OVERSIGHT METHODOLOGIES FOR ARTIFICIAL INTELLIGENCE IMPLEMENTATION IN THE ENERGY SECTOR

by

Cedric Alwyn Worthmann, BSc Eng, MSc Eng

Reg. No. 81599

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

FEBRUARY, 2025

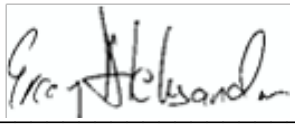
OVERSIGHT METHODOLOGIES FOR ARTIFICIAL INTELLIGENCE IMPLEMENTATION IN THE ENERGY SECTOR

by

Cedric Alwyn Worthmann

Supervised by

Dr Kamal Malik

APPROVED BY 

Dissertation chair

RECEIVED/APPROVED BY:

Admissions Director

Dedication

I dedicate this dissertation to my wife, Liesl, for giving me support, encouragement, and strength throughout my research and always helping us balance our lives, my job, and my studies.

Acknowledgements

I am extremely grateful for the continuous support and guidance that my Mentor, Dr Kamal Malik, has provided. Her continuous encouragement and guidance have made it possible for me to accomplish my research and to extend myself to deliver the research deliverables.

To the Swiss School of Business and Management faculty members, support staff, and committee members, thank you for your guidance throughout the process and for ensuring that I had all the necessary tools and access to undertake this research.

To Mr Wayne McKenzie, my friend and colleague, thank you for always being there to listen to me and indulge my curiosity. Thank you for sharing your expertise and knowledge with me and for always being available for me to bounce concepts and ideas off of you.

My CUC family and team, thank you for your encouragement and support as I undertook this research while working with you. Each of you has provided me with guidance and a gentle push to strive to complete this research and to better myself.

ABSTRACT

OVERSIGHT METHODOLOGIES FOR ARTIFICIAL INTELLIGENCE IMPLEMENTATION IN THE ENERGY SECTOR

Cedric Alwyn Worthmann
2025

Dissertation Chair: Aleksandar Erceg, Ph.D.

Humankind has been intrigued by Artificial Intelligence (AI) for decades (Fenwick and Molnar, 2022) as it is foreseen as a tool to improve efficiencies, further organizational interests, and improve societal well-being. Over the past years, there have been advancements in developing and deploying AI systems and tools in many critical services sectors, influencing organizations and people with a mixed level of benefits, successes, and risks.

The research aimed to understand the current knowledge base and gaps concerning the compliance oversight for the safe development, deployment, implementation, and use of AI systems within critical infrastructure services, specifically focusing on the energy or electricity sector. The driving factor for this research is that within the critical infrastructure and services sectors, incorrect decisions influence more than financial returns but can cause severe equipment damage, premature failure, and harm to people.

The outcome of this research is a unique AI compliance audit framework development procedure, AI compliance audit framework, and AI audit process, which is a fit-for-purpose compliance solution driven from a senior executive management level,

integrated into specific existing compliance or governance processes, which is more advanced than the existing AI governance frameworks, audit mechanisms, and checklists.

This research significantly contributes to the industry and sector by providing a practical structure that can be utilized to safely and sustainably implement AI systems in this critical sector. The AI compliance audit framework is a mechanism that allows the energy sector to place guardrails around the AI system that they are procuring or developing throughout its lifecycle. The framework considers a multi-dimensional audit regime, which can seamlessly integrate into existing quality, information technology, or environmental assurance processes. Notably, the framework ensures that the energy sector considers the integrated software systems collectively when auditing and ensuring compliance, not as individual components.

Lastly, the research provides a foundational structure for future researchers to expand on, focusing on gaps within the current governance structures, training regimes, and approaches to building a sustainable AI system environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I:

INTRODUCTION

1.1 Introduction

The energy sector, specifically the electricity sector, has been digital pioneers since the 1970s, using emerging technologies to facilitate grid management and operation (IEA, 2017). Digital transformation has become a key driver for the energy sector (Berger, 2018), with advancements in technology leading to constructive changes in how energy is produced, transmitted, consumed, and traded (Nazari and Musilek, 2023). Organizations and the world at large are experiencing a transition to a digital society where everything interconnects with each other (Laroussi et al., 2023). Over the past two decades, the energy sector's drive has been to digitize and de-carbonize (Światowiec-Szczepańska and Stępień, 2022), which provides profound benefits while creating a more complex environment, heavily dependent on large volumes of real-time data for system management and decision-making. As more granular and complex data became available to the energy sector, the utilities have had to secure new skills and develop or adopt more complex data analytic tools, including AI powered solutions.

As AI becomes more prevalent in the critical infrastructure and services sectors, such as healthcare, the military, and energy utility operations, the government, public and organizational objective has been to understand the benefits and risks of AI, with a focus on societal safety (Ozmen Garibay et al., 2023), organizational sustainability and impact on the workforce (Agrawal *et al.*, 2017; Fatima *et al.*, 2024). As AI becomes more complex and ingrained within the critical infrastructure and services sectors, it is essential to ensure the safety of the employees, the public, and infrastructure (Laplante and Amaba, 2021); this can be done by designing fit-for-purpose AI systems against a defined control structure and not retrofitting solutions after the fact. In the energy sector specifically, studies have

been conducted on how AI will improve the sector's performance, improve operations, and influence the workforce now and in the future (Lyu and Liu, 2021; Morris et al., 2022).

Due to the limited literature available focusing on AI within the energy sector, literature sources for other critical infrastructure and services sectors were considered in the initial literature portion of the review as they have similar risks, impacts, and possible mitigating factors. Throughout the literature review, there is a consensus that AI systems need mechanisms established to facilitate rules, guidance, or control for AI system development, deployment, implementation, and usage (Büthe et al., 2022; Dafoe, 2018; Taeihagh, 2021). Where there is a lack of shared vision is what format this should take and who should be driving the development and facilitation of these mechanisms (Munn, 2023). There is also a need for the future of leadership and the role of leaders (Johansson and Björkman, 2018; Jorzik et al., 2023; Shadman, 2023) to change during the integration and implementation phase and again during the deployment and operational phase to accommodate the new organizational and staff needs (Jorzik et al., 2023). With more intelligent solutions such as AI providing key functionality, it has become necessary for organizations, senior leadership specifically, to strategically plan the introduction, management, and socialization of these technologies and solutions into the organization (Peifer et al., 2022) to get staff acceptance.

Within the research undertaken in the past few years, there is a common thread, noting that establishing regulations, legislation, and governance principles without an auditing, compliance, or oversight structure gives limited protection or safety (Mökander et al., 2021; Sharkov et al., 2021). The lack of oversight perceived supports the premise that AI systems need to have some level of auditing or compliance checks (Roberts et al., 2022) in place to ensure that they are developed, deployed, and operated as per the prevailing regulations, governance or legislation. However, there does not appear to be a

consensus on whether this should be at an organizational, national, or international level. Furthermore, it is undecided whether the AI audit function should be done only during development or throughout the system lifecycle, and it is unclear how the requirements will change as AI matures from an assistant to a fully autonomous system (Raji et al., 2022). This is further complicated by the discourse between developers, researchers, and end-users on what level of human interaction or oversight is required within AI systems as they become more autonomous (Niet et al., 2021) and make life-impacting decisions.

The focus of the research for this thesis and the area of knowledge growth is on developing a compliance mechanism that safeguards that AI systems are safely introduced and used within the energy sector. The driving factor for this research is that within the critical infrastructure and services sectors, such as the energy sector, incorrect decisions influence more than financial returns but can cause serious equipment damage, premature failure, and harm people. The increased risk necessitates a structured approach to the protocols adopted or developed for governing regulation, legislation, principles, policies, and oversight for sustainable introduction of different maturity levels and complexity of AI within the sector.

To further the knowledge base from the existing literature review, a structured survey was undertaken with professionals within the energy utilities, regulators, information technology providers, to the energy sector, and AI software developers to better comprehend what governance, regulatory and/or oversight protocols already exist. This survey focused on identifying knowledge gaps in the industry regarding the safe and sustainable implementation of AI from an organizational, national, and international perspective to guide the outcome of this research. The survey questionnaire was issued to participants aligned with the energy industry in the North American and Caribbean region, with the participation being focused on energy sector information technology

professionals, AI system developers, regulators/legislators, technology specialists and decision-makers.

Of the one hundred and twenty-six participants that provided completed questionnaires, there was a fair dispersion of participants throughout the four key target groups, of which a sixty-seven percent majority represented the electricity sector or information technology system providers to the electricity sector. It was thought-provoking to see the level of knowledge of the existing AI regulations and governance structures in place to govern AI systems in the electricity industry but sobering to realize the overarching feeling that they were either not appropriately applied or insufficient to protect the organization, employees, equipment and the public. The participants indicate a need for a comprehensive AI regulatory and oversight compliance framework to be developed for the electricity sector and proposed that the most appropriate mechanism to develop, implement, and maintain this would be through a collaborative approach between the government, AI/information technology fraternity and the electricity sector. The parties further recommended a standardized compliance framework across the electricity sector, its support industries, and service providers to ensure that these entities build and operate compatible AI systems.

With the emergence of higher-risk AI and autonomous decision-making AI systems, the participants felt it imperative that a tiered approach be adopted for regulation and oversight compliance for the different levels of maturity of AI systems to ensure their effectiveness. Participants did not just want more stringent regulations and oversight implemented as AI levels of autonomy increased; they wanted a balanced approach that considered innovation, accountability, and autonomy in decision-making in setting the tiers. The participants further supported that human oversight, or involvement, was an essential factor in the compliance management process for AI systems. However, the level

of oversight should be dealt with using a risk-based methodology to modulate the human oversight requirements according to the deemed risk of the AI system. The participants acknowledged numerous risks and benefits of introducing an AI compliance framework but stated that the benefits outweighed the risks if a structured approach was established. However, they stressed that for an AI compliance framework to be beneficial and adopted, the AI compliance audit process should be included in relevant existing processes and frameworks within the organization. It was further acknowledged that for the AI compliance framework to be accepted and supported by the employees, the organization would need to provide targeted training for employees to understand how to work and collaborate with the AI systems.

These professional insights provided invaluable information to guide the proposed AI compliance audit framework development procedure, AI compliance audit framework, and AI audit process as the research outcomes. This is a fit-for-purpose compliance solution that can be adapted to all critical infrastructure sectors, to ensure that AI systems are safely and sustainably developed, deployed and used to support the current and future industry. Adopting the proposed risk-based approach to ensuring that the AI compliance audit framework is always aligned with the latest governing regulations, laws, AI system principles and industry standards, allows the energy sector to ensure compliance with the AI system and the safety of their organization, employees, the infrastructure and the public.

1.2 Research Problem

The energy sector is becoming complex, with deregulated models driven by real-time prices, intelligent technologies, distributed generation, and energy storage characterized by de-carbonization, decentralization, and digitization (Berger, 2018). Balancing demand and supply will require autonomous intelligence management systems

to optimize operations and decision-making (Laroussi et al., 2023) in near real-time. Energy organizations must innovate, develop, and adopt new intelligent solutions to manage the more complex digital energy system (Lyu and Liu, 2021). AI is quickly becoming one of the vital technological innovations providing promising solutions for the critical infrastructure and services sector, such as energy sector operations.

Problem 1 – Lack of an enabling environment to integrate and operate AI in the energy sector.

A standard gap in the prior research and studies is the enabling environment that the leadership and management team must establish for safe and sustainable implementation and assurance of varying maturity levels of AI within the critical infrastructure and service sectors, especially in the energy sector. The historical focus has been on leadership and organizational changes after the implementation of AI. However, no protocols have been developed to outline and manage the human intervention and oversight required for AI system outputs or decisions made throughout its maturity lifecycle as it evolves from a rudimentary level to a fully autonomous decision-maker.

Problem 2 – Lack of a structured governance/regulatory oversight mechanism for integrating and operating AI in the energy sector through its maturity lifecycle.

As AI technology matures and becomes more complex, it functions as a "Blackbox" where there is limited transparency, traceability, or explainability of the functioning or outputs of the system, which leads to a lack of confidence and distrust of the system operations, outputs, and decisions (Machlev et al., 2022). In the critical infrastructure and services sector, especially the energy sector, incorrect actions and decisions from an AI

system can influence more than financial returns; they can also cause severe equipment damage, premature failure, and harm to people. The dynamic nature of AI and its evolving role in energy organizations necessitate a sophisticated governance and compliance framework. These issues highlight the problem under investigation: the lack of a structured oversight in implementing governance and regulatory protocols for integrating AI into the energy sector, specifically the electricity utility sector. For AI systems to be accountable, explainable, and trustworthy for the energy sector and for the public to accept them and their decisions (Slate *et al.*, 2024), compliance with the AI system designs, and operations needs to be proved against regulations, policies, principles, standards, and norms (Raji et al., 2022; Walz and Firth-Butterfield, 2019). This problem becomes more prominent as AI matures in autonomy and requires governance and ongoing compliance protocols to evolve with the maturing technology.

Problem 3 – Lack of National and International collaboration on AI oversight protocols.

The lack of organizational, national, and international collaboration between the different energy providers, developers, governments, and countries to establish standardized governance, regulatory, and oversight protocol for the industry to ensure AI system replicability and compatibility is creating a platform for incompatibility, un-competitiveness, bias, and distrust.

1.3 Purpose of Research

This research aims to provide a comprehensive review of regulations, industry practices, and compliance audit frameworks. It will facilitate a controlled engagement with industry experts to identify existing knowledge and gaps. Thereafter it will propose future mechanisms that will be required to introduce a conceptual framework for the oversight of the safe and sustainable integration of AI into the energy sector.

1.4 Significance of the Study

The research study is significant as it contributes to the knowledge base by undertaking a regulatory, governance, and ethical impact gap assessment of AI in the energy sector as it evolves and matures. Furthermore, it proposes a methodological compliance framework that can facilitate and guide the safe and sustainable integration of evolving AI technologies into the energy sector. The outcome of this study will be valuable to the critical infrastructure and services sector, especially the energy sector, in establishing safe practices, policies, and procedures for integrating AI systems into the industry. It will also provide the related system software developers and AI system providers with measurable guiding principles in developing better AI tools for this sector.

1.5 Research Purpose and Questions

This research aims to gather information and undertake a gap analysis on four main areas related to the integration of AI into the energy sector:

1. Establish a baseline of existing governance structures, regulations, legislation, rules, design principles, and standards for the design, development, and deployment of AI in the energy sector. The focus will also be on what mechanisms and metrics have been established to measure whether the AI

systems are designed, developed, and deployed according to the existing governance structures.

2. Identify the industry's professional view on the responsibility and accountability matrix for governing and ensuring compliance with AI system design, development, and deployment in the energy sector. This outcome will guide the ultimate structure of the compliance framework to ensure that it is correctly framed to be accepted and adopted by the energy sector.
3. Identify knowledge gaps in the existing AI governance and compliance structures in the energy sector focusing on the human and compliance oversight perspectives. This will set the base for the development of the conceptual human oversight and compliance audit framework for the oversight of the design, development, and implementation of AI technologies, encompassing ethics, morale, safety, human oversight, regulatory and compliance auditing aspects over its maturity lifecycle in the energy sector.
4. Lastly, compare the proposed conceptual compliance framework against existing state-of-the-art techniques under crucial parameters, such as ethics, human rights, corporate law, safety, and operational compliance.

CHAPTER II:

REVIEW OF LITERATURE

2.1 Theoretical Framework

For the sake of this research, a deductive framework was utilized to quantify the original hypothesis regarding the use of AI in the critical infrastructure sector, with the specific focus being on the energy or electricity sector. The hypothesis for this research focuses on the need for a facilitated lifecycle compliance framework for the safe, reliable, and sustainable development, implementation, and usage of AI within the critical infrastructure sector. Researchers have explored the fundamental theories in many spheres before, but no clear construct or direction has been adopted or supported. The focal theories used to guide the research are human oversight in AI, AI regulation development and operationalization accountability, AI compliance metrics, and AI lifecycle oversight.

2.2 Literature Review Theories

AI has intrigued Humankind for decades (Fenwick and Molnar, 2022) as it is perceived as a tool to improve efficiencies (Berger, 2018), further organizational interests and improve societal well-being. Historical studies have reviewed the key drivers and barriers to implementing and accepting AI into organizations and society (Cubric, 2020). Numerous dispersed studies infer that organizations, management, and leaders have differing views on how ingrained AI will become in organizations. The area of greatest misalignment among leaders is whether they would accept AI transitioning from being an assistant to humans to being trusted as an autonomous, independent decision-maker (Johansson and Björkman, 2018). As AI has matured in the marketplace, more research has been undertaken to understand AI's actual benefits and risks (Ulnicane *et al.*, 2021) and to identify governance and regulatory requirements to mitigate risks (Dafoe, 2018).

A literature survey was undertaken through peer reviewed technical publishers, university press publishers, and research platforms, such as Google Scholar, Social Science Research Network (SSRN), Academia, and Science Direct to identify prior research on AI accountability, AI governance and regulation compliance, compliance or conformity auditing of AI, AI lifecycle functionality verification, and AI oversight in the energy sector. The survey identified hundreds of thousands of studies on similar topics, of which more than three hundred were identified as relevant to this proposed research. In a review of these relevant studies, there are five common themes or categories that guide and shape this proposed research. The statistical overview of the relevant literature for the five themes is shown graphically in Figure 1, and summarized below:

1. Twelve percent (12%) overview the complication of governing or regulating AI systems due to the lack of a universally accepted definition of AI, its fundamental principles, and norms.
2. Fifteen percent (15%) overview the risks, benefits, and opportunities to organizations, the public, and the global community when AI systems are integrated into organizations.
3. Thirty-three percent (33%) overview the structured governance and regulation requirements to safely and ethically implement AI in organizations.
4. Thirty-four percent (34%) overview proposed auditing mechanisms and frameworks to ensure that AI systems are developed, deployed, and used in compliance with standards, regulations, and industry norms.
5. Six percent (6%) of the overview proposed structures to include human oversight into the compliance audit protocol for AI, create a collaborative human/machine environment, and ensure that human centricity is maintained.

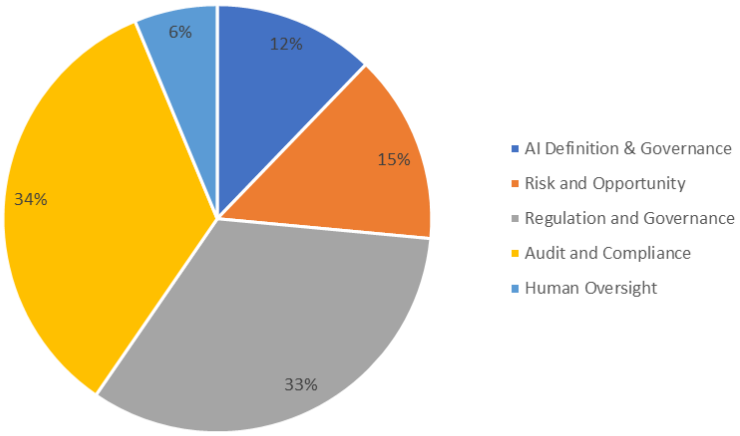


Figure 1: Graphical dispersal of literature per category

2.3 Literature Review Methodology

A systematic literature review process was chosen to undertake the literature review to identify gaps in the current knowledge base and to guide the proposed future research. This literature review type follows a structured review protocol and quality procedure to select relevant studies, extract relevant information from the selected studies, and analyse relevant information to answer structured research questions (Paul and Criado, 2020).

The introduction of AI tools and systems, into the energy sector, to facilitate extensive data processing, and facilitate decisions, makes it essential to better understand the structures already established for the safe and sustainable implementation of AI systems within the sector. This literature review aims to explore the benefits, opportunities, risks, and mitigation strategies established in the existing AI governing structures. The literature review was done as per the taxonomy captured in Figure 2.

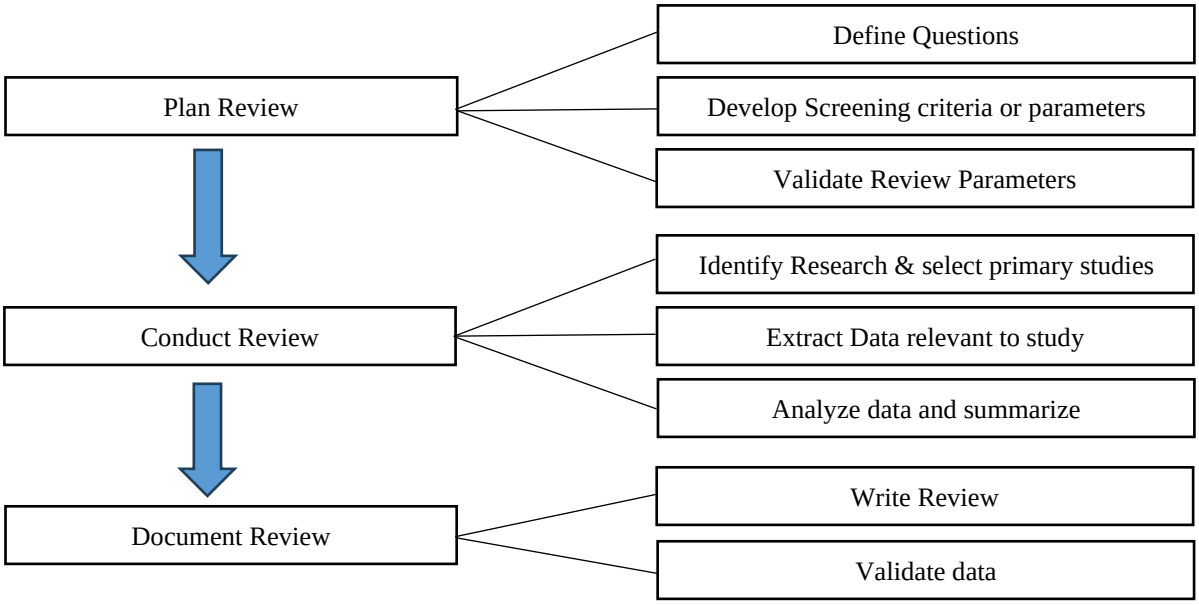


Figure 2: Literature review taxonomy

Screening criteria were developed to select appropriate and relevant articles for inclusion in the literature review. The research only focuses on oversight, assurance and compliance within the development, integration and usage of AI in the critical infrastructure sector, with specific focus on the energy sector. To focus the study to a representative sample, the following screening criteria was used to refine the search:

- Duplicates between different research questions were eliminated.
- Eliminate non-English publications.
- Include articles from the <= 5 years, screened according to:
    - Type of publication – article, conference paper, research article, standards, non-fiction book.
    - Study type – systematic literature review or scoping literature review.
    - Definition of terms – Specific to AI in energy, ethics in AI, AI governance, AI regulation.

    - Quality of evidence – leading journal, peer reviewed, linked to law, standards or regulations.
    - Reported Outcome – AI compliance or AI auditing framework or standard or AI compliance structure in critical infrastructure.
- Articles written within the past 10 years, cited within the relevant research in the past five years were included.

Figure 3 graphically depicts the identification and screening steps of the PRISMA methodology used. From the 63 databases queried, a list of 1,937,217 possible papers were compiled, with 75% of the results being from 18 top technical publishing houses, universities, or technical institutes. Of these articles and documents, 1,231,527 were removed as they were duplicates or in a foreign language. The author identified and characterized the contributions of the balance of the 705,689 articles and documents through an examination of the abstracts and summaries to gather the value added to this research. These articles and documents were screened as per the screening criteria described above, with an outcome of a collection of 344 articles or documents potentially eligible to guide in answering the research questions.

In reviewing this source material, several common threads and gaps were identified that are retarding the creation and deployment of a structured regulation and compliance regime for the development, implementation, and sustainable operability of AI in the critical infrastructure sector.

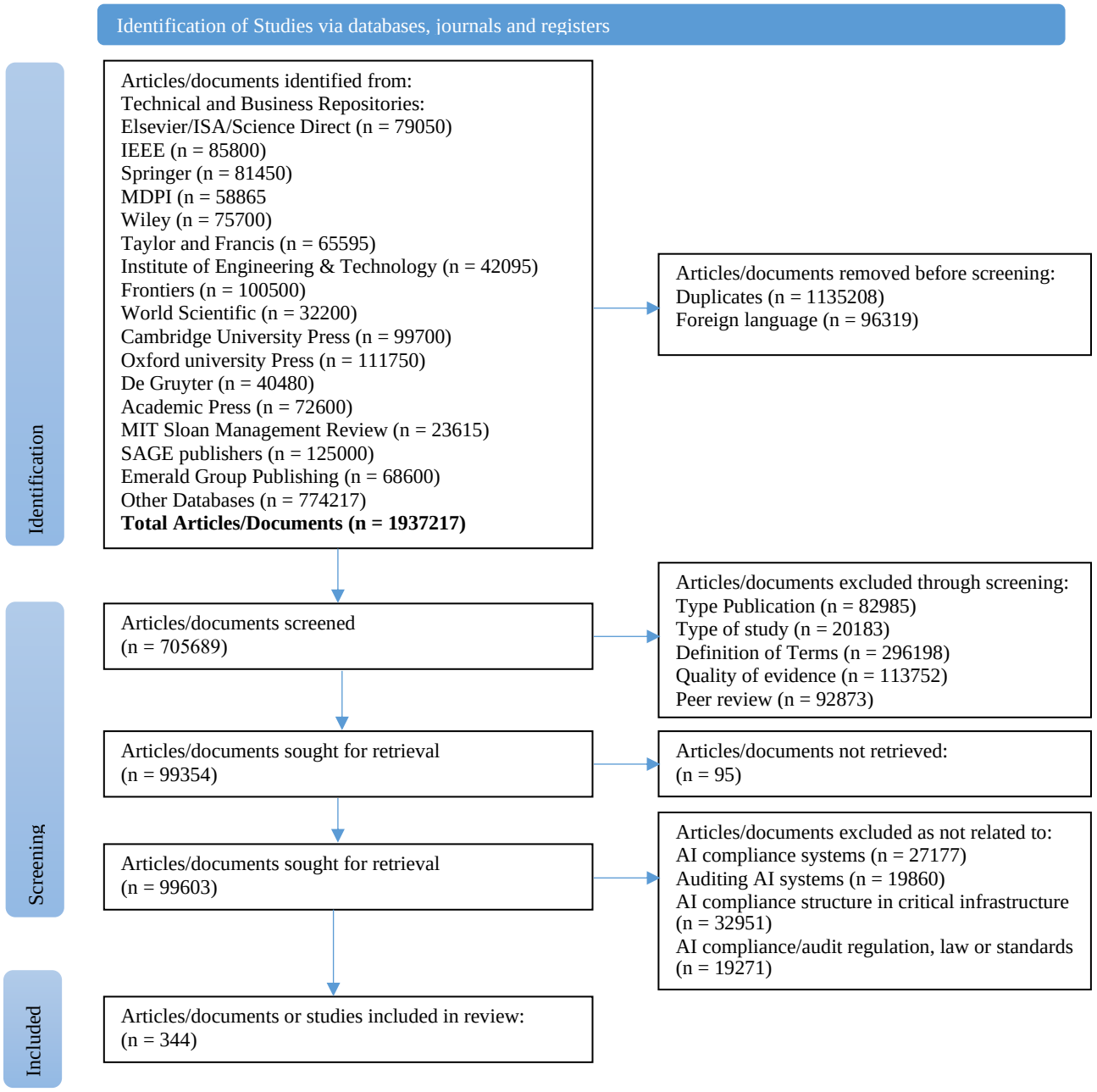


Figure 3: PRISMA Flowchart

The following subsections provide an overview of the literature reviewed in these critical themes and summarizes knowledge gaps in the current research and opportunities for further research.

2.4 Artificial Intelligence definition and governance

AI is still a complex fledgling technology, and there is still no clear, accepted definition for AI (Lyu and Liu, 2021). AI is inherently difficult to define due to its multifaceted nature, hence the various definitions, ranging from simple algorithms to complex systems that mimic human cognition, linked to decision-making and autonomy (Berente *et al.*, 2021). As researchers and developers are redirecting towards responsible, trustworthy, and explainable AI concepts, it is indicative that there is a lack of standardization and no clear definition (Williams *et al.*, 2022; Wilson and Van Der Velden, 2022; Zimmer *et al.*, 2022). This lack of consensus leads to confusion and miscommunication among stakeholders, hindering effective governance.

The concept of AI is not static; it evolves as technologies advance and change. Historically, techniques or systems once considered AI, such as expert systems, are now often categorized as traditional software solutions once their mechanisms are transparent and understood. This phenomenon, known as the "AI effect" complicates establishing a stable definition, as what qualifies as AI can shift over time even further mudding the scene (Lea, 2023).

An acceptable single definition of AI or its derivatives is needed to set up proper governance (Gasser and Almeida, 2017) and regulation structures (Pugliese *et al.*, 2021). Without a common understanding, regulations may be misaligned with the technologies they aim to govern. For instance, treating AI as a monolithic entity that will not evolve and change fast can lead to oversimplified regulations that fail to address the unique characteristics of different AI systems (Michel, 2023). Even when considering specific characteristics within the field of AI, such as bias, it highlights that there are different definitions and interpretations, which (Landers and Behrend, 2023) lead to inconsistencies in how AI systems are governed, developed, implemented and used. It is vital to harmonize

definitions of AI, its characteristics, and fundamental principles to govern, evaluate, and manage AI (Kharchenko *et al.*, 2022).

Many existing policies are based on entrenched assumptions about AI, such as its supposed intelligence and ethical capabilities. These assumptions can lead to policies overlooking potential risks and negative impacts, particularly for marginalized groups. A more nuanced understanding of AI is essential to create effective and equitable policies (Sheikh *et al.*, 2023). The ambiguity surrounding AI definitions also extends to ethical considerations in designing and using AI systems. Different interpretations of AI can lead to varying ethical frameworks, which may conflict. For example, a system defined as "intelligent" may be expected to adhere to ethical guidelines that do not apply to simpler, rule-based systems. This inconsistency can create challenges in establishing accountability and responsibility for AI systems (Lea, 2023).

In summary, the absence of a single definition of AI, along with the intrinsic ambiguity of AI, its evolving nature, and the varied interpretations of its capabilities, complicate structuring robust governance, policy development, and ethical considerations. To address the challenges, it is essential to foster evidence-based discourse that includes diverse perspectives, ensuring that governance frameworks are adaptable and reflect the complexities of AI technologies (Sheikh *et al.*, 2023). Ultimately, a clearer understanding of AI will facilitate the creation of effective policies that can guide the responsible development and deployment of these transformative technologies (Scott, 2024).

2.5 AI risks and opportunities

AI systems and tools in organizations offer many opportunities, benefits, and risks (Isensee *et al.*, 2021; World Health Organization, 2021), which may differ depending on the type of organization, geographical area, services, and culture (Fjeld *et al.*, 2020). The

impact of AI can be seen from the perspective of governments, organizations, cultures, and people (Kaminski, 2023), but due to the overlapping and interlinking effects, they have been reviewed as a collective historically.

Some key risks associated with the introduction of AI into organizations are that AI can lead to ethical and regulatory breaches (Floridi *et al.*, 2022; Morris *et al.*, 2022), which can lead to diminished public trust and loss of organizational reputation (Caner and Bhatti, 2020). If AI was designed with inadequate regulations and governance principles (Winfield *et al.*, 2019), or if it fails, there is a risk of the system causing physical harm to infrastructure, people and financial loss to the organization (Mannes, 2020). Should AI be poorly regulated and governed throughout its development, deployment, and implementation lifecycle, there is a risk of the systems being used for illegal or criminal undertakings (Walz and Firth-Butterfield, 2019). There is a risk of human rights violations (Leslie *et al.*, 2021), where systems will have an inherent bias towards cultures or races and can lead to victimization and inequality (Dignam, 2020). Furthermore, implementing AI will impact people's jobs (KILINÇ and Aslıhan, 2020), the best case scenario being that people need to reskill to stay relevant (Burton, 2019; Walz and Firth-Butterfield, 2019), and the worst case being outright job losses (Caner and Bhatti, 2020).

AI can affect an organization's data privacy and decisions made (Fjeld *et al.*, 2020), as well as increase the risk of cybersecurity breaches (De Silva and Alahakoon, 2022). One of the primary concerns is that if AI systems are not properly secured, this can introduce vulnerabilities that malicious actors can exploit to gain unauthorized access to critical systems (Leidy and Gerstein, 2024). Researchers have similarly explored AI-powered cyberattacks, such as AI-generated malware and AI-driven network intrusions, which can evade traditional security measures. AI systems rely on large amounts of data and computing infrastructure to function effectively, if this data or infrastructure is

compromised, it can lead to cascading failures in critical infrastructure systems. The centralization of AI data and computing resources can create single points of failure that can be targeted by attackers and increase risk of ransom-based attacks (Kelley, 2024).

The complexity and opaqueness of most AI systems make it a challenge to understand their decision-making processes, which is a significant concern when these systems are used in critical infrastructure applications. The lack of transparency and explainability in AI systems can lead to unpredictable and potentially harmful outcomes, making it challenging to hold AI systems accountable for their actions (Yigit *et al.*, 2024).

That being said, as much as AI poses risks within organizations, they also provide many opportunities and benefits, such as improved service offering range and quality, customer services, efficiency, and profitability (Bankins and Formosa, 2023). AI applications can improve living conditions and health, facilitate justice, create wealth, improve public safety, and mitigate the impact of human activities on the environment and the climate (Stahl and Stahl, 2021). As much as AI is noted as being a risk for loss of jobs and roles within an organization (Moldenhauer and Londt, 2018; Winfield and Jirotka, 2018), it is also an opportunity for the creation of more skilled jobs and new roles (Jarrahi *et al.*, 2023; Lyu and Liu, 2021), however the true impact of this will be dependent on the oversight and decision-making principles agreed for AI systems (Fanni *et al.*, 2023). By partnering AI and humans together, people can complete jobs faster, more accurately, and facilitate more complex tasks (Stahl and Stahl, 2021).

AI-powered systems can provide benefits by continuously monitoring critical infrastructure systems for anomalies and potential threats. As much as processing vast amounts of data is a risk, if used correctly data analysis from various sources can empower AI to detect patterns and identify potential issues before they escalate into more significant problems (Gudala *et al.*, 2019). AI systems can bolster critical infrastructure to become

more resilient and adaptable to changing conditions and unexpected events by using AI to analyse real-time data and make rapid decisions, allowing the system to respond faster to disruptions and recover more efficiently (Florkowski *et al.*, 2024; Franki *et al.*, 2023). Finally and most importantly, AI has many societal benefits and opportunities, including the potential to improve healthcare, finance, education, and surveillance (Nasim *et al.*, 2022).

