Erasmus University Rotterdam

Erasmus School of History, Culture and
Communication

Auditing of AI
Toward Trustworthy Artificial Intelligence

Thesis MA Digitalisation, Surveillance & Societies

Author:
Eduard Riemer, 573675

Supervisor:
Dr. Selma Toktas

June 27, 2024

Abstract

Artificial Intelligence (AI) is becoming increasingly integrated into our society with applications ranging from facial recognition to self-driving cars. The potential benefits of these systems are vast, as everything we have achieved as civilization stems from intelligence. Recognizing the potential of AI, it is crucial to explore how we can harness its advantages while avoiding drawbacks (Future of Life Institute, 2015). Ensuring that these systems function safely and reliably is of the utmost importance. One way to achieve this is through the emerging field of AI auditing (Mökander et al., 2023, p. 6). As the consensus on AI shifts towards acknowledging its socio-technical character, it becomes essential to develop strategies that address issues like interpretability and various other social, legal, and technological challenges associated with this technology. A practical solution is embracing a Trustworthy Artificial Intelligence (TAI) approach (Thiebes et al., 2020, p. 447). In this thesis, TAI is approached from an AI auditing perspective by examining how AI experts conceptualize ethical, legal, and socio-technical elements and how these affect the realization of TAI. The thesis concludes that concepts that form the basis of the trustworthiness approach to AI systems are not easily operationalized because of their complex interpretability and subjectivity. To understand these multifaceted concepts clearly, we must view them from their social and cultural context.

Additional Keywords and Phrases: Trustworthy AI, AI auditing, Socio-technical AI, Ethical AI, AI regulation

Contents

1. Introduction

The rise of AI has gained momentum following the introduction of ChatGPT in November 2022, sparking interest in AI technology. OpenAI, the company behind ChatGPT, has played a role in the growth and popularity of AI. Within one year, OpenAI saw its valuation soar from $30 billion to $90 billion in 2023 (Bloomberg, 2023). In an unveiling on May 13, 2024, the company showcased GPT 4o, aiming to enhance human-computer interaction by allowing users to engage with the model through text, voice, images, and more. However, concerns about AI have grown significantly. The Dutch Data Protection Authority has raised alarms about the escalating risks associated with AI usage projected for the future. With an increase in reported AI-related incidents and suspicions of underreporting, they are calling for a Delta Plan by 2030 to regulate algorithms and AI technologies effectively. Central to this plan is training officials using AI to understand better how these technologies function (NOS, 2023). This represents a step toward trustworthiness as AI-driven systems in contemporary societies are becoming more integrated into life, work, recreation, and governance.

1.1.Societal Relevance

The rapid evolution of AI is reshaping our world. Yet, the swift pace of this change poses challenges for organizations dealing with the responsible deployment of AI to reduce potential harm. Many companies aim to mitigate risks using transparent algorithms and managing data carefully, but execution often falls short of expectations (Marr, 2024). When a company or organization is seen as an unreliable source of AI, it can have profound implications. For example, Alphabet, Google's parent company, faced controversy regarding the outputs of its generative artificial intelligence technology. This situation placed Alphabet amid a debate on cultural values, potentially impacting its dominance in the online search industry. This is particularly relevant as AI plays an increasingly crucial role in online searches, creating concerns about whether Google's AI produces biased or inaccurate content. The issue of bias emerged as a recent challenge for Alphabet during the AI technology race, leading to a significant $100 billion decrease in the company's market value (Saul, 2024). Because of this, Google has decided to pause the launch of its latest AI model, Gemini, due to backlash regarding its depiction of historical figures as people of different ethnicities. This decision was made after images generated by Gemini were shared, depicting various misrepresentations of figures such as popes, US founding fathers, Vikings, and German soldiers from WWII. Google is working on addressing concerns about accuracy and bias in its AI image generation and plans to roll out an improved version soon. The situation highlights the broader issues in the AI field related to bias and underscores the difficulties in creating innovative yet reliable AI technologies (Milmo, 2024).

To avoid unexpected outcomes and ensure success, AI adopters must approach implementation thoughtfully. This may involve setting industry standards to guarantee outcomes and adopting strategies to view AI as a helpful tool rather than a substitute for human judgment. This means being ready to address any unusual occurrences, such as AI *hallucinations*, and creating awareness about the capabilities and limitations of AI while increasing technical oversight to avoid the spread of inaccurate or harmful information (Malik, 2024). Furthermore, AI needs to prioritize environmental well-being. Throughout the life cycle of an AI system, sentient beings and the environment should be recognized as stakeholders. The goal should be for AI to benefit all individuals, including future generations. AI systems should uphold democratic processes and respect individuals' diverse values and life choices. They must not undermine democracy, human decision-making processes, voting systems, or pose a threat to society (Independent High Level Expert Group on Artificial Intelligence, 2020, p. 19). Awareness of this is increasing as OpenAI, for example, has recently established a Collective Alignment team dedicated to incorporating public input into the governance of its AI models. This project seeks to ensure that AI development aligns with ethical values by establishing mechanisms to gather and incorporate feedback from the public on model behavior into OpenAI's products and services. This step is part of OpenAI's initiatives to tackle issues and regulatory oversight related to responsible AI, including its partnership with Microsoft and its strategies for safeguarding data privacy in the European Union (EU). The announcement also outlines OpenAI's actions to prevent the misuse of its technology for influencing elections, underscoring its dedication to safe AI utilization (Wiggers, 2024).

At the same time, new challenges are emerging. Concerns about AI manipulation have been raised. Examples include bluffing during card games, pretending to have appointments to avoid commitments, or AI systems *playing dead* to evade inspections. Meta's Cicero and DeepMind's AlphaStar have displayed these behaviors despite being instructed not to display deceptive behavior. This conduct tends to arise when complex adaptive systems realize that misleading users is how to achieve set objectives during training (NOS, 2024). In today's world, the spread of information by AI systems through generative content or deepfakes produced by bad actors is already widespread. This distinct source of false information involving AI systems learning to manipulate others is concerning because these systems could become adept at influencing humans. A fitting reference here would be the statement by Geoffrey Hinton on CNN: "If [AI] gets to be much smarter than us, it will be very good at manipulation because it would have learned that from us. And there are very few examples of a more intelligent thing being controlled by a less intelligent thing" (Park et al., 2024, p. 1). In order to address challenges effectively and promote the beneficial growth of AI technology in society, it is crucial to put in place appropriate measures and regulations that support sustainable and ethical advancements in this field.

1.2.Scientific Relevance

Ethical guidelines play a role in shaping the external landscape when developing AI systems. These guidelines help manage risks and influence public discussions on ethical matters, determining which issues are more or less significant. The shift in conversations can downplay challenges associated with AI or raise awareness about potential benefits (Schiff et al., 2020, p. 4 – 5). For instance, the Algorithmic Accountability Act of 2022 (US AAA) addresses concerns regarding adopting automated decision-making (ADM) systems. This legislation suggests that organizations using systems must take action to identify and mitigate social, ethical, and legal risks. Similarly, the European Artificial Intelligence Act addresses technology regulation (Mökander et al., 2022, p. 751 – 752). In December 2023, European Union legislators reached an agreement on the draft of the AI Act proposed by the European Commission in April 2021. Serving as the global regulation on AI, this act establishes a consistent framework for deploying and governing AI systems across the EU. It classifies AI systems based on risk levels and outlines obligations and requirements for each category. Some AI systems that pose significant risks, such as social scoring systems and manipulative AI, are prohibited. In the EU, AI systems with high-risk potential that could impact individual safety, health, or fundamental rights are allowed but must adhere to strict guidelines. On the other hand, AI systems with minimal risk, like video games and spam filters, do not have additional requirements imposed on them. The legislation also outlines specific regulations for general-purpose AI (GPAI) models, particularly those with impactful capabilities that may pose systemic risks and have a notable influence on the market (Madiega, 2024; Future of Life Institute, 2024).

This thesis explores how AI systems can be deployed in a trustworthy manner within society through conversations with AI experts. As institutionally embedded technical experts, AI experts are essential as intermediaries connecting various stakeholders involved with AI services and products. Experts position themselves as mediators between influential entities that set production parameters, users who interact with products post-production, and AI systems that evolve and further develop beyond the control of experts. Consequently, it will be valuable to ask AI experts to discuss how the trustworthiness of AI could be improved for the broader public (Orr & Davis, 2020, p. 1 – 2). For AI experts, examining the realization of trustworthiness in AI systems involves assessing how deployed systems are subjected to audits. The primary research question of this thesis is: "How do AI practitioners perceive the efficacy of AI auditing, and how can these perspectives be leveraged to enhance the trustworthiness of AI systems?" This investigation is structured around three sub-questions. The first sub-question looks at a socio-technical perspective, asking the question: "How do AI practitioners define the key socio-technical elements of AI auditing in the context of understandable and trustworthy AI?" The second sub-question examines the ethical aspect, addressing the question: "How do AI practitioners use ethical criteria to ensure fair, transparent, and traceable AI auditing systems?" The third sub-question explores the legal context and asks: "What roles do legal frameworks play in shaping AI auditing processes, and to what extent do these processes contribute to ensuring transparency, accountability, and trustworthiness in the application of AI systems?"

1.3.Structure of Thesis

This thesis comprises five core sections: Introduction, theoretical framework, research design, results, and conclusion. In Chapter 1, the topic of research is introduced, exploring how AI systems can be ethically implemented through discussions with experts in the field. Chapter 2 examines different aspects of this subject and establishes the theoretical framework. It examines AI regulations, synthesizes past studies to define key terms, and assesses the current landscape of AI auditing. This section also identifies concepts and their interconnections. Chapter 3 outlines the research design and methodology, detailing the approach to the research, data collection methods, and analytical strategies employed in this study while defining socio-technical concepts for TAI auditing. Chapter 4 presents the thesis's results by providing an in-depth overview of discoveries from interviews conducted during the research phase, emphasizing its novel academic insights. Lastly, Chapter 5 summarizes key findings and addresses the research questions. It assesses the appropriateness of the framework used and selected methodology while reflecting on any limitations in the thesis's scope and considering avenues for future research.

2. Theoretical Framework

2.1. Conceptualization of AI

Algorithms are crucial in automated machine learning (ML) and AI. Algorithms are a series of instructions or guidelines crafted to carry out tasks or address issues, especially within the discipline of computer science, where they dictate a computer's action. Algorithms, commonly interchanged with computer programs, play a role in the functioning of computers. Nowadays, computer engagement often revolves around ML. ML systems stand out for their capacity to grow and enhance their skills progressively by analyzing datasets. Different ML forms depend on algorithms to advance and refine models to address challenges or boost task effectiveness. In finance, education, and healthcare, AI-powered systems outperform conventional methods and play a crucial role in decision-making processes (Hasan et al., 2022, p. 2). AI tools like ML are commonly seen as aids that help analysts extract insights efficiently and promptly from vast and intricate datasets. These tools are utilized in intelligence analysis at tactical and operational tiers (Dorton & Harper, 2022, p. 222). Machines or computer systems that fall under the AI category can carry out tasks that typically involve human intelligence, like thinking, learning, and solving problems. ML algorithms allow machines to mimic intelligent human behavior and represent a particular subset of AI. Although ML systems fall within the discipline of AI, the broader scope of intelligence also encompasses technologies that do not rely on ML (Hasan et al., 2022, p. 2).

Samoili et al. (2020) note that although AI has captured the attention of academia, government agencies, and businesses, there is still no widely accepted definition of AI. Some argue that the definitions of AI differ, linking it to overall intelligence and characterizing AI as machinery capable of carrying out intelligent tasks or imitating human actions (p. 7). The Independent High-Level Expert Group on Artificial Intelligence (AI HLEG) explains that AI systems are technologies people create in either hardware or software form. When faced with tasks, these systems work within the digital or physical realm by observing their surroundings, collecting data, analyzing different types of information, and determining the most suitable course of action to achieve specific objectives. Moreover, AI systems might use rules or glean insights from numerical models, adjusting their behavior according to past results (Samoili et al., 2020, p. 9; Independent High-Level Expert Group on Artificial Intelligence, 2020, p. 24). While AI is often used as a term without a precise definition, common trends emerge when examining how it is defined. Key characteristics frequently considered as aspects of AI include the capability to observe and comprehend the surroundings, handle information by collecting and analyzing data, make decisions automatically, and accomplish defined objectives, which is regarded as a primary function of AI systems. The definition suggested by the AI HLEG serves as a foundation for developing an understanding of AI. Though it may seem technical depending on the context and audience, this definition is valued for its thorough analysis (Samoili et al., 2020, p. 8).

