentivizing mechanisms is the audit of AI.

An audit is a systematic assessment of (originally financial) records or other systems of an organization to verify them for integrity and regulatory compliance. These audits are typically conducted by external auditors who are employed by an accredited accountancy firm. This third-party assessment adds to the objective nature of the audit - although it should be noted that established accountancy firms have historically not been spared from scandals. Similarly, AI systems can also be subjected to external assessment through an AI audit. The AI audit can be regarded as an emerging subdomain of IT auditing (Boer, de Beer, & van Praat, 2023).

The IT audit is a standard practice - statutory for publicly traded companies - which over the past 50 years has co-evolved with technological developments. An IT auditor assesses for a client whether their IT system as well as its governance are compliant with set standards and established rules and regulations. The IT auditor will ask the client a list of predetermined questions, and the client is expected to respond by providing evidence that shows that they are in control of their IT systems. As such, weaknesses can be pinpointed and addressed, thereby mitigating potential risks and at the same time

generating stakeholder trust through fraud prevention (Stoel, Havelka, & Merhout, 2012). A key difference between AI and IT audits is that there are currently no established rules and regulations for AI audits.

The proposed European Union (EU) AI Act is the only announced piece of upcoming AI legislation, expected to come into effect in 2026. The EU AI Act in its current form has, however, been criticized for not concretely defining legal requirements for AI systems (Smuha et al., 2021). Meanwhile, organizations on the forefront of conscious AI development are already seeking external assurance on their AI system from established accountancy firms.

1.1.1 Knowledge Gap

This leaves the AI auditors at these firms facing a challenge. An increasing number of clients is requesting external assurance on their AI system while there are no concrete legal frameworks to be applied or anticipated. As a result, auditors report that their current approach to these AI audits lacks structure. Some AI auditing frameworks have been proposed by a number of institutions as well as in scientific literature. Commonly, the auditors aggregate audit questions from this variety of frameworks into a list of auditing questions for each particular AI audit engagement. This has lead to issues in determining the relevancy of questions as well as in finding a common language, as was brought to light in conducted interviews. **The core problem faced by AI auditors, the problem owners in this setting, is how they should audit AI.** It is this knowledge gap that the research that was conducted as part of this thesis aims to bridge.

1.2 Research Overview

1.2.1 Research Goal

Addressing this knowledge gap requires the design of an AI audit workflow. AI assurance professionals from a Big Four accountancy firm were approached for explorative conversations, as they are the intended product owners of such a workflow. During these conversations, the development of a general auditing framework for AI was proposed as a solution. This framework, which integrates the various existing frameworks and is intended to be applied in combination with a scoping approach, would add structure to and standardize elements of the AI audit process.

Its strengths lie in the standardization of the workflow, which reduces the need for a case-by-case customized auditing approach. Furthermore, the generalized auditing framework would be complete in covering all aspects of trustworthy AI through the incorporation of existing frameworks, proposed by various parties. Meanwhile, the scoping approach would ensure that only the most relevant questions are asked during the AI audit, thereby ensuring that the audit remains manageable and focused on the aspects of the AI system that pose the biggest risks. Potential weaknesses of the proposed solution are that in attempting to cover AI systems in a broad sense, the general auditing framework will not be adequate for specific cases, for example AI systems for niche applications or based in uniquely complex contexts. The combination of these strengths with the fact that the

problem owners proposed this solution lead to the decision to bypass an initial exploration of other solutions to the posed problem. The research goal was therefore formulated to be **the development of an executable workflow for the audit of AI, comprised of a general AI auditing framework in combination with a structured approach to scope the audit**.

1.2.2 Research Questions

Before developing such a workflow, it was first necessary to understand the AI audit from the perspective of the assurance professionals. Their view on the role of the AI audit is relevant as it furthers our understanding of their problem and its context. As the knowledge gaps originates from the current challenges faced in the AI audit, it is deemed important to understand those too. Ultimately, it is desirable that some design criteria and feedback are obtained, which can later be used to develop parts of the AI audit workflow. The first research questions were therefore formulated as follows:

RQ	**1:**	What design criteria and feedback do the assurance professionals prescribe for an executable AI audit workflow?
SRQ	**1a:**	What challenges do the assurance professionals currently experience in their approach to the AI audit?
SRQ	**1b:**	How do assurance professionals perceive the role of the AI audit?

Following the understanding gained from the point of view of the auditors, it is necessary to gain an understanding of the existing AI auditing frameworks. This requires first identifying the frameworks that have been published. To properly understand them, they will then be analyzed and compared to one another, as well as contrasted with related existing practices. For this, the following set of research questions was derived:

RQ	**2:**	What is the state of the AI auditing framework landscape?
SRQ	**2a:**	Which AI auditing frameworks have been published?
SRQ	**2b:**	How do the AI auditing frameworks compare to one another?
SRQ	**2c:**	How do the AI auditing frameworks compare to related established practices?

Once the published AI auditing frameworks have been examined and understood, their auditing questions can be recombined into a general auditing framework using the determined design criteria. Additionally, a scoping method will be developed based on a materiality assessment strategy used to determine which ESG topics an organization should report on. To validate this decision, the ESG reporting topic materiality problem will be compared to the AI audit scoping problem. Then, once the scoping approach has been derived using feedback from the assurance professionals, the complete AI audit workflow can be demonstrated through a case study. Combined this has lead to the third set of research questions:

RQ	**3:**	Can an executable AI audit workflow be constructed around a general auditing framework?
SRQ	**3a:**	Can a general auditing framework be created by recombining the identified auditing questions using the design criteria posed by the assurance professionals?
SRQ	**3b:**	How does the corporate ESG reporting topic materiality problem compare to the AI audit scoping problem?
SRQ	**3c:**	Can the ESG reporting materiality assessment be translated to a scoping method in the audit workflow, incorporating feedback from the assurance professionals?
SRQ	**3d:**	Does the proposed workflow prove effective in a case study?

1.2.3 Research Design

Assurance professionals will be interviewed to determine their view on the current state of AI auditing as well as to identify AI audit challenges and subsequent design criteria for the auditing workflow. Next, a literature search will be carried out to identify AI auditing frameworks. These will be analyzed by comparison to one another, both through a general typology as well as through an in-depth assessment of which topics the audit questions are focused on. A timeline will also be reconstructed to investigate the development of these frameworks over the past years. Additionally, the frameworks will be contrasted with existing IT auditing practices and an AI monitoring platform.

Then, the audit questions obtained from the identified frameworks will be recombined into a general auditing framework following the established design criteria. This will be followed by an analysis of the similarities between the ESG reporting and AI audit scoping problems. Next, the ESG reporting materiality assessment will be translated to an audit scoping approach using feedback from the interviewed assurance professionals. The proposed workflow will then be applied in a case study to show that the derived combination of a scoping approach and general auditing framework can guide an AI auditor to a set of auditing questions. A complete overview of the research design, subdivided into the research questions and described processes, is provided in Figure 1.1.

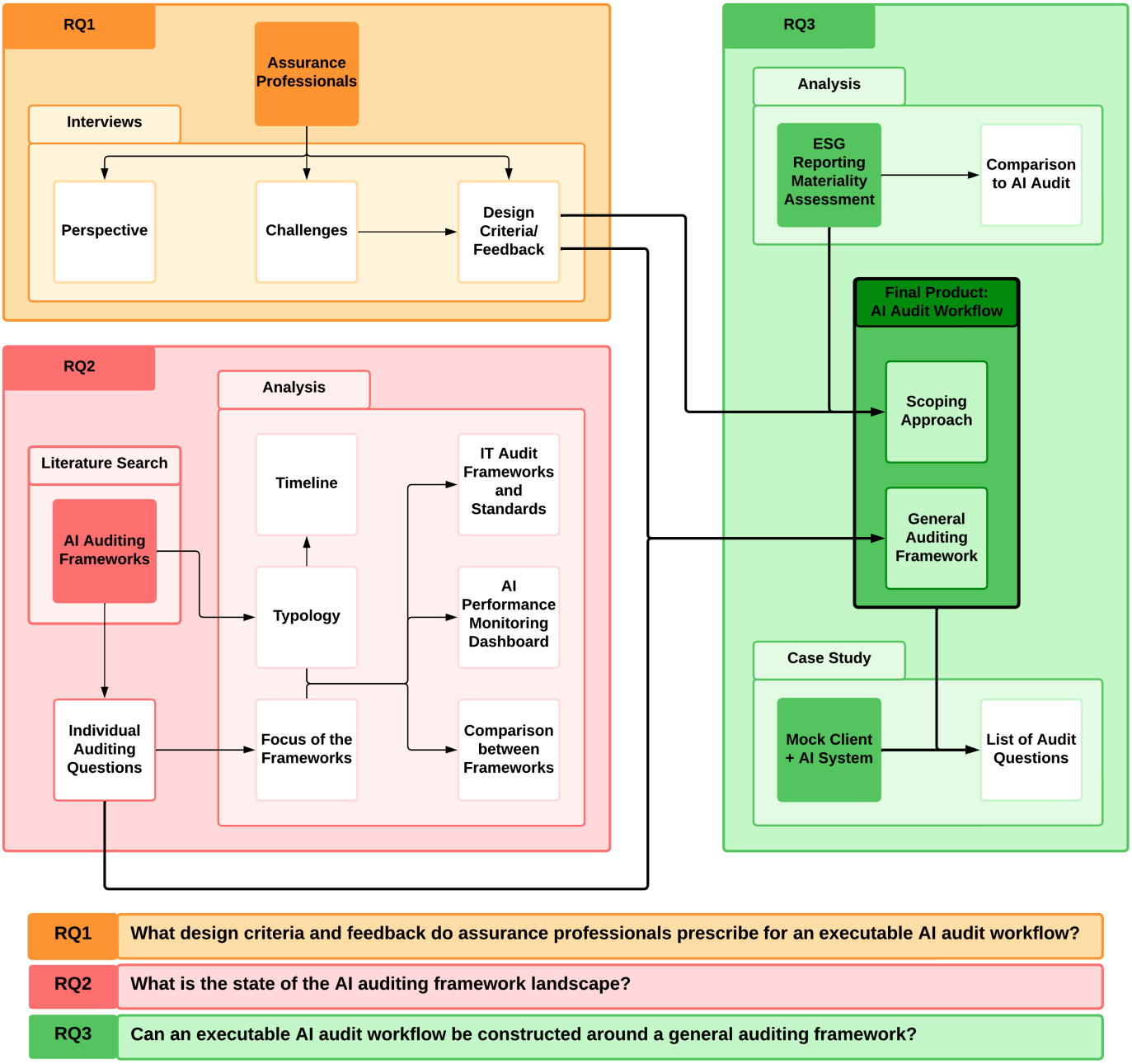


Figure 1.1: *Complete Research Overview.*
Diagram of all research components, subdivided into the posed research questions. The final product will be a complete AI audit workflow comprised of a scoping approach and general framework.

1.3 Relevance

1.3.1 Study Programme Relevance

This thesis is part of the requirements for the Management of Technology (MoT) curriculum at the Delft University of Technology. General thesis requirements have been specified by the faculty of Technology, Policy and Management (TPM), as well as specifically for MoT, all of which will have to be met in order for this thesis to be considered

relevant for the study programme.

The first general requirement is that the work contains an analytical component. This analytical component is present throughout the thesis: for example in the analysis of the auditing frameworks as well as the individual framework elements, the positioning of the auditing frameworks compared to other current practices, or the evaluation of the qualitative data obtained through interviews. Additionally, the work must be multidisciplinary in nature. The AI audit is situated at the cross-section of a variety of research domains, including ethics and stakeholder management, business, and technology - as will also become evident from the typology of AI auditing frameworks. Furthermore, this thesis required the integration of approaches from corporate ESG reporting as well as qualitative data analysis. Lastly, the work should be focused on a technical domain or application, which in this case are AI systems.

The MoT requirements should also be fulfilled. For one, the work has to report on a scientific study in a technological context. This is the case as the goal of this thesis is to develop a method to audit AI systems. Secondly, the work has to show an understanding of technology as a corporate resource or is done from a corporate perspective. This thesis acknowledges the strategic importance of AI for businesses, as it aims to aid in mitigating associated risks by proposing a structured approach to the audit of AI. It is also done from a corporate perspective as the proposed AI audit workflow is intended to be used by assurance professionals.

Lastly, scientific methods and techniques ought to be used to analyze a problem as put forward in the MoT curriculum. The development of a framework and scoping method, as well as applying both in a case study are related to the MoT curriculum as they exist at the intersection between technology and business, and are strongly tied to stakeholders. While building on the knowledge gained from the entire MoT curriculum, specifically the Research Methods and Digital Business Process Management courses have been relevant. Research Methods has provided the tools to carry out explorative research, for example through the coding analysis of large amounts of qualitative data and determining the need for a case study. Digital Business Process Management on the other hand offered project management and problem analysis tools related to business processes. The Digital Business Process Management assignment shared similarities with this thesis in how design criteria were obtained from stakeholders and used to develop a business process.

1.3.2 Scientific Relevance

The scientific relevance of this thesis lies in bridging the knowledge gap concerning the development of a new workflow for the audit of AI. The combination of various frameworks from academic, industry and auditing or regulatory sources into one has not been seen described in literature. Another contribution is connecting the corporate ESG reporting materiality assessment to the problem of scoping the AI audit. The research also further explores the implementation of AI audits through a case study and interviews. Whereas in literature, AI audits have mainly been described conceptually in the context of trustworthy AI. Furthermore, the proposed auditing workflow adds a perspective to the discussion on what constitutes trustworthy AI development.

1.3.3 Societal Relevance

The development of a structured approach for the audit of AI has societal relevance, as the current gap between AI capabilities and AI control, illustrated by algorithm scandals in the media, has caused public distrust. Through application of the developed workflow, those topics that are most material to various stakeholders (i.e. society) can be included in a standardized form of external assessment of AI systems. Additionally, the mitigation of risks through the developed AI audit workflow could protect members of society from being subjected to unfair or otherwise risky AI systems. Ideally, the trust of society in AI systems is also improved through these mechanisms. As public trust in AI is an enabling factor, this could then open the door to further AI advancements, for example in healthcare, which benefit society too.

1.3.4 Business Relevance

First and foremost the development of a structured AI audit workflow is relevant for auditors as they themselves predict the number of AI audit engagements to increase in the future. By adding a level of standardization to both the questions and scoping approach, their practice is thought to become both more consistent and efficient. This will also benefit the organizations that seek assurance on their AI system as the quality of the audit improves. Additionally, the proposed scoping approach could help the audited organization better understand which topics regarding trustworthy AI are most material to their stakeholders. This can in turn be used to tailor their AI system to stakeholder preferences. Lastly, the movement of accountants developing their own auditing approaches for AI systems is expected to also add pressure on lawmakers to propose more concrete rules and regulations than the EU AI Act in its current form.

1.4 Thesis Outline

This thesis consists of six more chapters. The next chapter will provide theoretical background on the IT audit, AI and the current concerns, which combine into the AI audit. This is information that was deemed necessary in order to understand the context of the described knowledge gap and subsequently this sets the stage for the research. Then, in the methodology chapter a detailed description will be provided of the various research strategies that were applied and decisions that were made in order to answer the established research questions. Following the methodology, the results obtained during the research will be presented along with a discussion of the findings in two separate chapters. The workflow development chapter covers all research up until the development of the sub-processes that constitute the final AI audit workflow. Next, the final product and validation chapter covers the complete AI audit workflow and its application through a mock audit. The research findings are wrapped up in the conclusion chapter, in which the results will also be linked back to the initial research questions. Lastly, a reflection on the full research process, the significance of the findings and the context wherein this research was conducted will be provided alongside future recommendations in the final chapter.

Chapter 2 Theoretical Background

This chapter will provide background information that describes the context in which the work of this thesis is situated. The research explores the future of the IT audit in which assurance on the trustworthiness of AI is expected to become an important domain. It is therefore deemed appropriate to first gain an understanding of past developments and the current state of both the IT audit and AI. In turn, that will set the stage for the developments in bridging the two topics as a way to manage the risks currently associated with the rapidly increasing power, prevalence and impact of AI systems (Sandu et al., 2022).

2.1 The IT Audit

The evolution of the IT audit will be explained first, as this enables a deeper understanding of the context in which AI auditing practices are emerging today. The field of IT auditing has shown to be able to adapt to both technological and legislative advancements while retaining its relevant role in the assurance domain.

2.1.1 Origin

IT auditing branched off of conventional financial accounting in the 1960s. The financial audit was at that time an established practice in which accountants would generate a report on whether the financial statements of a business were in compliance with regulatory standards. The financial audit is an objective assessment carried out by external auditors, thereby thought to safeguard the integrity and credibility of the financial reporting. This is essential for assurance to all stakeholders of the audited business on the reliability of the financial information as reported by the business, since the external auditor is expected to have no conflict of interest. The financial audit as such plays a crucial role in the capital market as its outcome is fundamental for the public trust in publicly traded companies (Rezaee, 2004).

Over time, businesses started to rely more heavily on newly available technology to store and process their financial data, as computers became more powerful and less expensive. Additionally, the development of Electronic Data Processing (EDP) software allowed businesses to automate data-related tasks and financial calculations, saving time and preventing human errors (Hafner, 1964). This meant that the financial statements of businesses increasingly relied on figures derived from computer processes. As a consequence, auditors had to familiarize themselves with the technology behind the numbers in order to be able give assurance on the reliability of those figures (Ajao, Olamide, & Temitope, 2016).

Due to the increasing complexity and diversity of EDP software, the need for standards and guidelines for EDP auditing arose. With that, it was required that some auditors specialize in EDP auditing - and thus the new profession of IT auditor came to be. This soon lead to the formation of professional associations for IT auditors such as the Information Systems Audit and Control Association (ISACA, then EDP Auditors Association) in

1969, and the publication of the first guidelines for structured EDP audits (Davis, Adams, & Schaller, 1968).

2.1.2 Development

The IT audit has since then adapted to technological developments. The dawn of the internet for example meant that the IT environment of a business was no longer an 'off-grid island.' Instead, if poorly protected, access to confidential information and control of the IT environment could be obtained through security breaches. This lead to the development of internationally recognized cybersecurity standards which cover both technological and management aspects, such as ISO 27032.

Other developments in IT auditing have been the result of evolving regulatory requirements following major accounting scandals. Most notably both in terms in recency and impact was the Enron scandal of 2001 in the United Sates. Enron was an energy company that achieved rapid growth in the 1990s, and was hailed as "America's most innovative company" by Fortune Magazine for five consecutive years (Bratton, 2002). However, in 2001 stories of fraud were publicly disclosed by whistleblowers and journalists. Through the abuse of loopholes in accounting standards, C-level management actively hid billions of dollars of company debt.

Their accounting firm at the time, Arthur Andersen, did not fulfill its duties as independent reviewer and signed off the financial statements whilst aware of these practices and the associated risks for the company as well as shareholders. This was likely the result of a combination of a lack of independency, rigor and oversight on the side of Arthur Anderson (Linthicum, Reitenga, & Sanchez, 2010). It turns out that while external auditors will declare themselves independent agents, it is in practice nigh impossible for auditors to be fully independent - after all, they are typically hired and paid by the very company that they examine (Bazerman, Morgan, & Loewenstein, 1997).

When the Enron scandal came to light, it lead to the downfall and subsequent bankruptcy of both Enron and Arthur Anderson. The scandal, together with other accounting scandals around the same time (i.e. WorldCom, Tyco), lead to the passage of the Sarbanes-Oxley Act (SOX) in 2002 (Lander, 2002). SOX was proposed to restore trust in the financial system and encompassed measures to prevent fraud, promote transparency, enhance auditor independence and improve internal controls (Gallegos, 2003). Some examples of these measures are: holding company executives responsible for financial reports, prohibiting the auditing firm from offering non-auditing services such as consulting, and mandatory reporting on the effectiveness of internal control on financial statements.

This last measure (Sec. 404 of SOX) had great implications for the IT auditor. It meant that as all financial information is processed, managed and reported through IT systems, it was vital for companies to demonstrate adequate IT controls (Gallegos, 2003). The focus of an IT audit would now also revolve around the design and effectiveness of these IT controls. The IT auditor was tasked with evaluating access controls, change management controls, and segregation of duties for which they rely on evidence provided by the audited firm.

While SOX was an American act, it also indirectly affected European legislation. In the

wake of the American scandals and SOX, the EU implemented a set of similar regulations such as the EU Eighth Company Law Directive in 2006. The Directive also refined the overall audit depth and transparency, thereby affecting the scope of the IT audit to encompass IT controls and risks management in the context of financial reporting. As such it was a way to further harmonize the statutory audit function across EU nations (Tiron-Tudor & Bota-Avram, 2013).

2.1.3 Current State

IT auditing practices nowadays are sometimes conducted outside the context of the financial statements. Three major areas of stand-alone IT auditing are cybersecurity, data privacy, and regulatory compliance (Gantz, 2014). In cybersecurity, IT auditing serves to assess the IT systems and practices in a firm in terms of how well they are protected. The audit provides insights into vulnerabilities and security risks which could compromise the integrity and reputation of a firm when abused. Compliance with privacy regulations has become especially important since the European General Data Protection Regulation (GDPR) came into effect in 2018 (Li, Yu, & He, 2019). The GDPR was created to protect the privacy of EU citizens and outlines how organizations ought to handle personal data. This resulted in the need for businesses to be able to indicate that they do indeed comply with these rules. Lastly, regulatory compliance with set IT standards, be it determined by industry or other regulatory bodies, is essential to guarantee proper IT practices are in place. In each of these cases, an IT auditor can provide assurance (Aditya, Hartanto, & Nugroho, 2018).

To illustrate the IT auditing relevance in the current digital world, a hypothetical example of such an audit will be walked through. Say company A, an online travel agency, seeks to improve the payment process of their online booking platform. Company B, a payment service provider, offers their services and claims to be able to provide a payment platform that meets all requirements as set out by company A. A safe and reliable payment process is critical to the business of company A as well as their relationship with customers and the organizations that are listed on their booking platform. Therefore, company A asks for assurance on the safety and security of the payment platform developed by company B.

Company A and B then hire auditor C, employed by an accredited accounting firm, to perform an IT audit of the payment platform. The findings of the independent IT auditor C will grant both parties information on whether the payment platform is compliant with current industry safety standards as defined by the financial sector as well as any applicable legal requirements. Secondly, the IT audit report will point out areas of improvement related to their payment platform to company B, for example through identification of weak risk management policies. Thirdly, the assurance provided by the audit report can be communicated by company A to their stakeholders to generate trust in the new payment platform (if they opt to make use of the payment platform). The overview of the relationships between these parties is shown in Figure 2.1.

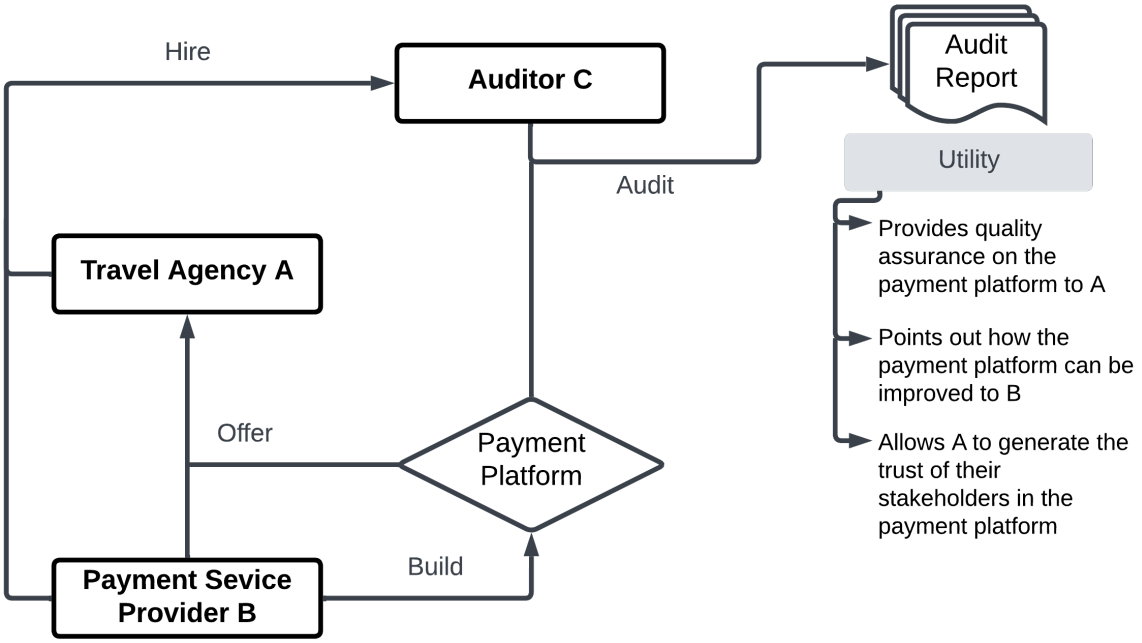


Figure 2.1: *Hypothetical IT Audit.*
Schematic overview of the parties involved in an IT audit.

This hypothetical example illustrates three of the main goals of an IT audit: to provide assurance on the compliance with regulations and standards; to identify risks and point out areas of improvement; and to generate trust from stakeholders (Stoel et al., 2012; Radovanović, Radojević, Lučić, & Šarac, 2010). These goals are evermore relevant as currently industries shift their operations to be increasingly more data-driven, and thereby more IT-dependent, as part of Industry 4.0.

Industry 4.0 refers to the increasing level of automation that can be achieved across industries due to the development of new technologies such as cloud computing, the Internet of Things (IoT), and AI. These new technologies allow for the integration of processes into an autonomous Cyber-Physical System (CPS) which ultimately improves process efficiency and information sharing by interconnecting all subparts of the system. The new role of the IT auditor in this automated industrial world is expected to include the evaluation of automated controls and analysis of process data within these CPSs (Albeda, 2020). It is against this backdrop that the IT audit is anticipated to also branch off into providing assurance on AI systems (Aditya et al., 2018).

2.2 Artificial Intelligence

AI is considered one of the key technologies of Industry 4.0 to disruptively redefine the way manufacturing processes and business models are structured (Peres et al., 2020). Specifically, it is a driver towards higher degrees of autonomy within a CPS (Santos & Martinho, 2020). It will do so by taking over increasingly complex tasks that would otherwise be performed by humans. Before diving deeper into AI, its applications and concerns, however, it is necessary to first establish a working definition.

2.2.1 Definition

Most members of the public are aware of AI (Kelley et al., 2021). Developments in the field of AI have been headline news since the previous decade. Notable examples include AI beating world champions in complex games such as chess or go, autonomous vehicles, and virtual assistants on smartphones. The most recent major breakthrough - the launch of chatbots such as ChatGPT, Bing AI and Bard, which are able to engage with the user and generate texts - has induced a surge of public interest into the technology.

Despite having become a familiar term across the globe, there is no true consensus on what AI entails exactly (P. Wang, 2019). To circumvent a semantic discussion, the definition proposed by the high-level expert group on AI of the European Commission (EC) for AI as a technology will be used:

> Artificial intelligence refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions.

Their definition is designed to be useful to experts and non-experts, and is constructed with discussions on AI ethics and policies in mind (AI HLEG, 2019). Since the audit of AI is related to these ethical discussions as well as policy and regulation, this broad definition is deemed suitable for the purpose of this thesis. For a more detailed overview of main the types of AI and their applications, the reader is referred to section A of the Appendix.

2.2.2 Increasing Complexity

The two main drivers of the rapid developments in AI, big data availability and computing power advancements, have enabled more complex AI systems to be accurately trained and have thereby broadened the scope of possible AI applications (Jordan & Mitchell, 2015; C. Zhang & Lu, 2021). Big data refers to information assets generally characterized by such a high volume, velocity and variety that it requires specific technology and analytical methods for its transformation into value (De Mauro, Greco, & Grimaldi, 2015). The availability and size of big data streams is increasing, through widespread adoption of technologies such as intelligent sensors that are interconnected through the IoT (Jagatheesaperumal, Rahouti, Ahmad, Al-Fuqaha, & Guizani, 2022).

The second part of the definition of big data - the requirement of specific technology and analytical methods - hints at the dependence on advanced computing power in order to valorize the data. Improvements in processing power have been pointed to as one of the main enablers in effectively using big data to train AI systems (L'Heureux, Grolinger, Elyamany, & Capretz, 2017). For example, even if extensive datasets had been available for model training in the past, the limitations in computational capabilities would have significantly hindered Deep Learning (DL) neural networks from attaining their current levels of accuracy and complexity (Hwang, 2018). Edge computing is one of the technolo-

gies currently developed that can enable more effective DL systems by shifting the core of computation from the cloud the the edge of the network, thereby reducing delays and alleviating the network from a data overload (X. Wang et al., 2020).

Applications of big data analytics through AI are diverse. In agriculture and the food industry it has enabled adaptive greenhouse monitoring, drone-based crop imaging, and food quality assessment automation amongst many others (Misra et al., 2022). Integration of AI in city management enables Smart Cities e.g. through data driven prediction and management of traffic or waste collection (Allam & Dhunny, 2019). And in healthcare whole genome sequencing, novel technologies through which vast amounts of genetic data can be generated, has enabled precise AI models to assist in accurate oncological diagnostics (Dlamini, Francies, Hull, & Marima, 2020).

The described progress does come with a trade-off, however. As the AI models grow more complex, generally their opacity increases as well (Angelov, Soares, Jiang, Arnold, & Atkinson, 2021). AI opacity refers to the lack of transparency or understanding in the decision-making processes of AI systems (Burrell, 2016). Opaque models are often referred to as black boxes: all that is truly known are their input and output, with the exact internal logic on how the output is derived remaining unknown (Castelvecchi, 2016).

The opacity associated with the latest generations of AI systems is the root of many of the concerns associated with AI (Burrell, 2016). The internal workings of these models are seemingly incomprehensible, which has raised many questions regarding the risks of (over)reliance on a technology which is on the one hand becoming more capable and valuable while on the other hand becoming harder to understand and therefore to trust (von Eschenbach, 2021).

2.2.3 Perception, Concerns and Risk Mitigation

Perception

A 2021 large-scale international survey (over 10,000 respondents, spanning six continents) in collaboration with Google revealed that most people think that "AI will have significant impact on society, but the overall nature of these effects is not yet determined, underscoring the importance of responsible development and use" (Kelley et al., 2021). The same study points out that the most prevalent sentiment toward AI, besides *futuristic*, is that of *worrying*.

Concerns

This sentiment is the result of multiple factors the make AI systems highly impactful but at the same time difficult to grasp. These can be reduced to not being able to understand how the AI system works, the biases that an AI system can demonstrate, and the evolving nature of some types of algorithms (Sandu et al., 2022).

The previously described opacity of complex AI systems makes it harder for those affected by the output of an AI system to understand how this output was derived (Burrell, 2016). Without an understanding of the considerations taken into account by an AI system, people find it difficult to both judge the AI system output as well as to put their trust

in it (Shin, 2021). This applies to the use of diagnostic AI systems in healthcare, for example. Diagnostic AI systems are being developed that assist healthcare professionals in identifying diseases or other medical conditions, e.g. the detection of early-stage tumor growth from an MRI scan. While the ability to predict the onset of tumor growth of some models is sometimes more accurate than professionals, both professionals and patients are reluctant to accept the AI system output at face-value as the lack of a clear explanation leads to distrust (Y. Zhang, Weng, & Lund, 2022).

Next, the data-intensive nature of AI systems has enabled higher accuracy in complex settings, but also means that existing biases in the data on which a model is trained are thereby ingrained in the model and perpetuated through its output (Ferrer, Nuenen, Such, Cote, & Criado, 2021). This is illustrated by a multitude of cases that have garnered attention from the news. For example, an AI system was used in the US to predict if defendants were likely to become repeat offenders in the future. This prediction would aid in deciding whether or not a defendant was allowed to be on probation. The AI did not rely on information about the defendant's skin colour or race for its input. However, journalists had discovered that black defendants were twice as likely as white defendants to be misclassified as being at higher risk of violent recidivism; on the other hand, white violent recidivists were 63 percent more likely to have been misclassified as having a low risk of violent recidivism, compared with black violent recidivists (Larson, Matt, Kirchner, & Angwin, 2016). Other examples include discrimination in the selection of prospective employees or students, racist biases in advertising, image searches, and price differentiation (Zuiderveen Borgesius & others, 2018).

Lastly, AI systems are not necessarily static once deployed. Instead, the AI system can be updated over time, with new data and in the case of Reinforcement Learning (RL) also through feedback on previous model output. This has in the past opened the door to abuse, for example in the case of the chatbot Tay, which Microsoft launched on Twitter in 2016. The idea behind Tay was that the chatbot would continually update itself based on messages received by other - human - Twitter users, as this would learn to mimic language patterns. Within a day the chatbot was retired as malevolent individuals on the platform had fed it misinformation, slurs and other explicit data. The nature of the messages posted by Tay had shifted towards socially unacceptable content, as that was the type of data it was fed (Bridge, Raper, Strong, & Nugent, 2021).

Risk Mitigation

AI and associated concerns have become a mainstream topic in the current public debate, and examples of AI incidents such as those described previously are just a small representation of many more cases. As such, the field of mitigating AI risks is also developing as part of the AI research agenda (Zuiderwijk et al., 2021). Various routes and combinations of measures have been proposed to enable verifiable development of trustworthy AI. Most of these measures serve to either increase the available options for AI developers to substantiate claims about their AI system, or increase the specificity and diversity of demands that can be made of AI developers by other stakeholders (Brundage et al., 2020).

The more technical approaches towards this relate to either software or hardware. Some examples of proposed technical mechanisms are: high-precision measurements of compu-

tational resource use (Liu et al., 2022); development of secure hardware enclaves dedicated to Machine Learning (ML) tasks (Stoica et al., 2017); development of privacy-preserving ML systems through encryption, differential privacy and federated learning(Al-Rubaie & Chang, 2019); and developing tools and dashboards to increase the interpretability of AI systems (Rudin, 2019).

Equally relevant as the technical approaches are institutional mechanisms, as all these measures support each other and become part of a broad toolbox for AI developers to aid in developing trustworthy AI (Brundage et al., 2020). Institutional mechanisms encompass processes that incentivize AI developers to be diligent in developing AI responsibly. One such mechanism is a push towards openly sharing information about AI incidents through published case studies, which will allow others to learn from past mistakes - as is the case in the aviation sector for example (McGregor, 2021). Anonymized reporting could then be a way to reduce negative effects of incident sharing due to the publicity harming the reputation of the developers.

Two other institutional mechanisms that are similar in nature are the use of either unaffiliated individuals or internal "red teams" to discover and explore AI system limitations and risks before they are taken advantage of (Avin et al., 2021). In IT, "bug bounty" programs are a successful way to entice independent individuals, such as ethical hackers, to report uncovered system weaknesses directly to the developing party through a reward system. These can then be patched before others are aware of the vulnerability. This prevents the exploitation of weaknesses when those who discover it instead spread this information to others who might take advantage of it. Such a reward system could be translated to AI risks, for example by offering compensation to those who report bias or safety issues related to an AI system to the developers.

While bounty programs are set up for people outside the developing organization who have their own motives to look into a system, so-called "red teams" are instructed by the developers themselves to stress test an AI system. These teams of experts take on the perspective of a malicious party with the goal of identifying vulnerabilities (Hua & Belfield, 2020). This allows the developers to then mitigate the discovered risks by patching discovered weaknesses.

Another institutional approach to mitigate AI risks is the audit of AI, a topic that will be explored in the following section.

2.3 The AI Audit

The audit of AI is thought to be a pivotal mechanism to encourage the development of trustworthy AI systems (Brundage et al., 2020; Sandu et al., 2022; Guszcza, Rahwan, Bible, Cebrian, & Katyal, 2018). The AI audit is structured similar to the IT audit. Both are carried out by an external auditor, who will subject their client to an array of questions. These questions serve to assess the control and mitigation of risks associated with an AI system throughout its lifecycle. It is then up to the client to provide evidence to the auditor which demonstrates this, such as documentation of development practices. Effectively, the audit ought to ensure that the audited party is in control of their AI

system and associated risks - thereby contributing to a robust and trustworthy business (The Institute of Internal Auditors, 2018).

The AI audit, however, differs from the IT audit on key areas. While the IT audit has been an established practice for over fifty years with institutionalized standards, laws and regulations in place, there are no official generally accepted or legal guidelines on the audit of AI (Radclyffe, Ribeiro, & Wortham, 2023; Sandu et al., 2022; Albeda, 2020). Additionally, IT auditors typically focus on business integration and IT governance (effective controls, mitigating fraud risks) which does not require a deep technological understanding of the audited IT system (ISACA, 2018). Due to the great variety in types and applications of AI, it is required that the AI auditor has a thorough grasp of the audited AI model on a technological level, as this will partly determine what risks are relevant for investigation (Brundage et al., 2020).