2.6 Regulation and governance

As AI gathered traction in organizations, governments, and societies as a whole, more focus has been placed on researching and developing mechanisms to regulate or govern AI (Perry and Uuk, 2019), along with proposing multitudes of principles, rules, guidelines, and other structures to ensure that AI is developed, deployed and implemented in a safe, human-centric and sustainable manner (Huang *et al.*, 2022). Throughout the literature reviewed, there is no clear consensus on how the AI lifecycle should be managed to ensure no harm; some studies recommend government regulation or legislation (Djeffal *et al.*, 2022; Ferretti, 2022; Taeihagh, 2021), others refer to industry-wide governance principles and regulations (Roski *et al.*, 2021), or organizational self-regulation through structured principles, standards and guidelines (Walz and Firth-Butterfield, 2019). It was stated that principles are the start of governance but not the end of it; policies, laws (human rights), regulations, professional practice, and everyday routines also need to be implemented in a structured manner (Huang *et al.*, 2022). In essence, it will take a collaborative approach between governments, organizations, non-profits, and society (Xue and Pang, 2022), with a hybrid governance process to ensure that AI is developed, deployed, and implemented safely through its lifecycle (Gasser and Almeida, 2017).

From an organizational perspective it was (Kitsios and Kamariotou, 2021) inferred that for AI to be successful, it is important for the information technology strategy to be aligned with the corporate strategy to get organization wide acceptance. AI in leadership research supports this premise and expands it to recommend that the Chief Information Officer should be a represented member of the organization's Board of Directors to provide strategic and technical direction (Li *et al.*, 2021).

Another school of thought is that AI regulation must be facilitated internationally to ensure inclusivity, diversity, and safety across the international markets (Clarke, 2019; Hickok, 2021). Studies of AI policies in the European Union, the United States of America, China, India, and Australia show that governments have developed AI strategies and other proposed governing policies, principles, and guidelines, however, they are focused on different outcomes for the people and the countries (Roberts *et al.*, 2021). For this to be a globally driven regulation and governance process (Ala-Pietilä and Smuha, 2021; Walter, 2024), there must be a shared vision and collaboration between countries, organizations, and society (Dafoe, 2018; Daly *et al.*, 2020).

Traditional regulatory frameworks are complex and fragmented, which poses risks of slow responses to challenges brought on by AI (Hadzovic *et al.*, 2024). Considering the inflexibility of the different regulatory, governance or management structures for AI (Schiff *et al.*, 2020; Schultz and Seele, 2023), it has been highlighted that AI systems, as they mature and broaden their scope, are becoming "Blackbox" systems, where no one knows how the inputs are processed in the system and how outputs are determined (Bankins and Formosa, 2023; Dwivedi *et al.*, 2021; Machlev *et al.*, 2022). This lack of understanding of the inner workings of AI systems, along with concerns on societal impact and human rights violations (Kop, 2021) has marked a step change in the focus on developing ethical and moral principles to guide AI development (Gutierrez and Marchant,

2021; Huang *et al.*, 2022; Ryan and Stahl, 2020). However, many studies have indicated that these principles are not easy to implement or action (Morley *et al.*, 2021; Stix, 2021a), which has led to the development of AI systems focused on specific ethical principles and attributes.

The majority of the AI systems developed to portray a specific attribute or characteristic (Baker-Brunnbauer, 2021; Haakman *et al.*, 2021; Mikalef *et al.*, 2022) focus on creating visibility on what and how the systems use data and make decisions, to ensure that there is a level of ownership of the AI systems operation (Stahl, 2022) and its outcomes. Explainable AI is the premise of developing an AI ecosystem that helps characterize model accuracy, fairness, transparency, and outcomes in AI-powered decision-making (Lima *et al.*, 2022). Trustworthy AI is designed to be lawful, ethical, and robust both from a technical and social perspective (Smuha, 2021), which leans toward the ethical pillars of human autonomy, preventing harm, fairness, and explicability (Hickman and Petrin, 2021). Responsible AI, often interchanged with Trustworthy AI, is not just about having a system that has no bias, is fair or ethical, but that it does what its design claims it will do in a replicable manner (Lu *et al.*, 2024; Schwartz *et al.*, 2022). Finally, Accountable AI aims to ensure the effectiveness of concepts such as transparency and explainability in automated systems and hold the developer, implementer, and user accountable against a prescribed set of standards, regulations, or legislation (Williams *et al.*, 2022). Organizations must ensure that AI is developed and aligned with these specific central principles, characteristics, and attributes, as it will build trust and confidence when putting AI models into production and ensure human flourishing (Stahl and Stahl, 2021).

2.7 Compliance and audit

In a world driven by economic growth, profitability, and organizational bottom line (Ulnicane *et al.*, 2021), there will always be a trade-off between organizational objectives, society, and the environment (Martin *et al.*, 2022). This same conundrum exists in the introduction of AI systems into organizations, it is thus crucial that mechanisms be introduced to ensure that systems are developed, deployed, and operated as per the prevalent principles, guidelines, regulations, and legislation (Falco *et al.*, 2021). Initial audit and compliance research focused on technical audits (de Laat, 2021) in the form of a risk and impact assessment during development and deployment to highlight and mitigate issues (Kazim and Koshiyama, 2020). Following this, researchers considered actual algorithm assessment against preset standards, norms and outcomes (Costanza-Chock *et al.*, 2022) to ensure accountability and mitigate bias and harm while noting that actionable policies are required to improve the quality and impact of audits (Stix, 2021b). The European Union and its members have been active first movers in proposing governance and regulation structures, which promote enforcement mechanisms, including conformity assessments and post-market monitoring plans (Mökander *et al.*, 2022), but this was focused predominately on high-risk systems. For low-risk systems, it was proposed that a voluntary labelling scheme be established to highlight the principles, regulations, and design criteria that systems are designed to conform to (Stuurman and Lachaud, 2022).

Many researchers and system developers have focused on specific aspects of the AI system or tools that they feel are most prevalent to monitor or control. One proposed option was to establish a risk-based approach to data protection and show that a system uses, processes, and stores the organization or person's personal data according to accountable principles (ICO, 2020). The Information Systems Audit and Control Association prepared a white paper to guide IT auditors on transitioning to auditing AI

systems. The proposal from the association is that the auditors should be auditing the IT governance of the AI systems and not focus on the algorithms, which should be left to the model specialists, who are predominately the developers (ISACA, 2018). The proposal is to focus on fairness, transparency of decision-making, and traceability of data and decisions rather than the wholistic system operability (Simbeck, 2024). An alternative approach proposed is establishing an organizational AI governance framework (Mäntymäki *et al.*, 2022), which is focused on establishing a governance process and not ensuring that the AI system works per the legislative and regulatory principles.

Other researchers proposed a third-party audit ecosystem establishment to move the market away from in-house audits and assessments to having independent auditors confront, verify, and subject to scrutiny performance claims, thereby creating transparency and building public trust (Raji *et al.*, 2022). To earn public trust, AI audit structures need to focus on system trust (Kaur *et al.*, 2022) and the integrity and trustworthiness of the organization about AI. This entails accurate, detailed AI documentation showing compliance with guidelines and legislation, which are audited by external auditors (Knowles and Richards, 2021). There is also a school of thought noting that the focus should be on applying ethical standards and psychological science in auditing AI systems to minimize harm and increase public trust (Landers and Behrend, 2023).

The common thread shown through the review of research notes that establishing regulations, legislation, and governance principles without an auditing, compliance, or oversight structure gives limited protection (Mökander *et al.*, 2021; Sharkov *et al.*, 2021). As much as there is agreement that a structured compliance mechanism is required, it is undecided whether the AI audit function should be done only during development or throughout the system lifecycle. It is unclear how the requirements will change as AI moves from an assistant to a fully autonomous system (Raji *et al.*, 2022).

2.8 Human Oversight

An ongoing debate between researchers, academia, AI developers and system users is on what role humans should play in the decision-making process regarding AI outputs and how they are used. The other fundamental discussion commonly raised is at which stage of the AI design, development, and deployment humans should have oversight of the systems, data, and algorithms, and at which level the oversight should be pinioned? The European Union's AI Act defines human oversight as a counterbalance between the aims of automation and authentic human reasoning. The Act further infers that the obligation to ensure human oversight implies that humans are expected to counterbalance some of the associated risks with AI systems (Enqvist, 2023). It is all good and well to propose that human oversight will assist de-risking AI systems, but that will be greatly dependent on the type of systems they are to oversee, the transparency of these systems, the capacity, knowledge, and capability of the individuals doing the oversight, as well as their mandates and working conditions. From research, much of the academic and research fraternity agree that human oversight can help to build public trust in the system and, if correctly implemented, can provide guardrails that safeguard the public.

Efficient human oversight in the training dataset is expected to provide developers with an important competitive advantage. However, this oversight's accuracy and quality depend on providing clear rules to the overseers and properly financially incentivizing them (Laux et al., 2023). In a Harvard Business Review article, the authors observe that AI has progressed to compete with the best human brains in many areas, often with stunning accuracy, quality, and speed, but raises the question of how to change the decision from being made purely from a cold, calculating judgment perspective. Often AI systems miss the big picture and cannot analyse the decision with the reasoning behind it as it does not include subjective experiences, feelings, and empathy in the decision-making process

(McKendrick and Thurai, 2022). It is stated that human oversight is tasking humans with oversight of AI algorithms that were put in place with the promise of augmenting human deficiencies, is this not creating a platform for false comfort and distracting from the inherent harmful uses of automated AI systems (Green and Kak, 2021)? One specific article notes that humans are regularly incapable of appropriately supervising AI in complex human-machine interactions which leads to the failure to protect the outcome (Beck and Burri, 2024). Some researchers purport that mandating the human oversight implementation is flawed as humans are unable to perform the oversight function due to the system's complexity and that it provides the government with a free pass to legitimize the use of controversial AI algorithms with no reprisal (Green, 2022).

Many questions have been raised regarding the European Union's AI Act's proposal that human oversight be used on high-risk AI systems, with the key concern being that the Act does not outline how this would be achieved and how to circumvent the limitations. Researchers propose that human oversight should only be implemented when it is effective or meaningful and that the Act should be revised to undertake an empirical test to ensure human oversight effectiveness before being implemented (Walter, 2023). Humans play an important role in the design, development, training, and deployment of AI systems, however, with the lack of definition of the meaningful role of humans in the oversight of this process and the collaboration between humans and AI, there is a responsibility gap when something goes wrong (Christen *et al.*, 2023). SGS Digital Trusts Services published a research paper proposing that the focus should be broader than just human oversight within AI but instead should focus on human agency as that is the foundation of trustworthy AI (Kopeinik *et al.*, 2023). The author states that Human agency maintains the human centricity of AI systems, as well as the autonomy of humans who use

and are exposed to the results of AI systems, and not just focusing on humans influencing the learning and actions of the system as is the case in human oversight.

Since the 1940's there have been an abundance of terms such as human-in-the-loop, human-on-the-loop, human-out-of-the-loop, and human-in-command used to define human oversight for automation systems, but almost all organizations and researchers have differing definitions for these. The question that is raised is why after centuries of successful alignment between human and machine is the community diverting from the age-old tested mechanism of aligning human and machine in automation systems, to something that is not ethical (Anderson and Fort, 2022). Many researchers and developers focus on including human oversight into the development and training of an AI system, but seem to forget that humans become users of AI systems, so they are not only looking at proper technical performance but also for the system to be easy to use, have transparent outputs and capable to achieve the objectives they have set (Mosqueira-Rey *et al.*, 2023).

One researcher notes that humans face three key issues when trying to regulate and oversee AI systems (Brown and Albert, 2023):

1. Machines cannot understand constraints beyond their programming and so do not act within human ethical or operational constraints.
2. Machines act without traditional notions of intent and causation that underlie core legal frameworks, making their frameworks inapplicable in large part to address actions taken by machines.
3. Machines will evolve to create unanticipated harm, and risk at the individual level, that backward-looking regulation cannot to keep up with or address.

The researcher further proposes that a new paradigm of legal regulation is required, not one of explicit rules but rather one that sets boundaries so that machines can evaluate as they develop new judgments and decisions. Successfully achieving this requires human

oversight, so that a human party with social and moral underpinnings can be held accountable and respond to the legal framework.

Human involvement is at multiple levels of design, development, training, and deployment of AI, and that involvement needs to be deliberately designed at each level to be effective. It is important to understand what aspects of the system process oversight will be aimed at when required, as this will set the risks or harms that humans need to try detecting. Lawyers, scholars, and the public are calling for human-machine teams with the necessary transparency and explainability so that the AI systems cause no harm; to do this requires meaningful human control (Davidovic, 2023). It is opined that there are five critical purposes of implementing meaningful or effective human control: establishing safety and precision, responsibility and accountability, morality and dignity, democratic engagement and consent, and institutional stability (van Diggelen *et al.*, 2024).

Finally, the question that all are avoiding in human oversight is whether AI should govern humanity. There are two opposing arguments, the first being that AI should not govern humanity, as it lacks the consciousness, emotions, moral compass, and decision-making abilities of humans, and AI is a tool created by humans, and as such, it should be used to serve human goals and objectives, not to govern them (Torrance and Tomlinson, 2023). The second argument is that AI should govern humanity due to its inherent characteristics, such as impartiality, fairness, and ability to simultaneously work with large amounts of data (Torrance and Tomlinson, 2023). As this debate unfolds, it is crucial that these decision-making systems are designed with an awareness of human values, and that they are transparent and explainable (Torrance and Tomlinson, 2023). These debates clearly indicate that there is a risk of extreme catastrophe if we do not have a structured approach to ensuring AI is implemented safely and that it functions within a set of structured guardrails.

2.9 Existing Oversight and Audit Frameworks

As AI matures and becomes autonomous, the public and organizations acknowledge the increasing risk to society and humankind. The increased risk has not only driven a flurry of principles, regulations, and policies development, but has focused governments, academia, and private organizations on the development of AI compliance audit methods. The main issue is that each entity focuses on different aspects of AI governance, risk management, and ethical compliance with no apparent convergence on method and approach. On reviewing mature academic, government, and developer research, white papers, regulations, and frameworks, it was possible to provide a summary outlining their core focus and methodologies of compliance monitoring, which provides a basis for comparison to the proposed outcome of this research.

The Information Commissioner's Office (ICO) developed a comprehensive framework for auditing AI systems, emphasizing best practices for data protection compliance. The framework is designed for compliance professionals and technology specialists, providing methodologies to audit and ensure fair processing of personal data, with guidance structured around accountability, lawfulness and fairness, security and data minimization, and individual rights (Kazim *et al.*, 2021).

The European Union’s AI Act is a comprehensive regulatory framework aimed to ensure AI technologies’ safe and ethical development and deployment. It categorizes AI systems based on risk levels and outlines specific compliance measures for each category, mainly focusing on high-risk applications. Providers of high-risk AI systems are subject to stringent compliance measures, including conducting detailed evaluations to identify and mitigate potential risks associated with the AI system, maintaining comprehensive records throughout the lifecycle of the AI system, from design to post-market monitoring, and conformity assessments to demonstrate compliance with the AI Act's standards,

allowing providers to affix a CE marking to their products, which signifies adherence to European Union's regulations (Musch *et al.*, 2023). This Act is one of the more comprehensive regulations that has been approved, however, it still has limitations due to its risk-based approach for managing AI systems being tied to naming specific risks and applications rather than a definition, which means it will be outdated swiftly and ineffective. The list-based approach is likely to be ineffective on the procedural complexities of AI system development, deployment, and use, which in turn might fail to duly acknowledge the influence AI has on people's daily lives, including realizing their fundamental rights (Beck and Burri, 2024).

The National Institute of Standards and Technology (NIST) developed the AI Risk Management Framework, which provides guidelines for managing AI-related risks. By providing a structured approach and emphasizing the socio-technical nature of AI systems, the framework aims to empower innovative and ethical development in AI. This framework is intended for voluntary use and aims to improve the incorporation of trustworthiness into AI systems (NIST, 2023).

The Chartered Institute of Internal Auditors (IIA) published an Artificial Intelligence auditing framework to guide internal auditors on approaching AI auditing. The framework addresses the unique challenges AI technologies pose and emphasizes the importance of governance, data quality, and performance monitoring in the auditing process. The focus on risk assessment, best practices, governance, and continuous learning, empowers internal auditors to effectively navigate the complexities of AI and provide valuable assurance to their organizations. Adopting these guidelines will enhance the auditing process and contribute to the responsible and ethical use of AI in various sectors (IIA, 2023). One of the key issues here is that this is more focused on governance than system operational compliance. Where compliance is focused on, they are relying on

internal audit to generate a report against metrics not set by them, against a technology that they and the people setting the metrics, may not be competent to understand due to the complexities.

Singapore's Personal Data Protection Commission (PDPC), in conjunction with the World Economic Forum, created the Model AI Governance Framework that focuses on the ethical and responsible use of AI technologies (PDPC and IMDA, 2020). The Model AI Governance Framework represents a proactive approach to AI governance, aiming to facilitate innovation while safeguarding consumer interests. By providing structured guidance and tools for implementation, the PDPC seeks to ensure that AI technologies are developed and deployed responsibly, fostering a sustainable digital economy (PDPC and IMDA, 2020).

The United States Government Accountability Office (GAO), developed an accountability framework specifically for federal agencies, providing guidelines on ensuring AI systems comply with existing regulations and ethical standards (GAO, 2021). The framework highlights the need for accountability mechanisms, including third-party assessments and audits, to foster trust in AI technologies. As AI continues to transform various sectors, including healthcare, transportation, and defence, establishing robust accountability practices is essential for safeguarding public interests and promoting ethical AI use. In short, the GAO's AI Accountability Framework provides a structured approach for federal agencies to implement AI responsibly, ensuring that these powerful technologies are used in ways that are transparent, accountable, and aligned with public values (GAO, 2021).

The Control Objectives for Information and Related Technologies (COBIT) Framework, particularly its 2019 iteration, though not solely focused on AI, offers a structured approach to auditing AI systems, ensuring alignment with organizational goals

and ethical standards. By leveraging this framework, organizations can ensure that their AI initiatives are practical but also ethical and compliant with regulatory standards. This holistic governance model positions organizations to harness the transformative potential of AI while mitigating associated risks, ultimately driving sustainable growth and innovation. The most significant shortfall of this framework for compliance auditing for AI is that it focuses on governance from an information technology perspective rather than on the complex AI environment.

The European Data Protection Board (EDPB) published a comprehensive checklist for auditing AI systems, developed by external expert Dr Gemma Galdon Clavell, to assess compliance with the General Data Protection Regulation (GDPR) and the European Union's AI Act. This checklist provides a structured methodology for conducting end-to-end audits of AI systems from a socio-technical perspective (Clavell, 2023). The EDPB AI auditing checklist predominantly focuses on data handling and protection. This lack of focus on the broader functionality of AI system compliance does not provide a sufficient compliance overview over the AI system's lifecycle.

The International Organization for Standardization developed and published a number of AI focused standards, but in 2023 published the first international standard dedicated to Artificial Intelligence Management Systems (AIMS), namely ISO/IEC 42001:2023. This standard provides a structured approach for organizations to manage AI systems throughout their lifecycle, emphasizing the integration of AIMS with existing organizational processes and ensuring that AI technologies are developed and used responsibly. It establishes a comprehensive framework to address the unique challenges posed by AI, such as ethical considerations, transparency, and the need for continuous improvement in AI practices (ISO, 2023).

As AI advances and becomes integrated into organizations and governments, AI auditing will be essential to ensure these systems are safe, ethical, and beneficial. To achieve this, organizations and governments need fit-for-purpose frameworks to be established, collaboration across the AI system value chain teams, and continuous innovation to harness the power of AI responsibly. The path forward requires diligence, precision, and an unwavering commitment to ethical principles to ensure a future where AI serves humanity. The existing frameworks summarized above provide a good foundation for private sectors and governments to base their fit-for-purpose compliance audit frameworks on to ensure that the AI systems they use are safe and sustainable for their specific functions over their lifespan.

2.10 Summary

Through the review of the relevant literature identified in the themes, it can be concluded that there is a consensus that AI systems need standardization and mechanisms established to facilitate rules, guidance, or control on how they are developed, deployed, implemented, and used (Ayling and Chapman, 2022). There is a lack of uniform and clear regulation and legislation regarding the use of AI in the energy sector (Arévalo and Jurado, 2024), no alignment as to the format this should take and who should be driving the development and facilitation of these mechanisms (Munn, 2023).

The research also supports that AI systems need to have some level of auditing or compliance checks (Roberts *et al.*, 2022) in place to ensure that they are developed and operated as per the prevailing regulations, governance, or legislation. However, there is no consensus on whether this should be done at an organizational, national, or international level. Furthermore, it is undecided whether the AI audit function should be done only during development or throughout the system lifecycle, and it is unclear how the

requirements may change as AI moves from an assistant to an autonomous system (Raji *et al.*, 2022).

The maturing of AI systems has driven a flurry of research and development of principles, regulations, and policies and has encouraged governments, academia, and private organizations to focus on developing AI compliance audit methods. Several mature regulations, policies, white papers and frameworks outline compliance mechanisms for AI systems, but the majority are focused on having a one-size-fits-all approach to auditing these systems. There is no clear convergence between the different governments, academia, and organizations on what and how AI systems should be audited to ensure that the systems operate safely, morally and that they are human-centric. Many of the existing frameworks focus on auditing only partial components within the AI system, such as the data, the learning process, or the algorithms (Koshiyama *et al.*, 2024).

Lastly, governments, academia, and organizations cannot agree on whether or how humans should play a role in ensuring that AI systems are developed, deployed, and operated safely (Priya *et al.*, n.d.). There is a discourse on what level of human interaction or oversight (Kazim *et al.*, 2021) is required within AI systems as they become more autonomous (Niet *et al.*, 2021), more complex and operate as a Blackbox, where no clear understanding of how the system makes decisions is available. To make it worse, as AI becomes autonomous, it can be capable of re-writing its code, which will make it very difficult to oversee unless there are structured guardrails established to guide its operation and to empower its human collaborators to understand its decisions and be able to make an informed decision on the outcome.

CHAPTER III:

METHODOLOGY

3.1 Overview of the Research Problem

Humankind has been intrigued with AI for decades (Fenwick and Molnar, 2022) as it is seen as a tool to improve efficiencies, further organizational interests and improve societal well-being. Studies have reviewed the key drivers and barriers to implementing and accepting AI into organizations and society (Cubric, 2020); the largest area of misalignment among leaders is whether they would accept AI transitioning from being an assistant to humans to being trusted as an independent, autonomous decision maker (Johansson and Björkman, 2018). Studies conclude that the future of leadership and the role of leaders will need to change during the integration and implementation phase and again during the deployment and usage phase to accommodate the new organizational and staff needs (Jorzik *et al.*, 2023). There is a discourse on the level of human interaction or oversight required within AI systems as they become autonomous (Niet *et al.*, 2021) and make life changing decisions.

Researchers and industry experts have a perceived consensus that mechanisms must be established to facilitate rules, guidance, or control on how AI systems are developed, deployed, implemented, and used. The preliminary research indicates that AI systems require some level of auditing or compliance checks in place to ensure that they are developed and operated as per the prevailing regulations, governance, or legislation. However, there is no clear consensus on whether this should be done at an organizational, national, or international level. Furthermore, it is undecided whether the AI audit function should be done only during the development phase or throughout the system lifecycle, and it is uncertain how the requirements should change as AI moves from an assistant to a fully autonomous system.

The key problem being focused on here is framing a structured risk-based approach to the protocols, governing rules, laws, policies, and oversight required to introduce different levels and complexity of AI within the critical infrastructure environment, such as the electricity sector.

3.2 Operationalization of Theoretical Constructs

This study's theoretical framework focuses on several critical areas related to the safe, reliable, and sustainable development, implementation, and usage of AI in the critical infrastructure sector. The key theories studied here have been explored by researchers in many spheres, but no clear construct or direction has been agreed upon. The key theories being studied to guide this research are Human oversight in AI, AI regulation development and operationalization accountability, AI compliance metrics, and AI lifecycle oversight. The research followed a survey methodology to gather quantitative expert opinions from regulatory bodies, AI experts, and electricity organizational leaders via a structured questionnaire on the current state of AI governance and compliance within the sector. These theories and expert advice will guide the structuring of a practical compliance oversight framework for AI within the critical infrastructure sector.

Human Oversight in AI – The big questions in this construct are whether the public will trust decisions made by AI if humans aren't involved in some way and whether humans can provide meaningful oversight as AI systems become more complex. There are many differing views on what role humans should play in decisions being made by AI systems, if any. The European Union's newly approved AI Act relies on human oversight to assess the quality of AI algorithm outcomes (Walter, 2023), but human oversight is not always reliable. As AI systems and the decisions taken become more complex, and the system begins to make autonomous decisions, the question arises of whether humans can provide

meaningful oversight. However, almost all researchers note that human oversight is pivotal in ensuring AI systems function and are used ethically and morally. The conundrum is how we achieve meaningful human control of AI (Davidovic, 2023).

AI Regulation development and operationalization accountability – Many researchers note that structured regulation or standardization is essential for ensuring safe and sustainable development, implementation, and use of AI (Janaćković *et al.*, 2024). It is further stated that the public sector, private sector, and NGO's are leading the discussion about developing AI regulations and policies, and not governments (Schiff *et al.*, 2020). In the past years, focus from governments, international bodies, civil society, organizations, and academia has been on developing frameworks, guidelines, and other guiding structures. However, the risk management and oversight of their implementation have largely been left to organizations and are uncoordinated (Ayling and Chapman, 2022). The big debate continuing is who should be responsible for the development, implementation, and oversight of these regulation and management structures for AI and whether it should be done on a national or international level.

AI compliance metrics – Governments, international bodies, civil society, organizations, software developers, and academia have proposed numerous principles and criteria that AI systems must comply with to be deemed safe, ethical, moral, and human-centric. However, as long as there are multitudes of differing principles and metrics being proposed and none being adopted as guiding principles, entities will build and operate AI systems in an uncontrolled manner. With the approval of the European Union AI Act, the world is seeing the first step towards creating a structured mechanism to formalize metrics to monitor AI to ensure that it operates safely and sustainably.

AI lifecycle oversight – Researchers, governments, organizations, and society have been focused on establishing regulations, legislation, and principles for how AI systems should be developed, implemented, and operated. Very little attention has been paid to establishing an all-encompassing compliance or assurance structure to ensure systems are designed, built and used per these guiding principles throughout their lifecycle. Some researchers, organizations, governments, and civil society have focused on limited sections of compliance as seen fit for their organizational needs but have not focused on compliance for the AI systems lifecycle (Thomas *et al.*, 2024). There is also discourse on whether compliance auditing should be undertaken as an internal function in organizations and government or if it should be undertaken by third parties (Hartmann *et al.*, 2024; Raji *et al.*, 2022).

3.3 Research Purpose and Questions

The research study aims to improve the understanding of the existing governance, regulatory, and oversight protocols and to identify knowledge gaps for the safe and sustainable implementation of AI into the electricity sector from an organizational, national, and international perspective. More specifically, the following research questions are sought to be addressed:

i. Are there existing governance and compliance protocols governing AI integration in the electricity sector, and how mature are they?
ii. What key considerations must be addressed in a comprehensive regulatory and oversight framework to address ethical and operational considerations at different maturity levels of AI in electricity organizations?
iii. Should the AI framework be standardized across the entire electricity industry?

iv. Does the electricity sector need to define what AI is at different maturity levels to enable proper governance and oversight?

v. What is the specific lifecycle oversight or audit requirements necessary for ensuring compliance with regulations and ethical standards across various AI maturity levels?

vi. What human oversight will be required at different levels of maturity of AI within the electricity sector?

vii. What challenges and benefits are associated with implementing the proposed oversight and compliance framework?

viii. Can the AI lifecycle compliance protocols be integrated into existing governance processes within the electricity sector, such as the quality, or, environmental, health, and safety audit framework?

3.4 Research Design

The proposed research design utilizes a mixed-methods approach, as shown in Figure 4, combining qualitative and quantitative techniques to gather data.

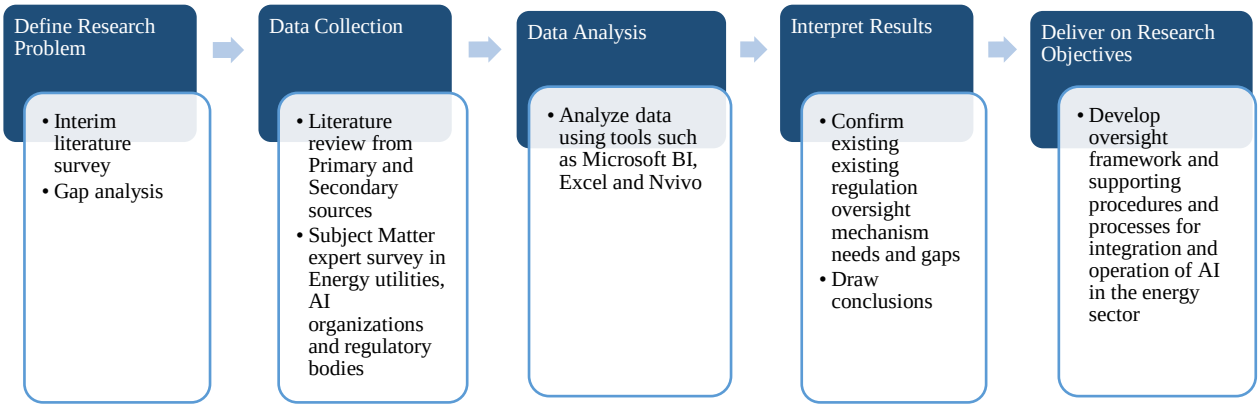


Figure 4. Research Approach

The research method used for the first phase was a qualitative content analysis literature review and gap analysis to identify and evaluate current research on the regulatory, governance, and ethical impact of AI in the electricity sector, as well as current industry practices, governance, regulatory, and oversight protocols or frameworks already established. This research will qualify the existing knowledge base and identify the key gaps in establishing a framework for overseeing the implementation of AI technologies in the electricity sector.

The second phase of the research will follow a survey methodology to gather qualitative and quantitative data from regulatory bodies, AI/information technology experts, and electricity organizational leaders on the current state of AI governance and compliance within the sector.

3.5 Population and Sample

The electricity sector, including regulators and service providers worldwide, is well established and interconnected through global standardization, environmental, and safety standards. The electricity providers' major distinctions are grid type, ownership, generation mix, and grid operability. Globally, there are over eight thousand electricity utility organizations (StatPlan Energy Ltd, 2020). For the sake of the research, the aim was to identify a representative sample of the international electricity sector. As such, the American and Caribbean region was selected, which consists of more than three thousand electricity utilities (Statista Research Department, 2024), made up of public utilities and private utilities of differing sizes that represent public generators, independent power producers, grid operators, market operators, and traders. Considering that these three thousand utilities have a high percentage of regional cooperatives and resellers, this can be broken down into a population base of between three hundred and four hundred unique

types of entities representing the market. Using a ninety percent (90%) confidence level and a smaller than seven and a half percent (7.5%) margin of error, this amounts to a sample size of between ninety and one hundred and ten responsive participants. Considering response rates from this sector being between fifteen and twenty percent, a sample of six hundred utilities was required to meet this response rate.

3.6 Participant Selection

The study focuses on the use of AI in the electricity sector, so participants were purposefully selected who were subject matter experts in the electricity sector, electricity regulation, and AI system development. The participants targeted within the North American and Caribbean Utilities, electricity regulators, and AI system development organizations were information technology professionals, regulators, legislators, and decision-makers within the target sector who influence policy and system adoption. This target group would ensure the participants have the requisite knowledge, influence, and exposure to provide structured and informed responses.

Utility participants were sourced via direct contact with utilities, through LinkedIn specialist groups, or through engagements with the key utility associations and institutes in the selected region, as listed in Table 1 below:

Table 1: List of source Utility associations and institutes

	Utility Association or Institute Name
1	American Public Power Association
2	North American Association of Utility Distributors (NAAUD)
3	Electrical Distributors Association (EDA)
4	Kansas municipal utilities
5	Municipal electric system of Oklahoma
6	North American Energy Markets Association (NAEMA)

7	Electric Power Supply Association (EPSA)
8	America's electric cooperative (NRECA)
9	United States Energy Association (USEA)
10	The American Clean Power Association (APA)
11	American energy engineers
12	Northern California power agency
13	North American power (NaPower)
14	International Energy Agency (IEA)
15	Smart Electric Power Association
16	Electric Power Research Institute (EPRI)
17	International hydropower association
18	Energy information administration (EIA)
19	Electricity Canada
20	International renewable energy agency (IRENA)
21	Edison electric Institute (EEI)
22	Caribbean Electric Utility Services Corporation (Carilec)

Regulator and Legislator participants were sourced via direct contact to regulators and government departments, as well as through engagements with the regulatory associations in the selected region as listed in Table 2.

Table 2: List of source Regulators, Government departments and Associations

	Regulator, Government Department or Association Name
1	OfReg
2	Federal energy regulatory commission (FERC)
3	North American Electric Reliability Corporation (NERC)
4	National Association of Regulatory Utility Commissioners (NARUC)
5	OECD
6	Canada Energy Regulator
7	Department of Energy USA

Finally, AI developer and information or operational technology participants were sourced via direct contact to developers internal to Utilities, as well as through engagements with the AI and software developers active in systems development for Utilities as listed in Table 3.

Table 3: List of AI systems Developers for Utilities

	AI solutions development organizations
1	Google / Alphabet
2	SiteSee
3	Blicker
4	Spark Cognition
5	mPrest
6	Bidgley
7	Autogrid
8	C3.AI
9	Uplight
10	ABB (Seven sense)
11	Microsoft Corporation
12	IBM Corporation
13	Amazon Web Services
14	Accenture PLC - UK
15	Oracle Corporation
16	Intel Corporation
17	Huawei Technology
18	SAP SE
19	General Electric Company
20	CISCO Systems
21	Rockwell Automation
22	HCC Technologies
23	Wipro Limited
24	Utilismart

3.7 Instrumentation

A Survey Monkey form was created for participants to take the survey and to collect all the data in a single place. The form was broken down into five main sections, beginning with the participant's personal details, and the other four sections focused on the key areas for this research. The survey was a combination of multiple choice and multiple response questions to ascertain a baseline of the participant's knowledge of the subject and for them to provide their professional inputs on key topics to guide the ultimate topic for this research.

3.8 Data Collection Procedures

The research was structured into two distinct phases. The first phase gathers research data through a qualitative and quantitative content analysis literature review and gap analysis for the identification and evaluation of current research on the regulatory, governance, and ethical impact of AI in the electricity sector, as well as current industry practices, governance, regulatory and oversight protocols and frameworks already established. This research qualified the existing knowledge base and identified the critical gaps in establishing a framework overseeing the implementation of AI technologies in the electricity sector.

The second phase of the research focused on gathering primary data via a structured questionnaire to collect quantitative data from regulatory bodies, AI experts, and electricity organizational leaders on the current state of AI governance and compliance within the sector.

3.9 Data Analysis

Once the survey questionnaire was closed, all participants' personal details supplied were checked to verify that they were in the target market. The data was checked for duplicates, errors, and other inconsistencies, and then it was combined into a single dataset and collated for analysis.