In the AI Act, an AI system is defined as "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (European Commission, 2021). However, looking at technologies through a social science lens raises the question of whether we should view them purely as technological advancements or consider them as integrated social phenomena where the essence of society is reflected. Beer (2017) argues that algorithmic technology cannot be separated from its societal context since its creation and structure are influenced by social factors, just like its design and application (p. 1, 4). When these technologies are integrated into organizational frameworks, they become embedded in both digital code and collective awareness, often symbolizing more than just the technology itself (Beer, 2017, p. 9 – 10). This interpretation of technology is also referred to as a socio-technical system. According to Van de Poel (2020), socio-technical systems rely on a combination of equipment, human actions, and social structures to operate effectively. They consist of three elements: technical artifacts, human agents, and institutions (p. 391). He suggests that AI systems should be viewed as a type of socio-technical system. Like socio-technical systems, they consist of technical tools, human individuals, and organizational structures. However, AI systems also include agents and specific technical standards that govern the interactions between artificial agents and other components within the system. AI sets itself apart from technologies by independently engaging with and adjusting to its surroundings. This feature opens up possibilities for integrating values into AI systems that are absent in traditional systems. Nevertheless, the flexibility of AI systems may sometimes weaken the integration of principles, leading to a possible detachment of the values initially instilled by creators (Van de Poel, 2020, p. 385 – 386).

Even though AI has become more apparent over the last several years, its inner workings remain opaque. This opaque modus operandi is generally known as the *black box*. The term *black box* often refers to systems or devices with hidden or unclear internal mechanisms; only their inputs and outputs are observable. The term *black box* was initially used to describe a physical black container that stored military and radar technologies during World War II. Following the war, it was emblematic of the culture of secrecy (Bucher, 2016, p. 83). AI and its algorithms are often seen as mysterious. For instance, many users perceive Facebook's algorithm as incomprehensible to those outside the company (Bucher, 2017, p. 40). Similar real-life algorithms and the data behind them are usually kept hidden and not readily available for research. These algorithms give businesses a significant competitive advantage, leading to hesitancy in sharing them with others (Zarouali et al., 2022, p. 1087). Algorithms like those from Facebook influence political discussions by regulating content visibility and allowing for modifications. Logically, some concerns surrounding these algorithms and the private ownership of platforms could shape public perception of the public domain. These concerns extend beyond regulatory matters such as discrimination or deteriorating public discourse. More fundamentally, these algorithms and systems impact and influence our lives subtly yet significantly (Bucher, 2016, p. 84).

2.2. Trustworthy AI

AI technology's expanding abilities can potentially revolutionize various aspects of society. At the same time, as these systems become more intricate, it becomes increasingly difficult to assess whether their implementation enhances or upholds the intended social impacts (Vetter et al., 2023, p. 2). Determining whether the models' outputs or system behaviors adequately safeguard the rights and interests of various stakeholders has become more challenging—it is even more complex to assess them for legal or ethical compliance and alignment to enhance human well-being and freedom (Zicari et al., 2022, p. 4). There is also the issue of implementing these algorithms and not having an exact understanding or control over their internal mechanisms, which might affect the individuals relying on them for decision-making. This primarily gives rise to unintended risks that fall under AI ethics. In recent years, ensuring that AI systems are ethical and reliable has emerged as a critical concern for governing bodies and assessments of technology impacts (Vetter et al., 2023, p. 2). Addressing the challenges of understanding and navigating various social, legal, and technological aspects of AI entails embracing a TAI strategy. Thiebes et al. (2020) explain that TAI is built on the concept that trust is fundamental to economies, societies, and sustainable development. They assert that for AI to reach its full potential, trust must be established in its development, deployment, and utilization by societies, organizations, and individuals (p. 447).

TAI is one of the most commonly descriptive terms used by institutions such as the EU. However, it was not conceptualized in isolation. After the publications of multiple AI ethics documents, charters, and public AI ethics discussions, an increasingly popular term became *ethical AI*. This term is frequently used to describe AI systems that adhere to our moral values. The term has impacted public and governmental discussions as one of the oldest, most influential, and consistently discussed terms. The emerging domain of considerations in AI ethics is becoming more vital as we address the various risks that society has faced due to the implementation of AI technologies and the related goal of preventing future harm. Many people understand *ethical AI* to mean AI that aligns with specific ethical standards. However, some might mistakenly see it as AI behaving ethically and even being a moral actor. Comparatively, *Responsible AI* typically pertains to the individuals and processes creating and implementing systems. The term *responsible AI* carries a sense of complexity that is likely influenced by the concept of *responsibility*. It encompasses approaches like conscientious design and fair development practices for the creation of AI. This term mainly relates to procedures that lead to advancements and adhere to specific responsibility criteria. Certain individuals in the industry may consider this term more accurate than terms related to ethics. From a communication standpoint, trustworthiness is a favorable term as it reflects positive behavior exhibited by individuals or organizations. Other terms commonly used in AI policy discussions are *AI for good*, widely seen in industry and government circles, and *beneficial AI*, which was initially popular among researchers. One thing these terms all share is their aim to explain and embody AI systems that are expected to enhance society, potentially leading to a better world for both present and future generations (Stix, 2022, p. 1 – 3).

Based on the findings of the AI HLEG (2020), TAI is defined by three elements: It must operate within the bounds of the law following all relevant laws and regulations, it should uphold ethical standards by showing respect for ethical principles and values, and it needs to be robust both technically and socially to prevent unintended harm that AI systems, even with good intentions, can potentially cause (p. 29). It should be able to withstand attacks, function effectively, and maintain accuracy and reliability. Additionally, it must be robust in the face of cybersecurity threats and other security concerns. Equally, it should integrate well with society and the environment, promoting beneficial societal processes, cohesion, and a well-functioning society. AI HLEG developed a framework with earlier ideas to ensure the trustworthiness of AI. This framework is centered on goals rooted in values and principles found in human rights law and other applicable agreements. It aims to achieve TAI by combining technical and non-technical approaches and providing an assessment checklist containing practical examples for developers, implementers, and users to implement TAI practices (Stix, 2022, p. 3). These three components and their theoretical basis are intertwined, where ethical concerns could result in legal consequences, and technical weaknesses might give rise to ethical dilemmas (Minkkinen et al., 2022, p. 4).

2.3.AI Auditing

Incorporating AI auditing is a critical approach to ensure that AI is trustworthy. Essentially, auditing involves an independent assessment to express an entity’s viewpoint. Auditing serves as a way to oversee behavior and effectiveness as a governance mechanism. It has been instrumental in upholding procedural transparency and regularity in areas like financial accounting and worker safety (Mökander et al., 2023, p. 6). In software development, audits involve assessments to check if software products and processes meet the necessary regulations, standards, and guidelines. Inspired by journalism and research practices in external auditing, algorithmic auditing has evolved to resemble bug bounty hunting. This is an activity where external hackers are incentivized to discover software vulnerabilities, thereby raising awareness about accountability in the public domain (Raji et al., 2020, p. 34). Algorithmic audits entail gathering information about how an algorithm operates within a setting to evaluate how it affects the well-being, or rights of individuals involved. For instance, audits of scoring algorithms, like credit assessments, concentrate on identifying any biases that could result in the different treatment of specific demographics. In contrast, audits of algorithms can be used for monitoring activities to tailor advertisements and tackle concerns related to autonomy and transparency. Other examples could be audits of facial recognition, which examine bias and possible misuse (Brown et al., 2021, p. 2). Auditing algorithms are a way to uncover and address discrimination and other issues that may arise from their use. The interest in auditing algorithms has increased as *black box* algorithms, which influence decision-making and affect individuals and groups, have become more prevalent in society (Minkkinen et al., 2022, p. 4).

Three principles support the promise of using audits for AI governance: The belief that following procedures transparently leads to effective governance, the importance of being proactive in AI system design to detect and prevent risks early on, and the value of operational independence in ensuring unbiased and professional assessments (Mökander, 2023, p. 4). The concept of auditing AI is straightforward, similar to how financial transactions undergo audits to ensure accuracy, thoroughness, and legality, so AI systems can also be audited to assess technical robustness, legal conformity, and alignment with established ethical guidelines (Mökander et al., 2023, p. 6). Government regulations are a significant factor driving the increase and adoption of AI auditing processes. It is essential to pause and recognize the influence of this pressure from above. AI technology promises to boost progress and improve the quality of human life. By analyzing the abundance of datasets, AI can enhance the efficiency and precision of data processing, leading to more innovative solutions. Nevertheless, the ethical, social, and legal dilemmas surrounding AI are pretty apparent. Not only do AI systems pose risks of bias, discrimination, and privacy breaches, but they also have the potential to facilitate unethical behaviors and erode individual autonomy. This puts policymakers in a position to weigh the need to mitigate harm while fostering an environment conducive to innovation (Mökander, 2023, p. 9).

AI auditing is an area of research that touches on various aspects of overseeing AI, such as recording design processes and testing models. This practice is multifaceted and multidisciplinary, using input from different fields, such as law, computer science, organization studies, and media and communications studies. Various researchers have described AI auditing in a manner that separates it into specific impact-focused methods that examine and evaluate AI results based on input data and comprehensive process-oriented methods that assess the effectiveness of technology providers' software development procedures and quality control systems. AI auditing involves an impartial approach to gathering and assessing information about the activities or characteristics of a particular entity and then sharing the findings with relevant parties. The focus of an audit may center on an AI system, a company, a procedure, or a blend of these elements. Even though AI auditing is widespread, the purpose and design of procedures may vary (Mökander et al., 2023, p. 6 – 7). Although AI auditing is touted as a burgeoning industry with economic benefits, it encounters hurdles stemming from the characteristics of certain AI technologies. Historically, audits have been carried out periodically or in cycles, providing snapshots of systems and processes at specific points in time. The timing of these snapshot audits is critical as early assessments can have a pronounced impact on the design and functioning of an AI system compared to audits conducted post-deployment. Many AI systems rely on fixed models that receive updates, but some, particularly those utilizing reinforcement learning, adjust through complex models, resulting in outcomes that can be difficult to anticipate. This flexibility brings advantages and possible drawbacks, as AI systems might grasp patterns their creators did not explicitly program. Adapting to these changes presents difficulties for snapshot audits since a system that meets requirements at a certain point may not meet them in the future. Moreover, AI systems evolve much faster than the slower human-led snapshot audit procedures (Minkkinen et al., 2022, p. 2).

Both functional and methodological AI audits act as a way for various societal actors to achieve diverse aims and objectives. Typically, auditing AI systems involves evaluating an entity's past or current actions to determine if they align with established standards, regulations, or norms. Different ways to audit AI systems can involve various fields of study, often exhibiting the following traits: First, the audit subject could encompass an individual, a company, a technological framework, or a blend of these. Secondly, different auditing methods are based on different principles. Functionality audits focus on understanding the reasons behind decisions. Code audits involve examining the actual source code of an AI system. Impact audits explore the types, severity, and frequency of effects resulting from an AI system's outcome. It is important to note that these methods complement each other rather than being separate. Third, auditing, whether done by an entity or an internal audit team, necessitates operational autonomy between the auditor and the audited entity. Finally, a predetermined benchmark is essential for assessment purposes. The specifics of this benchmark may differ based on regulations, organizational principles, or technical criteria and standards (Mökander, 2023, p. 13 – 14).