As AI adoption by businesses is increasing steadily (Enholm, Papagiannidis, Mikalef, & Krogstie, 2022), some of the earlier adopters are already looking for ways to obtain third party assurance on their AI system (Batarseh, Freeman, & Huang, 2021). A PwC survey amongst 500 business executives in various sectors (healthcare, finance, tech, media, production) revealed that nearly all participants are prioritizing AI initiatives in the near term, with over 40 % acknowledging the need for compliance in order to build trust (PwC, 2023). The audit of their AI system is expected contribute to that.

2.3.1 Pending Legislation

Compliance would require there to be established rules and regulations to comply with in the first place, which is currently not the case. The first step in this direction is taken by the European Parliament through the EU AI Act, which is expected to come into effect in 2026 at the earliest (Schuett, 2023). The proposal for the EU AI Act was launched in 2021, and its goal is to provide guardrails for the development and use of trustworthy AI in the EU (Tambiama Madiega, 2023). It aims to do so through a risk-based classification of AI systems, where each level is accompanied by mandatory requirements for the providers and users of AI systems in the EU to adhere to. Failure to comply could lead to penalties up to €40M, or 7% of the annual worldwide turnover of a company (whichever is higher) when rules regarding prohibited AI systems are disobeyed.

The EU AI Act proposes four risk categories for AI systems: low, limited, high and unacceptable. AI systems with unacceptable risks according to the draft EU AI Act include systems that exploit vulnerable groups, are used for social scoring or for real-time remote biometric identification. These systems are to be completely banned to protect citizens of the EU. High risk AI systems are defined as either systems that function as safety components in another regulated products (e.g. cars or medical devices), or that are deployed in eight specified areas, including education, law enforcement and employment. These high-risk systems require a conformity assessment prior to the system being placed on the market or put into service, as well as compliance with yet to be defined requirements and harmonized standards (Tambiama Madiega, 2023). The main types of AI system that are included in the limited risk category are generative AI systems and systems that interact with humans, such as chatbots. These systems are only subject to certain transparency obligations, where the human agent interacting with the AI system ought

to be properly informed about the fact that they interact with an AI system. Any AI system that does not fall in the prior categories (e.g. spam filters) is considered low risk and therefore is not subject to specific regulations. An overview of the proposed risk levels and subsequent regulatory implications is presented in Figure 2.2.

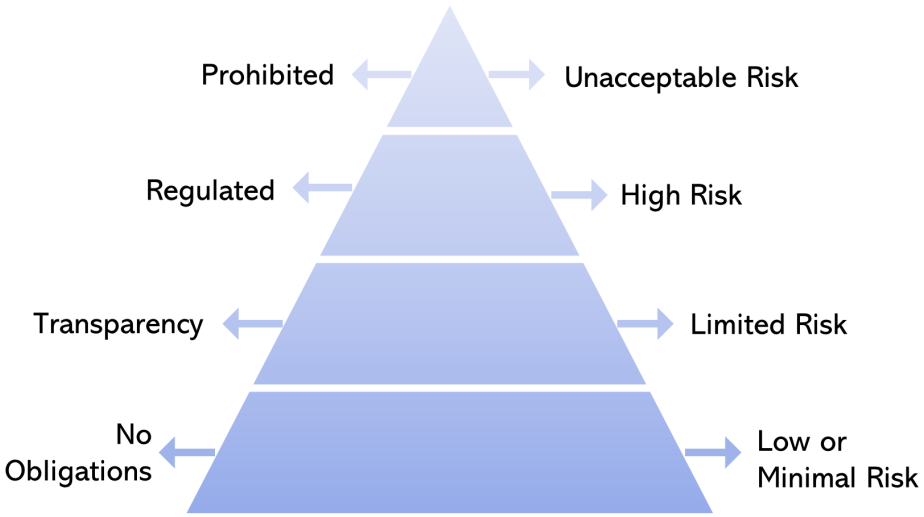


Figure 2.2: *EU AI Act Proposed Risk Levels and Requirements. Adapted from (Tambiama Madiega, 2023).*

The EU AI Act, being the first European legislative proposal for the regulation of AI, is subject to criticism. Critique related to the audit of AI stems from the ambiguity in the proposed regulatory framework. The AI Act in its current form does not concretely define the AI system requirements that accompany the high and limited risk categories, in which most AI systems are predicted to fall (Smuha et al., 2021). The proposed requirements are defined vaguely and leave room for interpretation, for example how "where appropriate, specifications of input data should be (partially) provided to users" (Veale & Zuiderveen Borgesius, 2021). While this ambiguity may be intentional and could be attributed to the AI Act being under development as the discussion on the regulation is still ongoing, this leaves a vacuum for organizations presently looking for a way to obtain assurance on their AI system through an external assessment.

2.3.2 AI Audit Vacuum

The lack of proposed concrete legal requirements for AI systems means that it is currently difficult to anticipate the future regulations for AI system developers and owners. However, organizations that are ahead of the curve in their conscious development of AI are already seeking external assurance on their AI systems. For this they have turned to established accountancy firms such as the Big Four, as the AI audit is arguably an extension of their regular line of work as (IT) auditors. The accountancy firms are interested in providing such assurance services, which is still an emerging practice. They will earn

revenue from the audit engagement while being able to develop their practice in a market which is expected to grow in demand - especially once regulations become established.

Currently, involved auditors describe the AI audit as an unstructured process in which the auditors need to aggregate questions from a variety of sources. This has lead to issues relating to the relevancy of specific questions and finding a common language. Both auditors and the organizations seeking assurance stand to gain from a more structured approach, which includes determining the most important themes to cover in the audit as well as a general auditing framework comprised of questions that match those themes. This insight gives rise to a design challenge for an auditing workflow that can be followed in every AI audit engagement.

Auditors and auditees will consequently benefit from the efficiency introduced by a standard set of auditing questions as well as a method to establish which themes are most relevant to cover in the audit. Importantly, this workflow should not define explicit rules, as this is the role of lawmakers, but should instead offer organizations a way to demonstrate their control over their AI system, verified by an external party. This, in turn, can be communicated to their stakeholders to improve their trust. Additionally, the audit may lead to the identification of opportunities for further AI system improvement.

This call to add structure to the AI audit approach is the starting point for the research conducted for this thesis, which revolves around the development of an executable workflow for the audit of AI, comprised of a general AI auditing framework in combination with a structured approach to scope the audit.

Chapter 3 Methodology

In order to answer the research questions laid out in chapter 1, a combination of a literature search and framework typology, coding of frameworks, comparison to related practices, interviews, and a case study will be used to establish an overview of the current state of the audit of AI, to derive a general auditing framework as well as a method to scope the audit, and ultimately demonstrate the applicability of the workflow. The overarching guide through these steps is the meta-framework proposed by Partelow, which describes the position of frameworks between theory and practice, as well as how they are used and developed (Partelow, 2023).

As such, it is deemed an appropriate framework to structure the methodology of this thesis, which revolves around framework analysis, development and subsequent application. A schematic overview of the meta-framework is shown in Figure 3.1. An important take-away from the meta-framework is how reasoning from observations to frameworks to theory is guided through inductive reasoning, whereas the reasoning from theory to frameworks to their application this done through deductive reasoning. This insight was used to aid in determining which research strategy would be appropriate at various stages throughout the research.

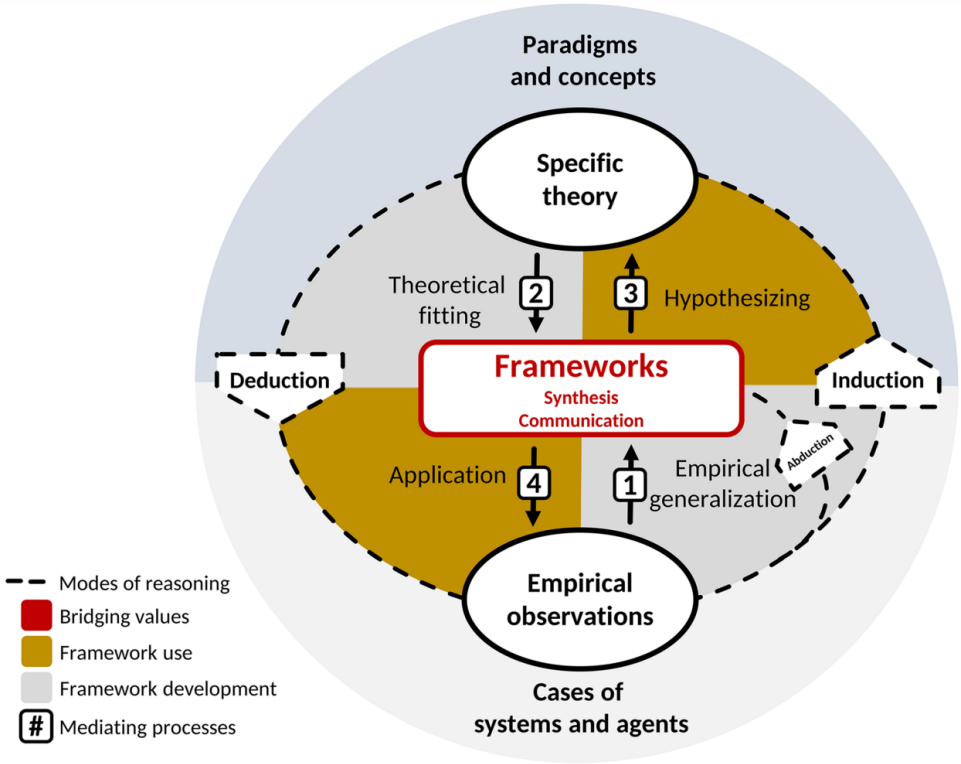


Figure 3.1: *Meta-framework for Frameworks.*
The meta-framework outlines the central role of frameworks in scientific advancement through their development and use (Partelow, 2023).

3.1 AI Auditing Frameworks

The analysis of existing AI auditing frameworks was comprised of three main steps. First the frameworks were identified through a literature search. Next, a typology of the frameworks was established and their individual questions analyzed. Finally, the frameworks were compared to one another, an AI performance dashboard, as well as conventional IT audit frameworks and standards.

3.1.1 Literature Search

The published AI auditing frameworks were to be obtained through a literature search. For the literature search, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were adapted, as this enabled a structured approach to obtain relevant sources (Moher, Liberati, Tetzlaff, & Altman, 2009). The applied literature search protocol consisted of three stages: identification, screening and assessment.

The identification stage was set up as a database search using a list of keywords and synonyms that was iteratively developed. An overview of the final keywords and synonyms is shown in Table 3.1. The databases included in the literature search were Google Scholar and Scopus. These were selected as the combination of the two was expected to achieve a sufficient coverage of reliable sources. In Scopus, the search criteria were used to search within article title, abstract and keywords. For Google Scholar, it was decided to limit the search to only match article titles (by placing "intitle:" in front of each search term in the query). Not doing so would result in tens of thousands of returns, most of which were likely beyond the scope of the literature search. Combinations of search terms were used such that one word of each of the first three rows of Table 3.1 was included, both with and without "Review" as final search term.

Table 3.1: *Literature Search Keywords and Synonyms.*

Keyword	**Synonyms**
AI	Artificial Intelligence, ML, Machine Learning, Algorithm(ic)
Audit	Assurance, Auditing, Assessment
Framework	Guidelines
Review	-

The identified articles would then be screened for relevancy. Their relevancy was determined through a number of inclusion/exclusion criteria. Articles published before 2018 were to be excluded, as at that time the regulatory agenda for AI was still being set (Black & Murray, 2019). Additionally, in 2018 the GDPR was implemented in the EU, which changed the data regulation landscape and with that the context for new (AI) regulations (Mitrou, 2018). Therefore, any literature published prior to 2018 is likely to lack relevancy. Furthermore, the titles of the identified articles were screened to determine whether they covered the audit of AI instead of other topics such as the use of AI in auditing and assurance or other AI applications. Lastly, some articles were not retrievable, behind a paywall, or not available in English, leading to their exclusion. A complete overview of the criteria is shown in Table 3.2.

Table 3.2: *Literature Inclusion and Exclusion Criteria.*

Property	**Included**	**Excluded**
Date of publication	2018 and later	before 2018
Language	English	non-English
Relevance	the audit of AI	use of AI in auditing other AI applications

The assessment stage of the literature search entailed determining whether the literature that passed the screening stage proposed an actual auditing framework. In addition to the database search through Google Scholar and Scopus, a list of frameworks that were referenced in the assessed sources was kept. Some of these referenced frameworks were developed and published through institutions rather than scientific journals, and thereby not retrievable through Google Scholar and Scopus. These frameworks were added to the literature search as they represent a different perspective on AI auditing than the academic frameworks.

3.1.2 Typology and Analysis

The framework typology was created based on the framework positioning factors proposed by Partelow. These are: who the developers are, what values or motives they uphold, what their goal or research question was, and the field in which the framework is embedded (Partelow, 2023).

This initial typology is the basis for a further substantive comparison of the frameworks. This required the analysis of a large amount of qualitative data, for which a coding strategy was developed. In accordance with the meta-framework of Partelow, this strategy involved inductive coding. Inductive coding meant effectively labelling each of the identified elements of the auditing frameworks and assigning them to one or multiple categories. These categories were derived through an iterative process of open and axial coding. The open coding served to determine emergent themes from the raw data, whereas the list of coding categories was refined through axial coding. The frameworks could then be compared to one another based on their typology as well as which codes would most frequently reoccur throughout each of the frameworks.

Selective coding, which would mean further reducing the derived codes to roughly five to seven overarching themes, had deliberately not been carried out. The explorative nature of the qualitative data analysis was expected to loose a level of nuance if all frameworks and their questions were labelled according to a limited number of selective codes. All coding was done using ATLAS.ti qualitative data analysis software.

3.1.3 Framework Positioning

The identified and characterized AI auditing frameworks were compared to an AI monitoring tool as well as IT audit practices. This allowed the positioning of the AI auditing frameworks in relation to those tools and practices. Thereby, strengths and shortcomings of IT auditing practices and the monitoring tool could be revealed within the context of

AI auditing. Additionally, it would provide insights in how monitoring tools such as the analyzed platform could in the future play a role within the AI audit.

For existing IT audit practices, four common IT audit frameworks and standards were identified in literature and subsequently compared to the AI audit frameworks. They were chosen specifically because of their widespread use within IT auditing. Additionally, an AI monitoring platform was suggested for comparison to the frameworks based on its perceived level of maturity and range of capabilities. As it was determined that this platform was amongst the most mature AI monitoring tools, it was deemed suitable for comparison to the AI auditing frameworks.

3.2 General Auditing Framework Development

In order to obtain a single auditing framework, deductive reasoning was followed - as also indicated by the meta-framework when developing a framework through theoretical fitting. The starting point for this phase of the development are the labeled qualitative data from the identified frameworks and the design criteria obtained from the assurance professionals. Insights from interviews with assurance professionals would guide how the general framework should be organized and categorized in order for it to be actionable. The labelled data from the identified frameworks was then recombined such that redundancies were removed and a final, harmonized framework was obtained. This process of fitting the data to the derived insights was conducted iteratively as to ensure that elements from the various frameworks were properly recombined into the general framework.

3.3 Scoping Approach

As scoping the audit was pointed out as one of the challenges faced in auditing AI, it was deemed necessary to also develop a way to determine which parts of the general framework were most relevant for a specific audit case. This problem was expected to share many similarities with the ESG reporting materiality assessment, a process already described in literature (Garst, Maas, & Suijs, 2022). Therefore, it was suggested to attempt to translate the steps identified in that process to a materiality assessment to guide the scoping of the AI audit. In order to also fit this materiality assessment to the perspective of auditors, the materiality assessment was also discussed in interviews with AI auditors in order to obtain their feedback. That way, the materiality assessment could both be further improved and it was ensured that the materiality assessment was aligned with their insights regarding its applicability. No other options were considered for the scoping process as the ESG reporting materiality assessment at face value was considered highly analogous with the AI audit scoping challenge. This would be confirmed through the comparison of the problems at hand in both ESG reporting and AI audit scoping.

3.4 Case Study

Once a general auditing framework and scoping approach were developed and combined into an AI auditing workflow, the final step within the scope of this design process is the

demonstration of the workflow. This is analogous with the demonstration of a prototype at the end of a design cycle, serving as a proof of concept as well as to base recommendations on for the further development of the AI auditing workflow. In the meta-framework this is represented by the application step and subsequent empirical generalization step from observations back to the framework. The goal of the case study was therefore to show that it is possible to use the suggested workflow in order to derive a list of questions to be used in the audit of a specific AI system. The case study was carried out under supervision of an assurance professional for guidance along the process.

To find a suitable AI system for the case study, the following criteria were used. There should be a substantive amount of detailed information available about the system, since that will be the starting point of the audit. Additionally, the system should be affecting stakeholders that are easy to reach out to as their input will be required in the materiality assessment. Lastly, if possible the cooperation of the developers or system owners would be beneficial as they can provide feedback on the final audit questions.

Based on these criteria, it was assumed that an AI system in the public sector would be most eligible. The Algorithm Register of the Dutch Government (Algoritmeregister van de Nederlandse Overheid) was therefore searched for a suitable AI system. Of the 167 entries in the register, the Public Eye AI system of the Municipality of Amsterdam was found to best match the set criteria.

The Municipality of Amsterdam was contacted for collaboration. As they asked for a level of control on which audit findings were to be published, the audit workflow was followed up to the point where a list of scoped audit questions was derived. That way the research findings would not be subject to any external control. The full process of scoping and translating questions from the general auditing framework to the client case could still be carried out. These subprocesses represent the novel aspects of the workflow and it is therefore deemed sufficient - given the limited time and cooperation - to demonstrate the application of these parts of the workflow specifically.

3.5 Interviews

Two sets of interviews were conducted as part of the research. First assurance professionals were interviewed in order to understand their perspective and to obtain design criteria and feedback for the AI auditing workflow. Later, a group of stakeholders was interviewed in order to determine which themes should fall within the scope of the case study. Approval to conduct these interviews was granted by the Human Research Ethics Committee TU Delft, as shown in Figure D.1 of the Appendix.

3.5.1 Assurance Professionals

The interviews with assurance professionals were relevant at multiple stages in the research: providing insights in current AI audit practices and challenges, as well as providing design criteria and feedback for both the scoping process and the development of the general audit framework.

The interview (full list of questions available in the Appendix, subsection D.1) was therefore set up to cover four topics: first the interviewee and their professional experience, then their view on the role of AI audits, next about the audit scoping challenges and solution, and lastly the current challenges in aggregating a list of audit questions and the general auditing framework. The interviews were designed to be semi-structured: the four topics guide the interview from general to detailed questions, with room for discussion between the researcher and interviewee. This semi-structured strategy follows from the explorative purposes of the interviews, as semi-structured interviews offer flexibility to both the participant and researcher to further explore the topics at hand (Knott, Rao, Summers, & Teeger, 2022). Prior to conducting the interviews, the interview protocol was once discussed with one of the participants on a separate occasion in order to obtain feedback on the clarity of the questions. This was done to prevent any potential misunderstandings. As they had indicated that all questions were clear, no alterations were made to the protocol.

The interviewees were three assurance professionals employed at a large accountancy firm that operates in the Netherlands, who were also professionally involved in providing AI assurance to clients. As this practice is novel, without legal guidelines, requiring in-depth knowledge on algorithms, and not commonplace or mandatory (unlike IT audits), only few auditors are currently involved in AI audits. This group of three auditors represents the auditors at this accountancy firm with the most relevant experience for the purpose of this thesis. They were connected with through a graduate internship position at this firm. Since the first language of all parties involved was Dutch, the interviews were conducted in Dutch as to enable the participants to best formulate their answers. The interviews were audio-recorded, transcribed and summarized in an excerpt. The participants were asked to verify the excerpt of the transcript of their interview for correctness and completeness to ensure the accuracy of their statements. This was also conditional for their inclusion in the research.

Product Owner Feedback

As these assurance professionals are the product owners of the developed AI auditing workflow, they were consulted throughout the development and application stages for feedback. This feedback, for example, included whether duplicate questions that fall into more than one of the final seven categories should be kept or placed in only one. Another example is feedback on whether the auditor or the audited party should be in the lead during each of the steps of the auditing process. Lastly, they verified the final design in terms of its applicability when it was used for the case study.

3.5.2 Stakeholder Inquiry

Stakeholders of the AI system were asked to indicate which AI auditing topics they considered most relevant to themselves, and should therefore be part of the scope of the audit. The goal of the interviews was to obtain enough input on the materiality of the auditing topics to be able to conclude which auditing topics would be relevant to include in the scope of the audit. As this is also an explorative investigation, these interviews were designed to be semi-structured to again allow flexibility to further explore topics

that might arise during the interview.

The participants were approached to participate in person at the site where the AI system of the case study was operational. A single afternoon was spent interviewing anybody who was present there at that time and willing to participate. While this constrained the sampling of participants to those who happened to be there at that specific moment in time, a sufficiently sized group of people (20) had cooperated such that the materiality assessment could be performed. People who were working in the area typically did not choose to participate as they prioritized their work.

The interview questions (full list available in the Appendix, subsection E.2) were designed to guide the conversation from the physical cameras at that location towards the materiality of auditing themes. The participants were asked to freely formulate themes they would find important prior to being presented with the seven topics that would encompass the general auditing framework. This allowed post-interview verification of their final ranking of auditing topics. The ranking was performed by asking the participants to order seven flash cards that matched the auditing themes and present them back to the interviewer. That way the risk of the participant losing track of the seven newly presented themes to them was mitigated. As indicated on the consent forms presented to participants, the interviews were audio-recorded, transcribed and deleted following the aggregation of the interview data in order to minimize stored personal data.

Chapter 4 AI Audit Workflow Development

4.1 Current AI Audit Practice

Interviews with AI assurance professionals at a big four accountant in the Netherlands revealed insights about the AI audit as well as design criteria for the AI audit workflow. An overview of the participants and their relevant experience, i.e. time at current employer working on AI related client engagements, is shown in Table 4.1. The interview excerpts are included in the Appendix, subsection D.1.

Table 4.1: *Overview of Interviewed AI Assurance Professionals.*

Ref.	Relevant Experience
I1	Five years in AI and data assurance and advisory roles
I2	Two years in AI assurance and advisory roles
I3	Two years in AI assurance and advisory roles

Reference ID will be used throughout this thesis to refer to the interviewees when used as source.

The interviewees noted that clients have only recently begun to solicit their assurance services for AI specifically, as each has worked on roughly three of these engagements. As one mentioned (I2): "Especially since the rise of generative AI, this demand has increased rapidly."

For all clients, the motivation for an audit of their AI system largely revolved around management of stakeholders and reputation (I1, I2, I3). The clients communicate the external assurance provided by an accountant to internal stakeholders as to demonstrate the trustworthiness of their AI system. Additionally, incidents that could harm the reputation of the client can be prevented through external assessment. It was also noted by the interviewees that their reputation as large accounting firm played an important role in adding value to the audit report findings (I1, I3).

Furthermore, a distinction could be made between clients from the public and private sector. Clients from the private sector were described to be more concerned with reputation management, as illustrated by a client who wished to prevent AI related incidents following one in the United States (I1). The use of algorithms in the public sector in the Netherlands has lately been under a lot of scrutiny, which is why public sector clients typically requested assurance in order to preempt public inquiry regarding their use of AI (I2). To a lesser extent the audit findings also had additional value to clients by serving as guidelines for the improvement of their AI system (I1).

Considering their opinion on the contribution of the AI audit to trustworthy AI, it should be noted that the interviewees were essentially asked to reflect on the relevance of their own profession. As such, their view could display a bias that favours the added value

of AI auditing. Nevertheless, since they are at the forefront of the developments in AI auditing, their opinion is relevant.

The interviewees agree that the AI audit is an important contributor to trustworthy AI, specifically as an incentive for risk mitigation (I1, I2) and to improve documentation and decision-making (I2, I3). They predicted that once laws and regulations come into effect, there will be an even higher demand for AI assurance services (I1, I2, I3), with stricter regulations resulting in greater importance of the audit (I1). Additionally, AI incidents in the news are also expected to affect this importance (I2), especially for the organizations in the public sector (I3). Furthermore, it was mentioned that certain platforms or dashboards could handle the monitoring of AI system performance and other technical aspects, thereby enabling the automation of a portion of the AI audit (I2, I3).

The current approach to AI audits could be mapped out based on the interview responses. The following steps and processes were identified: In the first stage the auditor and the client identify relevant regulations (e.g. GDPR) and industry-specific risk areas (I1, I2). Then the auditors will aggregate questions from existing AI auditing frameworks based on themes that were established with the client (I1, I3). These questions can be adapted to industry standards if needed (I2). The client is then asked to provide evidence for each question, after which all findings are combined in a report.

This process is described as unstructured (I3), and the interviewees experienced challenges in determining the relevancy of questions from various frameworks (I1, I2) as well as in establishing a final auditing framework by aggregating these questions (I3). While scoping the audit as it is currently done is not seen as a challenge by some (I1, I2), all did see merit in the proposed materiality assessment approach as it enabled the inclusion of more perspectives in the audit while adding structure in combination with the general auditing framework.

The proposed solution for the current challenges in AI assurance engagements is to use a general auditing framework consisting of auditing questions that cover the full range of themes related to trustworthy AI as proposed by the EC (I3), while allowing relevant questions to appear in multiple categories (I1, I2). Through a materiality assessment the framework can then be reduced to a list of most relevant questions that are to be reported on during the audit. Specific points of feedback on the scoping approach will be referenced in the related section (subsection 4.3.2).

4.2 AI Auditing Framework Analysis

4.2.1 Literature Search

A literature search was carried out following the identification, screening and assessment steps as described in subsection 3.1.1 of the Methodology chapter. Through this process, the 2000 documents that were initially identified could be reduced to 46 that covered the audit of AI. From the 40 studies that could be retrieved, nine auditing frameworks were obtained. Additionally, seven other frameworks were referenced in literature from which five were included based on the same accessibility and inclusion criteria. The complete overview of the literature search is shown in Figure 4.1.

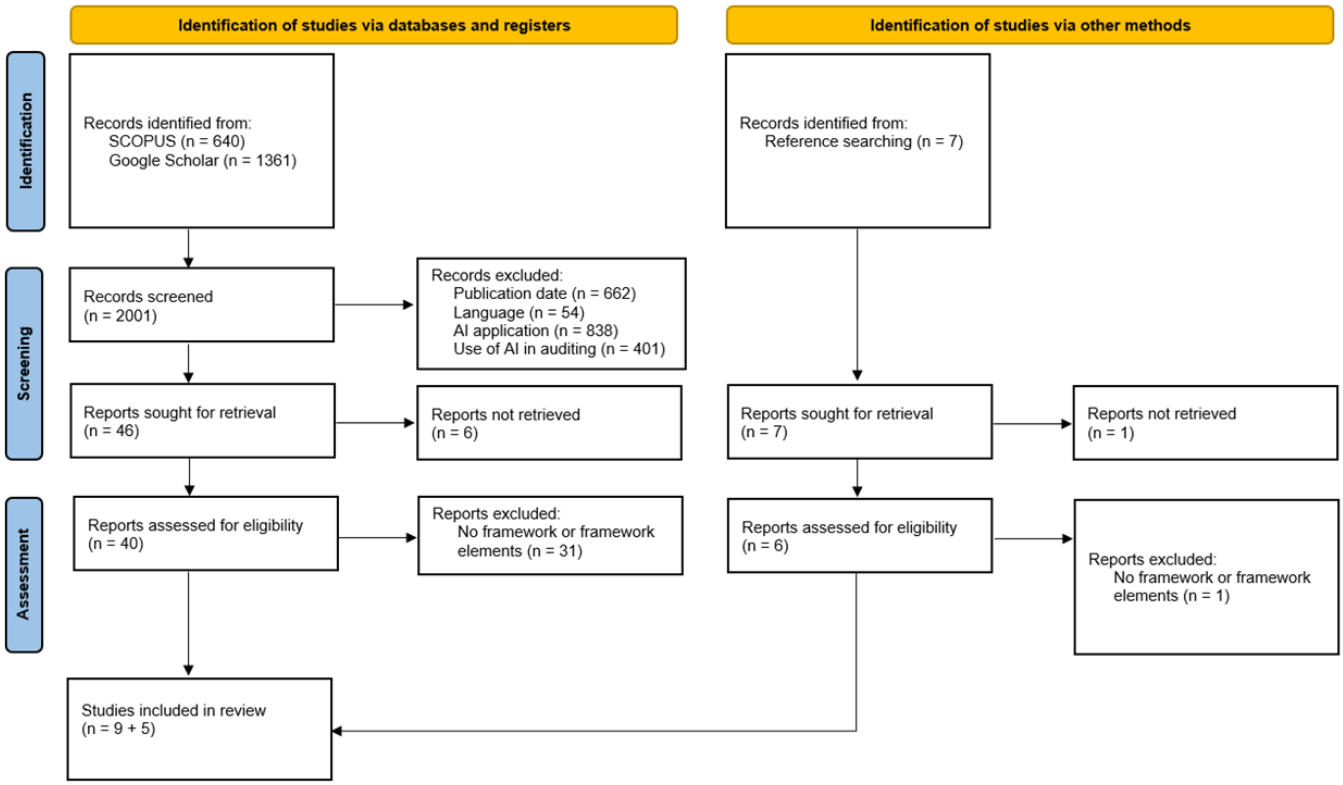


Figure 4.1: *Literature Search Results.*
Flowchart showing the stepwise filtering and selection of relevant frameworks for the audit of AI.

The literature search revealed that AI is a contemporary subject in the field of assurance, as most studies were published within the last five years. The exclusion criteria further showed that the focus of the assurance literature lies within the application of AI in current financial or IT audit practices. These studies revolve around the use of AI for task automation and decreasing the likelihood of human errors during audits (Hasan, 2022). Lastly, most studies that did cover the topic of assurance on AI systems were focused on proposing (high-level) initiatives for the advancement of trustworthy AI, but did not contain a framework to achieve this.

4.2.2 Framework Typology

A typology of the fourteen obtained frameworks was constructed; the frameworks and data related to their developers, values and goals are listed in Table 4.2. The typology revealed three distinct categories of framework based on their source type, namely those published by academic sources, by auditing or regulatory bodies, and by industry (i.e. organizations that develop AI systems).

Table 4.2: *Typology of Identified Frameworks.*
Reference label will be used to refer to the frameworks throughout the thesis.

Ref.	Framework	Year	Affiliation	Values/Motive	RQ/Goal	Field/Sector	Source type	Source
F1	Assessment List for Trustworthy AI	2020	Independent European Expert Group	Protection of fundamental human rights	Provide an initial approach to self-assess adherence of an AI system to guidelines for trustworthy AI, intended for flexible use.	multidisciplinary (legal, technology, management)	Auditing/ Regulatory	(AI HLEG, 2020)
F2	Attention to Algorithms	2021	Netherlands Court of Audit	Controlling algorithms in the public sector, transparency to civilians	Concretize existing normative frameworks and guidelines into aspects that need to be assessed.	Public/government	Auditing/ Regulatory	(Algemene Rekenkamer, 2021)
F3	Acces Depth Framework	2021	University of Central Florida	Sparking discussion in this emerging field of study and practice of AI audits	What aspects of an AI can be audited, based on the level of access to the AI system that the auditor has?	Computer Science	Academic	(Akula & Garibay, 2021)
F4	capAI	2022	University of Oxford	Ethics based auditing	Provide an assessment for conformity with the AI Act.	Governance	Academic	(Floridi et al., 2022)
F5	ML audit CRISP-DM framework	2018	Information Systems Audit and Control Association	Making the ML audit more accessible to more audit departments	Propose a high-level solution for ML auditing based on the system design cycle.	IT-auditing	Auditing/ Regulatory	(Clark, 2018)
F6	ESG Protocol for AI	2022	Ostfold University College	Unify AI and ESG	Propose a high-level tool for evaluating ESG related AI impacts.	Computer Science, Sustainability	Academic	(Saetra, 2022)
F7	GAFAI	2022	German Informatics Society	Enable more use cases of AI systems through establishing general requirements	Provide a set of general requirements for the audit of ML systems.	Machine learning	Academic	(Markert, Langer, & Danos, 2022)
F8	AI auditing framework	2018	Institute of Internal Auditors	Prepare internal auditors for the next digital frontier by furthering their understanding of the role of AI within an organization	Create a basis to provide AI-related assurance and advisory services within an organization.	Auditing	Auditing/ Regulatory	(The Institute of Internal Auditors, 2018)
F9	AI risk management framework	2023	United States National Institute of Standards and Technology	Protect individuals and communities from inequitable or undesirable AI outcomes	Offer a resource to the organizations designing, developing, deploying or using AI systems to help manage the many risks of AI and promote trustworthy and responsible development and use of AI systems.	Risk-management, standardization	Auditing/ Regulatory	(National Institute of Standards and Technology, 2023)
F10	Guiding Principles Trustworthy AI Investigation	2021	NOREA - Dutch Association of Chartered IT-Auditors	Providing trust in AI systems	Guide Dutch chartered IT-auditors in conducting investigations of AI systems based on leading practices for trustworthy AI.	IT-auditing	Auditing/ Regulatory	(de Boer & van Geijn, 2021)
F11	SLADA	2022	German Informatics Society	Reduce the gap between the rate at which AI systems evolve and risk-minimization strategies are developed	A framework to assess and analyze AI systems in terms of risks, adressing the structure and components of AI systems.	Computer Science	Academic	(Becker & Waltl, 2022)

Continued on next page

Table 4.2 – *Continued from previous page*

Ref.	Framework	Year	Affiliation	Values/Motive	RQ/Goal	Field/Sector	Source type	Source
F12	SMACTR	2020	Google	Conscious development of AI systems	Present a mechanism to check that the engineering processes involved in AI system creation and deployment meet declared ethical expectations and standards.	Management	Industry	(Raji et al., 2020)
F13	SMR	2021	University of Iowa	Attention to the larger socio-technical power dynamics that are inherent to AI	Propose a way to operationalize high-level ethical analyses of algorithms through an auditing instrument which translates those ethical analyses into practical steps .	Socio-technological	Academic	(Brown, Davidovic, & Hasan, 2021)
F14	Recommendations towards Trustworthy AI Development	2020	OpenAI	For AI developers to earn trust from users, customers, society, governments and other stakeholders that they are building AI responsibly	Create a robust alternative to self-assessment of claims as an institutional mechanism to verify the claims made by developers.	Management	Industry	(Brundage et al., 2020)

The typology enabled further analysis of the frameworks based on their perspective on the audit of AI. The two frameworks proposed by industry were developed by leading tech companies (Google and OpenAI) with a similar purpose, namely to "enable AI developments by generating public trust" (Brundage et al., 2020). Organizations developing AI systems have proposed auditing frameworks as they intend to gain public trust - which they require to limit public scrutiny on their AI systems. The frameworks proposed by auditing or regulatory organizations on the other hand have been developed from a perspective which aims to provide resources to guide the development and assessment of AI systems according to certain institutional values. Frameworks proposed in scientific literature have typically been developed based on a specific aspect related to trustworthy AI, such as ethics (F4), ESG (F6) or the underlying technology (F11).

To further understand the positioning of these frameworks with regards to the audit of AI, each framework was broken down into the individual questions of which they are composed. Each of the frameworks was then analyzed on a question-level through iterative open and axial coding. The 14 previously described frameworks contained a combined total of 595 auditing questions. Through this iterative open and axial coding strategy, 23 different coding labels were obtained from all of the questions. A list of the obtained coding labels is presented in Table B.1 of the Appendix. In turn, these coding labels were used to categorize each of the 595 auditing questions. A full overview of the coding of the individual questions per framework is provided in Appendix section B. This framework analysis on the question-level allowed for a further comparison of the frameworks to be carried out, both within and between the three source categories.

Frameworks from Academic Literature

Six of the frameworks were proposed in academic publications. Through the relative frequency of coding labels within each framework, their focus is quantified in Table 4.3.

Table 4.3 shows that besides an emphasis on system management, the frameworks differ in depth and focus. These differences can be explained from their typology. The ESG protocol (F6) aimed to provide a tool to evaluate ESG related AI impact, which is reflected in the prevalence of the environmental impact within the framework compared to other frameworks, which often lack this point of view. The SMR framework (F13) revolves around socio-technological dynamics, which explains the emphasis on stakeholders within the framework. GAFAI (F7) is focused on general requirements for ML algorithms, which resulted in a high-level framework with only ten auditing questions that revolve around control and documentation - two of the auditing cornerstones.

capAI (F4) is the most extensive academic framework with 40 questions, allowing it to cover many of the themes related to AI auditing. The Access Depth framework (F3) was derived from a computer science perspective with the goal of sparking discussion in the developing field of AI audits, which lead the researchers to zoom in AI quality. Lastly, the SLADA framework (F11) is a high-level framework revolving around the risk-minimization through explainability, as reflected in its focus and relatively few number of questions. Overall, it appeared that the emphasis of the academic frameworks varied from framework to framework, and that these differences could be explained through the established typology.