The survey questionnaire was structured to gather expert-informed opinions and proposals on critical factors in ensuring the implementation of sustainable AI in the electricity sector. This meant that the datasets were structured statistical data to be analysed and did not need complicated analysis software tools to process.

For this analysis, NVivo and Excel 365 were used as the key analysis tools to identify trends in the gathered data and visualize the patterns. The output data analysis was used to enforce the hypothesis for this research and used to guide the structuring of the proposed oversight framework for AI systems in the electricity sector.

3.9 Research Design Limitations

The research focuses on the oversight of sustainable lifecycle integration of AI in the critical infrastructure sector, specifically focusing on the electricity sector in the American and Caribbean regions. There are opportunities for the research to be broadened into the global arena to identify how the requirements for lifecycle compliance oversight differ in different markets. Furthermore, there is an opportunity to consider standardization of oversight frameworks and methodologies globally per sector while adapting the guidelines and regulations in the different regions to account for cultural and regional change requirements. Additional research would also be beneficial in identifying the uniqueness of compliance requirements for the different market sectors in the different regions, focusing on setting up a central system that can be calibrated to the market and

region it is utilized for. This would establish standardization for compliance and oversight of the technology globally.

3.9 Conclusion

The research aimed to identify the current knowledge of AI, its regulation and oversight within the electricity sector, with a specific focus on the electricity sector. Furthermore, it focused on identifying the state of governance, regulation, and oversight within the electricity sector for AI.

The research questionnaire was issued to more than six hundred entities constituted of electricity utilities, regulators, information technology providers and AI software developers. One hundred and twenty-six participants accurately completed the questionnaire surpassing the response requirement of between ninety and one hundred and ten. Of these participants, 88% claim to have at least basic knowledge of the development and impact of AI in the electricity sector, while 47% claim to have worked with and have an intimate knowledge of AI technology and systems. What is of interest, if we consider the results from the younger generations, less than forty-five years of age, is that these statistics change to 94% having basic knowledge and more than 62% having intimate knowledge of AI technology and systems.

From a research perspective, the participant's responses provide a balanced view of the area in question and a good platform for this research.

CHAPTER IV:

SURVEY RESULTS

The survey aimed to engage professionals within the electricity utilities, regulators, information technology providers, to the electricity sector, and AI software developers to better understand what governance, regulatory, and oversight protocols exist. Furthermore, it was planned to identify knowledge gaps for the safe and sustainable implementation of AI from an organizational, national, and international perspective to guide the outcomes of this research project. The survey scope was restricted to North American and Caribbean Utilities as a representative sample of the international sector, with the individual participant focus being on information technology professionals, AI system developers, regulators, legislators, and electricity sector specialists and decision-makers. Statistical models indicated that between ninety and one hundred and ten participants were required to respond to the detailed questionnaire for the research to be accurate and meaningful. To achieve this penetration, the research questionnaire was issued to more than six hundred participants, and an astounding twenty-one percent response rate was received, with one hundred and twenty-six completed questionnaires.

Of the one hundred and twenty-six participants, there was a fair dispersion of participants throughout the four key target groups, as shown in Figure 5. A sixty-seven percent majority represented the electricity sector, or Information Technology providers to the sector.

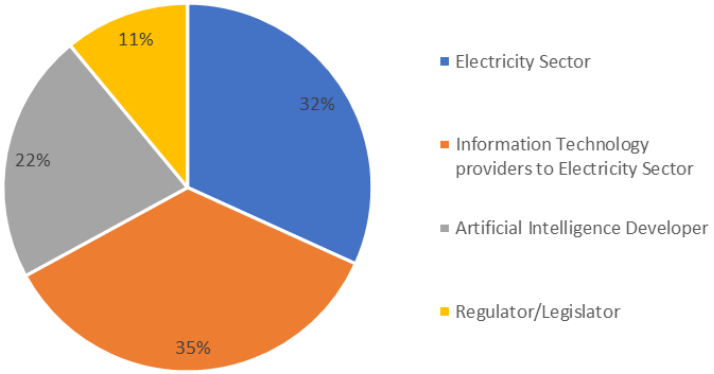


Figure 5: Participants per target grouping

Considering the technology fluency of the younger generations, it is promising to note that more than 70% of the respondents were under forty-four years of age, which provides confidence in the viability of the responses. The total age dispersion of the participants is provided in Figure 6 below.

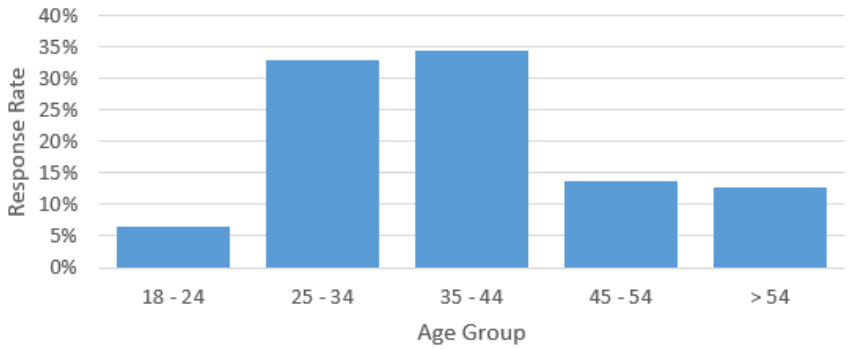


Figure 6: Age dispersal of participants

One of the questions raised in the questionnaire was designed to baseline the participants level of knowledge and personal experience with AI, especially in the electricity sector. As a supporting argument of technology fluency in the younger generations, ninety-four percent of the participants, aged between eighteen and forty-four

years, professed to have basic knowledge of AI or have worked with AI before, whereas, in the age group of forty-five and older, only fifty-two percent had basic knowledge or have worked with AI. Of the ninety-four percent of participants with basic knowledge, sixty-two percent profess to be familiar with AI in the electricity sector or have first-hand experience working with AI in the sector. The full summary of the AI knowledge base versus participant age group is shown in Figure 7.

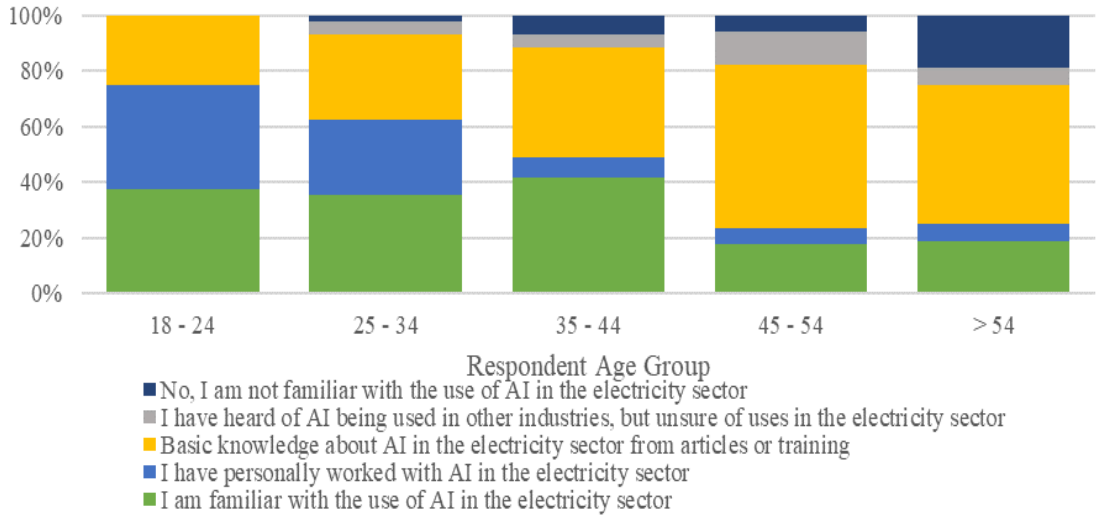


Figure 7: Level of AI knowledge according to participants age grouping

4.1 Research Question One

Are there existing governance and compliance protocols governing AI integration in the electricity sector and how mature are they?

The framing of this section of the research was to gather participants' understanding of whether regulations existed for AI in the electricity sector, how mature they are, and whether they are effective. The first question gauged the participant's knowledge of existing regulations or laws for governing the use of AI in the electricity sector. Forty-four

percent of participants acknowledge specific regulations are active for AI in the electricity sector, with many of those participants being aligned to the information technology specialists or AI developers providing services to the electricity sector. Figure 8 provides a graphical outline of the depth of knowledge of existing AI regulations per target grouping within the electricity sector.

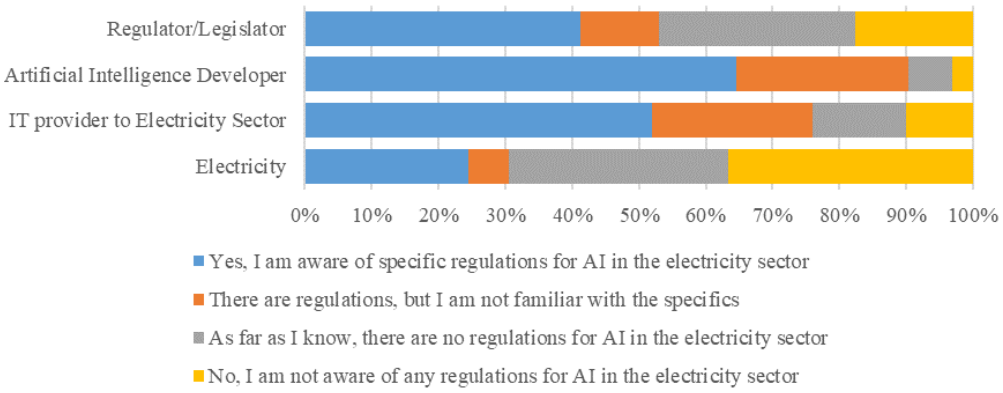


Figure 8: Industry knowledge base on existing AI regulations

A more in-depth review of the data, provided in the spider diagram in Figure 9, indicates that thirty-three percent of the participants recognize that there is some level of regulations implemented between governments and the electricity sector. A further fifty-one percent of the participants stated that ongoing collaborative discussions are underway between the government and the electricity sector to establish regulations. In comparison, seventeen percent indicated that no regulations were being adopted or established and that utilities are implementing AI systems without governance.

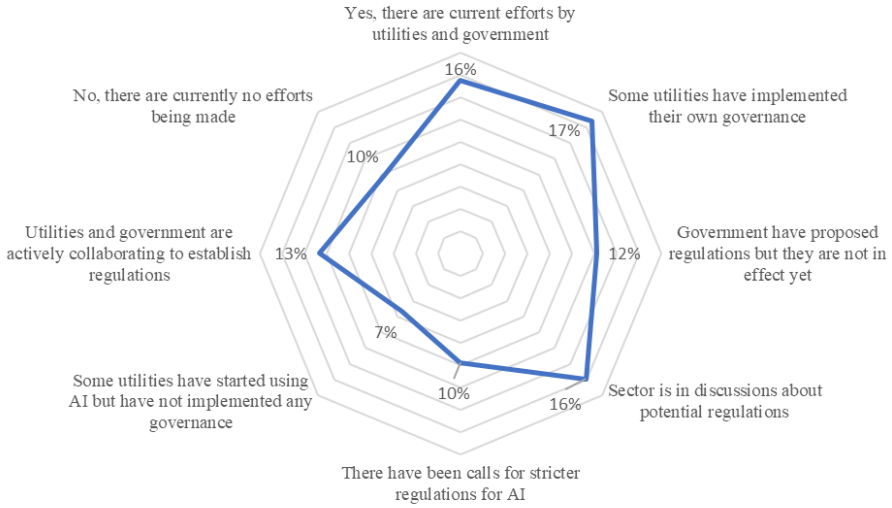


Figure 9: State of AI regulations in the electricity sector

The last request to the participants in this section was to establish whether the existing regulations or oversight protocols that were already enacted adequately addressed the potential risks associated with AI in the electricity sector. From Figure 10, it is observed that thirty-six percent of the participants affirmed that the existing structures were providing adequate mitigation. In comparison, twenty-two percent felt that further research was required to understand whether the regulations were sufficient as they were uncertain. Only nine percent outright stated that the current structures did not de-risk the use of AI systems in the sector. Of interest, the literature review noted that most research shows that when it comes to regulating new technologies, this is normally led by a need in industry while the government is playing catchup to formalize the regulations, which supports the expert opinions above.

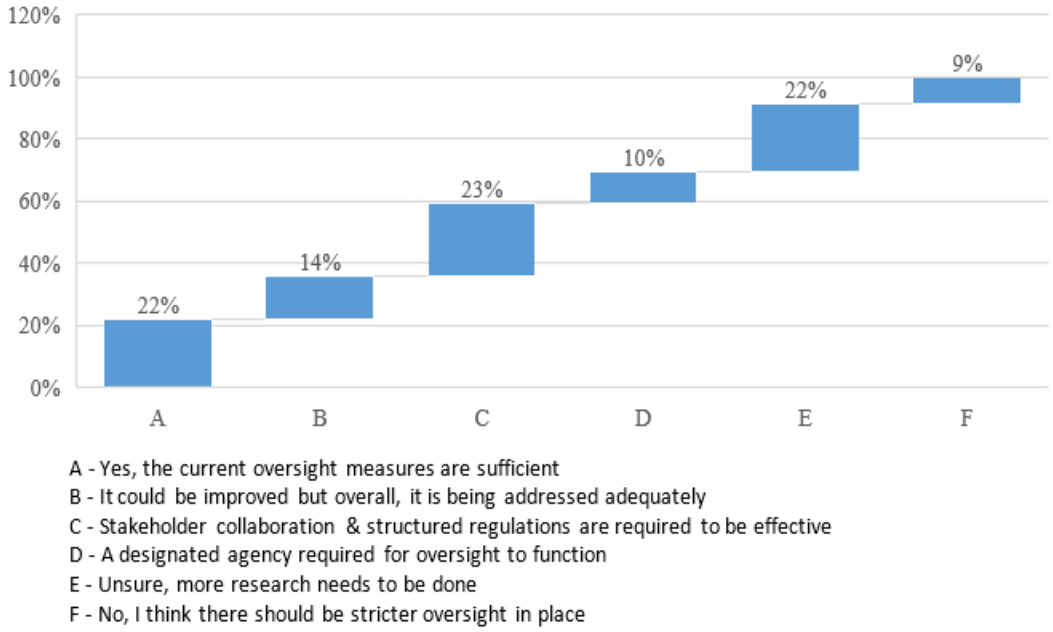


Figure 10: Waterfall diagram of opinions of the sufficiency of existing regulations

In summary, the participants have a fair knowledge of the existing regulations available or in use in the electricity industry. However, they do not agree on whether they are used or if they are sufficient to protect the organization, employees, infrastructure, or the public.

4.2 Research Question Two

What are the key considerations that must be addressed in a comprehensive regulatory and oversight framework to address ethical and operational considerations at different maturity levels of AI in electricity organizations?

The research question aimed to explore the attitude of the different target groups on the need for a comprehensive regulatory and oversight framework to be established for AI in the electricity sector, and to understand where the participants believe the accountability should lie in establishing and maintaining said structure. The last portion of this research question was to get a weighted average of the key considerations that the respondent's mandate as important to include in a framework.

As shown in Figure 11, a resounding consensus of seventy-four percent of the participants agreed that a comprehensive, structured regulatory and oversight framework was required for the ethical and safe implementation of AI in the electricity sector.

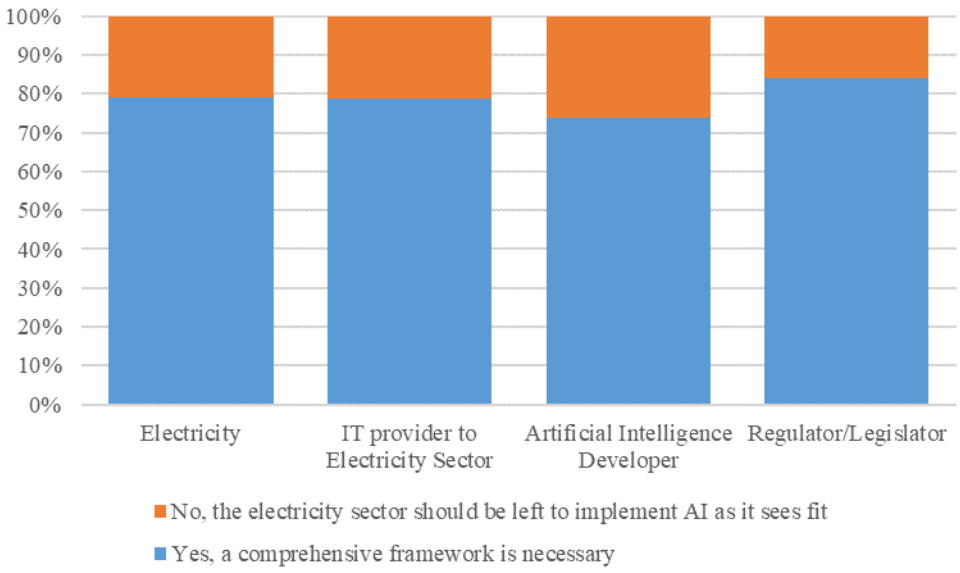


Figure 11: The need for a regulatory and oversight framework

As much as the target groups agree that a comprehensive regulation and oversight framework is required, they do not agree on which organizations or agencies should take accountability for developing and implementing this. From the results, shown graphically in Figure 12, there is no consensus between the target groups on whom should be accountable. However, the collective leaning is towards the government energy ministry, regulatory or standards bodies. Even though there is no clear consensus, most of the target groups recommended that the accountability for the development, implementation, and maintenance of the AI regulatory and compliance framework be done by an independent body rather than by the utility or software developer.

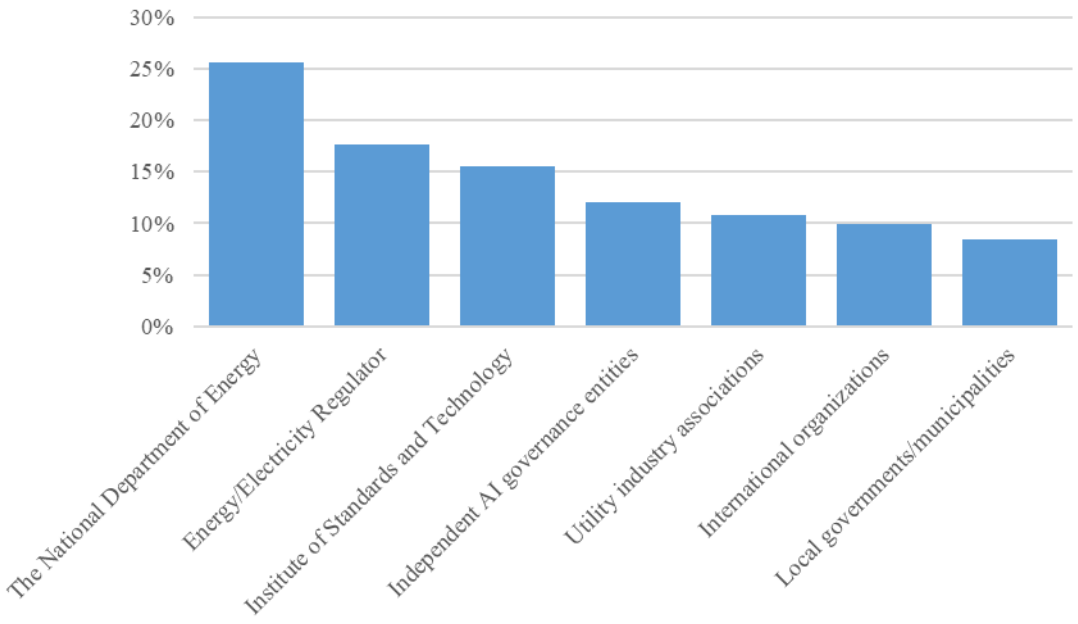


Figure 12: Graphical ranking of AI oversight framework accountability

The last portion of this survey question sought the target group's guidance on what the key considerations were that should be addressed in developing a structured, comprehensive regulatory and oversight framework for AI in the electricity sector.

The top fifty percent of the participants were concerned about transparency, accountability, data protection, ethics, and unintended consequences of decisions made by AI in the electricity sector. Having some insight into this topic makes the fifth-ranked item for consideration a surprise, as shown in Figure 13, in that the target groups request collaboration between all parties to establish a framework, as usually industry would prefer to self-regulate. Nevertheless, the consensus is that a comprehensive regulatory and oversight framework is required for AI in the electricity sector, and it should be collaboratively developed between government and private industry.

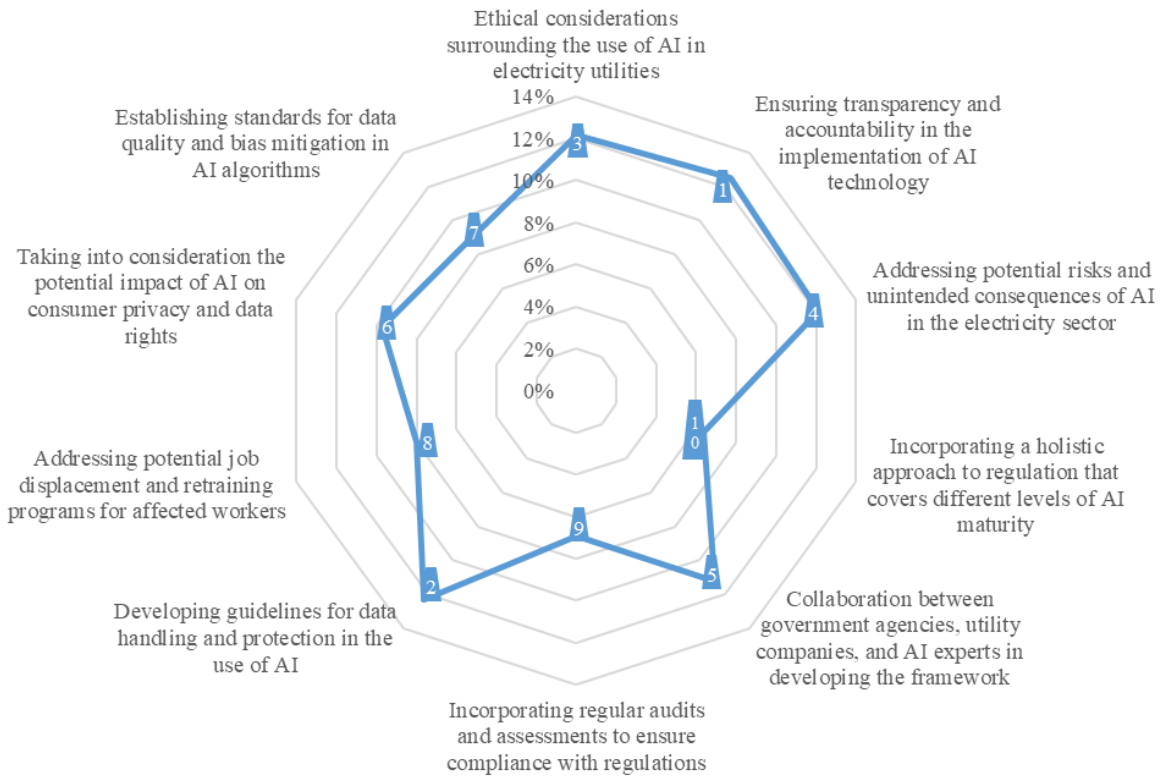


Figure 13: Graphical representation of framework key considerations

4.3 Research Question Three

Should the AI framework be standardized across the entire electricity industry?

Throughout the literature review, researchers referred to the need for global standardized principles, regulations, standards, and compliance structures to ensure that systems are compatible and free of biases. One researcher notes that unlocking the shortcomings of AI ethics principles will require a genuine inclusive global voice to review and learn from past mistakes and establish global languages, terminology, and principles (Hickok, 2021).

The participants were asked to weigh in on the need for standardized regulations to ensure ethical development and use of AI and to promote fair competition among the electricity sector entities using AI. Figure 14 shows that thirty-seven percent of the participants affirm that standardization is crucial, forty-five percent state that this should be undertaken on a balanced approach and be dependent on the AI systems impact on the organization, while only six percent negate the need for standardization.

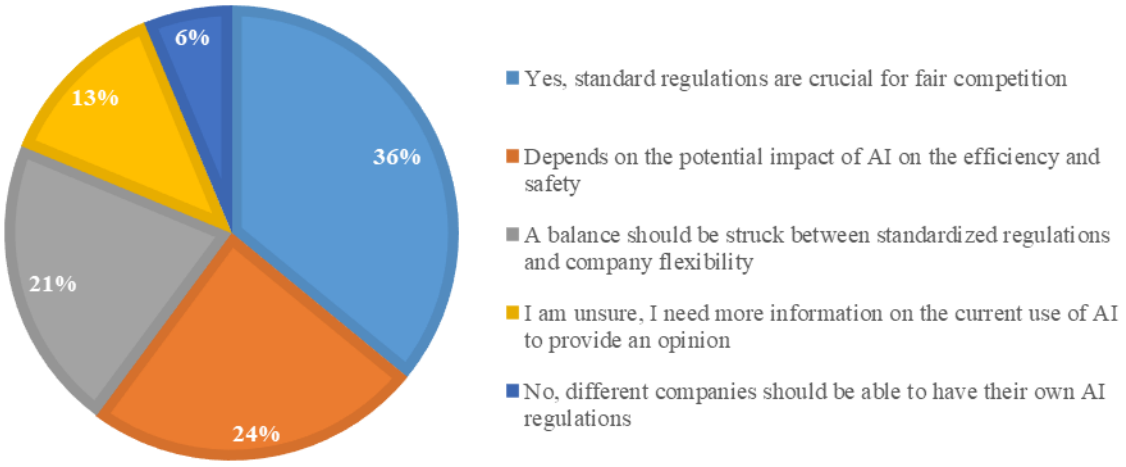


Figure 14: Opinions on standardizing the AI framework in the sector

Reviewing these responses against each of the target groups, as depicted in Figure 15, shows that the AI development organizations are the biggest supporter for having standardization in the frameworks, while the electricity sector is the lowest. This is likely driven by AI development organizations wanting a competitive operating environment in which to develop and sell their services. At the same time, electricity providers are more concerned with progressing innovation without too much red tape to provide cost-effective electricity to customers.

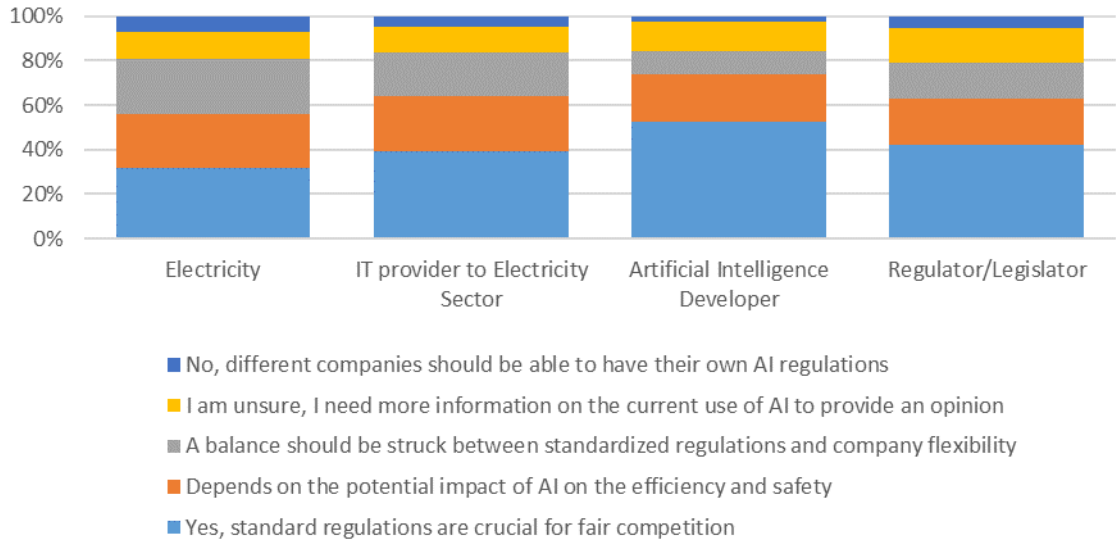


Figure 15: Target group view on AI framework standardization

The natural progression with the participants was to gauge the target group's attitude towards the government leading the process to standardize the regulation and oversight framework development, implementation, and maintenance for AI in the electricity sector. Interestingly, between thirty and thirty-six percent of the participants in the target groups supported the idea that the government should lead the standardization process. What was a surprise was that only seventeen percent of the electricity sector

recommended self-regulation, as shown in Figure 16, while twenty percent of the regulators supported self-regulation in a regulated market.

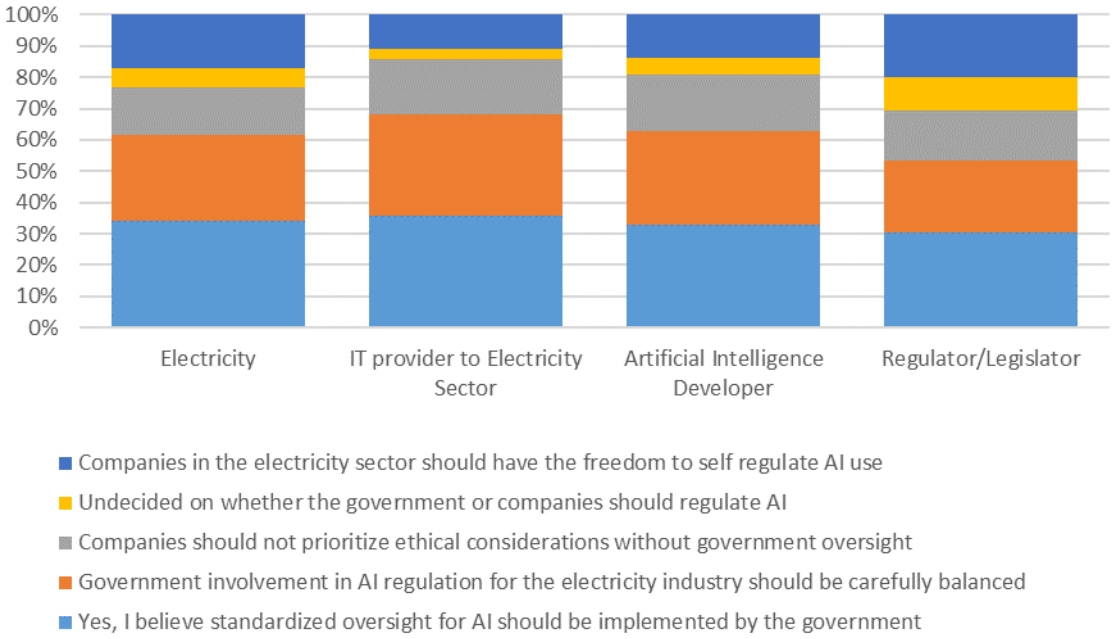


Figure 16: Government involvement in AI framework standardization

In summary, there is a consensus that standardization is required to regulate and oversee AI within the electricity sector. However, as with the previous questions on who should be responsible for developing and maintaining regulation for AI, the target group is divided on whether this should be the responsibility of the government or the organizations developing and operating the systems.

4.4 Research Question Four

Does the electricity sector need to define what AI systems are at different maturity levels to enable proper governance and oversight?

In the European Union's AI Act, AI systems are categorized into different risk levels considering the impact of the system to the public and the organization, the level of autonomy, and several other risk-based factors (Mökander *et al.*, 2022). This raised the question of whether the AI systems being introduced into the electricity sector should similarly be defined from a maturity and autonomy perspective, such that the regulation, governance, and oversight can be positioned to provide the necessary protection. A resounding fifty percent of the participants conferred that a tiered approach for regulations and oversight should be established for different levels of maturity of AI systems to ensure their effectiveness, as shown in Figure 17. A further twenty-five percent support this approach but include considering innovation and the risks when determining the regulation. Only fourteen percent feel the same regulation can be kept immaterial of the AI maturity level.

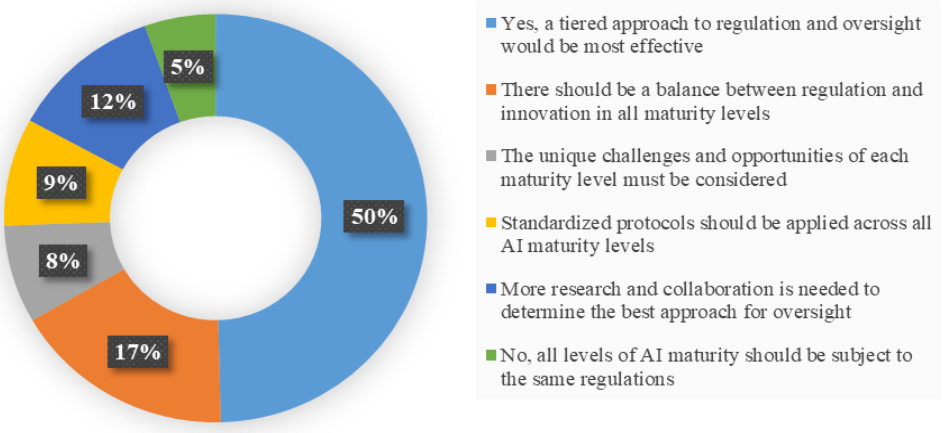


Figure 17: Survey of differing regulations for maturing levels of AI

Before delving into ranking the depth of regulation and oversight, the participants were requested to provide a view on the potential consequences of the organization's heavy reliance on AI systems as they become more autonomous. In the graphical representation of the results, shown in Figure 18, it is noted that eighty percent of the participants focused on the negative consequences of relying heavily on autonomous AI. This raises the question of whether the responses are driven by people not knowing what to expect, is it because there is no structured approach to how AI systems will be monitored and managed to be safe, or is it the myriads of research documents and articles that are providing conflicting views of what AI will and won't do?

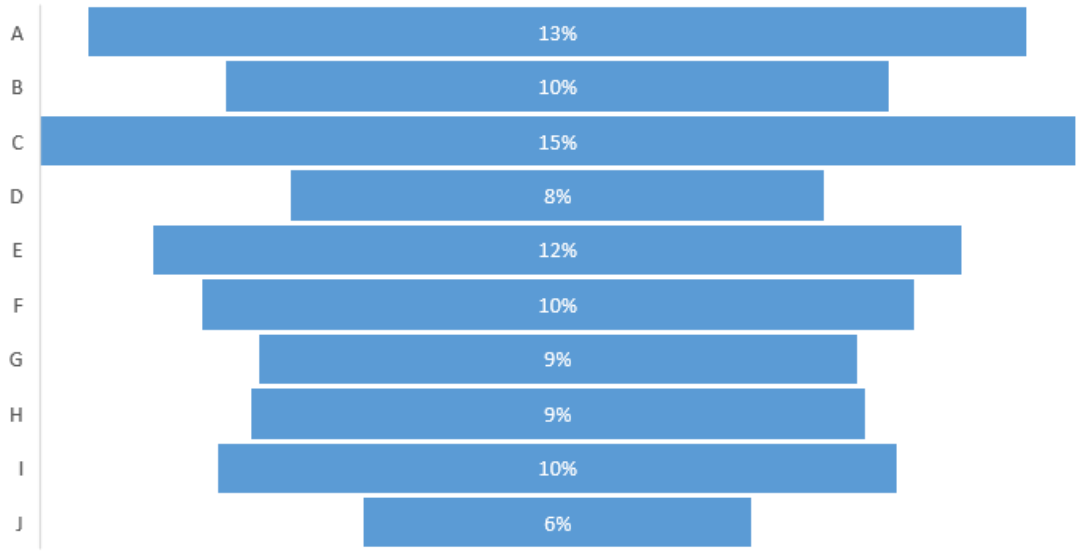


Figure 18: Potential Consequences on relying of autonomous AI

After identifying the participants views on the consequences of autonomous AI, it was imperative to understand the respondent's assessment on what level of autonomous AI required the most rigorous regulation and oversight structure, if any. The comparative summary provided in Figure 19, clearly indicates that more stringent regulation and oversight protocols are required as AI becomes more autonomous. Surprisingly, twenty-five percent of the participants felt that having standardized regulations would be sufficient for all levels of autonomy and that they did not need to be more stringent as AI became more autonomous.