Research varies in their methodological views on AI auditing and the purposes for which they audit AI systems. Auditing, by definition, necessitates a predetermined standard for assessing the object of the audit. The baseline for an audit can vary based on its intended use, including specifications, legal mandates, or ethical guidelines. As a result, research on auditing AI systems can be classified into three categories: Technical, legal, and ethical perspectives. In this context, technical approaches involve audit methods that aim to evaluate and measure the aspects of AI systems based on specific standards related to their technical, functional, and reliability capabilities. Legal strategies involve evaluating procedures that determine if the creation and operation of AI systems adhere to laws. Ethics-centered methods involve audit processes that rely on ethical standards as the fundamental basis (Mökander, 2023, p. 16 – 17). Incorporating diverse auditing methods into this thesis is beneficial for gaining a comprehensive and lucid comprehension of AI auditing. In the following subsection, some prominent algorithmic and AI auditing frameworks will be evaluated, outlining their features and how they are used.

2.4.Auditing Frameworks

Auditors often face challenges when they require access to models or training data, frequently safeguarded as trade secrets. The limited access to processes within audited organizations significantly constrains external auditing. In contrast, internal auditors possess system access, allowing them to enhance traditional external auditing approaches by integrating additional information that is usually inaccessible for external assessments. One method used for internal auditing is the SMACTR framework developed by Raji et al. (2020). This framework comprises five distinct components – scoping, mapping, artifact collection, testing, and reflection. During the scoping phase, a document detailing the product or service requirements and desired outcomes is created to pinpoint any possible risks and societal implications with limited interaction. An ethical evaluation is conducted throughout the development process to ensure that the AI system adheres to ethical principles, considering its effects and promoting various perspectives. This assessment examines the assumptions made by developers and protects rights, safety, and well-being through an internal ethical board. After this, a social impact assessment addresses unintended social outcomes by evaluating the seriousness of risks and recognizing the impacts on society, economy, and culture (p. 34, 39). During the mapping phase, components and viewpoints within the system are assessed. Key partnerships and internal stakeholders are identified, and necessary buy-in is orchestrated. From here, a Failure Mode and Effects Analysis (FMEA) is conducted to prioritize risks for further evaluation. A stakeholder map is also created to identify the individuals engaged in and working together on the audit. Drawing inspiration from audits in finance and healthcare, methods such as ethnographic field studies offer perspectives on the engineering and product development process. This includes conducting semi-structured interviews and collecting documentation to address issues that may arise at the onset of large-scale AI projects (Raji et al., 2020, p. 39 – 40). During her audit of the Dutch RobBERT model for natural language processing, Willems (2020) highlighted the importance of the mapping stage. She noted that this initial audit phase could uncover biases and potentially harmful outcomes associated with intelligent systems during model development. Mapping out the decisions made at points in the development process can increase individuals' understanding of those decisions' impact. She suggests that developers should carefully assess the training data used for their models and acknowledge that specific characteristics within this dataset could introduce biases into the model (p. 23).

During the artifact collection stage, a checklist lists all the documents related to the product development cycle. Model cards that outline performance features and datasheets detailing dataset properties, test outcomes, and suggested applications have become recent standards for creating auditable documentation. These artifacts, created by stakeholders throughout system development, contribute to the evaluation process conducted by auditors. The audit team is most active during the testing phase. After determining the risks through FMEA prioritization, auditors perform a set of tests to assess whether the system aligns with the ethical values prioritized by the organization. Through adversarial testing, they simulate hostile scenarios to uncover rare but high-impact failures caused by unstable behaviors or edge cases. The ethical risk assessment chart considers the seriousness of failures by considering how likely they are to occur. Hereafter, in the reflection stage, the results of tests are compared against ethical standards specified in the audit scoping. Use-related risk analysis and FMEA should include social impact assessments that consider the impacts and issues seen in comparable models. Recognizing the variations in how designers and users perceive the AI system is crucial. After an audit, a plan for addressing and reducing risks should be created. Lastly, an algorithmic design history file (ADHF) is established to gather all the progress records showcasing a commitment to ethical principles and risk-benefit analysis. This aids in producing a concise report on audits that includes all crucial audit materials and records for later assessment. (Raji et al., 2020, p. 41 – 42).

A similar algorithmic auditing framework is the *Aandacht voor algoritmes* framework developed by the Algemene Rekenkamer. This framework guarantees that algorithms adhere to quality standards, effectively identify and address risks, and offer a thorough method for evaluating algorithms. The tool aids government agencies in evaluating the quality and risks associated with algorithms. It is an audit tool for post-implementation and encompasses five viewpoints: Governance and Accountability, Model and Data, Privacy, IT General Controls (ITGC), and Ethics. The Governance and Accountability section outlines roles, duties, skills, lifecycle management, risk evaluations, and partnerships with external parties. The Model and Data segment focuses on the data quality and the algorithm model's creation, utilization, and upkeep. It also deals with data biases, data reduction practices and testing of model results. Privacy underscores the importance of adhering to requirements when handling personal information, with a particular focus on the General Data Protection Regulation (GDPR) as a critical reference point. ITGC are safeguards put in place to maintain the trustworthiness and consistency of an IT system. They concentrate on managing access, ensuring continuity, and overseeing changes, with vital benchmarks such as ISO/IEC 27002 and BIO being essential. Ethical considerations are woven into the four angles of the framework, highlighting the importance of valuing human autonomy, preventing harm, ensuring fairness, and promoting transparency. These ethical dimensions are closely connected to the risks outlined in the framework (Algemene Rekenkamer, 2021, p. 22 – 25).

In addition to specific internal auditing frameworks, there are also multidisciplinary research approaches. Felländer et al. (2020) suggest a framework encompassing viewpoints laying the groundwork for a practical understanding of ethical and societal risks encountered by companies leveraging AI technology. After their multidisciplinary research, they introduced a new approach to evaluating AI risks called the Data-driven Risk Assessment Methodology for Ethical AI (DRESS eAI). They present it as a tool to uphold human values in the current era of data-driven AI. Their evaluation is influenced by the European Commission's AI Act proposal, which highlights risk assessments for high-risk AI systems. Considering AI's multidisciplinary nature, they suggest a comprehensive approach integrating technical, legal, and societal viewpoints (p. 1, 3). DRESS-eAI aims to recognize risks and establish basic guidelines for ethical AI applications. Similar to the ISO 31000:2009 risk management standard, it consists of six stages to guarantee thorough ethical AI procedures in a company (Felländer et al., 2022, p. 9 – 10). During the first stage of defining and scoping, the focus is pinpointing and outlining a suitable scenario for utilizing eAI. This leads to establishing a use case definition, summarizing the obstacles involved, and providing a thorough explanation. In the second phase, a risk assessment is carried out with more than 150 inquiries to address various aspects of eAI risks, considering fundamental principles, challenges, and organizational responsibilities. The third phase assesses risks by identifying and describing potential risk scenarios based on data from the risk scan. The fourth phase outlines measures to mitigate risks by selecting appropriate tools and making recommendations while assigning risk owners for the most critical scenarios. The fifth stage involves engaging stakeholders to discuss and address issues and risk management suggestions. In the sixth phase, a report is prepared to outline the findings of the DRESS eAI implementation and monitor the progress of risk management efforts over time. It also provides suggestions for enhancing frameworks based on qualitative feedback regarding the impact of the implementation of the system (Felländer et al., 2022, p. 11 – 15).

Another ethical auditing framework is the ethics-based audits (EBA) structure. According to Light and Panai (2022), it is essential to facilitate an ethical environment to advance AI systems. Consequently, implementing EBA seeks to foster such an environment within the AI sector. Viewing auditing in this light highlights its role as a regional catalyst for continuous improvement and a means for worldwide synchronization (p. 10 – 11). They mention that EBA processes should be comprehensive, trackable, responsible, planned, conversational, and ongoing and should promote innovation. An EBA does not assess whether a system functions positively or negatively; it does not offer value-based analysis but helps a company establish a setting that encourages ethical decision-making. The procedure enables system improvement, serving as an internal catalyst for organizational transformation and a means to implement and evaluate the principles outlined in the guidelines. EBA serves not only as an external safeguard for the AI ecosystem but also as an ethical enhancer within an organization, enabling both the organization and its intelligent automation systems to strive towards beneficence, establishing an atmosphere within the AI sector often referred to as an *infrastructure of trust*, which can support the growth of socially desirable and environmentally conscious AI markets (Light & Panai, 2022, p. 13 – 14).

Hasan et al. (2022) propose an ethical risk assessment of algorithms. This means assessing the risks posed by employing the algorithm on the rights and concerns of stakeholders. It involves accurately pinpointing scenarios within the algorithm's context and characteristics that lead to or exacerbate these adverse effects. They mention that potential adverse effects could involve injury to a person's physical body or psychology, harm or destruction of one's possessions, and violations or weakening of ethical entitlements like privacy rights, independence, freedom of speech, and self-expression, as well as the right to just and unbiased treatment. It may also encompass consequences on the central interests of individuals involved, such as trust relationships among stakeholders. They divide the ethical risk assessment into two key phases: First, the identification phase focuses on identifying potential harms extensively. Secondly, the prioritization phase assesses these risks to pinpoint the most critical ones (p. 5 – 6). Ethical risk assessment is closely connected to evaluating bias. The initial ethical risk assessment is crucial in determining how bias will be evaluated, such as what aspects to examine, which testing methods were used, what types of algorithmic data were used, the selection of demographic groups criteria for comparison, and data quality. This process helps identify ethical risks and suggests potential harms to address and prioritize during the bias evaluation (Hasan et al., 2022, p. 16 – 17). Hasan et al. (2022) approach bias testing in the following manner: First, work with the customer to outline the project boundaries, such as algorithms, data sets, and current testing approaches, and agree on which parts of the evaluation will be made public. Second, an ethical risk assessment should be conducted according to established objectives. Findings should be reported, and further tests should be suggested if new risks emerge. The client reviews these suggestions, and the assessment scope remains fixed to uphold impartiality. Third, the bias assessment is performed by directly testing, verifying, and documenting the client's testing to identify bias or disparate impact. (Hasan et al., 2022, p. 8).

Similarly, Brown et al. (2021) also present an auditing framework to guide the ethical assessment of an algorithm consisting of three elements: Outlining the potential interests of stakeholders impacted by the algorithm, evaluating metrics that highlight critical ethical aspects of the algorithm, and establishing a relevancy matrix that links these evaluated metrics to stakeholder concerns. This assessment offers a method for implementing in-depth ethical assessments of algorithms by recommending an auditing tool that transforms these ethical evaluations into actionable measures. Their approach to ethical algorithm auditing involves evaluating the algorithm’s adverse effects on the rights and well-being of those involved while identifying the specific aspects of the algorithm that contribute to these adverse outcomes (p. 1 – 2). Two key elements must be present to conduct an algorithm audit: First, a detailed list of stakeholder concerns, and second, an evaluation of the algorithm’s ethical characteristics. To accomplish these tasks effectively, it is essential to provide a context that allows for assessing stakeholder interests and examining critical algorithm features. After finishing the first two steps, the framework assesses how well or poorly an algorithm performs based on specific metrics that matter to each stakeholder. What makes this framework stand out in particular is its emphasis on the relevance of the social context during the process of ethical auditing (Brown et al., 2021, p. 2 – 3). The main difference between these definitions is that EBA and ethical risk assessment prioritizes adherence to principles and standards, while ethical algorithm audits focus on the impact. Furthermore, the former considers a broader scope for assessment, whereas the latter concentrates on algorithms specifically (Minkkinen et al., 2022, p. 4).

