AI auditing:

Incorporating eXplainable AI in an auditing framework

Information System Science
Master's thesis

Author(s):
Anne Bran van Wingerden

Supervisor(s):
Dr. Hans Weigand – Tilburg University
Jordi Kerckhaert - KPMG

06.06.2024
Aix-en-Provence, Turku, Tilburg

Master's thesis

Subject: AI auditing
Author(s): Anne Bran van Wingerden
Title: AI auditing: incorporating eXplainable AI in an auditing framework.
Supervisor(s): Dr. Hans Weigand & Jordi Kerckhaert
Number of pages: 72 pages + appendices, 108 pages
Date: 06.06.2024

Abstract

The advancement of Artificial Intelligence (AI) technologies increased significantly in the last few years. Moreover, the application of AI models expanded to a broader range. Hence, auditors are progressively encountering AI like systems, models and algorithms during audit and assurance projects. The growing scientific domains of eXplainable AI (XAI) and Responsible AI raise concerns around the transparency, explainability, and other ethicalities. These concerns, in combination with upcoming legislation, demand audit statements on reliability, integrity, and other aspects of AI models. Where auditing is a formal well-established practice, AI auditing is a novel practice. This research includes literature research, exploration of AI audit cases, and interviews with AI experts in order to discover relevant methods and specificalities of AI audits. Through the methodology of design science, a first formalised AI Audit Process is developed and proposed in order to provide AI auditors with a flexible reference frame to conduct customised AI audits. This research is a step towards the advancement of an AI auditing method and offers valuable insights for science and practice.

Key words: AI, eXplainable AI, auditing, algorithm assurance, transparency, explainability, AI assurance, AI auditing

Acknowledgements

In the past two years I have studied in three different countries and this thesis internship at KPMG marks the end of the International Master in Management of IT (IMMIT), a study that has become very dear to me. The period at KPMG allowed me to learn a lot about artificial intelligence, auditing, assurance, and my own professional career. I am very proud to present my findings, but this thesis would not have been the same without the support and advice of several people.

A lot of KPMG employees were very open to meet me to discuss the topic of AI auditing and think along with my research. I want to thank the people that made the time and effort to provide me with feedback and allowed me to interview them. I want to specifically thank Jordi Kerckhaert who gave me the freedom to conduct a topic that was relatively new to the ITA department. Thank you for connecting me with the right people and providing very valuable feedback and insights to enhance the quality of my research.

I would like to thank Dr. Hans Weigand from Tilburg University who supervised this thesis project. Thanks for your excellent guidance and feedback. More importantly, thank you for the discussions and meetings we have had about AI auditing.

I would like to thank my friends and family who encouraged me to join the IMMIT program and supported me a lot during this research period. Lastly, I would like to thank the students of IMMIT cohort 16 who have become more than friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

A decade ago, experts foresaw major changes for traditional auditing. (Lombardi et al., 2014). An increasing demand to suffice information in a real-time manner was the driver. This was an interlude to auditing 3.0 in which Big Data and Data Analytics was included in auditing. (Alles et al., 2021). Another decade later the auditing field finds itself on the eve of auditing 4.0 in which Artificial Intelligence (AI) technologies are the object of focus during audits or even handle complete auditing activities. The predicted change to the domain of auditing seems imminent now. However, auditing 3.0 is not retired, rather it has become a step on a maturity ladder.

In November 2022, OpenAI released their generative AI, ChatGPT, which caused major changes and awareness around AI. Organisations, like the Big 4 Accounting & Consulting firms, had to update their policies and start innovating by researching opportunities to competitively use AI-systems. (Almufadda & Almezeini, 2022; PwC, 2023). The Future of Audit phrased it eloquently:

> *"The audit profession needs to stay in tune with the constantly changing world to effectively meet the needs of the users of information*"
>
> (Lombardi et al., 2014)

This research investigates how AI models are currently practices and present the findings in a formalised artefact developed with the design science methodology. The first chapter provides insights into the 'problem' and gives reason to initiate the research. The second chapter is an extensive literature background to provide the reader with sufficient knowledge to understand the context of AI and auditing by firstly describe the topics individually and then describe the joint field called algorithm assurance. The third chapter is the methodology section and describes how this research is executed and how the Design Science framework is applied. The fourth chapter presents the results of the research and the analysed findings by transforming the latter into requirements to AI auditing. This chapter presents the preliminary model to which further development led to the fifth chapter that showcases the artefact as the end result of this research. The sixth chapter discusses the theoretical and empirical findings and discusses the limitations of this research and potential future research directions. Lastly, the seventh chapter presents the conclusions of the research.

1.1 Problem indication

Innovations led to small and major applications of AI, such as preparing audit data, organising files, integrating data from multiple files, and concluding basic audit tests in Excel. (Zemankova, 2019). However, there are bigger applications which are, in a theoretical sense, already operational in the Big 4 accounting firms. Deloitte utilizes their own developed AI, GRAPA, to compare risk strategies, and KPMG has piloted an AI to evaluate inventory (controls) in combination with drones. (Almufadda & Almezeini, 2022). It concludes that the trade-off triangle of time, costs, and quality can become more benefical through the utilization of AI. (Almufadda & Almezeini, 2022). A lot of clients of the Big 4 have started to use AI systems in their core business operations as well and might even have IT application controls that are covered by complex automated algorithms. The working field of auditors will change significantly. Auditors need to start controlling, or auditing, whether the output of clients' AI systems is trustworthy. Furthermore, due to the complexity and required workload auditors need to consider what is going to be audited; the AI system, the coding, the input and/ or the output, the processes? AI audits might require more thorough knowledge or are IT auditors capable to address the new technology within the audits?

Despite proven benefits, three in five people are wary to trust AI systems according to a study amongst KPMG employees in the US, Canada, the UK, and Germany. (Gillespie et al., 2023). The study describes that it is challenging to create trust and acceptance amongst stakeholders. Particularly, AI systems tend to have a black-box nature which results in an unfavourable view upon the use of these systems in financial reporting. Managers have concerns around the quality of the output of AI, and ethical concerns such as fairness. (Estep et al., 2023). Providing assurance through AI audits seem to become pivotal in the near future.

1.2 Problem statement

The deployment of AI systems is raising critical concerns regarding transparency and explainability of these systems. There is a significant lack of documented real-world examples on either successful audits of AI systems or knowledge about faced challenges during those audits. Empirical research on explainability and transparency of AI systems encountered during audits is not overly present or shared cross-industrially. Unlike IT auditing there are no standardized best practices or formal ways of working. Drilling down into practice through the lenses from a management-, investor-, auditing-, or AI research level is important to gain comprehensive insight into the current practice. (Minkkinen et al., 2022;Berente et al., 2021).

Whether AI is to audit or the (output of) AI model of a client is audited, both realms are looking to be in-control of AI. This directs to AI governance, where organisational control mechanisms are being set up and researched. Through the lens of AI auditing, AI governance serves as an overarching mechanism that provides approaches and best-practices. In combination with the auditing domain there are still a lot of knowledge gaps, defined as: *"Uncertain effectiveness of ethical principles and regulations"* and *"Modest understanding of AI system design implications of transparency and explainability"*(Birkstedt et al., 2023). This research will explore these gaps by investigating the current practices of AI auditing in an attempt to carefully formulate an artefact that represents a formalised approach to this concept of auditing AI models.

1.3 Research questions

The main RQ:

How can AI auditors assure transparency and explainability in the practice of auditing AI systems?

To support the main research question above this research divided the RQ into several sub-questions:

- How do current AI auditing practices address the principles of transparency and explainability?
- What is the difference between the two XAI facets 'transparency' and 'explainability'?
- What challenges and limitations are currently faced in AI auditing for assessing transparency and explainability?
- How are AI auditing practices implemented in audit processes and what are their outcomes?
- What criteria should be used to evaluate whether an AI algorithm or system is considered responsible in the context of auditing?

The questions are answered through literature research and the applied empirical research method. This research uses a lot of specific terminology and abbreviations. In order to keep track a glossary is attached in appendix 1.

2 Background literature

The literature is scoped into 4 topics; the meaning of artificial intelligence, explainability and transparency of artificial intelligence, auditing, and the combination of auditing and artificial intelligence. Readers that are already familiar with artificial intelligence and auditing might opt to only read paragraph 2.2 Explainable AI and paragraph 2.4 AI auditing. Paragraph 2.1 provides a scoped understanding of the concept artificial intelligence and paragraph 2.3 is the equivalent for auditing.

The literature is structured in this order to provide understanding of the technology and build up to the auditing concept. eXplainable AI is a known research domain within the scientific research field of artificial intelligence and is included to provide a perspective on what is already done to investigate audit topics as reliability in the context of artificial intelligence.

2.1 What is Artificial Intelligence?

This first paragraph of the literature is dedicated to the tenet Artificial Intelligence, further to be referred as 'AI', and explores its origins and meaning. The conceptual beginning of AI is described to move towards the adopted definition in this research.

2.1.1 Formulating a definition

In 1995 a famous workshop took place at Darthmouth College where scientists investigated John McCarthy's question whether machines could think. This workshop led to the birth of the term and idea Artificial Intelligence and its yet to be discovered potential. (Veisdal, 2019). The main outcome of this workshop was an entire scientific field dedicated to McCarthy's question and followed the first initial definition of artificial intelligence:

> *"Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it".*
>
> (McCarthy et al., 1955).

This definition was not found through a scientific method, but rather the shared vision of the workshop attendees which was *"Computers can be made to perform intelligent tasks"* (Moor, 2006). However, it could be argued that according to the criteria, outlined by McCarthy et al, (1955), the mobile devices in our pockets can be considered AI since they are able to perform intelligent tasks. Alan Turing broadens the definition of AI through his famous ‘Turing Test’. Turing (1950) states that a machine can be considered intelligent if a machine is able to reproduce a human response well enough to fool a human judge. In the context of the Turing Test, a machine is an entity that can process information, execute calculations, follow instructions through programs, and could potentially pass the test. (Haenlein & Kaplan, 2019). A neuroscientist argued that these definitions should be broader than only the technology and argued to include various fields, such as neuroscience, philosophy and psychology besides computer science and mathematics. (Marr, 1977). Berente et al. (2021) outlines that AI research has moved to be not only a technical problem, but also a social problem. Showcasing the fact that AI research has become cognitive, questioning intelligence itself.

2.1.1.1 Intelligence

Discussing AI means to address intelligence. Hence this research cannot adopt a working definition without discussing intelligence. Consulting a dictionary, the definition of intelligence yields: *“The ability to learn, understand, and make judgement or have opinions that are based on reason”.* (Cambridge Dictionary, 2024). However, amongst intelligence scientist there appears to be no agreement on the exact definition of intelligence. (Tegmark, 2018, p. 49). For example, Breakspear (2013) wrote an issue that articulates a new definition of intelligence. Stating that intelligence is a capability to forecast change and do something about it, in time. More researchers, from different fields, have given their perspective upon the meaning of intelligence; *“the mental abilities necessary for adaptation to, as well as shaping and selection of, any environmental context*”. (Sternberg, 1997). All definitions share the same fundaments. Learning, reasoning, and problem solving are foundational concepts that formulate (Human) intelligence. (Colom et al., 2010). In concord terms, machines, software or algorithms undertaking complex tasks, associated with cognitive function, can be considered intelligent.

2.1.1.2 Types of intelligence

Non-biological intelligence, or artificial rather, is often perceived as a massive intangible superhuman concept in science fiction, like The Matrix, Wall-E, or even The Terminator. These portrayals align with interchangeable academic terms like human-level AI, Strong AI, Conscious AI, Superintelligence or the official overarching term called Artificial General Intelligence (AGI). Tegmark (2018) distinguishes two types of intelligence to differ AI and AGI by defining intelligence as *"intelligence is the ability to accomplish complex goals"* . Narrow AI (1), is able to apply knowledge and skills to achieve a single goal. Narrow AI excels in specific tasks like image classification, game playing, and language processing. Determining the next optimal move in chess is considered a complex algorithm with a narrow goal. (Haenlein & Kaplan, 2019). Broad AI (2) utilizes general cognitive skills to achieve a range of goals, including interaction and reasoning. General intelligence, or universal intelligence, refers to the ability to achieve virtually any goal, including learning. AI goals are measurable and explainable effects, allowing humans to model them into functions. (Tegmark, 2018, p. 39).

Within intelligence and AI research the terms intelligence and rationality are broadly discussed. Burgoyne et al. (2021) states that rationality and intelligence are correlated, but admits a strict difference; "*intelligence is the ability to achieve goals across a range of environments"*. (Russell, 2016). Whereas rationality involves: *"making the best decision given the available information and constraints".* (Russell, 2016). Thus, rationality requires specific cognitive abilities from entities with general intelligence. (Burgoyne et al., 2021). Through this reasoning it can be concluded that developing AI to emulate human intelligence is limited since humans are not always rational. (Russell, 2016). AI systems should incorporate both intelligence and rationality to be aligned with real-world applications.

Although this research distinguishes between these concepts it is not likely that the research will encounter AGI, but rather systems that are rational, intelligent or both. As per 2024, current AI applications exhibit narrow or broad intelligence. For example, ChatGPT is at its core purely a large language model. (OpenAI, n.d.).

2.1.1.3 Presenting a working definition

This research refrains from comprehensively covering a definition of the concept AI, as defining AI is a complex endeavour. However, a general working definition will be adopted to clarify the scope and context of the research. This research considers AI as a non-biological intelligence entity to which intelligence is the capability to accomplish complex goals. Algorithms are intelligent if they execute complex tasks, intelligent machines are able to learn and adapt, and gaol-oriented behaviour can be modelled and developed. Hence the following definition is adopted:

> "*Artificial Intelligence (AI) is a general term that implies the use of a computer to model intelligent behaviour with minimal human intervention"*
> (Hamet & Tremblay, 2017)

The next paragraph presents the reasoning of keeping the formulated definition abstract and relatively broad.

2.1.2 The moving frontier or AI

It is difficult to capture a clear and scoped definition of AI. The concept where thinking machines or algorithms perform complex tasks is no longer considered 'intelligent' once humans understand the process of the transformation. (McCorduck, 2004). This paradox is dubbed as the 'AI effect'. Kaplan (2021) argues that the moment humans gain an idea of what happens inside a black-box the term AI disappears with it. To conclude, the AI effect is an opaque philosophical discussion illustrating that AI-systems cannot be considered intelligent if they can be understood.

Due to innovations and continuous advancement the AI technology pushes the boundaries of current AI capabilities, applications, and general performance. , Berente et al. (2021) illustrates the idea of AI being a dynamic frontier. Implying increasing performance and an expanding scope related to AI's characteristics of autonomy, learning, and inscrutability. As a result, this research adopted a broad and somewhat 'future proof definition in order to provide limitations to the scope of the research. The aim of this is that intelligent algorithms apply for the same auditing principles as inscrutable AI systems.

2.1.3 Technical definition

As indicated in the previous paragraphs, learning makes an AI intelligent. This paragraph introduces some technical theory by outlining some of the AI sub-fields as: Neural Networks, Machine Learning, Deep Learning, and Natural Language Processing. These techniques are often utilized in AI applications. (Almufadda & Almezeini, 2022). Figure 1 below shows the relation of the sub-fields to AI as a general term.

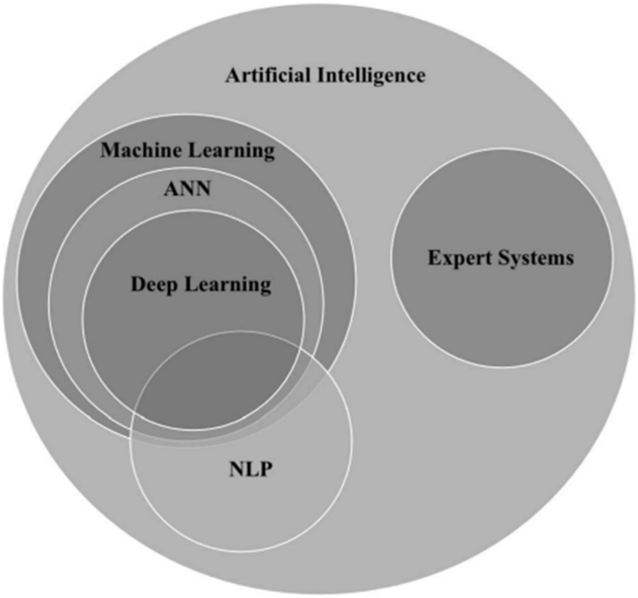


Figure 1. Positioning the AI sub-fields. (Almufadda & Almezeini, 2022)

Even though there are many more types than presented in figure 1, this theory provides sufficient technical understanding for this research. The technical perspectives from engineering, design, and development are considered as too detailed. Additional reading is recommended to fully understand the concepts.

2.1.3.1 Machine Learning

Machine Learning (ML) learns patterns from datasets without the need to define the data in advance or give explicit instructions and programming. (Lee & Shin, 2020). This pattern discovery is seen as learning. A ML model makes predictions about a certain phenomenon through trial-and-error in order to refine the model. The predictions are the output. (Baloglu et al., 2022). Figure 2 below showcases the iterative method that learns to predict the label of a datapoint based on tis features. (Jung, 2022). This research distinguishes three types of ML.

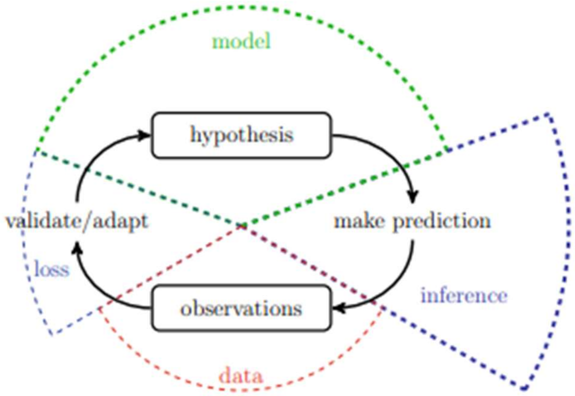


Figure 2. How Machine Learning works. (Jung, 2022).

Supervised ML uses labelled datasets to predict outcomes based on the provided datasets. (Jung, 2022). In contrast, Unsupervised ML finds patterns in data without labelled outputs and clusters the data to predict which datapoint belongs to which group. (Jung, 2022). Reinforcement Learning is not about finding the hidden patterns or structure, but maximizes rewards through interaction with its environment. It learns from rewards and punishments to achieve desired outcomes. (Pandey et al., 2023).

2.1.3.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are foundational to Deep Learning and serve as a machine learning algorithm, inspired by the human brains' neurons. (Almufadda & Almezeini, 2022). Initially aimed to replicate the brain, ANNs utilize interconnected processing elements (neurons) for computational tasks like forecasting. (Maier et al., 2023). ANNs are a good practice in recognizing faces and handwriting. (Almufadda & Almezeini, 2022).

2.1.3.3 Deep Learning

Deep Learning (DL) is a more advanced ML method to discover relations by combining the patterns of ANNs with computational capabilities form machines. (Almufadda & Almezeini, 2022). A DL is a neural network with a multitude of layers, each layer processes information with their own neurons. The number of layers determine the sophistication of the DL model and is ideal for analysing big data sets. (Almufadda & Almezeini, 2022; Jung, 2022).

2.1.3.4 Natural Language Processing

Natural Language Processing (NLP) entails various computational linguistic methods to analyse and represent plain and unformatted (natural) text. (Almufadda & Almezeini, 2022). Typically, every NLP follow the same pipeline of 5 steps. (Garousi et al., 2020). Furthermore, interaction between human and machine is characteristically to NLP. NLP models are inscrutable, or considered black-box, because of their lack of transparency on the predictions and decisions. (Liu et al., 2023). According to Tang & Kejriwal (2023), there is a possibility that language models find the right answer for the wrong reasons. The next paragraph shows how scientists address the complexity of models from which it is difficult to conclude they provide trustworthy outputs.

2.1.4 Responsible AI

The academic domain Responsible AI (RAI) emerged from several conferences about ethical issues in AI technologies. (Dennehy et al., 2021). Fjeld et al. (2020) from Harvard University describes about 8 themes among AI principles that would lead to RAI. Whereas Arrieta et al. (2020) sticks to ‘the’ 6 principles of RAI. Despite various perspectives, or rather areas of concern, there is no generalised framework or view. However, the core concepts are roughly similar. Principles as accountability, safety, privacy, and fairness could arguably be considered the same as the ‘ethics’ principle.

The EU commission’s guidelines stress the trustworthy AI principles: fairness, human autonomy, prevention of harm, and explicability (explainability) as fundamental human rights. (HLEG AI, 2019). This requires, accountability, robustness, privacy, hard-coded fairness, and transparency. (Dignum, 2019; Arrieta et al., 2020). Dignum (2019) proposes a RAI framework called ART (Accountability, Responsibility, Transparency) and puts the RAI guidelines around the value of human well-being. Accountability involves verifiability, replicability, traceability (Fjeld et al., 2020), and involves trade-offs on what is ethically acceptable. (Arrieta et al., 2020). Responsibility is required among all stakeholders. (HLEG AI, 2019). It emphasizes human-AI collaboration. (Dignum, 2019). Transparency is about openness in order to make AI decisions understood. (Dignum, 2019).

Fairness, a significant ethical demand, requires to integrate human norms into AI design in order to eliminate bias and discrimination. (Gillespie et al., 2020). However, it appears to be a technical challenge. Dignum (2019) calls for a method named 'Design for Values' in which human norm and values are integrated in the programmed functionalities during the process of design and development.

Achieving a competitive and responsible AI needs more research. However, organisations can start to prepare for the eventuality, referring to AGI, by starting with adopting ethical AI principles. (Minkkinen, 2023).

2.1.5 AI capabilities

Based on research on workplaces that integrated AI, which is also a scientific sub-field: 'AI in the workplace', AI capabilities can be divided into three different task types. 1) Mechanical tasks, 2) Thinking tasks, and 3) Feeling tasks. The first type, Mechanical tasks, involve equipment, repair and maintenance which are mechanized by machines and robotics in order to increase productivity. The second type, Thinking tasks, involve cognitive functions like processing, analysing, and interpreting information. (Tolan et al., 2021). AI technologies are increasingly becoming more efficient in these types of tasks. Lastly, Feeling tasks are defined by interpersonal communication which relies on emotional intelligence, a skill that is still unique to humans. (Huang et al., 2019). In a broad sense AI is capable of doing all kinds of tasks in type 1 and 2.

2.2 Explainable AI

Last year a distinguished AI-Safety summit brought AI safety research under the attention. (Jones, 2023). AI has become a widely discussed topic on online forums, where some even claim it to be a threat to human existence. (Perrigo, 2024). Science's response was the sub-domain Responsible AI. This field is often criticized through quotes as *"We are talking about AI that does not yet exist"* (Jones, 2023). However, it might be smart to start safety research to prepare for the eventuality of human-level AGI. (Tegmark, 2018, p. 42,). Hence, the need for regulations and general understanding is increasing. (Birkstedt et al., 2023). The sub-field eXplainable AI (XAI) investigates this 'general understanding' by researching practical methods to ensure ethics in AI.

2.2.1 The definition of XAI

eXplainable AI (XAI) exploded around 2018 along with the significant progress of AI in the past decade. (Tiainen, 2021), (Laato et al., 2022) & (Adadi & Berrada, 2018). The latest literature is considered to be the baseline for the definitions and scope adopted in this research. XAI emerged due to black-box AI systems making decisions without transparent reasoning. (Saeed & Omlin, 2023). Explainability is a result of the human need to explain the output of ML models (Arrieta et al., 2020), aiming to provide meaningful and trustworthy outputs in order to understand AI solutions. (Ali et al., 2023a).