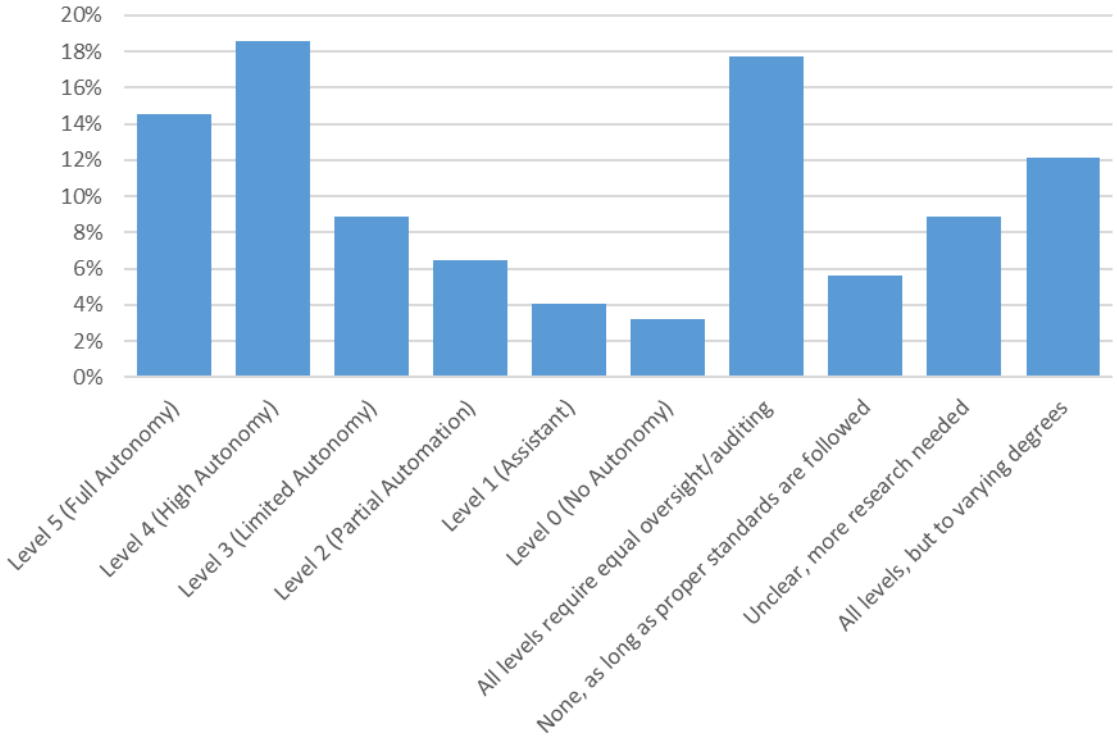


Figure 19: Level of oversight vs level of AI autonomy

4.5 Research Question Five

What is the specific lifecycle oversight or audit requirements necessary for ensuring compliance with regulations and ethical standards across various AI maturity levels?

After gathering information about regulations and governance for AI systems in the electricity sector, the next step was to focus on evaluating the need for compliance auditing or oversight against those regulations and governance structures to ensure that AI systems are designed, built, and used safely. As graphically shown in Figure 20, only thirty-two percent of the participants acknowledge that a formal oversight and audit process is required, while fifty-nine percent state that having a formal process may be beneficial but do not see it as imperative. However, it should depend on the level of autonomy and must balance innovation and accountability at the decision-making level.

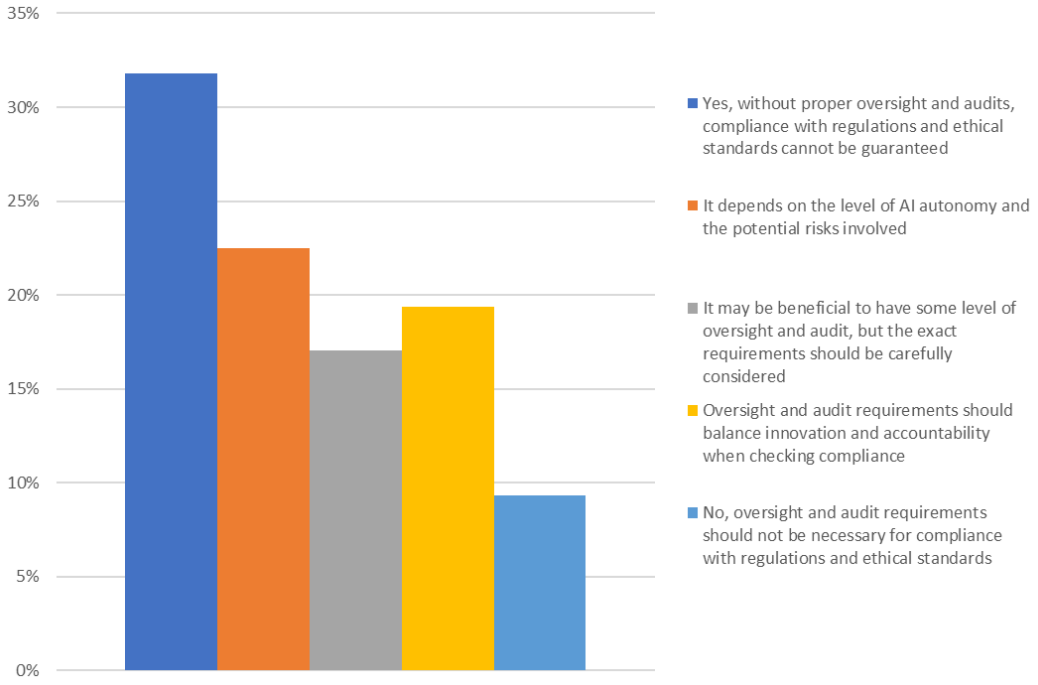


Figure 20: Preference for establishing AI compliance auditing

Reviewing this from a target group perspective, as shown in Figure 21, it is observed that the regulator team is split between having no compliance structure in place and establishing a structure dependent on the AI system's autonomy levels, while ensuring that it is fit-for-purpose. The electricity sector, information technology, and AI providers favour having a compliance oversight structure that considers the level of autonomy while balancing autonomy and accountability.

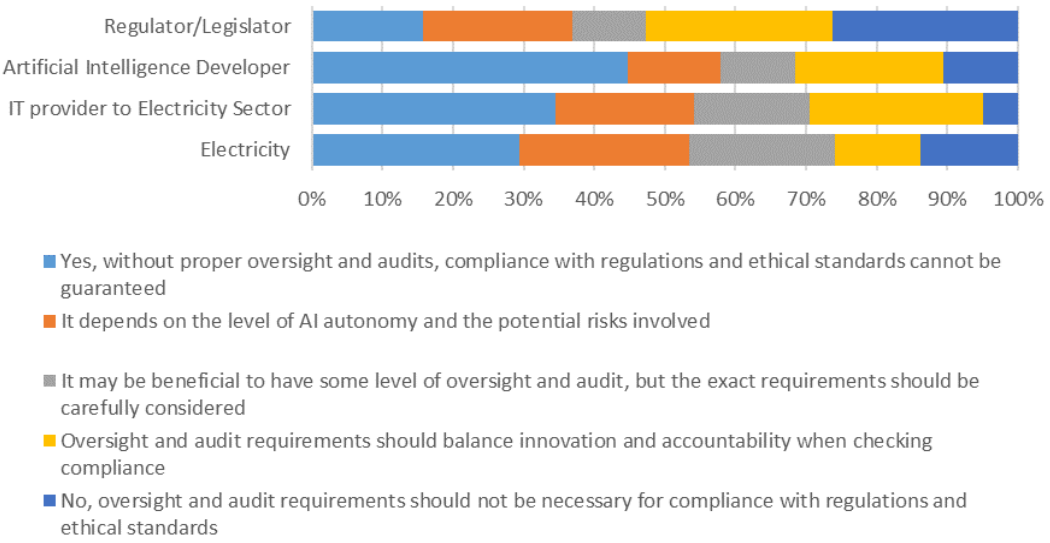


Figure 21: Target group ranking for establishing AI compliance auditing

With most of the participants supporting that some type of compliance auditing structure be established, the next area for investigation was what measures should be included in this compliance structure to provide safe operation of AI systems throughout its lifecycle. From the participant responses, summarized in Figure 22, the top measures to consider are 1) Clear policies and guidelines, 2) Regular audits and assessments, 3) Training and education, 4) Data privacy, and 5) Transparency and accountability. The balance of the measures mentioned are essential but are actual spinoffs of implementing these first key items correctly.

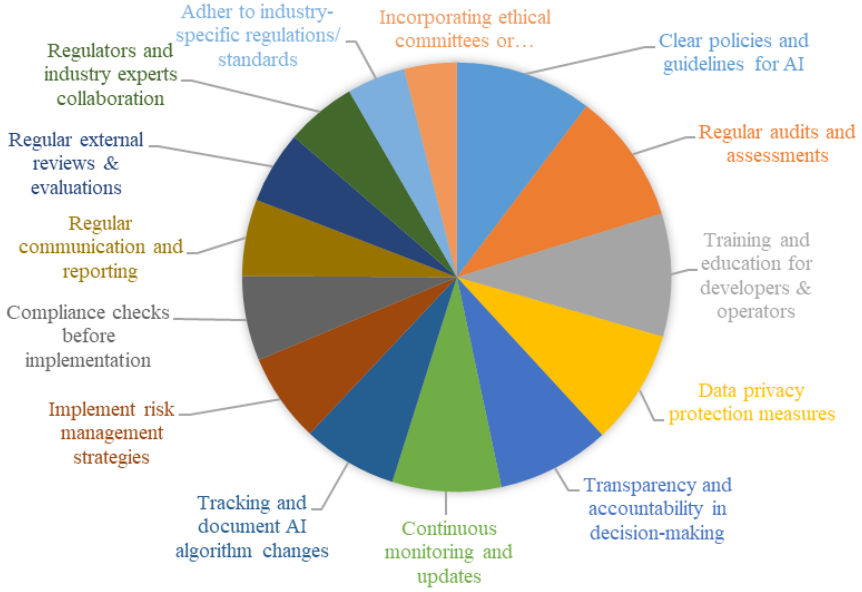


Figure 22: Key measures for AI compliance

Figure 23 shows that the target groups align closely on the top 5 key measures but differ on prioritizing the balance.

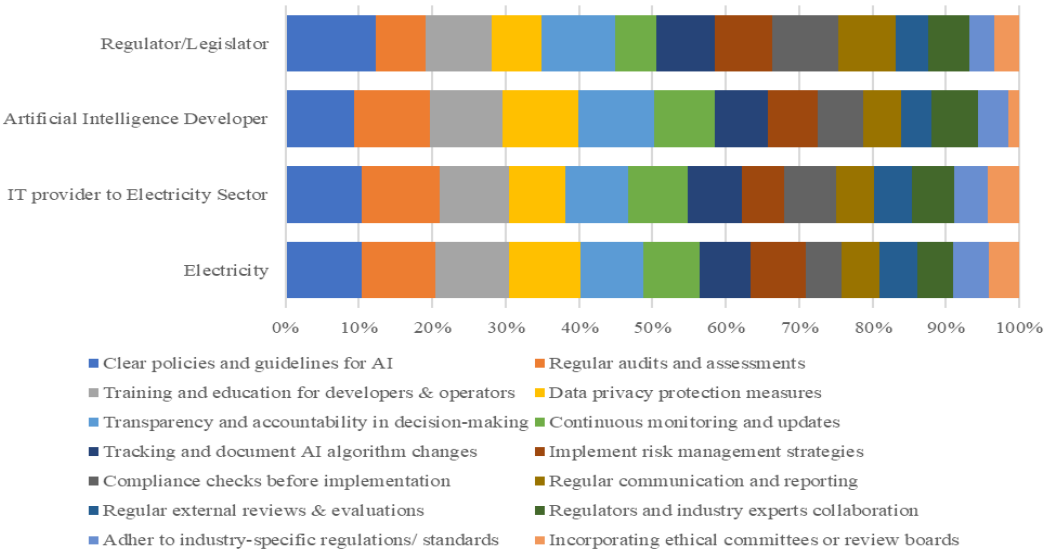


Figure 23: Target group rating of key measures for AI compliance

4.6 Research Question Six

What human oversight will be required at different levels of maturity of AI within the electricity sector?

As shared in the literature review on human oversight, an ongoing debate exists between researchers, academia, AI developers, and system users on the role humans should play in the decision-making process with regard to AI outputs and how they are used (Green and Kak, 2021). The other healthy discussion is at what stage of AI design, development, and deployment should humans be involved in the process to provide oversight of the systems, data, and algorithms and at what level should the oversight be pinioned. The research shows there is no clear consensus on the level of human involvement and oversight at the different levels of design, development, training, and deployment of AI. However, the debates indicate that by not having a structured approach for collaboration between humans and AI system's we cannot ensure that the AI system is implemented safely and functions within a set of structured guardrails (Nothwang *et al.*, 2016).

To close the gap between the literature review and industry knowledge, several probing questions were posed to the participants to understand the current level of human oversight and the need. When participants were requested to rate the current level of human oversight in implementing AI systems in the electricity sector, as shown in Figure 24, sixteen percent of the participants acknowledged that there was a high level of human oversight at all maturity levels of AI. In comparison, forty-three percent stated that human oversight is active but not at all maturity levels of AI systems. Only ten percent of the participants claim that there is no human oversight and a further eleven percent claim that there is limited human oversight.

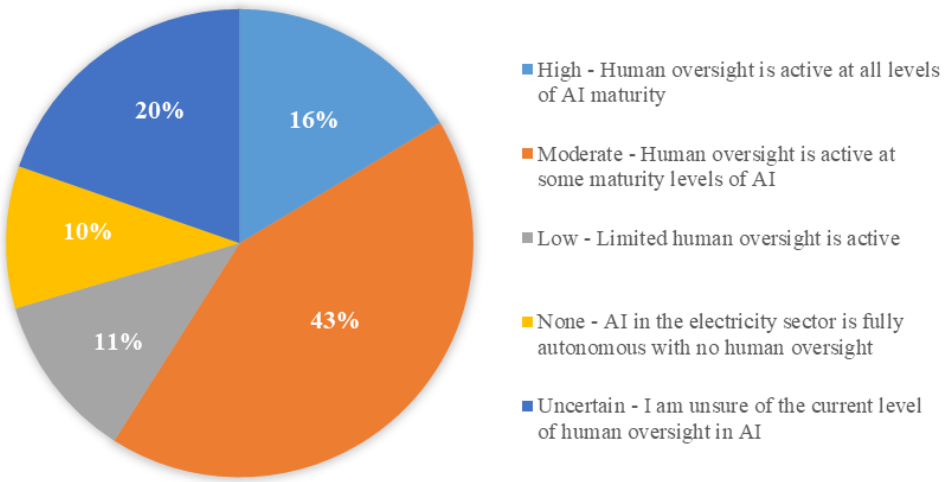


Figure 24: Knowledge of existing human oversight in AI

The next question posed was to ascertain, given the potential risks and benefits of AI in the electricity sector, how much control or inclusion humans should have in AI development, deployment, and operational processes. These results were normalized against the participants' age group to gain an understanding of the different generations' thoughts on how humans should be involved in AI systems governance.

Some key takeaways from the information shared in Figure 25 are that the generation between eighteen to twenty-four and those older than fifty-four do not believe that fully autonomous AI systems should operate without human oversight. Secondly, more than double the number of participants in the age group eighteen to twenty-four than in any other age group propose that a higher level of human oversight is required for critical decisions. Lastly, only participants in the age group of fifty-four and higher believe that human oversight should never be completely removed from the compliance process.

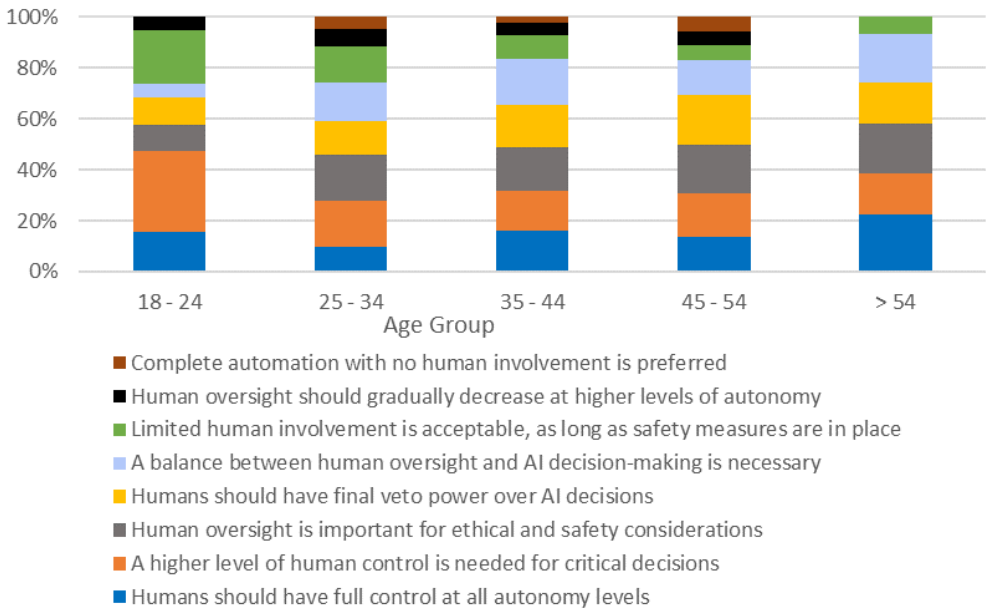


Figure 25: Age-normalized human oversight opinion

Finally, participants were asked to provide insight into the key considerations for determining the appropriate level of human oversight in the electricity sector. Figure 26 provides a graphical depiction of the ranking of aspects to be considered in determining the level of human oversight. The key factors are that the level should be dependent on 1) Type of AI technology being used, 2) The complexity of the task being undertaken, 3) The potential on safety and security, 4) The risk rating of the AI system, and 5) The level of decision-making authority. This leans towards having a risk-based approach to setting the human oversight level.

Figure 27 provides normalized responses according to target group, showing that the groups agree on most of the key considerations being proposed, which provides a good foundation for this research.

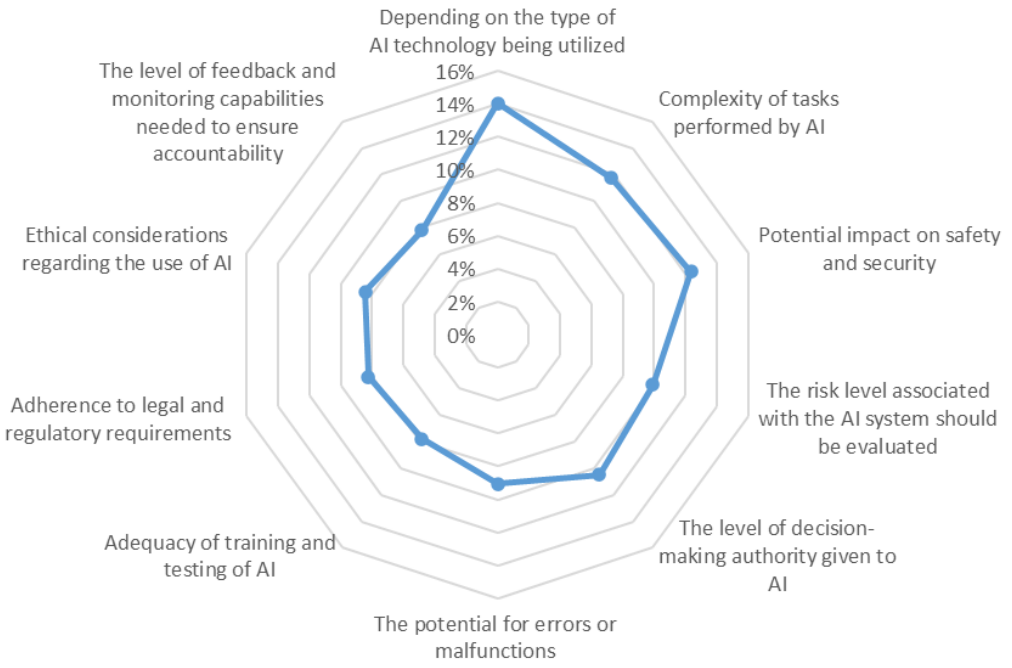


Figure 26: Key considerations for appropriate human oversight levels

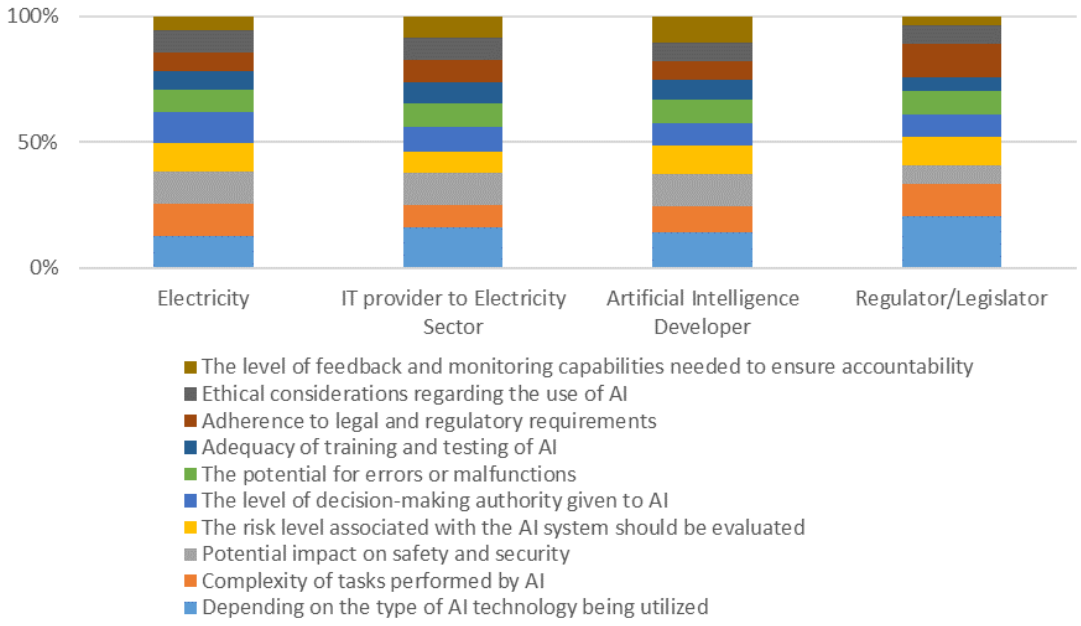


Figure 27: Key human oversight considerations per target groups

4.7 Research Question Seven

What challenges and benefits are associated with implementing the proposed oversight and compliance framework?

The aim of implementing an oversight and compliance auditing framework for AI in the electricity sector is to ensure that AI systems can be designed, built, implemented, and maintained safely and sustainably, as prescribed by relevant prevailing standards, regulations, governance, and policies. Many have raised the question of whether the benefits outweigh the challenges or risks of implementing a compliance audit framework for AI considering the lack of mature regulations and standards, no clear agreed definition of AI and no globally accepted guidelines. The other question is whom is capable of auditing AI systems that are designed to make autonomous decisions through multi-layered complex software and technology solutions that are difficult for humans to understand, especially as they become autonomous and can recode themselves to improve their performance and change how they operate with no human intervention.

The first portion of the information gathering was to understand what participants considered the key challenges with implementing an oversight and compliance auditing framework for AI system management. Figure 28 provides a normalized view from the target groups, rating their views of the key challenges in implementing the framework. The highest-ranking challenges raised were that the constantly changing technology, the lack of standardization and guidelines, and incompatibility with existing systems and processes, make it challenging to develop a framework. They also raised the concern of stakeholder resistance to adopting a new framework, limited availability and cost of skilled auditors, and privacy concerns, which were also key detractors to successfully implementing the proposed framework.

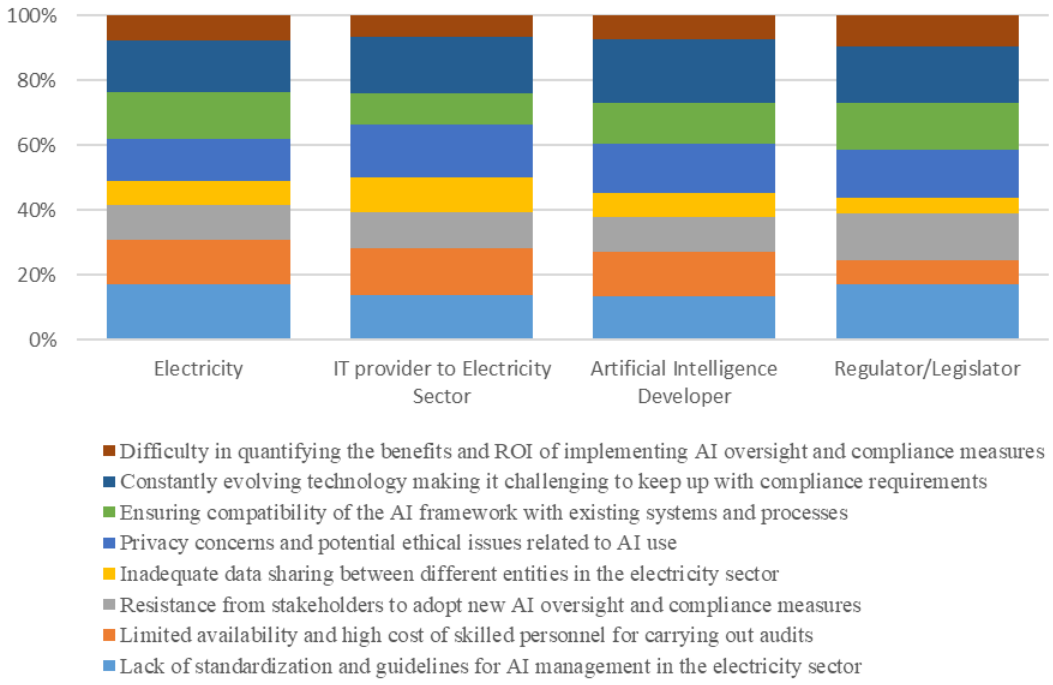


Figure 28: Challenges to implementing an AI oversight and compliance framework

The second question was raised to gauge what participants portrayed as the most important benefits of implementing the oversight and compliance auditing framework for AI system management in the electricity sector. The top-ranked benefits, as ranked in Table 4, are that compliance oversight improves AI system transparency and accountability in decision-making, and it enhances the safety and reliability of the AI systems, thereby improving the safety of the employees, infrastructure, and the public. The compliance oversight process will identify areas where the AI systems can be optimized and made more efficient. Finally, it can be used to identify risks early and mitigate them, thereby building trust in the operability of the system with employees and the public.

Table 4: Key benefits of AI compliance oversight audit

Benefits of AI compliance oversight	Percentile
Increased transparency and accountability in decision-making processes	13%
Enhanced safety and reliability of AI-powered systems in the electricity sector	13%
Mitigation of potential risks and ethical concerns associated with AI use	11%
Improved data governance and protection as AI systems handle sensitive information	11%
Identification of areas for optimization and efficiency improvements through auditing	10%
Cost savings through early detection and prevention of AI failures or errors	9%
Facilitation of regulatory compliance and adherence to industry standards	9%
Promotion of fair and non-discriminatory use of AI in the electricity sector	8%
Strengthening of consumer trust and confidence in AI-powered services	8%
Effective management of potential biases and unintended consequences of AI implementations	7%

The last item the participants were requested to consider was the potential risks and drawbacks of implementing an oversight and compliance framework for AI system management in the electricity sector. Of interest, most of the participants align the risks with the previously identified challenges of implementing an AI compliance audit framework. The participants note that if these risks aren't identified, managed and mitigated before developing, implementing and deploying the AI system, it could lead to a compliance framework that will aggravate the risks rather than minimise them. In Figure 29, the key issues are ranked as 1) high cost of implementation, 2) lack of clear guidelines or standards, 3) data privacy concerns, 4) resistance to changing from current policy, and 5) difficulty hiring qualified auditors that can action the framework.

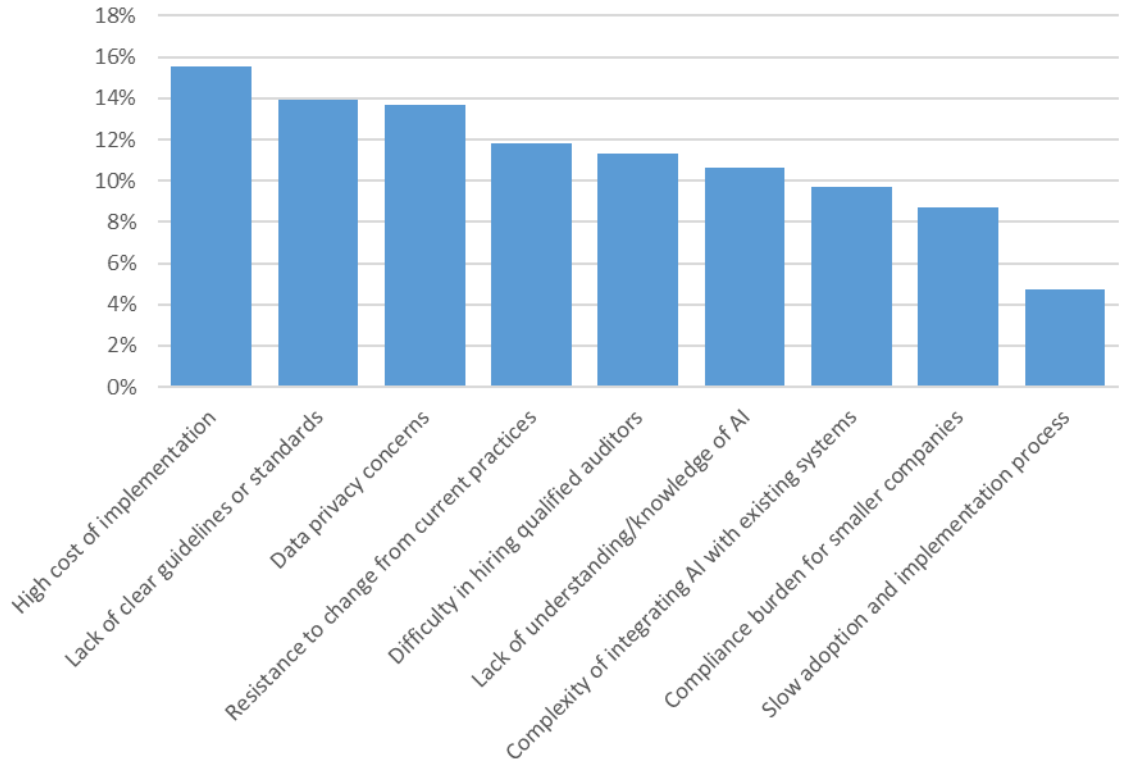


Figure 29: Risks of implementing an AI oversight and compliance Framework

The participants agree that there are benefits from implementing an AI compliance auditing framework for the electricity sector but recommend that an organizational supporting structure be established to make this a success. On review of the responses per target group, there is consensus on the key challenges that need to be addressed to ensure that they do not become a risk to the framework's successful implementation.

One key takeaway is that for this to be adopted and used, the solution needs to be accepted and supported by the employees; this requires targeted training for them to understand how to work and collaborate with the AI systems. It will also require establishing a change management process to ensure that all levels of employees understand and support the changing processes and policies.

4.8 Research Question Eight

Can the AI lifecycle compliance protocols be integrated into existing governance processes within the electricity sector such as the quality, environmental, or health and safety audit framework?

Most critical infrastructure sectors, such as the electricity sector, are heavily regulated to ensure the protection of the infrastructure, to manage quality, to protect the environment, employees, public, and to manage service affordability. One of the concerns in the electricity sector is that if additional regulations and compliance procedures are introduced as stand-alone structures and not aligned with existing policies and procedures, the employees may not support them and will boycott these additional procedures. As part of the research, the proposal was to investigate aligning the AI compliance auditing framework and regulations to existing compliance and auditing procedures in the organization to streamline the process and get buy-in from the organization and employees to the additional compliance needs.

The first question posed to the participants aimed to get their view on whether integrating AI lifecycle oversight and compliance audit protocols into existing governance or audit processes would be beneficial to the electricity sector and whether it would improve the employee buy-in. In Figure 30, it is shown that forty-five percent of the participants support the idea that integrating the AI compliance audit protocols would improve efficiency, accountability, and trust. In comparison, only ten percent vehemently oppose this, believing it will hinder progress and innovation. A further twenty-two percent of the participants perceive the potential of having the AI compliance auditing protocols aligned to existing procedures and processes, but they note that they need to be integrated with relevant protocols and be managed carefully.

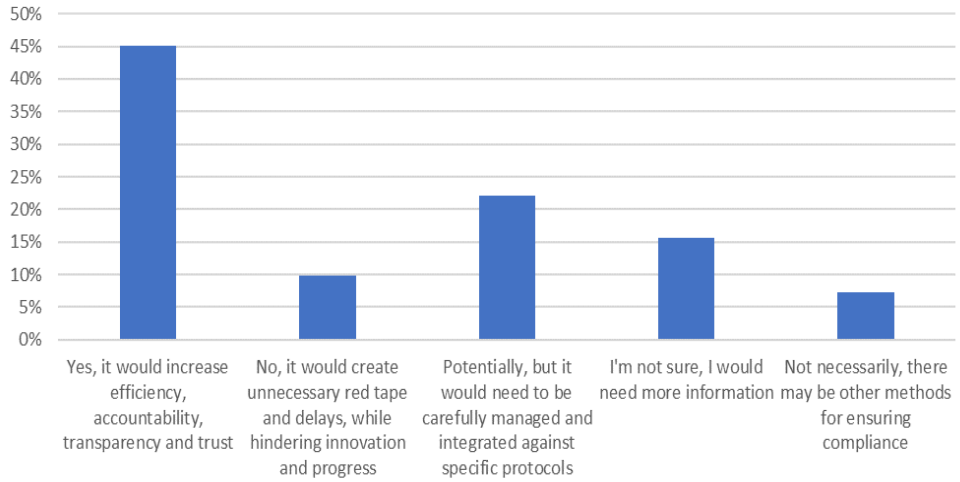


Figure 30: Opinion of integrating AI compliance audit into existing processes

As a benchmark, the responses were segregated into the electricity sector vs other target groups to gauge whether this sentiment was supported by the sector in question. As can be seen in Figure 31, the other sectors ranked the integration of AI compliance audit into existing protocols higher than the electricity sector, but the electricity sector rates this higher if integrated with specific relevant protocols and not just with general governance.