Table 4.3: *Academic Frameworks Coding Results.*

Coding Label / Framework	3. Access Depth	4. capAI	6. ESG Protocol	7. GAFAI	11. SLADA	13. SMR
Accountability	-	0.03	0.14	-	-	0.06
Autonomy	-	-	-	-	-	-
Control	0.06	0.15	-	**0.50**	-	-
Direct environmental impact	-	-	0.09	-	-	-
Documentation	0.22	**0.40**	0.14	**0.50**	0.14	0.11
Explainability	0.06	0.03	-	-	**0.43**	-
Fairness	0.22	0.10	0.05	-	-	0.06
Human involvement	0.06	0.03	0.09	-	-	0.11
Indirect environmental impact	-	0.03	**0.23**	-	-	-
Legality	0.06	-	0.05	-	-	0.11
Management of system	**0.33**	**0.45**	**0.32**	0.30	**0.43**	0.17
Objectives	-	0.13	0.05	0.10	0.14	-
Periodic assessment	-	0.15	-	-	0.14	-
Privacy	0.17	0.03	0.05	-	-	-
Quality	**0.33**	0.23	-	0.10	0.14	0.22
Reliability	0.22	0.08	-	-	0.14	-
Risk assessment	0.11	0.20	**0.23**	0.30	-	0.11
Robustness	0.28	0.08	-	-	0.14	0.06
Safety	0.06	-	0.05	-	-	-
Security	0.06	0.03	0.05	0.10	-	0.06
Stakeholders	0.11	0.08	0.09	-	0.14	**0.33**
System description	0.22	0.08	**0.23**	0.10	-	**0.28**
Transparency	0.11	0.10	-	-	0.14	0.11
Questions per Framework	18	40	22	10	7	18

Values indicate the fraction of questions within each framework that were labelled with the corresponding code, i.e. 40% of the questions in the capAI framework were linked to documentation. Values represented in grayscale. Top two labels with highest frequency for each framework marked in boldface.

Frameworks from Auditing or Regulatory Institutions

The frameworks proposed by auditing or regulatory organizations were broken down in the same manner as the academic frameworks. The overview of the frameworks and the relative frequency of the 23 coding labels throughout each framework is shown in Table 4.4.

Table 4.4: *Auditing and Regulatory Frameworks Coding Results.*

Coding Label	Framework: 1. ALTAI	2. Attention to Algortithms	5. ML Audit Framework	8. AI Auditing Framework	9. AI Risk Management Framework	10. Guiding Principles
Accountability	0.04	0.01	-	0.08	0.08	0.02
Autonomy	0.01	0.04	-	-	-	0.02
Control	0.10	0.14	0.14	**0.29**	0.21	0.18
Direct environmental impact	0.02	-	-	-	0.01	0.01
Documentation	0.10	**0.20**	-	0.08	**0.47**	0.20
Explainability	0.03	0.06	0.07	0.08	0.01	0.06
Fairness	0.19	0.05	0.07	0.13	0.01	0.06
Human involvement	0.09	**0.20**	-	0.13	0.22	0.12
Indirect environmental impact	0.02	-	-	-	0.01	0.01
Legality	0.05	0.08	0.07	-	0.06	0.06
Management of system	**0.30**	**0.42**	0.07	**0.46**	**0.38**	**0.28**
Objectives	0.02	0.15	0.29	0.13	0.13	0.12
Periodic assessment	0.10	0.03	-	0.08	0.17	0.05
Privacy	0.07	0.11	0.07	-	0.01	0.08
Quality	0.07	0.15	**0.43**	0.25	0.08	**0.26**
Reliability	0.05	0.04	-	-	0.03	0.03
Risk assessment	0.21	0.03	0.07	0.08	0.35	0.10
Robustness	0.06	0.01	0.07	-	-	0.04
Safety	0.07	0.01	-	-	0.06	0.02
Security	0.09	0.18	-	0.13	0.01	0.12
Stakeholders	**0.36**	0.14	0.07	-	0.21	0.18
System description	0.04	0.14	**0.43**	0.04	0.11	0.10
Transparency	0.13	0.08	-	-	-	0.06
Questions in framework	135	79	14	24	72	124

Values indicate the fraction of questions within each framework that were labelled with the corresponding code. Values represented in grayscale. Top two labels with highest frequency for each framework marked in boldface.

It was revealed that again most frameworks emphasize system management, which follows from the idea that auditing is a way to ensure proper technology governance. Furthermore, the focus of the frameworks could be explained through the established typology. The Attention to Algorithms guidelines (F2) were established by the Court of Audit of the Netherlands (Algemene Rekenkamer). Their work revolves around the evaluation of audit

trails in the public sector, which links to their guideline focus on documentation and human involvement respectively. The Guiding Principles (F10) as proposed by the Dutch guild of IT auditors (NOREA) emphasize system quality. This focus is logical as from an IT auditing point of view, a higher system quality translates to fewer and less severe technical risks. As one of the most detailed frameworks, they are able to incorporate all identified themes. The most detailed framework is the Assessment List for Trustworthy AI (F1), which specifically focuses on stakeholders. As the basis of the framework are fundamental human rights, this stakeholder focus enables the framework to cover a great variety of topics in relation to the people affected by an AI system.

The AI Risk Management Framework (F9) was developed by NIST, an American regulatory institute. Like the framework proposed by the Dutch Court of Audit (F2), this perspective of the regulator translates to a focus on documentation. The AI Auditing Framework (F8) was one of the first to be developed in 2018, at which time it was an exploratory work to provide a basis for AI related assurance and advisory. As the potential of AI nowadays through the latest RL/DL innovations were not as clear at that time, the framework lacks depth compared to the others in this source category. Furthermore, as the framework was developed by the Institute for Internal Auditors (IIA), the control perspective is prevalent - the IIA is not IT audit specific but rather oriented towards business process control. Lastly, the ML Audit Framework (F5) was proposed by the international organization for IT-auditors ISACA. Similar to the AI Auditing Framework (F8) of the IIA, it was proposed in 2018 and served as a first attempt to make the AI audit accessible, which is why the audit is the least detailed of the frameworks in this source category. Its focus on system description is in line with its defined goal to improve the level of understanding that IT auditors have of AI systems.

The overall focus of the frameworks in this source category seemed to be in line with their positioning in the typology. These larger frameworks were both broader and more detailed, covering many of the identified topics. This tendency towards larger, more detailed frameworks by the audit industry is in line with the preferences of the assurance professionals, the problem owners. This preference stems from a desire to standardize procedures as much as possible, in order to be able to produce auditing reports that are consistent and comparable between one another, as also indicated in the interviews (I2, I3). Ideally, this means the use of the same auditing framework for each client, thus requiring the frameworks to be both broad and detailed as there is a great variety in AI systems and subsequently in contexts in which they are deployed.

Frameworks from Industry

Lastly the two frameworks proposed through organizations that develop AI themselves were also analysed. The results are shown in Table 4.5.

Table 4.5: *Industry Frameworks Coding Results.*

	Framework	
Coding Label	12. SMACTR	14. Recommendations Towards Trustworthy AI
Accountability	-	0.11
Autonomy	-	-
Control	0.09	0.11
Direct environmental impact	-	-
Documentation	**0.30**	-
Explainability	0.04	-
Fairness	0.17	0.11
Human involvement	0.09	0.11
Indirect environmental impact	-	-
Legality	-	-
Management of system	**0.43**	-
Objectives	0.26	-
Periodic assessment	-	-
Privacy	0.04	0.11
Quality	0.13	0.11
Reliability	0.09	-
Risk assessment	0.22	-
Robustness	0.04	-
Safety	-	0.11
Security	-	0.11
Stakeholders	0.22	-
System description	0.09	0.11
Transparency	0.04	-
Questions in framework	23	9

Values indicate the fraction of questions within each framework that were labelled with the corresponding code. Values represented in grayscale. Top two labels with highest frequency marked in boldface for the first framework.

The SMACTR (F12) framework was proposed by Google as a high-level approach to meet societal expectations of ethical development of AI. To track the ethical considerations during development, their framework emphasizes documentation as well as the setting of objectives related to ethics. The Trustworthy AI Recommendations (F14), developed by OpenAI, was set up as broad, high-level recommendations which is why each of their questions fell into a separate coding category.

Source-based comparison

Following the analysis of the individual frameworks, the three source types could be compared. For this, the average coding label frequencies between the frameworks in each of the source types were calculated. These values are shown in Table 4.6.

Table 4.6: *Framework Coding Results per Source Type.*

Coding Label	**Framework Source**		
	Academic	Auditing/ Regulatory	Industry
Accountability	0.04	0.04	0.06
Autonomy	-	0.01	-
Control	0.12	0.18	0.10
Direct environmental impact	0.02	0.01	-
Documentation	**0.25**	0.18	**0.15**
Explainability	0.08	0.05	0.02
Fairness	0.07	0.08	0.14
Human involvement	0.05	0.13	0.10
Indirect environmental impact	0.04	0.01	-
Legality	0.04	0.05	-
Management of system	**0.33**	**0.32**	**0.22**
Objectives	0.07	0.14	0.13
Periodic assessment	0.05	0.07	-
Privacy	0.04	0.06	0.08
Quality	0.17	**0.21**	0.12
Reliability	0.07	0.02	0.04
Risk assessment	0.16	0.14	0.11
Robustness	0.09	0.03	0.02
Safety	0.02	0.03	0.06
Security	0.05	0.09	0.06
Stakeholders	0.13	0.16	0.11
System description	0.15	0.14	0.10
Transparency	0.08	0.04	0.02
Average number of questions	19	75	16

Values indicate the average of the fractions of questions within each of the frameworks that were labelled with the corresponding code per source type. Values represented in grayscale. Top two labels with highest frequency for each source category marked in boldface.

A significant difference in size exists between the frameworks developed by auditing and regulatory sources compared to academic or industry sources, with averages of 75, 19 and 16 questions respectively. This meant that frameworks by auditing or regulatory institutions were generally more complete, covering a greater range of topics as well as a greater level of detail. Compared to the other two categories, they put greater emphasis on control and quality of the AI system, which aligns with the sources being regulators and auditing parties - these topics are also essential in IT audits for example. The focus of the academics frameworks was shown to differ based on the various research fields the

frameworks were proposed in; the area of expertise of the researchers would be the focus of their framework. As mentioned by one of the interviewees, the strength of the academic frameworks lies in them being ahead of the curve in exploring a specific theme related to trustworthy AI (I3).

Remarkably, the two industry frameworks were more geared towards themes such as fairness, privacy and safety than the other frameworks. This is in part the result of the limited number of frameworks and questions skewing the data. It could also be explained given their stated purpose of enabling further AI development, for which they identified public trust as a key requirement. Public trust in turn closely relates to fairness, privacy and safety as those themes are directly linked to risks the AI system could pose to the public.

4.2.3 Timeline of Frameworks

The timeline of when the identified frameworks were published, displayed in Figure 4.2, shows a couple of trends. For one, the high-level exploratory frameworks precede the more detailed ones. This was to be expected as the field of AI assurance is maturing and developing over time, in parallel with technological advancements. Furthermore, what stands out is that AI developers were amongst the first to publish AI auditing initiatives. This can be explained from their awareness of the public sentiment regarding AI and associated risks, which is illustrated by the executives of Google and OpenAI supporting initiatives to call AI developments to a halt in order to allow laws and regulations to be developed (Future of Life Institute, 2023a, 2023b). Frameworks proposed by academic sources emerge latest in the timeline. While difficult to pinpoint the exact reason for this, it might be the result of the latency between submission and publication for certain journals. Only one framework was found to be proposed in 2023, which likely is the consequence of the literature search being carried out early in the year.

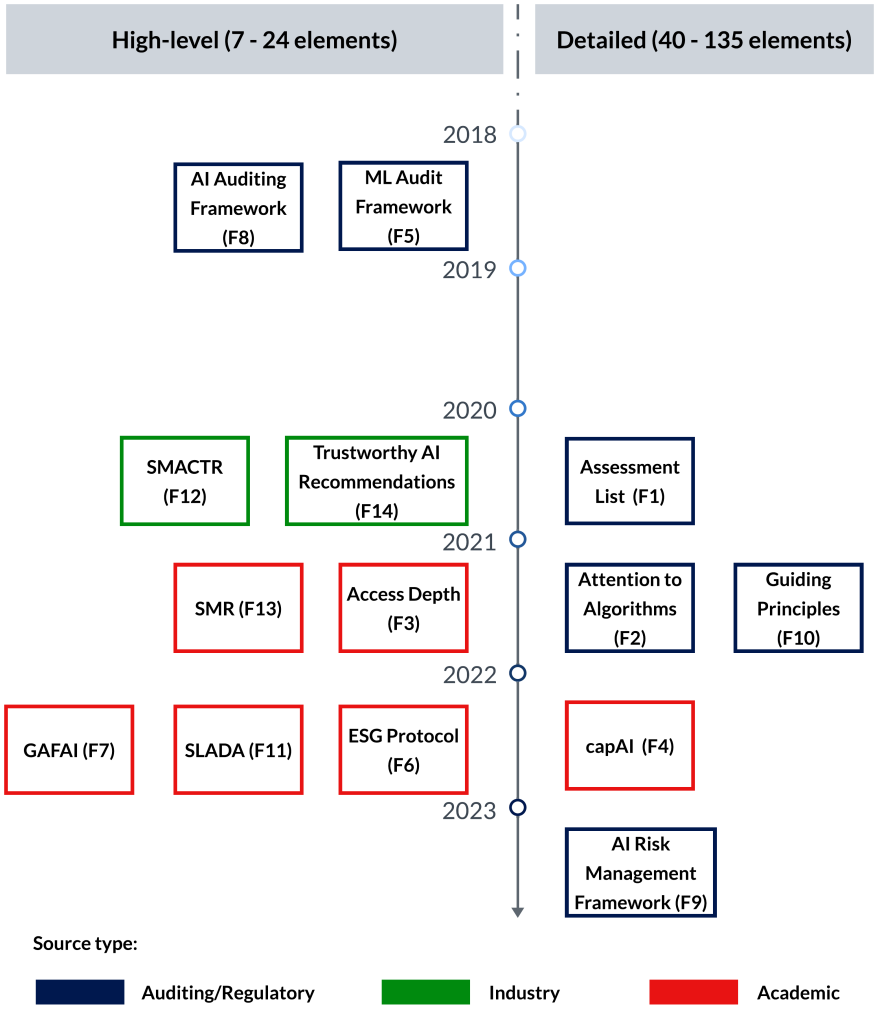


Figure 4.2: *Timeline of Identified Frameworks.*
Frameworks proposed by auditing or regulatory bodies shown in blue, by industry in green and academic sources in red. Distinction between high-level and detailed frameworks is based on number of framework elements. Reference IDs from initial typology included for all frameworks.

4.2.4 IT Audit Frameworks and Standards

The audit of AI is an extension of the field of IT auditing, a profession which has co-evolved with technological developments. The switch from traditional IT auditing to the audit of AI, however, is a great leap in terms of the impact of the system and associated risks. It is much greater than more traditional developments in the IT audit, such as the introduction of new security standards. This difference between AI and IT is reflected in the make-up of the examined AI auditing frameworks. To illustrate this, the AI auditing frameworks are compared to four of the key standards and frameworks that are regarded as leading practices in IT auditing: COBIT, ITIL, ISO 27001 and ISAE 3402 (Gantz, 2014). An overview of the four IT auditing frameworks and standards based on their purpose and focus is provided in Table 4.7.

Table 4.7: *Overview of Key IT Audit Standards and Frameworks.*

Name	Type	Purpose	Focus	Source
COBIT	Framework	Ensure IT processes and systems are delivering value to the organization, help align IT activities with business goals	Risk management, compliance, performance measurement and alignment with business objectives	(De Haes, Van Grembergen, Joshi, & Huygh, 2020)
ITIL	Framework	Assist businesses in aligning IT services with customer and business needs	Creating value for the stakeholders	(AXELOS, 2020)
ISO 27001	Standard	Demonstrate robust IT security practices, ensuring integrity within the organization	Risk management, compliance and security	(Hsu, Wang, & Lu, 2016)
ISAE 3402	Standard	Ensuring the security and robustness of (outsourced) IT services	Assessment of internal IT controls relevant for financial reporting	(Radulescu & Pestritu, 2011)

The goal of an IT audit is to show that management is in control of their information systems and thereby able to minimize risks to the business (Aditya et al., 2018; Stoel et al., 2012). At the core of IT audit, as is also apparent in the overview of the four IT audit standards and frameworks, therefore lie risk management, security, performance and value creation (Mancham, 2007). The AI auditing frameworks on the other hand cover a much greater variety of topics, and focus on more themes than these four. The lack of themes such as fairness, safety, environmental impact, and explainability illustrates the complex techno-social dimension which is not present in 'traditional' IT systems but rather prevalent in AI systems.

Some of the assessed AI auditing frameworks were reportedly developed with current IT auditing practices as starting point, such as NOREA's Guiding Principles (F10) and ISACA's ML Audit Framework (F5). While IT audits primarily emphasize risk and security, AI audits encompass a broader spectrum of considerations due to the unique characteristics of AI. Unlike IT systems, AI systems inherently contain a degree of unpredictability. This inherent unpredictability introduces uncertainties and associated risks to stakeholders that must be evaluated. Additionally, AI systems can be opaque, making it challenging to decipher their internal decision-making processes. Consequently, AI audit frameworks extend beyond the conventional IT auditing focus areas to encompass themes such as transparency, fairness, ethical considerations and bias mitigation.

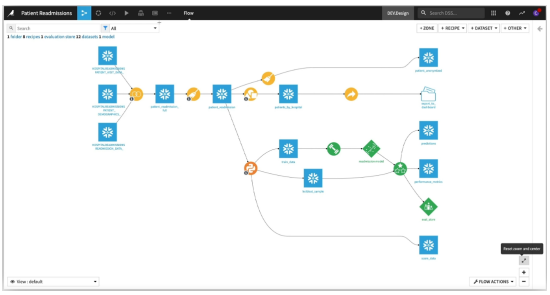
Moreover, the application of AI within complex socio-technological contexts involves multiple stakeholders, including regulators, data subjects, and AI system developers, each with distinct interests and concerns. This complexity requires a more comprehensive evaluation of the impact of AI systems on society, ethical implications, and legal compliance. Therefore, AI auditing frameworks address not only the traditional IT auditing aspects but also strive to cover the socio-ethical dimensions of AI.

4.2.5 AI Performance Dashboard

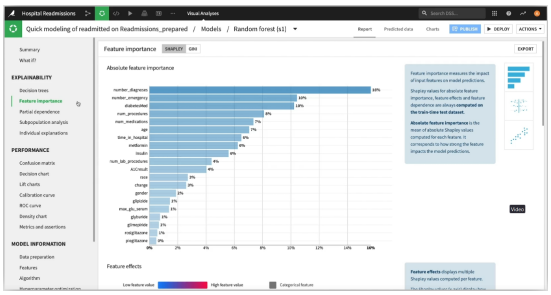
Two of the interviewed assurance professionals referred to already existing dashboards and platforms that can be used to monitor AI performance and track other technical specifi-

cations (I2 and I3). One of the largest companies that offer such a service are Dataiku, valued at $3.7 billion in 2022 (Dataiku, 2022). Their platform Dataiku Data Science Studio (DDSS) was launched in 2013, and reportedly over 200 of the listed companies in the Forbes Global 2000 are their customer (Dataiku, 2023a). DDSS was therefore chosen to serve as object of analysis in the context of AI audits. The main sources of information consulted for the capabilities of DDSS were a video released by Dataiku showcasing the DDSS functionality and the Dataiku webpage diving into the DDSS key capabilities.

The strengths of DDSS lie in its ability to provide insights in AI systems on various levels: the project, the model and the data. In each case DDSS relies on insightful visualization through its dashboard to clearly present breakdowns at the various levels of the AI system. On the project level, the flow of information and its processing are visually represented as a flowsheet like the one depicted in Figure 4.3a, thereby contributing to project transparency. This flowsheet can then easily be translated into an autogenerated project report. The model performance and reliability are presented through diagnostic panels in DDSS. The panels show performance metrics such as model deviations for data subgroups, and displays warnings in case the model is at risk of overfitting or data leakage. The model output can be broken down through a feature importance analysis such as the one shown in Figure 4.3b, which allows for a deeper level of explainability regarding what information for a given sample most heavily contributed to its associated model output. Furthermore, the training data quality can also be assessed through DDSS. Potential biases can be explored through statistical testing and subsequent visualization of correlations within the training data.



(a) DDSS Project Flowsheet.



(b) DDSS Feature Importance Interface.

Figure 4.3: *Dataiku Data Science Studio Interfaces.*
Screenshots of DDSS interfaces for a mock model used by Dataiku to illustrate their platform. From Dataiku (Dataiku, 2023b).

The benefits of platforms such as DDSS are thus derived from their technical depth, enabling developers to obtain technical details, as well as providing metrics and visualization to aid in model transparency and explainability. These figures and data derived from DDSS can for example be communicated to relevant stakeholders in order to reduce the opacity of the AI system to them. As such, DDSS and similar platforms could assist in covering the parts of the technical, transparency and explainability aspects of an AI audit.

It is, however, unlikely that they will be able to cover a complete AI audit. This is related to the shortcomings of platforms such as DDSS. For one, while its strengths lie in

'opening the black box' through data analytics and visualization, these insights are limited by what is known to the developers. For example, model performance can be visualized for various subpopulations in the data, which can in turn expose biases within the model. This can, however, only be done for known subpopulations in the data, meaning that this assessment of bias and fairness relies on the awareness of developers of the presence of subpopulations in their data - which should not be assumed.

An illustration of this is the case of ethnical profiling by the Dienst Uitvoering Onderwijs (DUO) (NOS, 2023), the Dutch executive agency of education. DUO employed an algorithm to mark students for fraud investigation concerning their right to receive study grants. Unintentionally, the algorithm had a strong bias towards students belonging to minorities even though ethnicity was not used as an input feature for the model. The algorithm had instead based its categorization on other features, such as living with relatives who are not their parents, which in turn highly correlated with subjects belonging to minority groups. Such insights would not have been uncovered through platforms such as DDSS, which rely on known subpopulations in the data.

Other platform shortcomings in the context of AI audits relate to the scope of the platforms. AI systems and their impact extend well beyond the technical details of the model. The dimensions such as safety, security, stakeholder perspectives, risk assessment and environmental impact all vary widely depending on the specific context in which an AI system is active, regardless of its technical specifications. As DDSS and other platforms alike cannot cover these topics, as these are too variable and context-dependent, this is where other mechanisms, such as the audit of AI, are needed in order to enable trustworthy AI.

In conclusion, existing platforms such as DDSS excel in providing insights in technical aspects of AI models, making them valuable for transparency and performance monitoring. Within the greater scope of AI audits they could be used to support the audit on those areas. Nonetheless an AI audit cannot solely rely on such platforms as the impact of an AI system depends for a large part on its socio-technological context. This requires active engagement of auditors in order to investigate potential risks related to this context instead.

4.3 Development of Workflow Sub-Processes

4.3.1 General AI Auditing Framework

The individual questions distilled from the fourteen identified AI auditing frameworks were recombined into a general AI auditing framework along the seven principles of trustworthy AI as defined by the EC - a design decision based on established criteria (I3).

As indicated by the interviewed assurance professionals (I1, I2), it is sensible for the framework to allow questions to appear within multiple categories when their relevancy spans across them. Therefore these were explicitly kept in place. Redundancies were removed within each of the seven categories. An overview of the number of questions included for each of the separate principles in the framework is provided in Table 4.8. The

final general AI auditing framework is included in the Appendix: Table C.1 to Table C.7 for each of the principles.

Table 4.8: *Questions per Principle in the General Auditing Framework.*

Principle	**Number of Questions within Framework**
1. Human Agency and Oversight	33
2. Technical Robustness and Safety	80
3. Privacy and Data Governance	22
4. Transparency	21
5. Diversity, Non-discrimination and Fairness	33
6. Societal and Environmental Well-being	13
7. Accountability	43

The data presented in Table 4.8 shows that while five of the categories are comprised of roughly 20 - 45 questions, the fewest can be found in the societal and environmental well-being category at 13, while the technical robustness and safety category has the most at 80. This could be the result of two factors. First of all, societal and environmental well-being is the broadest category while technical robustness and safety is more specific. This translates to the level of detail and consequently number of possible questions in the categories. As societal and environmental well-being are broad, immaterial topics that can vary greatly from case to case, the related questions will remain high-level. The technical robustness and safety category includes many more concrete questions and a greater level of detail, due to the more closed nature of these questions (e.g. Q2.18: "Has the training data annotation been validated?").

The second reason for this stems from the fourteen frameworks that were combined into this general framework. Many of these frameworks take on a technical perspective (F3, F11, F13) or are derived from an IT audit starting point - which extensively cover the technological quality and security aspects (F5, F8, F10). While all frameworks acknowledge the relevance of stakeholders to a certain extent, few actually cover the environmental (F6) or broader societal impact (F1) as one of the main focuses.

On the question level some further observations can be made. Each of the seven categories contains questions relating to communication with stakeholders, e.g. Q2.13: "Are model performance and limitations communicated to all stakeholders?", Q4.6: "Are stakeholders informed about the goal of the AI system?" This signifies the shared attitude across the sources towards the importance of stakeholder inclusion for trustworthy AI systems. If anything, stakeholder inclusion across the domains of AI trustworthiness will be an important building block to generate the trust of the stakeholders, which has also been pointed out in literature (Robinson, 2020). Other topics that reappear across most of the seven categories are value-trade-offs made during the development of the AI system and organizational guiding values.

Furthermore, some questions stand out as they have been derived from a great variety of the initial frameworks. These are the following: Q4.12: "Is AI system output clearly presented to all stakeholders?"; Q2.3: "Does the AI system perform at the defined ac-

ceptable level?"; Q7.12: "Has a risk assessment of the AI system been performed and documented?"; Q7.26: "Has the severity of the identified risks been assessed?" All four questions are based on question from six or more of the original frameworks. They signify core concepts for AI audits within the framework, namely the assessment of risks, the clarity and the quality of the AI system - each of which is also underlying many of the other questions within the framework.

4.3.2 Scoping Approach

ESG Reporting Materiality Assessment

In addition to a general auditing framework, a method was developed in order to be able to apply the framework in an audit setting. This required first establishing a materiality assessment practice, for which the observations from corporate ESG reporting were taken as starting point. The rationale for relying on the ESG reporting materiality assessment are the similarities between the scoping problems in ESG reporting and AI audits. ESG reporting is described as a practice that lacks standardized methodologies and that comes with inherent tensions between divergent stakeholder demands which result in an array of competing organizational goals (Garst et al., 2022). AI audits also currently lack standardization and entail the assessment of how AI design choices impact various stakeholders.

Furthermore, ESG reporting faces challenges in assessing the materiality of topics that extend beyond the boundaries of the organization due to three characteristics: their complexity, uncertainty, and evaluative nature (Garst et al., 2022). These same characteristics apply to AI auditing scoping challenges. The complexity for AI audit topics stems from both the interaction between many stakeholders and the interconnectedness of those topics. Next, the perceived importance of the various topics related to trustworthy AI is also subject to uncertainty. Changes in the public opinion or the development of regulations affect which topics are material for the AI audit. Lastly, the materiality of the topics is evaluative in nature, meaning that the materiality will vary as it is based on various perspectives and interests. This too is the case for the AI audit, where different stakeholders could prioritize different audit topics.

The main difference between ESG reporting and AI audits is the position of the party that executes the materiality assessment. Organizations conduct ESG reporting entirely on their own, whereas an AI audit is conducted by an external auditor. In ESG reporting, the absence of an external party in the development of a sustainability report, combined with the lack of standards, facilitates greenwashing (Garst et al., 2022). Effectively, this means that commonly ESG reports are focused on ESG topics which favour the firm while not mentioning ESG topics on which the organization underperforms (Kaplan & Ramanna, 2021).

AI audits on the other hand are conducted by an external auditor. Although their exact level of independence is up to debate as they are financed by the auditee (see the Enron scandal), their outside view will provide a more objective assessment in the audit report than a complete self-assessment. It also means that the auditor will need to become acquainted with their client and their AI system, which in turn is similar to how IT audits

are initiated. Therefore, the proposed AI audit scoping approach will blend elements from both the ESG reporting materiality assessment and standard auditing procedures.

AI Audit Materiality Assessment

The six steps for materiality assessment in corporate ESG reporting as identified in literature are: choosing a materiality perspective, specifying ESG topics, determining information sources based on the chosen perspective, collecting data to determine the materiality of topics, selecting the most material topics and finally deciding on a timeframe for when the materiality assessment should be conducted again (Garst et al., 2022). Additionally, the identified relevant IT audit processes are gaining an understanding of the client and the system under audit, as well as their purpose and stakeholders. The AI audit scoping approach was refined using feedback of the interviewed assurance professionals on an initial design, which is shown in Figure D.2 of the Appendix. The updated scoping approach is shown in Figure 4.4.

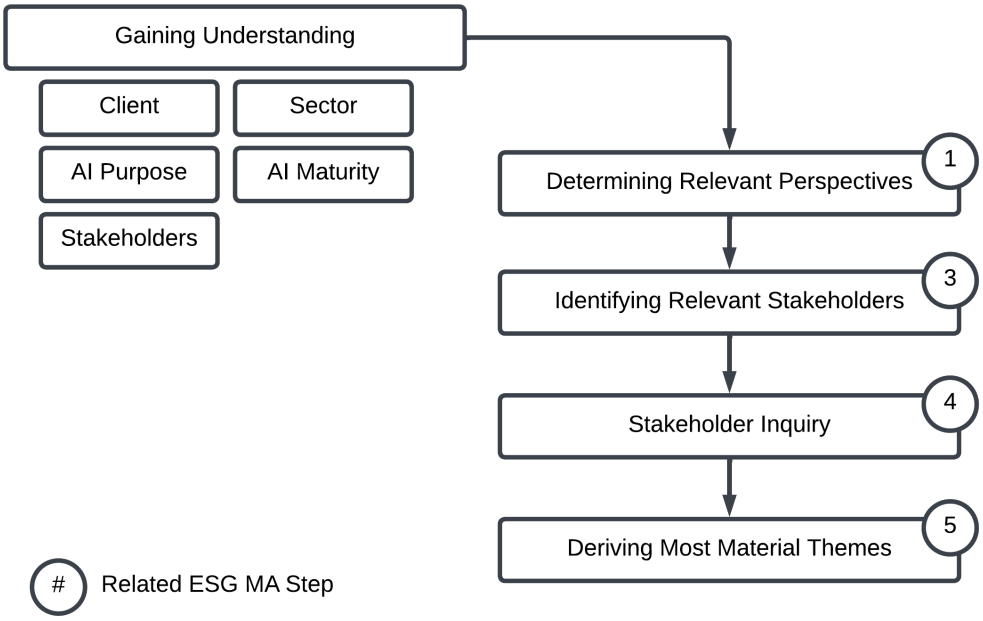


Figure 4.4: *Proposed AI Audit Scoping Approach.*
Associated numbers refer to corresponding ESG reporting materiality assessment step from the source material (Garst et al., 2022).

The second and sixth step of the ESG reporting materiality assessment are omitted from the AI audit scoping approach. The second ESG reporting materiality assessment step entails specifying relevant topics, whereas the AI audit will rely on the seven themes along which the general auditing framework was constructed. These themes are therefore predetermined and do not need to be specified for each audit separately; they are set up such that they cover all requirements for trustworthy AI (AI HLEG, 2019). Subsequently, the stakeholder inquiry serves to determine which of these topics the stakeholders deem most material. The final step, planning the next audit, is omitted from the scoping

approach as it will be the final step in the overall audit workflow, to be taken once the audit has been completed.

Feedback of the assurance professionals was incorporated such that the starting point of the materiality assessment, and the audit for that matter, is first gaining an understanding of the client and their AI system (I2), including the sector (I2), maturity (I3), purpose (I2) and the stakeholders (I1). These steps are similar to the start of a regular audit engagement, where the auditor also needs to first gain an understanding of the client. Additionally, it was pointed out that the auditor should be the one carrying out all the scoping steps as they will be more knowledgeable on the topics than the client (I2). The proposed scoping approach still respects key strengths of the initial design, such as sticking to the seven specified themes, and retaining the order of the steps as they were originally proposed (I1, I2).

The proposed scoping approach reveals that the AI auditor is expected to complete a number of preparatory steps before the actual audit can take place. This is expected to contribute to the objective nature of the audit overall, as it would be less convincing if these steps were carried out by the client themselves. That would allow for 'ethics-washing,' for example by limiting the audit scope to known strengths of the AI system (Wagner, 2019).

Chapter 5 Final Product and Validation

5.1 Complete Workflow

A complete AI auditing workflow was derived by combining the scoping approach with the translation of relevant questions from the general auditing framework to the specific case, and regular auditing steps relating to the client-side of the audit. The materiality assessment provides the auditor with a scope of the most material themes for the AI audit, for which auditing questions can be gathered from the general framework. These questions can then be assessed for relevancy for the specific client case based on the understanding that the auditor has of the client and their AI system. For example when the AI system is still in development and not yet deployed, some of the questions regarding the use of the AI system can be excluded as they are not applicable. Then, the auditor walks through the questions with the client, who are expected to provide evidence showing to what degree they comply with the question. The auditor can then assess the evidence for each question and finish the audit with a complete overview of their findings in an audit report. Lastly, the final step is to plan a future audit as the client further develops their AI practices and incorporates the feedback from the audit report. The complete workflow is shown in Figure 5.1.

The proposed workflow shows that the AI audit is in essence a broader version of the IT audit. A wider range of topics may be potentially relevant for the audit, which is why it is required to determine their materiality prior to the actual assessment. If this would not be done, and for example all the questions from the general auditing framework were to be submitted to the client, this would miss the point of the audit. The goal should be to "cover and mitigate the most important risks" (I1). IT audits on the other hand are typically scoped along what aspects and controls of the IT system are predetermined to be audited by laws and regulations - which are not in place for AI.

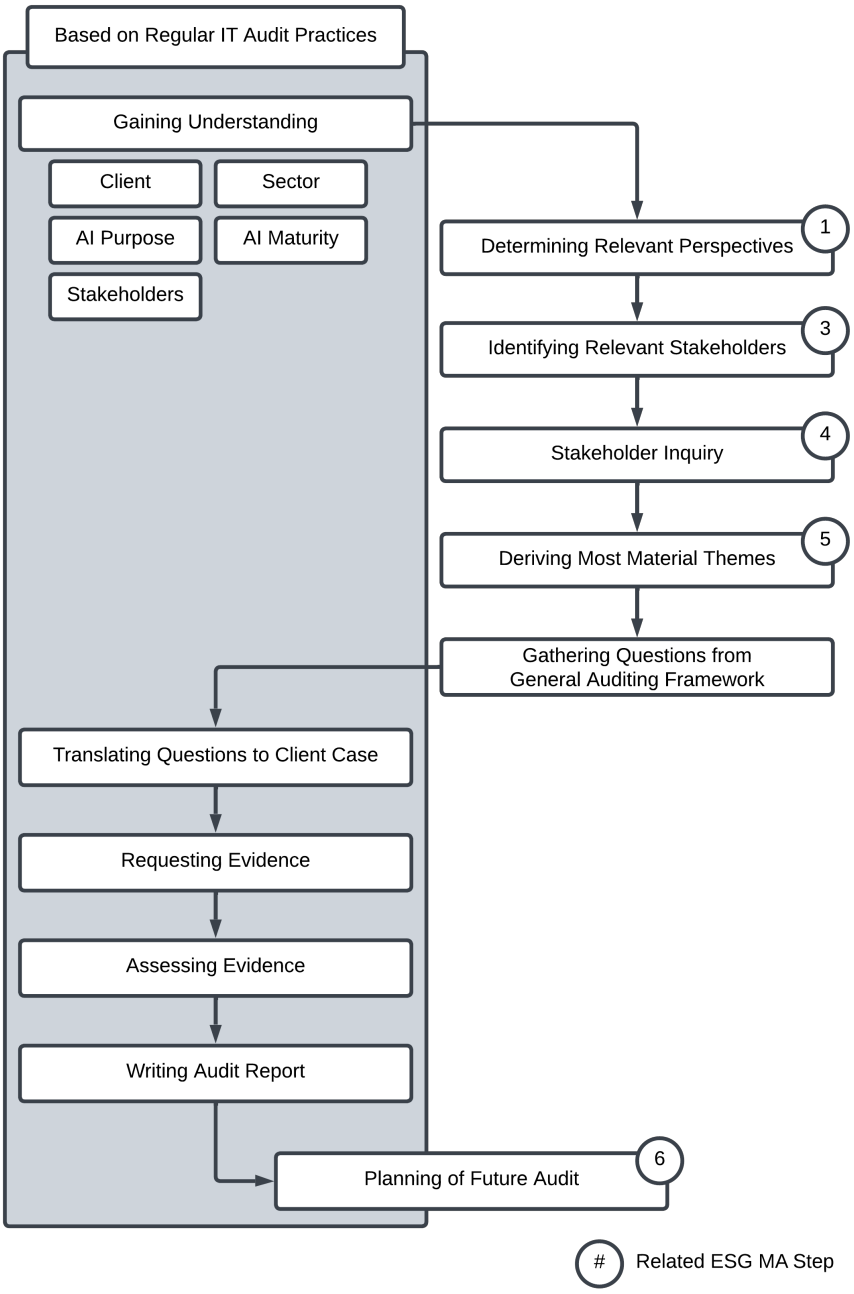


Figure 5.1: *Complete AI Audit Workflow.*
Associated numbers refer to corresponding ESG reporting materiality assessment step in the source material (Garst et al., 2022). Steps on the left side against the gray background correspond to steps that are comparable to conventional audit steps.