2.2.1.1 The black-box problem

An AI system transforms input data in various ways to create a certain output. Without knowledge of the transformation, inter alia due to its complexity, the system is labelled as a black-box. As a result, the model is considered the body of knowledge instead of the data. (Saeed & Omlin, 2023). Thus, concerns are raised about trust, fairness, accountability, privacy, security, transparency, and ethics in general. (Ali et al., 2023a). A system which inner workings can be observed in an unrestricted manner in order to identify undesired model behaviour is often referred to as a '*White-box'*. (Casper et al., 2024). Unfortunately, there is a trade-off between the model's accuracy of its predictions and its explainability. (Ali et al., 2023a). A 'hybrid-form' is the '*Grey-box'* which keeps an acceptable level of significance of the accuracy and includes possibilities to analyse internal workings.

2.2.1.2 Explainability as a solution

From social, scientific, industrial, model development, and regulation angles, a certain level assurance is demanded to trust whether the model work as intended, without bias. (Saeed & Omlin, 2023). Therefore, XAI aims to make AI systems and their results (output) more understandable to humans. (Nauta et al., 2023). Moreover, understanding is an enabler for trust (Ramon et al., 2021), if explainability can lead to understandability. Hence, explainability is crucial to practically deploy AI models. Furthermore, XAI can support vulnerability detection and protect the model from adversarial attacks that could manipulate the model. (Arrieta et al., 2020).

The research from Arrieta et al. (2020) outlines the foundational concepts of XAI and is extensively included in this research. Besides the beforementioned goal of XAI Arrieta et al. (2020) defines 9 more goals behind the creation of explainable models.
1) Trustworthiness is defined through the degree to which a model acts as intended.
2) Causality among variables since a ML model mainly discovers correlations.
3) Transferability and 4*) informativeness* emphasize the applicability and decision-making support of the model. *5) confidence* resembles generalization of how robust and stable a model is, together with *6) fairness* it covers a degree of reliability. *7) Accessibility, 8) interactivity, and 9) Privacy awareness* are goals that focus on user engagement, model-user interaction and the confidentiality of the data. As a collective the goals optimize the model whilst enhancing model clarity, ethical use, and the trust of users.

In general, XAI differentiates two types of models; models that are interpretable and models that can be explained through external XAI techniques. These are called Ante-hoc and Post-hoc models. (Arrieta et al., 2020). Only, the terms interpretability and explainability are often used interchangeably in the literature. (Zhang et al., 2022). According to Ali et al. (2023a) both concepts are defined to elucidate models, but with a slight nuance. ***Interpretability*** is about disclosing the internal working of the model. Understanding the intrinsic properties enhances the transparency of the model. ***Explainability*** however, is more about revealing the decision-making mechanism to verify whether the prediction is fair and ethical. Both Interpretability and Explainability are pillars that move towards the trustworthiness of the model. (Arrieta et al., 2020).

The XAI literature argues that Interpretability refers to the degree to which a human can understand the cause of a decision made by the model. (Arrieta et al., 2020). Whereas Explainability encompasses the broader context, that includes the methods and processes used to make the operations of an AI understandable. In that regard, Interpretability can be seen as a subset of Explainability. (Markus et al., 2021). Saeed & Omlin (2023) provide a more social definition in the XAI field. They argue that "*Explainability provides insights to a targeted audience to fulfil a need"* (Saeed & Omlin, 2023) and Interpretability is "*the degree to which the provided insights can make sense for the targeted audience's domain knowledge*". (Saeed & Omlin, 2023).

Science has not agreed upon a uniform meaning or representation. From the literature it can be concluded that Interpretability is of a psychological nature and Explainability comes closer to a provided explanation. For example, an IKEA manual to assemble furniture has a step-by-step explanation. The way how a human would interpret the steps is part of how good the explanation is. The manual itself is the explanation and whether it is understood or not, it will remain a provided explanation itself. Meaning, Explainability is an objective characteristic of an algorithm or model and Interpretability is relative towards an user/audience. Hence, the definitions marked in bold are adopted in this research in order to keep the definitions clear and aligned with the auditing field.

Revisiting XAI's distinction of two types of models it is pivotal to adopt more distinguished meaning in the terminology. The model type that is based on interpretability can be seen as a transparent model that allows for mathematical analysis from input to output. In contrast post-hoc explainability models require external XAI techniques to be elucidated. (Arrieta et al., 2020). On a general note, transparent model analysis is about the design of interpretable models. Whereas post-hoc analysis entails explaining decisions of black-box models.

Nauta et al. (2023) further divides explanations into three aspects: Reasoning, Functioning, and Behaviour. Reasoning refers to the process a model uses to reach a decision. A Functioning explanation focuses on internal workings and data structures. The third part of a model that can be explained is Behaviour which is how a model generally operates. Observing input and output is therefore sufficient.

2.2.2 Challenges in XAI

There are still some major knowledge gaps in the field of XAI and how it contributes to RAI. Almost all the challenges are traceable to the complexity of the systems and the lack of transparency of the black-box AI models. (Saeed & Omlin, 2023). Additionally, models cannot provide explanations that are relevant across different data distributions. (Arrieta et al., 2020). The data quality, communication around it, and the sparsity of analysis are challenging in the design phase. (Saeed & Omlin, 2023). Robustness is one of the bigger challenges. As stated before, regardless of the model type, AI models are often susceptible to adversarial attacks. These attacks manipulate the model into learning the wrong things. (Weber et al., 2023). Figure 3 below indicates a variety of challenges that are either currently researched or considered a knowledge gap in the scientific field of XAI.

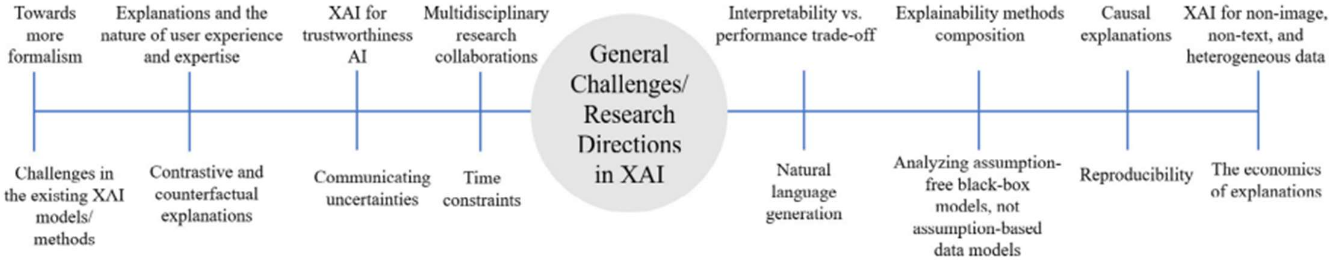

Towards more formalism
Explanations and the nature of user experience and expertise
XAI for trustworthiness AI
Multidisciplinary research collaborations
General Challenges/ Research Directions in XAI
Interpretability vs. performance trade-off
Explainability methods composition
Causal explanations
XAI for non-image, non-text, and heterogeneous data
Challenges in the existing XAI models/ methods
Contrastive and counterfactual explanations
Communicating uncertainties
Time constraints
Natural language generation
Analyzing assumption-free black-box models, not assumption-based data models
Reproducibility
The economics of explanations

Figure 3. Challenges of XAI. (Saeed & Omlin, 2023)

Weber et al. (2023) critiques that if all XAI methods are applied to gain insights into a model, the obtained insights do not necessarily result in better performing, more trustworthy, and/or more fair models. The reason behind this is probably the trade-off between accuracy and interpretability since they are polar opposites. (Saeed & Omlin, 2023; Arrieta et al., 2020). Investigating this topic remains a big challenge as well.

2.2.3 Explainability measures

XAI attempts to make sense of high-performing black-box systems. (Nauta et al., 2023). Nauta et al. (2023) proposes a framework that assesses explanations that are provided specifically by ML models. The assessment is based on the so-called CO-12 properties, which require no direct user participation and evaluates explanations beyond binary assessments. Table 1 below presents the CO12 properties that are based on a review of approximately 300 XAI studies.

CO-12 Property	Description of Explanation review measure
1. Correctness	Instead of predictive performance, correctness focusses on truthfulness. Measuring how accurately an explanation reflects the model's operations
2. Completeness	Evaluating the depth of an explanation which means whether the behaviour of the model is covered comprehensively in the explanation.
3. Consistency	Measures the explanation method's determinism and implementation invariance by testing if identical inputs yield identical explanations.
4. Continuity	Continuity is assessing the smoothness of the explanation function. Meaning, minor input variations should not result in significant explanation changes.
5. Contrastivity	Measures the ability of an explanation to distinguish different outcomes. Different populations might provide non-identical instances which should have dissimilar explanation.
6. Covariate Complexity	Analyses the covariates', or features', complexity that is used in an explanation in terms of semantic meaning and reciprocal interaction. Interpretable and simple functions are preferred by humans.
7. Compactness	This property emphasizes the brevity of explanations. This measure is motivated through the fact that human cognitive abilities are limited. Meaning, explanations that are too big might not be as well understood.
8. Composition	Measures the organization, presentation, and structure of the explanation. Some formats are more effective and clearer and therefore better interpretable. Focus on 'how' rather than 'what' is explained.

CO-12 Property	Description of Explanation review measure
9. Confidence	Explanation certainty and probability on information based on truthfulness, likelihood, and confidence of black-box predictions. However, the effectiveness remains debatable.
10. Context	Explanations should be tailored to user specific information needs and general expertise. Relevance and actionability are important measures for this property.
11. Coherence	The coherence property requires assessment on whether the explanation is aligned with exiting knowledge and beliefs in order to address plausibility and reasonableness.
12. Controllability	Science states that 'explanations have a social and interactive nature'. Therefore, this property assesses whether users are allowed to interact, control, or correct the explanation.

Table 1. Explanation evaluation properties. (Nauta et al., 2023)

Each CO-12 property can be analysed through a specific set of evaluation methods. The perceived explainability of a model increases when explanations are found of high-quality according to the systematic approach of Nauta et al. (2023). Naturally, it is very case-, user-, and client dependent which property should be evaluated.

XAI techniques mainly fit the earlier described type of Post-hoc explainability models. Post-Hoc Explainability can be categorized in Model-Agnostic and Model-Specific techniques. The former is adaptable to any model and the latter includes techniques that are tailored for more precise explanations. The knowledge from this paper and the continuous updated XAI techniques website from Nauta (2023), presented in figure 4 below, provide a well-constructed overview for all current XAI.

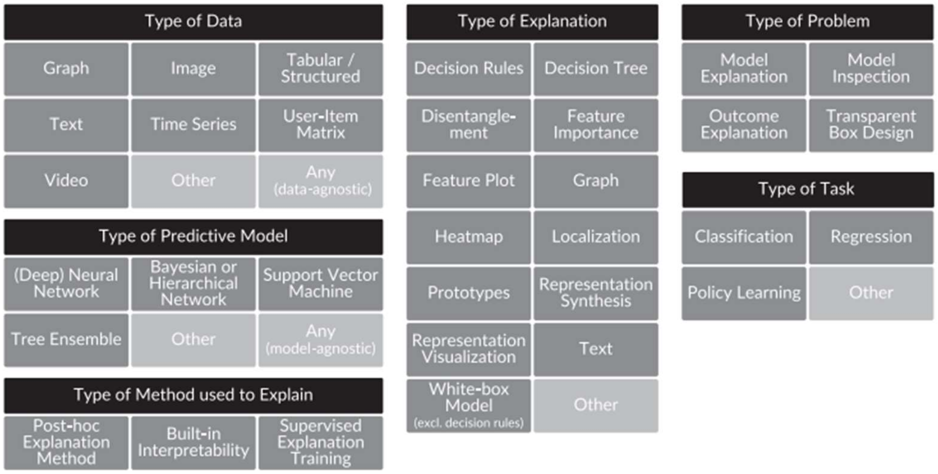


Figure 4. Current XAI techniques. (Nauta et al., 2023)

Ali et al. (2023) seconds the CO-12 framework. Despite agreements the framework is not an inducement to close the XAI terminology gap. The quality of explanations is determined through attributes as translucency, which defines how a method probes the model, generalizability, explanatory power, which is the number of events a method can explain, and lastly the algorithmic complexity that states that explanations can be as complex as the model itself according to Ali et al. (2023).

Worth mentioning is the recent development of a model-agnostic measure called Degree of Explainability, DoX. This metric is developed by Sovrano & Vitali (2023) and proposes an objective way to assess a model's explainability based on the so-called Theory of Explanations from philosophy. The higher the DoX score the higher the chance the assessed AI system's explanations are more accessible and understandable. This research considers DoX's readiness and reliability for widespread use as too pioneering, complex and out of scope given the timeframe and the lack of DoX tooling.

More established xAI metrics, or methods, are LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive exPlanations). LIME is a technique which is applied to explain complex ML models' predictions. (Chen et al., 2023). LIME was first proposed by Ribeiro et al. (2016) and aims to "*identify an interpretable model over the interpretable representation that is locally faithful to the classifier*". LIME is considered model-agnostic. Therefore, LIME is applicable to a multitude of ML, and potentially AI-system types. SHAP quantifiably examines the impact of features, or input, on the model's predictions through the calculation of Shapley values from the cooperative game theory. (Chen et al., 2023). The SHAP method was proposed by Lundberg & Lee (2017) to provide users with interpretation of predictions in order to lessen the tension of the accuracy and interpretability trade-off.

Both LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) are more technical and require a very thorough understanding of the to-be explained model's underlying mathematical and computational fundamentals. For this reason, it is not likely this research will encounter these techniques in AI audit cases. This holds true for DoX (Sovrano & Vitali, 2023) as well.

2.2.4 Transparency measures

Establishing trust in any system often relies on understanding the system's operational mechanisms, or inner works. Transparency provides a certain perception of control and predictability of the system. Therefore, transparency is one of the facets that is related to trust. (Laato et al., 2022; Schmidt et al., 2020). It is important to separate the definitions of Explainability and Interpretability to define transparency. Where transparency is about the model, the transformation, and the inner workings. Explainability is mostly post-hoc analysis closer to the output of the model. (Rendon, 2022). Other articles critique that transparency is part of Explainability. (Balasubramaniam et al., 2023). Arguably, it is difficult to determine a supported definition.

This research adopts a definition derived from the most recent and relatively leading articles in XAI. According to Adadi & Berrada (2018) both Interpretability and Explainability are an effort made to respond to AI transparency despite their nuanced difference. Dignum (2019) links accountability to interpretability and transparency. Commenting that XAI should openly include all stakeholders in its operational interpretations that are elicited. This aligns with the idea that transparency is primarily about communication. (Balasubramaniam et al., 2023). Therefore, transparency can be provided by the application of governance and provision of guidelines. For example, digital platforms can, and should, provide reports on their terms of service and data processing. (Suzor et al., 2019). This is increasingly mandated by law enforcement. (HLEG AI, 2019). Another example of providing transparency is to showcase data accuracy and data quality in a dataset. Laato et al. (2022) notes that transparency is measured by how effective an AI system can communicate its processes and decisions. Which are assessed through user understanding, trustworthiness, and decision control.

According to Arrieta et al. (2020), there are three levels, derived from XAI, in transparency that represent the characteristics of the models that influence explainability: 1) Algorithmic Transparency, 2) Decomposability, and 3) Simulatability.

Algorithmic Transparency is about the user's ability to grasp the process that the model is using to transform inputs into outputs. A model is algorithmically transparent if it is fully explorable and mathematically accessible.

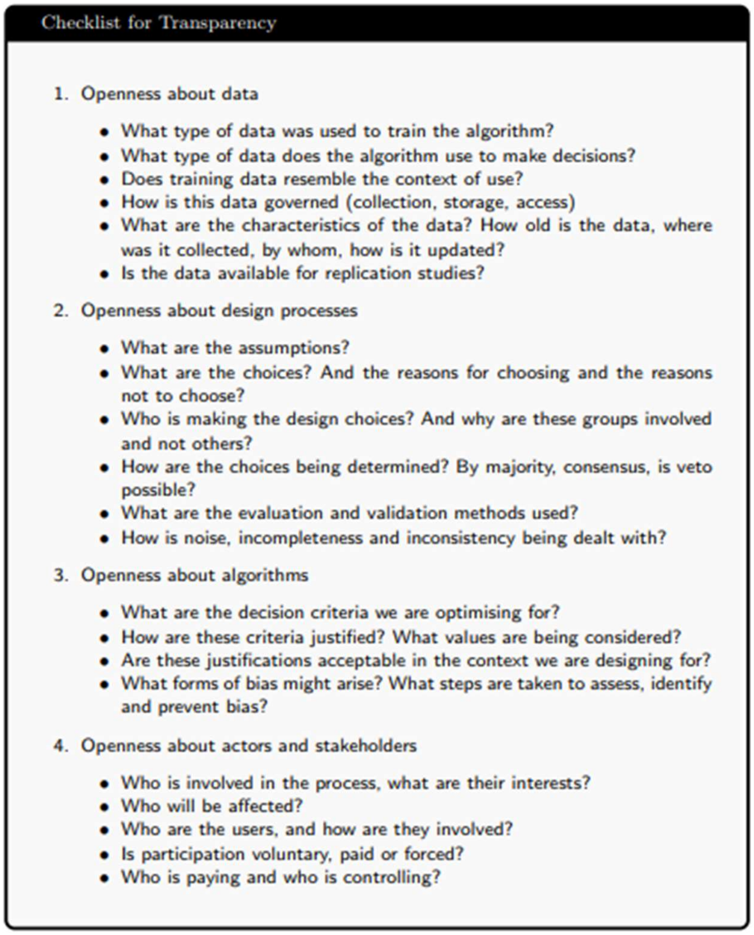
Decomposability requires that every input is immediately interpretable. An algorithmically transparent model can only be considered decomposable if every part of the model is humanly understandable without using tooling.

Simulatability refers to a model's capability to be simulated by a human. A model can be humanly simulated if it is not overwhelmingly complex and can be comprehended without tools.

Conversely, these three levels can be referred to as 'phases of interpretability'. (Ali et al., 2023). In the current stage of XAI research, interchangeable terms are inevitable. According to Ali et al. (2023) transparency could be achieved by using intrinsic methods that produce explanations on the decisions made by a model. Pivotal to safeguard the model against adversaries and evaluate the quality of the decisions. In conclusion, transparency is two-fold. Transparency can be provided through explanations or established and controlled through AI governance. The other perspective is that a transparent model is a model that can be assessed, whether it requires post-hoc explainability or it is a grey- or white-box model.

After theoretical distinction it is important to explore how transparency can be measured, even though its definition is relatively opaque. Ali et al. (2023) proposed three types of assessment: 1) System competence, 2) Compliance with the system, and 3) Understandability. Only these methods lack the quantitative exactness like LIME and SHAP. According to Bennetot et al. (2022) and Casper et al. (2024) determining model transparency for models that are not transparent-by-design (white-box or grey-box) is challenging. A reason could be that humans hold AI system transparency to a very high standard. Which is unfair according to Zerilli et al. (2019). Moreover, Dignum (2019) conveys that achieving algorithmic transparency is not as straightforward as making code and system's data fully open to inspection. As organisations often prioritise functional performance in the algorithmic design, black-boxes continue to exist. Hence, it is important to continue research on defining and measuring transparency in XAI.

A good practice might be to govern transparency over the entire learning and training process instead of inspecting, or even removing, the concept of a black-box model. (Dignum, 2019). This governance perspective presents a checklist for transparency presented in figure 5. This checklist provides a qualitative view on the auditability and traceability of a model which already provides information on the investigation on how transparent a model is. The definition of explainability might be more thorough, but the opaqueness around the concept of transparency forces to adopt a more governance-oriented definition. Meaning, the transparency of a model is often found through communication, (Balasubramaniam et al., 2023), and the environment of the model.



Checklist for Transparency

1. Openness about data
   - What type of data was used to train the algorithm?
   - What type of data does the algorithm use to make decisions?
   - Does training data resemble the context of use?
   - How is this data governed (collection, storage, access)
   - What are the characteristics of the data? How old is the data, where was it collected, by whom, how is it updated?
   - Is the data available for replication studies?
2. Openness about design processes
   - What are the assumptions?
   - What are the choices? And the reasons for choosing and the reasons not to choose?
   - Who is making the design choices? And why are these groups involved and not others?
   - How are the choices being determined? By majority, consensus, is veto possible?
   - What are the evaluation and validation methods used?
   - How is noise, incompleteness and inconsistency being dealt with?
3. Openness about algorithms
   - What are the decision criteria we are optimising for?
   - How are these criteria justified? What values are being considered?
   - Are these justifications acceptable in the context we are designing for?
   - What forms of bias might arise? What steps are taken to assess, identify and prevent bias?
4. Openness about actors and stakeholders
   - Who is involved in the process, what are their interests?
   - Who will be affected?
   - Who are the users, and how are they involved?
   - Is participation voluntary, paid or forced?
   - Who is paying and who is controlling?

Figure 5. Checklist for AI transparency. (Dignum, 2019)

2.3 Auditing

One of the Big 4 Companies explains an audit as a control of an organisation's financial report, part of the annual report, performed by an independent entity or body. The report entails a balance sheet, income statement, changes in equity, and more financial flows including policies. The goal of an audit is to assure whether the financial report represents the actual financial state. (PWC, n.d.). Besides the financial part it includes non-financial disclosers as well. In general, the primary goal of an audit is to provide an objective and unbiased view through assessments. According to Rittenberg et al. (2010) the audit process exists of steps such as audit planning, execution and reporting. Figure 6 presents the general flow of all (financial) audit phases. The beforementioned process steps could hypothetically serve as a structure for AI audits, mutatis mutandis.

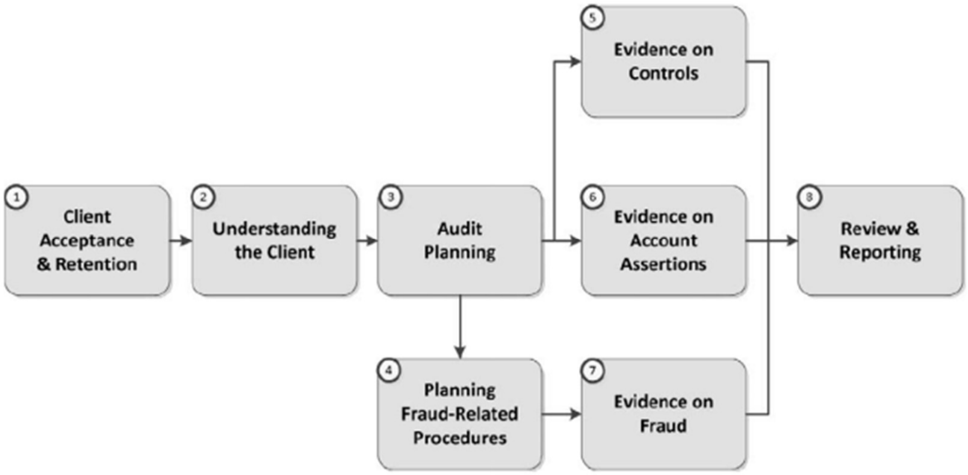

1 Client Acceptance & Retention
2 Understanding the Client
3 Audit Planning
4 Planning Fraud-Related Procedures
5 Evidence on Controls
6 Evidence on Account Assertions
7 Evidence on Fraud
8 Review & Reporting

Figure 6. Generalised auditing process. (Rittenberg et al., 2010)

Another report summarises the above to four distinct steps in the so-called risk-scenario-based audit process. It involves 1) Plan, 2) Define Scenarios, 3) Measure, and 4) evaluate. The last step represents the review and reporting, overlapping with Rittenberg's process. A subset of auditing is the IT audit profession. An IT audit is *"the examination and evaluation of an organisation's information technology infrastructure, applications, data use and management, policies, procedures and operational processes against recognized standards or established policies"* (Harvard University, n.d.). To be short, IT auditing focusses on the aspects of 'CIA'; Confidentiality, Integrity, and Availability and how they are aligned with the organisational goals, objectives or applicable standards such as COBIT for example. It is essential to ensure that IT systems' control design and effectiveness are adequately controlled to secure they are functioning as intended. Key elements for this are basing the audit plan on risk analysis, ensuring independent auditors, gathering comprehensive insights about the IT environment, developing control objectives, and carrying out a very thorough audit program. (Petterson, 2005). To provide a schematic overview, figure 7 was made to indicate IT auditing's position towards the entire financial audit.