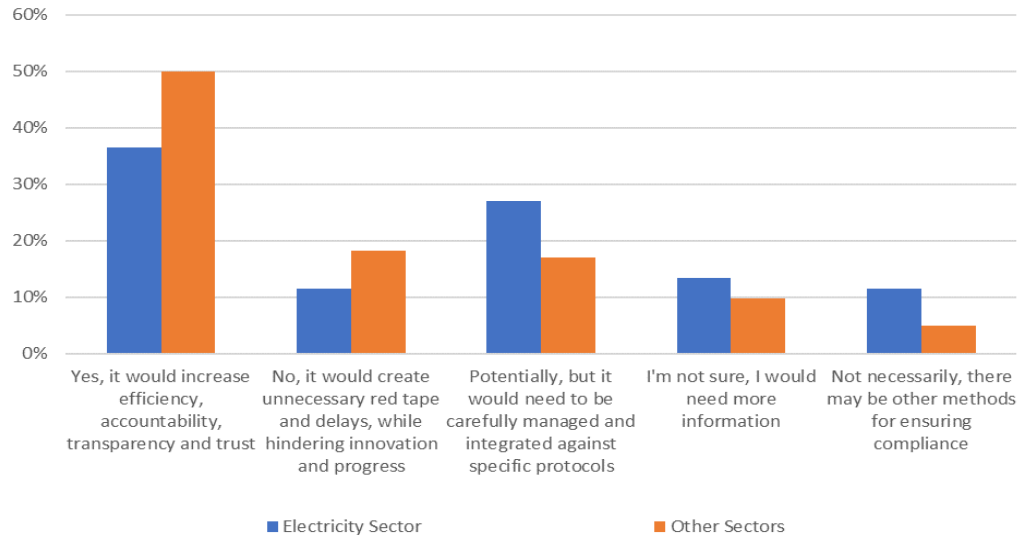


Figure 31: Existing vs new compliance audit process an industry comparison

The last question participants were asked, was to understand whether the industry would support the inclusion of AI lifecycle compliance oversight audit and governance protocols into the existing governance processes.

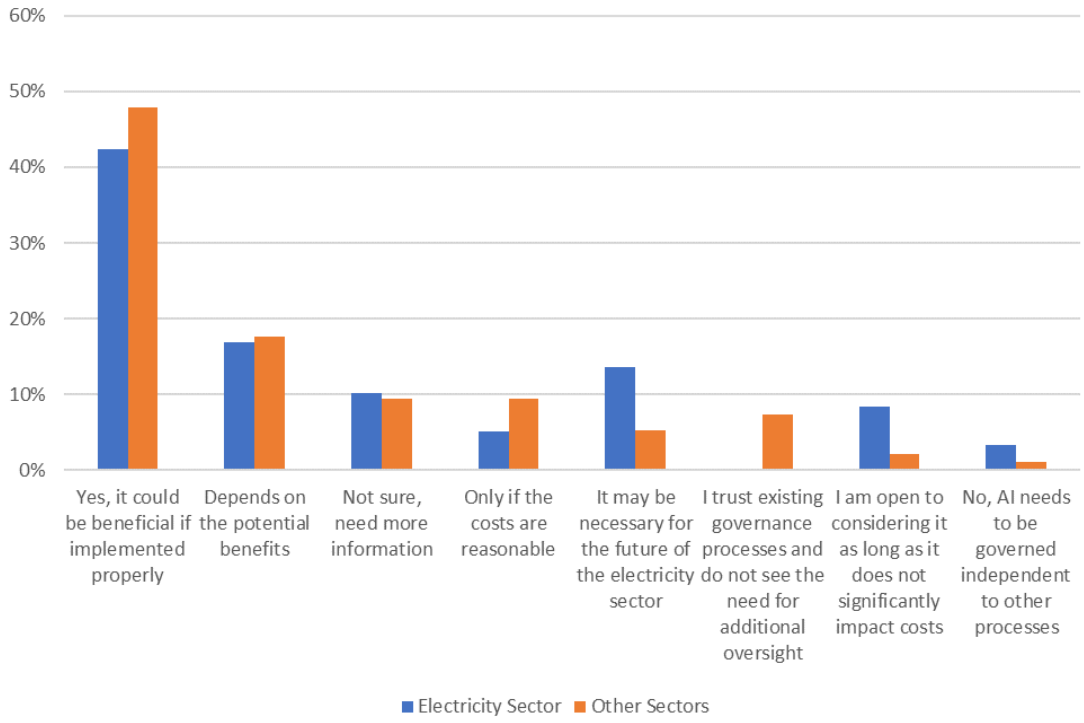


Figure 32: Opinion on benefits of integrated AI compliance

For the best understanding, the results were normalized between the electricity sector and the other target groups that participated, as shown in Figure 32. In general, the sectors agree that there are benefits to including the AI compliance audit process in existing processes, even if it costs more. The key outliers are that the electricity sector does not have trust in the existing governance processes, which is steering them to look for stand-alone oversight processes, and they believe that unless they have a proper combined operational governance and compliance process, they will not be able to build a sustainable electricity sector.

4.9 Survey Conclusion

The participants provided valuable insight and professional opinions on crucial factors that can guide the development of an AI compliance auditing framework that will help ensure that AI systems are safely and sustainably developed and deployed to support the current and future electricity sector. It is thought-provoking to see the level of knowledge of existing AI regulations and governance structures in place to govern AI systems in the electricity industry. However, it is sobering to realize the overarching feeling that they are either not appropriately applied or are insufficient to protect the organization's employees, infrastructure, and the public.

There is a clear need indicated for a comprehensive AI regulatory and oversight compliance framework to be developed for the electricity sector. There is also support from the parties that the most appropriate mechanism to develop, implement, and maintain this would be through a collaborative approach between the government, AI/information technology fraternity and the electricity sector. Where there is no apparent convergence between the parties is whether the government or the electricity sector should be responsible for developing and owning the AI regulatory and oversight compliance framework. However, the parties do agree that whatever the final framework is, it should be standardized across the electricity sector, its support industries, and service providers to ensure that these entities build and operate compatible AI systems.

With the emergence of higher-risk AI and autonomous decision-making AI systems, a tiered approach should be adopted for regulation and oversight compliance for the different maturity levels of AI systems to ensure their effectiveness. What was noted, though, is that participants don't just want more stringent regulations and oversight implemented as AI levels of autonomy increase; they want a balanced approach that considers innovation, accountability, and autonomy in decision-making in setting the tiers.

For an AI oversight compliance structure to provide safe and sustainable operation of AI systems throughout its lifecycle, the participants entreat that, at minimum, the following key measures be considered:

- Establishment and adoption of clear policies and guidelines,
- Undertaking regular audits and assessments,
- Training and education of the employees and public,
- Data privacy management, including cybersecurity, and
- Ensuring that the AI system process and decisions are transparent and that the system and developers are held accountable.

The different age groups had no clear consensus on human oversight or intervention within the AI compliance process. However, some key takeaways from the data collected are that the generations aged eighteen to twenty-four and those older than fifty-four think that fully autonomous AI systems should always operate with human oversight. To go further, the age group between eighteen and twenty-four propose that the level of human oversight should be higher for systems taking critical impacting decisions. This feedback and proposed approach lean towards introducing a risk-based methodology to set the human oversight level in the AI compliance framework.

The data gathered indicates numerous risks and benefits of introducing an AI compliance framework. However, by introducing a structured approach to compliance, the benefits outweigh the risks. The structured approach can improve AI system transparency and accountability in decision-making, identify areas where the AI systems can be optimized and made more efficient, improve data governance and protection by identifying and rectifying areas of concern early, thereby improving the safety of the employee's, infrastructure and the public, while building trust in the system's operability. One key

takeaway is that for this AI compliance framework to be adopted and used, the framework needs to be accepted and supported by the employees; this requires structured training for them to understand how to work and collaborate with the AI systems, which will entail the establishment of a change management process so that all levels of employees understand and support the changing processes and policies.

In general, the sectors agree that including the AI compliance audit process in existing organization audit and compliance processes is beneficial, even if it costs more. However, the sectors stress that unless they have a proper combined operational governance and compliance process aligned and integrated into relevant existing processes and frameworks, they will not be supported or used, which will hamper innovation and the long-term sustainability of the electricity sector.

CHAPTER V:

PROPOSED AI COMPLIANCE FRAMEWORK

The literature review and industry survey have confirmed the proposed research hypothesis, which notes that a facilitated lifecycle compliance audit framework is required for the safe, reliable, and sustainable development, implementation, and usage of AI within the critical infrastructure sector, such as the electricity sector. Figure 33 provides a high-level graphical overview of some areas in which AI is being integrated and used in the electricity sector, indicating the prospective scope of the impact of uncontrolled AI systems in the sector. As detailed in the following subsections, these applications impact the information technology system, operational technology system, the electricity system infrastructure, employees, customers, and the public.

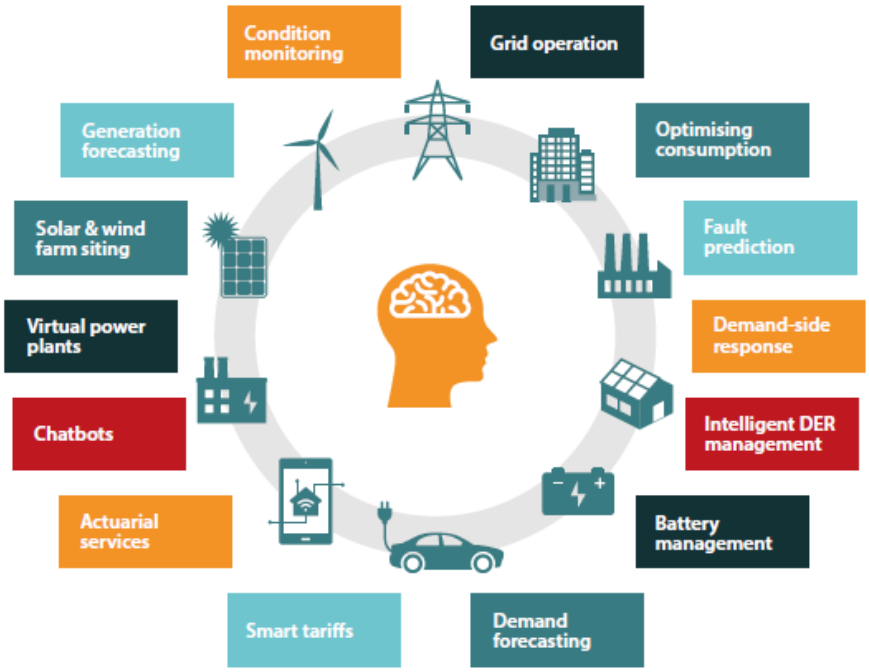


Figure 33: Uses of AI in the energy sector (Morris *et al.*, 2022)

Considering the organizational, operational, and human impact of introducing AI, along with the AI system's specific design, operation, and functionality, there are many factors to consider in defining a compliance audit framework throughout the lifecycle of complex AI solutions, especially when the structures that it is being checked against are not mature or well established. Considering the lack of maturity and ever-changing regulations, standards, guidelines, acts, and even the fast-paced growth and evolution of AI solutions, it is imperative to decide how often the compliance framework needs to be reviewed to ensure that it is meeting the latest requirements for the organization, government, and the public.

Throughout this chapter, the key consideration factors for structuring the proposed framework will be discussed, the structure of the proposed framework will be presented, and the operation of this framework will be explained.

5.1 AI Compliance Audit Framework Areas of Consideration

As noted by previous researchers, academia, governments, and organizations through their publications, blogs, books, and other media, along with the survey participants, there are many factors to consider when developing, adopting, or aligning a compliance audit framework for AI in the electricity sector. To complicate matters, these consideration factors may vary in importance in the compliance audit framework design and operations as the AI systems traverse through their lifecycle (Xia *et al.*, 2024), from cradle to grave. Before delving into the structure of the proposed AI compliance audit framework, it is essential to understand the critical items that need to be considered, monitored, measured, and de-risked to ensure that AI systems are operating safely and sustainably in the electricity sector.

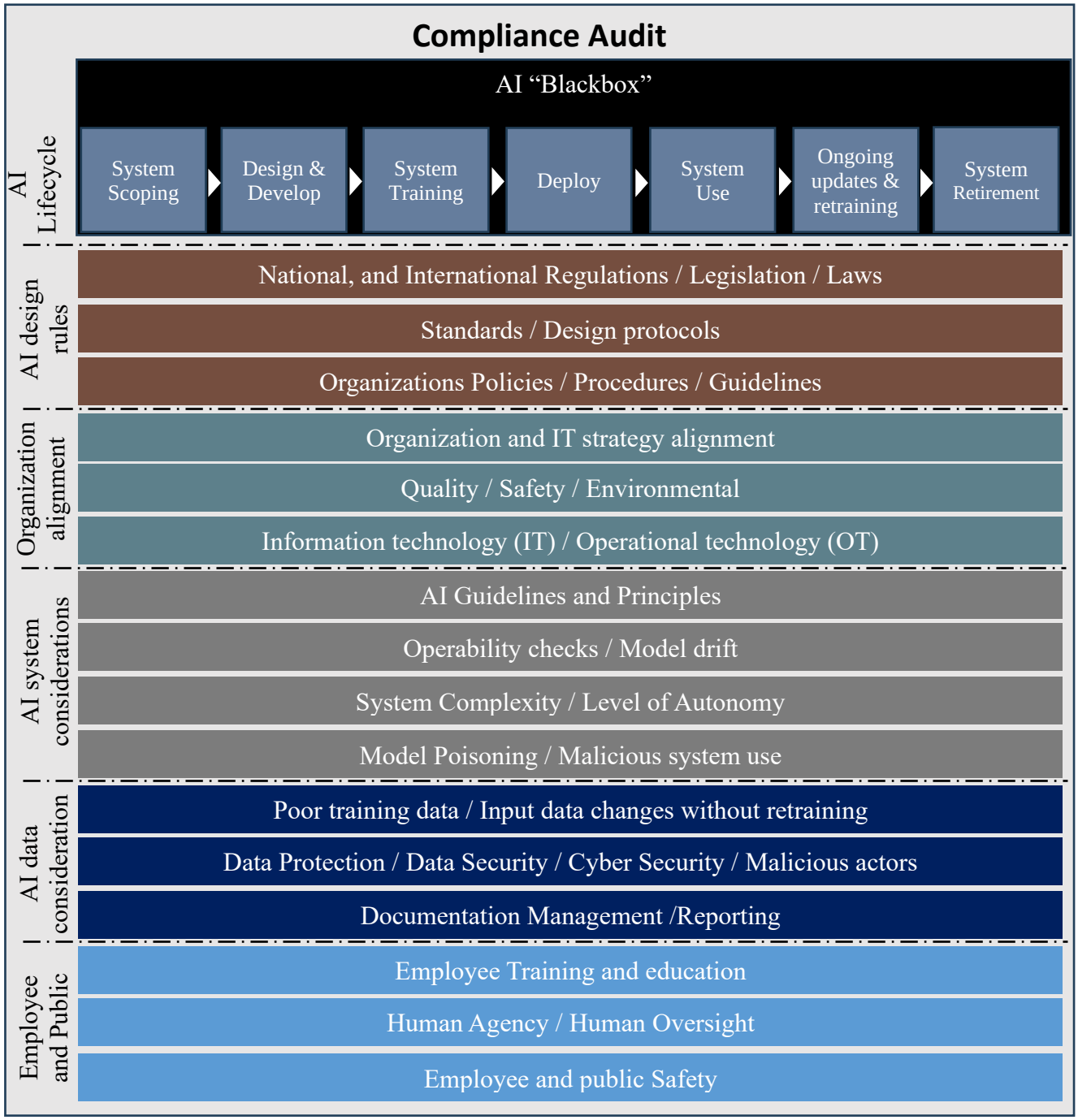


Figure 34: Key items to consider in establishing an AI compliance audit framework

The diagram shown in Figure 34 provides a high-level simplistic view of these critical factors, but in essence, it is more complicated than this illustration as many of these factors are interrelated and can compound their impact. The following subsections explain the details of these key areas of consideration to gain an understanding of the proposed treatment of the factors within the AI compliance audit framework.

AI Lifecycle

As inferred in the name, this contemplates at what stage of its lifecycle the AI system is in, as different measures and oversight is required in the different stages of its lifecycle. Different researchers and academia break this down into multiple stages depending on how finite the system design and training needs to be considered (De Silva and Alahakoon, 2022). For ease of explanation, these have been condensed into seven finite lifecycle stages, which are key milestones that need to be considered when monitoring the AI system and may be repeated as substages:

A. *System Scoping* – The system scoping stage, focuses on documenting what function the AI system will perform, what industry it is focused on, who the key stakeholders will be, what regulations, guidelines, and laws (both national and international, dependent on client geographical deployment area) it needs to abide by, and who the key stakeholders are for system development, training, testing, and deployment. In this stage the foundation is set for the system minimum principles, standards, functions and compliance requirements throughout the systems lifecycle. Considerations should be made as to whether this solution is being built for the information technology or operational technology environment, as this may impact the training environment, data use

and protection, and whether the system will be allowed to be actively linked to the internet.

B. *Design and Develop* – In this stage the scope is converted into a structured AI solution, framed to meet a specific user groups goal's or to achieve a specific function for an organization, which consists of software development and testing, hardware specifications, and system user and maintenance documentation. This includes specifying the AI system training methodology, data requirements, and the collaboration level between humans and the AI system once deployed.

C. *System Training* – This is a crucial step in the AI system process, as it sets the limitations on how the system is operated and whether biases and inaccuracies are embedded in the system operations. This stage includes collecting the suitable datasets, data pre-processing to clean the data to improve quality and relevancy, data annotation to ensure it is machine-readable, choosing the correct model architecture and algorithm, training the system, validation of the training with a new controlled dataset, and finally testing the system with a new dataset that has never been prepared for the system (Javaid, 2024). This stage is where we need to confirm that the system makes decisions per the requisite regulations, policies, and principles, that the decisions are repeatable, and that the decision process and outcome can be traced, deciphered, and understood.

D. *Integration into Organization Systems* –The critical factors here are to ensure that the system is installed correctly, staff are trained, and the system is tested to operate as per the design criteria. Furthermore, this is when the organization needs to ensure that the system is aligned with their strategic intent both from an organizational, informational and operational technology perspective, that it

meets the organization's minimum quality and safety standards, and that the implementation of this AI system does not negatively impact the complete system environment through its use.

E. *System Use* – As the system becomes part of business-as-usual, the organization needs to ensure that its employees have the correct training and certification to operate and understand the system, that the proper level of data protection and cyber security protocols are implemented, the correct people are partnered with the system to do periodic checks to ensure that the system is providing outputs per its design, that there is no biased outputs or decisions and that it is not suffering from model drift. It is also essential for the human overseer or agent to ensure that the operators are using the system as designed, the datasets are not being changed to poison the model, and that the operators are not maliciously feeding it incorrect information to skew the outputs and decisions. One critical risk here is that once the AI system has been operated successfully for a period, operators will become complacent and treat it like any other software package and not monitor it as required to ensure safe operations.

F. *Ongoing Upgrades and Retraining* – During any system upgrades or system calibration retraining, it is important that the system be treated as if it were new to the organization and run through the exact same process to ensure that it is developed and trained to perform the specific function as per all the prior considerations.

G. *System Retirement* – When an AI solution is being removed from the organization at the end of its useful life or even when it is being upgraded to a newer version, it is essential to ensure that no remnants of the prior system remain that could create data protection risks, initiate cyber security breaches,

or create safety risks to infrastructure, employees, or the public. It is also vital to ensure that the old system if not properly removed, does not corrupt the new system when installed and commissioned.

AI Design Rules

This section investigates the rules and structures that guide the safe and sustainable development, deployment, and usage of AI systems to identify their impact throughout the AI system lifecycle and how they should be considered in structuring the proposed AI compliance audit framework:

A. *Regulations, legislation, and laws* – When considering any lifecycle stage of AI systems, it is imperative to understand the regulations, legislation, and laws that govern the development, deployment, and use of the system in the country of development and the country of use. As an example, if we consider an AI system built in China but sold to an organization in the United States of America, it will have different core values and guiding principles due to the different country regulations, cultural identity, and economic systems (Michael *et al.*, 2020) and can introduce inaccuracies or biased decisions.

B. *Standards and design protocols* – As with regulations, legislation, and laws, most countries establish their own standards and have their own guiding standards organizations or committees to ensure that standards are fit-for-purpose for their country, conditions, and communities. This poses an issue when AI systems are sold across global boundaries, and the organization purchasing is in another jurisdiction governed by a mismatched standard. Secondly, almost every AI developer will have different sets of in-house design protocols and standards that they adhere to when scoping, building, and training

an AI system. This could create mismatches and biases between the multiple AI systems implemented in an organization if not developed under the same standards and design protocols. It is preferable that the end-user develops and publishes its own design criteria and proposes standards to adhere to, such that multiple developers follow the same structure to provide compatible AI solutions for their organization.

C. *Organization policies, procedures, and guidelines* – AI developers and end-users, have their own sets of policies, procedures, and guidelines for software system design, integration, and use. With AI systems, these considerations must be integrated into the initial scoping and development and agreed on between all parties, or there will be a misalignment of organizational values, ideals, and goals. For example, data protection, data privacy, and cyber security handling are hard coded into the fabric of the AI package; if the original system design does not align with the end-users' organization policies, procedures, or guidelines, it is almost impossible to rectify this with third-party software solutions. That means that this needs to be designed into the AI system, and the AI system needs to be continuously checked against that design to ensure that there has not been model drift or errors.

Organization Alignment

From an organizational alignment perspective, consideration will be given to aligning AI system compliance with end-users' organization strategies, goals, and core values. These will impact what should be validated, how often it should be checked, and where there is possible alignment to existing compliance protocols when structuring the proposed AI compliance audit framework:

A. *Organizational and information technology strategy alignment* – Every organization has a strategic plan focusing on the goals and objectives for the organization's sustainability and growth in the medium to long term. Most organizations, such as the electricity sector, also have information technology and even operational technology strategic plans outlining what hardware and software need to be implemented to provide the organization with the necessary tools and systems to achieve the strategic goals. AI systems are a significant investment for organizations and can change the work culture, impact how organizations are operated, and displace jobs; to realize appropriate value-add, the organization and information/operational technology strategic plans must align and inform each other (Li *et al.*, 2021). These plans will generally outline the minimum standards, functionality, and compliance requirements of AI systems that the organization will consider, and how they will introduce the system, use and manage them within the organization.

B. *Quality, Safety, and Environment* – Critical infrastructure organizations, like the electricity sector, have prescribed policies, guidelines, and structures around quality, safety, and the environment. In essence, this prescribes the quality of the product they will provide, the safety structures that will be implemented to ensure the safety of infrastructure, employees, and the public, and outline the plans to mitigate environmental impact from the organizations processes. Most electricity sector organizations align and certify to standards, such as ISO 14001, ISO 9001, and ISO 45001, to ensure that the organization follows a structured approach in implementation and monitoring of projects and solutions, and that they are aligned with industry best practices. Introducing AI systems into the organization can positively or negatively impact the company's

performance against these best practices and standards. It is thus essential to ensure that these AI systems and the changes in the company are factored into the audit metrics for these standards and best practices.

C. *Information Technology and Operational Technology* – Historically in the electricity sector, the information and operational technology environments were designed to operate independently and to have limited connectivity between the systems. However, the operational and information technology environments are converging due to the electricity sectors' drive towards smart grids, infrastructure interconnectivity, real-time monitoring, and system automation (Murray *et al.*, 2017). With the rise of cyber-attacks, the electricity sector is not just at risk of attacks on their organizational information platforms but also at risk of these bad actors impacting the safe, continuous supply of electricity. With the introduction of AI systems in either the information or operational technology environments, there is possible exposure of any inherent risks introduced by the AI system in both environments. The question raised is how the risk of AI systems being introduced into the information or operational technology environment be limited to only that environment in an interconnected system. Secondly, are two independent AI compliance audit frameworks needed to ensure AI system operational compliance in the two separate environments?

AI System Considerations

Some factors to consider when structuring the proposed AI compliance audit framework relate to the intrinsic design and use of an AI system:

A. *AI guidelines and principles* – As AI systems mature and broaden their scope, they are becoming "Blackbox" systems, where there is a lack of transparency in the system processing and decision making (Bankins and Formosa, 2023; Dwivedi *et al.*, 2021; Machlev *et al.*, 2022). This lack of understanding of the inner workings of AI systems, along with concerns about societal impact and human rights violations (Kop, 2021) has marked a step change in the focus on developing ethical and moral principles to guide AI development (Gutierrez and Marchant, 2021; Huang *et al.*, 2022; Ryan and Stahl, 2020). Many academia, AI developers, private organizations, and governments have developed guidelines and principles to simplify regulations and legislation into implementable options or bridge the gap to non-existent regulations. In many instances developers have focused on specific principles for system design, such as anti-bias, Ethics, Morality, Accountability, Transparency, Explainable, and Traceability to guide and distinguish their products. This mismatch of standards, principles and guidelines within the AI fraternity makes for systems that could be incompatible with the end-user if the principles do not align with the end-user organization's core principles, values, and objectives.

B. *Operability checks and software drift* – With any information or operational technology solution, it is good practice to undergo normative checks on operability to ensure that the software functions appropriately and accurately. It is common for software to experience errors, code corruption, or crashes due to numerous issues, and these need to be identified to ensure that the information or operational system performs its necessary function safely. With AI systems, this is further exacerbated as they become more complex,

automatically recode themselves, training datasets get skewed, and software or configuration drift occur.

C. *System complexity and level of automation* – As AI systems become more mature and progress towards autonomous decision-making, the systems become more complex and less transparent. This complexity makes defining AI in simple terms challenging, complicating the creation of structured regulations and governance for AI systems (Lyu and Liu, 2021). Furthermore, as the AI system's complexity increases, it becomes autonomous and can automatically generate code adjustments; it is more difficult to track and trace how input data is processed to generate a decision. It is driving many developers and organizations to build specific principles to make systems more explainable and traceable so that compliance can be checked. However, even if they are more explainable and traceable, they need compliance checking more frequently as they become autonomous.

D. *Model poisoning and malicious system use* – When AI systems are trained, they gather large quantities of data from different sources, such as the internet, Internet of Things device's, government databases, datasets from publications and studies, corporate data and specialized machine learning repositories (Hassan, 2024). Model poisoning occurs when malicious or corrupt data is introduced into these training datasets to cause the AI system to produce inaccurate outputs or perform poorly. These attacks can come from external agents that have access to the data or people maliciously damaging the data within the organization developing and training the system. The usual attacks come through mislabelling data, injecting inaccurate data, manipulating data in the dataset, and attacks on the dataset supply chain. They can also come

through backdoors planted into the AI systems. A secondary risk is people utilizing AI systems with malicious intent, which can cause threats to digital security (systems used for hacking), physical security (introduction of non-government controlled automated weapons), and political security (repressing people, running disinformation campaigns, and privacy elimination surveillance) (Brundage *et al.*, 2018). These pose issues for the electricity sector as misuse of AI systems can disrupt the supply of electricity and cause harm to infrastructure, employees, and the public. AI systems datasets must be audited to ensure no malicious intent is added, and when AI systems are retrained or upgraded, no backdoors are introduced. As part of the AI compliance audit process, checks are also imperative for cybersecurity, privacy, and data security to ensure that external forces cannot access the system for malicious purposes.

AI Data Considerations

With AI systems, processing big datasets is the key to training and using the system. The source of the data, accuracy of the data, and security of the data should be considered when structuring the proposed AI compliance audit framework, as they can seriously impact the system accuracy, safety, and performance:

A. *Poor training datasets and calibration for processing new datasets* – As noted in the previous section, inaccurate datasets used to train AI systems create an inaccurate and untrustworthy AI system where the decisions cannot be trusted. The lack of reflective datasets and scenario planning to train AI software is a risk to the sustainability of the AI platform (Laplante and Amaba, 2021). Inaccurate datasets can introduce biases into the AI system and poison the AI

model, which skews the system's decisions, increasing discrimination and decreasing transparency of the system and its decisions (Niet *et al.*, 2021). Similarly, if an AI system is trained using accurate datasets, but the standard input data template or structure is suddenly changed, biased and discriminatory decisions will be evident, as the system is not recalibrated or trained to the new datasets (Konidena *et al.*, 2024). This highlights the need to validate the initial training and data used and ongoing system compliance checks to ensure the AI system still operates as designed. Furthermore, it indicates that the data being processed should be checked for accuracy and compatibility to ensure that there is no misalignment.

B. *Data protection, data security, and cybersecurity* – AI is a complex software solution trained on big data to process big data to make requisite decisions. Much of the data used for training and that being processed includes confidential, personal data from organizations, governments, and individuals, which raises concerns about the security and privacy of the data (Morris *et al.*, 2022). When an AI system is trained, it intrinsically retains elements of the training data in its model, which is transferred with the AI system to the end-user, which can constitute a data protection, privacy, or security risk. The Information Commissioner's Office (ICO) recommends that the developer and system trainer minimize the amount of personal data used for training and that the individual's or organization's data being used should provide informed consent for said data to be processed (ICO, 2020). The second area of concern is cybersecurity; the goal is to prevent unauthorized persons from accessing the system who would potentially steal private information, manipulate the system to fabricate information, generate false decisions, or initiate cyberattacks

(Junklewitz *et al.*, 2023). Lastly, it is important to verify the source that the AI system gathers data from when generating an output, to qualify whether the data gathered was factual from a valid source and not fiction from a storybook or other unverified source.

C. *Documentation management and reporting* – As with all technology and software systems, the AI system development, training, and deployment scope, processes and procedures must be properly documented. This documentation process is important to enable the developer, end-user, and the compliance audit team to track the processes that were followed, what techniques and datasets were used, what standards, regulations, and legislations were used to guide the development and training, what the system was developed to do and how it has performed against calibrated verification. This allows the organization to have transparency regarding the system's operability, thereby empowering the end-user to structure human oversight with a defined outcome and collaborative expectation. Once operational within the end-user organization, the AI systems operation, performance, errors, and inconsistencies must be tracked and reported to guide the compliance audit process for calibration check requirements and to identify risks such as model drift, data incompatibilities, and bias.

Employee and Public considerations

Focusing more on the human psychological impact of AI, considerations need to be made on how humans and AI can coexist, what is required to build trust, and how to empower humans to work alongside and collaborate with AI systems. Several of these items raised in this subsection are an amalgamation of factors raised in prior subsections,

focused on ensuring that humans have the requisite training to collaborate with and understand AI systems, that they can trust decisions made by the AI systems, and that they have comfort that the AI system will not harm them physically, or through privacy breach:

A. *Employee training and education* – Many researchers and organizations have noted that introducing AI systems poses a risk of job losses. In contrast others negate this concept and state that for AI to be successful, the AI system must have a symbiotic relationship with humans (Jarrahi *et al.*, 2023), which would enhance the employees position and not lead to job losses. In essence, the author states that employees need to be reskilled and upskilled to understand the AI systems and understand how to use the AI system to enhance their emotional intelligence to make more informed and accurate decisions. Employees should be trained on AI system operations, how to trace the decision-making process, and how to provide the necessary oversight for the system to ensure that the decision is sound and accurate. Without the proper training or education in AI systems, employees will not be able to discern whether the systems are making safe and accurate decisions, which leads to them being incapable of undertaking the necessary checks for compliance against operational and regulatory requirements. The European Union AI Act emphasizes using human oversight as a key de-risking factor for AI systems; however, this is not possible if the human is not trained or provided the relevant tools to understand the system to make informed decisions (Enqvist, 2023).

B. *Human Agency and Oversight* – Many researchers, organizations, governments and developers have debated the capacity for humans to be involved with AI system operations. The other key debates are at which stage of the AI design, development, and deployment human oversight should be provided for the AI

system, and at what level should the oversight be pinioned? The questions are, if we add humans to oversee AI system decisions or override decisions, do they have the necessary training and understanding to make an informed decision? Is the AI system processes sufficiently transparent for the human overseer to track and verify decisions, or are we just adding a human overseer as a figurehead to provide the organization and public comfort that a person verified the AI decision? Recently, researchers have distinguished that focusing on human oversight alone does not provide for an ethical AI system but rather recommends that the focus should be on human agency, which underscores the broader necessity of AI systems to enhance rather than undermine human autonomy and decision-making (Kopeinik *et al.*, 2023).

C. *Employee and public safety* – As AI systems become more prevalent in the electricity sector, various safety and security risks can be associated with improperly designed, trained, or operated AI systems. When used in the basic information technology system, there is a risk of heightened cybersecurity breaches, data privacy breaches, supplier or vendor data breaches, incorrect customer billing, payment issues, and misinformation issuance, putting customers personal information at risk. More critical, though, is the risk of AI systems on the operational technology system, where there is a risk of disrupting electricity supply to large population groups, leading to public safety risks through the lack of access to medical care, lack of physical security due to security systems not functioning and lack of access to essential services. With the electricity sector's focus on establishing smart grid infrastructure, where everything is interconnected, there are heightened electrical contact risks to employees and the public, where a malfunctioning AI system could erroneously

energize isolated portions of the grid where employees are doing maintenance or upgrades. Malfunctioning AI systems that make inaccurate decisions on system faults or on circuit isolation and interconnects can cause infrastructure damage, which can cause harm to the employees and the public.

The factors to consider are heavily interrelated and codependent; they are ever-changing as AI technology matures, and they are complex, which can be daunting to organizations and individuals when trying to development an AI compliance audit framework. This highlights the reasons that so much research has been undertaken on the topic of ensuring that AI systems operate per their design criteria and why so many organizations have started to focus their attention on minor features of the systems lifecycle to identify key building blocks to determine how to gauge safe AI system operation, how to discern whether it is accurate and repeatable on an ongoing basis.

The lack of maturity and ever-changing regulations, standards, guidelines, and acts, along with the fast-paced growth and evolution of AI solutions, the system risk profile, and the safety benchmark of AI systems, makes it intricate to establish a simple mechanism to verify operability. This raises the question forming the foundation of this research, of how we implement a risk-based AI compliance audit framework that can be utilized to ensure that AI systems are operating safely and sustainably in the electricity sector against a volatile regulatory regime.