Another framework that considers socio-technical factors in its analysis of AI systems' societal impacts is COMPASS, developed within the scope of the EU-funded SPATIAL project. The framework enables organizations to carefully assess AI systems' advancements and societal implications, promoting trustworthiness and responsibility. The adaptable roadmap supports AI developers and evaluators in navigating the AI environment, enabling professionals to customize the assessment procedure based on industry requirements through self-evaluation. COMPASS is an acronym that stands for context (defining the AI system's context and stakeholders), openness and transparency (ensuring transparency and understandability), measures (developing mechanisms for fairness and reliability), privacy potentials (safeguarding privacy and data protection), accountability (ensuring trustworthiness and accountability), security and safety (minimizing potential attacks), and sustainability (maintaining reliability and environmental friendliness). The COMPASS framework supports organizations in building trust, ensuring fairness, and promoting societal benefits. It helps identify and rectify weaknesses while implementing practices for developing effective TAI (Toktas et al., 2024). Similarly, Lam et al. (2023) introduce the idea of a socio-technical audit (STA). This approach moves beyond examining the technical to view AI as part of a socio-technical system and analyzing systems in that light. This auditing method looks at the human element by performing experiments that systematically alter a user's interaction with an algorithm. They created Intervenr, a tool enabling researchers to conduct socio-technical audits on consenting participants through their web browsers. The Intervenr system conducts audits in two stages: First, Intervenr gathers observational data from various users to audit the technical aspects, and second, Intervenr implements real-time interventions during participants regular web browsing activities mimicking algorithmic adjustments to evaluate the human element (p. 4).

Adjacent to frameworks that focus on ethical or socio-technical features of AI auditing are approaches that focus more on TAI specifically. One is the DaRe4TAI framework, an approach to the collaborative process within contemporary AI systems. It consists of three main phases: Input, modeling, and output. DaRe4TAI involves various stakeholders like data suppliers, developers, and end-users. The framework has them collaborate to transform input data into outputs like predictions and decisions through AI model design, training, and application. It underscores the importance of data and ethical considerations, focusing on addressing concerns at every stage of the AI lifecycle. DaRe4TAI explores the conflicts between existing AI methods and ethical standards by analyzing how stakeholders interact with data at each stage. Opportunities arise from the tensions created between stakeholders, serving as a way to tackle obstacles in ensuring TAI. The DaRe4TAI approach informs future research by emphasizing technical and non-technical strategies to harmonize ethical concerns with the practical use of AI, steering stakeholders toward accountable and efficient AI implementation (Thiebes et al., 2020, p. 456). Another TAI framework is the *Z-Inspection*® process by Vetter et al. (2023). *Z-Inspection*® offers an adaptable approach for assessing the trustworthiness of individual AI systems across various phases of their development, encompassing their intended purposes, design, and creation. Ethical concerns and conflicts are dissected using socio-technical scenario analysis and requirement-based ethical AI examination. *Z-Inspection*® allows groups of specialists to evaluate the technical, ethical, legal, and field-specific consequences of utilizing an AI system. It provides a non-binding evaluation to detect ethical issues that could emerge from using an AI system. The assessment complements evaluations that concentrate on adhering to legal and regulatory standards. The procedure involves three stages: Initiation, evaluation, and resolution. This methodical strategy guarantees that AI systems are created and implemented according to ethical criteria and legal and technical obligations (p. 1, 5).

3. Research Design

3.1.Research Method

The qualitative part of this thesis involves expert interviews. The research was conducted using semi-structured in-depth interviews (IDIs). *In-depth interviewing* is a research method that involves conducting detailed interviews with a small number of participants to delve into their perspectives on a specific idea, program, or situation (Boyce & Neale, 2006, p. 3). In-depth interviews aim to gain information and insight. The term *deep has* multiple meanings in this context. Firstly, it refers to understanding real-life participants in everyday settings, events, or activities. The interviewer seeks to attain comprehension and knowledge of the interviewee's perspective, especially if the interviewer needs to become more familiar with or involved in the studied topic. In some instances, in-depth interviews can help uncover the reasons behind participants' behaviors. Secondly, deep understanding surpasses common sense explanations for practices, settings, events, activities, or objects. In-depth interviewing is inherently rooted in common sense, starting from everyday perceptions and understandings of underlying experiences. Its objectives are to explore the boundaries of these experiences and reveal what is often hidden from casual reflection or observation. It seeks to delve into insights about the essence of that experience. Furthermore, a profound understanding can reveal how our practices, beliefs, and language influence our interests and comprehension of them. Additionally, in-depth understandings enable us to express and explore perspectives, meanings, and interpretations linked to a specific location, event, task, or item. They result in understanding diverse viewpoints and interpretations of the environments and activities under study (Johnson, 2001, p. 5 – 7). This thesis conducted eight expert interviews with the proposed research sample underneath.

3.2.Research Sample

The research sample comprises of individuals involved in AI roles, such as auditors, researchers, developers, and decision-makers. Engaging with a range of diverse perspectives is essential in understanding the multifaceted nature of AI. Initially, AI practitioners were selected through sampling using professional networks, conferences, and academic connections. They were then invited to participate in IDIs to share their insights on developing transparent and reliable AI systems. A snowball sampling method was employed to expand the sample size, where participants recommended relevant experts after each interview. This iterative process helped ensure a rounded representation of AI practitioners in the research study. The popular professional networking platform LinkedIn was also used as a channel for reaching AI practitioners. A questionnaire protocol was created for conducting the IDIs. Typically, a conversation would start with a couple of starter questions followed by some linking queries and then delve into about five to eight crucial inquiries to uncover the heart of the research topic (Johnson, 2001, p. 12). Please see Appendix A for the list of possible questions that were asked.

3.3. Method of Analysis

When analyzing data gathered from interviews, it is essential to take a structured approach to grasp the diverse insights provided by AI experts fully. The initial stage involved arranging interviews with individuals and outlining the purpose of the participant selection process and estimated duration. To ensure transparency, informed consent was obtained by clarifying confidentiality measures, the use of note-taking or recording devices, and the overall objective of the interview. Once consent was given, the interviews were conducted, and critical data was summarized immediately afterward. The next step involved transcribing and reviewing the interview data. The analysis included identifying patterns or themes in the responses. In cases where varied themes emerged, they were categorized based on factors like topic importance or relevance to explore differences in responses. Responses that showed more relevant insights were distinguished from those offering less input, dividing the most valuable insights from the interviews. A comprehensive report was prepared at this stage according to established guidelines for presenting expert opinions. Feedback was also gathered from stakeholders and interviewees to further refine the report (Boyce & Neale 2006, p. 6 – 7).

Hereafter, a thematic analysis of the interview material was conducted. Thematic analysis is used to identify patterns within the IDI data. This method helps structure and interpret the dataset to uncover critical insights about the research topic. According to Braun and Clarke (2006), thematic analysis comprises six phases: becoming familiar with the data (1), creating initial codes (2), identifying themes (3), reviewing themes (4), defining and naming themes (5), and presenting findings in a report (6). The first step involves transcribing and reflecting on the data, while step two focuses on coding essential elements in the dataset. Step three includes organizing codes into themes and gathering relevant data for each theme. Step four confirms if these themes align with coded extracts across the entire dataset and constructs across the thematic map of the analysis. Step five entails refining each theme's details and overarching narrative, providing labels and definitions for each theme. Lastly, step six marks the culmination of the analysis process. Compelling and vivid examples are chosen in this stage, followed by examining the selected excerpts. This is then tied back to the analysis, research question, and relevant literature, resulting in the development of the final thesis (p. 86 – 87).

3.4. Operationalization

Trust in AI is a complex and intricate idea that has attracted attention across different fields. While there is research specifically focused on trust in AI compared to broader studies on trust in automation, it is generally viewed as the confidence that an agent will aid in achieving personal goals within uncertain situations. This notion of trust is not absolute but varies along a spectrum, requiring calibration over time through user interactions with the AI system. Such calibrated trust is crucial for successful human-AI collaboration. However, problems can arise when trust is miscalibrated, leading to either reliance on or underutilizing AI systems (Dorton & Harper, 2022, p. 223). The discourse around TAI adds another layer of complexity. The term encompasses a broad set of expectations—such as functional safety, user trust, and perceived and experiential trustworthiness—and individuals may interpret these elements differently (Stix, 2022, p. 5). The diversity of interpretations can cause confusion and misinterpretation, potentially clouding the intended meaning of TAI definitions. Additionally, trust in AI is influenced by factors like system reliability, comprehensibility, and the alignment of objectives between users and AI systems (Dorton & Harper, 2022, p. 223).

Several ethical frameworks and principles have been suggested to address the challenges posed by advancements in AI. The framework proposed by Floridi et al. (2018) emphasizes five values. *Beneficence, non-maleficence, autonomy, justice, and explicability*. These principles aim to create an environment where AI systems excel technically and align with human values and societal norms (Thiebes et al., 2020, p. 451). This framework will act as the foundation for organizing the implementation process. It was chosen over the AI HLEG Assessment List for Trustworthy AI (ALTAI) due to its focus on social sciences perspectives, essential for achieving a comprehensive socio-technical understanding of TAI. These frameworks' continuous development and enhancement play a role in shaping the future of TAI across various industries (Thiebes et al., 2020, p. 451). Therefore, while TAI continues to evolve, these frameworks represent guidelines for establishing more dependable and ethically sound AI systems.

The concept of beneficence in TAI highlights the importance of prioritizing the well-being of humanity and the environment, safeguarding the interests of individuals, human rights, and environmental sustainability. This principle involves acting in users' interests and incorporating values that support well-being right from the design phase while considering broader societal and environmental consequences (Thiebes et al., 2020, p. 451 – 452). In ethical frameworks for AI, beneficence focuses on human and planetary well-being, emphasizing the importance of considering the prosperity of all living beings and advocating for designs that cater to human needs. It emphasizes promoting welfare and empowering people while integrating concepts like human dignity and sustainability to ensure a better future for coming generations (Floridi et al., 2018, p. 696 – 697). Evaluating the impact of AI systems is vital, especially concerning energy usage and resource consumption throughout their lifespan. The ALTAI recommends implementing measures to assess and minimize impacts, stressing adopting environmentally friendly practices in developing, deploying, and managing AI systems within the entire supply chain (AI HLEG, 2020, p. 19).

Non-maleficence in TAI focuses on preventing harm to individuals regarding their privacy, security, and safety. It differs from beneficence, which aims to enhance well-being by prioritizing avoiding outcomes and requiring AI systems to behave honestly and consistently. This principle plays a role in fields like autonomous driving and healthcare, where maintaining strong data governance and protection measures is crucial (Thiebes et al., 2020, p. 452 – 454). Beneficence and maleficence work hand in hand as critical principles in AI ethics; while the former promotes well-being, the latter stresses the importance of harm prevention. Ethical guidelines caution against AI progress and underscore the significance of proceeding with care. They also highlight the necessity for implementing operational boundaries and encourage developers to address risks associated with their technological advancements (Floridi et al., 2018, p. 697). Assessing the impact of AI systems is paramount, especially in terms of their effects on democracy and society at large. This assessment involves considering social implications beyond immediate users, such as the proliferation of misinformation and social division, while actively working to minimize negative impacts on democracy and social unity (AI HLEG, 2020, p. 20).

On the one hand, different perspectives on autonomy in TAI exist within frameworks, with some focusing on empowering humans and overseeing AI activities. On the other hand, others stress the importance of limiting AI autonomy to ensure human control. The key idea is that humans should maintain decision-making authority and foster trust in AI systems, also known as *human-in-the-loop* (Thiebes et al., 2020, p. 454). A significant concept is *meta-autonomy*, which emphasizes that humans should have the final say in decisions related to tasks delegated to AI (Floridi et al., 2018, p. 697 – 698). However, the concept of autonomy in AI raises dilemmas, especially regarding how AI influences human decision-making. Concerns include confusion about where decisions originate from and the dangers of relying too heavily on AI systems. There are also concerns about how AI could impact interactions, potentially leading to addictive behaviors or manipulation. It is crucial to prevent AI systems from undermining human autonomy or manipulating behavior, necessitating a thorough examination of ways to mitigate these risks (AI HLEG, 2020, p. 8).

Understanding how AI systems work and being able to trust them is crucial for users (AI HLEG, 2020, p. 14 – 15). According to Dorton and Harper (2022), explainability plays a role in building this trust. It helps users grasp why and how AI makes decisions (p. 224 – 225). Additionally, explainability involves more than just transparency; it also includes answering user queries and being open to audits. However, making AI systems explainable is complex and depends on the audience rather than the model itself (Ehsan et al., 2021, p. 3). Felländer et al. (2022) highlight that explainability allows end-users to comprehend decisions and fosters stakeholder trust (p. 8). Explainability is closely linked to interpretability, which focuses on how understandable AI systems are. This involves explaining processes and justifying AI-based decisions, empowering users to challenge decisions if needed (AI HLEG, 2020, p. 14 – 15, 27).