5.2 Case Study

To demonstrate the applicability of what is essentially a prototypical workflow, a mock client was chosen to be audited. By going through the proposed steps, the feasibility of the workflow could be supported through proof of concept. The mock client of choice is the Municipality of Amsterdam, and their AI system called "Public Eye". This particular

AI system was chosen based on the accessibility and level of detail of publicly available information about it. The mock client was not actively involved in this process as they proposed to collaborate only on grounds that required a contract and their approval of any findings of the research.

5.2.1 Gaining an Understanding

It was decided that first all publicly available sources about Public Eye and the role of the municipality of Amsterdam were to be gathered in order to become more familiar with the AI system and its stakeholders. The publicly accessible sources that were identified as relevant for the auditors to gain an understanding of the client and their AI system are summarized in Table 5.1 below. The sources are entries in the Algorithm Register of the Government of the Netherlands and the Algorithm Register of the Municipality of Amsterdam, the open source repository on Github, the project overview webpages of the co-developers (Tapp and Life-Electronic) and site of deployment (Marineterrein), a blog post of one of the developers, as well as the Crowd Monitoring System Amsterdam (CMSA) sensor overview map and Marineterrein Busyness Dashboard.

Table 5.1: *Public Eye Publicly Accessible Sources.*

Name	**Source Location**
National Algorithm Register	algoritmes.overheid.nl/en/algoritme/38748497
Amsterdam Algorithm Register	algoritmeregister.amsterdam.nl/public-eye/
Public Eye Github Repository	github.com/Amsterdam/public-eye
Tapp Project Overview	tapp.nl/project-overview/public-eye
Life-Electronic Project Overview	life-electronic.com/diensten-1
Marineterrein Project Overview	living-lab.nl/experiments/open-source-crowd-monitor
Crowd Monitoring System Amsterdam Sensor Map	maps.amsterdam.nl/cmsa/
Marineterrein Busyness Dashboard	mt-dashboard.nl/
Developer Blog Post	amsterdamintelligence.com/posts/crowd-counting

An in-depth description of all findings related to the mock client and their AI system is included in the Appendix, subsection E.1. Main takeaways are that the Public Eye AI system is considered fully mature as it is operational and its output publicly accessible. Furthermore, the AI system is operational in the public domain where it is used to process camera images to count and track the number of people present at four locations in Amsterdam. This information can be used by municipal crowd managers to take measures against overcrowding, as well as by anybody intending to visit those locations in order to assess the busyness there as this information is freely accessible to the public. The model underlying the AI system is a pre-trained DL Computer Vision algorithm.

The final part of the first step of the audit is establishing a list of all stakeholders. This is needed as they will be used to provide insights in what themes in the scope of the AI audit are important to them, in order to improve their trust in the AI system. The list of relevant stakeholders was established based on the information available in the sources of Table 5.1. Besides the parties involved in the development and operation of the Public

Eye, other identified stakeholders are the users of the system, the visitors of the monitored areas, as well as people who are more frequently in the monitored areas due to work or living there. A full list of stakeholders is provided in Table 5.2.

Table 5.2: *Public Eye Stakeholders.*

Stakeholder	**Relevance**
Municipality of Amsterdam	Facilitating Development and Operations
CTO Innovation Team	Product Owner
Tapp	Co-Developer
Life-Electronic	Lead Developer
Marineterrein Amsterdam	System Deployer
Busyness Platform Users	Reliant on Model Output
Municipal Crowd Managers	Reliant on Model Output
Visitors of Monitored Area	Data Subject
Frequenters of Monitored Area	Data Subject

5.2.2 Audit Perspective

The next step in the proposed audit workflow was deciding on a materiality perspective for the audit. Due to time constraints and the expected convenience with which certain stakeholders could be reached, it was decided to limit the perspectives and approach the audit from the data subject perspective, a form of the societal impact perspective. Other perspectives to be considered were: the perspective of the user, the perspective of the system developers or of the Municipality of Amsterdam (the more traditional business perspective), and the perspective of the environmental impact.

5.2.3 Stakeholder Identification and Inquiry

It was then necessary to establish a method to conduct the stakeholder inquiry. This meant reaching out to people who are either incidentally or frequently in one of the monitored areas. They were interviewed in order to determine what they find important themes for external assurance on the AI system in which they are data subjects. The interview protocol is included in the Appendix, subsection E.2. The process through which the protocol was derived and how the interviews were conducted is described in the Methodology chapter, subsection 3.5.1.

A total of twenty stakeholders were interviewed at the Marineterrein, five of which indicated frequently visiting the area for work or nearby residence whereas the other fifteen were occasional visitors. The stakeholders were asked to rank the seven audit topics from most (1) to least (7) important to them. The average of the ranked importance of the seven audit topics was taken as indication of the materiality of those topics to the stakeholders. The resulting ranking is shown in Figure 5.2.

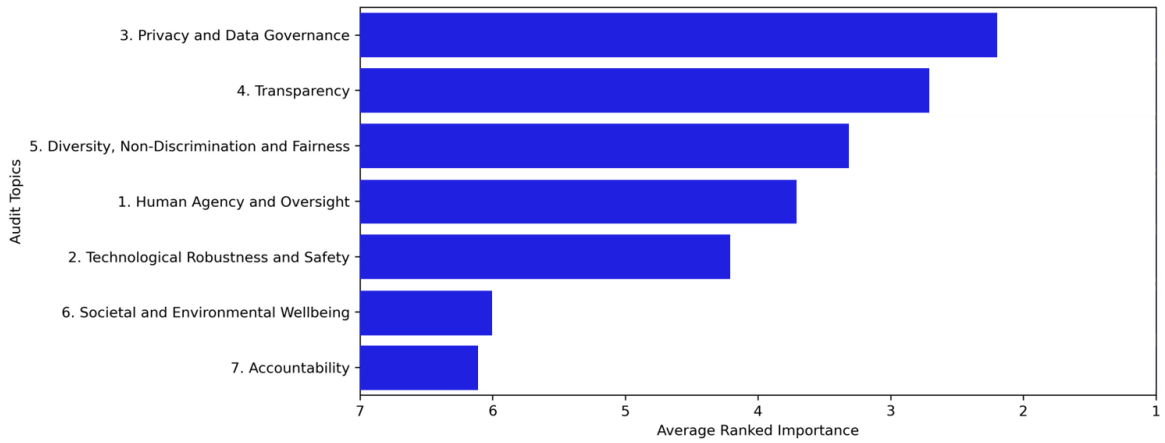


Figure 5.2: *Stakeholder Materiality Ranking of Audit Topics.*
Stakeholders were asked to rank each of the seven topics from most (1) to least (7) important to them. Topics shown from highest average importance to lowest.

Five of the stakeholders provided conflicting answers between the open question and ranking question regarding what topics they considered most important. Their input was therefore removed from the dataset, which did not result in significant changes in the final materiality ranking. The ranking with the input of all stakeholder including those who provided conflicting responses is included in Figure E.1 of the Appendix.

Additionally, when zooming in on the two subpopulations specifically, some differences between the two datasets were observed. The materiality rankings of the frequent and occasional visitors are included in the Appendix, Figure E.2 and Figure E.3 respectively. The two highest ranked topics are privacy and transparency for both groups. For the people who are frequently in the monitored area, on average the next most important topic was technical robustness and safety, which ranked fifth most important for the occasional visitors. Why they ranked the technical robustness and safety topic higher is speculative, but it could be the result of them prioritizing their own (data) security; since they are more frequently exposed to the AI system, they could consider themselves more at risk in case of security malpractices.

The lowest ranked topics were accountability and societal and environmental well-being. More often than the others, the accountability category had to be explained using examples or a description of what it entailed. It may be that because the interviewed stakeholders felt like they least understood the topic, they would rank it lowest. Societal and environmental well-being on the other hand did not need to be explained as often. However, as the stakeholders were asked to rank which audit topics would most improve their own trust in the AI system, the topics that impacted them as individuals were typically considered more material - as reflected by the three highest ranking topics. This in turn left topics such as societal and environmental well-being, which affect the stakeholders less directly, with a low ranking.

5.2.4 Translation to Audit Questions

This limit was chosen as it allows the auditor to focus their time and resources on areas that would have the greatest impact. The top two themes seemed too shallow of a scope whereas including four seemed too broad in terms of the balance between time and impact - a trade-off which will largely be made based on the experience of the auditor.

Taking the average ranking of Figure 5.2, which includes the ranking from both data subject subgroups proportional to their group size, was deemed a sufficient method to combine their materialities as there was a great degree of overlap between the results of the subgroups. In case different subgroups of stakeholders provide vastly different materialities to various audit themes, a more complex situation arises where the auditor may be required to assign different weights to the subgroup outcomes. Again this is a step that will have to be handled on a case-to-case basis, relying on the expertise of the auditing team to determine what approach is the best fit.

The questions from the three chosen categories in the general auditing framework were translated to the Public Eye case by applying the insights gained regarding the maturity of the AI system. Due to the full maturity of the AI system, little adjustments had to be made when translating the audit questions to the Public Eye case. If the AI system was still in development, questions concerning the lifecycle of the AI system would for example be limited to its development. The final list of questions to be used in the audit of the Public Eye AI system is included in Table E.1 of the Appendix.

The case study was concluded at this point, as it had been demonstrated that the proposed AI audit workflow could be used to scope the audit and derive a set of questions. The final steps of the workflow, which require the client to provide evidence and the auditor to write up a report, are established processes which do not warrant further exploration at this stage. Were the mock client to collaborate in the case study, their feedback on the final list of questions could have provided input for the refinement of the workflow and general auditing framework.

5.2.5 Generalization

The demonstration of the proposed AI audit workflow through this case study can be used to derive further insights on how to effectively use the workflow as presented in Figure 5.1. First of all, the auditor should strive to gain an in-depth understanding of their client and their AI system, as this will be useful when translating questions from the general auditing framework to the client case as well as in determining whether certain questions are applicable in the first place. In determining the relevant perspectives, the auditor should consider how each of the stakeholders fit into the societal, business or environmental impact perspectives. It is important here to also consider the context of the audited AI system, as it determines how these perspectives can be filled in. Additionally, the motivation of the client to subject themselves to an AI audit should be kept in mind - for example, do they wish to prevent negative press from potential scandals or are they mainly interested in the added value of the AI system to their business? Both require a vastly different auditing perspective.

The difficulty of identifying the relevant stakeholders for the chosen perspective can vary

greatly and depends on both the audited AI system and chosen perspective. Generally, the auditor ought to be thorough at this stage, as failing to include any stakeholder significantly weakens the scoping process and with that the final audit report. Next, when engaging with stakeholders the auditor should be cautious to prevent as much technical jargon as possible as this could easily confuse stakeholders that have limited knowledge on AI systems and associated risks. If concepts are unclear to stakeholders, this will negatively affect the materiality scores obtained from them. Lastly, in deriving the most material themes, the auditor should be in the lead and make a cut-off decision based on their experience and professional insight.

Chapter 6 Conclusion

The goal of this thesis was to develop an executable workflow for the audit of AI, comprised of a general auditing framework and a structured approach to scope the audit. To achieve this, a number of research questions had been posed related to AI assurance professionals, the auditing framework landscape, and the auditing workflow. This chapter is structured such that each of the research questions will be provided with answers based on the research findings. Ultimately, this research objective was found to be achieved.

6.1 RQ1: Assurance Professionals

Interviews were conducted with three assurance professionals at a Big Four accountancy who have worked on AI audits to obtain their view on the role of the AI audit, challenges therein, and subsequent design criteria for the general auditing framework and feedback on the scoping approach.

The interviews revealed that clients have only recently begun requesting their assurance on AI systems. The interviewees noted that demand is currently increasing, and set to grow even more as more AI incidents are reported in the news, and new laws and regulations come into effect. The motivation for clients to request an AI audit currently revolves largely around stakeholder management and preventing damage to their reputation through AI scandals. Additionally, the audit report adds value by serving as guidelines for further system improvements. As such, they see an important role for AI audits as an incentive for risk mitigation and to improve documentation and decision-making by developers.

The challenges experienced by the auditors relate to the unstructured nature of the current process in which questions are aggregated from various frameworks, causing issues in the determination of their relevancy and their recombination into a single framework. Design criteria that were obtained for the general auditing framework included structuring it according to the seven principles for trustworthy AI as proposed by the EC, and allowing questions to appear in multiple categories when their relevancy spans across them.

While not all interviewees regarded scoping the audit as an initial challenge, all saw merit in the proposed materiality assessment approach to scope the audit as it added further structure to the process as well as enabled the inclusion of multiple stakeholder perspectives. Their feedback on the first design iteration of the scoping approach included the need to first gain an understanding of the client and their AI system in terms of the sector, purpose, maturity and all stakeholders involved. Additionally, it included the need for the auditor to undertake all of the scoping steps, regardless of the client case as the auditor will be more knowledgeable on the full process. The order of subsequent steps based on the ESG reporting materiality assessment did not need to be changed.

6.2 RQ2: Auditing Frameworks

Existing AI auditing frameworks were investigated in order to understand the current state of the AI auditing framework landscape. This revealed which frameworks had been published, how they compared to one another as well as to related IT auditing and AI monitoring practices.

A literature search lead to the identification of fourteen proposed AI auditing frameworks. The framework typology revealed that the frameworks could be subdivided into three source categories: those proposed by academic sources, industry and auditing/regulatory organizations. Academic frameworks were typically developed from a specific aspect relating to trustworthy AI, whereas the industry frameworks were developed with the idea that public trust is a requirement for further AI developments in mind. The academic and industry frameworks were mostly high-level, whereas the frameworks proposed by auditing and regulatory organizations were typically more detailed. This was reflected in the average number of questions in the frameworks at 19, 16, and 75 respectively. It was furthermore revealed that over time, the published frameworks shifted form high-level and exploratory to more detailed as the field of AI assurance matured. Interestingly, Google and OpenAI were amongst the first to propose auditing frameworks, thereby illustrating their awareness of the need for public trust in their own AI systems.

The focus of the fourteen frameworks was further revealed through open and axial coding, through which 23 themes covered by the various questions were obtained. For each of the academic frameworks, the most prevalent themes largely aligned with the described research perspective. Most of the auditing/regulatory frameworks typically focused on documentation, quality, or control, all of which are pivotal topics in the field of assurance.

Comparison to four IT auditing standards and frameworks showed that the differences between IT systems and AI systems are also reflected in the frameworks. IT audit frameworks and standards revolve around risk management, security, performance and value creation. As AI systems are inherently unpredictable to a degree, and their application within complex socio-technological contexts involves multiple stakeholders, the AI auditing frameworks need to address a broader range of themes. These include fairness, explainability, transparency amongst other ethical topics.

Analysis of a state of the art AI monitoring dashboard revealed that these dashboards could be used to cover transparency and performance monitoring aspects of an AI audit. Its shortcomings in an auditing context relate to its insights being limited to what is already known about for example data subpopulations, as well as that the scope of an AI audit typically extends beyond the technical details of the model and includes many context-dependent topics. A combination consisting of just an IT audit framework and an AI monitoring dashboard would therefore not suffice for an AI audit.

6.3 RQ3: AI Audit Workflow

Finally, a workflow was proposed for the AI audit. This required the recombination of the questions gathered from the identified frameworks along determined design criteria

into a general auditing framework. Additionally, the corporate ESG reporting problem of deciding on the materiality of topics was compared to the AI audit scoping problem. Following this comparison, the ESG reporting materiality assessment could be translated to an AI audit scoping approach using feedback from the assurance professionals. The workflow was then demonstrated through a case study.

A general auditing framework was obtained through recombination of the 595 auditing questions gathered from the 14 frameworks along the seven principles of trustworthy AI in accordance with the design criteria. Questions spanning more than one of the principles were not reduced to a single principle, also in accordance with the design criteria. Societal and environmental well-being remained the least detailed category of the framework, likely due to it being the broadest category. Technical robustness and safety on the other hand contained the most questions due to the specificity of the category as well as the large proportion of source frameworks that took on a technical perspective or took IT auditing as a starting point.

The corporate ESG reporting materiality assessment was shown to be a problem that shared many similarities with the AI audit scoping problem, as both lack standardized methodologies and come with inherent tensions between divergent stakeholder demands, resulting in competing organizational goals. Additionally, the topics for both are complex, uncertain and evaluative in nature. As such, it was possible to construct a scoping process for the AI audit based on the ESG reporting materiality assessment steps proposed in literature, in combination with feedback from the assurance professionals. The importance of gaining an understanding of the client and their AI system, as well as the leading role of the auditor throughout the scoping process were highlighted.

The scoping process was combined with the general auditing framework and conventional auditing steps to derive a workflow for the audit of AI. This workflow was subsequently demonstrated to be applicable in a mock audit of the Public Eye AI system of the Municipality of Amsterdam, a crowd monitoring system currently operational at various locations. This system was chosen because of the extensive amount of publicly available information about it. An understanding of the mock client could be gained based on this information, after which the data subject perspective was chosen for the audit. Frequenters and occasional visitors of one of the monitored areas were identified as relevant stakeholders and subsequently interviewed in order to determine which audit topics were most material to them and should therefore be included in the audit. This resulted in a list of questions for the audit of Public Eye that covered privacy and data governance, transparency, and diversity, non-discrimination and bias.

Chapter 7 Reflection and Recommendations

Following the conducted research, some further reflection on the findings, the research process and their limitations is offered in this chapter. Points of reflection are subsequently translated to recommendations for future research. The structure of this chapter follows a division along three levels of specificity: the mock audit case study, the proposed audit workflow, and the AI audit in general. Lastly, a reflection on the relevance of the research will be offered.

7.1 Case study

During the case study, the audit workflow was limited to the step where a list of scoped audit questions was derived. Ideally, there would have been cooperation with the mock client as to obtain feedback on the client side of the proposed audit process. This could prove insightful, for example in refining the questions in the audit framework in future design cycles.

Furthermore, the case study was limited by the sample of stakeholders that was interviewed. A substantial number, five out of twenty, provided conflicting answers to the materiality assessment. It was also noted that the topics that the interviewed stakeholders found most material were not always as objectively relevant as they perceived them to be. For example, the fairness of the AI system ranked high in materiality; an auditor, who has gained an understanding of the AI system, would note that the risk of biases is minimal due to the way Public Eye has been designed. As such, there is a difference in the level of understanding of the audited AI system between the auditor and the inquired stakeholders which affects the scoping process.

Instead of interviewing people on location, organizing focus groups in order to obtain the materiality scores from stakeholders might be a better suited strategy for the audit. That way, the auditor is in a position to explain topics as well as to ask for clarification in case stakeholders provide conflicting statements regarding which topics they find material. This would also add a level of control over the demographic make-up of the group of participating stakeholders, which during the case study was random and circumstantial to the specific day the interviews were conducted.

7.2 Audit Workflow

The proposed audit workflow limits the scoping through a materiality assessment of the seven predetermined audit themes (one of the design criteria). It may however occur that those themes, while in principle sufficiently covering the aspects of trustworthy AI, do not exactly match the topics which stakeholders find most material. Additionally, determining a cut-off for which of the most material topics are included in the audit scope remains a decision which needs to be made by the auditor on a case-to-case basis.

A second layer of auditing subthemes may be added to the general auditing framework. These could then be used to fine-tune the scoping. The range of auditing topics becomes more detailed by including subthemes, each of which would relate to a smaller number of auditing questions within one of the seven principle categories. Through this, the materiality cut-off for the scoping will also become more subtle as the auditor does not have to decide on the inclusion of whole categories of auditing questions. It should be noted that this would add a layer of complexity to the process. This may be undesirable for the audit practice in its current state, given the indicated need for a structured audit workflow.

To further improve the audit process, the option to use an AI monitoring dashboard such as DDSS to automate a part of the audit could be explored. A monitoring tool could cover some of the technical questions of the audit, which followed from its comparison to the AI auditing frameworks and was mentioned by two of the interviewed auditors (I2, I3). Another aspect that could be further explored is the incorporation of industry standards, which the current workflow does not specify. This could be included as additional step in gaining an understanding of the sector at the start of the audit.

Another way to refine the audit process is through its application. In line with the inductive step in the meta-framework of Partelow, empirical observations when using the auditing workflow can provide new design criteria (Partelow, 2023). Additionally, through its use some exemplary cases can be obtained for future reference to aid auditors in using the framework - as was also a suggestion by one of the interviewees (I2).

The workflow is currently linear by design, which followed from the linear nature of the scoping approach and the conventional audit steps. It might, however, not be the best way to structure the audit - it may be beneficial to incorporate an iterative process whereby there is an option to expand the perspectives included in the scoping phase at a later stage in the audit. For this as well as the other suggestions, it is recommended that a future design iteration includes an even closer collaboration with the end-users - the AI auditors - to co-develop further improvements. This would add efficiency to the design process, and allows for group collaborations where differing perspectives can be aligned.

Lastly, the interviewed assurance professionals were all employed at the same accountancy firm, which narrows the included perspectives. It is therefore recommended that for future research the opinions of AI auditors from multiple accountancy firms are investigated for a broader perspective.

7.3 AI Audit

The starting point of the research was the indication by end-users that they would benefit from the development of a general auditing framework for AI. In combination with time limits, this meant that no alternative approaches to the AI audit were explored. For future research it is therefore recommended to explore other designs to the AI audit workflow, such as a (partly) continuous audit.

Furthermore, the AI audit should not be seen as a standalone solution for risk mitigation in AI systems. After all, an audit will always be "after the fact" (I3). Instead, it is

part of a larger toolbox which also includes complementary approaches such as developer transparency, pro-active stakeholder engagement and harmonized rules and regulations. These are thought to be most effective in tandem.

Since there are no concrete legal guidelines, no 'true assurance' can be granted (i.e. regulatory compliance) and the current AI audit is more akin to a third-party report of findings and observations. Once concrete rules and regulations are drafted, this will in turn require a change to the audit process in order to accommodate the assessment of regulatory compliance. This development is currently taking place in the related field of sustainability and ESG auditing, with the EU Corporate Sustainability Reporting Directive (CSRD) coming into effect in 2024, imposing specific reporting requirements. It is therefore recommended that for future iterations of the AI workflow lessons learned from the current developments in the CSRD audit process are taken into account.

A more fundamental point is that the value of an audit is closely tied to the level of trust stakeholders place in the auditor. As mentioned by the interviewed auditors, reputation of the accountants is important to their clients (I1, I3). Meanwhile, the Big Four accountants in the Netherlands are not spared of scandals as the Dutch Authority for Financial Markets (AFM) is currently in the midst of a large-scale investigation of fraud amongst accountants (Pols & van der Schoot, 2023). The added value of an AI audit may therefore also shift with the public sentiment towards accountants.

7.4 Relevancy

One scientific contribution of this thesis is providing an overview and characterization of the various AI auditing frameworks that had been proposed in literature. Additionally, their harmonization into a general auditing framework has been a way to bridge the various fields and sources that have brought these framework forward. Similarly, the commonalities between corporate ESG reporting and AI auditing with regards to scoping have been demonstrated. This was subsequently used to translate insights from the ESG reporting materiality assessment to an AI audit scoping approach, thereby bridging the two fields. Lastly, the research has been an exploration of the AI audit as a practice that is emerging from the IT audit, and how the complexity of the socio-technological aspects of AI demand a unique auditing approach.

Societal relevance of this work lies in the AI audit being one of the ways to incentivize the development of trustworthy AI systems. Currently, people have suffered from AI systems which propagate unethical biases - as illustrated in the news over the past year. An effective and structured AI audit is thought to be one of the ways to foster more conscious development of AI systems, thereby reducing AI related risks for society. Especially relevant is the proposed scoping approach, in which members of society, as stakeholders, are actively included in the identification of which topics related to trustworthy AI are most important to them.

The business relevance of this thesis is two-fold. On the one hand, the proposed AI audit workflow is relevant for accountancy firm that are starting to engage in AI audits. They are thought to benefit from this proposed approach, as it provides a uniform structure to

the audit approach - especially since its design has been based on input from AI auditors. If needed, the questions in the general auditing framework can be adapted to accommodate any further organizational preferences. Furthermore, it has been shown that an AI audit requires much more intensive stakeholder engagement than IT audits, which is relevant for the training of future AI auditors.

On the other hand, organizations that develop or operate AI systems can also benefit from the proposed approach. This is not just limited to when they seek external assurance on their AI system. They could also use the proposed workflow as a way to self-assess their AI system and governance, while also gaining insights in stakeholder concerns through the scoping approach.

References

Aditya, B. R., Hartanto, R., & Nugroho, L. E. (2018, 9). The Role of IT Audit in the Era of Digital Transformation. In *Iop conference series: Materials science and engineering* (Vol. 407). Institute of Physics Publishing. doi: 10.1088/1757-899X/407/1/012164

AI HLEG. (2019). *A Definition of AI: Main Capabilities and Scientific Disciplines.* European Commission. Available at: https://ec. europa. eu/digital-single˜....

AI HLEG. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment* (Tech. Rep.). Retrieved from `https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence` doi: 10.2759/791819

Ajao, O. S., Olamide, J. O., & Temitope, A. A. (2016). Evolution and development of auditing. *Unique Journal of Business Management Research*, *3*(1), 32–040.

Akula, R., & Garibay, I. (2021, 7). Audit and Assurance of AI Algorithms: A framework to ensure ethical algorithmic practices in Artificial Intelligence. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/2107.14046`

Albeda, J. (2020). Industrie 4.0: de risico's en aanbevelingen. *Audit Magazine*, 1–4.

Algemene Rekenkamer. (2021). Aandacht voor algoritmes. *Algemene Rekenkamer*, 1–66.

Allam, Z., & Dhunny, Z. A. (2019). On big data, artificial intelligence and smart cities. *Cities*, *89*(November 2018), 80–91. doi: 10.1016/j.cities.2019.01.032

Allen, G. (2020). Understanding AI Technology. *Joint Artificial Intelligence Center (JAIC)*(April), 20. Retrieved from `https://www.ai.mil/docs/UnderstandingAITechnology.pdf`

Al-Rubaie, M., & Chang, J. M. (2019, 3). Privacy-Preserving Machine Learning: Threats and Solutions. *IEEE Security and Privacy*, *17*(2), 49–58. doi: 10.1109/MSEC.2018.2888775

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. (2021). Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(5), 1–13. doi: 10.1002/widm.1424

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., ... Zilberman, N. (2021). Filling gaps in trustworthy development of AI. *Science*, *374*(6573), 1327–1329. doi: 10.1126/science.abi7176

AXELOS. (2020). *ITIL4: Direct, Plan and Improve*. London: TSO.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020, 6). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. doi: 10.1016/j.inffus.2019.12.012

Batarseh, F. A., Freeman, L., & Huang, C. H. (2021). A survey on artificial intelligence assurance. *Journal of Big Data*, *8*(1). Retrieved from https://doi.org/10.1186/s40537-021-00445-7 doi: 10.1186/s40537-021-00445-7

Bazerman, M. H., Morgan, K. P., & Loewenstein, G. F. (1997). The Impossibility of Auditor Independence. *Sloan Management Review*, *38*(4), 89–95.

Becker, N., & Waltl, B. (2022). Auditing and Testing AI – A Holistic Framework. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 13320 LNCS, pp. 283–292). Springer Science and Business Media Deutschland GmbH. doi: 10.1007/978-3-031-06018-2{_}20

Black, J., & Murray, A. (2019). Regulating AI and machine learning: setting the regulatory agenda (complementar). *European Journal of Law and Technology*, *10*(3), 1–17. Retrieved from http://eprints.lse.ac.uk/102953/4/722_3282_1_PB.pdf

Boer, A., de Beer, L., & van Praat, F. (2023). Algorithm Assurance: Auditing Applications of Artificial Intelligence. In *Advanced digital auditing* (pp. 237–256). doi: 10.1007/978-3-031-11089-4{_}9

Bratton, W. (2002). Enron and the Dark Side of Shareholder Value. *Tulsa Law Review*.

Bridge, O., Raper, R., Strong, N., & Nugent, S. E. (2021). Modelling a socialised chatbot using trust development in children: lessons learnt from Tay. *Cognitive Computation and Systems*, *3*(2), 100–108. doi: 10.1049/ccs2.12019

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, *8*(1), 2053951720983865.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... others (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, *3*(1), 1–12. doi: 10.1177/2053951715622512

Castelvecchi, D. (2016). The black box 2 0 —. *Nature*, *538*(7623), 20–23. Retrieved from http://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731

Clark, A. (2018). The Machine Learning Audit-CRISP-DM Framework. *ISACA Journal*, *1*. Retrieved from https://www.isaca.org/-/media/files/isacadp/project/isaca/articles/journal/2018/volume-1/the-machine-learning-audit-crisp-dm-framework_joa_eng_0118.pdf

Dataiku. (2022). *Dataiku Announces $200 Million Investment Led by Wellington Management.* Retrieved from https://www.dataiku.com/press-releases/series-f-pr/

Dataiku. (2023a). *Dataiku Hires Software Industry Veteran Krish Venkataraman as President as Everyday AI Leader Surpasses $230M ARR.* Retrieved from https://www.dataiku.com/press-releases/krish-venkataraman-president/

Dataiku. (2023b). *DSS Key Capabilities: Explainability.* Retrieved from https://videos.dataiku.com/watch/4wWBjCv8cnDQGRzKbKovU2

Davis, G. B., Adams, D. L., & Schaller, C. A. (1968). *Auditing & EDP.* American Institute of Certified Public Accountants.

de Boer, M., & van Geijn, H. (2021). NOREA Guiding Principles Trustworthy AI Investigations. *NOREA de beroepsorganisatie van IT-auditors*, 1–41.

De Haes, S., Van Grembergen, W., Joshi, A., & Huygh, T. (2020). COBIT as a Framework

for Enterprise Governance of IT. In *Management for professionals* (Vol. Part F574, pp. 125–162). Springer Nature. doi: 10.1007/978-3-030-25918-1{_}5

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. In *Aip conference proceedings* (Vol. 1644, pp. 97–104). American Institute of Physics Inc. doi: 10.1063/1.4907823

Dlamini, Z., Francies, F. Z., Hull, R., & Marima, R. (2020). Artificial intelligence (AI) and big data in cancer and precision oncology. *Computational and Structural Biotechnology Journal*, *18*, 2300–2311. Retrieved from `https://doi.org/10.1016/j.csbj.2020.08.019` doi: 10.1016/j.csbj.2020.08.019

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial intelligence and business value: A literature review. *Information Systems Frontiers*, *24*(5), 1709–1734.

Ferrer, X., Nuenen, T. V., Such, J. M., Cote, M., & Criado, N. (2021). Bias and Discrimination in AI: A Cross-Disciplinary Perspective. *IEEE Technology and Society Magazine*, *40*(2), 72–80. doi: 10.1109/MTS.2021.3056293

Floridi, L., Holweg, M., Taddeo, M., Silva, J. A., Mökander, J., & Wen, Y. (2022). *capAI: A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act* (Tech. Rep.). Retrieved from `https://ssrn.com/abstract=4064091`

Future of Life Institute. (2023a, 3). *Pause Giant AI Experiments: An Open Letter.*

Future of Life Institute. (2023b). *Policymaking in the Pause* (Tech. Rep.).

Gallegos, F. (2003). SARBANES–OXLEY ACT OF 2002 (PL 107-204) AND IMPACT ON THE IT AUDITOR. *The EDP Audit, Control, and Security Newsletter*.

Gantz, S. D. (2014). IT Audit Components. In *The basics of it audit* (pp. 105–128). Elsevier. doi: 10.1016/b978-0-12-417159-6.00006-7

Garst, J., Maas, K., & Suijs, J. (2022). Materiality Assessment Is an Art, Not a Science: Selecting ESG Topics for Sustainability Reports. *California Management Review*, *65*(1), 64–90. doi: 10.1177/00081256221120692

Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. (2018). Why we need to audit algorithms. *Harvard Business Review*.

Hafner, G. F. (1964). Auditing EDP. *The Accounting Review*, *39*(4), 979–982.

Hasan, A. R. (2022). Artificial Intelligence (AI) in Accounting & Auditing: A Literature Review. *Open Journal of Business and Management*, *10*(01), 440–465. doi: 10.4236/ojbm.2022.101026

Hsu, C., Wang, T., & Lu, A. (2016). The Impact of ISO 27001 Certification on Firm Performance. In *2016 49th hawaii international conference on system sciences (hicss)* (pp. 4842–4848). doi: 10.1109/HICSS.2016.600

Hua, S.-S., & Belfield, H. (2020). AI & Antitrust: Reconciling Tensions between Competition Law and Cooperative AI Development. *Yale JL & Tech.*, *23*, 415.

Hwang, T. (2018). Computational power and the social impact of artificial intelligence. *SSRN*, 1–44.

ISACA. (2018). *Auditing Artificial Intelligence* (Tech. Rep.). Retrieved from `https://www.pearson.com/us/higher-`

Jagatheesaperumal, S. K., Rahouti, M., Ahmad, K., Al-Fuqaha, A., & Guizani, M. (2022). The Duo of Artificial Intelligence and Big Data for Industry 4.0: Applications, Techniques, Challenges, and Future Research Directions. *IEEE Internet of Things*

Journal, *9*(15), 12861–12885. doi: 10.1109/JIOT.2021.3139827

Jordan, M., & Mitchell, T. (2015, 7). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. doi: 10.1126/science.aac4520

Kaplan, R. S., & Ramanna, K. (2021). How to Fix ESG Reporting. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3900146

Kelley, P. G., Yang, Y., Heldreth, C., Moessner, C., Sedley, A., Kramm, A., ... Woodruff, A. (2021). Exciting, Useful, Worrying, Futuristic: Public Perception of Artificial Intelligence in 8 Countries. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 627–637. doi: 10.1145/3461702.3462605

Knott, E., Rao, A. H., Summers, K., & Teeger, C. (2022). Interviews in the social sciences. *Nature Reviews Methods Primers*, *2*(1). doi: 10.1038/s43586-022-00150-6

Lander, G. P. (2002). The Sarbanes-Oxley Act of 2002. *Journal of Investment Compliance*, *3*(1), 44–53. doi: 10.1108/joic.2002.3.1.44

Larson, J., Matt, S., Kirchner, L., & Angwin, J. (2016, 5). *How We Analyzed the COMPAS Recidivism Algorithm.* Retrieved from `propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`

L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine Learning with Big Data: Challenges and Approaches. *IEEE Access*, *5*, 7776–7797. doi: 10.1109/ACCESS.2017.2696365

Li, H., Yu, L., & He, W. (2019). The Impact of GDPR on Global Technology Development. *Journal of Global Information Technology Management*, *22*(1), 1–6. doi: 10.1080/1097198X.2019.1569186

Linthicum, C., Reitenga, A. L., & Sanchez, J. M. (2010). Social responsibility and corporate reputation: The case of the Arthur Andersen Enron audit failure. *Journal of Accounting and Public Policy*, *29*(2), 160–176. Retrieved from `http://dx.doi.org/10.1016/j.jaccpubpol.2009.10.007` doi: 10.1016/j.jaccpubpol.2009.10.007

Liu, H., Wang, Y., Fan, W., Liu, X., Jain, S., Liu, Y., ... Tang, J. (2022). Trustworthy AI: a computational perspective. *ACM Transactions on Intelligent Systems and Technology*, *14*(1), 1–59. doi: 10.1145/nnnnnnn.nnnnnnn

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381–386. doi: 10.21275/ART20203995

Mancham, D. P. J. (2007). *Excuse me, do you speak ITGC?* (Tech. Rep. No. 1).