To validate that systems are utilized correctly and that they are not drifting from their design protocols is not a once-a-year audit; it needs to be a multipronged approach of ongoing verification exercises using automated systems to confirm processes and data against verified benchmarks, having human involvement in understanding and validating data submittals and decisions. It will require detailed quality assurance audits throughout

the year to ensure the system's design and operation against drift and that the necessary data protection mechanisms are in place, to ensure that the system is not a weak link to the sustainability of the business. Lastly, a decision is required on whether there will be independent AI audit frameworks for the information and operational technology environments or if a single framework can perform that function.

5.2 AI Compliance Framework for the Electricity Sector

After reviewing the literature review and the survey results from the subject matter experts, along with the key considerations in Figure 34, it became clear that this is a complex environment to establish a one-size-fits-all AI compliance audit framework or platform. In effect, differing regulations, legal requirements, decision matrix risks, and other key factors will influence how this should be structured for the different industries or sectors. Even within the electricity industry sub-sectors, different frameworks would be required, with differing organizational and deliverable focal areas, such as grid operators, traders, vertically integrated utilities, and pure generation facilities. However, key commonalities make it possible for a high-level AI compliance audit framework to be established as a baseline structure that can be re-aligned for the different sub-sectors to consider along with their intricacies and uniqueness.

The starting point for developing the AI compliance audit framework is categorizing the key factors that should be measured, tracked, and monitored per lifecycle stage of the AI system within the electricity sector. Figure 35 provides a summarized graphical representation of the factors and metrics that need to be monitored and integrated into the organization to ensure a successful compliance framework.

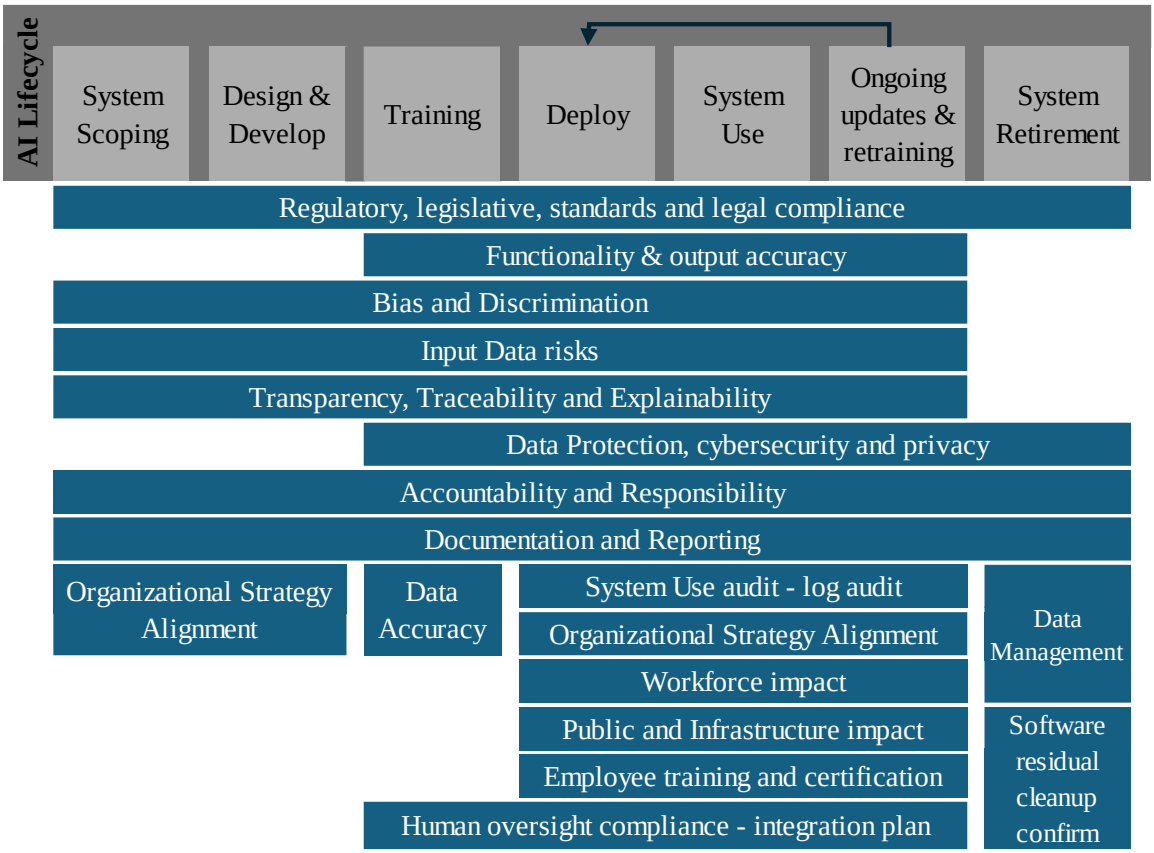


Figure 35: AI compliance audit framework factors through AI lifecycle

Each of these criteria forms the basis of the proposed AI compliance audit framework and needs to be considered according to the specific stage of the AI lifecycle it should be monitored for, how it should be monitored, and how often. As expected, several criteria overlap and should be monitored throughout the full lifecycle. However, there are also specific criteria that are tied to the specific stages of the AI lifecycle that need to be monitored and tracked on a more frequent basis due to the criticality of the system to cause data breaches and harm to people or infrastructure. Another area that is often overlooked when introducing a new AI system into an organization is the new AI system's interaction with existing information and operational technology systems, and even other AI solutions, including the interaction with data protection and cybersecurity systems. Suppose the

organization does not baseline the system's operation before introducing the new AI system; how does the company accurately monitor its performance to ensure that any underperformance or inaccuracy issues are from the new system and not because of incompatibility of the AI system and the complete amalgamated information, operational and AI systems.

Developing the AI compliance audit framework requires an iterative structured approach to ensure that the framework is aligned with the latest relevant governing regulations, laws, and legislation and that the framework is appropriately monitoring and reviewing key metrics of the AI systems to gauge ongoing compliance. The starting point of developing and implementing a compliance audit development procedure and, ultimately, the compliance framework is to nominate a sponsor from the organization's senior management team who will spearhead the process to ensure that this framework is accepted as a standard practice. The next step is for the organization to identify the relevant core set of regulations, legislation, laws, standards, and principles that will be adopted to govern the safe implementation and operation of AI systems within the organization. Thereafter, it is proposed that the procedure follow a parallel process, with the first portion being the initial development of the AI compliance audit framework, and establishment of the structure to undertake a periodic review of the organization's core governing documents and updates to the AI compliance audit framework. The second portion of the procedure is guided by a risk-based analysis undertaken on the AI system being implemented or updated to guide the ongoing core governing document reviews and updates in the framework. The proposed procedure for developing the AI compliance framework, as shown in Figure 36, adopts portions of the ISO9001 procedure, which is widely utilized and accepted within the electricity sector.

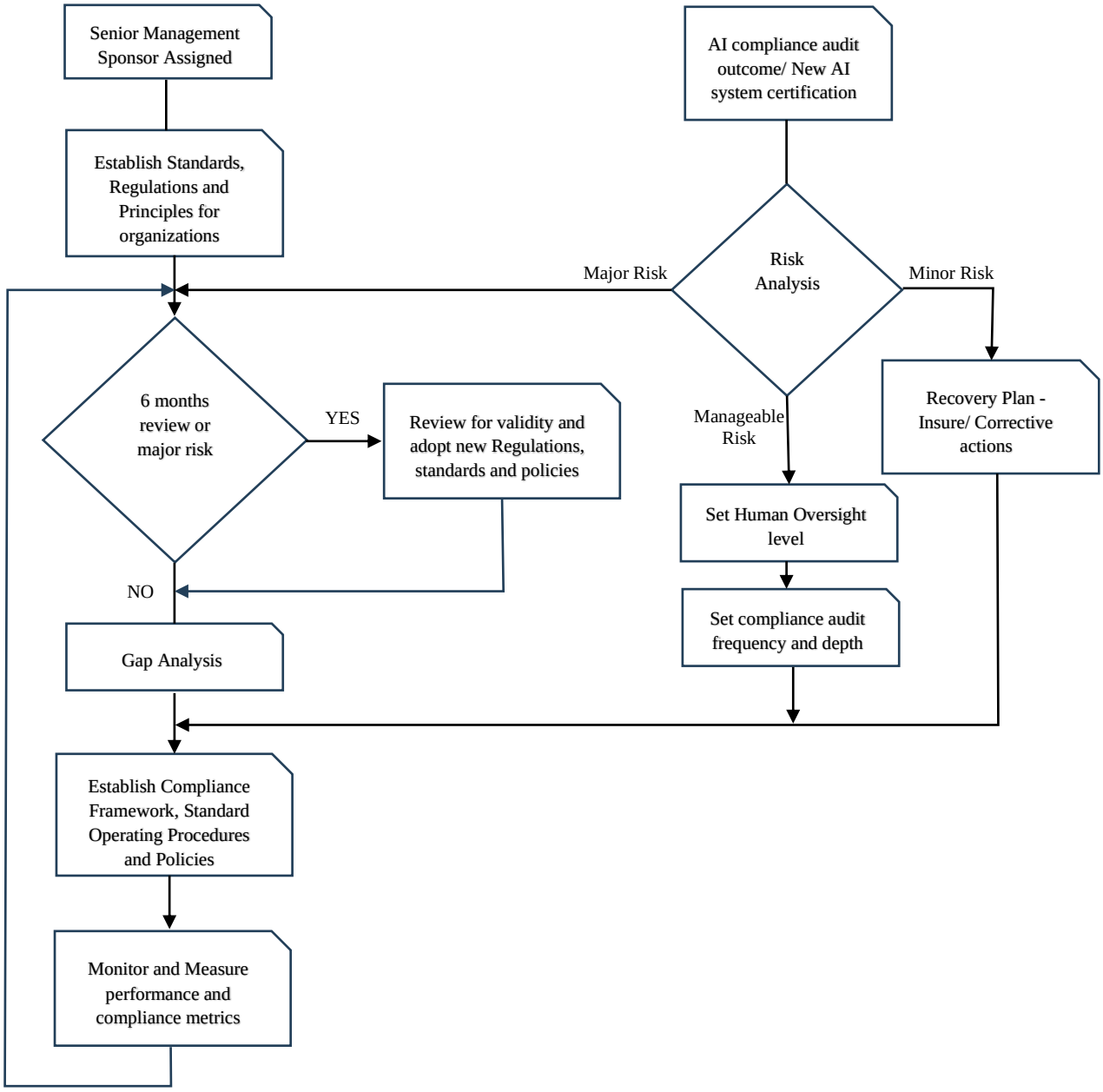


Figure 36: AI compliance audit development and review procedure

Once the core governance documents have been identified and adopted by the development team and approved by the senior management sponsor, this procedure comes into action. On the first implementation of the procedure, after the core governance documents are adopted, the organization will undertake a gap analysis to identify

shortcomings in the existing internal governance documents, policies, procedures, and systems to facilitate compliance management for the safe and sustainable AI system operations within their organization throughout its lifecycle. Once the gaps are identified, considering the critical compliance factors shown in Figure 35, the organization will be in a position to structure the AI compliance framework, operating procedures, and policies, outlining the types of system checks and audits, the frequency of the checks and audits, and ultimately the accountable parties. Once the AI compliance audit framework has been formalized and adopted, it will be implemented to monitor the system's performance and undertake the necessary compliance checks. On a six-monthly basis, unless triggered by a system risk audit sooner, it is recommended that the core governance documents be reviewed, and the procedure should be followed again to review and update the AI compliance audit framework.

The second portion of the procedure will be a risk analysis review of any new or upgraded AI systems to identify changes to the governance documents or unique governance or monitoring needs. The recommendation is that the bow tie risk model (Talbot, 2018) be utilized to identify the risk profile of the systems and to identify which risks can be mitigated or require additional governance structure to manage them, which risks require adjustments to the existing compliance framework to facilitate the increased risk and which ones require a recovery plan should the risk materialize. If a new or upgraded AI system is classified as high risk, it is recommended that the core governance documents be reviewed via the procedure to undertake a new gap analysis and revise the AI compliance audit framework. Should the system be classified medium to low-risk system, it is recommended that the level of human oversight be reviewed within the framework, along with the frequency and depth of critical checks and audits. And if the system is classified as very low risk, the recommendation is to develop a recovery plan,

which could be made up of system corrective actions should an issue be identified, system update reversals to remove unsafe additions, or to provide insurance to cover for the improbable occurrence of the risk.

Considering AI system integration into the information and operational technology environment in the electricity sector, using the development procedure discussed above, this research proposes the AI compliance audit framework, as overviewed in Figure 37, as the outcome of this research for the electricity sector. The proposed framework outlines the types and frequency of distinct checks and audits needed for the organization to manage the AI system safely and sustainably. Some of the audits are real-time system checks on a daily, weekly, or monthly basis. In contrast, others are done by an internal qualified audit team and verified by an external third-party audit team across the lifecycle of the AI system. What has become very apparent is that there is a need for collaboration between the AI developer, the organizations providing information and operation technology solutions, and the end-user organization to properly maintain the AI system compliance framework operational for the lifecycle of the AI systems.

It is also important to note that this AI compliance audit framework should be a live document and will continuously need revision, updating, and alignment as the AI systems mature, the national and international standards are updated, and organizations' processes and needs change.

The next subsections will discuss each component of the proposed AI compliance audit framework before the document proceeds to outline the proposed audit process, which will be implemented as part of the framework and aligned to the much-used ISO9001 or ISO14001 standards in the electricity sector.

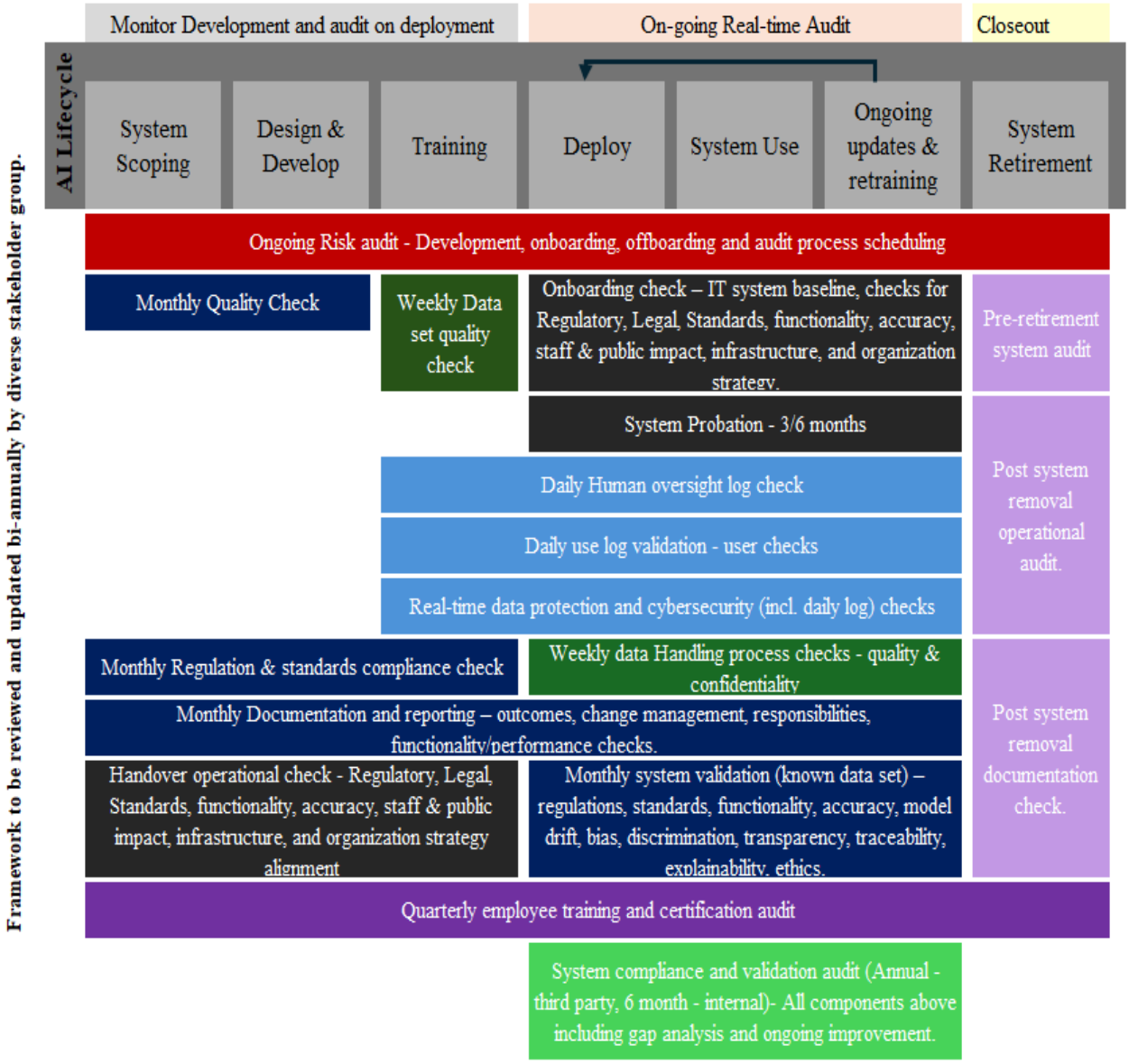


Figure 37: Proposed AI compliance audit framework for the electricity sector

To gain a better understanding of the proposed AI compliance audit framework, each of the checks and audits proposed are deconstructed and discussed below:

- *Ongoing Risk Audit* – Ongoing risk auditing is crucial to establish the necessary audit protocols, ensuring that organizations correctly incorporate the relevant checks and balances throughout the lifecycle of the AI system. During the

scoping, design, and development stages, the AI developer must include a continuous risk management system to ensure that any risks are identified early and mitigated or minimized. During the training stage, a risk profile will be undertaken with regards to the data used, the training methodology, and the verification procedures to ensure that the system operates as designed. When entering the deployment, usage, and upgrade stages, the system needs to undergo an onboarding risk assessment to ensure that the correct governance structures, system check, and audit structures have been integrated to monitor the system's performance and safe operation. Lastly, when a software package is obsolete and going to be replaced or removed, the organization needs to undergo an offboarding risk assessment to identify the impact on the collective software system when the AI system is removed, to ensure that the entire AI system is removed without any parasitic portions remaining that can create risks for the organization, employees, public or infrastructure.

- *Quality Check (Monthly)* – When the system is scoped, the AI developer will establish a minimum build quality and specify what quality standards and protocols the system will be built according to. As the system is designed and developed, the AI developer's quality assurance team must undertake monthly checks of the system's build quality according to the quality standards. Should there be any discrepancies, the AI developer should undertake a risk assessment to quantify the impact and decide whether the system needs to be reworked to re-align to the original design or whether it is still acceptable.
- *Dataset quality check (Weekly)* – A vital component of any AI system is the training data used during the training, update and retraining stages, as well as that used for the verification checks and audits during the deployment and usage

stages. It is recommended that a training data quality and management plan be established, along with a data usage log sheet, so that the AI developer and end-user can ascertain the data structure requirements, the level of confidential data to be included, identify who handled the data, record the purpose of the data being handled and record the period of use. During the training, updates and retraining stages, it is recommended that the training dataset quality be confirmed using the data quality and management plan structure and that detailed usage logs be kept. During the deployment and usage stages, the datasets that are built for system verification need to be audited weekly against the data quality and management plan to ensure that the system audits and checks are done against a known, controlled data baseline.

- *Human Oversight log validation checks (Daily)* – Including a person not just as a user but to provide a structured oversight role on the decision-making process has been an ongoing debate throughout the literature review and survey results. The European Union AI Act includes this as a mandatory requirement for high-risk systems to ensure that the system decisions can be reviewed and overruled if needed. However, this poses its own issues and additional risks, especially if the person providing oversight is not properly trained to do the function, becomes complacent in their check methods, or cannot trace and track how the system makes the decision. The proposal is that a human oversight plan be developed outlining the training requirements for the overseer, what the overseer will do to confirm the process and the decision, how the person will log observances and changes made to the decision, and any deficiencies identified within the AI system. As part of the AI compliance audit process, the human overseer is required to keep a daily log to capture all of the relevant

system data, as per the plan, which will be used for a daily check to ensure that the AI system is being tracked effectively, and to report any risks or issues to the information or operation technology environment for rectification or updating.

- *User Log validation checks (Daily)* – As with any other software solution, the user can impact how the AI system operates, provides decisions, manages data confidentiality, and generally how data is handled. The organization must maintain a processing and usage log to capture what the system was used for, by which user and note any concerns or abnormalities in the system outputs or decisions. It is recommended that this usage log sheet be maintained and checked daily throughout the training, deployment, usage, and upgrade stages so that incorrect, abnormal, or malicious use of the AI system, any data handling issues that could cause data privacy breaches or could create a back door for malicious parties to launch cyber-attacks are identified and thwarted early.
- *Regulation and Standards (Monthly)* - Each industry complies with specific governing laws, standards, and principles that guide the organization's facilities' ethical, legal and safe operation. At the establishment of the framework, the core governance structures for the AI system in the organization are identified and adopted. As part of the establishment of the AI compliance audit framework, the AI system must be monitored for compliance against these core governing structures throughout its lifecycle, but particular during the scoping, design and development and training stages, as this will ensure that these regulatory and standard operational structures form part of the core functionality of the AI system design.

- *Documentation and Reporting (Monthly)* – From the scoping stage to the updates and retraining stage, documenting and reporting on each action taken, decision made, and observance of the system usage and behaviour is essential for any audit to be properly undertaken. During the AI system scoping, design, and development stages, it is important that the scope, design, and development methodologies and process, along with the governance structure and standards utilized, are documented and categorized to enable the AI developer and end-user organization to baseline and track the system functionality. During the training stage, the datasets used, the methodologies implemented, the verification dataset structure, and system functionality tracking must be documented to record how the AI system operates under ideal conditions. During the deployment, usage, updates, and retraining stages, it is crucial to have a documented manual on how the system will be operated, who will be operating it, how human oversight will be employed, how decisions and outcomes will be monitored, how change requests will be processed and managed, who is responsible for each step within the value chain of usage and maintenance, and lastly how the functionality and performance will be monitored. This documentation should be updated each time an AI system is updated, new functionality is included, or new AI systems are integrated into the overall information or operational technology environment. Every month, monitoring and reporting should be issued by the responsible entity on the change management process, what performance and functionality checks were undertaken during the reporting period, how the system performed against those protocols, and list any changes in responsible parties in any portion of the operation and maintenance of the system.

- *Employee training and certification (Quarterly)* – Employees play a pivotal role throughout the entire AI lifecycle, which makes it a priority to ensure that the employees have the relevant skills and tools to do their job effectively, safely and successfully. For each AI system, it is recommended that an employee training and certification plan be established, outlining what upfront and ongoing training and certification will be required for the responsible employees to ensure that they can undertake their respective job functions, be it the software developer, the trainer, the system operator or the overseer. Without the necessary training, it is not fair to expect employees to carry out their job function properly, or to be held accountable for the deliverables and their accuracy. The recommendation is to undertake a quarterly check of the employee training and certification to ensure they are provided with the proper training and tools to do the job. If this shows a deficiency, additional training or certification should be scheduled for immediate action before said employee can continue with their function or they should work under supervision.
- *Real-time data protection and cybersecurity system (including daily log checks)* – Due to the large amount of sensitive customer and organizational data being processed throughout the training, deployment, usage, update, and system retirement stages of the AI system, data confidentiality, data management, data privacy, and cybersecurity are risks that need to be managed. This begins within the training stage by ensuring that the training datasets contain the minimal confidential information necessary, that the impacted individuals and organizations understand the risk and have provided their written consent to use the data, and that the data is controlled. In the deployment and usage stages, the AI system will be processing personal customer data, electricity usage data,

generation and grid operations data, as well as management data, which is confidential and sensitive to data breaches. A factor to consider is that the electricity sector digitization strategy creates a centrally monitored and controlled generation, transmission, and distribution system; when linked to a cloud-based AI toolset, this exposes the system to external actors. Should a breach occur, it is not just a data privacy risk but provides a potential back door to malicious actors to damage the organization through tampering with system operations. On the operational technology side, this could lead to the actors taking control of generation, transmission, and distribution devices, which could lead to power outages, infrastructure damage, and even harm to employees and the public if the actor turns on isolated grids that are being maintained. When it comes to the upgrades and retraining stage, the system will have data from the original training and all the processed data, so when it gets upgraded and retrained, the amount of confidential data available is higher and can increase the risk. Lastly, within the system retirement stage, the AI system will have retained all the confidential information from training and operation, and it is imperative that it is properly disposed of; it is also essential to ensure that when an old system is removed, that it does not leave any back doors into the system, for malicious actors to abuse. Considering this, the electricity sector must establish a comprehensive plan or policy on how data will be managed and protected, what real-time software platforms will be established to monitor data handling, how breaches will be detected and handled, and how the activity will be logged. What must be noted here is this should include how the data will be handled for all interconnected systems, including independent power producers, outsourced contact centres, payment

centres, and service providers. Once this is in place, it is recommended that, as part of the AI compliance audit platform, an integrated system cybersecurity compliance audit be undertaken as soon as a new AI system is installed, or a change is made to the baseline AI system. Within the implementation of the AI compliance audit framework, it is recommended that the daily data logs from the data management and cybersecurity system be checked for any anomalies in how the system has been used, against the baseline that could be deemed a breach, or if any external parties tried to gain access to the system. It is recommended that as part of this policy/plan, should any major breach be detected, the AI system should be isolated from the information and operational technology systems immediately until the defect is rectified.

- *Data handling process checks (Weekly)* – The data package, handling, and management are critical throughout the deployment, usage, updates, and training stages. Each AI system is designed and trained to accept data in a specific format with structured data packages. If the data packages are not appropriately structured and have inaccuracies or missing data, they could cause output inaccuracies and biased decisions, leading to discrimination against different racial groups, cultures, religions, governments, countries, income groups or organizations. Furthermore, if incorrect data is entered into the system, or the AI system is inappropriately fed with malicious information, there is a risk of a data privacy, confidentiality or cybersecurity breaches. As part of the AI compliance audit framework, it is required that the data quality and data package structure be checked every week, including the process to capture and utilize this data within the AI system to identify any risks and errors.

- *Handover Operational Check (On transfer to end user)* – This check encompasses the AI compliance audit framework's scoping, design/development, and training stages. This is an anticipated once-off check that will be undertaken after the AI system is designed, developed, and trained. In this check, the AI developer is required to perform a benchmark check of the final commercialized AI system alignment to the scoped regulations, legal requirements, and standards, as well as ensuring that the system is aligned to the end user organization's strategic intent, that it is accurate and fully functional as per the adopted principles. It is further required that the AI developer quantify that there is no harmful impact on employees, the public, and the organization's infrastructure as part of this baseline process. Should there be any misalignment, the AI developer should retract and rectify the system before deploying it to the end-user's organization.
- *Onboarding Operational Checks* – In the AI compliance audit framework, this check is relevant in the deployment, usage, and ongoing updates stages. The onboarding check is the end user organization's equivalent to the AI developer's handover check but from an organizational alignment perspective. The key differentiations are that within this check, the end-user needs to undertake a baseline of the complete information and operational technology systems before the AI system is installed and undergo a security baseline to ascertain if there are abnormalities and create a rollback point. Secondly, the system needs to undergo a functionality, safety, and security check after the AI system is installed to ensure no integrated system risks, clashes, or security breaches are added once the system is fully integrated. If there are risks, the AI system should be removed until this can be remedied.

- *Probation (3 to 6 months extendable)* – During the deployment, usage, and updates stages, it is recommended that the AI system, when first implemented be placed on probation as with any new employee. As much as this is not necessarily a pure audit function, it is placed here so that the organization and employees are cognizant that this is a new or updated system and that it needs to be monitored and overseen closely until everyone is comfortable with its functionality, accuracy, and outcomes. Adding in a probation period provides a sense of comfort to the employees as they know that should the system be dangerous the company is prepared to discontinue its use. It also creates a safe environment in which to socialize the system within the business and build trust with all stakeholders. During this period the involved stakeholders will have an opportunity to make changes to the proposed oversight levels and change the level and frequency of checks and audits depending on their confidence in the system.
- *System Validation (Monthly)* – Once the system has been deployed, it is fully functional, and continuous upgrades and betterments are performed, it is good practice to validate periodically that the system operates safely and sustainably per the system scope and design. Effectively, this will be an ongoing check using criteria like those undertaken in the AI developer handover operational check and the end-user organizations onboarding operational check. The aim is to perform a monthly revalidation of the systems, using a verified and known dataset (a checksum test, in essence), to ensure that the AI system is operating per the original design and the adopted governing structures, and it is still aligned with the end-user organization's needs. This entails checking the AI system operation against a known dataset, for alignment and verification

against regulations, legal guidelines or constraints, standards, system design functionality, output accuracy, model drift, bias, discrimination, transparency, traceability, explainability, ethics, and other organizationally adopted core governing principles.

- *Pre-retirement system audit* – When the AI system is at the end of its useful life and going to be retired, the organization must undertake a pre-retirement baseline of the complete information and operational technology systems before the AI system is removed. This retirement baseline, in conjunction with the baseline study taken during the onboarding operational check, provides a point of validation to detect whether the AI system has made any additional changes to the information or operational system environment since initial implementation that need to be considered when the system is uninstalled. The pre-retirement baseline will provide a view of what confidential and private data is held within the system or its repositories that need to be appropriately dealt with to ensure that they are correctly removed to mitigate data breaches.
- *Post Systems Removal Operational audit* – After the AI system has been removed, during the retirement stage, it is essential to audit the complete information and operational technology environment to ensure that all remnants of the system have been removed, no security holes have been introduced, and that the confidential and private data is properly removed and disposed of. Furthermore, it is important to confirm, that the complete information and technology systems operate per the operational baseline undertaken as part of the onboarding operational checks. If there are any abnormalities after the system is uninstalled, the organization needs to secure the necessary support to identify the issues and rectify the system to eliminate risks.

- *Post System Removal Documentation check* – The last important item during the system retirement stage is to check that the removal/uninstall process, validation process, and changes are properly documented. It is also recommended that all system documentation and reporting processes be updated to account for the removal of the AI system so future operations, change management, and audits can be properly undertaken with accurate information.
- *System Compliance and Validation Audit (Six monthly)* – During the deployment, usage, updates and retraining stages, a complete AI system audit must be undertaken by a qualified in-house audit team independent of the regular operational team. This will encapsulate auditing of the daily, weekly, monthly, and quarterly checks listed in the AI compliance audit framework, but from an independent standpoint. The aim is to confirm that the system complies with all the core governing structures, operates as per the original design, and performs as per the onboarding operational baseline. This will be critical in identifying model drift, biases, discrimination, training deficiencies, and other non-compliance issues that must be dealt with to ensure that the system is providing accurate, repeatable decisions and outputs in a safe, ethical, and sustainable manner.
- *System Compliance and Validation Audit (Annual)* – As with the six-monthly internal compliance and validation audit, this annual audit is aimed at confirming that the system continues to comply to all the core governing structures, operated as per the original design and performs as per the onboarding operational baseline. The only key difference is that this needs to be undertaken by an accredited third-party auditor, who will be able to provide

an independent review to ensure that the organization is not missing or ignoring essential items. Both the internal and third-party audit will be covered more in-depth in the next section, which outlines the proposed audit process.

The one caveat to note here is that for some electricity sector organizations, depending on their system setup, operational requirements, and risk assessment, the AI compliance framework may need to be split and framed slightly differently for AI systems integrated into the information and operational technology environment. However, the same procedure and base framework can still be utilized to adapt to the specific system needs and risks for compliance audits of the AI systems being integrated into the information and operational technology environments respectively.

5.3 AI compliance Audit Process

As with any audit, a formal structure must be created to guide and document the audit process so that everyone understands how this will be undertaken, who is responsible for the different steps in the audit, and how the outcomes will be reported and dealt with. With the electricity sector being a highly regulated and standards-driven sector, they already have many compliance auditing structures in place, which can be both beneficial and detrimental to the acceptance of another compliance process. For this reason, it was recommended that the AI compliance audit framework be integrated into existing frameworks in the electricity sector. Referring to the survey results from the subject matter experts, especially those representing the electricity sector, they ranked the integration of the AI compliance audit framework with specific relevant protocols within the sector as the best mechanism to gain employee and industry support of the additional requirements. In alignment with that suggestion, this research proposes that the AI compliance audit

framework and process be integrated into the ISO standardized audit functions within the sector, as many electricity sector organizations comply these structures.

Per the AI compliance audit framework, a formal audit process is required to undertake the six-monthly and annual system compliance and validation audits. The recommendation is to adapt the ISO 9001 quality management audit process, as it is closely linked to the proposed AI compliance audit framework and is widely accepted in the sector. The adapted ISO audit process proposed for auditing AI systems is shown in

The audit process starts with the AI developer and information or operational technology manager setting the technical and operational audit considerations aligned to the AI compliance audit framework, which will include, but not be limited to:

- Identify whether the system still complies with the design and operational standards.

Figure 38. The audit process breaks down the process accountabilities between four distinct stakeholder groups, the groups are:

- The AI developer or information/operational technology manager would be responsible for the integration, maintenance, and support of the AI system to ensure the system operates as designed within the organization.
- The auditor is an internally nominated cross-functional audit team for the six-monthly audits or an external independent third-party for the annual audit.
- The auditee would, in essence, be the operational employees that operate the AI system and provide oversight. In many instances, the auditee would also include the information technology specialist team members and
- The top management team, which would be made up of the senior management sponsor and other Executive or Board oversight members.

The audit process starts with the AI developer and information or operational technology manager setting the technical and operational audit considerations aligned to the AI compliance audit framework, which will include, but not be limited to:

- Identify whether the system still complies with the design and operational standards.