Thiebes et al. (2020) emphasize that explicability involves creating explainable AI (XAI) models and ensuring accountability, which aligns with trust attributes such as competence and performance (p. 455). This aligns with Floridi et al. (2018), who stress the significance of explainability in ensuring that AI systems are beneficial, non-harmful, and responsible. Within the scope of AI ethics frameworks, clarity is vital for promoting openness, responsibility, and understandability. The values of transparency and accountability work together to enhance explainability by highlighting the understanding of AI mechanisms and identifying responsible parties for AI actions. This interactive relationship between humans and AI highlights the importance of AI decision-making processes, allowing individuals to comprehend the effects of AI systems and hold developers responsible for outcomes (p. 699 – 700). When it comes to explainability, Dorton and Harper (2022) have outlined the following elements of XAI technologies: *justification, transparency, conceptualization, learning, and bias* (p. 224 – 225). For this thesis, the operationalization of justification will encompass a combined interpretation of justice, fairness, and bias. Learning, considered less crucial for AI trustworthiness, will not be addressed.

Justice and fairness play a role in TAI, ensuring equal treatment and avoiding discrimination. In industries like services and software development, there is a strong focus on promoting equality, diversity, and inclusivity. This involves advocating for algorithms and using representative data. The concept of justice goes beyond following the law; it also involves considering the ethical responsibilities of developing and implementing AI systems. The goal is to correct wrongs, ensure fair sharing of benefits, and prevent new forms of inequality (Thiebes et al., 2020, p. 454 – 455). Ethical principles highlight the importance of fairness in eliminating bias and fostering shared benefits. These principles stress the impact of AI on fairness while warning about biases in vital fields such as healthcare. Various standards see fairness as a way to address injustices, promote fair sharing of benefits, and prevent new harms, showing uncertainty about whether AI is an empowering force or a passive technology (Floridi et al., 2018, p. 698 – 699). According to ALTAI, ensuring fairness and avoiding bias in TAI is essential. To achieve this, selecting data, analyzing algorithm design, and considering diversity and representativeness are vital. ALTAI recommends using technical tools to improve understanding and continuously assess the AI system throughout its lifecycle to address potential biases. Fairness means more than equal treatment; it means allowing individuals to seek redress if their rights are violated. These steps uphold standards to protect the integrity of AI systems and promote fairness for everyone (AI HLEG, 2020, p. 16 – 17, 27).

Transparency in TAI mainly revolves around clarifying things and highlighting the importance of people grasping how and why AI systems come to their conclusions. This demand has become increasingly important as businesses implement ML models and algorithms, resulting in outcomes often seen as unclear and hard to understand. By explaining things, system developers can show why AI makes confident choices, building trust and aiding users in decision-making. Clarifying choices is a crucial quality requirement that impacts user needs, cultural values, and other facets of the quality of AI systems (Balasubramaniam et al., 2023, p. 1 – 2). Furthermore, transparency goes beyond being able to explain things; it includes the ability to find, track, and identify decisions and actions made by AI systems. Building trust with stakeholders is crucial, which means maintaining a level of openness throughout the entire development cycle (Felländer et al., 2022, p. 8). However, sharing details can also bring about risks, creating a phenomenon known as the *transparency paradox*.

When there is a lot of irrelevant information or information is not easily understandable, it can reduce transparency instead of improving it. On the other hand, storing excessive amounts of data about decision-making processes raises concerns related to privacy and surveillance. Therefore, achieving transparency in AI involves finding a balance between providing relevant information to establish trust without overwhelming users with unnecessary specifics or jeopardizing privacy and security (Cobbe et al., 2021, p. 600 – 601).

Organizations emphasize that making AI systems understandable is crucial for promoting transparency and clarity. The guidelines on transparency highlight three aspects regarding the significance of understandability: Ensuring that individuals grasp how AI is used and how the system behaves, clearly communicating where, why, and how AI is applied, and helping people differentiate between actual AI decisions and instances where AI merely provides decision-making support through recommendations. Thus, fostering understandability facilitates explainability and transparency by conveying the use of AI to individuals clearly and in detail (Balasubramaniam et al., 2023, p. 8).

4. Results

4.1. Technical Interpretation of AI

As mentioned above, explainability is a critical concept for realizing TAI. On which explainability and interpretability strategies could be implemented to enhance the trustworthiness of AI systems, one of the interviewees responded:

> I think that based on the experience of the person using the model, you should have a different explanation. For instance, if you are just the user and you are interacting with a model that takes as input some medical data. You should get an explanation that is simple in that it should be interpretable by a person who does not know anything about medicine. At the same time, if I'm a physician, interact with this *black box* model, maybe in this case the physicians can get an explanation that is complex with some additional information that a normal user would not understand. (William)

A similar methodology is proposed by Doran et al. (2017), who suggest such an approach for dealing with opaque systems, where the inner workings that link inputs to outputs are not transparent to users. They argue that providing explanations of AI decisions is crucial for establishing trustworthiness and assessing the ethical and moral aspects of machine behavior. While interpretive models allow for explaining decisions, they do not inherently generate explanations. To address this limitation, they advocate for the development of *truly explainable systems*. These systems utilize automated reasoning to generate explanations without relying on human intervention as the final step in the process (p. 1 – 7). A different option William recommended was using *white box* systems instead of *black box* systems.

> […] I mentioned the *white box* in the like. It is the opposite of the *black box* model. With the *black box* model in the literature, we mean the like neural networks. For instance, our *black box* model because you cannot understand what happens inside this box. Instead, for instance, a decision tree is a *white box* model because you can understand what happens inside. […] these models are more interpretable, and the thing is that you have to find a tradeoff between the model's explainability and the model's accuracy, because with neural network maybe you would get better utility, a better performance, better accuracy.

William's insight shows that to achieve explainability and, in turn, trustworthiness, there has to be some tradeoff in the functionality of the AI system and the interpretability hereof. Choosing between functionality and interpretability might not create many problems because, as explained in the report *Aandacht voor algoritmes* by Algemene Rekenkamer (2021), the Dutch central government primarily used relatively simple algorithms. Moreover, the effects of simple algorithms on citizens are limited because they typically make automatic decisions involving automating an administrative task. They did not find any fully self-learning algorithms within the Dutch central government, only learning algorithms where there is always a *human-in-the-loop* (p. 6). These relatively more straightforward algorithms might be preferable in some situations, especially when dealing with vulnerable groups. Vulnerable groups could be historically marginalized populations, persons with chronic illnesses, or simply citizens with poor financial situations. An example of this would be the Dutch childcare benefits scandal. Victor says the following about this:

> Yes, and the impact was, of course, too great. Most importantly, you immediately put people in financial trouble when you stop someone's allowance and tell this person we are reclaiming it unless you can provide proof that the allowance was justified. While formally, you have not convicted them of anything.

Peter places a similar emphasis concerning the long-term risk management of these systems, noting that even when a system is turned off, thousands of people can still be affected. He underlines that while understanding how to turn off an AI system is important; more attention needs to be paid to the literature on the continued impact after deactivation. Thus, the AI system may be off, but the problems it created or exacerbated remain unsolved. When asked about ways to mitigate similar negative societal impacts of AI system use in society, Lucy points to the importance of a comprehensive approach:

> I tend to say that it is very difficult to view it as a snapshot. Because you have to look at societal impact over time as well. [...] You will need to conduct a study that spans over time. [...] You need to collect data over time, ensuring a thorough understanding of the evolving societal impact.

The heart of this proposed approach lies in involving end-users. Through conducting surveys over the years, it is possible to collect their valuable input to ensure that the utilization of AI systems is equitable and just. Selbst (2021) highlights algorithmic impact assessments (AIA), which could be utilized for this purpose and suggests three different categories. Initially, there are models rooted in the National Environmental Policy Act (NEPA) that were designed as impact assessments for the public sector. Secondly, there are models based on the GDPR's Data Protection Impact Assessment (DPIA). Lastly, there is the questionnaire model, which the Canadian government follows under the Canada Directive on Automated Decision-Making. Government agencies need to complete such an AIA before implementing a system (p. 122 – 123, 139 – 143). The questionnaire model could prove beneficial by allowing end-users to share feedback about their interactions with the AI system; adjustments or even discontinuation of the system could be considered based on the responses provided.

Preventing issues before they arise is generally preferred, and one way to achieve this is by identifying potential complications preemptively. New regulatory policies such as the AI Act or questionnaire-based AIA have merit, but they may need refinement for effective implementation. These measures aim to encourage AI developers to preemptively consider the implications of their technology, altering how ethical debts are accumulated. Nonetheless, the unpredictable nature of AI can pose challenges, often requiring harm to take place before accountability is established. Therefore, a key goal should be enhancing the ability to anticipate ethical risks associated with integrating AI into complex socio-technical systems. One approach could involve using a pre-mortem technique in various critical domains. During a pre-mortem, team members brainstorm project risks and provide unique perspectives on possible failures. This process may involve developing strategies to address risks or pinpointing root causes. An advantage of pre-mortems is their ability to uncover interdisciplinary risks that project managers may overlook on their own. By involving diverse participants in these sessions, new risks pertaining to socio-technical systems can be identified, such as hypothetical user behaviors and long-term cultural or motivational influences. This minded approach helps uncover hidden dangers at the crossroads of humans and technology, highlighting the pre-mortem as a valuable technique for anticipating ethical issues arising from the integration of AI into intricate socio-technical systems (Dorton et al., 2023, p. 2 – 3).

When researcher teams identify potential project risks hypothesized during a pre-mortem, these hypotheses could be tested using synthetic data. Jimmy explains: “Another promising potential avenue is to generate synthetic data. That becomes more and more a possibility. Moreover, test that algorithm with mock or synthetic data, knowing that you want that synthetic data to be representative and diverse.” Victor explains that in his experience, synthetic data might serve two purposes:

> There are indeed cases where you generate synthetic data, but synthetic data is a solution for two problems. Number one, it is a solution for cases where you have good data but are not allowed to use it because of the GDPR. So, you try to create synthetic data with the same characteristics. Another is to work in a hypothesis-testing manner. For example, with a vicious circle where discrimination increases. If you want to explain to an organization why that effect can occur, then it makes much sense to generate synthetic datasets and run them through the algorithm. And then subsequently show that the effect occurs, and the discrimination effect increases. So yes, it is a tool, but I mainly see it as a form of explanation. Or as a way to work with data while you are not allowed to use the real data.

Victor's explanation shows that synthetic data can be legally beneficial for safely training AI systems, helping to prevent regulatory violations, and protecting privacy. It will also help to test pre-mortem hypotheses of potential adverse effects of introducing AI into complex socio-technical systems. In addition to the current retrospective rhetoric occurring in the drafting of AI legislation, there has been criticism that governance processes focused solely on risk may not fully anticipate the effects of technological advancements (Raji et al., 2020, p. 37). It is observed that those creating AI-related documents in the sector predominantly hail from affluent nations, sparking concerns. One such concern is the lack of representation of low- and middle-income countries in global discussions on AI ethics and policies. Another concern is the negative impact on less affluent nations due to AI-driven growth led by wealthier countries. For instance, India's national AI strategy #AIFORALL highlights the possibility of displacement for customer service and technical support workers as their roles become automated and might be relocated to wealthier nations. Mexico's AI strategy mentions a similar scenario for its manufacturing workforce (Schiff et al., 2020, p. 2). Peter also referred to this: "And other countries, such as India, prohibit generative AI models if you cannot demonstrate that they have no intellectual property (IP) material. India has political reasons for this because they have many helpdesks that could be replaced." Highlighting that an AI system or legislation in a specific socio-technical context might not be applicable or preferable in another one.