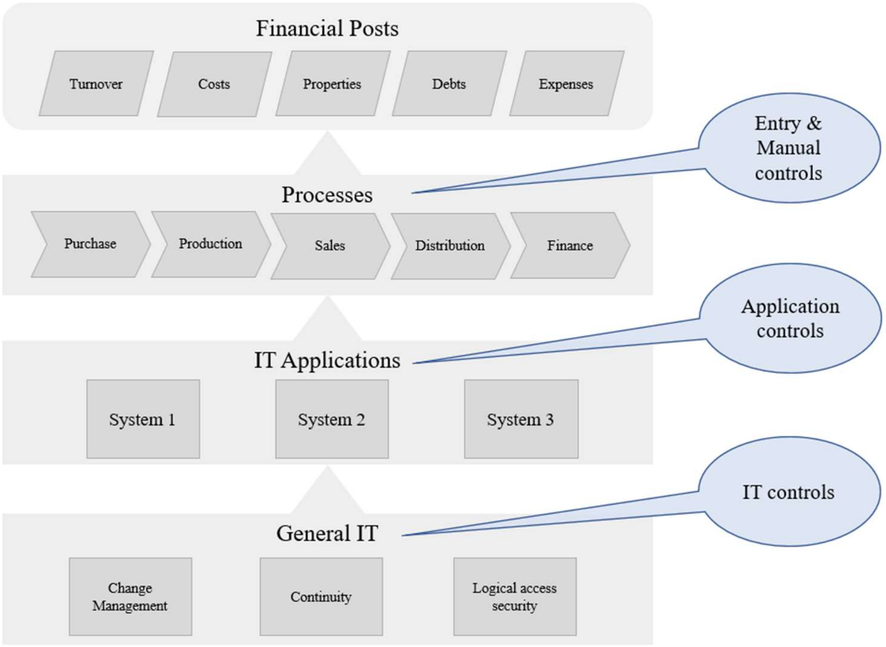


Figure 7. Clarification on IT auditing

In general, there are internal and external audits. Internal audits are set up by the organisation itself and seeks to identify and prioritise areas of risks from within. External audits are done by an external auditing party in regard to object of the audit. (Raji et al., 2020). In a general sense, an audit is meant to identify risks, but also investigate the way how these risks are mitigated and controlled. Through control testing, data analysis, and other data-oriented activities an audit attempts to state 'something' about the reliability, accountability, completeness, correctness, and integrity of the (financial) data and the supporting (IT) systems. Understanding the risks helps to better execute the audit and thus helps to improve the business in the long run. Financial data flows from humans to IT systems and back. If there are controls at place, the systems can be audited on whether they perform in accordingly. (Raji et al., 2020).

When information is utilized as audit evidence, its reliability and relevance must be evaluated. This is in line with KPMG's internal perspective upon auditing and the general knowledge from Rittenberg et al. (2010). Reliability means that the information is accurate and complete for auditing purposes. Hence, information reliability requires transparency about the sources (internal or external) and the nature (type of documents) of the data. Understanding the business context, source of the information, the nature of the information, and the circumstances allow auditors to determine the appropriate audit procedures.

KPMG identifies several risks, including input risks (inaccurate data entry), integrity risks (inappropriate data alteration), and extraction & manipulation risks (data loss or errors). According to Rittenberg et al. (2010), thorough risk assessment is crucial in the 'client understanding' phase of the audit process in figure 6. Referring back to the chapter on explainable AI, some explainability on the data is required to estimate whether risks are covered and mitigated by the right controls.

Note that this information is based on general auditing knowledge and KPMG's internal perspective and understanding. The following chapter, 2.4, provides a description of data explainability within the AI auditing setting.

2.4 AI Auditing

AI auditing is not a profession that is extensively present in practice or science. Nevertheless, organisations have an increased access to AI tools or even deploy their own complex models. (Raji et al., 2020). Thereupon, auditors might run into systems or algorithms that are opaque and difficult to assess. Through the lens of an auditor an AI or algorithm introduces considerable, digital or algorithmic risks that could result into a real-world problem. (Boer et al., 2023). Arguably not every AI model or algorithm is an entity of concern. If the model yields unacceptable results that are not accompanied with severe consequences and the problem has been well-tested and studied then there are arguments to leave 'it' in the (black-)box. (Saeed & Omlin, 2023).

However, the audit perspective requires some forms of compliance, especially after the announced release of this year's AI-Act, 2024, in which statements about explainability and transparency are included. The field of XAI describes the 'auditability' of AI systems. Auditability basically means that AI models, systems, and algorithms should be able to be independently audited. (Arrieta et al., 2020). Therefore, this paragraph will outline the current knowledge on AI auditing.

2.4.1 Algorithmic Assurance

Although there is not much documentation on AI auditing there is research on providing assurance over algorithms through the lens of IT auditing. The key auditable parts of algorithms are mainly the data (input, output and training-datasets), the model (parameters and formulation), and development (design phase and involved stakeholders). Each of these parts can be audited for completeness, correctness, bias, fairness, and other ethical concerns from the XAI field. Notably, the model can still be ignored and considered black-box. (Koshiyama et al., 2022).

Like (IT) auditing algorithm assurance, or auditing, addresses risks associated with AI applications since AI systems operate on algorithms of predictive models. (Boer et al., 2023). The identified risk factors by Boer et al. (2023) are autonomy (human oversight), complexity (technological and operational intricacies), and influence (the scope and impact of decisions). If an AI system or algorithm operates across all three dimensions and could be considered high risk on at least one of them, the AI entity is likely to be audited. (Boer et al., 2023).

The referred paper from Boer et al. (2023) dives deeper into the matter and further explains the differences between the risk dimensions. The algorithm audit process assesses whether algorithms are conformed to regulation, governance, and ethical standards. (Meßmer & Degeling, 2023). Figure 8 below presents an example of an algorithm auditing process and concentrates around the activities of development, assessment, mitigation, and continuous improvement to reinforce every step iteratively. (Koshiyama et al., 2022). Figure 8 is not exactly a flowchart or representation of a real auditing process. Upon comparison to the XAI theories and auditing field it can be concluded that the figure represents the client's perspective. The bullet points in the picture suggest areas of interest for an auditing party.

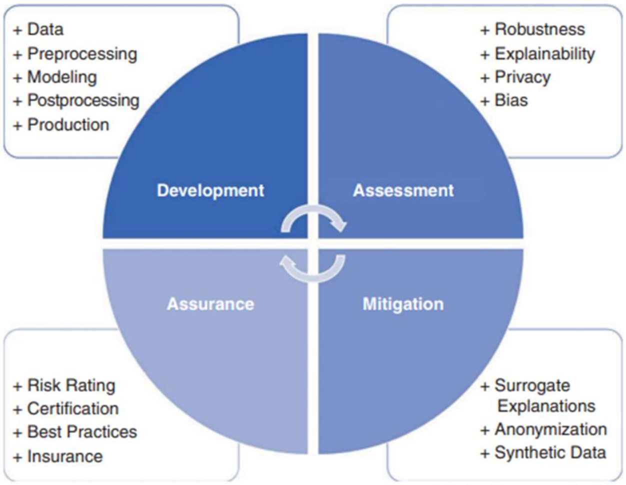


Figure 8. The activities of algorithm auditing. (koshiyama et al., 2022)

Roughly, algorithm assurance/auditing has significant overlap with the principles and theories of the XAI domain. (Zhang et al., 2022). The main difference is that it takes a more pragmatic approach since it could be considered as XAI's execution. It is important to consider the broader system environment as AI system decisions might be jointly made with traditional IT systems. (Boer et al., 2023). Solely auditing the algorithm is markedly insufficient. Although, a theoretical background and framework is provided, formalism remains a challenge.

2.4.2 Data explainability

Besides documentation made available throughout the design and development phases a model could still be considered a black-box and difficult to audit. The previous paragraph on (IT) auditing addressed the importance of investigating the data and the belonging controls. Data involves certain types of risks. One of the more tangible and auditable parts of a model is its data in terms of the input, output and datasets the model is trained on. Ali et al. (2023) articulates the importance of data quality as it is impossible to reach a well performing model on low data quality. Thus, after data collection rigor data examination is needed. A form of data explainability needs to be established as an addition to not only the field of algorithm assurance and AI auditing, but also XAI. Data explainability involves methods to better comprehend the datasets of AI models' design and training. (Ali et al., 2023a). Explaining data implies well organised meta data. Which involves information on the data sources, the data origins, pre-processing procedures, data bias, and variable features in the datasets. (Ali et al., 2023;Arrieta et al., 2020; Meßmer & Degeling, 2023; Boer et al., 2023). Data explainability is often implicitly reported and not (yet) a formalized concept.

2.4.3 Current guidelines on AI auditing

Legislative frameworks as the Digital Services Act (DSA), Digital Markets Act (DMA), and the EU AI Act (EUAIA) outline the primary audit guidelines. Meßmer & Degeling (2023) provide a report on a case in which recommender systems are audited by using the DSA. This Act requires platforms or organisations to make the algorithms, that their (AI) systems are using, more transparent and in compliance with due diligence obligations. The complete act, along with the DMA, can be retrieved from EU Commission (n.d.-b) and includes statements that are relatively auditable. The DMA aims to prevent market dominance abuse by large organisations. (EU Commission, n.d.-a). Compared to DSA the DMA has more of an indirect effect upon algorithm auditing. Mainly through its requirements for transparency and fairness in the digital environment. The DSA mandates external auditing on transparency on the consequences of algorithms and risk assessments. The EUAIA is per June 2024 not enacted long enough to provide literature on its effects.

2.4.4 AI Governance

Being in control requires organisational governance, the same accounts for ‘controlling’ AI. Especially in the light of XAI, RAI, and other concerns around AI it is pivotal to translate abstract ethical principles into executable governance practices. An ‘AI lifecycle’, as a case in point, includes AI design, development, deployment, and continuous monitoring (Mäntymäki et al., 2023). Forthwith, the overall quality and assurance will improve. Figure 9 positions AI Governance to all other organisational forms of governance.

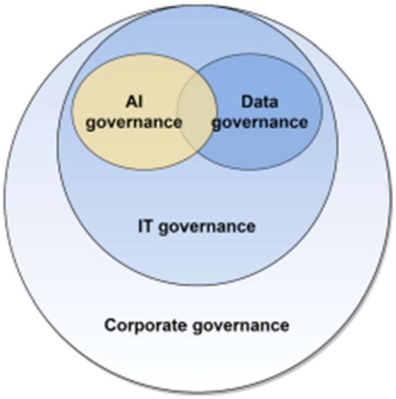


Figure 9 AI Governance. (Mäntymäki et al., 2022)

Data governance treats data as a strategic asset by establishing a formalised framework for data-related decisions through data policies and standards. (Mäntymäki et al., 2022). Data governance supports AI governance by enhancing explainability and transparency in AI models by addressing biases. Monitoring compliance with data quality rules is crucial for AI governance. AI governance specifically ensures organisational alignment between AI, organisational rules, strategies, values and objectives. (Mäntymäki et al., 2022). AI governance is important since there is a broader perspective than independently auditing an algorithm. (Boer et al., 2023). Furthermore, conducting (AI) audits is easier when an organisation has established thorough governance. It provides a reference frame to the audit.

3 Method

This chapter develops the methodological approach adopted in this thesis. The research is inherently explorative due to the nascent stage of the academic fields XAI and AI auditing. The method chapter firstly introduces the theoretical frame of this research and the second paragraph explains the application of the frame in the research process. The last paragraph ensures the trustworthiness of the research.

Although XAI is rapidly emerging it is concluded that it is without formalised rigorous evaluation metrics due to the absence of consensus on definitions and the lack of practical information from well documented cases. (Adadi & Berrada, 2018; Saeed & Omlin, 2023; Birkstedt et al., 2023). Furthermore, the auditing and assurance approaches for AI-like systems are not yet fully comprehensive. (Boer et al., 2023). Given the current status of the academic field and the exploratory nature of this research, it is both logical and appropriate to adopt Design Science Research as theoretical framework.

3.1 Design Science Research

The relation between auditing, AI systems, and the XAI facets like transparency and explainability is underexplored. (Birkstedt et al., 2023). Therefore, the primary objective of this research is to explore and formulate novel insights that contribute to the understanding of how AI-systems could be audited. This research aims to provide insights in the practical profession of auditing AI systems by investigating current practices and comparing them to XAI literature. Through the application of the Design Science Research methodology new theories and ideas can be found through the creation and analysis of Information System (IS) objects. (Wieringa, 2014). An object, or according to the Design Science Research methodology; an artefact is designed within a problem context in order to improve something in that context. (Wieringa, 2014).

Design Science Research, further to be referred as DSR, entails three main blocks, as presented in figure 10 below; 1) Environment, 2) Knowledge base, and 3) IS Research. Block 1, the environment, defines the entire space of the problem. (Hevner & Park, 2004). In this case this so-called ‘problem of interest’ can be loosely described as the unformatted endeavour to audit complex (black-box) algorithms. In addition, the environment represents the place in which the to-be developed solution, or artefact rather, operates. (Wieringa, 2014).

As can be seen in the figure below, the environment includes all people, processes, and existing or envisioned technologies. (Hevner & Park, 2004).

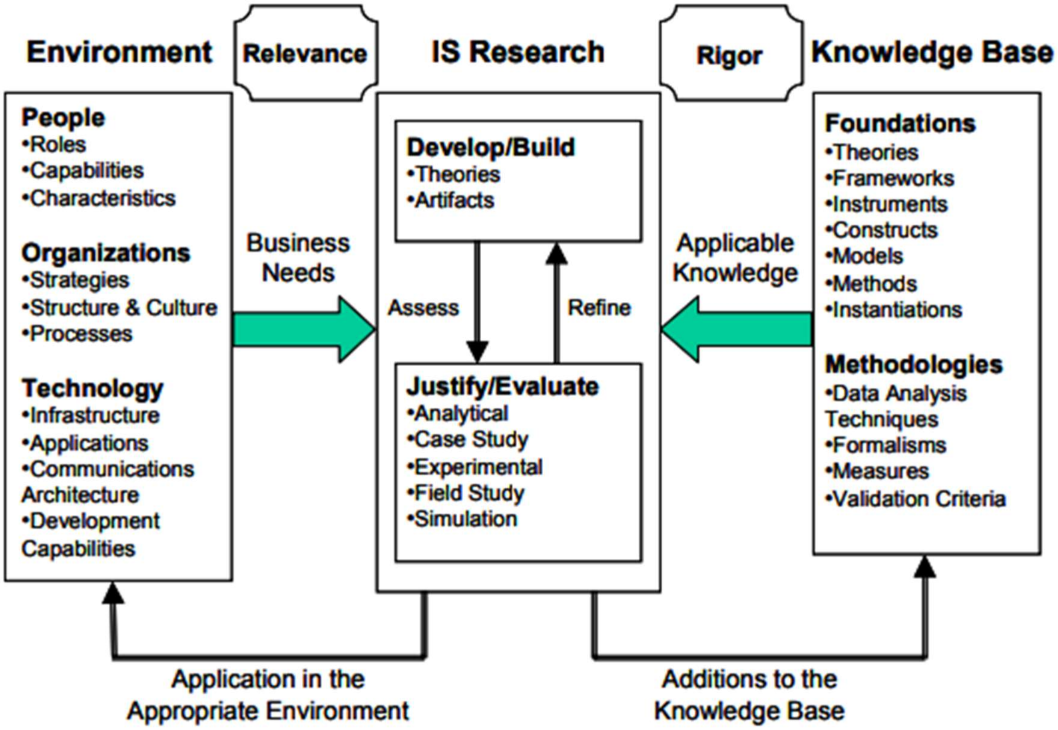


Figure 10. Design Science Research Method. (Hevner & Park, 2004)

This research is empirically conducted during a work placement at KPMG Netherlands. Hence, KPMG will serve as the environment, or organisation, of interest. The departments included in the study are IT Assurance and Responsible AI / Trusted Analytics. The people of interest are predominantly the ones that encountered algorithms, AI-systems or black-box like entities during audit or assurance assignments at KPMG's clients. The people of interest are (IT) auditors, IT Assurance consultants, (Responsible) AI consultants and analysts, or managers and researchers within the organisation.

The second block, on the right side of figure 10, is the Knowledge Base. This represents all the material which is found through literature research. It involves all earlier researched foundations and methodologies presented in XAI and AI auditing literature. The Knowledge Base formulates the input basis onto which the proposed solution will be developed. (Gregor & Hevner, 2013).

Lastly, the third block in the middle of figure 10 is the IS research block in which the so-called artefact is build and developed upon the gained knowledge. Creating an artefact is the centralised point of attention for DSR since it is the outcome of this theoretical research frame. (Hevner & Park, 2004). An IS artefact can be developed based on the findings of current ways to audit and assess AI-systems. In line with the DSR, an artefact is an object that is made with the purpose to tackle a practical problem. (Weigand et al., 2021). IS DSR research from Hevner & Park (2004) describes that through these artefacts IT related problems in organizations are understood. IS artefacts can appear as models, methods, systems, constructs, or instantiations. (Hevner & Park, 2004). Weigand et al. (2021) includes algorithms, and modelling languages to the definition. Concluding, an artefact is an object made by humans and used in practice. (Wieringa, 2014).

The to-be developed artefact should interact with the Environment, which is KMPG's auditing departments such as Responsible AI and IT assurance. The people in these departments run into clients that make use of complex algorithmic systems. Leading to all kinds of discussions and difficulties in IT and regular financial statement audits. A potential artefact could be a type of maturity model, roadmap, or general approach which explicates which type of organisations and their AI-system require which auditing method. The general fundamental idea would be a set of guidelines and requirements that an AI-system must meet to be audited. Partly these ideas are met through literature research and the other part through empirical research.

As can be seen in figure 10**,** there is one other block, 'Justification & Evaluation'. The process of DSR is iterative. Meaning, the development of the artefact will be refined by testing it in the environment. However, this research includes only one iteration, from preliminary model towards a refined model. This will be further explained in the research process paragraph and the limitations in the discussion.

3.2 Research process

This paragraph describes the exact research activities and utilized research method in line with the theoretical framework of DSR. First the general research process is explained, then the utilized methods are emphasized in sub-paragraphs.

In this research process the main method is expert interviews, which is the most viable option due to the fact that AI auditing simply needs to be explored and the visions and ideas of experts could provide interesting findings. Figure 11 below summarises the conducted research activities. The process in figure 11 is based on the DSR theory from Hevner & Park (2004) and the formalized process from Johannesson & Perjons (2014). The process is slightly adapted to the time constraints, project limitations, case specifics and availability, and organizational dependency. The content schema of this research document follows the identified sections from Gregor & Hevner (2013).

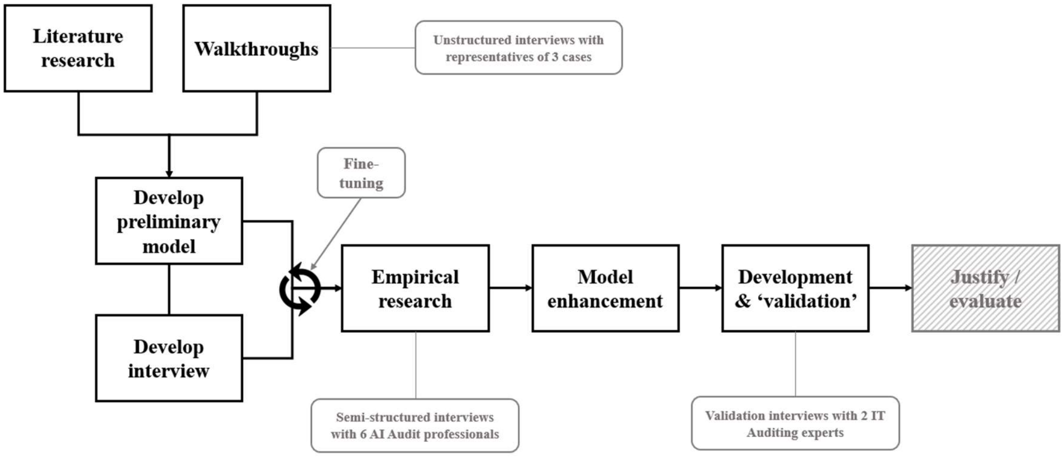

Literature research
Walkthroughs
Unstructured interviews with representatives of 3 cases
Fine-tuning
Develop preliminary model
Develop interview
Empirical research
Model enhancement
Development & 'validation'
Justify / evaluate
Semi-structured interviews with 6 AI Audit professionals
Validation interviews with 2 IT Auditing experts

Figure 11. The research process

The research started with literature research in order to gain a baseline of reasoning (the Knowledge Base). Additionally, the literature provides an overview on what is known about AI auditing. Simultaneously, the Environment was explored to gain understanding of the 'problem' and investigate the relationship between theory and practice. The exploration of the environment were orientating unstructured interviews with representatives of 3 KPMG client cases in which an AI or algorithm audit had taken place. The unstructured interviews were practically case walkthroughs in which each representative explicated how AI like systems or algorithms were addressed. Paragraph 3.2.2. describes how the cases were selected and analysed.

The combination of literature and walkthroughs led to a preliminary model or graphical representation of a first line of thinking to address AI auditing. A preliminary model or conceptual draw of an idea helps to make unstructured conversations about the novel topic slightly more structured. Thus, the unstructured interviews provided initial understanding and input for a set of more in-dept interview questions about AI auditing with AI experts to enhance the preliminary model. Between these steps and the semi-structured interviews with (responsible) AI experts there is continuous fine-tuning. All the people that are interviewed are pioneers in AI auditing, meaning new insights are found daily and shared mutually. Therefore, the interview questions are slightly adapted per interview and per case (if the interviewee happened to work within one of the designated cases). After interviewing AI audit professionals from KPMG, the model will be enhanced and after that developed into an artefact.

As can be seen in the research process, figure 11, the justify and evaluate step from the theoretical DSR frame is greyed out. During this step and actual case study should be done on the implementation of the artefact to validate its relevance, acceptance, and workings. However, due to a limited scope and time constraints it is difficult to execute. Another, more vital problem is the fact that the clients do not allow such thorough research. A lot of the information around clients' AI and complex algorithms is concealed and require too much disclosure to carry out proper research. Hence, the cases that were explored during the walkthroughs via unstructured interviews were used as input rather than factual objects of research. (Paragraph 3.2.2.). To compensate the justification part of DSR two validation interviews were conducted. The artefact will be validated through feedback sessions with IT auditors. Although, IT auditors lack significant experience with auditing AI's, the perspective IT auditors bring is considered as very valuable since IT auditing exists far longer than AI auditing and is thus familiar with best practices and standards. As a result, the newly developed artefact is checked upon clearance and sensibility by two professionals who are familiar with auditing artefacts like approaches, practices and standards.

3.2.1 Literature study

As discussed in the overarching paragraph, the first part of this research was to study the literature in order to accustom with the current knowledge and theories on the topics of AI, XAI, auditing, and AI/algorithm assurance/auditing. The literature research partly answers the research question and sub questions. The acquired knowledge serves as general reasoning and input. The approach to finding literature was mainly snowball sampling from top tier articles in the field of AI and XAI. Where AI research has been around longer the idea of making it explainable or responsible really started around the year 2018. The leading articles provide a lot of references to other well-established articles. Furthermore, the literature research relied on the following keywords: *transparency, explainability, AI ethics, Responsible AI, AI systems, AI auditing, auditing, AI assurance, and explainable AI*.