Markert, T., Langer, F., & Danos, V. (2022). GAFAI: Proposal of a Generalized Audit Framework for AI. In *Lecture notes in informatics (lni), proceedings - series of the gesellschaft fur informatik (gi)* (Vol. P-326, pp. 1247–1256). Gesellschaft fur Informatik (GI). doi: 10.18420/inf2022{_}107

McGregor, S. (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, *17B*, 15458–15463. doi: 10.1609/aaai.v35i17.17817

Misra, N. N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R., & Martynenko, A. (2022). IoT, Big Data, and Artificial Intelligence in Agriculture and Food Industry. *IEEE Internet of Things Journal*, *9*(9), 6305–6324. doi: 10.1109/JIOT.2020.2998584

Mitrou, L. (2018). Data protection, artificial intelligence and cognitive services: is the general data protection regulation (GDPR)'artificial intelligence-proof'? *SSRN*, 1–90.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of clinical epidemiology*, *62*(10), 1006–1012. doi: 10.1016/j.jclinepi.2009.06.005

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework. *NIST*, 1–48.

NOS. (2023). *DUO mag algoritme niet gebruiken totdat meer bekend is over mogelijke discriminatie.*

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. doi: 10.14445/22312803/ijctt-v48p126

Partelow, S. (2023). What is a framework? Understanding their purpose, value, development and use. *Journal of Environmental Studies and Sciences*.

Peres, R. S., Jia, X., Lee, J., Sun, K., Colombo, A. W., & Barata, J. (2020). Industrial Artificial Intelligence in Industry 4.0 -Systematic Review, Challenges and Outlook. *IEEE Access*, 220121–220139. doi: 10.1109/ACCESS.2020.3042874

Pols, M., & van der Schoot, E. (2023). *Zelfs de toets van de eigen beroepsorganisatie is voor accountants niet heilig.* Retrieved from `https://fd.nl/financiele-markten/1493585/zelfs-de-toets-van-de-eigen-beroepsorganisatie-is-voor-accountants-niet-heilig`

PwC. (2023). *PwC 2023 Trust Survey* (Tech. Rep.). Retrieved from `https://www.pwc.com/us/en/library/trust-in-business-survey-2023.html`

Radclyffe, C., Ribeiro, M., & Wortham, R. H. (2023). The assessment list for trustworthy artificial intelligence: A review and recommendations. *Frontiers in Artificial Intelligence*.

Radovanović, D., Radojević, T., Lučić, D., & Šarac, M. (2010). IT audit in accordance with Cobit standard. *MIPRO 2010 - 33rd International Convention*, 1137–1141.

Radulescu, R., & Pestritu, L. (2011). Termination of Supplier-Buyer Relationship. *Global Conference on Business and Finance Proceedings*, *6*(2), 142–149. Retrieved from `http://www.researchgate.net/publication/233370687_Tantalite_Production_in_Zimbabwe_and_the_Role_of_World_Price_Trends_in_the_Past_Three_Decades/file/79e4150eda9943a203.pdf#page=1125`

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... Barnes, P. (2020, 1). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Fat* 2020 - proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 33–44). Association for Computing Machinery, Inc. doi: 10.1145/3351095.3372873

Rezaee, Z. (2004). Restoring public trust in the accounting profession by developing anti-fraud education, programs, and auditing. *Managerial Auditing Journal*, *19*(1), 134–148. doi: 10.1108/02686900410509857

Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society*, *63*(October), 101421. Retrieved from `https://doi.org/10.1016/j.techsoc.2020.101421` doi: 10.1016/j.techsoc.2020.101421

Rudin, C. (2019, 5). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead* (Vol. 1) (No. 5). Nature Research.

doi: 10.1038/s42256-019-0048-x

Saetra, H. S. (2022, 4). The AI ESG protocol: Evaluating and disclosing the environment, social, and governance implications of artificial intelligence capabilities, assets, and activities. *Sustainable Development*. doi: 10.1002/sd.2438

Sandu, I., Wiersma, M., & Manichand, D. (2022, 9). Time to audit your AI algorithms. *Maandblad voor Accountancy en Bedrijfseconomie*, *96*(7/8), 253–265. doi: 10.5117/mab.96.90108

Santos, R. C., & Martinho, J. L. (2020). An Industry 4.0 maturity model proposal. *Journal of Manufacturing Technology Management*, *31*(5), 1023–1043. doi: 10.1108/JMTM-09-2018-0284

Schuett, J. (2023). Risk Management in the Artificial Intelligence Act. *European Journal of Risk Regulation*, *2017*, 1–19. doi: 10.1017/err.2023.1

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human Computer Studies*, *146*(April 2020), 102551. Retrieved from `https://doi.org/10.1016/j.ijhcs.2020.102551` doi: 10.1016/j.ijhcs.2020.102551

Shyam, R., & Singh, R. (2021). A Taxonomy of Machine Learning Techniques. *Advancement in Robotics*. Retrieved from `http://computers.stmjournals.com/index.php?journal=JoARB&page=index` doi: 10.37591/JoARB

Smuha, N., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU can achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. *SSRN*, 1–17.

Stoel, D., Havelka, D., & Merhout, J. W. (2012, 3). *An analysis of attributes that impact information technology audit quality: A study of IT and financial audit practitioners* (Vol. 13) (No. 1). doi: 10.1016/j.accinf.2011.11.001

Stoica, I., Song, D., Popa, R. A., Patterson, D., Mahoney, M. W., Katz, R., ... Abbeel, P. (2017, 12). A Berkeley View of Systems Challenges for AI. *arXiv preprint*. Retrieved from `http://arxiv.org/abs/1712.05855`

Tambiama Madiega. (2023). BRIEFING - EU Legislation in Progress - Artifical intelligence act. *EPRS — European Parliamentary Research Service*(June).

The Institute of Internal Auditors. (2018). *Global Perspectives and Insights: Artificial Intelligence - Considerations for the Profession of Internal Auditing* (Tech. Rep.). Retrieved from `www.theiia.org/gpi.`

Tiron-Tudor, A., & Bota-Avram, C. (2013). European Union Directive - The 8th Company Law Directive on Disclosure & Transparency. In *Encyclopedia of corporate social responsibility* (pp. 1097–1104).

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. doi: 10.9785/cri-2021-220402

von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philosophy and Technology*, *34*(4), 1607–1622. Retrieved from `https://doi.org/10.1007/s13347-021-00477-0` doi: 10.1007/s13347-021-00477-0

Wagner, B. (2019). Ethics as an escape from regulation. From "ethics-washing" to ethics-shopping? *Amsterdam University Press*, 84–89. doi: 10.2307/j.ctvhrd092.18

Wang, P. (2019, 1). On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, *10*(2), 1–37. doi: 10.2478/jagi-2019-0002

Wang, X., Han, Y., Leung, V. C., Niyato, D., Yan, X., & Chen, X. (2020). Convergence of Edge Computing and Deep Learning: A Comprehensive Survey. *IEEE Communications Surveys and Tutorials*, *22*(2), 869–904. doi: 10.1109/COMST.2020.2970550

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. *Journal of Industrial Information Integration*, *23*(March), 100224. Retrieved from `https://doi.org/10.1016/j.jii.2021.100224` doi: 10.1016/j.jii.2021.100224

Zhang, Y., Weng, Y., & Lund, J. (2022). Applications of Explainable Artificial Intelligence in Diagnosis and Surgery. *Diagnostics*, *12*(2). doi: 10.3390/diagnostics12020237

Zuiderveen Borgesius, F., & others. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. *Council of Europe*.

Zuiderwijk, A., Chen, Y. C., & Salem, F. (2021, 7). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3). doi: 10.1016/j.giq.2021.101577

Appendices

A AI Overview

Broadly generalizing, all AI systems can be reduced to three technical components: some form of input data, a model and the output data generated by the (trained) model based on the input it received (Allen, 2020). There are many degrees of complexity in how an AI model is structured, with the most basic ones consisting of rules that are written directly by their developers. This bypasses the need for the model to generate its own data insights. Such rule-based or knowledge-based AI models typically consist of a sequence of if-then expressions. Depending on whether the model input satisfies the predetermined criteria, the model will produce the associated hard-coded response.

More complex models use an algorithm which is first trained using a training dataset, from which it distills patterns - this is referred to as Machine Learning (ML) (Shyam & Singh, 2021). The training data will be formatted and chosen such that it is akin to the input data that the AI system is intended to process once the system is deployed. The method by which the model is trained can vary based on the underlying algorithm.

Supervised learning algorithms are trained on data which is labeled, meaning that it is fed examples of input data and the desired associated output data (Osisanwo et al., 2017). The algorithm will then (through statistical techniques) within the training dataset attempt to derive a relation between characteristics of the input data and the corresponding output data. This relationship is then supposed to hold in the context in which the AI system is deployed. Supervised learning models are used to solve problems related to classification and predictions based on regression among others.

Unsupervised models on the other hand are trained on unlabeled data, where there is no indication beforehand on how datapoints are supposed to relate to one another (Mahesh, 2020). This means that the link between (subgroups of) datapoints needs to be uncovered through the statistical method on which the algorithm is based. The connections derived from the training data are then used once the AI system is deployed to characterize new input data. Unsupervised learning models are used in problems revolving around data clustering and reducing data complexity, among others.

More recent developments in the successes achieved in the AI field include reinforcement learning and deep learning algorithms. While both are not novel concepts, their success is more recent as their strengths have been enabled by the availability of increasingly larger datasets and advancements in computer processing power (Jordan & Mitchell, 2015).

Reinforcement learning differs from supervised and unsupervised learning as the model receives feedback based on the quality of its output (Shyam & Singh, 2021). When the feedback is positive, this will then be processed such that the internal pathways that lead from the given input to the correct output are reinforced. If the model receives negative feedback, this will result in the internal pathways being restructured with the expectation that this will lead to a better output for the next input. RL models are used in situations where there are many outputs possible due to the complex nature of the environment in which the system operates, therefore requiring a trial and error approach.

Deep learning (DL) does not refer to a method of learning, but rather describes the way the model is structured. DL models are based on artificial neural networks that are inspired by the way neurons are connected and transmit information in nature. The DL model consists of various levels of depths which allows for the extraction of information on increasing levels of abstraction (Barredo Arrieta et al., 2020). This allows for DL to be used for tasks which relate to the recognition of complex, abstract patterns and processing of large amounts of data.

B Reviewed Frameworks

For each of the analysed frameworks, a separate table is included that contains the elements as identified during the open coding phase. The coding labels derived through axial coding are shown in Table B.1 and labelling results are also included per framework in Table B.2 through Table B.15. Framework elements are numbered to facilitate referencing.

Table B.1: *Axial Coding Labels.*

Coding labels	
Accountability	Periodic Assessment
Autonomy	Privacy
Control	Quality
Direct environmental impact	Reliability
Documentation	Risk assessment
Explainability	Robustness
Fairness	Safety
Human involvement	Security
Indirect environmental impact	Stakeholders
Legality	System description
Management of system	Transparency
Objectives	

B.1 Assessment List for Trustworthy Artificial Intelligence

The Assessment List for Trustworthy AI was published in 2020 (AI HLEG, 2020). Its identified framework elements are shown in Table B.2 below.

Table B.2: *Assessment List for Trustworthy Artificial Intelligence Codes.*

Ref	Code	Code Groups		
1	Does the AI potentially negatively discriminate?	Fairness		
2	Is bias assessed during development, deployment and use of the AI?	Fairness	Periodic Assessment	
3	Are processes in place to address and rectify potential bias?	Fairness	Management of system	
4	Does the AI respect the rights of the child?	Legality	Safety	
5	Are processes in place to address and rectify harm to children by the AI?	Safety	Management of system	
6	Is the AI system tested and monitored for potential harm to children?	Risk Assessment	Control	Safety
7	Does the AI system protect personal data relating to individuals in line with GDPR?	Privacy	Security	
8	How is the need for a DPIA assessed?	Risk Assessment	Privacy	
9	How is personal data protected in the development, deployment and use of the AI?	Periodic Assessment	Security	Privacy
10	Does the AI system respect the freedom of expression and information and/or freedom of assembly and association?	Legality	Stakeholders	

Continued on next page

Table B.2 – *Continued from previous page*

Ref	Code	Code Groups			
11	Are processes in place to test and monitor for potential infringement on freedom of expression and information, and/or freedom of assembly and association, during the development, deployment and use phases of the AI system?	Legality	Stakeholders	Periodic Assessment	
12	Have processes been put in place to address and rectify for potential infringement on freedom of expression and information, and/or freedom of assembly and association, in the AI system?	Control	Legality	Management of system	Stakeholders
13	Is the AI system designed to interact, guide or take decisions by human end-users that affect humans or society?	Stakeholders	Objectives	System description	
14	Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision?	Stakeholders	Transparency		
15	Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision?	Transparency	Stakeholders		
16	Could the AI system generate confusion for some or all end-users or subjects on whether they are interacting with a human or AI system?	Stakeholders	Transparency		
17	Could the AI system affect human autonomy by generating over-reliance by end-users?	Autonomy	Stakeholders		
18	Have procedures been put in place to avoid that end-users over-rely on the AI system?	Stakeholders	Control	Management of system	
19	Could the AI system affect human autonomy by interfering with the end-user's decision-making process in any other unintended and undesirable way?	Stakeholders	Risk Assessment		
20	Is theree any procedure to avoid that the AI system inadvertently affects human autonomy?	Risk Assessment	Management of system		
21	Does the AI system simulate social interaction with or between end-users or subjects?	Stakeholders	System description		
22	Does the AI system risk creating human attachment, stimulating addictive behaviour, or manipulating user behaviour?	Risk Assessment	Stakeholders		
23	Have the humans been given specific training on how to exercise oversight?	Management of system	Human involvement	Control	Accountability
24	Is there any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?	Risk Assessment	Control	Stakeholders	
25	Is there a 'stop button' or procedure to safely abort an operation when needed?	Control	Security		
26	Are there any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?	Periodic Assessment	Control	Autonomy	
27	Could the AI system have adversarial, critical or damaging effects (e.g. to human or societal safety) in case of risks or threats such as design or technical faults, defects, outages, attacks, misuse, inappropriate or malicious use?	Risk Assessment	Stakeholders	Safety	
28	Is the AI system certified for cybersecurity or is it compliant with specific security standards?	Security	Legality		
29	How exposed is the AI system to cyberattacks?	Security			
30	Have potential forms of attacks to which the AI system could be vulnerable been assessed?	Security	Risk Assessment		
31	Have different types of vulnerabilities and potential entry points for attacks been considered?	Security	Risk Assessment		

Continued on next page

Table B.2 – *Continued from previous page*

Ref	Code	Code Groups			
32	Have measures been put in place to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle?	Security	Robustness	Periodic Assessment	Management of system
33	Has the system been tested for security?	Security			
34	Have end-users been informed of the duration of security coverage and updates?	Stakeholders	Security	Transparency	
35	What length is the expected timeframe within which security updates for the AI system are provided?	System description	Security		
36	Have risks, risk metrics and risk levels of the AI system been defined in each specific use case?	Risk Assessment	Documentation		
37	Are processes put in place to continuously measure and assess risks?	Periodic Assessment	Risk Assessment		
38	Have end-users and subjects been informed of existing or potential risks?	Risk Assessment	Transparency	Stakeholders	
39	Have the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible consequences been identified?	Risk Assessment			
40	Has the risk of possible malicious use, misuse or inappropriate use of the AI system been assessed?	Risk Assessment	Security		
41	Have safety criticality levels (e.g. related to human integrity) of the possible consequences of faults or misuse of the AI system been defined?	Safety	Documentation	Risk Assessment	
42	Has the dependency of a critical AI system's decisions on its stable and reliable behaviour been assessed?	Risk Assessment	Reliability	Robustness	
43	Are the reliability/testing requirements aligned to the appropriate levels of stability and reliability?	Reliability	Robustness		
44	Is there a duplicated system in order to remain operational when the original system fails?	Safety	Management of system		
45	Has a mechanism been developed to evaluate when the AI system has been changed to merit a new review of its technical robustness and safety?	Robustness	Safety	Periodic Assessment	Management of system
46	Could a low level of accuracy of the AI system result in critical, adversarial or damaging consequences?	Quality	Risk Assessment		
47	Have measures been put in place to ensure that the data (including training data) used to develop the AI system is up-to-date, of high quality, complete and representative of the environment the system will be deployed in?	Quality			
48	Has a series of steps been put in place to monitor, and document the AI system's accuracy?	Quality	Documentation	Control	Management of system
49	Has been considered whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects?	Safety	Risk Assessment		
50	Have processes been put in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated?	Transparency	Stakeholders	Quality	
51	Could the AI system cause critical, adversarial, or damaging consequences (e.g. pertaining to human safety) in case of low reliability and/or reproducibility?	Risk Assessment	Reliability	Stakeholders	
52	Has a well-defined process to monitor if the AI system is meeting the intended goals been put in place?	Documentation	Objectives	Control	Management of system
53	Has been tested whether specific contexts or conditions need to be taken into account to ensure reproducibility?	Reliability	Robustness		

Continued on next page

Table B.2 – *Continued from previous page*

Ref	Code	Code Groups			
54	Have verification and validation methods and documentation (e.g. logging) been put in place to evaluate and ensure different aspects of the AI system's reliability and reproducibility?	Documentation	Reliability	Robustness	
55	Have processes for the testing and verification of the reliability and reproducibility of the AI system been clearly documented and operationalised?	Documentation	Management of system	Reliability	Robustness
56	Have tested failsafe fallback plans been defined to address AI system errors of whatever origin and put governance procedures in place to trigger them?	Safety	Management of system	Documentation	
57	Has a proper procedure been put in place for handling the cases where the AI system yields results with a low confidence score?	Management of system	Control	Reliability	
58	Does the AI system use (online) continual learning?	System description			
59	Have potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function been considered?	Risk Assessment			
60	Has the impact of the AI system been considered on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection?	Privacy	Safety	Risk Assessment	Stakeholders
61	Has a mechanisms been established that allows flagging issues related to privacy concerning the AI system?	Privacy	Stakeholders	Transparency	
62	Is the AI system being trained, or was it developed, by using or processing personal data (including special categories of personal data)?	Privacy	System description		
63	Is the system compliant with GDPR?	Privacy			
64	Has the right to withdraw consent, the right to object and the right to be forgotten been implemented into the development of the AI system?	Stakeholders	Management of system		
65	Have the privacy and data protection implications of data collected, generated or processed been considered over the course of the AI system's life cycle?	Management of system	Privacy		
66	Have the privacy and data protection implications been considered of the AI system's non-personal training-data or other processed non-personal data?	Privacy			
67	Has the AI system been aligned with relevant standards or widely adopted protocols for (daily) data management and governance?	Management of system			
68	Have measures been put in place that address the traceability of the AI system during its entire lifecycle?	Periodic Assessment	Documentation	Transparency	
69	Have measures been put in place to continuously assess the quality of the input data to the AI system?	Periodic Assessment			
70	Can it be traced back which data was used by the AI system to make a certain decision(s) or recommendation(s)?	Explainability	Transparency		
71	Can it be traced back which AI model or rules led to the decision(s) or recommendation(s) of the AI system?	Explainability			
72	Have measures been put in place to continuously assess the quality of the output(s) of the AI system?	Control	Periodic Assessment	Quality	
73	Have adequate logging practices been put in place to record the decision(s) or recommendation(s) of the AI system?	Documentation	Transparency		
74	Are the decision(s) of the AI system explained to the users?	Explainability	Stakeholders	Transparency	

Continued on next page

Table B.2 – *Continued from previous page*

Ref	Code	Code Groups			
75	In cases of interactive AI systems (e.g., chatbots, robo-lawyers), is it communicate to users that they are interacting with an AI system instead of a human?	Transparency	Stakeholders		
76	Have mechanisms been established to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system?	System description	Transparency	Stakeholders	Objectives
77	Have the benefits of the AI system been communicated to users?	Transparency	Stakeholders		
78	Have the technical limitations and potential risks of the AI system been communicated to users, such as its level of accuracy and/ or error rates?	Quality	Stakeholders	Transparency	
79	Have appropriate training material and disclaimers been provided to users on how to adequately use the AI system?	Stakeholders	Management of system		
80	Has a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system been established, both regarding the use of input data as well as for the algorithm design?	Management of system	Fairness		
81	Has diversity and representativeness of end-users and/or subjects in the data been considered?	Fairness	Quality		
82	Has been tested for specific target groups or problematic use cases?	Robustness	Fairness		
83	Have publicly available technical tools been researched and used, that are state-of- the-art, to improve understanding of the data, model and performance?	Control	Quality	Management of system	
84	Have processes been assessed and put in place to test and monitor for potential biases during the entire lifecycle of the AI system (e.g. biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)?	Periodic Assessment	Fairness		
85	Were diversity and representativeness of end-users and or subjects in the data considered?	Fairness	Stakeholders		
86	Have educational and awareness initiatives been put in place to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system?	Human involvement	Management of system	Fairness	
87	Is there a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?	Transparency	Quality	Fairness	
88	Are clear steps and ways of communicating established on how and to whom issues can be raised?	Accountability	Management of system	Transparency	
89	Have the subjects that could potentially be (in)directly affected by the AI system been identified, in addition to the (end-)users and/or subjects?	Stakeholders			
90	Is the definition of fairness commonly used and implemented in any phase of the process of setting up the AI system?	Fairness			
91	Were other definitions of fairness considered?	Fairness			
92	Have impacted communities been consulted about the correct definition of fairness, i.e. representatives of elderly persons or persons with disabilities?	Stakeholders	Fairness		
93	Is a quantitative analysis or metrics ensured to measure and test the applied definition of fairness?	Fairness			
94	Have mechanisms been established to ensure fairness in the AI system?	Fairness			

Continued on next page

Table B.2 – *Continued from previous page*

Ref	Code	Code Groups			
95	Has it been ensured that the AI system corresponds to the variety of preferences and abilities in society?	Stakeholders			
96	Has it been assessed whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion?	Stakeholders			
97	Has it been ensured that information about, and the AI system's user interface of, the AI system is accessible and usable also to users of assistive technologies (such as screen readers)?	Transparency	Stakeholders		
98	Were end-users or subjects in need for assistive technology consulted during the planning and development phase of the AI system?	Stakeholders			
99	Was ensured that Universal Design principles are taken into account during every step of the planning and development process, if applicable?	Stakeholders	Fairness		
100	Has the impact of the AI system on the potential end-users and/or subjects been taken into account?	Stakeholders			
101	Was assessed whether the team involved in building the AI system engaged with the possible target end-users and/or subjects?	Stakeholders	Human involvement	Management of system	
102	Was assessed whether there could be groups who might be disproportionately affected by the outcomes of the AI system?	Fairness	Stakeholders	Risk Assessment	
103	Was the risk of the possible unfairness of the system onto the end-user's or subject's communities assessed?	Fairness	Risk Assessment	Stakeholders	
104	Was a mechanism considered to include the participation of the widest range of possible stakeholders in the AI system's design and development?	Stakeholders	Management of system		
105	Are there potential negative impacts of the AI system on the environment?	Indirect environmental impact	Direct environmental impact		
106	Where possible, were mechanisms established to evaluate the environmental impact of the AI system's development, deployment and/or use (for example, the amount of energy used and carbon emissions)?	Management of system	Direct environmental impact	Indirect environmental impact	
107	Were measures defined to reduce the environmental impact of the AI system throughout its lifecycle?	Documentation	Direct environmental impact	Indirect environmental impact	Management of system
108	Does the AI system impact human work and work arrangements?	Human involvement	Stakeholders		
109	Have impacted workers and their representatives been informed and consulted in advance of the introduction of AI into the organisation?	Human involvement	Management of system		
110	Have measures been adopted to ensure that the impacts of the AI system on human work are well understood?	Management of system	Human involvement		
111	Has it been ensured that workers understand how the AI system operates, which capabilities it has and which it does not have?	Human involvement	Management of system		
112	Could the AI system create the risk of de-skilling of the workforce?	Human involvement	Risk Assessment		
113	Were measures taken to counteract de-skilling risks?	Risk Assessment	Management of system		
114	Does the system promote or require new (digital) skills?	Human involvement	Accountability		
115	Were training opportunities and materials for re- and up-skilling provided to workers?	Human involvement	Management of system		
116	Could the AI system have a negative impact on society at large or democracy?	Risk Assessment	Stakeholders		

Continued on next page

Table B.2 – *Continued from previous page*

Ref	Code	Code Groups				
117	Has the societal impact of the AI system's use been assessed beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large?	Stakeholders				
118	Were actions taken to minimize potential societal harm of the AI system?	Risk Assessment	Management of system	Stakeholders		
119	Were measures taken that ensure that the AI system does not negatively impact democracy?	Management of system				
120	Have mechanisms been established that facilitate the AI system's auditability?	Documentation	Management of system			
121	How is it ensured that the AI system can be audited by independent third parties?	Management of system	Documentation			
122	Was any kind of external guidance or third-party auditing processes foreseen to oversee ethical concerns and accountability measures?	Fairness	Stakeholders	Management of system	Accountability	
123	Does the involvement of third parties on ethical concerns and accountability go beyond the development phase?	Management of system	Accountability	Fairness	Stakeholders	
124	Was risk training organised and, if so, does this also inform about the potential legal framework applicable to the AI system?	Human involvement	Legality	Risk Assessment		
125	Has it been considered to establish an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas?	Management of system	Fairness	Accountability		
126	Has a process been established to discuss and continuously monitor and assess the AI system's adherence to guidelines?	Control	Periodic Assessment			
127	Is there a process of identification and documentation of conflicts between algorithm requirements or between different ethical principles and explanation of the 'trade-off' decisions made?	Quality	Explainability	Documentation	Fairness	
128	Was appropriate training provided to those involved in the monitoring process and does this also cover the legal framework applicable to the AI system?	Human involvement	Fairness	Control	Management of system	Legality
129	Is there a process for third parties (e.g. suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system?	Stakeholders	Fairness	Documentation		
130	Does the third-party reporting process foster revision of the risk management process?	Risk Assessment	Periodic Assessment	Management of system	Stakeholders	
131	For applications that can adversely affect individuals, have redress by design mechanisms been put in place?	Accountability	Management of system	Stakeholders	Fairness	

B.2 Attention to Algorithms

The Attention to Algorithms framework was published in 2021 (Algemene Rekenkamer, 2021). Its identified framework elements are shown in Table B.3 below.

Table B.3: *Attention to Algorithms Codes.*

Ref	Code	Code Groups		
1	What is the name of the algorithm or the system that it is part of?	System description		
2	What is the business process in which the algorithm plays a role?	System description	Management of system	Objectives

Continued on next page

Table B.3 – *Continued from previous page*

Ref	Code	Code Groups			
3	Are GDPR regulations being adhered to?	Legality	Privacy		
4	What is the product or service in which the algorithm plays a role?	Objectives	System description		
5	Does the algorithm advice or support actions or decisions made by human agents?	Human involvement	System description		
6	Does the algorithm function autonomously without human intervention?	Autonomy	Human involvement	Control	
7	What type of algorithm is used?	System description			
8	What type of applications and software are used?	System description			
9	Which data sources are used?	System description			
10	Is the algorithm a learning algorithm, which develops and improves over time with data and/or experience?	System description			
11	Has a purpose for the algorithm been determined?	Objectives	Management of system		
12	Is there a documented consideration of risks concerning the use of the algorithm?	Risk Assessment	Documentation		
13	Does the organisation possess appropriate professional skills to utilize the algorithm?	Management of system	Human involvement		
14	Does the organisation possess sufficient skilled personnel to utilize the algorithm?	Management of system	Human involvement		
15	Has the full algorithm life cycle been documented?	Documentation	Management of system		
16	Are roles, tasks, responsibilities and powers documented and practiced?	Human involvement	Documentation	Management of system	Accountability
17	Is there a documented agreed-upon approach regarding quality and performance goals for the algorithm?	Objectives	Quality	Documentation	Management of system
18	Have third-party agreements been made and documented?	Management of system	Documentation		
19	Does periodic monitoring take place?	Periodic Assessment			
20	Has the goal of the algorithm been usefully operationalized through the model and data used?	Objectives	Management of system	Quality	
21	Is there a common goal for the algorithm between owner, developer and user?	Stakeholders	Objectives		
22	Is the common goal clear and explainable for owner, developer and user?	Transparency	Explainability	Objectives	Stakeholders
23	Is the algorithm explainable?	Explainability			
24	Has a consideration taken place between model explainability and performance?	Management of system	Quality	Explainability	
25	Have design considerations been documented?	Documentation	Management of system		
26	Have implementation considerations been documented?	Documentation	Management of system		
27	Is there a document describing the design?	Documentation	System description		
28	Is there a document describing the implementation?	Documentation	Management of system		
29	Have hyperparameter choices been substantiated?	System description	Management of system		
30	Is the model publicly accessible to stakeholders?	Stakeholders	Transparency	Documentation	
31	Is (a description of) the used data publically accessibly for stakeholders?	Stakeholders	Documentation	System description	Transparency
32	Is the algorithm used in compliance with legislations and regulations applying to automated decision-making?	Legality			
33	Have stakeholders/end-users been involved in the development?	Stakeholders	Management of system	Human involvement	
34	What controls have been implemented in order to guarantee the accuracy and completeness of the data processing?	Quality	Control		

Continued on next page

Table B.3 – *Continued from previous page*

Ref	Code	Code Groups				
35	Is the model periodically updated in line with applicable laws and regulations?	Legality	Management of system			
36	Is there training/testing data choices quality control?	Management of system	Control	Quality		
37	Is bias prevented through choices made regarding the model?	Fairness	Management of system			
38	Does the data contain undesirable bias?	Fairness				
39	Have training, test and validation data been processed separately?	Quality	Reliability			
40	Is the data used representative for the application?	Quality	Explainability	Management of system		
41	Is there complete ownership of the data used for the model?	Control	Management of system			
42	Is dataminimization done proportionally?	Management of system				
43	Have performance metrics been documented?	Objectives	Documentation	Quality	Robustness	Reliability
44	Is there target leakage within the model?	Quality	Reliability			
45	Are performance metrics in place?	Objectives	Management of system	Control	Quality	
46	Is the model output monitored?	Human involvement	Quality	Control		
47	Does external coommunication take place concerning the model and its limitations?	Quality	Objectives	Documentation	Management of system	Stakeholders
48	Are maintenance and management of the algorithm in place?	Management of system	Control			
49	Is the use of personal data kept track off in a register?	Documentation	Privacy			
50	Is data protection by design in place?	Privacy				
51	Has a data protection impact assesment been caried out?	Privacy	Risk Assessment	Management of system		
52	Has the automatic decision-making been authoritized?	Legality	Human involvement	Autonomy		
53	Are those involved offered an option to not be subjected to automated decision-making?	Stakeholders	Human involvement	Autonomy	Control	
54	Has dataminimization been applied?	Privacy				
55	Is data processed on a legal basis?	Legality				
56	Is the processing of (special) personal data through the algorithm compatible with the original goal?	Objectives	Privacy	Management of system		
57	Has the responsibility to process personal data been appointed?	Human involvement	Privacy			
58	Is there discrimination due to the used data or model?	Fairness				
59	Has the degree of profiling and its legal basis been tested?	Legality	Fairness			
60	Have those whose data has been processed/used been informed proactively/upon request?	Stakeholders	Privacy	Transparency		
61	Is the logic of the used algorithm and data clear to those involved?	Explainability	Stakeholders			
62	Are the consequences of the application of the used algorithm clear to those involved?	Objectives	Transparency	Stakeholders		
63	Is there a public privacy policy which covers used data and algorithms?	Privacy	Transparency	Stakeholders	Management of system	
64	Is there an accessible audit trail?	Documentation	Management of system			
65	Are the access rights to the environment in which the algorithm operates checked for being up-to-date?	Security	Periodic Assessment	Human involvement		
66	Are access rights changed as soon as an employee leaves office or changes position?	Security				
67	Are access rights handed out by the authorized personel?	Security	Human involvement	Management of system		
68	Is separation of duties prevented at the access of users to the algorithm?	Security	Human involvement			
69	Are generic management accounts used?	Human involvement	Security			

Continued on next page

Table B.3 – *Continued from previous page*

Ref	Code	Code Groups				
70	Does the number of management accounts match the number of managers?	Security				
71	Are the access rights named and ordered systematically?	Security	Documentation			
72	Are naming conventions used such that users and managers can be identified?	Documentation	Management of system	Security		
73	Are tasks executed under the appropriate accounts?	Security	Human involvement			
74	Do users have access to underlying elements of the algorithm?	Security				
75	Is there separation of duties between requesting, authorizing and processing changes in accounts and access rights?	Security	Human involvement			
76	Is password management interactive?	Security				
77	Are passwords of the appropriate strength?	Security				
78	Are changes to code executed in a controlled manner?	Management of system	Control			
79	Is the algorithm protected from unauthorized access, changes, damage or data loss?	Security	Control	Safety		
80	Is the algorithm backed-up?	Control	Management of system			
81	Can the back-up be recovered?	Control	Management of system			
82	Did security by design take place?	Management of system	Security			

B.3 Access Depth Framework

The Access Depth framework was published in 2021 (Akula & Garibay, 2021). Its identified framework elements are shown in Table B.4 below.

Table B.4: *Access Depth Framework Codes.*

Ref	Code	Code Groups				
1	Is the use of the algorithm in compliance with regulations and corporate policy?	Legality	Management of system			
2	How does the algorithm respond to fabricated input?	Robustness				
3	How well does the model perform?	Reliability	Robustness	Quality		
4	How well does the model perform, assessing a range of inputs and associated outputs?	Quality	Robustness	Reliability	Fairness	
5	How consistent is the model during perturbation testing?	Reliability	Robustness			
6	What is the network size of the algorithm?	System description				
7	Has the algorithm been stress tested?	Reliability	Robustness	Fairness	Documentation	Quality
8	What trade-offs have been made between bias, privacy and performance?	Privacy	Quality	Fairness	Risk Assessment	Management of system
9	What risk mitigation practices are put in place?	Risk Assessment	Documentation	Management of system		
10	Does governance ensure quality and integrity of the data utilized?	Privacy	Quality	Management of system	Accountability	
11	What is the relevance of the data in the area where the algorithm will be deployed?	Control	Quality			
12	What data access procedures are in place?	Human involvement	Security	Management of system		
13	Is data handled such that privacy is respected?	Privacy				
14	Is there bias in the system?	Fairness				
15	Are procedures within the system transparent?	Transparency	System description			

Continued on next page

Table B.4 – *Continued from previous page*

Ref	Code	Code Groups			
16	Have system capabilities and purposes been publicly disclosed?	Transparency	System description	Stakeholders	Documentation
17	Are system outputs explained to stakeholders?	Explainability	Stakeholders		
18	Is there a backup strategy?	Safety	System description	Management of system	

B.4 Conformity Assessment Procedure for Artificial Intelligence

The Conformity Assessment Procedure for AI was published in 2022 (Floridi et al., 2022). Its identified framework elements are shown in Table B.5 below.