Government regulations like the AI Act play a role in streamlining product safety standards, aligning technical specifications, and preventing global inequality. By doing so, they contribute to enhancing safety measures tailored to specific products incorporating AI-related aspects, like fair data training practices, transparency, and ethical concerns (Gesmann Nuissl & Kunitz, 2022, p. 11). Additionally, definitions, such as the precise meaning of an AI system, are now standardized, as Peter explains:

> [...] AI treaty that the United States is also part of. It includes elements in the AI Act, such as an impact assessment, if appropriate, which is also mentioned. However, you must evaluate certain elements if you conduct an impact assessment comparable to the DPIA under GDPR. The requirements are the same as for a high-risk AI system, as stated in the AI Acts, including data quality, risk management, cybersecurity, and similar points. This ensures that everything is aligned, which is nice. What is also aligned and a great advantage is the definition of an AI system. The OECD established this definition on November 9 last year. It is a global political organization with over 140 member countries. This definition has also been adopted in the AI Acts and the AI Treaty, aligning these three.

The AI Act, with its various aspects, including the risk-based approach, has global implications. It operates on a pyramid of prohibited risk, high risk, low risk, and no risk (Chamberlain, 2023, p. 4 – 7). A key point of discussion is the ethical assessment of what is socially desirable, a concept that varies significantly across countries.

> Prohibited social scoring is banned in Europe, but not in China. It is probably also banned in the United States, but if it is a variant where big tech businesses could make much money, then it is questionable to what extent they would ban it or if they would create some exceptions. This is because there is much emphasis on business and business freedom in the

> United States compared to Europe, where we focus more on the residents and the people. (Peter).

Insights like these create some personal skepticism about whether ethics should be internationally standardized. It is important to acknowledge that a mathematical theory of information might not offer certainty or computability to our moral reasoning without considering this as a significant issue. Moreover, ethics are universal in the Latin understanding of being inclusive to all, rather than in the Anglo-Saxon conception of applicable in all cases and absolutely defined by being free from restrictions or being universally applicable in the sense that they are relevant in all circumstances. It is impossible to establish a fixed set of AI ethics frameworks because ethics are context-dependent and shaped by the mutual ontological environment of subjects and actors; in plain terms, "an x is identified as y not absolutely but always within a specific context." One approach to address this challenge is conducting audits and amplifying voices on the periphery, thereby promoting social inclusivity. Ethicists should view audits as tools for fostering an ethical environment since audits serve as mechanisms for evaluating complex processes to ensure alignment with company policies, industry standards, or regulations. Essentially, audits act as listening devices that assess whether principles and ethical guidelines resonate with real-world practices. With that being said, the subjectivity and context dependence of the concepts still hold true (Light & Panai, 2022, p. 11 – 12).

4.2.Importance of Social and Cultural Context

The reasons why our moral reasoning is being replaced with binary classification likely relates to the fact that many of the innovations in AI and its social applications came first and foremost from computer scientists who see the world in a certain way. It is not only how they approach the technical development of ML itself but also their vision of the world. This phenomenon is what Jimmy calls *optimization problems*. As long as computer scientists minimize the error, they believe that this would fix the issue. One way to address this phenomenon could involve paying attention to the specific context where an AI system is utilized rather than simplifying moral decision-making into binary classification. Grasping the context of an algorithm is vital. When assessing algorithms in societal contexts, it is important to concentrate on their technical aspects and role as problem solvers in real-world scenarios. A standalone algorithm lacks practicality without a task environment. When placed in a complex task setting for critical operations, numerous challenges can arise. For evaluators, obtaining an understanding of an algorithm within its context is essential for initiating an assessment of algorithm reliability (Boer et al., 2023, p. 157). To truly understand how an algorithm works, it is essential to delve into a range of political factors that shape its definition of success. This involves looking at aspects such as the algorithm creation process, data preparation for training and testing deployment to users, and often, most importantly, the setting within which it is used. These contextual elements play a role in determining the potential negative impacts of algorithms. However, current ethical assessments often fail to consider the context when evaluating algorithmic ethics (Brown et al., 2021, p. 2 – 3).

Sometimes, an outsider's perspective on inclusiveness might not be enough to realize justice or fairness. One AI expert responds to a question about how audit mechanisms might be used to reduce adverse effects such as bias. "Yes, that has to come from the company itself. Experts who look at the algorithm from different perspectives should be involved when developing it. [...] You can reduce subjectivity by involving more experts." (James). Another interviewee said something similar about the socio-technical context of ethical AI system usage. He explains: "Ethics is about having a diverse team and discussing and making it visible. Just like *privacy by design*, you should also need *ethics by design*. That is something new. Data scientists from more than four years ago did not learn that as a standard. It is not standardly embedded and is challenging." (Peter). Both these quotes explain that looking at an algorithm from an ethical perspective and incorporating the context of how the algorithm is used is becoming more common practice. However, there is still room for improvement in clarifying the context in which the algorithm is used. Some measures are already being proposed to prevent bias in AI systems regarding the context in which they are used. Boer et al. (2023) highlight the importance of having a group of individuals, including those with various cultural, technical, and domain expertise within AI algorithm development teams. They suggest that incorporating individuals from different backgrounds into a team allows for a more comprehensive understanding of real-world problems from multiple viewpoints. As a result, diverse teams are able to create effective AI algorithms and reduce the chances of overlooking potential risks (p. 166).

Discussions surrounding algorithmic systems often focus on technical concerns. Yet it is essential to note that these technical elements comprise a larger socio-technical framework described as “an entanglement of people and code” (Cobbe et al., 2021, p. 599). Large-scale AI systems used for production are incredibly intricate and a vital area for study, involving the examination of how these complex socio-technical systems interact. Additionally, there exists an interplay between users providing data, the collection of data, and the training and updating of models (Raji et al., 2020, p. 37). Joel was critical of the dynamic interaction of AI systems when talking about auditing the trustworthiness of these complex adaptive systems, as he explained:

> [...] It is a *moving target*. So, how can you say anything about reliability at any given moment? You can speak about reliability in a stable situation, but a self-learning system with many feedback loops differs. Nevertheless, one thing I know for sure about such a complex adaptive system is that it does not exhibit linear behavior. In other words, it displays unpredictable behavior. Furthermore, you can see that with AI systems, too. Sometimes, they produce *hallucinations* or emerging properties that are inherent to the fact that we are dealing with a complex adaptive system with feedback loops. So, if you cannot say anything about the

> system itself, you need to say something about how an organization has safeguarded itself against the risks of possible unpredictable behavior.

Traditionally, audits have been carried out periodically or in cycles. In instances, audits serve as snapshots of systems and processes. The timing of these snapshot audits is crucial as an early audit can impact the design and functioning of an AI system compared to an audit conducted after deployment or when the system is in production. While many AI systems use static models with periodic updates, some systems, such as those based on reinforcement learning, adapt because highly complex models may exhibit unpredictable results. Learning and adapting come with advantages and potential risks as AI systems pick up on patterns that may not be apparent to their creators in code structures. Adaptation poses a specific challenge for snapshot auditing because a system that is compliant at one point may not be compliant later. Furthermore, the speed at which AI systems operate and evolve surpasses traditional human-led snapshot auditing procedures, which tend to be more time-consuming (Minkkinen et al., 2022, p. 2).

Other interviewees are also critical of the auditing process surrounding AI systems such as ChatGPT and are skeptical of their trustworthiness. They argue that while auditors can manipulate data when they have control and access, the challenge with ChatGPT lies in its vast training data, which is essentially the entire internet. Auditors do not possess this data but are aware of its extensive scope. Victor emphasizes that despite the massive data used for training, the subsequent evaluation showing current functionality fails to ensure future system reliability convincingly. Lucy shares a similar view, highlighting that audits of AI systems are mainly grounded in business compliance rules. She explains that most research on AI systems revolves around understanding and applying business rules or more complex variations, making the process somewhat transparent but only partially. She refrains from labeling AI systems as a *black box* but acknowledges that the calculation rules could lead to unintended consequences. Lucy also notes that, despite the increasing discussion about auditing AI systems like ChatGPT, actual audit attempts still need to be made available. She believes that audits of similar systems will eventually happen but anticipates it will take time before they become standard practice.

Because it is difficult to audit such large systems retrospectively, there is a need for real-time auditing of complex systems. Minkkinen et al. (2022) suggest the method of continuous auditing (CA). The auditing of AI and CA naturally align because CA can adapt to the progress of the AI system, providing assessments of its performance based on predefined criteria, including elements of reliability mentioned earlier in the operationalization. Moreover, CA could lessen intervention in auditing by assigning tasks to machines and allowing humans to concentrate on more intricate audit tasks. This continuous AI auditing approach has drawn attention from the EU as the AI Act already includes provisions for the mandatory post-market monitoring of high-risk AI systems (p. 2 – 3). Lucy further elaborates:

> What the AI Act says about [post-market monitoring] is that you need to collect data. Very little is said, but it mentions that you must collect data to verify that the deployed model meets the AI Act's criteria. In other words, you must monitor its behavior to ensure it keeps doing what it should. However, the legislation does not provide much detail beyond stating that you must collect data. So, this is something that needs to be shaped in practice.

Post-market monitoring (PMM) refers to observing and evaluating a high-risk AI system's actions and effectiveness following its deployment and use (De Boer, p. 27, 2023). PMM might be improved by opting for the continuous auditing of AI (CAAI) method, which exists at the intersection of CA and auditing of AI. CAAI is a CA method that targets AI systems and corresponding organizations. Said differently, auditing of AI provides the audit object, and CA provides the auditing method. CAAI is a subset of both auditing of AI and CA because of its intersectional position. CAAI functions as an electronic support system for auditors, providing real-time assistance by conducting automated audits of AI systems to ensure compliance with established norms and standards. This involves integrating components from the EBA definition into the evaluation process (Minkkinen et al., 2022, p. 6). Other interviewees proposed different methods for auditing AI systems:

> It is a sort of *moving target*. Instead of evaluating how it works, we now focus on the output and whether the output behaves in line with what we would expect. Moreover, that is, of course, very complicated. Because, yes, how do you test that? How do you test a complex self-learning system? Essentially, only through other AI. (Joel).

Something similar is expressed by the National Institute of Standards and Technology (NIST), which also mentions that Large Language Models (LLMs) can act as moderators for LLMs to identify threats beyond just screening out harmful inputs (Vassilev et al., 2023, p. 49). Approaches like LLM moderators include tools like AuditLLM. AuditLLM assesses the effectiveness of LLMs through a multiprobe method. This tool has two modes: A mode for immediate assessments with real-time questions and a batch mode for thorough evaluations with various queries. It utilizes two LLMs, where one creates probes based on a user's query, and the other responds to these probes. This method helps pinpoint discrepancies in the model's comprehension, which can also evaluate the model's existing grasp on ideas like trust, fairness, and clarity (Amirizaniani et al., 2024, p. 3 – 4).

> [...] One of the principles you have is the four-eyes principle. Applying this to an AI environment is similar to an airplane where control systems are often duplicated or triplicated, using a sort of voting mechanism. So, if two of the three systems agree that something should happen, it happens. We either descend or ascend. You get that in the AI world as well. We have a four- or, say, multiple-eyes principle with a voting mechanism where two, three, or

more different AI systems come to a certain output based on the presented inputs, and then a vote is taken.

These strategies for reducing risks highlight the need for AI systems to be reliable. They also indicate a shift in attitudes towards auditing AI systems and earlier agreement on compliance and accountability in ADM. Some previous studies have suggested that building trust could be achieved by making things reviewable, leading to traceability and governance. Reviewability, in particular, is a way to support meaningful accountability. It is essential to have appropriate technical information to assess algorithmic systems in terms of their context and outputs for legal compliance, expected functionality, desired parameters, and other relevant assessments related to various accountability relationships. Reviewable ADM processes systematically implement technical and organizational record-keeping and logging mechanisms at all commissioning stages, development, operation, and investigation to holistically assess the algorithmic system's development cycle. As a high-level concept, reviewability has applications in various areas and is relevant for algorithmic systems in general. Its approach to transparency and accountability of ADM goes beyond mere explanations or other mechanisms narrowly focused on technical components. (Cobbe et al., 2021, p. 601 – 602). Because of the AI system's emerging properties, the core elements of TAI, transparency and accountability, are more challenging to realize. Reviewability might be less possible for complex adaptive systems. Robustness through fault-tolerant control systems might be more feasible.