3.2.2 Cases

As presented in the research process, unstructured interviews, or case walkthroughs, were pivotal input to the outcome of this research. There is a difference between using a case as input for DSR and doing an actual case study. A lot of case information was made available by KPMG, such as documentation and people to be interviewed. A case study requires an in-dept and highly detailed examination of a case with a narrow focus. (Yin, 1981). Case Study Research (CSR) aims to find generalisable results through a systematic investigation of a multitude of cases. Whereas the goal of the case studies in DSR, in this research, is to contribute to the design by describing and analysing cases. Hence, the cases provided contextual information and do not require further analysis than the description in the appendices (appendix 2). These cases can be seen as possible use cases of the artefact.

The IT assurance and RAI departments of KPMG provided 3 cases. Each represent large multinational organisations in which either complex algorithms were encountered or formulated one of the audit objectives. The organisations are all anonymised with the 'CO' (Case Organisation) abbreviation. CO1 provided a very straightforward case in which a ML tool was audited in order to see whether assurance could be provided. CO2 and CO3 are currently undergoing DSA / DMA audits, part of these audits are complex algorithms and black-box like systems. Both CO2 and CO3 need to be compliant to the legislation. These cases were added because a DSA / DMA audit happen to be the first AI like audits for which actual lawful articles are provided. Table 2 below gives a description on what type of company is looked into. The cases were acquired through convenience sampling after brief orientation at the referred KPMG departments. Table 2 below gives a description of the types of case organisations.

Organisation reference	Type of organisation
CO1	Major Dutch insurance company
CO2	Very Large Online Platform (VLOP)
CO3	Multinational e-commerce company *(VLOP)*

Table 2. Type of case organisations

3.2.3 Interviews

This research conducted two types of interviews. Unstructured interviews to describe the cases and semi-structured interviews with AI audit professionals to gain more in-depth information. The latter is dubbed as ‘main interviews’. The interview questions were derived from the research questions in chapter 1 and the combination of literature and walkthroughs. Besides asking the AI professionals questions there was space for follow-up questions. Hence, the main interviews were semi-structured. Because of the distance and personal planning of the interviewees it was not always possible to do the interviews face-to-face. According to Saarijärvi & Bratt (2021) online video interviews are a good alternative. A benefit of online interviews is to ask follow-up questions through the online platform that has been used for the interviews. (Curasi, 2001). KPMG makes use of Microsoft Teams. One of the main reasons this thesis made use of online interviews is the release of Co-Pilot in Teams which enables interviewer and interviewee to transcribe everything automatically. Only one of the main interviews was face-to-face (with interviewee A).

In order to establish a personal relationship with all the interviewees an introductory ‘coffee-meeting’ was scheduled prior to the actual interview. Furthermore, the orientating interviews (walkthroughs) were important to level with the AI experts. Based on the interviewees’ answers, attitude and reactions some interview adjustments were made.

While the main interviews were often related to the cases that are attached in this document, the questions aimed to a deeper level. Whereas the unstructured interviews took place in advance and were meant to simply describe the case itself. Naturally, there is some overlap in the findings from both type of interviews since some of the interviewees have worked with one of the cases which adds to their professional experiences and opinions. Table 3 describes the details of the 3 unstructured interviews that were conducted. It was found sufficient to have only 1 interview per case due to the informativeness of the interviews and the limited availability of human resources.

Interviewee name	Role towards case	Department	Date	Walkthrough #
CO1 representative	Client engagement	IT Assurance	15-04-2024	1
CO2 representative	AI consultant / analyst	Responsible AI	29-04-2024	2
CO3 representative	AI consultant / analyst	Responsible AI	30-04-2024	3

Table 3. Overview of the 3 unstructured interviews

The unstructured interviews are in this research also referred to as ‘walkthroughs’ since the interviewees literally walked the researcher through the cases in an unstructured manner. The knowledge gained from the cases determined, together with the literature, an initial understanding of AI auditing which led to the preliminary model in the results. The table below represents the main interviews which were based on the earlier acquired knowledge in order to achieve empirical and deeper understanding.

Interviewee name	Function	Department	Date	Time	Interview #
A	Senior Manager RAI	Responsible AI	30-04-2024	11:00 – 12:00	2
B	RAI consultant	Responsible AI	01-05-2024	10:00 – 11:15	3
C	Senior RAI consultant	Responsible AI	22-04-2024	16:00 – 17:00	1
D	RAI consultant	Responsible AI	02-05-2024	11:00 – 12:00	4
E	Senior ITA consultant	IT Assurance	14-05-2024	13:30 – 14:00	5
F	Director	IT Assurance (RAI)	22-05-2024	15:00 – 16:00	6

Table 4. Overview of the 6 interviews with experts

In order to provide a form of evaluation and validation of the model two validation interviews were planned with managers that have strong affinity with IT auditing. The most important reason for validation interviews was to receive feedback on the artefact from an IT auditing perspective. Within the IT auditing field best practices are already determined and every now and then AI systems are encountered. IT auditors are able to recognise activities or pinpoint certain gaps in AI auditing that would be very logical to conduct during IT audits.

Interviewee name	Function	Department	Date	Time	Validation #
G	Manager ITA	IT Audit & ITA	23-05-2024	13:30 – 14:30	1
H	Manager ITA	IT Audit & ITA	27-05-2024	11:00 – 11:30	2

Table 5. Overview of the 2 artefact validation interviews

3.2.4 Data analysis

As indicated before, throughout the interview process some adjustments have been done. This research is exploratory in nature because it follows the DSR framing. Hence, this nature allows for adjustments during the data processing and collection. (Wieringa, 2014). In other words, the interviews were analysed preceding the next interview in order to fine-tune the data collection method. The result of this is that the analysed interviews provide insights that are interesting to further investigate during another interview. Furthermore, the unstructured interviews served as a knowledge bridge to dive deeper into the matter with the interviews structured through questions. After the unstructured interviews a close to 1-page summary of the case was made and sent to the case representatives for approval. Since the semi-structured interviews were online, a transcription from Microsoft Teams' Co-Pilot could be obtained. The transcription of each interview was used as input for the interview summaries that are attached in appendix 3. It was often not necessary to validate the summary of the transcription afterwards because all interviewees are able to see the transcription live. The only summary that received some feedback was the summary on the Dutch insurance case (CO1). The other case summaries were agreed upon. The interviewees for the structured interviews indicated to be content with a summary over the transcription without a review.

3.3 Trustworthiness of the research

There are several factors that determine the quality of this qualitative research that is conducted according to the DSR methodology. Empirical research is required to comply to quality standards as credibility, dependability, confirmability, transferability, and reflexivity. (Stenfors et al., 2020). This paragraph explains how the beforementioned research quality concepts are ensured in this research.

3.3.1 Credibility

In a general sense, credibility is about the internal validity of the research. (Rolfe, 2006). It refers to the alignment between the chapters of the research; theory, research question, method, and results. (Stenfors et al., 2020). The credibility of the research is, inter alia, increased by thick description. This research adopted a thorough introduction to the related topics by including extensive literature research. The literature is the scoping and context of the research. Furthermore, multiple interviews are conducted with different types of people in order to gain better understanding of the problem(s). Although, the unstructured case interviews happened earlier in the process they serve as contextual and additional information to the results. Tracy (2010) adds 'triangulation' to the term of credibility. This thesis draws data from different sources via literature and interviews which is in line with the theoretical framing of DSR.

3.3.2 Dependability and confirmability

According to Rolfe (2006) dependability is about the reliability of the research and confirmability is more about the presentation and objectivity. By following the DSR framework the undertaken scientific research steps were methodological and appropriate. The research steps are visually described in such a way that another researcher is able to follow the exact same procedure. Hence, the research is replicable. (Stenfors et al., 2020). In addition, there is a description on how the findings have been found making this research thesis confirmable. The findings from the literature research are mostly all from well-established academic journals with peer reviewed articles, increasing the research's credibility.

3.3.3 Transferability

Transferability is the ability of the findings to be transferred, or generalised, to another environment. (Stenfors et al., 2020). In a general sense, the empirical research is done at KPMG which theoretically means that the research can be exactly replicated at clients of another Big 4 company. Although, the legislation discussed in this research is mainly applicable to European countries and the interviewees were all Dutch, the client cases have a highly international nature. Despite this limitation, the findings could serve as a fundamental beginning of other research projects, making the research generalisable. This research attempts to build the "next layer" rather than filling a "research gap", it is a contribution to the progression of the field instead of plugging a discrete and isolated hole. For that reason, the research is as limited as it is relevant. Another important argument to show the transferability is the fact that DSR combines theory and practice which leads to an outcome that contributes to the scientific body of knowledge as presented in figure 12. (Wieringa, 2014).

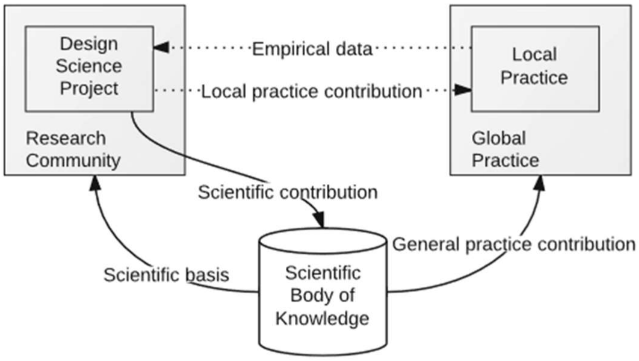


Figure 12. Transferability of Design Science. (Wieringa, 2014)

3.3.4 Reflexivity

Researcher reflexivity is about the objectivity of explorative research and researcher himself. (Stenfors et al., 2020). The setting of the internship was mainly Dutch since both the research as the interviewees have the Dutch nationality and linguistics. As a result, there were no language barriers. The researcher has no professional experience in both AI or auditing. The only experience and potential reasoning bias stem from the knowledge obtained from university and work placements. As for sincerity, this research is conducted in an objective nature since there was not much research available to be biased on. Furthermore, the theoretical framework is widely used in the Information Systems research field and thus ensures sincerity. (Tracy, 2010).

4 Results

This chapter presents the findings that emerged from the literature study and interviews. The result analysis was conducted with a strict focus on the creation of the artefact. In other words, the information from the literature, walkthroughs, and interviews are transformed into requirements for an AI auditing approach. The first paragraph showcases the requirements to the to-be developed artefact that are distilled from the literature. The walkthroughs (unstructured interviews about the cases) are used to orientate upon the empirical matter and formulate, together with the literature, the preliminary model. This preliminary model is presented in paragraph 4.2 and is enriched with the results from empirical research. The interviews with the AI professionals will formulate an additional set of requirements to the artefact. Paragraph 4.3 will showcase a general analysis of the interviews and is continued with paragraph 4.4 which presents the requirements distilled from the interviews with the AI experts.

4.1 Requirements from the literature

This paragraph solely displays the requirements to AI auditing that were derived from the literature. Not all the requirements are included in the artefact or copied into the model 1 on 1. Some coding is used to indicate this. The 'Requirement type' column in table 6 represents preparatory ranking in order to structure the artefact development. The requirements can be a consideration, abbreviated with 'C', which means that this 'finding' should be kept in mind and somehow involved in the model. The artefact will be a overview of steps and not all requirements are directly related or translatable to AI audit activities. However, all requirements are valid points an AI audit should meet. Another coding example is 'Framing (F)' which means that the entire artefact should operate within the set boundaries of that requirement. If a model is not aligned with legislation, it does not make sense to follow. For a "procedure (P)' code it is most likely that this requirement becomes a concord entity within the artefact. The chapters and sources are included to provide an option to find more details.

The coding is no exact science, especially since the development of the model was somewhat iterative. Hence, this list of requirements (and the other list in the next paragraph) should be considered as a structured line of thinking to develop an artefact.

Artefact aspect from literature	Definition	Requirement type	Chapter reference	Sources
Autonomy	Automated processing of systems without human collaboration.	Consideration (C)	Ch. 2.4.1	(Boer et al., 2023)
Complexity	Advanced algorithms which outcome is hard to predict.	C	Ch. 2.4.1	(Boer et al., 2023)
Impact/influence	A system Decisions' effect on the financial statements or ethics	C	Ch. 2.4.1	(Boer et al., 2023)
Legislative alignment	Regulative demands to algorithms or AI systems.	Framing (F)	Ch. 2.4.3	(EU Commission, n.d.-a) (EU Commission, n.d.-b) (Meßmer & Degeling, 2023)
Standardization & formalism	Formalised best practices or approaches to XAI (and AI auditing)	F	Ch. 2.2.2	(Saeed & Omlin, 2023)
Ante- and post-hoc XAI	Intrinsic as external methods to explain model decisions	C & F	Ch. 2.2	(Arrieta et al., 2020) (Nauta et al., 2023)
IT Application controls (ITAC)	Controls that ensure integrity of processed information.	C	Ch. 2.3	Walkthrough
XAI transparency	Algorithmic transparency, Decomposability, and simulatability	F	Ch. 2.2.5	(Arrieta et al., 2020) (Ali et al., 2023)

Artefact aspect from literature	Definition	Requirement type	Chapter reference	Sources
Risk assessments	Each AI model could involve unique risks	Procedure (P)		Walkthrough
Audit process	An AI model should have familiarities with a regular audit process model mutatis mutandis	F	Ch. 2.3	(Rittenberg et al., 2010)
Data governance	Comprehensive metadata support	C	Ch. 2.4.4	(Boer et al., 2023) (Mäntymäki et al., 2022)
Data quality controls	Garbage in – garbage out relates to bias in – bias out.	C	Ch. 2.4.2	(Ali et al., 2023)

Table 6. Overview literature requirements

The requirements are distilled from the literature based on what is written in the literature chapter. All points that were recurring or a point of attention according to researchers were adopted. In a general sense the requirements are a generalised summary of the literature study. Since the unstructured interviews led to the case descriptions in appendix 2 some of the generalised points are included in the overview as well and referred to as 'walkthrough' in the table. The reason to include this with the literature requirements is the fact that both the walkthroughs, or cases, and literature are input to the preliminary model. The preliminary model in paragraph 4.2 was then used to explore further with AI experts which led to the requirements in paragraph 4.4.

The blocks of text on the next page further explain the artefact aspects presented in table 6 above. For a complete description of the artefact aspects the sources are included.

The first artefact aspects stem from the algorithm assurance paper in which the risk areas of an AI system are narrowed down to Autonomy, Complexity, and Influence. Logically reasoning these three might help to categorise an AI or algorithm into a group with techniques or approaches that are a best-practice for those determined characteristics. However, literature does not provide information on how to address an algorithm that is very autonomous, very influential, but not very complex or any other mix between these three. Be that as it may, the characteristics provide knowledge on assessing potential risks and are therefore important to include as a consideration.

Developing an artefact that completely deviates from legislation, guidelines and law makes no sense in the auditing practice. Although, the regulation appears to be quite behind on what is needed it is important to put it into the perspective of the artefact by including it as a frame. Meaning, the artefact should be adaptable to future demands from the law creating an artefact that can be used under stricter rules and laws as well. Moreover, it is likely that the current regulations are not fully developed yet which already insinuates that the to-be developed artefact is not a one-time-fits-all solution.

The literature describes a lack of templates and standards to refer to during audits. Despite that, there are no known records on best practices or formalised ways of working. Meaning, the artefact aspect that is required here refers to the need of formalism and standardization requiring the artefact to be generalisable and scalable to other cases.

Ante- and post-hoc explainability is a requirement in the form of framing and consideration. Depending on the model that is being audited a choice has to be made whether to look into the model or not. Since these are only two techniques it formulates a certain frame to an artefact. The XAI field provides a body of knowledge and thus a frame and consideration.

The walkthroughs indicated that AI auditors are investigating the operating effectiveness of controls by looking at integrity, accuracy, and completeness of the data and system performance. Organisations might have set up events or metrics within or around AI like systems that could be considered to check.

(XAI) Transparency is an artefact aspect since transparency is required for auditors to be able to at least look into the matter. Depending on the level of transparency operations can be done. Hence, XAI transparency determines a working frame.

The walkthroughs made it obvious that risks need to be controlled and assurance is sought on how well the risks are being controlled. In like manner the utilization of AI or algorithms bring along certain risks. Risks could be different and unique compared to regular IT audits. Hence, a risk assessment step is imperative in the to-be developed artefact.

As the literature pointed out several times, AI auditing is not an isolated endeavour and should be part of something bigger. Therefore, the regular audit process serves as a guiding golden thread, or frame, for developing the artefact. It is logical that with some adaptions the regular auditing process fits the premise of AI auditing.

A somewhat underexplored topic in the literature is the organisation of data around AI. Data governance is lightly touched upon in the last chapter. The literature insinuated that looking into data (controls) is one of the options whenever a system is considered of black-box nature. Meta data and comprehensive data support will therefore grow more important since information on the input and output data is needed in auditing AI's. For this reason, it should be considered in the artefact.

Ethical concerns as fairness, privacy, security, bias have a strong relation to data quality. As pointed out in the literature study whenever the dataset already contains faulty data or contains a bias the simple natural law from data management 'garbage in, garbage out' changes into bias in, bias out. Consequently, it is pivotal to consider the quality of the data and how it is maintained in the artefact.

4.2 Preliminary model

The first set of artefact aspects formulate the initial insights in how to navigate the AI auditing industry to a potential approach, or best practice. The literature and walkthroughs of the cases in appendix 2 provided a first line of thinking on how AI auditing could be characterised. This paragraph introduces the preliminary model that has been made according to the required aspects in table 6. The main goal of the preliminary model is to provide understanding a mutual ground between researcher and interviewee.

Before displaying the preliminary model, some background reasoning needs to be established to understand how the preliminary model originated. Figure 13 is a graphical representation on the type of cases that were made available by KPMG for this research project. CO1 is a case in which an AI model was investigated thoroughly with the objective to explore whether the AI model was reliable enough to lean on assurance wise. CO2 and CO3 are larger corporations which encountered new legislation; DSA and DMA. Both types are considered to be part of auditing with a strong relation to AI.

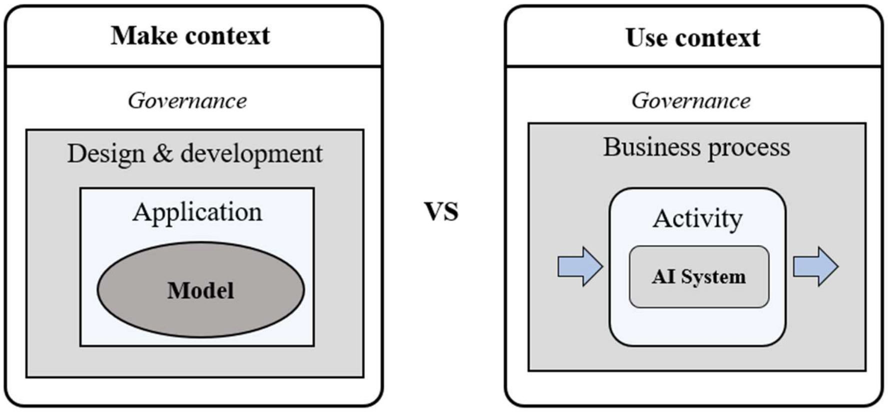


Figure 13. Context behind AI auditing

Figure 13 depicts two types of contexts within the auditing field. The ‘Make context’ refers to an audit in which the focus relies solely on the AI model and its direct environment. It could be part of an overarching application where controls are at place. Furthermore, around a model’s design and development there are a lot of procedures and documentation which might ‘say something’ about the application of the model. Theoretically, governance mechanisms set up are around this, but considered out of scope for this research. The other option is the ‘Use context’ in which the performance of an AI is audited in the broader context. The environment is investigated more than the model itself. An AI System is part of a certain activity or ‘task’ in the business context. A series of activities make a business process.

This background led to the preliminary model presented in figure 14 below. The preliminary model is a quadrant diagram that describes the type of AI audit on the axis of the complexity of an AI model (or algorithm) and the maturity of the organisation.

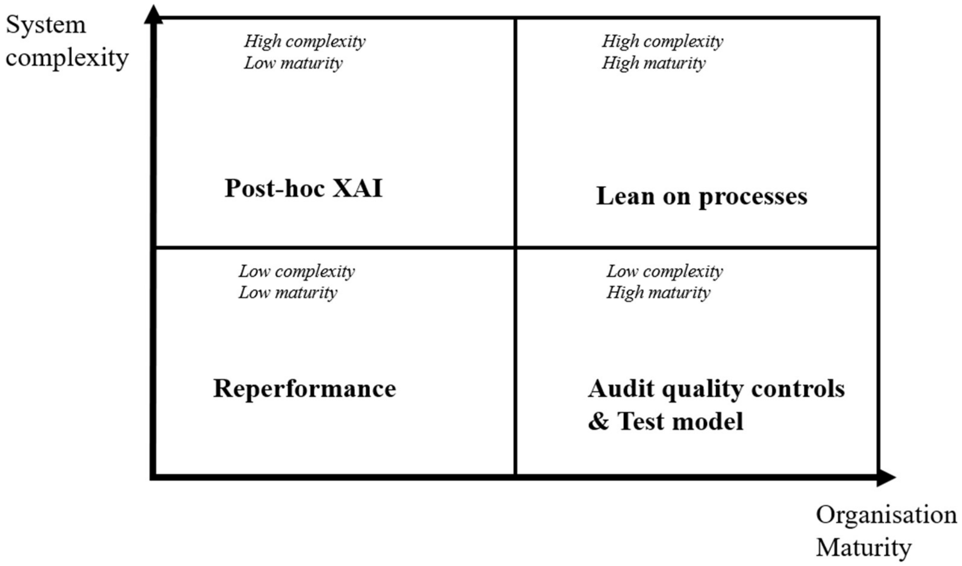


Figure 14. The Preliminary model

The vertical axis represents the 'System complexity' and is about 'how much' of a black-box the AI system is. It is likely that a multi-layered ANN is harder to investigate than a ML model with a clear set of business rules. This was an important aspect derived from the literature. Furthermore, as the walkthroughs made clear, there is a big difference in how organisations are organised around the models. The 'Use context' in figure 13 could have very predictable business processes, management layers that are in control, audit controls at place, and proper documentation available from operations making the organisation quite mature. An example are the Very Large Online Platform corporations that qualify for the DSA / DMA. These VLOPs require mature processes and complex AI systems to perform complex tasks. With this logic, the upper right quadrant is selected.

In the upper right quadrant, and auditor could opt to completely rely on the processes (around the AI model) in the Make and Use context. As the literature pointed out, the AI audit is bigger than the AI entity itself and the cases show that a lot of environmental controls can be utilized. Controls as implemented SCRUM reviews, developmental documentation, and 4-

eyes-principles could be sufficient to provide some degree of assurance. Obviously, it is hard to put it down as black and white, but the quadrant is mainly meant to help structuring.

In the bottom-left quadrant, organisations are likely new to using AI models, resulting in under validated maintenance and control mechanisms. The AI model in the context is not too complex so it allows for deep inspection. The organisational maturity lacks defined reliability. The auditor could opt for reperforming activities to replicate the training phase of the model. Another option is to statistically test the model, such as CO1, for assurance purposes.

For cases in which the maturity is high enough to rely on the processes and environment around the AI entity, but the entity is not too complex allows to check quality controls that are build-in and set up in the environment. The model can be tested as well. For example, an independent test to audit the data and performance of the AI system on the datasets. Optionally the code could be reviewed as well, which might provide information on the integrity and efficiency of the model.

The upper left quadrant is harder to define since an AI model is comparatively complex with an organisation or environment of lower maturity. It is more likely advisory projects around assurance are initiated here since the only possibility to make a statement on the model is by applying the XAI post-hoc techniques described in the literature study. Although, solely providing explanations are no guarantee for a thorough compliance audit or even an assurance project.

A noteworthy addition to the preliminary model came from a validation session in which the director of the RAI department questioned the difference between the meaning of controls and processes. The presence of control mechanisms determines the maturity, but also the maturity of the business processes. The more predictable an outcome of a business process is, with clearly defined actors and documentation, the more it could be considered as mature. Processes are the broader perspective where controls are monitoring points.

4.3 Overview of the interview findings

This paragraph provides a summary on the findings from the 6 interviews with AI experts. The summaries from each individual interviewee are attached in appendix 4. This overarching summary presents the general findings and is an interlude to the next paragraph in which the empirical findings are presented in the form of required artefact aspects.