Table B.5: *Conformity Assessment Procedure for Artificial Intelligence Codes.*

Ref	Code	Code Groups			
1	Has the organisation has defined the set of values that guides the development of AI systems?	Documentation	Management of system		
2	Have guiding values have been published/ communicated externally?	Transparency	Documentation	Stakeholders	Management of system
3	Have guiding values have been communicated to internal AI project stakeholders?	Transparency	Human involvement	Stakeholders	Management of system
4	Has a governance framework for AI projects been defined?	Management of system	Documentation		
5	Has the responsibility for ensuring and demonstrating that AI systems adhere to defined organisational values been assigned?	Accountability	Management of system		
6	Have the objectives of the AI application been defined and documented?	Documentation	Objectives		
7	Has the AI application been assessed against the ethical values?	Fairness			
8	Have performance criteria for the AI application been defined?	Documentation	Objectives	Quality	
9	Has the overall environmental impact for this AI application been assessed?	Indirect environmental impact	Management of system		
10	Has the data used to develop the AI application been documented?	Documentation	System description		
11	Has data used in the development been checked for representativeness, relevance, accuracy, traceability (e.g., external data) and completeness?	Quality	Fairness	Transparency	
12	Have the risks identified in the data impact assessment been considered and addressed?	Risk Assessment	Management of system		
13	Is the system legally compliant with respect to data protection?	Privacy	Security		
14	Has the source of the model been documented?	Documentation	System description		
15	Has the selection of the model been assessed with regard to fairness, explainability and robustness?	Fairness	Explainability	Management of system	Robustness
16	Have the risks identified in the model been considered and addressed?	Risk Assessment	Management of system		
17	Has the strategy for validating the model been defined?	Quality	Documentation		
18	Did the organisation document the AI performance in the training environment?	Quality	Documentation		
19	Has the setting of hyperparameters been documented?	Documentation	System description		
20	Does the model fulfil the established performance criteria levels?	Objectives	Quality		

Continued on next page

Table B.5 – *Continued from previous page*

Ref	Code	Code Groups				
21	Has the strategy for testing the model been defined?	Documentation	Control			
22	Has the organisation documented the AI performance in the testing environment?	Quality	Documentation	Control		
23	Has the model been tested for performance on extreme values and protected attributes?	Robustness	Control			
24	Have patterns of failure been identified?	Control	Reliability	Robustness		
25	Have key failure modes been addressed?	Management of system	Control			
26	Has the deployment strategy been documented?	Documentation	Management of system			
27	Has the serving strategy of the system to end-users been documented?	Documentation	Management of system			
28	Have the risks associated with the given serving and deployment strategies been identified?	Risk Assessment				
29	Have the risks associated with the given serving and deployment strategies been addressed?	Risk Assessment	Management of system			
30	Does the model fulfil the established performance criteria levels in the production environment?	Periodic Assessment	Quality	Objectives		
31	Have risks associated with changing data quality and potential data drift been identified?	Risk Assessment	Periodic Assessment	Quality		
32	Have the risks associated with model decay been identified?	Periodic Assessment	Risk Assessment	Reliability		
33	Has the strategy for monitoring and addressing risks associated with data quality and drift; and model decay been defined?	Management of system	Documentation	Reliability	Risk Assessment	Periodic Assessment
34	Have periodic reviews of the AI applications with regard to the ethical values been set?	Periodic Assessment	Fairness			
35	Does the organisation have a strategy for how to update the AI application continuously?	Periodic Assessment	Management of system			
36	Has a complaints process been established for users of the AI system to raise concerns or suggest improvements?	Transparency	Stakeholders			
37	Has a problem-to-resolution process been defined?	Documentation	Management of system			
38	Have the risks of decommissioning the AI system been assessed?	Risk Assessment	Management of system			
39	Is there a strategy for addressing risks associated with decommissioning the AI system?	Management of system	Control			

B.5 Cross-Industry Standard Process for Data Mining Auditing Framework

The CRISP-DM Framework for the Machine Learning Audit was published in 2018 (Clark, 2018). Its identified framework elements are shown in Table B.6 below.

Table B.6: *Cross-Industry Standard Process for Data Mining Auditing Framework Codes.*

Ref	Code	Code Groups	
1	What is the business use case?	System description	Objectives
2	What attributes of the use case should be included in the AI model?	Objectives	System description
3	Where is the data stored?	System description	
4	What are the input variables and do they conflict or introduce bias?	Fairness	System description

Continued on next page

Table B.6 – *Continued from previous page*

Ref	Code	Code Groups		
5	How do variables correlate and vary in response to one another?	Quality	Robustness	
6	What are the nature and idiosyncrasy of the data?	System description		
7	Is the degree of interpretability required for the given use case met?	Explainability		
8	Does the AI system operate in compliance with GDPR?	Privacy	Legality	
9	Has data been split into training and test sets?	Control	Quality	
10	Can model accuracy be validated?	Quality	Control	
11	Does the determined accuracy meet the goals of the model?	Objectives	Quality	
12	Does the model violate the principles of the business?	Objectives	Management of system	
13	Does the model produce any unintended effects?	Risk Assessment	Quality	Stakeholders
14	Is any technical debt integrated in the model?	System description	Quality	

B.6 Environmental, Social and Governance Protocol for Artificial Intelligence

The ESG Protocol for AI was published in 2022 (Saetra, 2022). Its identified framework elements are shown in Table B.7 below.

Table B.7: *Environmental, Social and Governance Protocol for Artificial Intelligence Codes.*

Ref	Code	Code Groups		
1	How does the AI consume energy and generate emissions?	Direct environmental impact	Indirect environmental impact	
2	Does the application of the AI result in positive results for the environment, either directly or indirectly?	Management of system	Indirect environmental impact	
3	What is the material basis of the computing infrastructure?	System description		
4	What are the materials used in and environmental impact of the machinery used in the AI system?	Indirect environmental impact		
5	What are the materials used in and environmental impact of machinery used in the supply chain i.e. regarding data sources?	Indirect environmental impact		
6	Where in the entity is AI used?	System description		
7	What sort of data does the entity control?	System description		
8	What sort of AI and data related capabilities does the entity have?	Human involvement	Accountability	
9	How is AI and data used in the entity?	System description		
10	Who is operatively in charge, and who holds responsibility?	Accountability	Management of system	
11	What are the relevant strategies, plans, and governance documents?	Management of system	Documentation	
12	Is there an ethics policy, and/or does the entity subscribe to any ethics/sustainability standard?	Fairness	Documentation	Management of system
13	What are the main identified risks and opportunities?	Risk Assessment	Objectives	Management of system
14	Has a risk analysis and matrix been constructed?	Risk Assessment		

Continued on next page

Table B.7 – *Continued from previous page*

Ref	Code	Code Groups				
15	Has a materiality analysis and matrix been constructed?	Risk Assessment	Stakeholders			
16	Has organizational readiness been assessed?	Management of system				
17	Have AI risks and opportunities been identified?	Risk Assessment				
18	Who is responsible for implementation and overseen implementation?	Accountability				
19	How many and what type of computers are part of the system?	System description				
20	What are the power demands of the AI system?	Indirect environmental impact				
21	How are datasets protected?	Security	Safety			
22	How is dataset privacy guaranteed?	Privacy				
23	Are workers exposed to environmental harms?	Risk Assessment	Human involvement			
24	Are workers exposed to harmful data?	Risk Assessment	Human involvement			
25	What is the energy cost of training the AI?	Indirect environmental impact				
26	Are there negative impacts related to source data and privacy?	Privacy				
27	Are there positive impacts related to data and priacy?	Privacy				
28	Is there documentation of origin and legality of data used?	Legality	Documentation			
29	What are the AI readiness evaluation results?	Management of system				
30	How much electricity was bought for own computers AI related?	Indirect environmental impact				
31	What type of electricity was bought?	Indirect environmental impact				
32	What insights does the LCA of system equipment provide?	Indirect environmental impact	Direct environmental impact			
33	What are the workers' rights implications of systems used?	Stakeholders	Human involvement			
34	What is the positive environmental impact of using the AI?	Indirect environmental impact	Direct environmental impact			
35	What is the positive social impact of using the AI?	Stakeholders				
36	What is the negative social impact of using the AI?	Stakeholders				
37	What are the positive sustainability related economic impacts of the AI?	Indirect environmental impact	Direct environmental impact			
38	What are the negative sustainability related economic impacts of using the AI?	Indirect environmental impact	Direct environmental impact			
39	Are ESG risks assessed?	Risk Assessment	Management of system	Indirect environmental impact	Direct environmental impact	Stakeholders

B.7 Generalized Audit Framework for Artificial Intelligence

The Generalized Audit Framework for AI was published in 2022 (Markert et al., 2022). Its identified framework elements are shown in Table B.8 below.

Table B.8: *Generalized Audit Framework for Artificial Intelligence Codes.*

Ref	Code	Code Groups	
1	Have claims and the functionality of the AI system been defined?	System description	Documentation
2	Have potential threats and hazard that might occur during the AI life cycle been examined?	Security	Risk Assessment
3	Have system requirements been defined that sufficiently cover the examined threats and hazards?	Risk Assessment	Management of Control system
4	Can the auditee provides relevant evidence to the auditors?	Control	Documentation
5	Does the provided evidence reflect the information provided during scoping phase?	Control	Management of Documentation System
6	Is the provided evidence sufficient to support the defined requirements?	Documentation	Management of system
7	Have thresholds and boundary values for testing been defined?	Quality	Objectives
8	Are results accompanied by an estimation of the residual risk?	Risk Assessment	Control
9	Does the available documentation meet control criteria?	Documentation	Control

B.8 Institute of Internal Auditors Artificial Intelligence Auditing Framework.

The IIA AI Auditing Framework was published in 2018 (The Institute of Internal Auditors, 2018). Its identified framework elements are shown in Table B.9 below.

Table B.9: *Institute for Internal Auditors Artificial Intelligence Auditing Framework Codes.*

Ref	Code	Code Groups			
1	Has an AI strategy been documented?	Documentation	Management of system		
2	Are effective cyber threat defenses and responses in place?	Security			
3	Is a sound process for determining staff and budget needs to support AI in place?	Human involvement	Management of system		
4	What are the existing assessments of AI threats and opportunities?	Risk Assessment	Objectives		
5	Do business models and organizational structure reflect the organization's AI strategy?	Management of system			
6	Do organizational policies and procedures clearly identify AI roles and responsibilities related to AI strategy, governance, data architecture, data quality, ethical imperatives, and measuring performance?	Accountability	Management of system	Quality	Fairness
7	Do those with AI responsibilities have the necessary competencies to be successful?	Human involvement	Accountability		
8	Do AI policies and procedures sufficiently address AI risks?	Risk Assessment	Management of system		
9	Do audit trails provide sufficient information to understand what decisions were made, and why?	Documentation	Management of system		
10	Are access policies in place and access controls effective?	Control	Security		
11	Is the organization preparing for compliance with new technology regulations, such as the EU's General Data Protection Regulation (GDPR)?	Management of system			

Continued on next page

Table B.9 – *Continued from previous page*

Ref	Code	Code Groups		
12	Do organization's disaster recovery protocols include AI failures, including the breakdown of controls that maintain the rules set forth by AI governance?	Control	Management of system	
13	Is the system infrastructure capable of handling structured and unstructured data?	System description		
14	What is the quality, completeness, and consistency of use for the enterprise-wide data taxonomy?	Management of system		
15	Has the organization implemented methodologies to validate AI outcomes with actual, real-world outcomes?	Control	Quality	
16	Are procedures in place to continuously measure, monitor, escalate, and rectify inconsistencies between AI outcomes and real-world?	Periodic Assessment	Quality	Control
17	Are policies and procedures in place to continuously measure, monitor, escalate, and rectify data accuracy and integrity issues?	Periodic Assessment	Quality	Control
18	Is the organization consistently following and monitoring a formalized data reconciliation framework, which includes a rationale for differing methodologies and results should they exist?	Management of system	Control	
19	Are policies and procedures in place to limit data input bias?	Fairness		
20	Have those responsible for decision-making received and considered explanations on material exceptions related to data quality?	Human involvement	Explainability	Quality
21	Is there variance between the intended results of the AI activities (strategic objectives) and actual results, and was bias the cause?	Objectives	Quality	Fairness
22	Does the meaning derived from AI outputs follow from the AI outputs?	Quality	Objectives	
23	Has black box data been identified and is it understood?	Explainability	Management of system	
24	Have AI vulnerabilities been stress-tested?	Security	Control	

B.9 Artificial Intelligence Risk Management Framework

The Artificial Intelligence Risk Management Framework was published in 2023 (National Institute of Standards and Technology, 2023). Its identified framework elements are shown in Table B.10 below.

Table B.10: *Artificial Intelligence Risk Management Framework Codes.*

Ref	Code	Code Groups		
1	Are legal and regulatory requirements involving AI understood, managed, and documented?	Legality	Documentation	Management of system
2	Are the characteristics of trustworthy AI integrated into organizational policies, processes, procedures, and practices?	Management of system		
3	Are processes, procedures, and practices in place to determine the needed level of risk management activities based on the organization's risk tolerance?	Management of system	Risk Assessment	Objectives
4	Are the risk management process and its outcomes established through transparent policies, procedures, and other controls based on organizational risk priorities?	Transparency	Management of system	Risk Assessment

Continued on next page

Table B.10 – *Continued from previous page*

Ref	Code	Code Groups				
5	Are ongoing monitoring and periodic review of the risk management process and its outcomes planned and organizational roles and responsibilities clearly defined, including determining the frequency of periodic review?	Accountability	Periodic Assessment	Management of system	Risk Assessment	
6	Are mechanisms in place to inventory AI systems and are they resourced according to organizational risk priorities?	System description	Management of system	Documentation		
7	Are processes and procedures in place for decommissioning and phasing out AI systems safely and in a manner that does not increase risks or decrease the organization's trustworthiness?	Safety	Management of system			
8	Are roles and responsibilities and lines of communication related to mapping, measuring, and managing AI risks documented and are they clear to individuals and teams throughout the organization?	Documentation	Management of system	Accountability	Risk Assessment	Human involvement
9	Have the organization's personnel and partners received AI risk management training to enable them to perform their duties and responsibilities consistent with related policies, procedures, and agreements?	Human involvement	Management of system			
10	Does executive leadership of the organization take responsibility for decisions about risks associated with AI system development and deployment?	Accountability	Management of system	Risk Assessment		
11	Is decision-making related to mapping, measuring, and managing AI risks throughout the lifecycle informed by a diverse team?	Risk Assessment	Human involvement			
12	Are policies and procedures are in place to define and differentiate roles and responsibilities for human-AI configurations and oversight of AI systems?	Accountability	Management of system	Documentation	Human involvement	Control
13	Are organizational policies and practices in place to foster a critical thinking and safety-first mindset in the design, development, deployment, and uses of AI systems to minimize potential negative impacts?	Safety	Management of system	Human involvement		
14	Do organizational teams document the risks and potential impacts of the AI technology they design, develop, deploy, evaluate, and use, and do they communicate about the impacts more broadly?	Documentation	Human involvement	Risk Assessment	Stakeholders	
15	Are organizational practices in place to enable AI testing, identification of incidents, and information sharing?	Management of system				
16	Are organizational policies and practices in place to collect, consider, prioritize, and integrate feedback from those external to the team that developed or deployed the AI system regarding the potential individual and societal impacts related to AI risks?	Stakeholders	Management of system	Risk Assessment		
17	Have mechanisms been established to enable the team that developed or deployed AI systems to regularly incorporate adjudicated feedback from relevant AI actors into system design and implementation?	Periodic Assessment	Stakeholders	Management of system		
18	Are policies and procedures in place that address AI risks associated with third-party entities, including risks of infringement of a third-party's intellectual property or other rights?	Risk Assessment	Legality			
19	Are contingency processes in place to handle failures or incidents in third-party data or AI systems deemed to be high-risk?	Risk Assessment	Safety			

Continued on next page

Table B.10 – *Continued from previous page*

Ref	Code	Code Groups			
20	Are intended purposes, potentially beneficial uses, context-specific laws, norms and expectations, and prospective settings in which the AI system will be deployed understood and documented?	Documentation	System description	Objectives	Legality
21	Do interdisciplinary AI actors, competencies, skills, and capacities for establishing context reflect demographic diversity and broad domain and user experience expertise, and is their participation documented?	Documentation	Stakeholders	Human involvement	
22	Are the organization's mission and relevant goals for AI technology understood and documented?	Objectives	Documentation		
23	Are organizational risk tolerances determined and documented?	Documentation	Objectives		
24	Are system requirements (e.g., "the system shall respect the privacy of its users") elicited from and understood by relevant AI actors?	Stakeholders	Human involvement	System description	
25	Are the specific tasks and methods used to implement the tasks that the AI system will support defined?	System description	Documentation		
26	Is information about the AI system's knowledge limits and how system output may be utilized and overseen by humans documented?	Human involvement	System description	Documentation	
27	Are scientific integrity and TEVV (test, evaluate, verify, validate) considerations identified and documented?	Quality	Control	Documentation	
28	Are potential benefits of intended AI system functionality and performance examined and documented?	Documentation	System description		
29	Are potential costs, including non-monetary costs, which result from expected or realized AI errors or system functionality and trustworthiness – as connected to organizational risk tolerance – examined and documented?	Risk Assessment	Documentation		
30	Is the targeted application scope specified and documented based on the system's capability, established context, and AI system categorization?	System description	Documentation		
31	Are processes for operator and practitioner proficiency with AI system performance and trustworthiness – and relevant technical standards and certifications – defined, assessed, and documented?	Documentation	Human involvement	Control	Management of system
32	Are processes for human oversight defined, assessed, and documented in accordance with organizational policies?	Human involvement	Documentation	Management of system	
33	Are approaches for mapping AI technology and legal risks of its components – including the use of third-party data or software – in place, followed, and documented, as are risks of infringement of a third party's intellectual property or other rights?	Legality	Documentation	Risk Assessment	
34	Are internal risk controls for components of the AI system, including third-party AI technologies, identified and documented?	Documentation	Control		
35	Are likelihood and magnitude of each identified impact (both potentially beneficial and harmful) based on expected use, past uses of AI systems in similar contexts, public incident reports, feedback from those external to the team that developed or deployed the AI system, or other data identified and documented?	Documentation	Risk Assessment		
36	Are practices and personnel for supporting regular engagement with relevant AI actors and integrating feedback about positive, negative, and unanticipated impacts in place and documented?	Stakeholders	Documentation	Periodic Assessment	

Continued on next page

Table B.10 – *Continued from previous page*

Ref	Code	Code Groups				
37	Are approaches and metrics for measurement of AI risks selected for implementation starting with the most significant AI risks?	Risk Assessment	Management of system			
38	Are the appropriateness of AI metrics and effectiveness of existing controls regularly assessed and updated?	Periodic Assessment	Control			
39	Are internal experts who did not serve as front-line developers for the system and/or independent assessors involved in regular assessments and updates?	Control	Human involvement	Periodic Assessment		
40	Are test sets, metrics, and details about the tools used during TEVV (testing, evaluation, verification, validation) documented?	Documentation	System description			
41	Do evaluations involving human subjects meet applicable requirements (including human subject protection) and are they representative of the relevant population?	Stakeholders	Human involvement			
42	Are AI system performance or assurance criteria measured qualitatively or quantitatively and demonstrated for conditions similar to deployment setting(s)?	Objectives	Control			
43	Are the functionality and behavior of the AI system and its components monitored when in production?	Objectives	Periodic Assessment			
44	Is the AI system to be deployed demonstrated to be valid and reliable?	Reliability				
45	Is the AI system evaluated regularly for safety risks?	Periodic Assessment	Safety			
46	Are AI system security and resilience evaluated and documented?	Security	Documentation			
47	Are risks associated with transparency and accountability examined and documented?	Risk Assessment	Transparency	Accountability	Documentation	
48	Is the AI model explained, validated, and documented, and is the AI system output interpreted within its context to inform responsible use and governance?	Objectives	Explainability	Documentation		
49	Is the privacy risk of the AI system examined and documented?	Privacy	Documentation			
50	Are fairness and bias evaluated and results documented?	Documentation	Fairness			
51	Are environmental impact and sustainability of AI model training and management activities assessed and documented?	Documentation	Direct environmental impact	Indirect environmental impact		
52	Are effectiveness of the employed TEVV (test, evaluate, verify, validate) metrics and processes evaluated and documented?	Quality	Documentation			
53	Are approaches, personnel, and documentation in place to regularly identify and track existing, unanticipated, and emergent AI risks based on factors such as intended and actual performance in deployed contexts?	Periodic Assessment	Risk Assessment	Quality	Documentation	Human involvement
54	Are risk tracking approaches considered for settings where AI risks are difficult to assess using currently available measurement techniques or where metrics are not yet available?	Risk Assessment	Control			
55	Are feedback processes for end users and impacted communities to report problems and appeal system outcomes established and integrated into AI system evaluation metrics?	Stakeholders	Management of system			
56	Are measurement approaches for identifying AI risks connected to deployment context(s) and informed through consultation with domain experts and other end users?	Human involvement	Risk Assessment	Stakeholders		

Continued on next page

Table B.10 – *Continued from previous page*

Ref	Code	Code Groups			
57	Are measurement results regarding AI system trustworthiness in deployment context(s) and across the AI lifecycle informed by input from domain experts and relevant AI actors to validate whether the system is performing consistently as intended?	Human involvement	Periodic Assessment	Reliability	
58	Are measurable performance improvements or decreases based on consultations with relevant AI actors, including affected communities, and field data about context-relevant risks and trustworthiness characteristics identified and documented?	Documentation	Stakeholders	Quality	Risk Assessment
59	Has a determination been made as to whether the AI system achieves its intended purposes and stated objectives and whether its development or deployment should proceed?	Objectives	Management of system	Quality	
60	Is the treatment of documented AI risks prioritized based on impact, likelihood, and available resources or methods?	Risk Assessment			
61	Are responses to the AI risks deemed high priority developed, planned, and documented?	Management of system	Risk Assessment	Documentation	Control
62	Are negative residual risks (defined as the sum of all unmitigated risks) to both downstream acquirers of AI systems and end users documented?	Risk Assessment	Stakeholders	Documentation	
63	Have resources required to manage AI risks been taken into account – along with viable non-AI alternative systems, approaches, or methods – to reduce the magnitude or likelihood of potential impacts?	Control	Risk Assessment		
64	Are mechanisms in place and applied to sustain the value of deployed AI systems?	Objectives	Periodic Assessment		
65	Procedures are followed to respond to and recover from a previously unknown risk when it is identified.	Management of system	Control	Risk Assessment	
66	Are mechanisms in place and applied, and responsibilities assigned and understood, to supersede, disengage, or deactivate AI systems that demonstrate performance or outcomes inconsistent with intended use?	Quality	Management of system	Control	Accountability
67	Are AI risks and benefits from third-party resources regularly monitored, and risk controls applied and documented?	Control	Risk Assessment	Documentation	Stakeholders
68	Are pre-trained models which are used for development monitored as part of AI system regular monitoring and maintenance?	Control	Management of system		
69	Have post-deployment AI system monitoring plans been implemented, including mechanisms for capturing and evaluating input from users and other relevant AI actors, appeal and override, decommissioning, incident response, recovery, and change management?	Periodic Assessment	Management of system	Control	Stakeholders
70	Are measurable activities for continual improvements integrated into AI system updates and do they include regular engagement with interested parties, including relevant AI actors?	Periodic Assessment	Management of system	Stakeholders	Quality
71	Are incidents and errors communicated to relevant AI actors, including affected communities?	Stakeholders	Management of system		

B.10 Guiding Principles for Trustworthy Artificial Intelligence Investigations

The Guiding Principles for Trustworthy Artificial Intelligence Investigations was published in 2021 (de Boer & van Geijn, 2021). Its identified framework elements are shown in Table B.11 below.

Table B.11: *Guiding Principles for Trustworthy Artificial Intelligence Investigations Codes.*

Ref	Code	Code Groups		
1	Has the organization defined and documented common language for the development, implementation and operation of its AI systems?	Documentation	Management of system	
2	Does the use of the algorithm negatively impact existing governance mechanisms related to data processing?	Management of system		
3	Is use of AI in line with organization's risk acceptance?	Risk Assessment	Management of system	
4	Does use of the AI potentially impact human rights?	Legality	Safety	Stakeholders
5	Is the broader societal impact of the AI system assessed?	Stakeholders	Management of system	
6	How do the duties of the AI relate to the duties of humans in the system?	Human involvement	Autonomy	
7	What human controls are in place?	Human involvement	Control	
8	Have the risks of harm and damage to stakeholders been assessed?	Risk Assessment	Stakeholders	
9	Has a goal been set for the algorithm accuracy?	Objectives	Quality	
10	Are steps of the development processed documented?	Documentation	Management of system	
11	To what extent can the AI output be explained?	Explainability		
12	To what extent does the AI affect decision-making processes?	Autonomy		
13	Why was the system deployed?	Objectives		
14	What is the AI's added value?	Management of system		
15	How is the AI explained to stakeholders?	Stakeholders	Explainability	
16	How are biases in the AI prevented?	Fairness		
17	Has fairness been defined and applied in the development of the system?	Fairness	Management of system	
18	Is the AI accessible/usable to a wide range of individuals?	Human involvement	Control	
19	Have relevant stakeholders been involved during AI development and implementation?	Stakeholders	Management of system	
20	Have trade-offs between values as a result of the algorithm been documented?	Reliability	Transparency	Explainability
		Fairness	Quality	Documentation
21	Is the AI and underlying technology legally permitted?	Legality		
22	Has a legal assessment been performed on applicable legislation?	Legality		
23	Have all direct and indirect types of personal data of the algorithmic system been identified	Privacy		
24	Is there a lawful basis for all the purposes of the algorithmic system and the use of personal data?	Privacy	Legality	
25	Can stakeholders object to automated processing of their data?	Stakeholders	Control	
26	Is the personal data collected proportional, relevant and necessary for the purpose of processing?	Privacy	Objectives	
27	Have alternatives been considered using fewer personal data to achieve the same objectives of processing?	Privacy	Objectives	

Continued on next page

Table B.11 – *Continued from previous page*

Ref	Code	Code Groups		
28	Does the AI lead to decisions with major effects on data subjects?	Stakeholders		
29	Is the right to not be subjected to solely automated decision making been accommodated?	Control	Autonomy	Stakeholders
30	Is the AI transparent regarding the basis for decisions/conclusions?	Transparency		
31	What is the decision the algorithmic system is designed to support?	System description	Objectives	
32	Was the algorithmic system designed specifically to support this decision	Management of system	Objectives	
33	Is there evidence of the rationale and the scoping of the algorithmic system concept?	Documentation	Management of system	
34	Have requirements for the HR involved with the AI been defined?	Documentation	System description	Human involvement
35	Is there an exit/change strategy within the development plan for the AI that considers the dependency on external parties?	Stakeholders	Management of system	
36	Are policies in place to address AI security risks, attacks and threats?	Security	Documentation	Management of system
37	Is there a Data Protection Officer and Data Protection Impact Assessment?	Risk Assessment	Human involvement	Privacy
38	Has the organization facilitated the algorithmic system to be auditable?	Documentation	Management of system	
39	Does the organization provide training and education to help develop accountability practices?	Human involvement	Accountability	
40	Has the organization assessed whether unlawful bias can occur in the algorithmic system (input and output)?	Legality	Fairness	
41	Is the AI designed with appropriate user management funcitionality?	Management of system	Human involvement	Security
42	Does the AI process any special categories of personal data?	Privacy		
43	Can the personal data processed be used to profile or discriminate data subjects?	Stakeholders	Privacy	
44	Is the data collected for development, training and implementation of the AI limited to the scope of the solution?	Management of system	Objectives	
45	Is there a technical guide that demonstrates the logical flow of the algorithmic system?	Documentation	System description	
46	Is the data in the algorithmic system of good quality?	Quality		
47	Is the data used for the development, training and implementation of the AI representative of the task?	Objectives	Quality	
48	Is the data the AI uses derived from other models?	Transparency		
49	What processes does the algorithmic system use to handle input data?	System description		
50	Is the lineage of the data used for the development, training and implementation of the algorithmic system documented so that sources, changes and alterations can be traced?	Documentation	System description	
51	Has a (data) access control policy been established, documented and reviewed based on AI security requirements?	Documentation	Security	Human involvement
52	Has a process of user management been implemented on data and the algorithmic system?	Human involvement	Management of system	Security
53	Are data collection, storage, processing and use oversight mechanisms documented?	Documentation	Security	Management of system
54	Is the system aligned with relevant standards regarding data management and governance?	Quality	Management of system	
55	What protocols, processes and procedures did the organization follow to manage and ensure proper data governance?	Management of system	Documentation	
56	Has data been de-identified?	Privacy		
57	Is the data correctly annotated for the intended purpose of the AI?	Objectives	Quality	

Continued on next page

Table B.11 – *Continued from previous page*

Ref	Code	Code Groups				
58	Has a process been implemented to ensure the quality and integrity of the data?	Quality	Management of system			
59	Have all used data sources been verified?	Control	Quality			
60	When data is sourced from a third party, has it been ensured that this third party has strong security practices?	Security	Stakeholders			
61	Can malicious input data be detected?	Security				
62	How does the algorithmic system interact with decisions by human (end) users?	Human involvement	System description			
63	Is there a self-learning/autonomous AI? Are specific mechanisms of oversight and control in place?	System description	Control			
64	Can AI decisions explained to all stakeholders?	Explainability	Accountability			
65	How is model performance against prohibited discrimination grounds measured, monitored and mitigated?	Fairness	Quality	Documentation	Risk Assessment	Control
66	Are all types of personal data identified and is there a lawful basis for the AI purposes concerning this data?	Privacy	Objectives	Legality		
67	Has a Data Protection Impact Assessment (DPIA) been carried out to assess the data protection of the algorithmic system?	Stakeholders	Privacy	Security		
68	Is there a profound understanding of the algorithmic system?	Human involvement	Quality			
69	To what degree could the algorithmic system be misused for unintended purposes?	Objectives	Risk Assessment			
70	Has the organization ensured that the algorithmic system has a sufficient fallback plan if it encounters adversarial attacks or other unexpected situations?	Safety	Stakeholders	Security		
71	Are mechanisms established to measure the environmental impact of the algorithmic system's development, deployment and use	Direct environmental impact	Management of system	Indirect environmental impact		
72	Is a strategy in place to monitor and test if the algorithmic system meets the goals, purposes and intended applications?	Objectives	Quality	Stakeholders	Documentation	
73	Does the model respond logically to basic changes being made to the algorithmic system inputs?	Robustness				
74	How accurate is the detail of the algorithmic system?	Reliability	Quality			
75	Are the details of algorithmic system assumptions recorded and justified?	Documentation	Quality	Reliability		
76	Are the AI and training data within scope of the risks assessments?	Risk Assessment	Objectives			
77	Are controls in place to safeguard interaction between the AI system and other entities that could alter input or output data?	Stakeholders	Quality	Control		
78	Are any measures or systems implemented to ensure the integrity and resilience of the algorithmic system against potential attacks?	Security				
79	Are the AI development and training environment protected through access control?	Security	Management of system	Human involvement		
80	Have quality of models and their third-party providers been verified?	Quality	Stakeholders	Control		
81	Are processes in place for third-parties to report potential vulnerabilities, risks or biases?	Stakeholders	Transparency	Risk Assessment	Fairness	Security
82	Did the organization estimate the likely impact of a failure of the algorithmic system?	Management of system	Risk Assessment	Quality		
83	On which aspects does the organization monitor the AI systems?	Control	System description			
84	How accurate does the algorithmic system perform against historical data?	Quality				
85	Has the algorithmic system been subject to external review during or after development?	System description	Management of system			

Continued on next page

Table B.11 – *Continued from previous page*

Ref	Code	Code Groups			
86	Has the status of the assumptions been critically compared to third party sources, or benchmarked against industry norms?	Quality	Control		
87	What are the uncertainties of the algorithmic system?	Risk Assessment			
88	Has a sensitivity analysis been performed to calculate the likelihood of outcomes occurring?	Quality			
89	Are appropriate relationships between variables and hypotheses defined?	Quality	Control	Documentation	
90	Do changes in the inputs/assumptions have a material or significant impact on outputs?	Robustness			
91	Have issues over poor-quality data and assumptions and other identified risks been addressed?	Quality	Robustness		
92	Are AI outputs validated?	Control	Quality	Human involvement	
93	Are decisions based on the AI output proportionate to the robustness of the model?	Control	Robustness	Risk Assessment	
94	Is the model output processed outside of the AI?	System description			
95	Does the model output meet the requirements and aims of the algorithmic system as outlined in the algorithmic system concept?	Objectives	Quality		
96	Are dynamic learning AIs monitored to prevent undesirable and runaway behaviour?	Reliability	Periodic Assessment	Control	
97	Have (KPIs and KRIs) metrics been defined to monitor the algorithmic system's performance and are these adequate?	Control	Quality	Documentation	
98	Have different methods/approaches been selected to evaluate the algorithmic system performance?	Control	Quality		
99	Are AI deployment versions tracked?	Documentation	System description		
100	Is there a periodic review with stakeholders to identify any significant missing items and is reasonableness of targets and tolerances redefined?	Periodic Assessment	Stakeholders	Quality	
101	Is there a clear dashboard available which shows performance results that are easy to understand for stakeholders?	Documentation	Explainability	Quality	Stakeholders
102	Are the results from the algorithmic system presented correctly and understandably so as to ensure all involved parties are adequately informed and are able to understand the core aspects of the algorithmic system?	Stakeholders	Explainability	System description	
103	Based on which interval is the algorithmic system's performance revaluated?	Periodic Assessment			
104	Is an override process in place for exceptions (controllability)?	Control	Quality		
105	Is a process in place to assess exceptions in the AI systems performance?	Risk Assessment	Quality	Control	
106	Has a process been implemented in order to detect input attacks and poisoning of training data?	Security	Management of system		
107	Are users aware they interact with an AI?	Stakeholders	Transparency		
108	Has the algorithmic system been published?	Transparency	Documentation		
109	What documentation and processes are in place to ensure a corporate memory for the algorithmic system exists?	Documentation	Management of system		
110	What process is used to change/update assumptions?	Management of system			
111	What is the process for the routine review of outputs?	Quality	Periodic Assessment		
112	How are model outputs presented to decision makers?	Human involvement	Explainability	Transparency	
113	Are model outputs responsive to organization needs?	Objectives	Periodic Assessment		

Continued on next page

Table B.11 – *Continued from previous page*

Ref	Code	Code Groups			
114	Are predictions compared to actual outputs to validate results?	Quality	Control		
115	Did the organization put in place ways to measure whether its system is making an unacceptable number of inaccurate predictions?	Quality	Control	Management of system	
116	Have measures been taken to ensure that only authorized users have access to the algorithmic system, data and output?	Security	Human involvement	Management of system	
117	Has a separation of development, training and operational environments been implemented?	Quality			
118	Is there a senior responsible owner who approved the AI before deployment?	Management of system	Documentation	Accountability	
119	Does the organization have sufficient documentation in place on governance and quality assurance for their algorithmic system?	Documentation	Quality	Management of system	Control
120	How does the organization regularly monitor the outcome of the algorithmic system against unlawful bias?	Periodic Assessment	Legality	Fairness	
121	Are roles and responsibilities on system and data governance defined?	Accountability	Documentation	Management of system	
122	Is the information security risk management process documented?	Security	Management of system	Documentation	Risk Assessment
123	Is the algorithmic system within the organization business critical?	System description	Management of system		
124	How are algorithmic system outputs challenged and used within the organization?	Management of system			

B.11 SLADA Artificial Intelligence Auditing Framework

The SLADA Artificial Intelligence Auditing Framework was published in 2022 (Becker & Waltl, 2022). Its identified framework elements are shown in Table B.12 below.

Table B.12: *SLADA Artificial Intelligence Auditing Framework Codes.*

Ref	Code	Code Groups			
1	What are the motives for the deployment of AI?	Objectives	Management of system		
2	Is the system described from the point of view of all actors?	Documentation	Stakeholders		
3	Has the complete life cycle of the system been thought out?	Periodic Assessment	Management of system		
4	Have the algorithm, the data and the infrastructure of the system been considered regarding model behaviour?	Management of system	Explainability		
5	Is the implemented algorithm accurate, robust, and interpretable and explainable?	Quality	Transparency	Explainability	Robustness
6	Is the confidence with which a decision is made calculated?	Reliability			
7	Are input, output and metadata considered in the context of the decision?	Explainability			

B.12 SMACTR Framework for Internal Algorithmic Auditing

The SMACTR Framework for Internal Algorithmic Auditing was published in 2020 (Raji et al., 2020). Its identified framework elements are shown in Table B.13 below.

Table B.13: *SMACTR Framework for Internal Algorithmic Auditing Codes.*

Ref	Code	Code Groups			
1	What are the product's requirements/expectations?	Documentation	Objectives		
2	What are the intended use-cases?	Documentation	Management of system	Objectives	
3	What are the ethical objectives, standards and AI principles of the team?	Human involvement	Fairness	Objectives	Management of system
4	Does the technology align with a set of ethical values or principles?	Fairness	Management of system		
5	What is the social impact of the use of AI?	Stakeholders	Risk Assessment		
6	Has an assessment of the severity of the risks been caried out?	Risk Assessment			
7	Have relevant impacts of the AI, applied in its context, been identified?	Risk Assessment			
8	Have the parties involved in the system audit and collaborators in the execution of the audit been outlined?	Human involvement	Documentation		
9	How do the metrics specified in the design of the AI reflect the core values?	Objectives	Control	Management of system	Quality
10	Do any aspects of the algorithm fall outside the scope of the defined measurements and metrics?	Risk Assessment	Control		
11	What assumptions and values underly the metrics?	Objectives	Explainability	Management of system	
12	Are all expected documentations from the development in place?	Documentation	System description	Management of system	
13	Have details on how the model was built been made public?	Documentation	Stakeholders	Transparency	
14	Have assumptions made during model development been made public?	Documentation	Stakeholders	System description	
15	What kind of bias could different groups of people experience?	Stakeholders	Fairness		
16	What mechanisms or procedures were used to collect the data?	Management of system			
17	Was any ethical review process conducted in the data collection?	Management of system			
18	Does the dataset relate to people?	Stakeholders	Privacy		
19	Does non-statistical testing using tailored inputs to the model result in undesirable outputs?	Quality	Fairness	Reliability	
20	What is the importance of each risk?	Risk Assessment			
21	Are there gaps between the intended and actual use of the algorithm?	Objectives	Management of system		
22	What is the threshold for acceptable performance?	Reliability	Quality	Robustness	
23	Are all activities related to the development of the algorithm documented?	Documentation	Management of system		

B.13 Stakeholders-Metrics-Relevancy Auditing Instrument

The Stakeholders-Metrics-Relevancy Auditing Instrument was published in 2021 (Brown et al., 2021). Its identified framework elements are shown in Table B.14 below.