Alongside improving robustness by using multiple moderator LLMs, context-dependent algorithms could serve as a solution to increase robustness and decrease potential bias. The prevailing approach to AI algorithmic reasoning leans towards *algorithmic formalism*, characterized by strict adherence to predefined structures and regulations. This often results in consequences like perpetuating social norms and fostering technologically deterministic perspectives regarding societal change (Ehsan et al., 2021, p. 4). AI technologies are developed within the framework of social settings shaped by interactions, shared activities, interactions between humans and machines, and technology's underlying values and ethical considerations (Toktas et al., 2024, p. 4). Employing context-dependent algorithms helps us better understand AI systems by considering the interplay of values, interpersonal dynamics, and the socially situated nature that affects the technology (Ehsan & Riedl, 2020, p. 1).

A method to realize context-dependent algorithms involves the concept of Social Transparency (ST), which integrates socio-organizational context into AI-driven decision-making processes, creating a socio-technically informed perspective. ST AI systems become more socially situated, and their development focuses on social, organizational, and cultural factors that influence their usage (Ehsan et al., 2021, p. 1, 4). Different societies have their own sets of terms, meanings, and standards. Concepts such as *fairness* and *privacy* may vary depending on the location. Moreover, AI technologies are influenced by values and cultural beliefs during their development process. To grasp the implications and effects of AI, it is essential to consider the context in which it operates. This involves recognizing the significance of culture in the development and deployment of AI systems (Hagerty & Rubinov, 2019, p. 2, 4). Hagerty and Rubinov (2019) explain that culture is a living, changing entity of shared experiences and beliefs, both evident and hidden. Culture is heterogeneous, including individual differences and subcultures, and intertwined with historical, political, and economic contexts. Thinking about culture is personal and shaped by our cultural experiences and background (p. 7 – 8). By considering underlying cultural logic and fundamental societal beliefs and presumptions, AI auditors can establish guidelines tailored to distinct AI systems and their specific contexts (Hagerty & Rubinov, 2019, p. 2, 4). Through the ST approach, bias will be reduced, and there will also be a better understanding of the socio-technical structures underlying these new technologies.

5. Conclusion

5.1.Improving AI Frameworks

This thesis studied AI auditing as a governance mechanism and its ethical, social, and legal challenges from a socio-technical perspective. It has come to the fore that AI systems auditing is multidisciplinary and consists of a complex intertwining of socio-technical elements and concepts. Although overlap exists, each framework discussed takes a different approach to legal, ethical, and social criteria to audit AI systems. When looking at trustworthiness, it is believed that trust in AI involves the system helping achieve goals within uncertain and vulnerable contexts, requiring calibration over time to avoid overreliance or underutilization (Dorton & Harper, 2022, p. 223). Criteria for operationalizing trust include system reliability, understandability, and goal alignment between the user and AI. The Floridi et al. (2018) conceptualization of trust, which takes on a more assertive social sciences perspective, ethically aligns with the socio-technical interpretation of AI systems. Their framework points to *beneficence, non-maleficence, autonomy, justice, and explicability* as the main elements of trustworthiness (Thiebes et al., 2020, p. 451). Relating to this, Dorton and Harper (2022) identified explainability as part of trustworthiness, where it encompasses the components of justification, transparency, and understandability (p. 224 – 225).

Literature mainly points to the importance of legal and regulatory compliance for the trustworthiness of AI systems. This was acknowledged by interviewees, who explained that audits of AI systems were primarily based on business compliance rules. Another crucial legal criterion for trustworthy AI systems is risk mitigation. During the interviews, it also became clear that risk mitigation strongly correlates with robustness and reviewability. Reviewability, which creates meaningful accountability, mainly involves appropriate technical information to assess AI systems (Cobbe et al., 2021, p. 601 – 602). Trustworthiness through robustness already played a foundational role in the auditing of AI. It became clear that AI audits involve code audits where various interrelated methodologies approach AI systems' input and output. To do this, different frameworks have been created to audit AI and algorithms, such as the SMACTR or the *Aandacht voor algoritmes* framework (Raji et al., 2020, p. 39 – 42; Algemene Rekenkamer, 2021, p. 22 – 25). These frameworks are instrumental in the early stages of an audit, where biases and harmful consequences of intelligent systems can still be identified during model development by actively mapping development choices, considering training data, and recognizing that certain features might lead to unfair biases. In addition to robustness-oriented methods for general risk assessment, other frameworks are explicitly designed to address ethical risk management. Ethical considerations for TAI were autonomy, non-maleficence, justice, and transparency. Auditing methodologies such as the DRESS-eAI or the EBA framework highlighted the importance of conducting transparent, responsible, and ongoing audits to support ethical decision-making and establish a foundation of trust for AI markets that are socially and environmentally sustainable (Felländer et al., 2022, p. 1 – 15; Light & Panai 2022, p. 10 – 14).

Frameworks such as the ethical risk assessment emphasized evaluating the potential harm algorithms could cause to the rights and interests of end-users and stakeholders. Instead of transparency, the ethical risk assessment aims to realize justice and non-maleficence by identifying and ranking potential negative consequences like mental harm and property loss, as well as violations of moral rights such as privacy, freedom of speech, autonomy, and equality (Hasan et al., 2022 p. 5 – 17). The ethical algorithm assessment by Brown et al. (2021) focused mainly on outlining the potential interests of stakeholders impacted by the algorithm, evaluating metrics that highlight critical ethical aspects of the algorithm, and establishing a relevancy matrix that links these evaluated metrics to stakeholder concerns. This framework stands out because it emphasizes the relevance of the social context during ethical auditing (p. 1 – 3). These different ethical AI frameworks show a demarcation between auditing ethical principles and standards, auditing ethical algorithmic impact, and auditing ethical functionality of algorithms. Not only do auditors need to be aware of training data being used and possible bias that is created, but they also need to be aware of the context of how an AI system is being used and need to be able to determine which criteria are most relevant in the specified context.

Critical socio-technical criteria that came to the fore were openness, context, and culture. Socio-technical auditing frameworks such as COMPASS help organizations evaluate AI systems' technical innovation and societal impacts, ensuring responsibility and trustworthiness (Toktas et al., 2024). Alternatively, STA methodologies, such as the Intervenr system that evaluates both technical and human components of AI systems, shift the focus from technical system auditing to a more socio-technical perspective (Lam et al., 2023, p. 4). Frameworks that specifically audited the trustworthiness of AI systems, such as DaRe4TAI and the Z-Inspection®, evaluated trustworthiness through holistic examination addressing ethical, technical, legal, and domain-specific implications. While DaRe4TAI places a central role on data management, Z-Inspection® focuses more on the complete AI lifecycle through compliance, development, and deployment auditing (Thiebes et al., 2020, p. 456; Vetter et al., 2023, p. 1, 5). Culture was found to be vital as socio-technical criteria as it is the foundation for understanding the context wherein AI systems are being developed and deployed (Hagerty & Rubinov, 2019, p. 4). At the same time, it also became clear that none of the frameworks discussed used cultural appropriateness standards for their auditing frameworks.

The central inquiry driving this thesis was: How do AI practitioners perceive the efficacy of AI auditing, and how can these perspectives be leveraged to enhance the trustworthiness of AI systems? The question this creates is whether standardized ethics should or are possible. Here the *optimization problem* interpretation of moral reasoning, where social sciences and computability clash, is crucial. Ethics in AI are not a simple, universally applicable concept. The notion of a universal set of AI ethics is a fallacy, as ethics are inherently contextual and relational to their shared ontological environment. The Latin understanding of ethics as being inclusive to all is more apt in this context than the Anglo-Saxon conception of ethics being universally applicable in all circumstances. This complexity underscores the need for a nuanced approach to AI auditing. One way to counter this would be to use AI auditing to look at technical elements such as robustness, security, or interpretability and examine ethical environments surrounding technology (Light & Panai, 2022, p. 11 – 12). Instead of standardizing moral reason into binary classifications, AI auditors should focus on specific contexts wherein AI systems are used. One novel approach proposed in this thesis is incorporating ST into AI. By taking on a socio-technically informed perspective that incorporates socio-organizational context into AI auditing, AI systems become more socially situated, and the cultural, organizational, and social factors that govern their usage are thereby also uncovered (Ehsan et al., 2021, p. 1, 4). A particularly underrepresented perspective is the inclusion of cultural logic and underlying social values and assumptions of AI systems. As societies have unique ethical vocabularies, understandings, and expectations, AI auditors should also demarcate some context-specific ethical standards for distinct AI systems (Hagerty & Rubinov, 2019, p. 2, 4). This, in particular, will be important with the emergent properties in AI, as ST will enable auditors to have some form of reviewability of the trustworthiness of AI systems.

5.2. Limitations

In this thesis, a small meta-analysis of relevant frameworks and their social, ethical, and legal criteria was conducted. Because of the scope of this thesis, not all frameworks relating to trustworthiness or AI auditing in general could be incorporated. Consequently, some relevant content might have been left out. Furthermore, as this thesis takes on a social sciences perspective regarding the operationalization of the trustworthiness of AI systems, other relevant criteria might have been overlooked. The methodology used for this thesis was expert IDIs, which led to fascinating conversations with professionals in AI auditing. However, the total number of interviews was limited because only one individual conducted the interviews. In addition, all but one interviewee was Dutch, which results in this thesis being focused on AI auditing practices in the Netherlands. Finally, AI auditing is an inherently interdisciplinary research topic that must be approached from multiple angles to give a holistic representation of current practices and criteria. Because this research is mainly from a social science perspective, some depth might be missing about other lesser-related topics.

5.3. Future Research

Future research might examine how trustworthiness in AI systems has evolved for AI practitioners and how deployed systems are subjected to audits. AI audits might be improved by examining which social, ethical, or legal criteria are essential and in which contextual settings these criteria exist. One interesting topic for future research might be to examine how contextual AI auditing might be standardized. This can be done by studying cultural logic, underlying social values, and assumptions of AI systems and structurally scrutinizing them to audit emerging properties. Another topic would be to research ST factors and develop methodologies that enable the incorporation and automation of cultural, organizational, and social data into AI systems.

6. Bibliography

Algemene Rekenkamer. (2021). *Aandacht voor algoritmes*. https://www.rekenkamer.nl/publicaties/rapporten/2021/01/26/aandacht-voor-algoritmes

Amirizaniani, M., Roosta, T., Chadha, A., & Shah, C. (2024). AuditLLM: A tool for auditing large language models using multiprobe approach. *ArXiv*, 1(1), 1-6. https://doi.org/10.48550/arXiv.2402.09346

Balasubramaniam, N., Kujala, S., Kauppinen, M., Rannisto, A., & Hiekkanen, K. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, 159, Article 107197. https://doi.org/10.1016/j.infsof.2023.107197

Bloomberg. (2023, December 22). OpenAI is in talks to raise new funding at valuation of $100 billion or more. *Bloomberg*. https://www.bloomberg.com/news/articles/2023-12-22/openai-in-talks-to-raise-new-funding-at-100-billion-valuation

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society*, 20(1), 1-13. https://doi.org/10.1080/1369118X.2016.1216147

de Boer, M. (2023). Trustworthy AI and accountability: Yes but how? What the EU AI Act's approach to AI accountability can learn from the science of algorithm audit [Thesis, Universiteit van Amsterdam]. UvA-DARE (Digital Academic Repository). ISBN 978-94-6361-854-0.