As 5 out of 6 interviewees were either part or had a strong role within the Responsible AI department of KPMG a clarification might be needed. This department serves primarily as an advisory body. Technical expertise is leveraged during audits at cases like CO1. However, after the release of the DSA and DMA the department is getting more involved in audits. As interviewee A indicated, an important role from his expertise is in the initial phases of the (AI) audit, in which the focus is put on understanding the systems and involved processes. Interviewee E emphasized the aim of IT assurance, which is providing confidence in the reliability, security, and compliance of IT systems, including AI systems under the DMA and DSA. Looking at CO1 and the answers from interviewee A the goals are none but the same. Interviewee E made a remark that for defining the audit object, an underlying framework, referring to ISO standards or other common frameworks, is often used. Furthermore, the purpose and control of the measures should be defined at the start. Interviewee A, B, C, and D talked about the importance of understanding every unique case and regard them as such. However, interviewee A indicated that in the end, due to professional experience, there are always a select number of risks that need to be controlled or audited. In a general sense, for every case one could select the risks and the additional metrics that are pivotal for that specific case. Arguably every case requires its own assessments and investigations.

Challenges in ensuring transparency and explainability remain, mainly because they are hard to define. In addition, interviewee A, B, C, and D indicated there are no standards to compare measurable outcomes, such as precision, recall, and accuracy. Interviewee A sees them more as a tool and not as a quantifiable means that declares a model as fair or ethical. Interviewee B outlines that, even for audits with a standard, the audit process heavily relies on professional judgement. The representative for the CO1 case stated during the walkthrough that AI auditing remains a "*format-less activity in which it is a best-effort of all people involved*". Another problem, due to the lack of legislation and standards, is the recurring topic of profiling in the DSA and DMA related interviews. Interviewee B illustrated that the theoretical implementation of profiling differs in practice compared to legislative definitions.

Interviewee B and E both underline that under de DSA, in which areas as profiling and transparency are dictated, the areas of focus are often seen as vague and lacking detailed guidelines. Interviewees C, D, and E add that the demands on the extent of explanations, that are actually measurable, on data usage and decision-making processes in AI systems vary a lot per case.

Despite challenges brought up by the DSA / DMA, the cases differ amongst each other as well. Interviewee C indicated that for CO3 the process rather than the technology was audited. As interviewee D adds, the maturity of the organization's processes played a significant role in determining the effectiveness of the audit, underlining the fact that the AI entity is not the audit object, but simply part of it. Interviewee C explained that it might be easier to immediately dive into the model, when it turns out to be a simple and uncluttered model consisting of prepared data. Theoretically, information about controls and processes might then not be necessary.

According to interviewee C it could be considered as a sufficient control if there is proof that a review has taken place in the AI lifecycle. However, C and D add that receiving a statement from a client that says '*we are considering the concern'* is not enough, nor purely focusing on the provided controls. An AI audit should go further than that. Interviewee D critiques that an AI audit can only be fully covered when a multidisciplinary team, consisting of all stakeholders and domain experts, is involved. However, this would be impossible for an audit. Interviewee B reasoned that the maturity of an organisation might not always be that beneficial to AI audits because currently there is an ongoing trend of involving the term AI into everything and whether it is truly a specialised system or not, clients appear to be cautious in providing detailed information.

Interviewees A, B, C, and D admitted that the AI audit process has a lot of similarities with the regular audit process. Interviewee C provided the insight that there might be considerable overlap between the initial AI audit phases and CRISP-DM since looking into the data is imperative in AI auditing. Interviewee A, B, and D described that in hindsight they unconsciously walked through the 'understanding' steps. Interviewee E reasoned from his background and added the importance of risk analysis and broader perspectives as well. From the cases CO2 and CO3 the classic IT audit process change management and identity and access management remain relevant to AI audits as well.

In closure, the interviewees provided a lot of similar answers and emphasize the evolving nature of the AI auditing field. Which leads to a need for adaptive methodologies with an incorporation of both technical and ethical aspects in order to look into more than 'just' the AI object.

4.4 Requirements from the interviews

Table 7 below follows the same structure as the artefact aspect list derived from the literature study presented in table 6. The overview in table 7 presents the required artefact aspects that are derived from the empirical results achieved through interviews with AI (audit) professionals. If the interviewees reasoned from a case that is included in this research a designated column refers to that specific case. Sometimes the interviewees reasoned from other experiences and their general professionalism. In that scenario, the column is left empty. In order to structure the empirical artefact aspect table below the same 'Requirement type' classifications are used. Furthermore, the interviewees related to the aspect are referred via a designated column.

Empirical artefact aspects	Definition	Requirement type	Interviewee reference	Case reference
Intake	Every AI audit starts with an initial exploration to investigate the scope and objective.	Procedure (P)	A, B, C, D, E	CO2
Audit objective	Like regular audits the objective and purpose need to be determined	P	A, C, E	CO1
Audit entity	What the audit is about, the AI model, the processes, the controls.	P	A, E	CO1
Maturity	How mature processes are and the presence of monitoring and controls.	Consideration (C)	A, B, C, D	-
Complexity	Whether the model can be investigated or should be considered black-box.	C	A, B, C, D, E	-

Empirical artefact aspects	**Definition**	**Requirement type**	**Interviewee reference**	**Case reference**
Continuous evaluation	What worked what did not work.	P	A, D, E	-
Professional judgement	The model must allow for professionals to apply judgement from their own expertise.	C	A, B D, F	CO1, CO3
Limited standardization	In all cases other elements are encountered, requiring a somewhat customized approach on the detailed levels	Framing (F)	A, D, F	CO1
Ethical risk framework	Every AI audit has their own risks and ethical concerns.	P, C & F	A, F	CO1
Regulatory compliance	Alignment with compliance audits such as DSA / DMA.	F	B, C, D	CO2, CO3
Controls *(or pillars of control)*	KPMG identified five pillars; *reliability, resilience, explainability, accountability, and fairness.*	P & F	A, C, F	CO1, CO2
Documentation	Comprehensive design documentation and reports on performance	P	A, C, D, F	CO1, CO3
Stakeholders	Human in the loop for validation or review through administrative separation of functions.	P & C	A, F	CO1, CO2
CRISP-DM	The initial phases of CRISP-DM provide understanding which is very important for every individual case	P	C, *(A, B, D)*	CO2
AI-Lifecycle	Ethical considerations need to be incorporated throughout the lifecycle	P	A, B, C, D, E	-

Table 7. Overview empirical requirements

The interviews pointed out that there are significant similarities between AI auditing and regular (IT) audits. A recurring topic was the initial phase of any type of audit in which the auditor and the client determine the scope, plan, objective, and what is being audited. Meaning, some exploration and understanding from both sides has to be reached in order to determine if a black-box should be explored and/or the environment, so the processes and controls, should be included. Since these are pivotal steps in the audit process the first three requirements of the table above are procedural.

Interviewee C started the conversation about maturity of controls, processes, and documentation which happened to become a recurring topic. Maturity is twofold. In a way maturity is an enabler of being able to support the audit and rely on the processes when there are built-in reviews, validation mechanisms, and a certain predictability of the outcome of the processes and all actors involved. Furthermore, maturity of the controls itself is an important point to consider since an auditor cannot rely on a control that performs well, but is not using the right data. Therefore, the reliability of the data is audited as well which is based on data input, data integrity, data accuracy, and data completeness. Auditors should consider this during audit and assurance projects.

Complexity is a topic that is always discussed during interviews. All interviewees (A, B, C, D, E, F) admit that one of the bigger differences to IT auditing is the required expertise to grasp AI systems in order to fully understand the risks they might bring. According to interviewee A and F the level of complexity is a crucial aspect in choosing the audit strategy, method, and techniques.

The artefact aspect of continuous evaluation is not explicitly mentioned in the interviews, but is often implied. Interviewee D highlighted the need for adaptive methodologies because of the evolving nature of AI, and interviewee E mentioned the evolving regulations. Interviewee A added that due to the complexity of AI systems, variability in AI audits is imminent. Hence, the importance of ongoing refinement is required to be procedural.

Leveraging the professionalism and expertise of the AI auditors is a highly required aspect, that should be considered according to interviewee D and B. A standardized method cannot cover the complex technical and ethical issues since every case is considered as unique with different (types of) risks. Interviewee F acknowledged this need for flexibility to address the varying situations, systems, and clients.

From the CO1 walkthrough and interviewee A it was understood that a certain formalised guidelines is used within KPMG. This guideline is a knowledge base that includes all types of AI risks that could be considered during an audit. This guideline was basically a 'shopping-list' in which auditors could shop the risks and belonging ethical concerns and measures per distinct case. Due to strict disclosure this list could not be included or copied into this research document. However, the artefact creation should include this requirement as a procedural step, frame and consideration to be not only aligned with the previous mentioned requirements about customization, but also to acknowledge the fact that an ethical frame determines part of the scope of an AI audit.

A lot of the interviewees that were involved in DSA / DMA audits (B, C, D, E) comment that a large part of the AI audit scope and objective is dependent on what information the client delivers and not necessarily what the law and legislative bodies prescribe. On one hand regulatory compliance is of importance on the other hand the articles allow for interpretation. The artefact should adopt a frame that aligns mainly with the headlines of regulation.

Interviewees B and C noted that CO2 and CO3 differ in the presence of controls, such as IT application controls, change management, and identity & access management. Investigating these types of controls can provide sufficient information to provide a reasonable form of assurance according to interviewee E. Thereupon, evidence of controls is imperative to AI audits and say something about the performance, integrity, and working of not only the AI model, but also the environment according to interviewees B, E, and F. Gathering the evidence is a procedural step in any audit approach. The ethical pillars of KPMG provide a certain frame to the AI audit.

Another requirement mentioned is documentation, referring to reports, policy, design notations, and other kinds of records that provide insights into performance, design and ways of working. Investigating these elements are a pivotal procedural step to make statements about transparency (B, C, D) and explainability of systems (A).

Interviewee F mentioned the element people, or stakeholders, as validation employees or a separated and aside department that 'controls' or monitors the delivered work on and from AI models. The artefact in the next chapter should include this requirement as a consideration in terms of validating its presence at a client, but also include it as a procedural inspection step. The separation of roles is severely stressed by interviewee F.

As explained in paragraph 4.1.2, the initial steps of CRISP-DM are recognised by interviewee C since AI's strong relation with data and the fact that auditing requires some exploration to define and scope the audit. Especially for AI audits, in which (technical) expertise is essential, steps that initiate 'understanding' should be included as a procedural step in the artefact.

Interviewee C and D accentuated the need to understand the AI system's lifecycle stages such as design, development, and implementation. Interviewee A and F added that these activities contain a lot of information that provide, for example, how ethical an AI is being developed. Some systems include ethics-by-design methodologies which is solid evidence and documentation utilize in assurance statements. Auditing the AI lifecycle becomes a required procedural artefact aspect.

5 Developed artefact

Clear insights are obtained from the requirements, based on the literature, walkthroughs, and interviews, to develop a formalised approach to AI auditing. Figure 15 on the last page of this chapter displays the developed AI auditing artefact. The presented approach is not a comprehensive standard. Nevertheless, the graphic presents a deep understanding of the current practice of AI auditing and is a first formalised process model. The following paragraphs explain the phases of the AI auditing process.

5.1.1 Phase 1. Understanding

The first part of the AI auditing process is iterative since thorough understanding of the situation is required since AI audits are challenged by the complexity and novelty of this field. This paragraph explains the circle, that is based on CRISP-DM per activity.

Step 1 Business Understanding is about understanding the business and the contribution of the system. The primary goal of this step is to gain comprehension on the objectives and requirements of the audit project from the business perspective. Business understanding results in audit alignment with the organisation's goals and eventual applicable regulation. Activities are often intake interviews, determining the audit objectives, and assessing the business situation.

Step 2 Data Understanding aims to achieve a thorough understanding of all the data related to the audit objectives. In terms of an AI audit, a deep understanding of all the datasets that form the input of the AI model is required. This step involves the collection, description, exploration, and verification of all the data that will be audited. The data cannot be understood without the business context. Thus, sometimes the auditor needs to refer back to the business.

Step 3 Model Understanding is about investigating the type of AI model in order to comprehend its architecture, behaviour, and performance in the business setting. The model is judged on inter alia, its complexity, autonomy, and influence on the business. As depicted in figure 15, Business, Data, and Model Understanding go hand in hand and require information from each other. Through this iterative interaction the business objectives can be refined, the data needs can be reassessed and improved, and the model analysis can be deepened further.

Step 4 Risk Assessment is a result of the preceding steps and can be executed after the business context, utilised data, and model purpose is understood. Through the previous understanding steps an estimation of the risks that apply for this particular situation can be made. As learnt from the cases some recommender systems do not impose great risks of bias whilst others do.

5.1.2 Phase 2. Research object

The research object phase determines the focus of the AI audit and refers back to the preliminary model. The research object can be the Make Context and/or the Use Context. This phase is configured by determining **step 5 Ethical frame** and eventually **step 6 the audit planning**. The ethical frame part might misleadingly imply that an auditor defines the ethicalities for the client case. However, the ethical frame mainly means that a selected number of risks are related to this specific AI audit and require investigation. All the other existing AI risks and ethical concerns are then considered out of scope for this particular audit. Upon the ethical frame the audit planning can be made and presented, in consultation with the client. The audit planning is a common step in regular audits. The difference with AI audits is what precursory to the audit planning, the initiation simply requires more (unique) knowledge.

5.1.3 Phase 3. Environment and/or Model

The third phase of the AI audit process is divided by the Make Context and Use Context. In simplified audit terms, the Environment and the Model. An auditor could opt to either investigate the one and/or the other.

Step 7a Environment focusses more on the environment, controls and monitoring mechanisms around the AI model. For example, the AI lifecycle can be audited by looking into whether a development methodology was utilized in which for example (sprint) reviews took place. Through the lifecycle documentation is recorded. Like IT audits controls as change management and identity and access management contain data that is valuable to look into in the same manner as IT audits are addressed. An important facet is data explainability which refers to clearance and transparency of data usage, insights in how the model is trained and it includes data quality controls.

Step 7b Model puts the focus on auditing the model through methods as inspection or reperformance, where activities of the AI lifecycle are reperformed by the auditors. Another option is independent testing, in alignment with data explainability, to test the model output with another external input dataset from the auditors. An auditor could opt to review the code and investigate the main variables that are used for profiling or decision making. However, the sensibility and effectivity is broadly discussed. If the model is not mathematically approachable some functionality could be replicated in line with the selected ethical frame for the audit. The literature of XAI provided a lot of post-hoc techniques that provide insights in the decisions of a black-box and support understanding. It is seen as part of the assurance, but not something that could be assured itself let be an argument of assurance. The last technique is to apply statistical calculations to investigate how much the AI model has missed which it should not have missed. This provides relative insights in reliability. Without a standard or agreed upon quantitative output, the statistical metrics as recall cannot add more than so-called 'reasonable assurance'.

5.1.4 Phase 4. Assurance

After validation interviews, appendix 5, the evidence steps are split accordingly.
Firstly **step 8 Evidence on AI Model** is to gather the evidence from the AI model audit. When the evidence is found negative, the auditor is obliged to cycle back one step to gather more reliable environmental controls. As pointed out, the controls directly related to the AI model can cover risks together with controls that were set up in a more business-oriented manner. In that case the evidence of the AI model controls can be considered as sufficient coverage. From an IT audit perspective an IT system (read AI system) can cover a risk point. Hence, the evidence on these controls precedes the evidence that are more general and set up around the environment. This is **step 9 Evidence on Risk points**, which also covers the evidence gathered from step 7a Environment. **Step 10 Review & Reporting** is to present the findings in a typical assurance report. Step 10 is directly derived from the literature on auditing as well. The reason to emphasize reviewing is to denote lessons learnt in the novel AI auditing field.

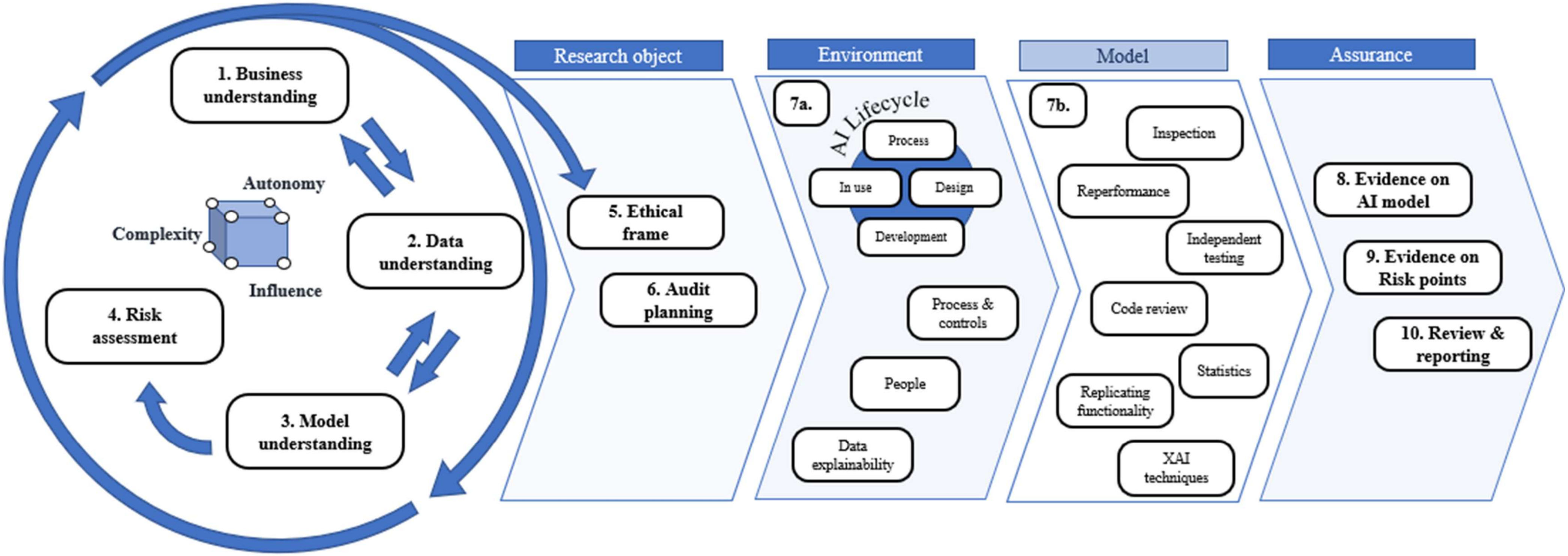


Figure 15. The AI Auditing process

6 Discussion

The research was initiated with a research question on how AI auditors can assure transparency and explainability in the practice of auditing AI-systems. The aim was to reason from XAI theory and investigate the current practice empirically. This chapter discusses the findings, how they relate to the theory, the limitations of the research itself, and lastly provides recommendations to extend the research.

6.1 Interpretation of findings

This paragraph briefly indicates how the empirical results, and part of the literature, address the problem statement and research questions.

A large part from the research question was answered through the literature study and points out that the difference between explainability and transparency is nuanced, but significant. Where transparency is about the disclosure of internal workings, is explainability about providing clarity on the outcomes in an understandable and interpretable way. The main challenge for AI auditors is the complexity and inscrutability of the encountered AI systems. Unlike IT audits there are no standards, best practices or known approaches to address the AI audit. Standardisation and a formalised way of working is missing. This requires continuous advancements in XAI research and AI auditing research. The literature defines certain frameworks that help define an algorithm as responsible or not. However, the empirical results point out that frameworks are not really used in the auditing practice which leads to the fact that governance is more of a record of lessons learnt.

The main research question is answered solely by the artefact, the AI auditing process, since it showcases ‘how’ AI auditing works in a generalised and abstract way. Be that as it may, there is still a difference between how XAI understands transparency and explainability and how the auditing field looks at the concepts. Transparency is regarded as a prerequisite so an auditor is able to investigate and the legislation defines it in a more communicative way. Explainability is not something that is really audited or assured, it is stressed as important to have clarity on model outputs. It is rather a part of an audit then the focus.

This research is the first formalisation of AI auditing, within KPMG, and possibly a prelude to more scientific research on AI auditing since topics as algorithm assurance and the combination of audits and AI is not yet considered a scientific domain.

6.2 Comparison theory and practice

Theoretically audits are to be held up to standards, guidelines, frameworks, and prescribed laws. As found in the literature study, chapter 2, there are not yet formalised standards available to set up AI compliance audits. Interviewees A, B, C, D, E, and F fully underscore this finding. However, it should be noted that currently in practice efforts are made to do AI auditing with, or without strict guidelines. Interviewee A outlined that for CO1 an AI audit could be part of the general (financial) statement audit, in which is controlled whether companies' statements are valid. 2 findings can be made from this answer and walkthrough observation. As the algorithm assurance literature indicated (paragraph 2.4.1), the audit entity is not solely the AI system. Meaning more concepts require investigations. The other finding is that the audit at CO1 was completely based upon the organisation's statements instead of a law or standard. Even the interviewees that were involved in a DSA / DMA audit at CO2 and CO3 (B, C, and D) admitted that with the current provided law there are a lot of discussions on definitions and compliance matters which led to providing assurance on statements of the organisation itself. Hence, an audit checks either whether an organisation is complaint to official statements or statements of their own. This finding is currently pivotal for AI audits. Figure 16 below showcases how an (AI) audit draws information from the two possible areas.

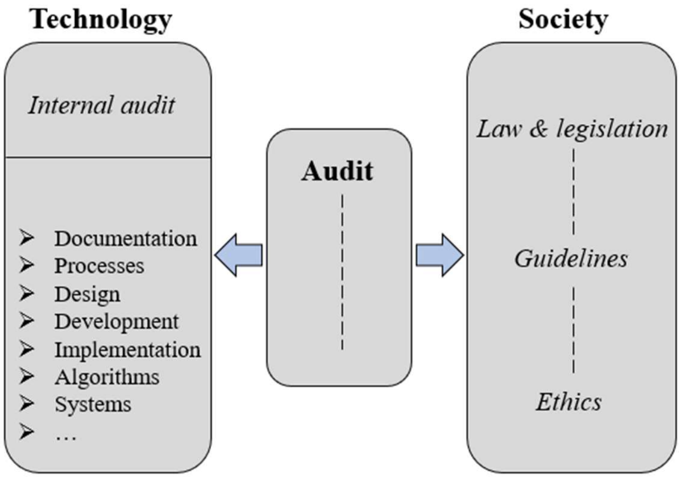


Figure 16. Positioning the AI audit

Interviewee A stated that an AI audit follows a general audit-like procedure with an intake that determines the objective and availability of required intel. An organisation's statement could be held against the information from internal audits, documentation, processes, the development cycle (of an AI), AI entities themselves, and (IT) control mechanisms. As the field of XAI emerges from ethical concerns the law and legislative bodies formulate official guidelines and standards. However, it appears that there is a lot of knowledge from the scientific domain which is not yet included in auditable terms. Meaning, quantitative outcomes such as recall, precision, and accuracy (CO1, interviewee A) cannot be really part of assurance. Another consequence is that inspecting the black-box is not always as sensible. Especially for AI audits that follow a DSA or DMA like frame. It is made clear from all interviews, and observations that, as how interviewee C it puts; the process is audited more than the technology.

6.3 Limitations

The main limitation of this research is that the AI Auditing Process, in figure 15, misses the justification from DSR since it could not be tested through an implementation and a thorough case study. This limitation is compensated by two validation interviews with IT auditors to test the artefact's sensibility.

The novelty of this research is a limitation since topics as algorithm assurance and the combination of audits and AI is not extensively researched. Instead, angels as AI research, XAI, and auditing provided a baseline.

The preliminary model and eventually the AI auditing process (artefact) is based on three cases and 6 expert interviews. This research was limited to time constraints and the resources that could be provided by the research organisation.