Table B.14: *Stakeholders-Metrics-Relevancy Auditing Instrument Codes.*

Ref	Code	Code Groups		
1	Have all relevant stakeholders and their interests that might even just plausibly be affected by the use of some algorithm been numerated?	Stakeholders	Documentation	Management of system

Continued on next page

Table B.14 – *Continued from previous page*

Ref	Code	Code Groups			
2	Is a complete description of the algorithm available?	Documentation	System description		
3	What is the model's statistical bias?	Quality			
4	Does the model have a societal bias?	Fairness			
5	What is the model's accuacy?	Quality			
6	How robust is the algorithm?	Robustness			
7	How efficiently does the AI use input data?	Quality			
8	How well is the structure of the AI known to stakeholders?	Stakeholders	System description		
9	How transparent is the fact that the algorithm is being used?	Stakeholders	System description	Transparency	
10	How transparent is the collection and processing of data to stakeholders?	Transparency	Management of system		
11	What is the potential for the AI to be used to infringe on stakeholder rights or be used in other dangerous ways?	Legality	Stakeholders	Risk Assessment	
12	Does the very use of the AI violate stakeholder rights?	Legality	Stakeholders		
13	Who and what has the ability to use the AI and access the associated data?	Accountability	Management of system	Human involvement	System description
14	Who can use the algorithm?	Human involvement	System description		
15	How secure is the data associated with the AI?	Security			
16	For each stakeholder interest, how much could each metric threaten that interest if the algorithm performs poorly with respect to that metric?	Stakeholders	Quality	Risk Assessment	

B.14 Recommendations Toward Trustworthy Artificial Intelligence Development

The Recommendations Toward Trustworthy Artificial Intelligence Development were published in 2020 (Brundage et al., 2020). Its identified framework elements are shown in Table B.15 below.

Table B.15: *Recommendations Toward Trustworthy Artificial Intelligence Development Codes.*

Ref	Code	Code Groups	
1	What level of privacy protection can be guaranteed?	Privacy	
2	How well is the AI system tested for safety?	Safety	
3	How well is the AI system tested for security?	Security	
4	How well is the AI system tested for ethical concerns?	Fairness	
5	What are the sources of data?	System description	
6	What are the sources of labor?	Human involvement	Accountability
7	Can the accuracy of previous claims made by the developers be confirmed?	Control	Quality

C General AI Auditing Framework

Sources in the table denote which questions from which frameworks were the basis for the question. The referencing is set up as "$n.m$" where "n" denotes the original framework and "m" the specific framework element. Additionally, each question in the general auditing framework has its own reference number as well (leftmost table column) for reference to specific questions in the Results and Discussion chapter.

Table C.1: *1. Human Agency and Oversight.*

Ref.	Question	Source(s)				
1.1	Has a legal assessment of the complete AI system been performed and documented?	10.22	9.1	9.33	9.20	
1.2	Is the AI system periodically assessed for legal alignment and preparation for new regulations?	8.11	1.45	1.126	1.11	
1.3	Have all stakeholders and their rights been considered in the assessment?	13.12	13.11	1.11	1.4	9.18
1.4	Has data been handled in accordance with applicable laws and regulations?	2.55	10.66			
1.5	Has the legality of data use been documented?	6.28				
1.6	Has the GDPR been adhered to?	2.3	5.8			
1.7	Has legal training been provided to those involved in the development of the AI system?	1.124	1.128			
1.8	What is the extent of decision-making autonomy of the AI model?	2.6	2.52	10.12		
1.9	How do the duties of humans in the system relate to those of the AI system?	10.6	1.26	1.17	10.7	
1.10	Are roles and responsibilities clearly defined and documented?	8.6	9.12	9.32	6.10	
1.11	Has somebody been appointed to ensure adherence to organizational values?	4.5	1.125			
1.12	Has senior ownership of the AI system been assigned?	10.118	9.10			
1.13	Have data governance roles and responsibilities been assigned?	10.121				
1.14	Have monitoring and risk management review roles and responsibilities been assigned?	9.5	9.8			
1.15	Has management considered a strategy for the complete lifecycle of the AI system?	2.48	11.3	4.4		
1.16	Does the strategy include the situation where the AI system is no longer aligned with its intended use?	9.66	10.72			
1.17	Has the complete lifecycle strategy of the AI system been documented?	2.15	10.109	2.26	2.28	
		4.26	6.11	4.27	8.1	
1.18	Is the lifecycle documentation sufficiently detailed to ensure auditability: is it clear what decisions were made and why?	2.64	10.38	1.120		
		8.9	10.119	1.121		
1.19	Is there a strategy to retire the system, and has it been documented?	10.35				
1.20	Have decommissioning risks been assessed and addressed?	4.38	4.39	9.7		
1.21	Is the AI system periodically monitored, maintained and updated as to prevent quality decay?	2.35	9.68	1.48	4.33	
1.22	Do updates to the AI system follow a change protocol?	2.78	4.35	10.110		
1.23	Are all stakeholders involved in the continual improvement of the AI system?	9.70				
1.24	Are stakeholders able to withdraw from the AI system?	1.64				
1.25	Are subjects able to object to being part of an AI system if it facilitates automated decision-making?	2.53	10.29			

Continued on next page

Table C.1 – *Continued from previous page*

Ref.	Question	Source(s)				
1.26	Has organizational readiness for using an AI system been assessed?	6.16	6.29			
1.27	Is the use of an AI system in line with organizational principles?	5.12	8.5	3.1		
1.28	Is the AI system periodically monitored for alignment with predefined organizational ethical values?	4.34				
1.29	Does the organization possess adequate technological proficiency in AI development?	2.13	1.83	9.1	2.14	9.31
1.30	Has resource allocation been adequately executed?	8.3	9.6			
1.31	Does the organization follow a common language within the AI system?	2.72	8.14	10.1		
1.32	Are there gaps between the intended and actual use of the AI system?	12.21	10.32	10.113		
		8.22	9.48	8.21		
1.33	Is the AI system monitored for alignment with organizational needs?	10.113	9.43			

Table C.2: *2. Technological Robustness and Safety.*

Ref.	Question	Source(s)				
2.1	Have performance metrics been defined for the AI system?	2.45	2.43	10.98	10.97	4.21
2.2	Has an agreed-upon threshold for acceptable performance levels been determined and documented?	10.9	4.8	12.22	7.7	2.17
		10.115	7.3	2.43	12.1	4.6
		2.45				
2.3	Does the AI system perform at the defined acceptable level?	10.72	3.3	4.20	2.20	5.13
		8.21	5.11	4.20	10.95	10.92
		9.59	8.22			
2.4	Can the model performance be validated?	2.46	14.7	10.114		
2.5	How well does the AI system score on the performance metrics?	13.5	10.74	13.3	11.5	4.17
		8.15	10.84	11.6	5.10	
2.6	Has the responsibility for performance measurement been assigned?	8.6				
2.7	Does the model performance validation occur periodically and is it documented?	1.48	1.69	10.111	8.16	8.17
		10.119	10.65			
2.8	Are processes in place to identify risks based on actual performance?	9.53	4.31	1.46		
2.9	Have bias metrics been defined and documented?	12.9				
2.10	Is the AI system periodically assessed for bias?	4.34	10.120			
2.11	Has bias been assessed through deliberate input manipulation?	3.4	12.19	3.7		
2.12	Are stakeholders included in the evaluation of bias?	10.100	1.87	1.47	13.16	
2.13	Are model performance and limitations communicated to all stakeholders?	1.50	10.101	2.47	9.27	
2.14	Has the quality of the training data been assessed?	10.58	4.11			
2.15	Have the data sources been verified and documented?	10.59				
2.16	Has the training data been assessed for alignment with organizational values?	1.81				
2.17	Has the training data been assessed for alignment with model purpose?	10.47	3.11	2.40		
2.18	Has the training data annotation been validated?	10.57	10.46			
2.19	Is the model input data assessed for quality?	1.78	10.77	3.10		
2.20	What controls are in place to ensure consistency in data processing?	2.34				

Continued on next page

Table C.2 – *Continued from previous page*

Ref.	Question	Source(s)				
2.21	Has the model been trained in line with sound development practices?	5.14				
2.22	Have train/validate/test data been processed separately?	2.39	2.44	5.9	2.36	10.117
2.23	What trade-offs have been made between performance and other values?	2.24	3.8			
2.24	Have these trade-offs been documented and communicated to stakeholders?	10.20				
2.25	Do developers demonstrate a profound understanding of the model?	10.68	1.83			
2.26	Are all assumptions underlying the model substantiated and documented?	10.75	1.127	10.91	10.86	
2.27	Has the model performance been tracked and recorded during development?	4.18	4.22	9.52		
2.28	Has the quality of the model been verified, when supplied by a third-party?	10.80	9.67			
2.29	Has the choice of input variables been based on sound hypotheses?	10.89	13.7	5.5		
2.30	Has a user management policy been defined, documented and practiced?	10.79	2.67	10.52	10.41	
2.31	Are fallback plans in place in case of system failure or other unexpected malfunctioning?	1.56	10.70	9.19		
2.32	Has the AI system been backed up?	3.18	1.44	10.70		
2.33	Are adversarial effects due to AI system failure assessed?	1.27	1.49	14.2		
2.34	Have AI system failure metrics been determined and documented?	10.115	4.25			
2.35	Is the AI system safety reassessed periodically?	1.45	9.45			
2.36	Is the AI system access secured?	2.79				
2.37	Are policies in place to ensure safe system retirement?	9.7				
2.38	Are policies in place to foster a safety-first mindset within the organization?	9.13				
2.39	What data access policies are in place?	10.51	8.10	3.12	10.52	10.41
2.40	Are access rights periodically checked for being up-to-date?	2.65	2.70			
2.41	Are access rights changed as soon as an employee leaves office or changes position?	2.66				
2.42	Are access rights handed out by authorized personnel?	2.67				
2.43	Has a data protection impact assessment (DPIA) been carried out?	10.67	1.7	4.13		
2.44	Is there a separation of duties policy in place for access right management?	2.75	2.73			
2.45	Are levels of access and associated privileges defined and documented?	2.71	2.69	2.72	3.12	
2.46	Is the user environment separated from the developer environment?	2.74	2.68	10.79	10.116	
2.47	Are password policies in place?	2.76	2.77			
2.48	Are procedures in place to detect malicious input?	10.61	10.106	1.40		
2.49	What security practices are in place to protect input and output data?	10.53	10.60	1.9	13.15	6.21
2.50	Is there a procedure to immediately abort and cease the operations of the AI system?	1.25				
2.51	Have cybersecurity risks been assessed and documented?	8.24 1.29	14.3 1.33	7.2 1.31	9.46	1.30
2.52	Are policies in place to address the identified cybersecurity risks?	10.36	10.122	10.78	1.32	

Continued on next page

Table C.2 – *Continued from previous page*

Ref.	Question	Source(s)				
2.53	Are cybersecurity defenses and responses in place?	8.2	2.79	2.82		
2.54	Are specific cybersecurity standards adhered to?	1.28				
2.55	Is there a way for third parties to report security vulnerabilities?	10.81				
2.56	Are stakeholders informed about security updates?	1.34				
2.57	Has the model been subjected to varying inputs in order to assess its reliability?	3.5	9.44			
2.58	Have patterns of failure been identified?	4.24	1.53			
2.59	Are procedures in place to patch patterns of failure?	1.57				
2.60	Have risks associated to AI system reliability been assessed?	1.51	1.42			
2.61	Has the model robustness been assessed and documented?	13.6	11.5	1.55	1.54	
2.62	Has the model consistency been assessed through controlled inputs?	3.5	3.4	4.15	3.2	10.73
		10.90	3.7			
2.63	Have acceptable levels of robustness been defined?	12.22				
2.64	Does the AI system meet robustness requirements?	3.3				
2.65	Is the robustness of the AI system aligned with its goals?	1.43	1.42	10.93		
2.66	Has output consistency been tested for known problematic use cases or cases with a higher risk of introduced bias?	1.82	10.91	1.53	4.23	4.24
2.67	Is the robustness reevaluated following system changes?	1.45	1.32			
2.68	Have safety critical levels of possible system impact been defined?	1.41				
2.69	On which aspects is the AI system periodically monitored?	2.19	9.64			
2.70	At what interval does the monitoring of those aspects take place?	10.103	9.5			
2.71	Are independent assessors or internal experts not directly involved with the AI system involved in the regular monitoring?	9.39				
2.72	Is the AI system monitored to prevent the development of undesired bias?	10.96	1.2	10.120	1.84	
2.73	Is the AI system output monitored to prevent decreased performance?	1.26	4.33	10.111	4.32	4.20
		1.69	9.57	10.96	9.58	
2.74	Is the AI system input data monitored for system suitability?	1.69	4.31			
2.75	Are system metrics evaluated for effectiveness?	9.38	9.57			
2.76	Is the AI system monitored for safety?	9.45	1.37	1.32	1.9	
2.77	Is the periodic assessment documented?	9.53	9.36			
2.78	Is there a protocol to mitigate newly identified risks and rectify issues found?	9.53	8.17	8.16		
2.79	Do updates to the AI system follow a change protocol?	4.35				
2.80	Are all changes tracked and logged to ensure traceability?	1.68				

Table C.3: *3. Privacy and Data Governance.*

Ref.	Question	Source(s)				
3.1	Has data been handled lawfully throughout the AI system lifecycle?	2.55	10.66			
3.2	Has the legality of data used been documented?	6.28				
3.3	Has the GDPR been adhered to throughout the AI system lifecycle?	2.3	5.8	1.63	1.7	
3.4	Are data management policies in place and documented?	12.16	10.53	10.55		
3.5	Do data management policies cover the complete AI system lifecycle?	1.65				

Continued on next page

Table C.3 – *Continued from previous page*

Ref.	Question	Source(s)				
3.6	Are data management policies in line with rules and regulations, such as the GDPR?	10.54	1.67			
3.7	Has the used data been assessed for quality, such as its representativeness, alignment with goals, consistency across datasets, and correct annotations?	2.40	3.10	12.17	10.47	10.58
		10.57	8.18			
3.8	Has a DPIA been carried out and documented?	2.51	2.63	10.66	10.37	10.67
		1.8	9.49	1.65	1.66	
3.9	Have DPIA findings been addressed?	4.12	8.23	6.26	6.27	
3.10	Has the role of data protection officer (DPO) been assigned?	10.37	2.57			
3.11	Can the DPO be contacted by any stakeholder to raise data and privacy related issues?	1.61				
3.12	Does the organization have complete ownership of the data used for the AI system?	2.41				
3.13	Has data minimization been applied?	10.44	2.42	10.27	10.26	
3.14	Have data choices been substantiated and documented?	2.49				
3.15	Does any of the used data include personal data?	12.18	10.23	1.62	10.66	10.42
3.16	Are policies in place to minimize and de-identify the data?	2.54	10.26	10.56	10.43	
3.17	Are those who own the personal data informed that their data is used?	2.60				
3.18	Has the need to use personal data been assessed?	2.56	10.27	3.8	10.24	
3.19	Are data processing procedures documented and publicly accessible?	2.63				
3.20	Are data storage procedures in place and documented?	2.63				
3.21	Is the data storage policy compliant with GDPR regulations?	4.13	3.13	2.50	1.9	1.60
3.22	Does the organization communicate a guarantee of privacy protection to stakeholders?	6.22	3.10	14.1		

Table C.4: *4. Transparency.*

Ref.	Question	Source(s)				
4.1	Is the model underlying the AI system publicly accessible?	2.30	1.97	3.15	10.108	
4.2	Is the data used to develop the model publicly accessible?	2.31	4.11			
4.3	Have model development practices been published?	12.13	1.68	1.73	2.27	
		12.12	9.40	10.99		
4.4	Is a complete description of the AI system publicly available?	13.2	13.8	9.6	11.2	
4.5	Is there a guide describing the logical flow of the AI system?	10.45	10.49	10.102	3.15	
4.6	Are stakeholders informed about the goal of the AI system?	2.22	1.77	2.62	3.16	
		2.21	1.76	1.14	1.74	
4.7	Are stakeholder informed about the logic behind the AI system?	2.61	10.15	13.8	12.13	12.14
4.8	Are stakeholders made aware of the fact that they interact with an AI?	10.107	1.15	1.16	13.9	1.75
4.9	Are stakeholders informed about AI system limitations?	2.47	1.78	1.76	1.38	1.50
4.10	Are stakeholder informed about AI system performance levels?	1.50				
4.11	Are stakeholders informed about organizational guiding values?	9.24	4.2	4.3	10.20	
4.12	Is AI system output clearly presented to all stakeholders?	10.101	3.17	1.74	10.102	10.64
		10.11	2.23	1.71	11.5	2.61
		9.48	10.30	10.112	11.7	8.20
4.13	Are stakeholders informed about what data is used for the AI system?	2.31	12.18	2.63	10.43	
		2.60	1.70	10.48	13.10	

Continued on next page

Table C.4 – *Continued from previous page*

Ref.	Question	Source(s)				
4.14	Are incidents and errors communicated to all stakeholders?	9.71	9.14			
4.15	Can stakeholders report any (perceived) issue with the AI system?	10.81	1.61	1.130	9.55	
		4.36	9.16	1.129	1.24	
4.16	Can stakeholders object to being subjected to the AI system?	2.53	2.60	10.25	1.64	10.29
4.17	Are stakeholders informed about AI system data security?	1.34				
4.18	Are data choices explainable in the context of the intended purpose of the AI system?	2.40	2.61	11.4		
4.19	Is the AI system output explainability sufficient for the intended purpose of the AI system?	5.7	4.15			
4.20	Have assumptions and design choices been substantiated and documented?	12.11	2.29	12.14		
		3.6	4.19	5.14		
4.21	Have value trade-offs made during the AI system development been documented?	10.20	1.127	2.24		

Table C.5: *5. Diversity, Non-discrimination and Fairness.*

Ref.	Question	Source(s)				
5.1	What mechanisms are in place to prevent undesirable bias in the AI system?	10.16	1.94	1.131		
5.2	Has fairness of the AI system been defined, e.g. through objectives/principles/standards/policy?	10.17	12.3	6.12		
5.3	Were other definitions of fairness considered?	1.91				
5.4	Is the definition of fairness in line with laws and regulations?	2.59				
5.5	Is the AI system periodically reviewed for alignment with fairness as defined?	4.34	10.120	12.4	4.7	1.90
5.6	Has the AI system been assessed for bias during all stages of its lifecycle?	1.2	1.99	1.84		
5.7	Have roles and responsibilities been assigned regarding fairness of the AI system?	8.6	1.125			
5.8	Have all stakeholders been involved in the assessment of the potential for biases in the AI system?	12.15	1.92	12.15	1.103	1.109
		2.33	10.19	13.1	1.101	1.104
5.9	Were third-parties consulted for assurance on ethical concerns at any stage?	1.122	1.123			
5.10	Can perceived biases be reported by stakeholders?	10.81	1.87	1.129	1.88	9.55
5.11	Has the input data been assessed for biases?	8.19	5.4	1.80		
5.12	Have system development choices been assessed for biases (e.g. type of model used)?	2.37	1.80	4.15		
5.13	What value trade-offs have been made between values during model development?	3.8				
5.14	Have the value trade-offs been documented?	10.20	1.127			
5.15	Have assumptions and design choices during system development been substantiated and documented?	2.29	12.14	3.6	4.19	5.14
5.16	Have metrics been defined to quantify bias in the AI system?	1.93	9.50	10.40	14.4	10.65
5.17	Have input manipulation tests been performed to assess system bias for known subpopulations in the data?	3.4	12.19	1.102	1.82	3.7
5.18	What bias rectification policies are in place?	1.3	10.65			
5.19	Does the AI system display undesirable bias?	3.14	13.4	1.1	2.58	8.21
5.20	Has the training data been evaluated for undesirable biases?	2.38	4.11	1.81	1.85	

Continued on next page

Table C.5 – *Continued from previous page*

Ref.	Question	Source(s)				
5.21	Are all stakeholders informed about AI system capabilities and limitations?	2.47				
5.22	Are all stakeholders informed about data management practices?	13.10				
5.23	Is policy in place to inform all stakeholders in case of incidents?	9.71				
5.24	Is training offered to those within the organization whose work is related to the AI system, covering ethical and technical aspects?	1.115	1.79	9.9	1.23	1.128
		1.113	1.111	9.13	1.86	
5.25	How is stakeholder feedback incorporated in the AI system?	9.17	9.69	1.130	1.131	9.16
5.26	Has a set of organizational guiding values been defined and documented?	12.4	4.1	12.3	6.12	10.17
5.27	Are the guiding values complete in covering the characteristics of trustworthy AI?	9.2				
5.28	Do specified fairness metrics reflect these values?	12.9	9.42	12.11		
5.29	Have the guiding values been communicated to all stakeholders?	4.2	4.3			
5.30	Have stakeholders been involved for input in the complete lifecycle of the AI system?	1.101	1.11			
5.31	Have stakeholders been involved in the development of the AI system?	2.33	10.35			
5.32	How are stakeholders included in periodic review and improvements of the AI system?	10.100	9.70	9.17	9.36	9.69
		1.12	9.58	1.34	1.131	
5.33	How are stakeholders included in the implementation of the AI system?	10.19	9.56			

Table C.6: *6. Societal and Environmental Well-being.*

Ref.	Question	Source(s)				
6.1	Has the overall environmental impact and risks of the AI system been assessed and documented?	10.71	1.105	1.106	4.9	
		6.32	9.51	6.2	6.39	
6.2	Have measures been defined, documented and practiced to reduce the environmental impact of the AI system?	1.107				
6.3	How does the AI system output affect the environment?	6.34	7.38	6.2	6.37	6.38
6.4	What is the energy demand of the AI system across its lifecycle?	6.20	6.25	6.30	6.31	6.1
6.5	What is the environmental footprint of the infrastructure required to operate the AI system (e.g. data storage and processing)?	6.4	6.32			
6.6	What is the environmental footprint of the supply chain of the AI system (e.g. hardware, training data)?	6.5				
6.7	Has the impact of the AI system on society as a whole been assessed?	10.5	1.13	1.12	1.110	1.119
		12.5	1.116	12.7	1.118	1.113
		1.112	6.35	6.36	6.39	1.117
		1.27	1.108	1.103		
6.8	Could the use of the AI system potentially affect human rights?	10.4	1.5	1.6	1.4	1.60
6.9	Has the impact of the AI system on human workers been assessed?	1.108	1.110	1.112		
6.10	Has the AI system been assessed for undesirable discrimination?	3.14	13.4	1.1	2.58	8.21
6.11	Has it been made clear to relevant stakeholders that they interact with an AI?	10.107	1.15	1.16	13.9	1.75
6.12	Is the model underlying the AI system publicly accessible?	2.30	1.97	3.15	10.108	

Continued on next page

Table C.6 – *Continued from previous page*

Ref.	Question	Source(s)				
6.13	Has the privacy of stakeholders been respected in accordance with the GDPR?	2.3	5.8	1.63	1.7	

Table C.7: *7. Accountability.*

Ref.	Question	Source(s)				
7.1	Have roles and responsibilities on the AI system been defined and documented?	2.16	14.6	8.6		
7.2	Has final/executive responsibility been assigned?	10.121	9.10	11.118	6.10	
7.3	Has a support representative for stakeholders to raise issues to been assigned?	1.88	1.131			
7.4	Have roles and responsibilities regarding system retirement been assigned?	9.66				
7.5	Have roles and responsibilities regarding system implementation been assigned?	6.18				
7.6	Have roles and responsibilities regarding system monitoring been assigned?	9.1	9.8	4.5	9.10	9.11
7.7	Have roles and responsibilities regarding human-system configurations (i.e. operating the AI) been assigned?	9.12	13.13	6.10		
7.8	Has an ethics review board been assigned?	1.125				
7.9	Are all roles and responsibilities practiced as defined?	2.16				
7.10	Are those with responsibilities in possession of necessary competencies?	8.7	6.8			
7.11	Is training and education provided to develop necessary competencies?	10.39	1.23	1.114		
7.12	Has a risk assessment of the AI system been performed and documented?	6.39	2.12	1.39	6.13	
		9.14	6.17	8.4	4.16	
7.13	Does the risk assessment cover the complete lifecycle of the AI system?	12.10	10.76	7.2		
		4.38	4.28	9.56		
7.14	Does the risk assessment cover organizational strategy and policies?	9.4	9.3	7.5	7.6	9.15
7.15	Have societal risks been assessed?	10.5	1.13	1.12	1.110	1.119
		12.5	1.116	12.7	1.118	1.113
		1.112	6.35	6.36	6.39	
		1.27	1.108	1.103	1.117	
7.16	Have environmental risks been assessed?	6.2	1.106	4.9	10.71	6.39
7.17	Has a DPIA been carried out?	2.51	1.8	10.37	1.60	
7.18	Have legal risks been assessed?	9.33				
7.19	Have the risks for unlawful discrimination been assessed?	10.65	1.103			
7.20	Have the risks and consequences of system failure/maluse been assessed?	1.40	10.82	10.105	10.69	
		1.46	1.27	1.51	4.32	
		1.42	10.87	9.19		
7.21	Have technical risks been assessed?	1.49	4.31	1.59	4.33	
		5.13	10.93	9.34		
7.22	Have risks following from system transparency and accountability been assessed?	9.47				
7.23	Have security risks been assessed?	10.122	1.30	1.31		
7.24	Have the risks of overreliance on the AI system been assessed?	1.17	1.18	1.19		

Continued on next page

Table C.7 – *Continued from previous page*

Ref.	Question	Source(s)				
		1.21	1.22	10.122		
7.25	Is the risk assessment carried out periodically?	1.37	9.53	9.5	9.54	
7.26	Has the severity of the identified risks (chance x impact) been assessed?	12.6	6.14	1.36	6.15	12.20
		9.35	9.29	7.8	13.11	
		6.15	13.16	10.28	9.60	
7.27	Have the risks of the AI system been considered for all stakeholders?	10.8	1.6	13.16	1.102	1.24
		1.19	9.62	6.23	6.24	13.11
		1.22	1.20	13.1	1.10	
		13.12	10.4	1.89	1.60	
		1.51	9.58	6.33	1.100	
7.28	Do the stakeholders included in the risk assessment reflect demographic diversity?	9.21	12.15	1.92	1.102	
		1.95	1.96	1.97	1.98	
		1.85	1.104	9.41	1.99	
7.29	Are the identified risks communicated to the stakeholders?	1.38				
7.30	Is there a publicly accessible way for people to voice any concern regarding the AI system?	10.81	1.61	1.88	1.87	
		4.36	1.130	9.16		
7.31	Has organizational risk tolerance been defined and documented?	9.3	1.41	7.3	9.23	
7.32	Is the AI system aligned with the organizational risk tolerance?	10.3	9.4			
7.33	Is the added value of the AI system periodically assessed?	9.64	9.43			
7.34	Has risk training been offered to employees?	1.124				
7.35	Are processes in place to respond to identified risks?	1.3	1.5	3.9	8.8	
		4.37	10.36	9.37		
7.36	Have the identified risks been responded to?	1.107	1.118	9.65	4.16	
		9.61	4.12	9.18	9.63	
7.37	Have third parties been involved for external review?	1.123	1.122	10.85	12.8	
7.38	Have third-party agreements been made and documented?	2.18				
7.39	Have the motives for and against using AI for its intended purpose been substantiated and documented?	11.1	10.3	10.33		
		10.13	8.4	6.13		
7.40	Are there gaps between the intended and actual use of the AI system?	12.21	10.32	10.113		
		8.22	9.48	8.21		
7.41	Has the development process been documented (e.g. through change logs)?	2.27	12.12	9.40	10.99	1.55
7.42	Have assumptions and design choices been substantiated and documented?	2.29	12.14	3.6	4.19	5.14
7.43	Are all system changes logged to ensure traceability?	1.68				

D Interviews

Approval to conduct these interviews was granted by the Human Research Ethics Committee TU Delft on 30-06-2023. Their letter of approval is included in Figure D.1.

Date 30-Jun-2023
Contact person Grace van Arkel, Policy Advisor Academic Integrity
E-mail E.G.vanArkel@tudelft.nl

TU Delft Delft University of Technology

Human Research Ethics Committee TU Delft
(http://hrec.tudelft.nl)

Visiting address
Jaffalaan 5 (building 31)
2628 BX Delft

Postal address
P.O. Box 5015 2600 GA Delft
The Netherlands

Ethics Approval Application: Auditing AI
Applicant: Sewandono, Tijn

Dear Tijn Sewandono,

It is a pleasure to inform you that your application mentioned above has been approved.

Thanks very much for your submission to the HREC which has been conditionally approved. Please note that this approval is subject to your ensuring that the following condition/s is/are fulfilled:

IC:
i: For the street IC, please add 'voluntary withdrawal' to it.

In addition to any specific conditions or notes, the HREC provides the following standard advice to all applicants:
• In light of recent tax changes, we advise that you confirm any proposed remuneration of research subjects with your faculty contract manager before going ahead.
• Please make sure when you carry out your research that you confirm contemporary covid protocols with your faculty HSE advisor, and that ongoing covid risks and precautions are flagged in the informed consent - with particular attention to this where there are physically vulnerable (eg: elderly or with underlying conditions) participants involved.
• Our default advice is not to publish transcripts or transcript summaries, but to retain these privately for specific purposes/checking; and if they are to be made public then only if fully anonymised and the transcript/summary itself approved by participants for specific purpose.
• Where there are collaborating (including funding) partners, appropriate formal agreements including clarity on responsibilities, including data ownership, responsibilities and access, should be in place and that relevant aspects of such agreements (such as access to raw or other data) are clear in the Informed Consent.

Good luck with your research!

Sincerely,

Figure D.1: *Human Research Ethics Committee TU Delft Letter of Approval.*

D.1 Assurance Professionals

The interviews were conducted in Dutch as all this was the mother tongue of all participants. The following (translated) set of questions was used to guide the interview:

1. **Interviewee**
    - What is your professional experience in IT assurance?
    - What is your background in AI?
    - How many AI assurance jobs have you done?
2. **Auditing AI**
    - What motivates clients to have you audit their AI system?
    - How would you define trustworthy AI?
    - Do you think the audit of AI can contribute to trustworthy AI, and why?
    - What role do you expect the audit of AI to play in the future?
3. **Scoping the AI Audit**
    - How do you with the client derive the themes that fall within the audit scope?
    - Do you encounter difficulties in this process?
    - Could you reflect on the utility of the translated ESG materiality assessment framework as presented to you for the scoping of an AI audit?
4. **General Framework**
    - How do you currently translate the determined audit scope to a concrete set of questions to be covered in the audit?
    - What is the role of the client in this process?
    - Do you encounter any problems or hardships when aggregating the list of questions?
    - Imagining the general framework from the materiality assessment to pick questions from, would this solve some of the issues in this step of the audit?
    - What would then be the requirements for this general framework for it to be actionable?
    - In the general framework, some questions might appear multiple times as they are relevant for multiple themes within the framework. Can you reflect on the strengths or weaknesses of this situation?

The materiality assessment as discussed in the interviews (question 3.3) was presented as shown in Figure D.2 below.

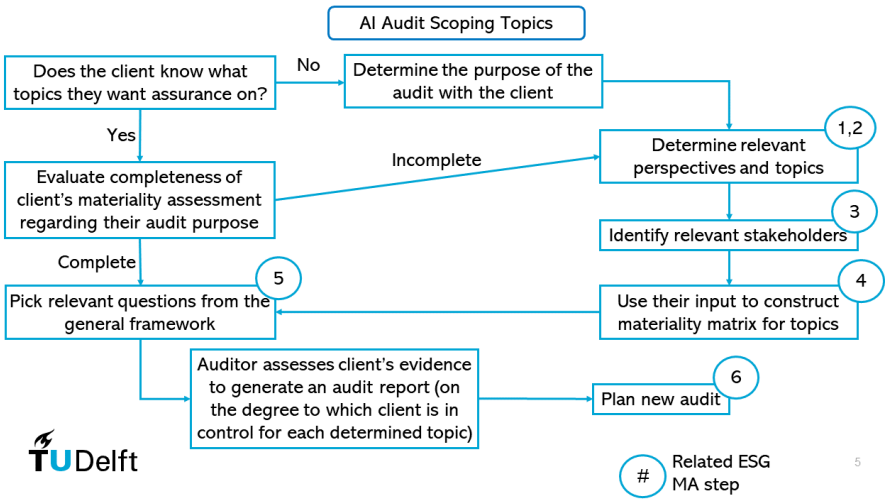


Figure D.2: *Draft Materiality Assessment as Discussed in Interviews. The shown draft of the audit scoping approach is a modification of the ESG reporting materiality assessment found in literature (Garst et al., 2022).*

Excerpt Interview 1

Reported answers are summaries of the discussion had with the interviewee that followed each of the questions. The interview excerpt has been validated by the interviewee prior to publication.

Q: What is your professional experience in IT assurance? A: I am part of the data and technology team within digital trust at this Big Four firm, focused mainly on data assurance. This includes the validation of data streams and models or actual scripts/lines of code. It also encompasses contributing to making the audit more data-driven instead of reliant on sampling by the auditor. Other than it also involves advisory roles on data quality or anything data related. I have worked there for five years now.

Q: What is your background in AI? A: Our team has spearheaded responsible AI, so occasionally we will aid clients in benchmarking their AI environment. This means not just checking the literal code but can include everything around that. There are currently no legal frameworks for that, which is why we do benchmarking instead – comparing the client's AI environment to best practices from leading research or alignment with other laws and regulations. In the end these jobs revolve around mitigating the client's risks.

Q: How many AI assurance jobs have you done? A: The AI assurance jobs are relatively new, so there have not been many – about a handful. I have not been involved with each of them, but personally I have been/am currently part of three jobs in various roles.

Q: What motivates clients to have you audit their AI systems? It often comes down to stakeholder management. For example, we receive a client that processes a lot of data and must report to their stakeholders that they do so responsibly. Stakeholders could be management or other organizations with whom they collaborate. They solicit our services to preempt questions about handling data and operating their AI responsibly. That way they can indicate that a reputable external party has checked the AI systems.

Then, based on the audit findings, the client can demonstrate control and that their AI system is up to par to their stakeholders. Another example was a client who had garnered negative press in the United States. In order to prevent AI related incidents in Europe they solicited our assurance services. So predominantly clients want to anticipate questions regarding the trustworthiness of their AI system and be able to demonstrate this to stakeholders. In lesser extent our findings will also be used internally at the client for improvements of course.

Q: How would you define trustworthy AI? A: That is hard, as I do not think AI will ever be 100% trustworthy. From an assurance perspective, it means that the margin of error of the model is small enough that the AI is robust and accurate. For me trustworthy AI means that there is sufficient technical information while the processes around the AI system are also guaranteed to be working as intended. It is broader than just the code and the person who wrote it.

Q: Do you think the audit of AI can contribute to trustworthy AI, and why? A: I do think so yes, at least to a certain extent. When a third party goes over your AI system as part of an audit it will most likely prevent major incidents. Of course an audit can never go over every single detail and thereby offer a 100% guarantee. But the most important risks will be checked and mitigated. Meanwhile this field is still developing so rapidly, and comes with many pitfalls and challenges. So perhaps only those with in-depth technical knowledge have a proper overview of the risks of the AI system. Compared to the financial audit, the challenge here is that the scope of risks is much wider than in a financial statement for example because the technology is more of a 'moving target'.

Q: What role do you expect the audit of AI to play in the future? A: I do expect it to become a bigger thing in the future. Especially when laws and regulations come into effect, this is always a huge motivation for companies to solicit the services of an external party. Some clients do have more of an intrinsic motivation to 'do everything right.' It could also entail a more advisory based role on how a client could effectively put their AI to use. But I can not imagine that the audit of AI will not become a larger topic. It is just unclear at this point at which rate and how exactly – this depends on developments regarding laws and regulations. If these are put into effect quite strictly then the audit will become especially important, and if not then it will still be very relevant for companies and institutions with a great societal impact as they ought to demonstrate being in control of their AI system to their stakeholders.