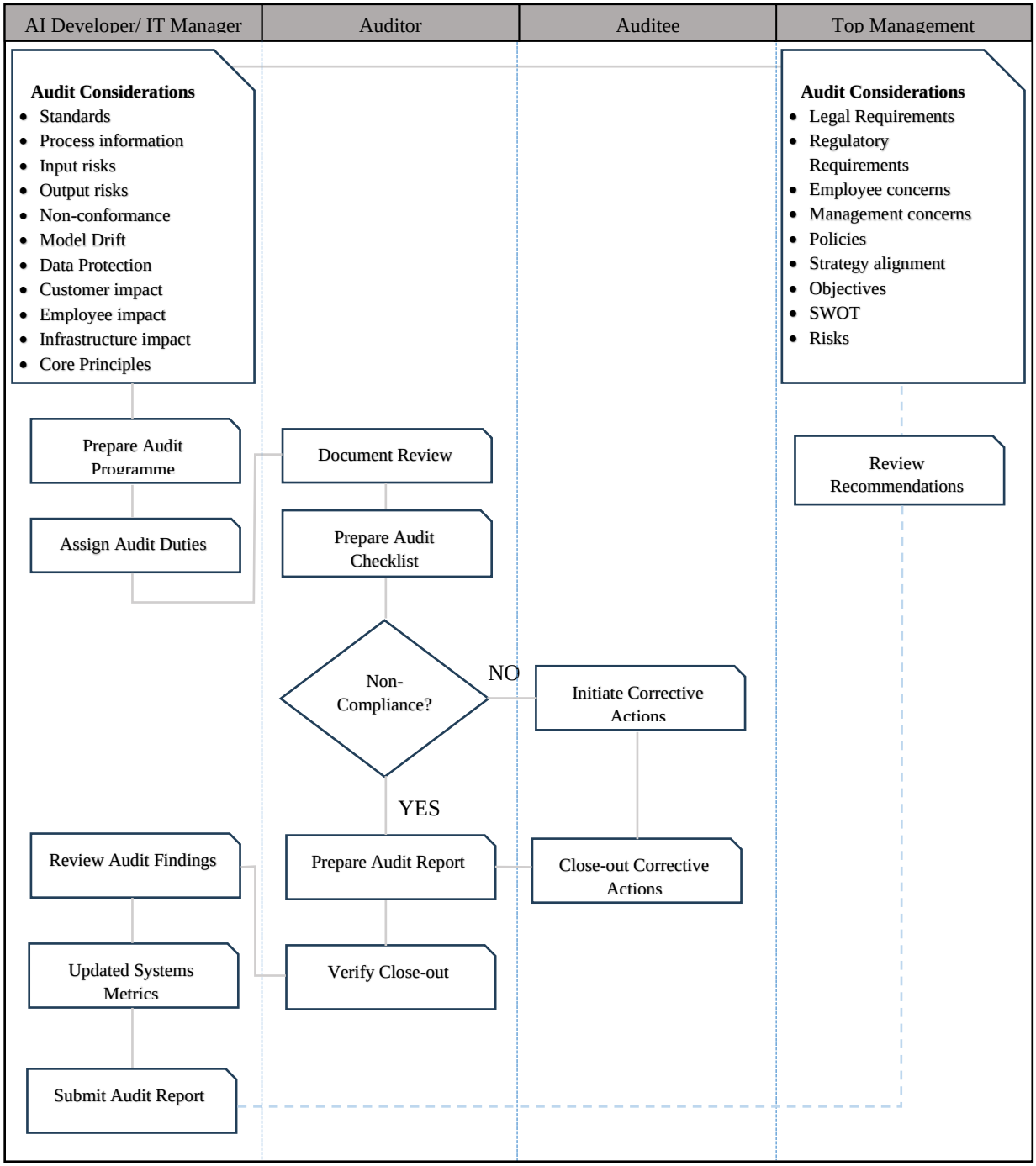


Figure 38: Proposed AI Audit Process

- Process information, outlining how the system processes data and what checks and balances should be reviewed.

- Input risks, such as data package format, data type, and confidentiality.
- Output risks focus on the decision's accuracy, consistency, and repeatability.
- Non-conformance is a broad review to ensure the system aligns with the organization's policies and procedures.
- Model drift focuses on identifying changes that may have occurred in the AI system that may have altered its structure, processes, or decision-making.
- Data Protection focusing on the existing protection structures to ensure that the AI system processes the data without break privacy rules and that it does not share data without consent of the people or organization involved. This will also focus on the compliance and operability of the cybersecurity mechanisms to protect the system from malicious attacks.
- The customer impact review is focused on identifying any system decisions-making process changes that may detrimentally impact the customers or public.
- The employee impact review focuses on understanding whether the AI system implementation negatively impacted the employees and if it is a safety or security risk.
- The infrastructure impact review focuses on understanding whether model drift changes, system decision-making changes, or updates heighten the risk of damage to the information technology, operational technology software systems or physical infrastructure.
- The core principles review is focused on ensuring that the system functions within the ethical and moral boundaries it was designed, that the process and decisions made are transparent, traceable, and explainable and that no bias or discrimination is introduced.

The second portion of the audit process is for the top management team to set the audit considerations from a legal, governance, and organizational perspective, aligned to the AI compliance audit framework, which will include, but not be limited to:

- Legal requirements that need to be considered as part of how data is processed, how decisions are made, and how that impacts the customers and public. It will also consider other legal alignments from a licensing and trading perspective.
- Regulatory requirements focus on compliance with the core regulatory, legislative, and other governance structures used as guardrails in the AI compliance audit framework. This will also focus on reviewing whether the core governance structures being used are still valid for the AI system.
- Any valid employee concerns that were raised during the audit period on the AI system functionality, risks, and other deficiencies will be raised for checking and rectification.
- As with the employee concerns, any valid management concerns raised during the period on the AI system functionality, risks, and other deficiencies will be raised for checking and rectification.
- A policies and procedures review will be carried out to ensure that the internal management documentation is correctly aligned for the AI system and that they provide the necessary guardrails to manage the AI system operability.
- Strategy alignment reviews the alignment of the AI system decision process, outputs, and decisions against the organizational strategic intent and deliverables.
- From an objectives perspective, it is imperative that the system is reviewed to validate that it is still meeting the established design objectives, from a functionality, efficiency improvement, and employee support perspective.

- The senior management team will review all the prior and current audit outcomes to identify whether the changes in the system and outcomes of the audit have made any changes to the strengths, weaknesses, opportunities and threats from an organizational perspective. This will allow them to decide whether changes are required on the AI system to better suit the organization or if it has added new opportunities for them to pursue.
- Throughout the year, a company risk register is kept; any key findings that pertain to the AI system, added during the audit period, will be scoped into the audit requirements, and any findings from the audit will be added to the risk register so that they are not lost.

Once all of the audit considerations are agreed upon between the parties, the AI developer, information or operational technology manager will prepare the audit program, outlining the audit metrics, how the monitored metrics will be audited, the auditing duration, the makeup of the auditing team, how the audit findings will be categorized, how they will be documented, how they will be reported and how feedback will be provided to the respective stakeholders. This audit program will also outline what needs to be done should deficiencies or risks be identified and who needs to be informed if high-risks items are identified that need immediate attention. When the audit program is complete and approved by all parties, the AI developer, information or operational technology manager will assign the audit functionalities and requirements to the relevant stakeholders, with a list of critical outputs and timelines.

Once the audit duties have been assigned, the auditor will review all documentation provided for the system as part of the audit initiation, and all the audit requirements outlined in the audit program. Using this information as a baseline, along with the previous years' audit structures, the auditor will prepare an audit checklist which will, at minimum, outline

what documentation, logs, or reports will be required, what system access will be required, what datasets will be utilized for baseline verification, what scenario's need to be run on the system, what policies, procedures, standards and other core documents need to be made available for the audit. It will also outline who from the auditee, AI developer, information/operational technology team, or management team should be available for different portions of the audit.

The auditors will utilize the checklist to guide the information gathering and strategize how they will confirm specific functionality and system compliance. At this point, the complete system audit will commence with the support of relevant stakeholders. The outcome of the audit process will provide a qualified view of whether the system is compliant or non-compliant with the audit program requirements. If the system is compliant, minor deficiencies or corrective actions should still be listed even if they are deemed non-material in the system risk operability profile. However, suppose a system is partially compliant but has significant risks or material deficiencies identified; it should follow the route of non-compliance shown in

The audit process starts with the AI developer and information or operational technology manager setting the technical and operational audit considerations aligned to the AI compliance audit framework, which will include, but not be limited to:

- Identify whether the system still complies with the design and operational standards.

Figure 38 to allow for the necessary rectification and corrective actions to be undertaken to meet the compliance requirements.

If the system is compliant, the auditor will prepare a detailed audit report outlining the audit process, the audit metrics, the details of the audit team, details of the audit findings, any non-compliances identified along with the corrective actions undertaken, any

low-risk corrective actions to be undertaken, and any recommendations for metric changes for the next audit program. If the system is non-compliant or has significant deficiencies, the auditors will issue a list of corrective actions to the auditee for immediate rectification. Once the auditee has undertaken all the corrective actions, the auditor will confirm the actions taken, and they will then prepare the detailed audit report. It should be noted that it is the prerogative of the auditor, dependent on the risk level identified if non-compliant, to recommend that should the system corrective actions not be instituted in a set period, that the system should be locked out until rectification is complete to ensure a safe and sustainable system.

Once the full audit report is drafted, the auditors will undertake a final verification process to ensure the quality of the report and that all factors have been included before the report is issued to the AI developer or information/operational technology manager. Once issued to the AI developer or information/operational technology manager, they will review the report for any corrective actions already taken, as well as system updates or minor corrective actions that still need to be actioned; this will include any recommended changes to the ongoing metrics to be measured for the AI system during its normal operations. They will provide a timeline for all the minor corrective actions and system metric updates before finalizing the report and issuing it to all relevant stakeholders for review. The senior management team will have an opportunity to review the report and make recommendations for additional actions or reviews to be undertaken. Furthermore, senior management will use this report to review the compliance of the audit considerations they put forward and ensure that the necessary governance and oversight guardrails are in place. Should there be deficiencies, they will action an AI compliance audit framework review to ensure that the correct core governance structure is correctly aligned to the latest regulations, standards, legislation and principles.

5.4 Summary

Over the past decades, the energy sector's drive has been to digitize and de-carbonize, which provides profound benefits while creating a more complex environment heavily dependent on large volumes of real time-data for system management and decision-making. As more granular and complex data becomes available to the electricity sector, the sector have had to secure new skill sets and develop or adopt more complex data analytic tools, which, in recent years, have included AI driven solutions. As these AI systems become more prevalent and make autonomous decisions within the critical infrastructure sector, such as the electricity sector, the risk profile increases, primarily if the AI system is not regulated or governed. In the electricity sector, mistakes or errors from an AI system, such as incorrect decisions or incorrect grid operations, influence more than financial returns but can cause severe equipment damage and harm to people.

The proposed AI compliance audit framework development procedure, the AI compliance audit framework, and the AI audit process outlined in this chapter provide a fit-for-purpose compliance solution that can be adapted to all critical infrastructure sectors. Employing this risk-based approach to ensuring that the framework is always aligned with the latest governing principles for the AI systems and the industry, allows the electricity sector to ensure the AI system compliance, and the safety of their organization, employees, infrastructure, and the public.

Lastly, aligning the proposed AI compliance auditing procedure, framework, and process to specific existing compliance or governance processes within the electricity sector gains more straightforward organizational implementation and support. The linkage to existing compliance or governance processes provides the employee's comfort and

builds trust as they already have intimate knowledge of the existing process and the benefit it provides the organization, achieving seamless buy-in from the employees.

In the next section, a comparison of existing AI audit processes and the one proposed through this research will be reviewed. This will highlight the robustness and uniqueness of the proposed AI compliance audit framework compared to those proposed by other researchers, governments, and standards organizations historically.

CHAPTER VI:

FRAMEWORK COMPARISON AND DISCUSSION

As AI matures and becomes autonomous, such as self-driving vehicles and credit evaluation systems, the public and organizations have acknowledged the increased risk to society and humankind. This has not only driven a flurry in the development of principles, regulations, and policies but has had governments, academia, and private organizations focused on the development of AI compliance audit methods. This chapter compares these existing audit frameworks or methods to those proposed in this research to highlight the value-add and the unique positioning and structure of the proposed AI compliance audit framework.

6.1 Information Commissioners Office (ICO) AI Audit

The ICO developed an AI auditing framework to ensure compliance with data protection laws while addressing the unique risks AI systems pose. The ICO identified auditing as playing a pivotal role in educating and assisting organizations to meet their obligations under the Data Protection Act and will enhance public trust in AI technologies, which are increasingly integrated into organizations in various sectors (ICO, 2022). In effect, the framework provides organizations with best practices for data protection compliance for AI systems, whether developed in-house, when implementing bespoke AI systems from third-party developers or utilizing AI as a service solution (ICO, 2023).

The AI framework aims for the ICO to undertake independent audits on organizations' AI systems. The AI auditing framework provides a structured audit methodology for the ICO's assurance investigation teams to evaluate organizations' compliance with data protection obligations when developing and using AI systems (Birhane *et al.*, 2024). The framework also guides organizations on implementing best

practices for facilitating the development and procurement of AI systems, ensuring that the AI system processes personal data fairly, legally, and transparently. The critical components of the guidance framework can be categorized into four sections:

- Accountability and Governance - this section outlines the importance of data protection impact assessments to identify and understand the impact of AI systems on the organization and public, allowing the organization to develop an informed data protection and individual rights mitigation strategy.
- Fair, Lawful, and Transparent Processing – this section unpacks the lawful foundation required for processing and retaining personal data and mechanisms to undertake performance assessments for AI systems. It further outlines strategies to identify and mitigate discrimination risks from the AI systems process and decision-making.
- Data Minimization and Security - this section focuses on ensuring that only the data necessary for processing is collected for the AI system to limit data exposure. It focuses on establishing or adopting robust security measures to protect personal information.
- Facilitating Individual Rights – this section outlines how organizations should uphold individual rights, especially for AI systems making automated decisions, ensuring that affected individuals can exercise their rights and understand the risks involved.

The ICO emphasizes a risk-based approach to AI auditing, which involves identifying potential risks to individual rights and expressing freedom for individuals to choose how their data is processed and retained when associated with AI applications. Organizations are encouraged to implement appropriate technical and organizational measures to mitigate these risks, which aligns with general requirements under data

protection law, ensuring that organizations do not overlook legal obligations even when risks appear minimal.

The ICO's AI auditing framework provides methodologies to audit and ensure fair processing of personal data, with guidance structured around accountability, lawfulness and fairness, security and data minimization, and individual rights (Kazim *et al.*, 2021). This framework focuses predominately on fair and safe handling of personal data and does not delve into the broader sense of oversight for AI systems. In contrast, the AI compliance audit framework proposed in this research goes much broader than just data protection in that it considers other vital items, such as:

- Integrating the compliance audit functionality with relevant existing frameworks to get quicker organization and employee acceptance of the additional requirements.
- National and international regulations and laws alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.
- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organizational accepted core governance structures, including regulations, laws, standards, and generally accepted principles.
- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.

- Safety of employee's, public, and infrastructure.
- Accuracy, repeatability, and traceability of decisions made
- Human involvement in AI system oversight, dependent on system risk profile and autonomy.
- Skills and training of employees to ensure they can perform their jobs safely and thoroughly.
- Continuous system validation against a known dataset and baseline.
- Alignment to organization strategy and deliverables.
- Mechanism to ensure that AI systems being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.
- Auditing of documentation and reporting for AI systems to ensure that all operational information, changes, and risks are properly identified and captured, not just focusing on creating the documentation.
- Provides a dual auditing procedure with internal and third-party audits to drive accountability.

6.2 European Union's AI Act

The European Union's AI Act represents a significant regulatory framework (Scannell *et al.*, 2024) structured to ensure AI systems' safe, responsible, and ethical development and deployment. It categorizes AI systems based on risk levels and outlines specific compliance and audit measures, focusing on high-risk applications (Bhuvan, 2023). The AI Act subjects providers of high-risk AI systems to stringent compliance measures, including conducting detailed evaluations to identify and mitigate potential risks associated with the AI system, maintaining comprehensive records throughout the lifecycle of the AI system, from design to post-market monitoring, and conformity assessments to

demonstrate compliance (Thelisson and Verma, 2024), allowing providers to affix a CE marking to their products, which signifies adherence to European Union's regulations (Musch *et al.*, 2023).

The European Union's AI Act classifies AI systems into four key categories (European Commission, 2024), which it utilizes to set specific compliance metrics for system development and deployment (KPMG, 2024):

- Prohibited systems are AI applications that pose unacceptable risks to organizations and the public or contravene the European Union's values. Examples are systems that utilize behavioural manipulation, exploit the vulnerable characteristics of people, provide social scoring by public authorities or provide real-time remote biometric identification for law enforcement.
- High-risk systems are AI systems that pose a high risk to health, safety, environment, and fundamental rights and should be subject to strict regulations. Examples include AI systems used in financing, health or life insurance benefits evaluations, biometric identification, analysis of job applications for recruitment, and those in critical infrastructure management.
- Limited-risk systems are AI systems at risk of impersonation or deception and are thus mandated to comply with transparency obligations, such as informing users that they are interacting with an AI system and not a person. Examples are AI systems that interact with consumers, such as chatbots, and generative AI for image, audit, or video manipulation.
- Minimal-risk systems are effective AI systems that do not pose any risk to the government, organizations, the public, or infrastructure. No proposed regulation is put forward for these systems within the Act. Examples are AI-based spam filters, video games, and entertainment software.

The AI Act provides auditing requirements but only for high-risk AI Systems. The AI Act deems that prohibited AI systems should be outlawed completely, so no audit regime is required. In contrast, the limited risk systems should be self-governed by organizations to ensure that they are transparent in all they perform. For the high-risk AI systems, the AI Act breaks the audit requirements into two sections, namely:

1. Conformity Assessments (Pre-market assessment) – all high-risk AI systems must undergo conformity assessments before being placed on the market or deployed within organizations. These AI Act designed assessments are structured to evaluate whether the systems comply with established regulatory, legal, and ethical design standards.
2. Post-Market Monitoring – providers of high-risk AI systems must establish post-market monitoring plans structured to continuously evaluate the performance of AI systems throughout their lifecycle. The plan should include establishing mechanisms for tracking incidents and performance issues, reporting serious incidents to relevant authorities within a specified period, and implementing corrective actions based on the monitoring outcomes.

The AI Act emphasizes utilizing internal and third-party auditors to ensure AI system compliance and notes that both play a pivotal role. The internal auditors play a crucial role in verifying that high-risk AI systems meet the necessary assessment criteria set out in the compliance plan. The third-party audits are slated to enhance accountability by providing independent evaluations of compliance and performance.

The AI Act is one of the more comprehensive solid foundations for AI governance approved to date (Novelli *et al.*, 2024). However, it still has limitations because its risk-based approach is tied to listing or naming specific risks and applications rather than

creating system risk definitions, which means it will be outdated quickly. The list-based approach is likely to be ineffective on the procedural complexities of AI system development, deployment, and use, which in turn might fail to suitably acknowledge the influence AI has on people's daily lives, including realizing their fundamental rights (Beck and Burri, 2024). Some organizations and researchers have proposed and developed frameworks using the AI Act to create a more structured and actionable auditing approach. One such example is the researchers from the University of Oxford who have developed tools like capAI to help organizations translate high-level ethical principles into actionable compliance measures aligned with the European Union's AI Act, which have made them more robust (Floridi *et al.*, 2022).

The AI Act provides a structure to perform broad compliance oversight for AI systems, but only if they are classified as high-risk. The Act completely discounts "prohibited" systems and outlaws their use, which creates a market for uncontrolled AI systems to be sold or implemented illegally. In contrast, the AI compliance audit framework proposed in this research provides a risk-based compliance audit mechanism, which adapts the level and complexity of the audit dependent on the risk levels of the AI system. Furthermore, the proposed AI compliance audit framework goes much broader than the AI Act in that it considers the following:

- AI systems of all risk categories, not just high-risk.
- Integrating the compliance audit functionality with relevant existing frameworks to get quicker organization and employee acceptance of the additional requirements.
- National and international regulations and laws alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.

- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organization accepted core governance structures, including regulations, laws, standards, and generally accepted principles.
- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.
- Safety of employee's, public, and infrastructure.
- Skills and training of employees to ensure they can perform their jobs safely and thoroughly.
- Continuous system validation against a known dataset and baseline.
- Alignment to organization strategy and deliverables.
- Mechanism to ensure that AI systems being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.

6.3 NIST AI Risk Management Framework

The National Institute of Standards and Technology (NIST) has developed the AI Risk Management Framework, a comprehensive guideline designed to assist organizations in managing the risks associated with AI systems. The Risk Management Framework serves as a guideline for organizations and governments seeking to navigate the complexities of AI risk management by providing a structured approach to identifying, assessing, and mitigating (Miles, 2023) AI risks throughout the lifecycle of AI systems.

By providing a structured approach and emphasizing the socio-technical nature of AI systems, the framework aims to empower innovative, ethical, and responsible development of AI systems (Dotan *et al.*, 2024). This framework is intended for voluntary use and aims to improve the incorporation of trustworthiness into AI systems (NIST, 2023).

The NIST AI risk management framework is structured around four core functions that it recommends be implemented by organizations:

- The establishment and adoption of governance structures and processes to foster a culture of AI risk management within the organization. This includes defining roles and responsibilities, ensuring accountability, and promoting transparency in AI practices.
- Identify, assess, and map the risks associated with the AI systems and their usage within the organization.
- Evaluate and analyse the organization's exposure to the identified AI system risks. It is recommended that the organization implement metrics and measures to assess the performance and reliability of AI systems, focusing on their trustworthiness characteristics.
- The organization is recommended to implement risk management controls to mitigate identified risks. This core function emphasizes that resources must be allocated to regularly address the mapped and measured risks, ensuring that the organization can respond effectively to incidents or emerging threats.

The specific audit requirements covered in the risk management framework can be summarized as follows:

- The audit process begins with identifying all AI systems within the organization.

- The auditors should evaluate existing risk management practices through interviews with personnel, documentation reviews, and observations of AI system development processes. This assessment helps determine current practices alignment with the risk management framework's requirements.
- Establish a documentation standard to ensure comprehensive documentation throughout the lifecycle of AI systems. Auditors must ensure that critical processes, calculations, and models are well-documented.
- Test the effectiveness of risk management controls during audits. This involves evaluating the efficacy of these controls in mitigating the identified risks and whether they are consistently applied across the organization.

The framework promotes continuous evaluation and adaptation of risk management practices, which are essential as the field evolves rapidly. The NIST AI risk management framework focuses predominately on compliance with the AI systems models and outcomes against its governing structures, establishes a clear documentation trail, and ensures that the identified risks have a functional mitigation plan. This covers only a tiny portion of the actual risk profile of AI systems during their lifecycle. The proposed AI compliance audit framework from this research goes much broader than the NIST framework in that it considers the following:

- AI systems of all risks are much broader than just the model, algorithm, and decisions made.
- Integrating the compliance audit functionality with relevant existing frameworks to get quicker organization and employee acceptance of the additional requirements.

- National and international regulations and laws alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.
- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organizational accepted core governance structures, including regulations, laws, standards, and generally accepted principles.
- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.
- Safety of employee's, public, and infrastructure.
- Skills and training of employees to ensure they can perform their jobs safely and thoroughly.
- Continuous system validation against a known dataset and baseline.
- Alignment to organization strategy and deliverables.
- Mechanism to ensure that AI systems being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.

6.4 Chartered Institute of Internal Auditors AI Auditing Framework

The Chartered Institute of Internal Auditors (IIA) published an AI auditing framework to guide internal auditors on approaching AI auditing. Their framework

addresses some of the unique challenges posed by AI technologies and emphasizes the importance of governance, data quality, and performance monitoring in the auditing process. However, it is still predominately focused on the financial outcomes of the organization and not the AI system operation. The focus on risk assessment, best practices, governance, and continuous learning empowers internal auditors to effectively navigate the complexities of AI and provide valuable assurance to their organizations (IIA, 2023).

The IIA AI auditing framework is made up of five primary components that it deems essential for the proper auditing of AI systems, and these are (Institute of Internal Auditors, 2017):

1. Organizational strategic alignment: The IIA recommends that the organization establish an AI strategy to guide how AI systems are developed, utilized, and audited within the organization. It emphasizes that the AI strategy needs to create alignment between the proposed AI system's goals and the organization's strategic objectives. The strategy should place special care in creating cyber resilience to ensure that the organization can resist, react to, and recover from cyberattacks.
2. Governance: The framework stresses the need to create structures, processes, and procedures aligned to relevant regulations to effectively direct, manage, and monitor AI system interactions within the organization. This includes defining roles, responsibilities, and accountability mechanisms to manage AI risks effectively and ensuring that the people filling the roles have the relevant skillsets. The auditing framework aligns with almost all financial frameworks in that it recommends the establishment of three lines of defence:
    - The first line of defence is focused on the operational managers taking ownership to manage the AI systems risks on a day-to-day basis.

- The second line of defence is the establishment of compliance, ethics, risk management, and information privacy policies and procedures to have a structure to audit the system's operations.
- The third line of defence uses internal auditors to provide independent assurance over AI risks, governance, and controls. It also recommends using external third-party auditors, to check materiality of the AI functions.

3. Data Architecture and Infrastructure: This focuses on data handling and infrastructure requirements to manage the large data quantities the AI systems require. The three areas of emphasis are data accessibility, data privacy and security, and data ownership and usage roles.
4. Data Quality: The completeness, accuracy, and reliability of the dataset used to build AI algorithms are critical. This is especially imperative when using two AI systems that are not communicating with each other but using the same datasets and making co-dependent decisions.
5. Use of Standards: The framework recommends that internal auditors conform to the IIA's applicable standards when planning or undertaking AI audits.

When it comes down to the actual audit process for AI systems within the IIA audit framework it involves several critical steps:

- Auditors must clearly define the scope and objectives of the audit plan, outlining what aspects of the AI system they will evaluate, including compliance with legal standards such as GDPR.
- Undertake a risk assessment to identify potential risks associated with the AI system's operation and its impact on organizational strategic objectives.

- Evaluate the effectiveness of the controls implemented for mitigating the identified risks.
- Provide a comprehensive report outlining observations, areas for improvement, and recommendations for enhancing governance.

By focusing on governance, strategy, human factors, and ethical considerations, auditors can help organizations leverage AI responsibly while ensuring compliance with legal and regulatory standards. One of the critical issues here is that this is more focused on governance than system operations compliance. In contrast, the AI compliance audit framework proposed in this research goes much broader than just the considerations in the IIA framework in that it considers other vital items, such as:

- National and international regulations and laws alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.
- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organizational accepted core governance structures, which includes all regulations, laws, standards, and generally accepted principles, and not just financial facing regulations.
- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.
- Safety of employee's, public, and infrastructure.

- Accuracy, repeatability, and traceability of decisions made
- Human involvement in AI system oversight, dependent on system risk profile and autonomy.
- Continuous system validation against a known dataset and baseline.
- Mechanism to ensure that AI systems that are being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.
- Auditing of documentation and reporting for AI systems to ensure that all operational information, changes, and risks are properly identified and captured, not just focusing on creating the documentation.

6.5 Model AI Governance Framework

Singapore's Personal Data Protection Commission (PDPC), in conjunction with the World Economic Forum, created the Model AI Governance Framework focusing on the ethical and responsible use of AI technologies. The Model AI Governance Framework emphasizes the importance of governance, risk management, and compliance in deploying AI technologies. By providing structured guidance and tools for implementation, the PDPC seeks to ensure that AI technologies are developed and deployed responsibly, fostering a sustainable digital economy (PDPC and IMDA, 2020).

First and foremost, this is a governance guidance framework to assist organizations in building, integrating, and utilizing AI systems transparently, fairly, and responsibly. However, it does provide some oversight and risk management controls for ongoing compliance checks. The Model AI Governance Framework provides guidance on establishing governance principles and internal governance structures as outlined below:

1. Governance Principles: The framework aligns itself with eleven core governance principles recommended for organizations to adhere to:
    - Transparency: Clear and concise communications to all stakeholders.
    - Explainability: Decisions made by AI systems must be understandable.
    - Repeatability/Reproducibility: Processing and outcomes should be consistent and repeatable across similar conditions.
    - Safety and Security: AI systems must be developed and operated to be resilient against failures or attacks.
    - Robustness: AI systems should perform reliably under various conditions.
    - Fairness: Avoidance of bias and discrimination.
    - Data Governance: Management and protection of data used in AI systems.
    - Accountability: Clear assignment of responsibility for AI processing and outcomes to all stakeholders.
    - Human Agency and Oversight: Ensuring that the appropriate human involvement is included in automated AI decision-making.
    - Inclusive Growth: Ensuring stakeholder inclusion and equitable benefits from AI technology development and usage.
    - Societal and Environmental Well-being: Considering broader impacts on society and the environment.
2. Internal governance structures: Organizations are encouraged to establish standard operating procedures for monitoring risks associated with AI systems and training staff on ethical AI practices. Clear stakeholder roles and responsibilities should be set for using, managing, and maintaining the AI.

3. Compliance: The framework involves several critical risk control steps to ensure compliance with the established governance principles. The summary of the audit process within the framework are:
    - Risk assessment: Organizations must conduct a comprehensive AI system review to identify and classify risks associated with their systems.
    - Documentation and record-keeping: Detailed records of design processes, data lineage, and algorithmic decisions must be kept for the AI systems.
    - Technical audits: The framework recommends that organizations may need to perform technical audits to verify that their AI models function as intended. This includes algorithm performance, data accuracy, and standards compliance.
    - AI verify toolkit: As part of the Singapore AI initiative, they launched the AI Verify toolkit, which provides a structured approach for organizations to test their AI systems (AI Verify Foundation, 2023).

Singapore's Model AI Governance Framework establishes comprehensive governance requirements to promote responsible AI deployment. By observing the outlined governance principles, conducting thorough risk assessments, maintaining rigorous documentation, and utilizing tools like the AI Verify toolkit, organizations can ensure that their AI systems operate ethically and effectively within a robust governance structure. However, the Model AI Governance Framework does not provide in-depth guidance to ensure that systems operate per these governance principles throughout their lifecycle. In contrast, the AI compliance audit framework proposed in this research goes much broader than just governance of the AI systems, as it considers other items, such as:

- Integrating the compliance audit functionality with relevant existing frameworks to get quicker organization and employee acceptance of the additional requirements.
- National and international regulations and laws alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.
- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organizational accepted core governance structures, including regulations, laws, standards, and generally accepted principles.
- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.
- Safety of employee's, public, and infrastructure.
- Accuracy, repeatability, and traceability of decisions made.
- Skills and training of employees to ensure they perform their jobs safely and thoroughly.
- Continuous system validation against a known dataset and baseline.
- Alignment to organization strategy and deliverables.
- Mechanism to ensure that AI systems being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.

- Auditing of documentation and reporting for AI systems to ensure that all operational information, changes, and risks are properly identified and captured, not just focusing on creating the documentation. The model framework covers documentation but does not consider how to audit the documentation and records to ensure quality and accuracy.
- Provides a dual auditing procedure with internal and third-party audits to drive accountability.

6.6 GAO AI Accountability Framework

The United States Government Accountability Office (GAO) plays a crucial role in ensuring accountability, responsibility, and transparency in the use of AI in government programs, processes and in federal financial management through its auditing framework. The GAO AI accountability framework was explicitly developed for federal agencies, providing guidelines on ensuring that AI systems comply with existing financial management regulations and ethical standards. The framework highlights the need for accountability mechanisms, including third-party assessments and audits, to foster trust in AI technologies. GAO developed this framework through a collective forum of AI experts across the federal government, industry, and nonprofit sectors. This framework has a core structure made up of four critical interdependent components (GAO, 2021):

1. Governance - focused on promoting accountability through the development of processes to manage, operate, and oversee the implementation of AI systems.
2. Data - focused on ensuring the quality, reliability and representativeness of data sources and data processing for the AI system.
3. Performance - focused on ensuring that the AI system produces results consistent with the program or federal objectives.

4. Monitoring – focused on ensuring reliability and relevance of the system and its outcomes over time.

As AI development continues transforming in various sectors, including healthcare, transportation, and defence within the federal government, establishing robust accountability and oversight practices is seen as essential for safeguarding public interests and promoting ethical AI use. The GAO's AI accountability framework integrates rigorous auditing standards that are essential for maintaining public trust in government financial reporting. The emphasis on quality management, leadership accountability, and adherence to contemporary auditing standards reflects a commitment to transparency and effective governance in federal financial management. In short, the GAO's AI Accountability Framework provides a structured approach for federal agencies to implement AI responsibly, ensuring that these powerful technologies are used in ways that are transparent, accountable, and aligned with public values (Bignami, 2022).

The GAO AI accountability framework provides a comprehensive governance and oversight framework for the government, but its core focuses is on financial reporting and alignment with financial regulations. In contrast, the AI compliance audit framework proposed in this research goes much broader and deeper than that proposed in the GAO AI accountability framework, as it considers other vital items, such as:

- AI compliance for the critical infrastructure sector, both public and private.
- Integrating the compliance audit functionality with relevant existing frameworks to get quicker organization and employee acceptance of the additional requirements.

- National and international regulations and laws alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.
- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organizational accepted core governance structures, including regulations, laws, standards, and generally accepted principles. This mechanism focuses on more than just financial governance and regulations compared to the GAO framework.
- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.
- Safety of employee's, public, and infrastructure.
- Human involvement in AI system oversight, dependent on system risk profile and autonomy. This goes beyond purely having supervision but focuses on intervention where required
- Skills and training of employees to ensure they can perform their jobs safely and thoroughly.
- Continuous system validation against a known dataset and baseline.
- Alignment to organization strategy and deliverables.
- Mechanism to ensure that AI systems being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.

- Auditing of documentation and reporting for AI systems to ensure that all operational information, changes, and risks are properly identified and captured, not just focusing on creating the documentation.
- Provides a dual auditing procedure with internal and third-party audits to drive accountability.

6.7 COBIT Framework

The COBIT (Control Objectives for Information and Related Technologies) framework, developed by ISACA, provides a structured approach to governance and management of enterprise information technology, including AI systems (ISACA, 2018). As organizations increasingly adopt AI technologies, effective auditing mechanisms are critical to ensure compliance, risk management, and alignment with business objectives. Applying COBIT for auditing AI provides a structured approach to managing the complexities and risks associated with AI technologies (ISACA, 2024). By leveraging this framework, organizations can ensure that their AI initiatives are practical but also ethical and compliant with regulatory standards. The COBIT framework focuses on six main pillars in its implementation (Synergist Technology, 2024):

1. Alignment with Business Objectives – one of the fundamental requirements in auditing AI within the COBIT framework is ensuring that AI systems design and outcome align with the overall organizational strategy. This involves:
    - Defining an AI strategy - outlining the AI initiative's objectives and ensure they are integrated with the broader organizational goals.
    - Stakeholder engagement – identification and involvement of relevant stakeholders to understand their needs and expectations from AI initiatives.

2. Governance and Management Objectives - establishing governance structures tailored for AI systems and not just using structures for information technology systems, these include:
    - Defining governance objectives - organizations should identify and adopt specific governance objectives related to AI.
    - Role Clarity –create clear roles and responsibilities using tools such as RACI matrices to create a clear structure.
3. Risk Assessment and Control – to effectively audit AI systems, a thorough assessment of risks associated with the system implementation is required. Key steps include:
    - Identifying Risks - evaluate potential risks related to data integrity, algorithmic bias, and compliance with regulatory frameworks such as the GDPR.
    - Control Mechanisms – implementing controls to mitigate identified risks to ensure safe AI system integration.
4. Continuous Monitoring and Improvement –promote ongoing oversight of AI systems to ensure they remain compliant and effective, this involves:
    - Monitoring Processes –establishment of processes and procedures for continuously monitoring AI systems.
    - Performance Evaluation - regular AI system assessments against performance metrics will identify areas for improvement in governance practices.
5. Transparency and Accountability – it is imperative that a structure be created that provides transparency of the AI system use and decisions, along with a clear view of who is accountable within the AI system process, this should include:
    - Clear Communication - maintain open lines of communication regarding AI development and usage to foster accountability with all stakeholders.