Along with the previous limitation, this research was solely based on findings found within KPMG. Even though the methodology and its outcome allow for transferability, there is still a possible company bias.

Although, the cases represent international organisations, the organisation of the thesis internship KPMG Netherlands. As a result, only Dutch people were interviewed. Although, the orientation of KPMG Netherlands is international, the research was limited to only one nationality regarding the interviewees.

This research attempts to build the next layer rather than filling a knowledge gap, it is a contribution to the progression of the field instead of plugging a discrete and isolated hole. For that reason, the research is as limited as it is relevant due to its explorative character.

6.4 Recommendations

This paragraph presents some future research recommendations to follow up on this research. The long-term goal of a formalised AI audit framework is to enhance the trust in the use of AI through audits and assurance. However, more research on this topic is highly recommended.

Some of the limitations are a next step in this research. Hence, a future research direction could be to test the AI audit process in practice and conduct a case study on it.

As stated in the introduction on this paragraph, longitudinal studies are interesting to discover whether the trust in the use of AI systems is affected through audits and assurance that are based on a formalised approach.

As the literature pointed out, there is no coherence on the terminology of Transparency and Explainablity (and interpretability). Recording cases and combining theory and practice might help to distinguish the terms. Furthermore, the field of XAI could study broader aspects than these model characteristics in order to improve transparency and explainability techniques.

Explainability in the review and reporting stage of the AI audit could be a future research direction since a lot of expertise is required from the auditor. Determining what needs to included and what aspects of the AI audit require explainability and to whom in an assurance report is a recommended research area, according to a validation interview.

More thorough research on the topics presented in this research is important in order to improve regulatory compliance, but also improve legislation itself. The AI auditing process is not set in stone yet either, it requires more testing and research in order to be adopted in broader settings. Furthermore, the differences in types of AI models (Like ANN or ML) might demand different approaches after thorough testing. Continuous model improvement and research is recommended.

7 Conclusion

The rapid evolvements in AI fundamentally changed the field of auditing. This thesis explored the integration of auditing AI models with an emphasis on the need for transparency and explainability, defined by the research domain of XAI. Through extensive literature research, empirical case studies and interviews with AI audit experts a comprehensive understanding of AI auditing was gained and resulted in a first formalised AI Audit Process.

The auditing field has undergone significant changes through technological advancements. It was found that for AI auditing the need to ensure the accuracy, reliability, completeness, integrity, accountability, correctness, and ethicalities remain the same. Hence, the development of the AI Auditing Process is supported by the regular auditing process since it is an extension. The main contribution of the AI Auditing Process is that AI auditors now have a formalised ‘standard’ to refer to, which could potentially enhance assurance and create trust.

The field of XAI provides perspectives on transparency and explainability. Through the lens of AI Auditing both terms are not necessarily audited. Transparency is seen as an enabler / prerequisite for the audit and is about openness and communication of the model decisions and involved processes in the environment. AI auditors conclude Explainability as an instrument. XAI defines techniques for making the outcomes of AI models understandable. Moreover, Explainability is more than a characteristic of a system, but involves a broader perspective because the model itself is not the only focus during AI audits. This conclusion enriches the scientific field of XAI as well.

This research contributes to the growing body of knowledge on AI auditing, algorithm assurance and XAI and provided a practical application of AI auditing and a foundation for future research. Continuing research and collaboration between researchers, AI auditors, and public administration will become crucial to shape AI auditing.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Ser, J. Del, Díaz-Rodríguez, N., & Herrera, F. (2023a). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Elsevier: Information Fusion*, *99*. https://doi.org/10.1016/j.inffus.2023.101805

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Ser, J. Del, Díaz-Rodríguez, N., & Herrera, F. (2023b). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Elsevier: Information Fusion*, *99*. https://doi.org/10.1016/j.inffus.2023.101805

Alles, M. G., Dai, J., & Vasarhelyi, M. A. (2021). Reporting 4.0: Business reporting for the age of mass customization. In *Journal of Emerging Technologies in Accounting* (Vol. 18, Issue 1). American Accounting Association. https://doi.org/10.2308/JETA-10764

Almufadda, G., & Almezeini, N. A. (2022). Artificial Intelligence Applications in the Auditing Profession: A Literature Review. *Journal of Emerging Technologies in Accounting*, *19*(2), 29–42. https://doi.org/10.2308/JETA-2020-083

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Articial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inus.2019.12.012

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*. https://doi.org/10.1016/j.infsof.2023.107197

Baloglu, O., Latifi, S. Q., & Nazha, A. (2022). What is machine learning? *Laaketieteellinen Tiedekuntakirjasto. Protected by Copyright. on March*, *107*, 2024. https://doi.org/10.1136/archdischild-2020-319415

Bennetot, A., Franchi, G., Ser, J. Del, Chatila, R., & Díaz-Rodríguez, N. (2022). Greybox XAI: A Neural-Symbolic learning framework to produce interpretable predictions for image classification. *Knowledge-Based Systems*, *258*. https://doi.org/10.1016/J.KNOSYS.2022.109947

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). SPECIAL ISSUE: MANAGING AI MANAGING ARTIFICIAL INTELLIGENCE 1. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. In *Internet Research* (Vol. 33, Issue 7, pp. 133–167). Emerald Publishing. https://doi.org/10.1108/INTR-01-2022-0042

Boer, A., de Beer, L., & van Praat, F. (2023). Algorithm Assurance: Auditing Applications of Artificial Intelligence. In *Advanced Digital Auditing* (pp. 149–183). Springer, Cham. https://doi.org/10.1007/978-3-031-11089-4_7

Breakspear, A. (2013). A New Definition of Intelligence. *Intelligence and National Security*, *28*(5), 678–693. https://doi.org/10.1080/02684527.2012.699285

Burgoyne, A. P., Mashburn, C. A., Tsukahara, J. S., Hambrick, D. Z., & Engle, R. W. (2021). *Understanding the relationship between rationality and intelligence: a latent-variable approach*. https://doi.org/10.1080/13546783.2021.2008003

Cambridge Dictionary. (2024). *The word Intelligence*. Retrieved March 3, 2024, from: https://dictionary.cambridge.org/dictionary/english/intelligence

Casper, S., Ezell, C., Siegmann, C., Curtis, T. L., Csail, M., Bucknall, B., Haupt, A., Wei, K., Law School, H., Scheurer, J., Research, A., Hobbhahn, M., Sharkey, L., Von Hagen, M., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., … Hadfield-Menell, D. (2024). *Black-Box Access is Insufficient for Rigorous AI Audits*. https://doi.org/https://doi.org/10.48550/arxiv.2401.14446

Chen, Y. W., Chien, S. Y., & Yu, F. (2023). An overview of XAI Algorithms. *IEEE CACS*. https://doi.org/10.1109/CACS60074.2023.10326174

Colom, R., Karama, S., Jung, R. E., & Haier, R. J. (2010). Human intelligence and brain networks. *Dialogues in Clinical Neuroscience*, *12*(4), 489–501. https://doi.org/10.31887/DCNS.2010.12.4/rcolom

Curasi, C. F. (2001). A critical exploration of face-to-face interviewing vs. computer-mediated interviewing. *International Journal of Market Research*, *43*(4), 361–375. https://doi.org/10.1177/147078530104300402

Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Pappas, I., & Mäntymäki, M. (2021). *Responsible AI and Analytics for an Ethical and Inclusive Digitized Society* (D. Dennehy, A. Griva, N. Pouloudi, Y. K. Dwivedi, I. Pappas, & M. Mäntymäki, Eds.; Vol. 12896). Springer International Publishing. https://doi.org/10.1007/978-3-030-85447-8

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6

Estep, C., Griffith, E. E., & MacKenzie, N. L. (2023). How do financial executives respond to the use of artificial intelligence in financial reporting and auditing? *Review of Accounting Studies*. https://doi.org/10.1007/s11142-023-09771-y

EU Commission. (n.d.-a). *The Digital Markets Act: ensuring fair and open digital markets - European Commission*. Retrieved April 12, 2024, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en

EU Commission. (n.d.-b). *The Digital Services Act package | Shaping Europe's digital future*. Retrieved April 12, 2024, from https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Fjeld, C., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. *Harvard Library*. https://dash.harvard.edu/handle/1/42160420

Garousi, V., Bauer, S., & Felderer, M. (2020). NLP-assisted software testing: A systematic mapping of the literature. *Information and Software Technology*, *126*, 106321. https://doi.org/10.1016/j.infsof.2020.106321

Gillespie, N., Curtis, C., Bianchi, R., Akbari, A., & Fentener van Vlissingen, R. (2020). *Achieving Trustworthy AI: A Model for Trustworthy Artificial Intelligence*. The University of Queensland and KPMG. https://doi.org/10.14264/CA0819D

Gillespie, N., Lockey, S., Curtis, C., Fisher, R., Gobbi, L., Malta, J., & Mabbot, J. (2023). *Trust in Artificial Intelligence: A global study*. https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2023/09/trust-in-ai-global-study-2023.pdf

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly*, *37*(2), 337–355. https://www.jstor.org/stable/43825912

Haenlein, M., & Kaplan, A. (2019). A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925/FORMAT/EPUB

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, *69*, S36–S40. https://doi.org/10.1016/J.METABOL.2017.01.011

Harvard University. (n.d.). *What is an Information Technology (IT) audit?* . Harvard University. Retrieved April 9, 2024, from https://rmas.fad.harvard.edu/faq/what-does-information-systems-audit-entail

Hevner, A., & Park, J. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105. https://www.researchgate.net/publication/201168946

HLEG AI. (2019, April 8). *Ethics Guidelines for Trustworthy AI*. European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Huang, M. H., Rust, R., & Maksimovic, V. (2019). The Feeling Economy: Managing in the Next Generation of Artificial Intelligence (AI). *California Management Review*, *61*(4), 43–65. https://doi.org/10.1177/0008125619863436/FORMAT/EPUB

Johannesson, P., & Perjons, E. (2014). An introduction to design science. In *An Introduction to Design Science* (1st ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-10632-8/COVER

Jones, N. (2023). The world's week on AI safety: powerful computing efforts launched to boost research. *Nature*, *623*(7986), 229–230. https://doi.org/10.1038/D41586-023-03472-X

Jung, A. (2022). *Machine Learning: The Basics*. Springer Nature Singapore. https://doi.org/10.1007/978-981-16-8193-6

Kaplan, A. (2021). Artificial Intelligence (AI): When Humans and Machines Might Have to Coexist. In *AI for Everyone? Critical Perspectives* (pp. 21–32). University of Westminster Press. https://doi.org/10.16997/BOOK55.B/

Koshiyama, A., Kazim, E., & Treleaven, P. (2022). Algorithm Auditing: Managing the Legal, Ethical, and Technological Risks of Artificial Intelligence, Machine Learning, and Associated Algorithms. *IEEE Computer Society*, *55*(04), 40–50. https://doi.org/10.1109/MC.2021.3067225

Laato, S., Tiainen, M., Islam, A. K. M. N., & Kauppakorkeakoulu, T. (2022). How to explain AI systems to end users: a systematic literature review and research agenda. *Internet Research*, *32*(7), 1066–2243. https://doi.org/10.1108/INTR-08-2021-0600

Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Elsevier | Business Horizons*, *63*(2), 157–170. https://doi.org/10.1016/j.bushor.2019.10.005

Liu, Y., Huang, Y., & Cai, Z. (2023). AED: An black-box NLP classifier model attacker. *Neurocomputing*, *550*, 126489. https://doi.org/10.1016/J.NEUCOM.2023.126489

Lombardi, D., Bloch, R., & Vasarhelyi, M. (2014). The Future of Audit. *Article in Journal of Information Systems and Technology Management*. https://doi.org/10.1590/S1807-17752014000100002

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advancec in Neural Informatoin Processing Systems*, 4765–4774. https://github.com/slundberg/shap

Maier, H. R., Galelli, S., Razavi, S., Castelletti, A., Rizzoli, A., Athanasiadis, I. N., Sànchez-Marrè, M., Acutis, M., Wu, W., & Humphrey, G. B. (2023). Exploding the myths: An introduction to artificial neural networks for prediction and forecasting. *Environmental Modelling & Software*, *167*, 105776. https://doi.org/10.1016/J.ENVSOFT.2023.105776

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics*, *2*(4), 603–609. https://doi.org/10.1007/s43681-022-00143-x

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2023). Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance. *AIGA*. https://doi.org/10.48550/arXiv.2206.00335

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, *113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Marr, D. (1977). Artificial intelligence—A personal view. *Artificial Intelligence*, *9*(1), 37–48. https://doi.org/10.1016/0004-3702(77)90013-3

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *A Proposal for the Dartmouth Summer Research project on Artificial Intelligence*. https://www-formal.stanford.edu/jmc/history/dartmouth.pdf

McCorduck, P. (2004). *Machines Who Think: A personal Inquiry into the History and Prospects of Artificial Intelligence* (2nd ed.). A K Peters.

Meßmer, A.-K., & Degeling, M. (2023). *Auditing Recommender Systems -- Putting the DSA into practice with a risk-scenario-based approach*. https://doi.org/https://doi.org/10.48550/arXiv.2302.04556

Minkkinen, M. (2023). Roadmap to competitive and socially responsible artificial intelligence. *Turun Yliopisto Library*. https://www.utupub.fi/handle/10024/174307

Minkkinen, M., Niukkanen, A., & Mäntymäki, M. (2022). What about investors? ESG analyses as tools for ethics-based AI auditing. *AI and Society*. https://doi.org/10.1007/s00146-022-01415-0

Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine (AAAI)*, *27*(4), 87–91. https://ojs.aaai.org/aimagazine/index.php/aimagazine/issue/view/165

Nauta, M. (2023). *Overview of XAI Methods* . https://utwente-dmb.github.io/xai-papers/#/

Nauta, M., Trienes, J., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., Van Keulen, M., Seifert, C., Trienes, J., Schmitt, Y., Schlötterer, J., Seifert, C., Nguyen, E., Peters, M., van Keulen, M., Pathak, S., & van Keulen, M. (2023). From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*, *55*(13), 1–42. https://doi.org/10.1145/3583558

OpenAI. (n.d.). *Models - OpenAI API*. Retrieved May 10, 2024, from https://platform.openai.com/docs/models/overview

Pandey, S., Agarwal, R., Bhardwaj, S., Singh, S. K., Perwej, Dr. Y., & Singh, N. K. (2023). A Review of Current Perspective and Propensity in Reinforcement Learning (RL) in an Orderly Manner. *International Journal of Scientific Research in Computer Science,* *Engineering and Information Technology*, *9*(1), 206–227. https://doi.org/10.32628/CSEIT2390147

Perrigo, B. (2024, March). *AI Poses Extinction-Level Risk, State-Funded Report Says | TIME*. TIME. https://time.com/6898967/ai-extinction-national-security-risks-report/

Petterson, M. (2005). The keys to effective IT auditing. *Journal of Corporate Accounting and Finance*, *16*(5), 41–46. https://doi.org/10.1002/JCAF.20134

PWC. (n.d.). *What is an audit?* Retrieved April 10, 2024, from https://www.pwc.com/m1/en/services/assurance/what-is-an-audit.html

PwC. (2023, March 15). *PwC announces strategic alliance with Harvey, positioning PwC's Legal Business Solutions at the forefront of legal generative AI*. https://www.pwc.com/gx/en/news-room/press-releases/2023/pwc-announces-strategic-alliance-with-harvey-positioning-pwcs-legal-business-solutions-at-the-forefront-of-legal-generative-ai.html

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Ramon, Y., Farrokhnia, R. A., Matz, S. C., & Martens, D. (2021). Explainable AI for Psychological Profiling from Behavioral Data: An Application to Big Five Personality Predictions from Financial Transaction Records. *MDPI*. https://doi.org/10.3390/info12120518

Rendon, L. G. (2022). An Introduction to the Principle of Transparency in Automated Decision-Making Systems. *2022 45th Jubilee International Convention Information, Communication and Electronic Technology, MIPRO*, 1245–1252. https://doi.org/10.23919/MIPRO55190.2022.9803417

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rittenberg, L. E., Johnstone, K. M., & Gramling, A. A. (2010). *Auditing: a business risk approach* (7th ed.). South-Western College Pub.

Rolfe, G. (2006). Validity, trustworthiness and rigour: Quality and the idea of qualitative research. *Journal of Advanced Nursing*, *53*(3), 304–310. https://doi.org/10.1111/J.1365-2648.2006.03727.X

Russell, S. (2016). Rationality and Intelligence: A Brief Update. In *Fundamental Issues of Artificial Intelligence* (Vol. 376, pp. 7–28). Springer. https://doi.org/https://doi.org/10.1007/978-3-319-26485-1_2

Saarijärvi, M., & Bratt, E. L. (2021). When face-to-face interviews are not possible: tips and tricks for video, telephone, online chat, and email interviews in qualitative research. *European Journal of Cardiovascular Nursing*, *20*(4), 392–396. https://doi.org/10.1093/EURJCN/ZVAB038

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Elsevier: Knowledge-Based Systems*, *263*. https://doi.org/10.1016/j.knosys.2023.110273

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278. https://doi.org/10.1080/12460125.2020.1819094

Sovrano, F., & Vitali, F. (2023). An objective metric for Explainable AI: How and why to estimate the degree of explainability. *Knowledge-Based Systems*, *278*. https://doi.org/10.1016/j.knosys.2023.110866

Stenfors, T., Kajamaa, A., & Bennett, D. (2020). How to … assess the quality of qualitative research. *Clinical Teacher*, *17*(6), 596–599. https://doi.org/10.1111/TCT.13242

Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist*, *52*(10), 1030–1037. https://doi.org/10.1037/0003-066X.52.10.1030

Suzor, N. P., Myers West, S., Quodling, A., York Licensed, J., York, J., & Myers, S. (2019). What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation. In *International Journal of Communication* (Vol. 13). https://ijoc.org/index.php/ijoc/

Tang, Z., & Kejriwal, M. (2023). Evaluating deep generative models on cognitive tasks: a case study. *Discover Artificial Intelligence*, *3*(1), 1–19. https://doi.org/10.1007/S44163-023-00067-3

Tegmark, M. (2018). *Life 3.0: Being Human in the Age of Artificial Intelligence*. Penguin Books.

Tiainen, M. (2021). *To whom to explain and what? : Systematic literature review on empirical studies on Explainable Artificial Intelligence (XAI)* [Turun Yliopisto]. https://www.utupub.fi/handle/10024/151554

Tolan, S., Pesole, A., Martínez-Plumed, F., Fernández-Macías, E., Hernández-Orallo, J., & Gómez, E. (2021). Measuring the Occupational Impact of AI: Tasks, Cognitive Abilities and AI Benchmarks. *Artificial Intelligence Research*, *71*, 191–236. https://doi.org/https://doi.org/10.1613/jair.1.12647

Tracy, S. J. (2010). Qualitative quality: Eight a"big-tent" criteria for excellent qualitative research. *Sage Qualitative Inquiry*, *16*(10), 837–851. https://doi.org/10.1177/1077800410383121

Turing, A. M. (1950). COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/MIND/LIX.236.433

Veisdal, J. (2019, September 12). *The Birthplace of AI. An essay about the 1956 Dartmouth workshop*. Medium.Com. https://www.cantorsparadise.com/the-birthplace-of-ai-9ab7d4e5fb00

Weber, L., Lapuschkin, S., Binder, A., & Samek, W. (2023). Beyond explaining: Opportunities and challenges of XAI-based model improvement. *Information Fusion*, *92*, 154–176. https://doi.org/10.1016/j.inffus.2022.11.013

Weigand, H., Johannesson, P., & Andersson, B. (2021). An artifact ontology for design science research. *Data & Knowledge Engineering*, *133*. https://doi.org/10.1016/J.DATAK.2021.101878

Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer. https://doi.org/10.1007/978-3-662-43839-8

Yin, R. K. (1981). The Case Study Crisis: Some Answers. *Administrative Science Quarterly*, *26*(1), 58–65. https://doi.org/https://doi.org/10.2307/258557

Zemankova, A. (2019). Artificial Intelligence in Audit and Accounting: Development, Current Trends, Opportunities and Threats-Literature Review. *Proceedings - 2019 3rd International Conference on Control, Artificial Intelligence, Robotics and Optimization, ICCAIRO 2019*, 148–154. https://doi.org/10.1109/ICCAIRO47923.2019.00031

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy and Technology*, *32*(4), 661–683. https://doi.org/10.1007/s13347-018-0330-6

Zhang, C., Cho, S., & Vasarhelyi, M. (2022). Explainable Artificial Intelligence (XAI) in auditing. *International Journal of Accounting Information Systems*, *46*, 100572. https://doi.org/10.1016/j.accinf.2022.100572

Appendices

Appendix 1 Glossary

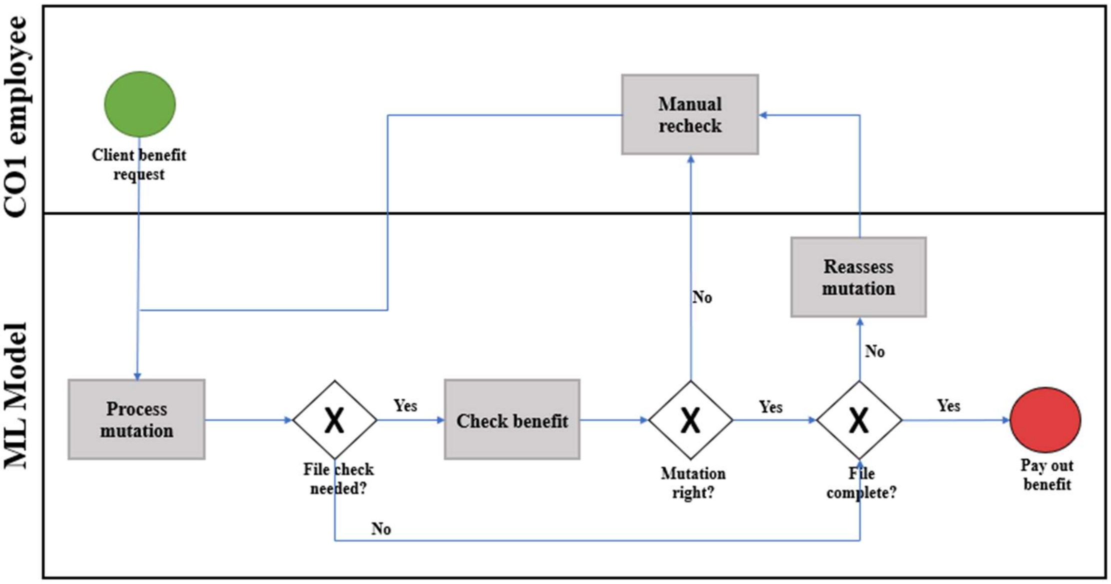
Term	Explanation
AI (Artificial Intelligence)	Non-biological intelligence entity
ABI (Artificial Broad Intelligence)	AI that covers multiple complex goals
AGI (Artificial General Intelligence)	Superintelligence General AI Human-level AI
AIG (Artificial Intelligence Governance)	Policies, regulations, and procedures to oversee the AI lifecycle
ANI (Artificial Narrow Intelligence)	AI for a specific goal or narrow range of tasks.
ANN (Artificial Neural Networks)	Computational model inspired by structure of biological neurons
Black-box	Algorithm or AI entity that is complex and difficult to understand since the transformation of input and output is fully conducted, without human intervention, by an algorithm/AI
DL (Deep Learning)	Neural network with many layers
GPT (Generative Pretrained Transformer)	Large language model that leverages DL to interpret human text
Grey-box	Combination between white-box (transparent and understandable) and black-box (opaque, no knowledge on internal workings).
Intelligence	The ability to accomplish complex goals. (Tegmark, 2018)
ML (Machine Learning)	AI that develops models to perform tasks without instructions
NLP (Natural Language Processing)	Interaction between computers and humans through natural language
RAI (Responsible AI)	Integration of ethical concerns in AI research
RL (Reinforcement Learning)	ML that learns through rewards and punishments
White-box (glass-box)	(ML) models that are easier to comprehend and self-interpret without the need for an (external) explanation. (Ali et al., 2023a).
XAI (eXplainable AI)	Research to make AI output explainable

Appendix 2 Case descriptions

Case 1. Dutch insurance company – CO1

At an insurance company (CO1), during a regular financial statement audit, a control was found which had an almost fully automated execution with Machine Learning (ML) at its core. This case ran from 2020 to 2022. The control on which assurance was required is a so-called four-eyes principle.