Q: How do you with the client derive the themes that fall within the audit scope? A: We first check which regulations are currently relevant for the client. We also check if we can specify towards the purpose of the AI system. Then based on practices in the client's sector and papers we aggregate those into list of themes to investigate. We coordinate this with the client at this stage.

Q: Do you encounter difficulties in this process? A: Not many, but it is important to keep the job manageable in the sense that it should not cover all topics as this will not be reasonably possible given the allocated time and resources.

Q: Could you reflect on the utility of the translated ESG materiality assessment framework as presented to you for the scoping of an AI audit? A: The first

step makes sense conceptually as it is logical for the client to take this step. So far when we did the benchmarking we would just take everything into account. But when broadening the audit scope this becomes impossible, so this framework would be a good starting point to engage into that scoping discussion. Furthermore I think that the stakeholder inclusion part of the framework is very relevant. To clients that are focused on data processing it is important to take the concerns of their stakeholders into account, like being GDPR-proof and proper data security. And for clients with a public role their societal impact will be relevant for the audit. So it is good that these perspectives are taken into account during the scoping phase. The challenge will be what to do when stakeholders have very different perspectives on what is relevant, will you then still audit everything? Then you have to consider which stakeholder is more important and should therefore be given more weight in the materiality assessment. The setup makes sense. It may be hard for an outsider to evaluate whether the assessment is complete, but the framework will help explain how the scoping has been derived. By reporting the method and steps taken in this materiality assessment, one can cover their bases as they can demonstrate having put in adequate effort in uncovering all relevant stakeholders and their perspectives. You could also include a more risk-based approach, as through that the auditor can indicate what the biggest risks are that they cover through an audit. In the end risk mitigation is the goal of an audit.

Q: How do you currently translate the determined audit scope to a concrete set of questions to be covered? A: We basically aggregate questions from existing frameworks based on the themes that were agreed upon with the client. In that we also make a professional judgement on what questions are relevant and if they are redundant. That way we try to limit the length of the list. In the end you end up with a list of questions per theme within the scope. For example under data and AI ethics you will find that it is best practice to have an ethics committee within the organization. We also include references for each question. The we ask the client to provide evidence on each question and in the end report our findings.

Q: What is the role of the client in this process? A: We usually just do this by ourselves. We then do the testing of the questions together with the client as it involves Q&A. When we share our findings with the client, on points where the client has room for improvement they sometimes then indicate that that point is not as relevant to them. In the end it is up to them what they do with it as there are no laws and regulations, so there is room for this discussion. Typically these points are not removed from the report but included as lower priority and the client can then decide to improve on that themselves.

Q: Do you encounter any problems or hardships when aggregating the list of questions? A: Determining what questions are relevant from the various source materials is difficult, some similar form of materiality assessment could be helpful in that process.

Q: Imagining the general framework from the materiality assessment to pick questions from, would this solve some of the issues in this step of the audit? A: Yes of course this would simply the process. It usually takes a lot of time and discussion within the audit team before a final set of questions is obtained. Additionally, it also allows for a more standardized form of auditing as you will be basing the audit of all clients on the same greater set of questions.

Q: What would then be the requirements for this general framework for it to be actionable? A: Ideally it is the commonly accepted framework that everybody uses to audit AI, as otherwise there is room for discussion on the questions included. What we encounter with current frameworks is that they provide a high-level guidance only, but do not concretize what questions should be asked. While for IT accountants everything in the audit is prescribed, for example for certain risks there are guidelines on how many samples of a transaction should be checked. That does not exist for the AI topics, so we will base our approach on best practices. Such a framework that is not as high-level and more prescriptive will move the AI audit towards a more standardized format, which is where the field ought to want to move towards. As professionals we can even take the initiative in this, for example through developing a framework, instead of waiting for laws and regulations to be shaped and become into effect. This will in a sense force regulatory bodies to position themselves and ultimately propose laws and regulations.

Q: In the general framework, some questions might appear multiple times as they are relevant for multiple themes within the framework. Can you reflect on the strengths or weaknesses of this situation? A: I think that it is better that those questions are coupled to the multiple themes for which they are relevant for a complete assessment for each theme. The number of double questions will depend on how strictly these themes are defined of course. It may not be ideal but will most likely be workable. And then for a specific audit if there is discussion on the inclusion of one or two questions within a theme, this can then be solved through tailoring these last points to the specifics of the audit based on professional insights.

Excerpt Interview 2

Reported answers are summaries of the discussion had with the interviewee that followed each of the questions. The interview excerpt has been validated by the interviewee prior to publication.

Q: What is your professional experience in IT assurance? A: Within assurance I do both advisory and audit-related work, particularly related to AI. For example, assessing how AI is implemented, where we look from governance to implementation and the conditions for responsible deployment. I also develop AI for audits as a managed service to automate processes or link certain information to regulatory standards.

Q: What is your background in AI? A: Before I worked on AI here, I had already had experience working with AI from my education background in philosophy and econometrics.

Q: How many AI assurance jobs have you done? A: I do not know exactly. These jobs are a recent development, and the number is growing – especially since the recent rise of generative AI it has become more prevalent. Before that it did not take up 100% of my time but now it is taking up more and more.

Q: What motivates clients to have you audit their AI system? A: This varies of course. For clients from industry it is often about managing their reputation. From the public sector this is also the case, but they have other values at stake. When they are checked by journalists or regulatory bodies and those discover that something is wrong,

for example when it comes to an algorithm that discriminates unethically, then they are basically done for – that is something that they will wish to prevent. But also from industry, for example the banking sector, we see that this topic has gained a lot more attention. There are already regulations which apply to AI, albeit not specifically, such as the GDPR, so they already have to comply with that. Sometimes it is also in anticipation of upcoming laws and legislation like the AI Act.

Q: How would you define trustworthy AI? A: There are of course many factors which make it trustworthy or not. Trust to me is related to not having to know everything and despite that still having faith in the AI system's functioning and the way it was built. This trust needs to be built through the way we interact with the AI, as well as how it is depicted in the news and checked by journalists. This trust is of course easily lost. To ensure the building of trust then, it is essential that an AI system is tested for alignment with laws and regulations by a third party. Trust in this case is a result of testing, and in the case of AI systems that change their behavior due to feedback on previous outputs or a change in training data, this will require periodic testing.

Q: Do you think the audit of AI can contribute to trustworthy AI, and why? A: As just mentioned, the audit of AI is important in gaining trust. Organizations do not inherently have a direct incentive to develop trustworthy AI. Of course, they do want to protect their reputation, but their upfront incentive is to have an effective AI system, whether its goal is fraud detection or personalized advertisements. In pursuing their goals certain risks arise, and the incentive to also mitigate these risks is just not as great in these organizations. That is why, in a market, you need independent parties who test these AI systems.

Q: What role do you expect the audit of AI to play in the future? A: How big its role will become is hard to say. Many AI systems are provided by large vendors such as openAI, and in the case of pretrained generative AI the EU plans to hold the vendors accountable for the AI systems they distribute. Unless they are able to lobby against this, this is one of the areas that the audit will then be focused on. The buyers and users of these AI systems will still have to be controlled in some way, as they can adapt and retrain models. We will have to see; I do think it will play an important role. The question is also if it should be limited to the technological aspects or be more use-case specific. I think the latter. The AI act is a nice start for that, but although I believe it will become an important process, how big it will become is hard to predict. If there are a couple of huge scandals in the news, it will certainly lead to a more important role for the audit as well. There are already platforms that allow the monitoring of your AI system's performance and technical specifications – so these may well cover the technical aspects of the AI system in the audit.

Q: How do you with the client derive the themes that fall within the audit scope? A: With every client we assess the most impactful areas, and in terms of AI this is essentially human rights. Based on that risk-based insight, which differs per AI system, we determine the scope. We also do benchmarking of AI systems to industry standards. Right now we barely look at any code, as the field is not mature enough to demand that level of specificity in our research.

Q: Do you encounter difficulties in this process? A: Not really, this works fine for us.

Q: Could you reflect on the utility of the translated ESG materiality assessment framework as presented to you for the scoping of an AI audit? A: When evaluating for completeness this also should include the quality of the materiality assessment by the client. So it is key for the auditor to first gain an understanding of the position of the client and its AI system in the AI landscape – including its purpose and the sector in which it will be operational. Only then can the auditor assess whether a materiality assessment is complete. The auditor should probably execute the steps on the right anyway, concerning gathering perspectives and determining the materiality of specific themes from various stakeholders, as the auditor will be much more knowledgeable on this than the client. This is something the auditor will be experienced in, and as such this will lead to a better materiality assessment than the one proposed by the client. Determining the purpose of the audit will also have to be done anyway, as it follows from the auditor gaining an understanding of the AI system and the client.

Q: How do you currently translate the determined audit scope to a concrete set of questions to be covered in the audit? A: We typically work with a risk management framework in which we include industry standards and elements from other frameworks. So, starting from existing frameworks we aggregate elements which we deem relevant based on the set scope.

Q: What is the role of the client in this process? A: In principle we do this ourselves. The client may have done a self-assessment separately, which we will then check for completeness.

Q: Do you encounter any problems or hardships when aggregating the list of questions? A: It is important to not rely too much on the set list, it should be an assisting tool in determining what risks may be present and what controls should be in place to mitigate those. The auditor should understand the specific use-case and this list of questions is a tool that is useful to indicate to the auditor which information should be gathered. But there will still be room for interpretation given that there are no set laws and regulations. So that should always be kept in mind when aggregating the list of questions.

Q: Imagining the general framework from the materiality assessment to pick questions from, would this solve some of the issues in this step of the audit? A: I think those standards do exist, but there will always be some need for interpretation from the auditor – for example when it comes to explainability and bias. As a dataset is biased per definition, it will need to be up to the expertise of the auditor to assess what level of bias is tolerable. One cannot simply say that it is never permitted to include personal data when training a model, as in some cases the inclusion of personal data allows us to pinpoint the bias. When not including the explicit data for ethnicity, simply leaving it out will not mean that the bias is removed as it will be present in other data categories through cross-correlation. So explicitly leaving it in can afterwards allow for a clear correction of the bias in the output, for example.

Q: What would then be the requirements for this general framework for it to

be actionable? A: As mentioned before the questions in the framework should leave room for interpretation by the auditor. Additionally, probably some level of jurisprudence, exemplary cases and cross-industry setting of standards will be beneficial for the auditor to better interpret those questions and translate them to the case they have at hand.

Q: In the general framework, some questions might appear multiple times as they are relevant for multiple themes within the framework. Can you reflect on the strengths or weaknesses of this situation? A: It does not have to be an issue, as long as the framework is consistent. I do not think that the list of questions even needs to be watertight – this is probably impossible. Rather than being a true list of concrete questions, it will likely look like a list of dimensions within the categories on which the questions will address general risks and controls. For example in the case of governance, the questions could indicate that governance should be set up correctly on the level of the organization, the team and the individual. The auditor can then assess how this applies to their current case, and consequently what evidence should be provided by the client. So the list ideally exists on a high enough level that allows for the auditor to fill in case-specific levels of detail.

Excerpt Interview 3

Reported answers are summaries of the discussion had with the interviewee that followed each of the questions. The interview excerpt has been validated by the interviewee prior to publication.

Q: What is your professional experience in IT assurance? A: I have worked at my current employer for two years now, where I am predominantly working on Responsible AI. This means advising clients on how to responsibly operate their AI system, through both a controlling and an advisory role. As there are no established laws and regulations, we often benchmark a client's AI system against frameworks we develop as part of an advisory job.

Q: What is your background in AI? A: My background is in econometrics, from which I am familiar with AI.

Q: How many of these AI assurance jobs have you done? A: I have previously worked on three of these jobs. At the moment I am doing one that is more focused on control. Besides those there have also been a couple of advisory jobs. These are more about setting guidelines for data scientists, so that they know what steps to take for responsible AI use.

Q: What motivates clients to have you audit their AI systems? A: Our earliest clients were really ahead of the curve in their conscious development of AI. They wanted to show to their stakeholders that the AI that they use is being developed and used responsibly. I also think that that is why they chose us to control this, as we are one of the Big Four – a big name adds value. Additionally, we have a team that has demonstrated to be very knowledgeable on the topic.

Q: How would you define trustworthy AI? A: That really comes down to the principles that have been developed by the High Level Expert Group. But in the first place

trustworthy means that the system is transparent. This will depend on the type of AI model, but the end user should be aware of the fact that they are dealing with an AI system and should understand in general terms how the model works and the underlying data. That way there is some clarity about which variables an output is based on. This very specific, but in general a model should not discriminate or be biased.

Q: Do you think the audit of AI can contribute to trustworthy AI, and why? A: Well in a sense an audit will be after the fact, so it should not be necessary if everybody does their job the way they are supposed to. But the reality is that if one knows that their AI system may be audited in the future, they will ensure better documentation and better consideration of choices to be made. Data scientists usually do not want to focus on documentation but rather keep programming. All the while it is important that decisions are tracked, especially when a data scientists leaves the organization or something goes wrong with the model. Then an organization can also learn from past mistakes. In short, if audits become practice, this leads to people improving their documentation and decision making. This documentation does not have to just be a Word document, there are many tools available that can be integrated in the AI and automate some of the documentation. For example, through dashboards that track performance or data quality – an alert can be set when input data starts to deviate from the data the model was tested on, which could lead to less reliable output. Those types of things can be automated.

Q: What role do you expect the audit of AI to play in the future? A: I think that many organizations will feel the need for an external party to check their work, especially with the AI Act coming up. Specifically in the public sector I have noticed that the government is aware of this and interested in having an external party check their work. Given recent developments and news reports that have not always been as positive it should not be a surprise that the government is more aware of these issues. So I expect the public sector to be the first to mature into this field, since they are more used to being held accountable and documenting their work as they are subject of society's scrutiny. Other organizations are still only busy with reporting to stakeholders within their organization and their own interests, and not so much with other stakeholders such as end-users or society as a whole.

Q: How do you with the client derive the themes that fall within the audit scope? A: In principle this is mainly done by us. We base the scope on good practices, these are a couple of frameworks such as the NIST AI RMF, the AI Act and others depending on the case. We then evaluate all the aspects that are addressed in the frameworks and use them to assess our client.

Q: Do you encounter difficulties in this process? A: There are some challenges which stem from having to combine multiple frameworks. As they differ, it takes time to establish consistency in the final framework while also taking into account client-specific risk areas.

Q: Could you reflect on the utility of the translated ESG materiality assessment framework as presented to you for the scoping of an AI audit? A: The themes as determined by the EC HLEG are a good starting point, as all of them are relevant for trustworthy AI and you probably want to cover them all as it is hard to see

them separately. One thing we currently face is that based on where a client is in the life cycle of their AI system, not everything related to trustworthy AI is immediately relevant. Different issues are relevant in the development phase versus when the system is active. So when scoping the themes this will have to be taken into account.

Q: How do you currently translate the determined audit scope to a concrete set of questions to be covered? A: This process is not that structured. We take the frameworks and the elements that they consist of, and just shove them together. Sometimes we first have a list of questions and following that we establish themes to group them under. I can imagine this changing in the future, when we are more aware of the various themes that can fall within the scope, this process becomes more streamlined and takes the themes as a starting point.

Q: What is the role of the client in this process? A: We do this ourselves basically.

Q: Do you encounter any problems or hardships when aggregating the list of questions? A: As previously mentioned, tailoring the list to the specific situation of the client is sometimes difficult. The challenge also lies in finding the right frameworks, as we always want to base ourselves on sources. Sometimes the scientific literature is quite ahead of the curve, so it useful to incorporate academic frameworks.

Q: Imagining the general framework from the materiality assessment to pick questions from, would this solve some of the issues in this step of the audit? A: Yes, I do believe this would be useful. In the end, it will be a general framework and not focused on a specific case so there will always be some challenges in that regard.

Q: What would then be the requirements for this general framework for it to be actionable? A: The questions will need to be sufficiently high-level that they can be translated to any specific case. This should then be done by the auditor through their expertise. It might be nice to also have some cases or examples to make it more clear how the questions can be translated to a specific case.

Q: In the general framework, some questions might appear multiple times as they are relevant for multiple themes within the framework. Can you reflect on the strengths or weaknesses of this situation? A: I do believe it is avoidable, we can remove any overlap by assigning questions to the category they best fit in.

Informed Consent Form

As agreed upon in the Data Management Plan, the template informed consent form is included in the appendix, shown in Figure D.3.

Participant Information/Opening Statement

You are being invited to participate in a MSc thesis research study titled Auditing Artificial Intelligence. This study is being done by Tijn Sewandono from the TU Delft.

The purpose of this research study is to develop a holistic framework to audit artificial intelligence, and will take you approximately 20 minutes to complete. The data will be used to support the findings of the thesis. We will be asking you to validate the presented results concerning auditing frameworks and artificial intelligence, and pinpoint missing elements based on your expertise.

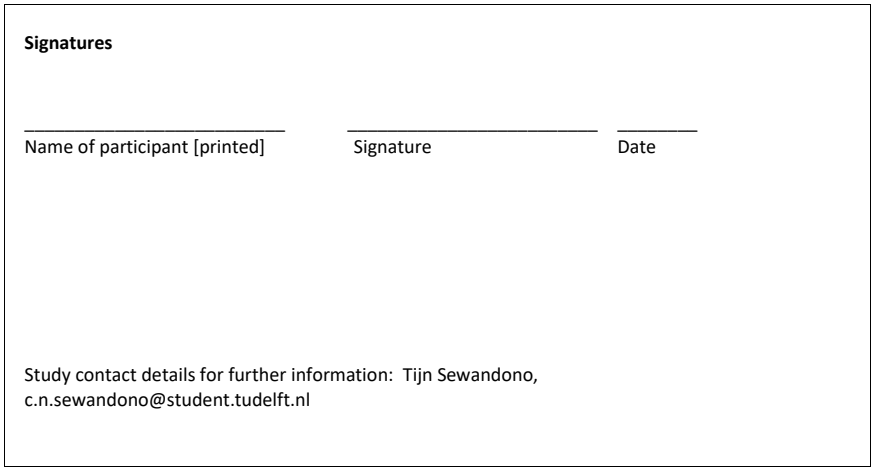
As with any (online) activity the risk of a breach is always possible. To the best of our ability your answers in this study will remain confidential. Audio recordings will be destroyed immediately following transcription. Transcripts are stored on TU Delft institutional storage that is only accessible to Tijn Sewandono and his graduation committee (Sander Renes and Aaron Ding, both TUDelft TPM staff). The thesis will be publicly accessible but raw interview data (i.e. transcripts) will not be published. Any data included will be fully anonymized.

Your participation in this study is entirely voluntary and you can withdraw at any time. You are free to refuse to answer any question, and reach out afterwards to rectify information.

Corresponding and responsible researcher can be contacted through c.n.sewandono@student.tudelft.nl and s.renes-1@tudelft.nl respectively.

Explicit Consent points

PLEASE TICK THE APPROPRIATE BOXES	Yes	No
A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICPANT TASKS AND VOLUNTARY PARTICIPATION		
1. I have read and understood the study information above, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	☐	☐
3. I understand that taking part in the study involves: an audio-recorded interview, which will be transcribed as text after which the recording will be destroyed. An anonymous summary of the interview will be produced and included in the MSc thesis.	☐	☐
4. I understand that the study will end at the time of defense of the thesis, anticipated to be late September 2023.	☐	☐
B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)		
5. I understand that taking part in the study involves the following risks: not feeling comfortable to answer interview questions for any reason whatsoever. I understand that these will be mitigated by having complete freedom to refuse any interview question or abandon the interview altogether at any time.	☐	☐
6. I understand that taking part in the study also involves collecting specific personally identifiable information (PII) – profession – and associated personally identifiable research data (PIRD) – professional views – with the potential risk of my identity being revealed leading to damage to my professional reputation.	☐	☐
8. I understand that the following steps will be taken to minimise the threat of a data breach, and protect my identity in the event of such a breach: audio files are deleted following transcription; all data is stored on an institutional storage with restriced access for only the researcher and their graduation committee.	☐	☐
9. I understand that personal information collected about me that can identify me, such as my name or employer, will not be shared beyond the study team.	☐	☐
10. I understand that the (identifiable) personal data I provide will be destroyed at the latest one month after the end of the project.	☐	☐
C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION		
12. I agree that my responses, views or other input can be quoted anonymously in research outputs.	☐	☐
D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE		
13. I give permission for the anonymous interview summary to be archived in a public repository so it can be reused for future scientific work. I understand the summary will be sent to be for review before publication.	☐	☐



Signatures

_____ _____ _____
Name of participant [printed] Signature Date

Study contact details for further information: Tijn Sewandono,
c.n.sewandono@student.tudelft.nl

Back to text

Figure D.3: *Template Informed Consent Form for Auditors.*
Three-page document signed by all interviewees whose interviews are included in this thesis.

D.2 Stakeholder Inquiry

As the stakeholder inquiry is part of the case study, the full interview protocol is included in that part of the Appendix, subsection E.2.

Informed Consent Form

Similar to the interviews with auditors, the stakeholders on the streets in Amsterdam who were asked to participate were presented with an informed consent form. This form is depicted in Figure D.4.

Do you consent to answering a couple of questions about artificial intelligence? It will only take about five minutes. Your answers will be used to illustrate the public impression regarding an AI system for my master thesis performed at Delft University of Technology. I will not ask for or store any personal information. This short interview will be recorded; this recording will be deleted immediately upon transcription later today. You can voluntarily withdraw at any time.

Your answer will be aggregated with the input I obtain from other people in the result section of my thesis.

If you want more information on this project, you can contact me at the following email address: c.n.sewandono@student.tudelft.nl. The thesis will be made publicly available once I am done, and you will be able to find it in the Delft University of Technology thesis repository.

Can I quote you? (yes/no)

Figure D.4: *Informed Consent Form for Stakeholders.*
Was printed and provided to anybody who participated and wished to do keep it.

E AI Audit Case Study

E.1 Client Understanding

Client Goal and Collaborations

Relying on the sources of Table 5.1, an understanding of the client, their goal with the AI system, and relevant collaborators could be derived. The primary product owner is the Municipality of Amsterdam, in which the Chief Technology Office (CTO) Innovation Team spearheads innovative solutions to problems in the city. One of these solutions is Public Eye. Public Eye falls under the CMSA initiative, which is an open-access monitoring system for the busyness at various hotspots within the city of Amsterdam. Through the CMSA dashboard, anybody can access live figures of business at these locations. The Marineterrein site is the only one at which busyness figures as determined by Public Eye are fully publicly accessible.

The problem for which Public Eye was developed is that originally crowd monitoring on behalf of the municipality was done manually, by a person continuously checking live camera footage. Crowd monitoring was considered essential to avert safety hazards caused by overcrowding through proactive crowd control measures. The manual monitoring needed to be replaced as it was deemed invasive, inefficient and unquantifiable. Currently, the goal of crowd monitoring also includes making this data directly available to the public, such that visitors and inhabitants can take live busyness figures into consideration when moving around through the city of Amsterdam. This was evermore relevant during the COVID-19 pandemic, when people were advised to avoid crowded areas and maintain 1.5 meters of distance from one another.

To solve the issues with manual monitoring, the Public Eye AI system was developed by the CTO Innovation Team in collaboration with two external partners: Tapp and Life-Electronic. Life-Electronic is mentioned as lead developer of the project, while the exact role of Tapp is less clear in the found sources. However, since Tapp is also mentioned to be co-maintaining the Public Eye Github repository, they are considered to be of significant relevance in the development and maintenance of the Public Eye AI system. The Marineterrein is a central location in Amsterdam which is open to the public for leisure (sports, hotel and catering). It also functions as adaptive and innovative hub where experiments in the field of learning, working and living are conducted by businesses, universities and other institutions in collaboration with the Municipality of Amsterdam.

As such, it is evident that the Public Eye algorithm is operational in the public sector: serving a public goal in a public place for the Municipality of Amsterdam.

Public Eye AI System

The identified sources provided enough information to characterize the AI system and reconstruct its workflow. The Public Eye AI system is based on a pre-trained Vision Crowd Counting Transformer, a Deep Learning Computer Vision algorithm that is able to classify and count the number of heads in a provided image. The novelty of the Public Eye algorithm lies in its ability to assign importance to areas within images, helping better

distinguish people from backgrounds.

The model was pretrained on the ShanghaiTech Crowd Counting dataset, which is publicly available large-scale dataset of 1200 annotated crowd images. Public Eye was then further trained on a set of hundreds of annotated camera images for each of the areas in which it is operational. The reported accuracy of the model lies around 90%, which satisfies the predetermined minimal accuracy of 70% required to obtain relevant busyness insights. The model is not additionally trained or fine-tuned while operational, meaning that the model has remained exactly the same while deployed. The developers also mention that the system is in compliance with the GDPR and Tada, a set of principles for the responsible use of data and technology developed by the Municipality of Amsterdam.

Connected surveillance cameras send an image to a central server through end-to-end encrypted transmission. There the Public Eye model will process them and send the number of heads determined from the image to a publicly accessible dashboard as well as a dashboard used by municipal crowd managers. Upon analysis by Public Eye the images are deleted from the server. Furthermore, the CTO Innovation Team periodically assesses whether the accuracy of the model is not deteriorating.

At the Marineterrein site, the only site from which Public Eye data is publicly accessible, four cameras monitor the busyness. As the data from the four cameras is reported on a fully functional and user-friendly dashboard, and all model documentation is available open-source, it was concluded that the Public Eye AI system is fully mature.

E.2 Stakeholder Inquiry

Interview Questions

To determine which of the two stakeholder groups they fall into:

- Do you frequent this location?

To introduce the topic of crowd monitoring and activate the interviewee:

[Interviewer can point to one of the cameras.]

- Are you aware of the municipal cameras monitoring this area?

To introduce the topic of trustworthy AI:

The Municipality of Amsterdam actually uses an algorithm to automate counting the number of people in this area based on the camera footage. You can access this data yourself, one of the goals of the algorithm is to provide the public with live busyness figures.

- Knowing this, would you say you place a level of trust in this algorithm?

To introduce the topic of assurance and allow the interviewee to freely formulate themes that are important to them:

- Say there were a knowledgeable independent party, that was in a position to expertly and thoroughly assess the algorithm as well as the developers, what would you like them to check up on in order for your trust in the algorithm to improve?

To introduce the themes from the general framework and obtain materiality scores:

- Furthermore, could you rank how the assessment of any of the following themes would further improve your trust in the algorithm?

  [Hand the participant flash card with the following seven themes: Human oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability.]

Materiality Results

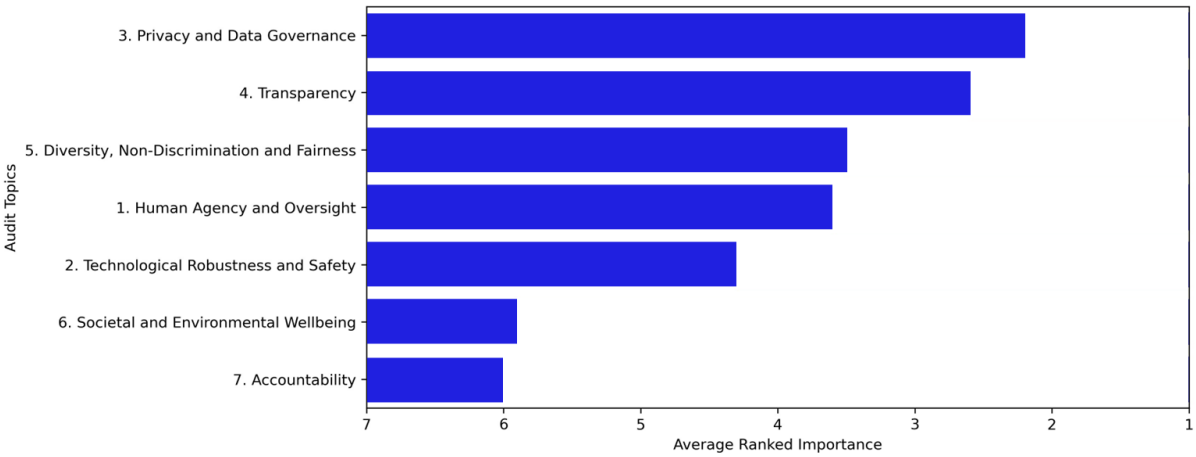


Figure E.1: *Consistent Stakeholder Materiality Ranking of Audit Topics. Only stakeholder input from those consistent with their interview responses is included. Topics shown from highest average importance to lowest.*

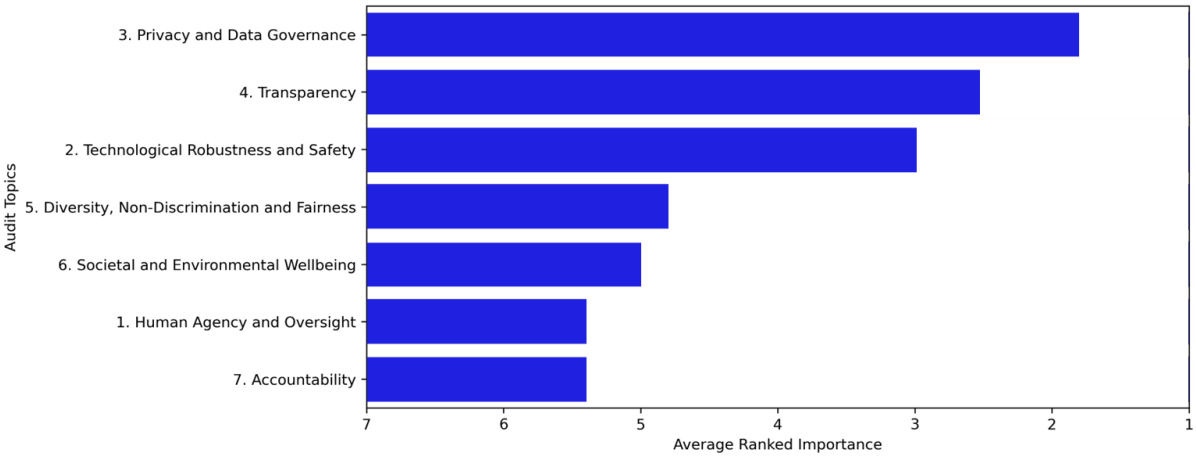


Figure E.2: *Frequent Visitor Stakeholder Materiality Ranking of Audit Topics. Stakeholder input from those who are frequently near the Marineterrein. Topics shown from highest average importance to lowest.*

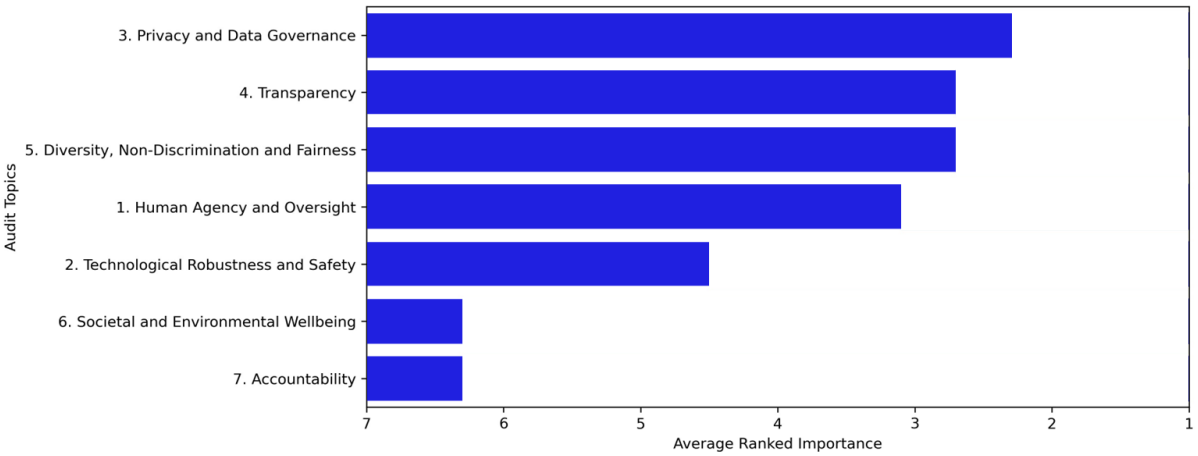


Figure E.3: *Occasional Visitor Stakeholder Materiality Ranking of Audit Topics. Stakeholder input from those who are occasionally near the Marineterrein. Topics shown from highest average importance to lowest.*

E.3 Final Audit Questions

Table E.1: *Public Eye Audit Questions*

Privacy and Data Governance
Has data been handled lawfully throughout the development and use of Public Eye?
Has the legality of data used been documented?
Has the GDPR been adhered to throughout the development and use of Public Eye?
Are data management policies in place and documented?
Do data management policies cover the complete lifecycle of Public Eye?
Are data management policies in line with rules and regulations, such as the GDPR?
Has the used data been assessed for quality, such as its representativeness, alignment with goals, consistency across datasets, and correct annotations?
Has a DPIA been carried out and documented?
Have DPIA findings been addressed?
Has the role of data protection officer (DPO) been assigned?
Can the DPO be contacted by any stakeholder to raise data and privacy related issues?
Does the Municipality of Amsterdam have complete ownership of the data used for Public Eye?
Has data minimization been applied?
Have data choices been substantiated and documented?
Does any of the used data include personal data?
Are policies in place to minimize and de-identify the data?
Are those who own the personal data informed that their data is used?
Has the need to use personal data been assessed?
Are data processing procedures documented and publicly accessible?
Are data storage procedures in place and documented?
Is the data storage policy compliant with GDPR regulations?
Does the organization communicate a guarantee of privacy protection to stakeholders?
Transparency
Is the model underlying Public Eye publicly accessible?
Is the data used to develop the model publicly accessible?
Have model development practices been published?
Is a complete description of Public Eye publicly available?
Is there a guide describing the logical flow of Public Eye?
Are stakeholders informed about the goal of Public Eye?
Are stakeholder informed about the logic behind Public Eye?
Are stakeholders made aware of the fact that they interact with an AI system?
Are stakeholders informed about the limitations of Public Eye?
Are stakeholder informed about the performance levels of Public Eye?
Are stakeholders informed about organizational guiding values?
Is Public Eye output clearly presented to all stakeholders?
Are stakeholders informed about what data is used by Public Eye?
Are incidents and errors communicated to all stakeholders?
Can stakeholders report any (perceived) issue with Public Eye?
Can stakeholders objected to being subjected to Public Eye?
Are stakeholders informed about the data security of Public Eye?
Are data choices explainable in the context of the intended purpose of Public Eye?
Is the output explainability sufficient for the intended purpose of the Public Eye?
Have assumptions and design choices been substantiated and documented?
Have value trade-offs made during the development of Public Eye been documented?
Diversity, Non-Discrimination and Fairness
What mechanisms are in place to prevent undesirable bias in Public Eye?
Has the fairness of Public Eye been defined, e.g. through objectives/principles/standards/policy?
Were other definitions of fairness considered?

Continued on next page

Table E.1 – *Continued from previous page*

Is the definition of fairness in line with laws and regulations?
Is the Public Eye periodically reviewed for alignment with fairness as defined?
Has Public Eye been assessed for bias during development and deployment?
Have roles and responsibilities been assigned regarding fairness of Public Eye?
Have all stakeholders been involved in the assessment of the potential for biases in Public Eye?
Were third-parties consulted for assurance on ethical concerns at any stage?
Can perceived biases be reported by stakeholders?
Has the input data been assessed for biases?
Have system development choices been assessed for biases (e.g. type of model used)?
What value trade-offs have been made between values during model development?
Have the value trade-offs been documented?
Have assumptions and design choices during system development been substantiated and documented?
Have metrics been defined to quantify bias in the AI system?
Have input manipulation tests been performed to assess system bias for known subpopulations in the data?
What bias rectification policies are in place?
Does Public Eye display undesirable bias?
Has the training data been evaluated for undesirable biases?
Are all stakeholders informed about the capabilities and limitations of Public Eye?
Are all stakeholders informed about data management practices?
Is policy in place to inform all stakeholders in case of incidents?
Is training offered to those within the organization whose work is related to Public Eye, covering ethical and technical aspects?
How is stakeholder feedback incorporated in Public Eye?
Has a set of organizational guiding values been defined and documented?
Are the guiding values complete in covering the characteristics of trustworthy AI?
Do the specified fairness metrics reflect these values?
Have the guiding values been communicated to all stakeholders?
Have stakeholders been involved for input in the complete lifecycle of Public Eye?
Have stakeholders been involved in the development of Public Eye?
How are stakeholders included in periodic review and improvements of Public Eye?
How are stakeholders included in the implementation of Public Eye?