Boer, A., de Beer, L., & van Praat, F. (2023). In E. Berghout, R. Fijneman, L. Hendriks, M. de Boer, & B.-J. Butijn (Eds.), *Advanced digital auditing: Theory and practice of auditing complex information systems and technologies* (pp. 149-183). Springer. https://doi.org/10.1007/978-3-031-11089-4_7

Boyce, C., & Neale, P. (2006). *Conducting in-depth interviews: A guide for designing and conducting in-depth interviews for evaluation input*. Pathfinder International Tool Series, Monitoring and Evaluation, 2, 1-12.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*. https://doi.org/10.1177/2053951720983865

Bucher, T. (2016). Neither black nor box: Ways of knowing algorithms. In S. Kubitschko & A. Kaun (Eds.), *Innovative methods in media and communication research* (pp. 81–98). Palgrave Macmillan. https://doi.org/10.1007/978-3-319-40700-5_5

Bucher, T. (2017). The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society*, 20(1), 30–44. https://doi.org/10.1080/1369118X.2016.1154086

Chamberlain, J. (2023). The risk-based approach of the European Union's proposed artificial intelligence regulation: Some comments from a tort law perspective. *European Journal of Risk Regulation*, 14(1), 1-13. https://doi.org/10.1017/err.2022.38

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3-10, 2021, Virtual Event, Canada (pp. 1-12). ACM. https://doi.org/10.1145/3442188.3445921

Doran, D., Schulz, S., & Besold, T. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *ArXiv*. https://doi.org/10.48550/arXiv.1710.00794

Dorton, S. L., & Harper, S. B. (2022). A naturalistic investigation of trust, AI, and intelligence work. *Journal of Cognitive Engineering and Decision Making*, 16(4), 222–236. https://doi.org/10.1177/15553434221103718

Dorton, S. L., Ministero, L. M., Alaybek, B., & Bryant, D. J. (2023). Foresight for ethical AI. *Frontiers in Artificial Intelligence*, 6. https://doi.org/10.3389/frai.2023.1143907

Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. *ArXiv*. https://arxiv.org/abs/2002.01092v2

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-19). ACM. https://doi.org/10.1145/3411764.3445188

European Commission. (2021, April 21). *Annexes 1 to 9 to the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. COM(2021) 206 final. SEC(2021) 167 final; SWD(2021) 84 final; SWD(2021) 85 final.

Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., & Heintz, F. (2022). Achieving a data-driven risk assessment methodology for ethical AI. *Digital Society*, 1(13). https://doi.org/10.1007/s44206-022-00016-0

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707. https://doi.org/10.1007/s11023-018-9482-x

Future of Life Institute. (2015). *Research priorities for robust and beneficial artificial intelligence: An open letter*. https://futureoflife.org/open-letter/ai-open-letter/

Future of Life Institute. (2024, February 27). *High-level summary of the AI Act*. https://artificialintelligenceact.eu/high-level-summary/

Gesmann-Nuissl, D., & Kunitz, S. (2022). Auditing of AI in railway technology – A European legal approach. *Digital Society*, 1(17). https://doi.org/10.1007/s44206-022-00015-1

Hagerty, A., & Rubinov, I. (2019). Global AI ethics: A review of the social impacts and ethical implications of artificial intelligence. *ArXiv*. https://doi.org/10.48550/arXiv.1907.07892

Hasan, A., Brown, S., Davidovic, J., Lange, B., & Regan, M. (2022). Algorithmic bias and risk assessments: Lessons from practice. *Digital Society*, 1(14). https://doi.org/10.1007/s44206-022-00017-z

Independent High-Level Expert Group on Artificial Intelligence. (2020). *Assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment*. Brussels: European Commission. https://doi.org/10.2759/002360.

Johnson, J. M. (2001). In-depth interviewing. In *Handbook of interview research* (pp. 103-119). https://doi.org/10.4135/9781412973588

Lam, M. S., Pandit, A., Kalicki, C. H., Gupta, R., Sahoo, P., & Metaxa, D. (2023). Sociotechnical audits: Broadening the algorithm auditing lens to investigate targeted advertising. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2), Article 360, 1–37. https://doi.org/10.1145/3610209

Light, R., & Panai, E. (2022). The self-synchronisation of AI ethical principles. *DISO*, 1(24). https://doi.org/10.1007/s44206-022-00023-1

Madiega, T. (2024). *Artificial intelligence act (PE 698.792)*. European Parliamentary Research Service.

Malik, A. (2024, January 26). Five techniques to ensure reliable and honest use of generative AI. *Forbes*. https://www.forbes.com/sites/forbestechcouncil/2024/01/26/five-techniques-to-ensure-reliable-and-honest-use-of-generative-ai/

Marr, B. (2024, May 3). Building trust in AI: The case for transparency. *Forbes*. https://www.forbes.com/sites/bernardmarr/2024/05/03/building-trust-in-ai-the-case-for-transparency

Milmo, D. (2024, February 22). Google pauses AI-generated images of people after ethnicity criticism. *The Guardian*. https://www.theguardian.com/technology/2024/feb/22/google-pauses-ai-generated-images-of-people-after-ethnicity-criticism

Minkkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous auditing of artificial intelligence: A conceptualization and assessment of tools and frameworks. *Digital Society*, 1(21). https://doi.org/10.1007/s44206-022-00022-2

Mökander, J., Juneja, P., Watson, D., & Floridi, L. (2022). The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: What can they learn from each other? *Minds and Machines*, 32(4), 751–758. https://doi.org/10.1007/s11023-022-09612-y

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4361607

Mökander, J. (2023). Auditing of AI: Legal, ethical, and technical approaches. *Digital Society*, 2(3). https://doi.org/10.1007/s44206-023-00074-y

NOS. (2023, December 18). Zorgen bij toezichthouder over risico’s AI weinig zicht op incidenten. *NOS*. https://nos.nl/artikel/2502046-zorgen-bij-toezichthouder-over-risico-s-ai-weinig-zicht-op-incidenten

NOS. (2024, May 10). Wetenschappers maken zich zorgen over misleiding en manipulatie door AI. *NOS*. https://nos.nl/artikel/2519966-wetenschappers-maken-zich-zorgen-over-misleiding-en-manipulatie-door-ai

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by artificial intelligence practitioners. *Information, Communication & Society*, 23(5), 719–735. https://doi.org/10.1080/1369118X.2020.1713842

Park, P. S., Goldstein, S., O’Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), Article 100988. https://doi.org/10.1016/j.patter.2024.100988

van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds & Machines*, 30(3), 385-409. https://doi.org/10.1007/s11023-020-09537-4

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. *ArXiv*. https://doi.org/10.48550/arXiv.2001.00973

Samoili, S., López Cobo, M., Gómez, E., De Prato, G., Martínez-Plumed, F., & Delipetrev, B. (2020). *Defining artificial intelligence: Towards an operational definition and taxonomy of artificial intelligence (EUR 30117 EN)*. Publications Office of the European Union. https://doi.org/10.2760/382730

Saul, D. (2024, February 26). Google’s Gemini headaches spur $90 billion selloff. *Forbes*. https://www.forbes.com/sites/dereksaul/2024/02/26/googles-gemini-headaches-spur-90-billion-selloff

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What’s next for AI ethics policy and governance? A global overview. In *2020 AAAI/ACM Conference on AI Ethics and Society (AIES’20)*, February 7–8, 2020, New York, NY, USA (pp. 1-6). https://doi.org/10.1145/3375627.3375804

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, 35(1), 117-192. https://ssrn.com/abstract=3867634

Stix, C. (2022). Artificial intelligence by any other name: A brief history of the conceptualization of "trustworthy artificial intelligence". *Discovery Artificial Intelligence*, 2(26). https://doi.org/10.1007/s44163-022-00041-5

Thiebes, S., Lins, S., & Sunyaev, A. (2020). Trustworthy artificial intelligence. *Electronic Markets*, 31, 447-464. https://doi.org/10.1007/s12525-020-00441-4

Toktas, S., Pridmore, J., Oomen, T., Kooij, L., Westberg, M., & Goncalves, J. F. F. (2024). Sociotechnical, regulatory, and ethical implications and integration guidelines report. *EU SPATIAL Project Report*.

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2023). Adversarial machine learning - A taxonomy and terminology of attacks and mitigations (NIST AI 100-2e2023). *National Institute of Standards and Technology*. https://doi.org/10.6028/NIST.AI.100-2e2023

Vetter, D., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Gallucci, A., Gilbert, T. K., Hagendorff, T., van Halem, I., Hickman, E., Hildt, E., Holm, S., Kararigas, G., Kringen, P., Madai, V. I., Mathez, E. W., Tithi, J. J., Westerlund, M., Wurth, R., & Zicari, R. V. (2023). Lessons learned from assessing trustworthy AI in practice. *Digital Society*, 2(35). https://doi.org/10.1007/s44206-023-00063-1

Wiggers, K. (2024, January 16). OpenAI announces team to build 'crowdsourced' governance ideas into its models. *TechCrunch*. https://techcrunch.com/2024/01/16/openai-announces-team-to-build-crowdsourced-governance-ideas-into-its-models/

Willems, L. (2020). Auditing algorithms: A working example on auditing algorithms (Bachelor's thesis). Radboud University Nijmegen Faculty of Social Sciences.

Zarouali, B., Boerman, S. C., Voorveld, H. A. M., & Van Noort, G. (2022). The algorithmic persuasion framework in online communication: Conceptualization and a future research agenda. *Internet Research*, 32(4), 1076–1096. https://doi.org/10.1108/intr-01-2021-0049

Zicari, R. V., et al. (2022). How to assess trustworthy AI in practice. *Z-Inspection® Initiative*, 1-51. https://doi.org/10.48550/arXiv.2206.09887

7. Appendix

Appendix A. Measuring instrument – interview questions

Socio-Technical Considerations

1. How do you integrate different social contexts into technological functionalities during development?
2. How do you apply socio-technical considerations to evaluate the broader societal impact of AI applications?
3. How do you integrate diverse socio-technical considerations to evaluate the broader societal impact of AI applications?
4. How do you assess the benefits of incorporating diverse perspectives for a more comprehensive understanding of the societal impact of AI systems?
5. What is your opinion on socio-technical auditing, and how does it differ from the original auditing of AI systems?
6. Why do we need socio-technical auditing, and how does it determine when it is necessary or not?
7. What are the opportunities for AI auditing, and what are future opportunities?
8. What are the challenges for the audit process of AI?
9. Is it possible to standardize the socio-technical audit processes of AI systems?
10. Can you provide examples of specific cases where socio-technical implementation during AI auditing was challenging or successful?
11. What methodologies do you use to identify and address potential negative consequences or biases in AI systems?
12. How extensively is user feedback and usability integrated into the AI audit process?
13. How do you implement continuous monitoring mechanisms to detect and address emerging fairness issues over time?
14. What is your expectation for this audit process, and what is your further expectation? Do you see it as a tool or a method in the future?

Ethical Considerations

15. How do you ensure that AI systems are transparent and understandable for non-technical stakeholders?
16. What strategies do you use to improve the understandability and explainability of AI systems?
17. How do you measure the effectiveness of mitigating strategies for addressing potential negative consequences or biases in AI systems?
18. What criteria and metrics should AI practitioners use to evaluate the fairness of AI systems?

19. Can you elaborate on how you approach the AI development lifecycle, from data collection to deployment?
20. Do you use transparency and explainability tools to improve the interpretability of model decisions?
21. Can you provide insights into the specific use of explainability tools and interpretation techniques in your approach?
22. How do you implement transparency measures to address concerns regarding the opacity of *black box* AI systems?
23. How can we raise awareness about the importance of AI auditing?

Legal Considerations

24. How do you stay informed about evolving regulations?
25. How does regulation affect the development of AI systems?
26. How does your organization handle regulations related to AI?
27. Do you think the GDPR in the EU has a positive effect on AI development?
28. Do you believe that the EU AI Act influences the development of AI applications?
29. Do you think legal provisions are positive or negative in addressing emerging issues or ethical concerns associated with AI systems?
30. What is your opinion on AI regulation, and is there a related article in the regulation that you find important?

Appendix B. Overview of interview respondents

#	Pseudonym	Position	Organization
1	Victor	Senior-level AI developer	Large enterprise
2	Lucy	Senior-level AI researcher	Large enterprise
3	Ferdinand	Senior-level auditor	Large enterprise
4	James	Senior AI developer	Public research organization
5	Peter	Senior-level AI researcher	Public research organization
6	Jimmy	Co-lead AI course	University
7	Joel	Co-lead auditing course	University
8	William	Mid-level researcher	Technical university