CO1 has an insurance system in which mutations occur. By 2022, whether or not to check a benefit was determined by a predetermined probability of a benefit mutation being wrong. The high-risk mutations, based on variables as benefit amount, employee experience, fraud, are identified by a ML model. The selection of mutations, made by the model, must go through a 4-eyes (sometimes 6-eyes) control. For example, the riskiest 20% are pushed forward. The picture below present gives a rough idea on the process flow. Note that this is not the official representation of the process, merely an oversimplified adaption to indicate the checkmarks of the model.



In general, there are 4 classifications in the data. True positives, the positive benefit mutations. False Negatives, the cases where the model predicts benefit mutations as negative when they are positive. In addition, there are true negatives and false positives. Providing assurance over this control (the ML model) was mainly done through testing how the model performs based on precision, accuracy, and recall. It was discovered that of all positive benefit mutations 68% was correctly identified. This means that of all benefit mutations that have to go through the 4-eyes principle, 68% actually went through. Although, it was found that the 32% of the errors that were not found had no direct impact on the financial statements, the question for assurance remained on how much risk was left. This is the point where the case ended since KPMG cannot determine whether 32% is an acceptable margin of error. Partly because this risk could not be statistically supported. If it was a manual control, that is based on business rules, the risk is often implicitly accepted.

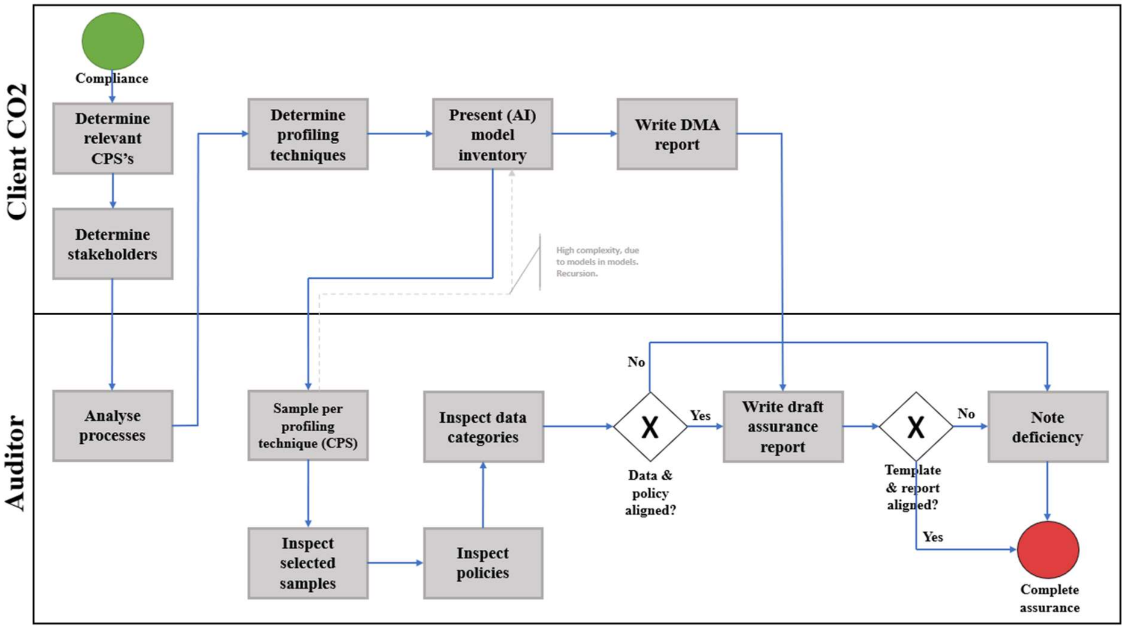
During the algorithm audit, training data was considered, but not the input data. CO1 draws a 5% sample on the total population as a training set for a renewed model version. The audit also looked into general risks which includes changes that happen on logical access security. The availability of the model was not in question as it posed little to no risk. Changes to the model, change management, is also an item that was investigated. The process was commented to be ‘OK’, but it was not organised well enough to be auditable. There were no model shortcomings or flaws detected.

KPMG concluded for CO1 that there was a substantiated business case for a clear benchmark of the model sufficient exploratory data analysis had taken place, there were no methodological errors in the model, the approach to training and validation is sound, and CO1 had set up sufficient roles and responsibilities regarding management and continuity of the model. The main comment to this model was on the fact that CO1 required trainings to employees who were more likely to make mistakes on certain mutations. The ML model is a feedback loop. Meaning people with small errors are checked more often and thus small errors will be flagged even more often.

This case description is conducted through unstructured interviews with stakeholders that were involved. One of the interviews was with a senior manager who was responsible for the engagement with this particular client. This person stated at the end of the unstructured interview that there is nothing prescribed for the audits of algorithms, and called it a free-format activity with a best-effort of all people that were involved. According to this person it will be very difficult to argue whether a model is considered ‘sufficient’ based on recall, accuracy, and precision.

Case 2 & 3. Very Large Online Platforms – CO2 & CO3

Both presented cases from Case Organisation 2 (CO2) and Case Organisation 3 (CO3) are an algorithm, AI, audit to be compliant to the European legislation DSA and DMA. CO2 is a so-called large digital platform within the social media landscape and CO3 is a multinational e-commerce and technology company. Due to the sizes of these clients, the novelty of the audit type, and the sensitivity of the available data and information not a lot of details are allowed to be documented. Despite significant differences, the execution of the audit is relatively similar. It should be noted that CO3 had included some type of (IT) controls in the recommender system whilst this was different for CO2. Hence, CO2 was able to provide some more explanations on what was done. The figure below is an indication of the DMA audit process. Note that due to strict disclosures this graphical representation is not at all representative for the exact steps. The process is a generalised adaption of the, by KPMG provided, internal source. In combination with the unstructured interview with the CO3 representative this flowchart can be regarded as basic case documentation for both CO2 and CO3. Because both cases were currently auditing on profiling aspects it was included in the picture in order to stay close to the actual source material. However, the basis was CO2.



Where there is no opportunity to describe clear explications on the activities it is allowed to highlight some aspects of the above process. A significant difference with both cases is that in CO2's model inventory exists from a multitude of ML models that feed one and another. This model in a model structure is heavily recursive, increasing the complexity. Furthermore, the articles about profiling and transparency are interpreted differently per client. Since CO3 is a large digital social media platform profiling is one of the main goals with little negative impact due to the corrective nature of an algorithm. The possible philosophical and ethical concerns that can be raised fall out of the legislative scope. At CO3 transparency is considered to be sufficient if there are clear instructions about video reporting, profiling goals, and data collection. For CO2 the data collection works somewhat differently since the recursive structure of models is trained upon personal data. However, the personal details are not 'profiled' back to the individual when the 'profiling pop-up' is rejected by that specific individual user. The information can be used for other users, which is not considered as profiling according in the DSA and DMA. The brief article about transparency is as interpretable for CO2 as it is for CO3. CO2 covers this article through for example the terms of use agreement.

For both DMA audits one of the first steps is for the client to determine the most relevant core platform services that is to be controlled / audited by a third party according to the law. Together with the other step that are executed on the side of the client there is some kind of interplay to which the audit can be hung up to. The two parts are between 'what the law prescribes' and what is said by the client. For CO2 and CO3 what the client states is as minimal as possible, in alignment with the law. In a general sense the algorithm audit refers back to the client statements whilst being operational under the flag of the DSA and DMA. The auditor is responsible for checking, for example, the profiling techniques and algorithms beneath it. For CO3 this meant that transparency is provided about the variables the recommender system is basing its recommendations on. This was done even on the code level. CO3 had some built-in controls to be checked whereas CO2 mainly had to rely on certain parameters to make understanding of the models in models' structure. Leading to the fact that the recommender system was considered to be more of a black-box than the recommender system at CO2.

Besides the procedural representation above, this is the very abstract information distilled from the walkthroughs. The walkthroughs are purely contextual input next to the literature.

Appendix 3 Interview questions

Most of the interviews are an adaptation from the questions below. The first interview was not only used as input, but also as a test of the interview questions. Small additions have been done for this final set.

Current role

1. As a Responsible AI consultant, what is your role in an AI audit at Amazon?
2. What are the process steps in an AI audit compared to a normal audit?

Concept of AI auditing

3. What risks are covered with algorithm/AI assurance?
4. What standards and frameworks does the audit rely on?
5. If there are (IT-application) AI controls, how are they audited, how are they different from normal controls?
    a. Or what if there are no controls?

DMA/DSA specific

6. DMA / DSA requires transparency from all AI systems. How is transparency defined at CO2?
7. What metrics are in place to measure and ensure this transparency?
8. DMA / DSA is mainly about providing transparency around the system, how do you ensure ethical principles, like fairness, in the training sets and how can you ensure that a Black-box is compliant with such principles?
9. How do you assure that the decisions of a (recommender) system are fair, ethical, and explainable?

Maturity and guidelines

10. What was challenging and what was successful at the CO2 case?
11. What features should definitely be included in (new) guidelines?
12. Do you foresee some kind of maturity path in AI auditing? Not every customer might have controls, how will the process be changed?
13. What criteria should an auditable AI meet to start an assurance project?
14. How do you foresee an AI audit when the business maturity is relatively low and the AI system is a complex Black box?
15. A known standard is CRISP DM (business to data understanding), is this line of thinking applied in DSA/DMA audits?
    a. Is there such a thing as the AI life cycle?

Although, the questions above represent the final set of questions. There were interviewees with more ties to IT assurance than Responsible AI (auditing). As a result, some additional questions were included for those interviewees. Furthermore, during the Design Science phase some readjustments were made creating a certain need for more practical insights in the assurance perspective of AI as well. Below, the set of questions for the more assurance-oriented interviews.

Additional questions about algorithm assurance

1. What is (IT / AI) assurance?
2. How would you compare IT assurance to algorithm / AI assurance?
3. What criteria should an auditable AI meet to start an assurance project?
4. Is there a general approach to providing AI assurance?
5. The context around an algorithm is of importance. When is the context reliable and sufficiently mature to consider the algorithm a black-box?
6. If the model or algorithm cannot be reviewed by itself on what elements will the assurance engagement focus?
7. On what exact elements is the assurance focussed? What will be written in the assurance report of an AI audit compared to IT assurance.
8. How do you foresee an AI audit when the business maturity is relatively low and the AI system is a complex Black box?
9. Do you agree that the ethical framing of each client case will be different, which leads to different audit activities?

Is the risk assessment for AI audits different than for IT audits?

Appendix 4 Interview summaries

Interviewee A. CO1 – Interview 2

(1) Basically, the case at CO1 is not defined from Responsible AI, but from a regular audit. The Responsible AI team is more of an advisory team. However, consultants are brought to cases like CO1 because of the technical expertise they bring. Concrete process steps are difficult to define. (2) On the one hand, it always starts with an intake and exploration of the systems and processes. (3) However, KPMG is careful not to draw its own hard conclusions or define standards. Basically, the auditor is always looking for external standards or a review model from the client itself, and this is especially true in the new area of AI auditing or algorithm assurance. Other than that, interviewee A defines a standard procedure during the intake where questions like:

- What has been built?
- How is it put together?
- Is there (design) documentation available?
- Are there controls internally or around it?

During the intake, the scope of the audit is determined and thus basically the 'audit object'. (4) The audit object is the system, the AI entity, or the environment. The main focus at that instance is on what is 'feasible' to do. KPMG itself has a 'control framework' for that, which is more or less based on the 5 pillars (reliability, resilience, explainability, accountability, fairness). (5) That framework can then be used to sort of "shop" which types of "controls" are appropriate for the risks found at the client so that they can be tested. Based on the shopping list, a kind of appropriate control framework is created on the case. (7) There were no AI-specific controls. There was nothing, as controls, established to which the situation could be audited.

Interviewee A says that providing assurance is difficult at the moment. (6). With CO1, the question was whether it was reliable enough to lean on as a control. Basically, you often end up with the 5 (to 7) pillars devised by the high-level expert group based on about 4 ethical values, which forms a blueprint for the AI act. You should then read the AI act as mitigating measures on infringement of the 26 fundamental human rights. The upcoming AI Liability Directive is about claiming for damages and who, or what, is then accountable. However, the purpose of an AI audit is always aligned with the object, the scope, and is traceable back to those 5 pillars.

In CO1, the extent to which the model is transparent was not really measured per se. While a qualitative statement can be made about that, transparency mainly says 'something' so that an auditor can look at it. One can look at the General IT Controls and, for example, logging the number of pushes done. Another example is going through the code and checking the public libraries used. (8) Further, to provide a level of assurance at CO1, about relying on the ML model, a number of performance tests were done, such as measuring recall, accuracy and precision. Now, auditors are mainly interested in recall while an organisation wants high precision. An auditor is looking for what is potentially missed and to what extent the missed errors have material value. In a model, we perform a performance metric and the outcome of such a metric is a percentage to be transposed to euros. However, there is no baseline of performance defined. Simply because a metric like recall is very domain-specific. The environment should be included in the consideration, and perhaps also something like this in a defined standard (12). Because there is no standard, one often avoids to formulate a statement as to whether the model (that is issuing benefits) is fair and ethical. An assurance or audit report is then more likely to note 'no findings'.

(14) Ultimately, you look at how a business case translates to CO1's ML solution. That documentation was present and then you can judge by a piece of explainability and fairness. However, explainability is different for everyone according to interviewee A. Basically, the insurance model does not use personal data so in that sense, fairness was not necessarily an issue. Interviewee A states that the problem with organisations that are 'more mature' is that they want to provide less information because anything the auditor sees can be considered as a potential competitive risk. In a way, that complicates questions like transparency. In general, companies are looking more for model validation than to put it in auditable terms. Model validation is what was done at CO1. Interviewee A states that in the end you are looking for a description of purpose, an ROI, and the measurability of success. Auditing AI is also about what the market values.

Interviewee B. CO2 – Interview 3

Microsoft copilot summarized the interview from the Teams transcript as follows:
The meeting was a conversation between B, a consultant at KPMG who specializes in AI, and Bran van Wingerden, a student who is writing a thesis on AI audit. The meeting had the following main points:

- *Bran asked B about her role and experience in the Amazon case, where she was involved in a compliance audit for the DMA and DSA regulations, which are related to recommender systems and profiling.*
- *B explained the difference between the DMA and DSA audits, and how they are based on a template from the EC that specifies what information should be provided by Amazon in their reports. She also mentioned some of the challenges and limitations of the audit, such as the lack of clear guidelines, the difficulty of defining materiality, and the black-box nature of the AI systems.*
- *Bran asked B about the frameworks and standards that she uses for the AI audit, and she said that there are not many specific ones, but they use the KPMG audit methodology and some ISA standards as a basis, and also rely on their professional judgment and experience.*

Bran asked B about the concept of transparency and how it is defined and measured in the audit. Angelica said that for the DMA, it is mainly about the data that is used and how it is described, and for the DSA, it is about the compliance with the relevant articles of the regulation, such as the right to opt out, the right to explanation, and the transparency report.

The following part is a summary made by the author of this thesis.

As a responsible AI consultant interviewee B was brought to the case as an expert on AI and the story around those type of systems that are mentioned in the legislation of the DMA and DSA. (1) Interviewee B refers to the audit case at CO2 as more of having AI as a part of the compliance audit than the audit being an absolute AI audit. Access and change management are typical for IT audits, but are also checked for AI audits. These controls are often set up maturely and testable. Through the lens of AI there are differences since financial statement audits are looking for Risk of Miss statements whereas an AI audit is not interested in that. (2) Very often it is not that material either. In every audit an important part are the walkthroughs and they are even more vital in audits that involve AI's. Especially around the DMA/DSA where a lot of compliance is required around the AI system. Those walkthroughs determine where some deep-dives are required.

(2)(4) The main different thing is that for DMA/DSA the audit relies on a framework that is provided from the European Commission. Therefore, the report from CO2 on their profiling technique is important. Despite some of the provided templates around the legislation it remains to be a best-effort of all the people involved since there are no concord standards or frameworks.

Some examples of guidelines that could be used are the ISAE3000 and ISAE3402. Important to realise about (6) DSA and DMA is that the DSA is a point in time and for the DMA a period in time is investigated. Despite some regulations interviewee B still acknowledges that the audit remains, more or less, a format less activity. Obviously, it is not that black and white, only the entire endeavor tends to rely more on the professional judgement from the auditors than an actual measurable fact provided by authorities. (5) An AI system, or algorithm, of just a system viewed as a black-box if you will, might be part of a system landscape. For that reason, it might be interesting to look at GITC's (General IT Controls). Change management is then interesting to look at since it formulates a point in time. In a general sense, the audit looks whether the recommender system is truly correct and works as it intends to. However, DMA and DSA just lightly touches upon AI related systems. Mainly article 15 of the DMA is interesting, but most articles are not seen relevant by CO2. CO2 is leading in the matter of being compliant and KPMG is the controlling part which leads to discussions, especially since the guidelines from the EC are unclear on what they exactly want to see. As a result, both auditor and client (CO2 in this case) come to an agreement on what is considered to be sufficient to the regulation. For transparency the system was decided to keep as a black-box, so the transparency was considered whatever CO2 delivered and showed. In Dutch the 'sixes-culture' is then considered to be sufficient.

The transparency could not be actively measured, but what was provided by CO2 was a list of ML models that were part of a system or even something that could be a control. (7) The list was more or less an excel sheet about name and description on personal data to investigate where it was utilized. With that information it could be concluded whether or not a model was profiling. So basically, the backend of a model was viewed.

The interviewee provided an example of a report which showed a dedicated chapter that explicated the used features of a model. Each feature consists of different strategies that are responsible for the product's essence. From the features and their essence (which is basically a huge list in a list) it can be derived whether profiling was actually used by the model. Using the subscriber status, like some type of premium account, is considered to be profiling since that is a customer account information category.

(7) If a model is trained on personal data, but is not utilizing it to individual cases then the model is not considered as a model that is profiling. Interviewee B gave a very explicit explanation on this topic. The conclusion is enough to indicate the reasoning that more ethical principles are out of the scope for DMA / DSA. If it were in fact a profiling system, the regulation would have been different.

All the conclusions that are derived from the findings are highly qualitative. In addition, everything that falls outside the set frames of the DSA/DMA regulation is not considered to be relevant. The only thing the auditor can do is provide so called 'comments' to the conclusion which can either be positive or negative. It should be noted that the DMA and DSA came up quite fast, meaning that not all organizations were ready for it by default. Furthermore, this is the first year it is in effect.

(7) (8) At CO2 the ML models make use of heuristics which require a deeper level to understand the decisions. Transparency is required on where the data originates from and how it is extracted. Afterwards, the code and databases are reviewed. However, the focus was on the input and not on how the model came to a decision. For the DMA / DSA it is solely interesting to look into the certain data categories (features). Besides, CO2 makes use of neural networks as well which makes it even more difficult to formulate a statement on topics as fairness. (9).

(11) Interviewee B states that it is hard to outline features that should be included in guidelines. Reasoning from something as open as the DMA / DSA the audits are dependent on the level of transparency that is provided to which the question is asked if reasonable assurance can be given. The assurance is therefore mainly based on professional documentation. Auditing a black-box itself is difficult. Performance metrics are interesting for an organization as well, since they will not benefit from a model that is biased either. Laws and regulation that go deeper in that topic would be nice according to interviewee B.

(12)(13) when a model is not that mature or even highly complex one should consider the maturity of the controlling mechanisms, process-wise, around it. You want to see if there were some steps undertaken which you would see back in the controls. For example, bias is taking into account. In addition, cyber security is another concern. Looking at language models like ChatGPT, it requires moderation on questions such as 'how do I make a bomb' and the workarounds users come up with. Interviewee B states that one expects some degree of maturity around that.

In conclusion, interviewee B states that there are, apart from what is described in the DMA/DSA, no strict guidelines from the European Commission which leads to the necessary discussions with clients. The data is one of the only things that can be further looked into and the system itself is simply regarded as a black-box due to the highover scope from the regulations and the simple fact that the systems might be complex. The client is very much in the lead and determines what is in the scope and what not. Meaning, the client determines the level of transparency given over everything around the algorithms, including the processes and documentation. Measuring the compliance is done through the described report-requirements and the direct interaction between systems.

Interviewee C. CO3 – Interview 1

Interviewee C, in line with the DSA, distinguishes three main types of AI systems:

1. Content moderation systems
2. Advertisement systems
3. Recommender systems

Each system has specific transparency requirements under the DSA. Interviewee C provided this information to outline part of the job of a RAI consultant and a current form of algorithm audits that audited against a form of regulation.

Content moderation systems - art. 15(1)(e)
Includes automatic content assessment, such as detecting violence in videos. Regular IT systems are not suitable for analysing video images or sounds. Manual controls also apply here. Violence can be reported manually. All automated means of content moderation are considered AI.

Advertisement systems - art. 26
Advertisements use explicit profiling to target users based on demographics such as age and location. Transparency in ad systems includes explaining to the user why an ad is shown. Advertisers want to be able to define their audience, which is a specific profiling step and does not mean analysing videos and deciding that an individual user likes cars, for example.

Recommender systems - art. 27
These do not use explicit profiling prior to recommendations. Instead, they base recommendations on user behaviour and the preferences of similar users. Here, videos are analysed and the user's behaviour is concluded as 'likes car'. Then another car video is recommended because similar users also liked this video about cars.

When asked how transparency is defined in the CO3 case, Interviewee C replies that Article 15E of the DSA sets out the transparency requirements for automated content moderation systems, including the need to report performance and accuracy measures. In short, the transparency here consists of declaring performance by describing the purpose with the measures of accuracy. A description about the operation is not important. The story and transparency around the system is.

Initially, KPMG's interpretation was that for each automated means of content moderation, which has its own individual and distinguishable purpose, a measure of accuracy should be included for transparency. CO3 does not do so because DSA Article 15E only defines transparency reporting. CO3 reports a number across all their automated means of content moderation through a kind of performance number using assigned objections. A performance metric consists of Precision, Recall, and Accuracy. An objection is a reassessment request such as a video that has been reported for violence. If a performance metric is based on objections, there is no representative sample, only the samples of objections. In doing so, you only have the positives, because the only people who could object are those who owned the content. This gives a very distorted picture of precision. In theory, you can post violent videos, by hoping a large proportion slips through, by not drawing attention to your account because you don't object yourself.

The DSA defines transparency for advertising systems as demonstrating, ad by ad, why this ad is displayed to the user. The transparency requirement is an explanation of why an ad is targeted to you. Whereas with advertising systems it is purely about linking the outcome, per ad and why this individual. With Recommender Systems, it's about the terms and conditions. A general story that applies to everyone. These are different parameters.

The DSA provides limited guidelines for transparency and is therefore a relatively scant regulation. This is a disadvantage for auditors because they want to define the requirements themselves as little as possible. This would lead to taking too much responsibility. Currently, auditors check the bar while there are still few requirements for this AI bar. Simply because the subject is so new and complicated.

In this sense, the audit scope is relatively limited because the main concern now is the accuracy and completeness of the information. If the information is complete and accurate and CO3 decides to adopt a very high level of summary in generalities on that, it still meets the DSA requirement.

The AI audit at CO3 is very close to a regular audit. Therefore, the process is audited more than the technology. In the AI audit, with the DSA, 2 things happen:

1. The audit looks at what CO3 itself undertakes, what the process is to arrive at the main parameters of the Recommender System. The Responsible AI team then reviews the setup in CO3's design.
2. The audit looks into points in the process, Process Risk Points, where things can go wrong. They then look for management measures, or control activities. Automated controls are then definitely looked at

An AI audit differs in the sense that this type of audit requires more expertise AI systems to fully comprehend the bigger picture.

Controls can be automated workflows, literally flags. However, because there are still so few frameworks, auditors still do very little in this. Because one does not yet go into the depth of the technology very much, with a DSA audit, this form of AI audit is very similar to the regular audit process and can in principle also follow an ISA standard.

With CO3, the scope and context of the audit is agreed on and managed beforehand, all details are hammered. In doing so, KPMG itself still gives its own opinion on what is stated. CO3's interpretation is measured against KPMG's. One example is that CO3 has a support page explaining adjustments. There is an option in the app to customise your location, however, it is not included in the help centre article. This way, one's own opinion is formulated about CO3's compliance.

An example of controls are the design documents. These are prepared for every change in the development process. The design document requires a review that functions as a 4-eye principle control. In fact, this is an IT application control that is not about the recommender system itself. For Content Moderation systems, no application controls are tested.
It is only about transparency as to why the output is actually there. Whether the recommender system works properly is outside the scope. In fact, for current DSA audits, with the narrow scope, the system is considered a black-box and hardly touched.

For both DSA audits as general AI/algorithm audits, 5 measurable boxes are important:

1. Reliability
2. Resilience
3. Explainability
4. Accountability
5. Fairness (ethics)

For advertising systems, the simplest form of explainability applies. Content Moderation systems are in the spectrum of reliability, based on measures of accuracy. Recommender systems are one big jumble of dozens, if not 100, machine learning models all feeding each other. To make a statement about reliability you really have to enter the black box. However, there are no reliability requirements for Recommender systems. These 5 characteristics serve as a sub-objective to the predetermined, leading, audit objective.

On the question (12) whether interviewee C sees a certain maturity within auditing AI systems, she replies as follows: CO3's environment is very mature, they spend almost all day working on AI models. As a base, then, there is a relatively controlled process. Development has to be meticulous because it is a platform with many users. Mature processes are then a requirement. In terms of technology, collaboration, and supporting tools, there are good quality standards at CO3. Before anything is implemented, there are many test procedures and you can measure these using the DSA. Because of the professionalism of the processes, you can also lean more easily on the quality controls that are in there.

However, there is also the other way around. If the model turns out to be simple and uncluttered then sometimes it is easier and faster to dive in. With prepared data from a dataset that is not too large, you can run it as an AI model trainer on a laptop. A performance audit is then more effective and efficient. When an organisation is not mature at all, then one does need to dive into the content to make a judgement on DSA transparency. There is some sort of a trade-off according to interviewee C.

Lastly, interviewee C mentioned a framework that is often referred to in AI auditing and assurance. The framework stems from Data Management, and is called CRISP DM. In the CRISP model, you start with business understanding and make sure the business stakeholders are revived. This is a control according to interviewee C. Data understanding involves examining the data and looking for errors and outliers. The solutions and techniques based on the missing data reports are also a kind of control. The development phase checks whether it has been implemented as envisaged by the data understanding phase. Interviewee C states that if this is maturely put in place then only evidence is needed that the review has taken place. Which could serve as some kind of control as well. Another example is that process controls are spread across several departments. Privacy department checks design documentation and the Legal team looks at key DSA requirements such as profiling.

Interviewee D. CO3 – Interview 4

Interviewee D introduced herself as a Responsible AI consultant that got involved with the CO3 case because the DSA / DMA defines articles that involves statements about AI components. Reasoning from the regular IT auditing perspective would be sufficient and according to Interviewee D an IT auditor would pick up on what is defined. However, it is extremely helpful for such an audit to bring along a certain level of expertise on AI systems. Including background knowledge into these types of audits ensures that no risks are overlooked. (1)

(2) interviewee D answered that it is still difficult around standards and frameworks. She describes that the Responsible AI team never originated from a standard or envisioned way of working. There were simply complex systems that required a certain level of auditing. The team was an answer to look into the formulated guidelines and what is possible with those guidelines. Characteristically to AI, things are not as simple or always the same. Therefore, you will always have a case-by-case approach. Interviewee D compares AI auditing to a standardized approach for SAP ERP systems which has components you will see everywhere. An AI system contains often way more than that and is applied in a highly specific situation. Interviewee D concludes that she cannot provide concord process steps.

(3) In the AI audits cases there are some classical components that always find their return. There are 5 certain pillars onto which every risk can be hung. Transparency will, for example, always be discussed because of the black-box problem. Very often the same kind of elements will be looked into, but it is never the same set of elements. Because of the different cases and AI systems the combination of elements changes. In a sense, the audit relies on the professionalism of the AI auditor as well. From experiences they know what they can, or actually should, expect during AI systems that are investigated. This to some sort of knowledge list of risks that could be found in AI systems. Interviewee D describes that based on this inhouse knowledge the right combination of expected risks will be investigated. These risks are expected to be covered by the clients. (4) There is no standard that determines how or what exactly should be audited. It is mainly checking the homework of the client to see if there are control mechanisms. However, for any DSA / DMA audit it is not necessary to rely on a knowledge bank since it, more or less, defines the aspects that are required to be looked into. The European Commission wants to know whether CO3 is compliant to the law. (4)(5) Other standards are ISA. There is a professional benchmark for all clients at KPMG, which creates general understanding for AI professionals to refer to. This general understanding formulates the professionalism, but is not something that could be applied to a client. For example, there is a formulated expectation for a topic as recommender systems. There is no formulated baseline to audit on. Interviewee D states that all standards always represent relatively the same. In that sense AI professionals know where the alarm bells ring.

Definitions are defined, but require to be made more concord. (6) So, for example, illegal content needs to be able to be reported by a user. It is defined that the report mechanisms need to be easy to access. However, CO3 might define user friendliness differently than the auditor. Furthermore, it opens conversations about the risks that are brought along. Interviewee D indicates that the discussions always come back to the fact that it all needs to be about the defined law. So, every time something appears to be not as transparent and clear the question is asked how it affects the European audience. The goal of the law is to protect that specific audience.

Interviewee D explains that the DSA and DMA is sometimes poorly written if reasoned from a RAI or XAI perspective. It is mainly a first formal law around complex algorithms and sometimes AI.

Interviewee D describes this as one of the reasons for the problem that the other side, the client, does not always expect they can talk about the complex technicalities. (7) This is the reason we bring people with AI expertise according to interviewee D. A sidenote is that the audit still remains highly qualitative. Interviewee D states carefully that there could be a potential risk in attempting to capture and scope AI audits into a set framework. In the current practice there is not such a thing to refer back to for an audit, unlike regular auditing regulations. If a stamp is put on the AI, that was audited, and some shallow tests are used the meaning of a such a stamp and standard is debatable. Interviewee D states that it is, at this point, the best to orchestrate a multidisciplinary group (containing, data scientists, RAI professionals, policymakers, directors and other stakeholders) to involve all topics in one discussion to investigate what is relevant for the particular case. Interviewee D describes that this would be very hard to cover with a simple questionnaire. She refrains from making something too simple or standardised.

For the audits that rely on the DSA and DMA it could be concluded that for an article a lot of work has been put it, but it is not in their own hands yet. Since it is the first year of this audit this would mean that there is not a strict bad judgement yet. At this point the involved auditors, AI professionals, and client give their best work.

(6-8) transparency is mainly defined around the systems at CO3. Although, components of the code have been reviewed and general IT controls were looked into. Interviewee D states that if the algorithm is extremely complex, but its defined goal is quite simple and not that impactful then it might not be needed to open the black-box. However, it is probably another consideration when there is no human included in the entire loop.

(5)(12) If the AI system is truly a black-box then the only concord thing the audit can lean on are the processes around it. In that sense a way of working can be auditable. The processes need to be somewhat mature, which means that they include controls.

(9) Interviewee D talks about bias-in-bias-out, stating that when there is something wrong with the input data then the AI model will not meet its prescribed goals or even perform differently than intended. Over the data some degree of transparency and explainability could be given. In the case of the CO3 audit the regulation defines 3 boxes for recommender systems. The first one is the content itself, the second one is the user, and the third box is the interaction between user and content. These 3 boxes determine a lot of variables (data) that could be looked into. For example, box 1 has variables such as the topic of the content, if it is a video, it has variables as length, sound, and creator. Because of the number of variables for this specific case it is hard to determine which datapoints go in and how they get transformed into a recommendation. Furthermore, bias was not an area of concern for this specific case as well. Furthermore, interaction with the content is not something qualitative, it is quantitatively measured. Interviewee D states that it is therefore a representation of real life which gives another perspective to some of the ethical concerns that are placed at these types of systems. (13) This AI auditing domain might be more ethical, and even philosophical, than one might think. The moderation processes around the systems are indeed at place and provide guidelines, but

moderation of digital platforms might contradict laws as freedom of speech. Auditing an AI on ethical principles is different for every case. Interviewee D says that having one person ‘check’ the ‘checklist’ will never be enough for a guideline or regulation, it always requires more people to think about the to-be controlled features. Interviewee D adds to the discussion that agreement on the norms and values is a start, but acknowledges that it is a philosophical debate which cannot contribute much to current practice. Interviewee D cycled back to the asked questions and answered them briefly (after a discussion); maturity? Yes. Controls? Often no. In the end an objective statement is provided about the AI system and its environment.

Interviewee D did not really see something as CRISP DM being applicable, but agrees with the overall idea of initiating some type of a business and data understanding. It could help shaping the ethical thinking for each specific case. She gave earlier her ideas on standardising AI audits, but provided a nuance by saying that there are most likely certain key moments in the audit that could be checked. The moments are then more or less standardised.

Furthermore, interviewee D added that receiving a statement from a client that ‘they are thinking and considering the concern’ is not enough proof either. She states that staring blindly on controls is not a good practice either. The AI entity is not necessarily the audit object, but rather a part of it.

Interviewee E. ITA Department – Interview 5

This interviewee had a slightly different angle since interviewee E has an IT assurance background whilst all previous interviewees are placed in a responsible AI department. The questions were slightly adapted to fit the reference frame of the interviewee. Despite the ITA background the interviewee could reason from one experience which was a DSA / DMA audit at a large online retail service company.

Interviewee E described that (IT) assurance is about providing some level of confidence about the reliability, security, and compliance to law and legislations of IT systems. A major part is risk management. Assurance aims to ensure risks are covered to a certain degree. It involves developing and maintaining control frameworks, performance measurement and compliance as well.

Interviewee E indicated that there are always two important control objectives to look into.

1. Control of the purpose
2. Control of the measures

This determines a lot about an IT system. Initially the ‘object of research’, or to-be audited entity, should be defined as well with the underlying framework. With underlying framework some ISO standards are meant, or other common frameworks. Apart from determining what is going to be audited and to what frame it is going to be compared input from the business is required. Within an assurance report chapter 3 is always “*Description organisation & services*” which is basically a huge list that provides oversight onto everything that is relevant. The next chapter is the “*Control framework & test results*” in which the services and entity are reported towards the articles. Interviewee E presented an example chapter from an experience. The chapter 4 is basically a large table with statements and if needed some comments about the level of assurance that could be given. Interviewee added to this that instead of a list of services on billing and financial statement systems the AI audit & assurance field will likely run into services connected to recommender systems. Interviewee E said that the object ‘that goes into the assurance report’ will be measured, which requires information inquiry. This happens mainly through interviews. An interview is often not enough to assure something. Hence, an inspection, observation, or reperformance follows.

The audit object also involves peripheral matters and broader perspectives as well since it is often part of organisational structures. Therefore, the assurance puts a focus on stakeholders and processes as well. For example, in a DSA audit, change management on a recommender system has a log of around 200 executed changes to the model. In IT assurance there are risk matrixes that determine the sample amount on how many instances are logged since checking all 200 changes is too extensive. The risk matrix could, for example, prescribe to draw 15 changes as a sample. This will say something about the working of the system. Assurance is more or less a frame. Such as DSA. It offers framing to the auditing work and make a statement of confidence about the research/audit object.

During the interview, interviewee E provided an example of a current case he was working on. This happened to be a DSA / DMA assurance case. Interviewee E indicated that there are a lot of conversations about the article that describes the concept ‘profiling’ and ‘grouping’. The law demands that users of the digital platform have the option to select ‘prevent’ profiling in their privacy settings. This is auditable. However, the definition is tough according to interviewee E, since the simple fact that the settings allow to not profile, but the training-dataset is still using the personal data of the individual that deselected the profiling option. Interviewee E states that it mainly means that the personal data of individual X will not be ‘profiled’ over individual X himself.

Furthermore, for the advertisement systems it is hard to assure that the provided ads on the digital platform are not based on the personal details of the individual. Interviewee E gives an example by saying; ‘how is it possible to know whether a system decided that a 35-year-old European male fits the ad of a 37-year-old European male. It raises questions on whether both classifications are the same and whether geographical data is used or not. On one hand the law prohibits it, on the other hand the advertisements one sees are all purchasable in the current country an individual is living in. Somehow, both systems are still aware of the person’s nationality and geographical location, even if it is theoretically not considered to be profiling. Interviewee E gave this example to demonstrate the difficulties of interpreting the current written laws on AI and algorithms. Furthermore, interviewee E distinguished a term that is called ‘grouping’. The DSA describes certain parameters for advertisement systems. Those parameters belong to a certain group and are able to group a set of data accordingly. However, the assurance question is how those parameters got to that specific group. It remains difficult to showcase all the data or trace it down to the individual level. Hence, there will always be a form of so-called ‘reasonable assurance’ which basically means that 100% ensuring risk coverage is simply impossible. There will always be a level of uncertainty. The benefit of the DSA, in contrast to regular algorithm audits, is that the DSA actually provides the risks, or rather outlines the most impactful ones. Meaning that there is already framing for the risk assessment, making these discussions slightly easier.

Interviewee F. ITA/RAI Department – Interview 6

Interviewee F is the director of the RAI department at KPMG Netherlands, which is a subdepartment of IT assurance. Interviewee F describes how AI assurance can be seen as a specialism within IT assurance and might probably be the next step since he outlined that AI technology is used more often in a more impactful way. As a result, there is an increasing need for certainty about the working of IT systems. Interviewee F said that IT auditing not automatically translates to AI auditing because of the highly specific domain knowledge and expertise that is required to understand AI’s.

Interviewee F indicates that the concepts of risks and controls are comparable between AI and IT assurance, but the specific risks and controls differ since AI introduces new risks as explainability and the necessity to use data for system development. Examples of controls are validations though validation employees or separate teams and continuous monitoring of AI systems. Although, the profession of AI assurance, or AI auditing if you will, is still on a relatively small scale the expectation is that it will grow significantly according to interviewee F. Interviewee F repeated that assurance is about continuity and a control is functional separation of roles.

Interviewee F admits that looking into an equivalent of IT application controls is very different per type of model and algorithm is very difficult to standardise since all these controls are most likely to be different. Interviewee F continues with stating that it is very difficult to come to an audit standard that is applicable in a generalised situation and indicates that it is very dependent on the choices that were made during the development. That determines which controls are in force.

Only on a higher level it might be possible to do so.

However, interviewee F also described that making a specific, case depending, analysis is a recurring activity. Basically, each case is looked into in a somewhat customized way. The customization depends on the risks of the AI or algorithm for that specific case.

For very complex AI systems, black-box like models, it is very hard to provide direct assurance over the model itself. Often one needs to trust on the environment controls and processes around the development and usage of the AI model. However, it is never the optimal choice to rely on the latter type of controls since it is a very indirect way to say something about a model. Interviewee F hesitates to call it an AI audit when the audit cannot be detailed enough. Although, he stated that it depends on the audit objective and the scope as well. If a model cannot be audited itself then the elements of focus are the people who develop the AI, how they were trained, how the model was trained, how were the developers selected and accepted, was there a clear policy on AI usage and a developmental plan and policy?

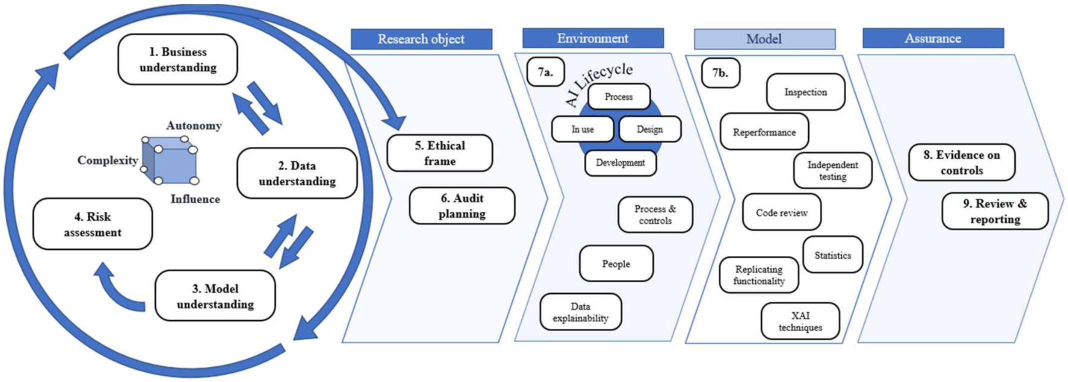
The responsibility of ethical AI's often lies on the developer's side, who has to explain which choices were made and why. Assurance over AI systems can include methods like data-oriented controls and output testing.

Interviewee F refers to something like an attestation perspective where an organisation has set up their own control framework to which an auditor investigates in what way the AI model is compliant to the organisation's statements. System descriptions and business practices are important information. Furthermore, AI assurance reports follow the somewhat same order and structure as IT assurance reports. The difference is mainly the type of risks and the controls that cover them.

Appendix 5 Validation interviews

Validation 1. Interviewee G

On 23-05-2024 at 13:30 – 14:30 interviewee G provided feedback on the developed artefact on AI auditing. Interviewee G is a manager at the IT assurance department of KPMG and was selected as a validator because of his expertise in IT auditing projects.



Interviewee G looked to the model from the IT audit lens instead of a Responsible AI oriented lens. His feedback was mainly upon the Environment, Model, and Assurance part. The state of the model presented to interviewee G is depicted below.

Interviewee G commented on the output of the investigations and audit techniques as ‘data explainability’. From an IT audit perspective, it is quite a relative and qualitative business to investigate training-datasets and what it might or might not say about an AI model. Furthermore, if a model is investigated, he found it important to get valid statements on the reliability since IT systems (in IT audits) cover a certain risk in which the belonging controls cover a ‘Process Risk Point’. Interviewee G added to this that when a system has found 68% of the faulty values it sounds like one third is missing which might ring immediate alarm bells. After the explanation that the values are going, in the case of CO1, through another control he admitted that this is not something you could possibly derive from the model above and emphasized the context.

Continuing on that comment, the most important feedback from interviewee G was to split step 8 in the assurance block. Step 8 was worded as ‘Evidence on controls’ to be in line with the literature. However, interviewee G suggested that to him there is a distinction between the output of the model ‘controls’, whatever these might be, and the controls based on the environment which are either set-up around the model or indirectly related to it. He added to the example of CO1 that if 68% of the risk is covered by the model and the other third is automatically pushed to another type of control such as a 4 to 6-eyes principle then the risk point is assumingly sufficiently covered. On the question which one would precede the other interviewee answered ‘definitely the evidence on the model controls’ since it would not make sense to continue if the output is really bad because then the result would be to select more type of controls to sufficiently audit something and make a statement on the environment. His suggestion was to split step 8 into the following:

8. Evidence on AI model controls
9. Evidence on process risk points

Interviewee G’s last comment was that most likely the output of the audit needs to be sensibly reported since higher management will receive the audit/assurance report. He pointed out that the researcher of this thesis might already have acquired more knowledge whereas an explainable item might not be as ‘explanatory’ to the non-tech-savvy people in higher management.

Validation 2. Interviewee H

Interviewee H is a Manager at the ITA department and was interviewed on Monday 27-05-2024 at 11:00 – 11:30 in order to validate the artefact through an IT auditing perspective. Interviewee H was asked for a validation interview because of his expertise in IT auditing projects at the ITA department of KPMG Eindhoven.

Interviewee H had no remarks on the visualization of the model nor on the choice of the activities after deciding to split the evidence on the controls into two parts after the suggestion of interviewee G. However, interviewee H had some clarifying questions on the contents and reasoning which were explicated during the interview.

The main remarks from interviewee H were that he did not really see the difference with IT auditing and suggested to describe this explicitly in the description part of the artefact. On the other hand, he admitted that it is ‘probably positive’ that the relative same activities take place since it is a proven profession in that way. The reasoning of interviewee H was that at its core both AI and IT auditing should be investigating how well risks are covered so it is about the reliability of the systems.

Furthermore, interviewee H mentioned to explain him the added value of this artefact and what would be the contribution of it because from seeing such a model itself it was not exactly clear to him. After an explanation he understood more and he suggested to put the focus on the parts that add value, or are different than IT auditing. Meaning, putting emphasis on the iterative first part and the difference between environment and model.

Appendix 6 AI usage

For this thesis two AI tools were utilized in the research. Besides the earlier mentioned **Co-Pilot** from Microsoft (Teams), the generative AI tool **ChatGPT** was used as well. The integrated tool Co-Pilot was mainly used to transcribe the interviews live in order to make the processing of the interviews more efficient. ChatGPT was used more broadly, but in an indirect way. This appendix explains per chapter how ChatGPT was used.

Introduction

Sentences that required more fluency in the abstract, problem indication, and problem statement were reviewed by ChatGPT. The output of ChatGPT was used to change the existing, self-written texts in order to improve the fluency, grammar, and general structure of the sentences.

Example questions are:
"Could you improve this sentence: {…}"
"What entails an introduction of a master thesis"
Furthermore, the introduction represents the phase in which the research proposal was written. ChatGPT was used, in combination with search engines on the internet and the University library, to orientate and play around with the topics of interest. For example, ChatGPT was asked the following questions "*What would be the best theory to measure explainability and transparency?"* Even though none of the suggestions were applied in this thesis, ChatGPT stimulated the researcher's creative thinking process. Questions of this nature shaped the orientation of the research and made the exploration on what is possible and viable to research more efficient. The topic itself was formulated together with KPMG and enabled questions like the example above.

Literature

For the literature research a module inside ChatGPT was used to make the research process proficiently more efficient. The module is called Consensus and derives knowledge from academic research banks, unlike ChatGPT. Consensus never provided texts that could be directly copied into this thesis, but rather recommended articles for the researcher to read by providing a small introduction. An example question for this is: "*Can you provide me with papers that describe what LIME and SHAP are?*"
If a research paper was not understood by the researcher, ChatGPT could explicate this

further and suggests the key findings from the article. ChatGPT could also process paragraphs of certain papers and if put in, ChatGPT automatically yields (without asking a question) a brief explanation on what the paragraph is about. This supported the researcher in judging the relevance and relation between articles. Lastly, ChatGPT was used to improve sentences through questions as:

"Could you improve this sentence: {..}"

"Can you rewrite this text in a formal and academic way: {...}"

The output of these questions was used to adapt and improve the existing texts.

For the literature exploration itself

Method

The use of ChatGPT in the Method section was mainly by asking ChatGPT questions like *"how would you commonly structure a master thesis' methodology"*. Thus, ChatGPT was of creative assistance to structure the text and of course improve the sentences in the same ways as the other chapters.

Results

A combination of Co-Pilot and the generative AI inside KPMG, that is based on ChatGPT modules, were used to transcribe and summarise the interviews. The interview summaries are not a direct output of AI since they were manually written and not only based on the transcription, but also the researcher's handwritten notes. AI significantly improved the efficiency of processing the results of the interview and transforming them into summaries and findings. Often, the transcription was put in the language model which resulted in a readable overview. Both the summary and the transcription were then used to write the interview summary.

Discussion and conclusion

The discussion and conclusion were supported by ChatGPT since it provided suggestions and improvements for sentences, structure, and grammar. ChatGPT was asked for example: "*What are common paragraphs in a discussion / conclusion chapter of a master thesis?"*.