- Documentation - detailed records of AI processes, decisions, and audits should always be kept, promoting accountability and facilitate compliance checks.

6. Compliance with External Standards - organizations must ensure their AI systems comply with relevant legal and regulatory standards. This includes:
    - Regular Compliance Audits – carrying out audits focused on compliance with laws such as GDPR ensures that data privacy is upheld during AI operations.
    - Adapting to New Regulations - organizations should be prepared to adjust their auditing practices accordingly as regulations evolve.

The COBIT framework provides a robust structure for auditing information technology systems, with a carve-out for AI systems, emphasizing alignment with business objectives, effective governance, risk management, continuous monitoring, transparency, and compliance. The most significant shortfall of this framework for compliance auditing for AI is that it focuses on governance more in an information technology perspective than in the complex AI environment. In contrast, the AI compliance audit framework proposed in this research focuses much broader than just information technology and data management in that it considers other vital items, such as:

- National and international regulations and laws alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.
- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organizational accepted core governance structures, including regulations,

laws, standards, and generally accepted principles. The focus is broader than just the alignment of information technology and data management regulations shown in COBIT.

- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.
- Safety of employee's, public, and infrastructure.
- Accuracy, repeatability, and traceability of decisions made
- Human involvement in AI system oversight, dependent on system risk profile and autonomy.
- Skills and training of employees to ensure they can perform their jobs safely and thoroughly.
- Continuous system validation against a known dataset and baseline.
- Alignment to organization strategy and deliverables, broader than just from an information technology perspective.
- Mechanism to ensure that AI systems being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.
- Auditing of documentation and reporting for AI systems to ensure that all operational information, changes, and risks are properly identified and captured, not just focusing on creating the documentation.
- Provides a dual auditing procedure with internal and third-party audits to drive accountability.

6.8 European Data Protection Board AI Auditing Checklist

The European Data Protection Board (EDPB) published a comprehensive checklist for auditing AI systems, developed by expert Dr Gemma Galdon Clavell, to assess compliance with the GDPR and the European Union's AI Act. This checklist provides a structured methodology for conducting end-to-end audits of AI systems from a socio-technical perspective (Clavell, 2023). The aim of the EDPB's AI Audit is to assess the impact of AI systems on data protection and ensure that these systems adhere to legal and ethical standards throughout their lifecycle. The key objectives of the EDPB AI auditing checklist are to:

- Help organizations and regulators to understand and evaluate data protection safeguards in AI systems.
- Provide a framework for Data Protection Auditors to inspect AI systems and assess their GDPR and EU AI Act compliance.
- Address potential biases in AI algorithms, datasets, and AI systems outcomes.
- Enhance accountability by promoting transparency and responsibility among developers and deployers of AI systems throughout the AI system lifecycle.
- Ensure fairness by promoting equitable treatment of all data subjects affected by AI systems to minimize discrimination or skewed outcomes.

The EDPB's AI auditing checklist outlines several essential elements that must be addressed during any AI audit (Lisievici, 2024):

1. The compilation and collation of information on the AI system's training and testing. It should include items such as data protection impact assessments, data sharing agreements, and relevant data protection approvals.

2. A system map illustrating the relationships between the algorithm, the technical system, and the decision-making process.
3. Identify potential bias sources and instances within the AI system, as well as conducting analyses to evaluate the impact of different biases on the demographics of individuals, groups, and society and the efficient operation of the AI system.
4. Adversarial audits to challenge the system's robustness and identify any vulnerabilities under real-world conditions to uncover vulnerabilities.
5. Generate comprehensive audit reports that document findings, proposed mitigation measures, and recommendations for ongoing compliance.

The AI auditing checklist emphasizes an end-to-end socio-technical methodology for auditing AI systems, which recognizes that algorithmic systems operate within complex social contexts and interact with diverse data sources. The EDPB's AI auditing checklist represents a significant advancement in ensuring that AI technologies align with fundamental rights and freedoms under EU law. By establishing precise requirements for auditing AI systems, the EDPB aims to enhance transparency, accountability, and fairness in AI deployment, ultimately fostering public trust in these technologies.

The EDPB AI auditing checklist is predominantly focused on handling of data and data protection under the EU law. This lack of focus on the broader functionality of AI system functionality does not provide a adequate compliance overview of the AI system. In contrast, the AI compliance audit framework proposed in this research goes much broader than just focusing on data, data protection, and EU regulations in that it considers other vital items, such as:

- Integrating the compliance audit functionality with relevant existing frameworks to get quicker organization and employee acceptance of the additional requirements.
- National and international regulations and laws alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.
- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organizational accepted core governance structures, including regulations, laws, standards, and generally accepted principles.
- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.
- Safety of employee's, public, and infrastructure.
- Accuracy, repeatability, and traceability of decisions made
- Human involvement in AI system oversight, dependent on system risk profile and autonomy.
- Skills and training of employees to ensure they can perform their jobs safely and thoroughly.
- Continuous system validation against a known dataset and baseline.
- Alignment to organization strategy and deliverables.

- Mechanism to ensure that AI systems being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.
- Auditing of documentation and reporting for AI systems to ensure that all operational information, changes, and risks are properly identified and captured, not just focusing on creating the documentation.
- Provides a dual auditing procedure with internal and third-party audits to drive accountability.

6.9 ISO/IEC 42001

The International Organization for Standardization has developed a number of Standards that pertain to information and operational technology data and system management, which includes basic AI requirements as they have matured. It is only recently that they developed and published the first international standard, ISO/IEC 42001:2023, which establishes guidelines for establishing an AI specific management system within organizations. This standard provides a structured approach for organizations to manage AI systems throughout their lifecycle, emphasizing integrating AI management systems with existing organizational processes and ensuring that AI technologies are developed and used responsibly (Coglianese and Crum, 2024). It establishes a comprehensive framework to address the unique challenges posed by AI, such as ethical considerations, transparency, and the need for continuous improvement in AI practices (ISO, 2023).

The primary aim of an ISO 42001 audit is to evaluate an organization's adherence to the standards for responsible AI management. This involves assessing both the AI management systems' conformity to established requirements and its implementation effectiveness (Barr Advisory, 2024). Auditors systematically review processes,

documentation, and practices to ensure alignment with organizational objectives while managing AI-related risks and opportunities. The critical components of the audit process within ISO 42001 (Benraouane, 2024) are:

1. Planning - establishing a structured approach for the audit, which includes:
    - Defining and establishing the audit scope, objectives, and critical criteria based on the control objectives outlined in Annex A of ISO 42001.
    - Resource allocation for conducting the audit effectively and efficiently.
    - Outline how relevant evidence will be collected, by whom, and how the collection process will be recorded.
    - Plan how the evidence will be evaluated, and by whom, and how results will be documented and reported.
    - Conducting AI system assessments, which includes monitoring, measurement, analysis, and evaluation of AI management system performance.
2. Collection of evidence - organizations must collect and verify evidence for the AI system, according to the planned collection process, which will include:
    - Systematic reviews of the AI management system documentation.
    - Interviews with employees to assess their competence and understanding of AI roles and responsibilities.
    - Observations of AI system operations to validate findings from interviews and documentation reviews.
3. Assessment criteria - auditors will evaluate the systems as per the established audit plan; some critical criteria of the audit include (Bufe, 2024):
    - AI risk assessment – entails identifying and evaluating the identified risks which may prevent the organization from achieving the AI strategy objectives.

- AI risk treatment plan – evaluating if the necessary controls to prevent or mitigate the assessed risks have been implemented and are effective.
- AI impact assessment –understanding the potential broader implications for individuals, groups, and societies that could materialize due to AI systems.
- Statement of application – documentation of all necessary controls to prevent or mitigate the risks identified in achieving the AI strategy objectives.

ISO/IEC 42001 audits play a vital role in promoting responsible AI system management by ensuring organizations adhere to ethical standards and effectively manage risks associated with AI technologies. By establishing a structured approach to auditing AI management systems, organizations can demonstrate their commitment to ethical practices and enhance stakeholder trust in their AI systems. It is important to note that this standard is not a compliance audit platform, but rather a standard for guiding how a compliance audit platform should be established and managed. This standard has some shortcomings in that it focuses more on alignment with standards than regulations and governing law. In contrast, the AI compliance audit framework proposed in this research goes much deeper than just providing a standard to comply with, it actually develops an AI management system, as outlined in the ISO standard, with a broader scope including additional vital items, such as:

- Integrating the compliance audit functionality with relevant existing frameworks to get quicker organization and employee acceptance of the additional requirements.
- National and international regulations and laws alignment, not just standard alignment, dependent on the source of the AI system and the placement of the organization developing and using the system.

- Pre- and post-AI system baselining for data security, cybersecurity, and operability.
- Lifecycle auditing mechanism to ensure that from scoping to AI system retirement, the system is developed, trained, deployed, and used per the organizational accepted core governance structures, which includes regulations, laws, standards (such as ISO 42001), and generally accepted principles.
- Ongoing lifecycle risk-based compliance framework and governance structure alignment review.
- Integrated organizational information technology, operational technology, and AI system holistic performance checks.
- Safety of employee's, public, and infrastructure.
- Accuracy, repeatability, and traceability of decisions made
- Human involvement in AI system oversight, dependent on system risk profile and autonomy.
- Skills and training of employees to ensure they can perform their jobs safely and thoroughly.
- Continuous system validation against a known dataset and baseline.
- Mechanism to ensure that AI systems being uninstalled are safely removed and that the system is properly restored to its previous operational baseline.
- Auditing of documentation and reporting for AI systems to ensure that all operational information, changes, and risks are properly identified and captured, not just focusing on creating the documentation.
- Provides a dual auditing procedure with internal and third-party audits to drive accountability.

CHAPTER VII:

SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

7.1 Summary

As the electricity sector has accelerated its digitization and modernization process over the past two decades, it has created a complex data-dependent industry requiring accurate, trustworthy large quantities of data to facilitate system operations and guide decision-making. To achieve this, the electricity sector, as with other industries, has adopted intelligent tools, such as AI systems, to process large quantities of data and make real-time decisions. As AI systems have matured and evolved in this sector, it has become even more critical to better understand the safe and sustainable implementation parameters and guardrails required for AI systems within the sector. The research for this thesis and the area of knowledge growth is focused on developing a fit-for-purpose compliance framework that will provide the guardrails needed to ensure that AI systems are safely implemented within the electricity sector. The driving factor for the research is that within the electricity sector, incorrect decisions influence more than financial returns but can cause severe equipment damage, premature failure, and harm people.

A review of literature on AI systems, their governance, and compliance mechanisms within the electricity sector, made it possible to identify gaps, opportunities, and risks within the existing structures and research. Through the review of the relevant literature identified for this subject, it was concluded that researchers agree that AI systems require standardization and mechanisms established to facilitate rules, guidance, and control on how they are developed, deployed, implemented, and used (Ayling and Chapman, 2022). Prior research supports that AI systems need to have a level of auditing or compliance checks (Roberts *et al.*, 2022) in place to ensure that they are developed and operated as per the prevailing regulations, governance, or legislation. However, there is no

consensus on whether this should be done at an organizational, national, or international level. Furthermore, there is no consensus between governments, academia, and organizations on whether or how humans should play a role in ensuring that AI systems are developed, deployed, and operated safely (Priya *et al.*, n.d.). There is an ongoing discourse on whether humans can even oversee AI system operations and check their outputs as AI systems become more complex and operate without clear visibility or understanding of how the system processes a decision. Researchers further note that as AI becomes autonomous, it may be capable of restructuring its own code, which will make it very difficult to oversee unless there are structured guardrails established to guide its operation and to empower its human collaborators to understand its decisions and be able to make an informed decision on the outcome.

This rapid maturing of AI systems has driven a flurry of research and development of principles, regulations, and policies and has gotten governments, academia, and private organizations focused on developing AI compliance audit methodologies. Several mature regulations, policies, white papers, and frameworks outline compliance mechanisms for AI systems, but the majority are auditing only partial components within the AI system, such as the data, the learning process, or the algorithms. There is no apparent convergence between different governments, academia, and organizations on what and how AI systems should be audited to ensure that the systems operate safely, morally and that they are human-centric.

To better understand the level of knowledge within the electricity sector on AI systems, a structured survey was undertaken with subject matter experts within the electricity utilities, regulators, information technology providers, to the electricity sector, and AI software developers to gain a better understanding of what governance, regulatory and oversight protocols already exist. This survey focused on identifying knowledge gaps

in the industry regarding the safe and sustainable implementation of AI from an organizational, national, and international perspective to guide the proposed framework.

It was inspiring to realize the level of knowledge on the existing AI regulations and governance structures in place to govern AI systems in the electricity industry but sobering to understand the overarching feeling that they were either not appropriately applied or insufficient to protect the organizations, employees, infrastructure and the public. There was a clear need indicated by the participants for a comprehensive AI regulatory and oversight compliance framework to be developed for the electricity sector, and support from the parties that the most appropriate mechanism to develop, implement, and maintain this would be through a collaborative approach between the government, AI and information technology fraternity and the electricity sector.

The participants further supported the adoption of a tiered approach for regulation, oversight compliance, and human oversight or involvement for the different maturity levels of AI systems using a risk-based methodology to ensure their effectiveness. Their guidance was that the industry did not need more stringent regulations and oversight implemented as AI levels of autonomy increased; they needed a balanced approach that considered innovation, accountability, and level of autonomy in decision-making in setting the tiers. They also stressed that for an AI compliance framework to be beneficial and adopted, the AI compliance audit process should be included in relevant existing processes and frameworks within the organization.

The insights and professional sentiments from the participants provided invaluable guidance to developing the AI compliance auditing development procedure, framework, and audit process, which is a fit-for-purpose compliance solution that can be adapted to all critical infrastructure sectors to ensure that AI systems are safely and sustainably developed, deployed and used to support the current and future industry. By adopting the

proposed risk-based approach to ensuring that the AI compliance audit framework is always aligned with the latest governing regulations, laws, and principles for AI systems and the industry, the electricity sector can ensure compliance of the AI system and the safety of their organization, employees, the infrastructure and the public.

In summary, the outcome of this research is a unique AI compliance auditing procedure, framework, and process that is driven from a senior executive management level, aligned and integrated to specific existing compliance and governance processes within the electricity sector, which is more advanced than the existing AI governance frameworks, audit mechanisms, and checklists.

7.2 Implications

The proposed compliance framework provides a unique alternative to existing compliance frameworks, which will impact how systems are designed, deployed, used, and retired. This will influence the mechanisms that developers, end-users, and maintenance organizations implement to perform checks and balances in their dealings with AI systems. The AI compliance framework mandates procedures that end-user organizations procuring AI systems, outsourcing the development of AI systems, or developing the AI system in-house need to follow to begin ensuring that the AI systems are compliant from the system conception stage and not only when the system is deployed within their organization. This has major impact on the AI developer, the end-user organization and independent auditors as summarized below:

- **Implications to AI Developers** – under the proposed AI compliance framework, the AI developer will be required to subject their AI system development and training processes and procedures to continuous risk auditing, which will be used to establish

the specific level and frequency of compliance audits during the development stages. Other specific areas of impact to the AI developer are:

- Systems being developed and trained must comply with specific national and international regulations, laws, standards, and generally accepted principles per the end-user's organizationally accepted core governance structures. They will also be required to adopt any changes to these governance structures should the risk audit indicate that changes or new adoptions are required.
- The developer will need to furnish training registers for their employees to confirm that the team has the necessary skills, certification, and training to perform their jobs safely and thoroughly. Periodically, they will need to comply with an audit from the end-user organization or their appointed audit team to ensure that ongoing training and re-certification is occurring.
- The end-user organization will require the AI developer to ensure that the AI system is scoped, designed, and developed in full alignment with the organization's AI strategy, business strategy, and key deliverables. This ensures the systems value add and alignment with the core values for the organization.
- The developer will need to follow a documentation and reporting regime throughout the AI system scoping, design, development, and training. The end-user will audit this to identify functionality, design, and risk compliance alignment. This will also allow the end-user organization to track and trace any changes that may have been implemented outside of the original scope.
- The developer will need to develop a structured data management and reporting procedure to facilitate the sourcing, handling, verification, and disposal of datasets used in the development and training of the AI system. These records will need to be transferred to the end-user upon deployment of the AI system.

- Finally, under the AI compliance audit framework, the developer will be required to furnish a full handover portfolio to the end-user, outlining the development cycle compliance process, verification process and documentation, which the organization will use to confirm compliance before fully deploying within its organization.

- **Implications to End-User Organization and information/operational technology service provider** - under the proposed AI compliance framework, the end-user and their information or operational technology service provider will be required to align their auditing processes, procedures, and operational regime to include the continuous risk auditing procedure required under the framework. The specific areas of impact to the end-user and information or operational technology service providers are:
  - Senior management must take ownership of the AI compliance procedure in providing leadership and sponsorship for the framework and process. From a leadership change management perspective, it is imperative that this process be facilitated from a senior management or Board perspective to get the required buy-in from the organizational teams.
  - The organization will be required to establish an AI strategy and align it with its organizational strategy and deliverables. This document needs to be a live document which is updated continuously, and owned by the senior management sponsor, to guide the development and maintenance of any AI systems.
  - The organization will be required to develop a procedure for onboarding any new or updated AI systems, which will include checks to be performed on the system compatibility, functionality, design, compliance to core governance structures and alignment to the organizational strategies. This onboarding

procedure should include mechanisms for socializing the AI system with the employees and the probationary operation of the AI system during this period.

- The end-user will need to adopt the AI compliance framework development procedure and use it to develop their organizational specific AI compliance audit framework. This will include undertaking the necessary risk assessments for the AI system to be deployed to guide the identification and adoption of the core regulations, laws, standards, and generally accepted principles that will be used to govern the AI system operations. The ongoing lifecycle risk-based compliance framework and governance structure alignment review will be integrated into the organization through this process.
- The finalized framework will need to be integrated into an existing compliance or audit framework, such as ISO9001, within the organization, and the employees will need to be trained to perform the updated audit requirements.
- A procedure will need to be established and implemented to undertake a baseline study for the information and operational technology platforms' data security, cybersecurity, and operability before and after an AI system is deployed. This should include ongoing performance checks for the holistic organizational information technology, operational technology, and AI system post-deployment.
- A policy should be established for human oversight or involvement in AI system decision-making and functioning. The level of human involvement in AI system oversight will be defined by the system risk profile and autonomy level as per the ongoing risk assessments. The policy will guide what training is required for people undertaking oversight, how they will perform oversight, and how the involvement or interventions will be recorded and reported.

- The organization is required to establish a protocol or procedure for dealing with AI systems that are being retired or uninstalled. This protocol needs to ensure that AI systems are safely removed, and that the full information and operational systems are properly restored to its previous operational baseline to prevent any data protection or security risks.
- The organization will need to develop an internal audit team with supportive information and operational technology specialists capable of building the necessary checks and audit structures and undertaking the ongoing audits for the AI systems. Lastly, they will need to establish a training regime for the auditors to keep abreast of the continuously changing environment and ensure that they are capable of meeting the organizational needs.
- The end-user will need to develop a procedure to guide how they will deal with an AI system when a severe risk or material deficiency is identified. This should overview whether the system will be disabled or un-installed until corrective actions can be made. If the deficiency is through an employee deficiency, what training or certification process will be followed to ensure their capability to perform their task. Or if it is a data security risk, or cyber security breach, how will this be managed, reported and who will be informed.

- **Implications to third-party auditors** – independent auditors that provide auditing services to the electricity sector and the information technology services organizations will need to build up the necessary training regime, skillsets, and capabilities to carry out the audits per the AI compliance audit framework and procedures established by the industry.

7.3 Recommendations for Future Research

This research provides a foundational structure for future researchers to expand on, focusing on gaps within the current governance structures, training regimes, and approach to building a sustainable AI system environment. The industry knowledge on AI system development, operations and maintenance, AI governance and AI compliance, still has a long way to go and will always have gaps due to the rapid growth of the technology. Some key future research areas that can empower organizations and governments to make more informed decisions are:

- *Optimized governance principles study:* structuring a methodology for choosing the most appropriate governing principles to guide the safe and sustainable development and operation of AI systems in different organizations and sectors. It should include a guideline on how organizations should perform continuous improvement reviews on the governing principles to ensure that the latest, most up-to-date governing principles are used to protect the organization.
- *Regulations and law standardization study:* undertake a study to identify how global standardization could be achieved for AI regulations, legislation and principles (Manheim *et al.*, 2024). What would it take to align the dispersed governments and lawmakers to develop a global collaboration to mitigate bias, discrimination, and AI system disparity between organizations, countries, and communities.
- *AI system development framework:* All AI developers follow their in-house processes, standards and procedures to develop AI systems, which makes it relatively difficult to manage compliance. Further research is recommended on mechanisms and platforms that can be established to standardize the development of AI systems and align them to a national or international

minimum standard. Considerations would be to create international baseline development platforms that can be shared with all developers to build their systems on, thereby creating a known AI system.

- *Unified design approach:* As a slight diversification or even extension to the previous item, further research is required to create a central control body, rather than just a central framework, either nationally or internationally, that can facilitate a collaborative, unified approach to AI system scoping, development, and training. By having a central control body, there is an opportunity for AI systems to be designed and developed faster, with less risk of bias, discrimination, or other detractors, as a biodiverse stakeholder grouping will provide input to the system from the inception phase.
- *Defining AI functionality and operational risk profiles:* AI systems are either being risk classified once they are built, and the system's complexity is ascertained or classified only once something goes wrong. The recommendation is that research should be undertaken on how best to define AI systems against functionality, design, and other vital attributes so that the developer and user know the system's risk profile upfront, making it easier to classify them for management and compliance.
- *Compliance skills development:* Many of the regulations, standards, compliance audit frameworks, and laws, including the one proposed in this research, propose the inclusion of human oversight or collaboration and training for the operators and the auditors. A great deal more research is required to properly determine what the relevant training requirements should be for this function, what ongoing certification would be required, and whether additional training

and certification would be required as the systems become more complex, more intertwined with other AI systems or becomes autonomous.

- *Information and Operation Technology impact:* Many sectors operate information and operational technology systems, which they continuously strive to keep independent to ensure that cybersecurity attacks or malicious use of the information technology system cannot impact the operational technology system. However, with the introduction of intelligent systems, such as AI, these systems have fast become intertwined. Additional research is required on how to manage AI systems being introduced into the information technology or the operational technology system, to ensure that they do not create cross system risks. This includes looking at the compliance requirements to detect and manage these risks in a fully interconnected system without compromising either platform.
- *Data Protection and cybersecurity management:* Additional research is required to ascertain how data management systems and cybersecurity systems need to morph as AI systems become more complex and autonomous. One of the risks identified during this research was that as more AI systems are integrated with other AI systems, information, and operational technology systems, it is becoming more challenging to track what is happening, how information flows, and whether there are risks. So, the question to answer here is how do we protect highly dynamic systems from cyber risks and data breaches? Do we need fit-for-purpose solutions to be developed, can we use AI solutions to protect integrated platforms, or will that add in additional risks?

The development and implementation of AI solutions is such a dynamic space, with new AI algorithms and models being developed at an astronomical pace. The benefits, opportunities, risks, and impact are still broadly unknown and will evolve as systems mature and become more integrated with other AI systems, information and operational technology systems. As long as this technology is so volatile and dynamic, there will be no shortage of areas for investigation. The proposed future investigation area's above only touches on one small portion of research that is required for humans to understand and accept the full potential and risks of AI solutions.

7.4 Conclusion

In conclusion, this research provided a unique AI compliance auditing development procedure, framework, and audit process, which is a fit-for-purpose compliance solution for the electricity sector. This significantly contributes to the industry and sector by providing a practical structure that can be utilized to safely and sustainably implement AI systems in this critical sector. The AI compliance audit framework is a mechanism that allows the electricity sector to place entire lifecycle-focused guardrails around AI systems they are procuring or developing. The framework considers a multi-dimensional audit regime, which can be integrated seamlessly into existing quality, information technology, or environmental assurance processes. Notably, the framework ensures that the electricity sector considers the integrated software systems as a collective when auditing and ensuring compliance, not as individual components.

APPENDIX A

COVER LETTER AND QUESTIONNAIRE

COVER LETTER

Dear Participant

I am in the process of studying towards my Executive Doctor of Business Administration. The focus of the research and the area of knowledge growth centres on the safe and sustainable introduction and use of Artificial Intelligence within the energy/electricity sector. The objective of this research is to undertake a regulatory, governance and ethical impact assessment of Artificial Intelligence in the energy sector as it evolves and matures, and to develop a compliance framework for overseeing the implementation of Artificial Intelligence technologies, encompassing alignment to ethical, morale, safety, human oversight and regulatory aspects over its lifecycle, as it progresses from an entry-level administrative assistant to a fully autonomous decision maker, in the energy/electricity sector.

I am reaching out to subject matter experts in electricity, electricity regulation and Artificial Intelligence system development, to request their participation in a survey to assist in gaining a better understanding of the current Artificial Intelligence regulatory and compliance oversight landscape in the energy/electricity sector and to identify any gaps. I kindly request your support in sharing this survey broader with your relevant team members for them to share their knowledge and thoughts on the subject.

Your assistance is greatly appreciated. I look forward to hearing back from you.

QUESTIONNAIRE

1) Region

a) Americas and Caribbean
b) East Asia & Pacific
c) Eastern Europe and Central Asia
d) Middle East and North Africa
e) South Asia
f) Sub-Saharan Africa
g) Western Europe

2) What industry do you work for

a) Electricity
b) Information Technology
c) Artificial Intelligence Developer
d) Regulator/Legislator
e) None of the above

3) Job Description

a) Executive/Director
b) Manager
c) Project Manager
d) Engineer
e) Information Technologist
f) AI Developer
g) Legal/Regulatory

h) Legislator
i) Other – Specify

4) Age Group

a) 18 - 24
b) 25 - 34
c) 35 - 44
d) 45 - 54
e) > 54

5) Are you familiar with the use of Artificial Intelligence in the electricity sector? (Select all that apply)

a) I am familiar with the use of AI in the electricity sector
b) I have personally worked with AI in the electricity sector
c) I have heard of AI being used in the electricity sector, but I am not very familiar with it
d) I am familiar with AI, but not how it is specifically used in the electricity sector
e) I have some knowledge about AI being used in the electricity sector from articles, media or training
f) I have heard of AI being used in other industries, but I am not sure if it is used in the electricity sector
g) No, I am not familiar with the use of AI in the electricity sector

6) Which organizations or agencies do you believe are responsible for setting and enforcing governance and compliance protocols for AI in the electricity sector? (Select all that apply)

a) Energy/Electricity Regulator
b) The National Department of Energy
c) Utility industry associations
d) Government agencies at the state level (e.g. Department of Energy and Environment)
e) Institute of Standards and Technology
f) Independent organizations specializing in AI governance and compliance
g) International organizations (e.g. International Energy Agency)
h) Local governments and municipalities
i) Other

7) Are you aware of any specific regulations or laws that govern the use of AI in the electricity sector? (Select 1)

a) Yes, I am aware of specific regulations for AI in the electricity sector
b) I am not sure about the regulations for AI in the electricity sector
c) There are regulations, but I am not familiar with the specifics
d) As far as I know, there are no regulations for AI in the electricity sector
e) No, I am not aware of any regulations for AI in the electricity sector

8) Are there any current efforts being made by electricity utility companies or government agencies to regulate the use of AI in the electricity sector? (Select all that apply)

a) Yes, there are current efforts from both electricity utility companies and government agencies
b) Some electricity utility companies have implemented their own governance and compliance protocols for AI
c) Government agencies have proposed regulations for AI use in the electricity sector, but they are not yet in effect
d) The industry is currently in discussions about potential regulations for AI in the electricity sector
e) There have been calls for stricter regulations for AI in the electricity sector
f) Some electricity utility companies have started using AI but have not implemented any governance or compliance protocols yet
g) Both electricity utility companies and government agencies are actively collaborating to establish regulations for AI use
h) No, there are currently no efforts being made

9) Do you think there is a need for a comprehensive regulatory and oversight framework to govern the implementation of AI in the electricity sector? (Select 1)

a) Yes, a comprehensive framework is necessary to ensure proper regulation and oversight for the ethical and safe implementation of AI in the electricity sector
b) No, the electricity sector should be left to implement AI as it sees fit so as not to stifle innovation

10) Do you believe the potential risks associated with AI in electricity sector are being adequately addressed by existing oversight measures? (Select 1)

a) Yes, I believe the current oversight measures are sufficient
b) I'm unsure, more research needs to be done on the possible risks
c) It could be improved, but overall, I think it is being addressed adequately
d) I believe more collaboration between stakeholders is needed for effective oversight
e) There should be specific regulations in place for different levels of AI maturity
f) There should be a designated agency responsible for overseeing AI implementation in utilities
g) No, I think there should be stricter oversight in place

11) What are the key considerations that must be addressed in a comprehensive regulatory and oversight framework for AI in the electricity sector? (Select all that apply)

a) Ethical considerations surrounding the use of AI in electricity utilities
b) Ensuring transparency and accountability in the implementation of AI technology
c) Addressing potential risks and unintended consequences of AI in the electricity sector
d) Incorporating a holistic approach to regulation that covers different levels of AI maturity
e) Collaboration between government agencies, utility companies, and AI experts in developing the framework
f) Incorporating regular audits and assessments to ensure compliance with regulations
g) Developing guidelines for data handling and protection in the use of AI

h) Addressing potential job displacement and retraining programs for affected workers
i) Taking into consideration the potential impact of AI on consumer privacy and data rights
j) Establishing standards for data quality and bias mitigation in AI algorithms

12) Are standardized regulations necessary to ensure fair competition among companies using AI in the electricity sector? (Select 1)

a) Yes, standard regulations are crucial for fair competition, to establish accountability and transparency, as well as prevent unethical use of AI in the electricity sector
b) It depends on the potential impact of AI on the efficiency and safety of the electricity sector
c) I believe a balance should be struck between standardized regulations and allowing flexibility for companies in the electricity sector
d) I am unsure, I need more information on the current use of AI in the electricity sector
e) I am not familiar with AI in the electricity sector and cannot provide an opinion on standardized regulations
f) No, different companies should be able to have their own AI regulations, so as not to hinder innovation, and to allow them to address unique challenges in the electricity sector

13) How important is it to you that companies using AI in the electricity sector follow the same rules and regulations? (Select all that apply)

a) It is very important for companies to follow standardized rules and regulations for AI in the electricity sector
b) Standardized regulations for AI in the electricity sector are necessary for consumer protection
c) Standardized requirements for AI in the electricity sector will ensure safety, reliability, fairness and transparency
d) Companies should not be allowed to have different regulations for AI use in the electricity industry

14) Are you in favour of the government implementing standardized oversight for AI in the electricity sector, or should it be left up to individual companies to regulate themselves? (Select all that apply)

a) Yes, I believe standardized oversight for AI should be implemented by the government
b) Standardized oversight for AI in the electricity sector would ensure consistency and fairness across the entire industry
c) Companies in the electricity sector should have the freedom to regulate AI use, as long as it aligns with ethical standards and guidelines
d) I believe there should be a combination of government oversight and self-regulation by companies for AI in the electricity industry
e) Companies may not prioritize ethical considerations without government oversight for AI in the electricity industry

f) Self-regulation by companies may lead to unequal levels of oversight and potentially harmful consequences in the electricity sector
g) Government involvement in AI regulation for the electricity industry should be carefully balanced to avoid hindering innovation
h) I am undecided on whether the government or companies should regulate AI in the electricity sector
i) No, I think companies in the electricity industry should be responsible for their own regulation of AI

15) Do you think there should be different regulation and oversight protocols for different levels of AI maturity in the electricity sector? (Select 1)

a) Yes, there should be a graduated approach to regulation as AI technology advances
b) I believe a tiered approach to regulation and oversight would be most effective
c) There should be a balance between regulation and allowing for innovation in all maturity levels
d) It's important to consider the unique challenges and opportunities of each maturity level
e) I think a standardized set of protocols should be applied across all AI maturity levels
f) More research and collaboration are needed to determine the best approach for oversight
g) No, all levels of AI maturity should be subject to the same regulations

16) Do you believe that current oversight protocols are sufficient to handle the potential risks associated with highly autonomous AI systems in the electricity sector? (Select 1)

a) Yes
b) Unsure
c) Possibly, more research is needed
d) More stringent oversight may be necessary for higher levels of autonomy
e) Current protocols could be improved to better address AI risks
f) The electricity sector may require unique regulations for AI systems
g) No, oversight protocols should be consistent regardless of autonomy level

17) Do you believe there should be specific oversight or audit requirements in place for ensuring compliance with these regulations and standards? (Select 1)

a) Yes, without proper oversight and audits, compliance with regulations and ethical standards cannot be guaranteed
b) It depends on the level of AI autonomy and the potential risks involved
c) It may be beneficial to have some level of oversight and audit, but the exact requirements should be carefully considered
d) Balancing innovation and accountability, oversight and audit requirements should be implemented to ensure compliance with regulations and ethical standards
e) No, oversight and audit requirements should not be necessary for compliance with regulations and ethical standards

18) Which level(s) of AI autonomy do you believe require the most rigorous oversight or auditing? (Select 1)

a) Level 5 (Full Autonomy)
b) Level 4 (High Autonomy)
c) Level 3 (Limited Autonomy)
d) Level 2 (Partial Automation)
e) Level 1 (Assistant)
f) Level 0 (No Autonomy)
g) All levels require equal oversight/auditing
h) None, as long as proper regulations and ethical standards are followed
i) Unclear, more research needed
j) All levels, but to varying degrees

19) What measures should be in place to ensure AI systems remain compliant with regulations and ethical standards throughout their lifecycle? (Select all that apply)

a) Regular audits and assessments of AI systems
b) Clear policies and guidelines for AI development and operation
c) Continuous monitoring and updates to ensure compliance
d) Training and education for developers and operators on regulations and ethics
e) Collaborating with regulators and industry experts for guidance
f) Transparency and accountability in AI decision-making processes
g) Incorporating ethical committees or review boards
h) Regular external reviews and evaluations
i) Data privacy protection measures
j) Tracking and documenting any changes made to AI algorithms

k) Compliance checks before implementation of AI systems
l) Regular communication and reporting to stakeholders
m) Implementation of risk management strategies
n) Adhering to industry-specific regulations and standards

20) Are you aware of any current or potential future regulatory changes that could impact the oversight or auditing of AI autonomy? (Select 1)

a) I am aware of potential changes to regulations regarding AI autonomy oversight
b) I am aware of recent changes to regulations related to AI autonomy oversight
c) I believe there will be updates to regulations concerning AI autonomy oversight in the near future
d) I am not aware of any current or future regulations that could impact AI autonomy oversight
e) No, I am not currently aware of any changes to regulations related to AI autonomy oversight

21) How would you rate the current level of human oversight in the implementation of Artificial Intelligence within the electricity sector? (Select 1)

a) High - Significant human oversight is required at all levels of AI maturity in the electricity sector
b) Moderate - Human oversight is needed at some levels of AI maturity, but not all
c) Low - Limited human oversight is needed as AI has advanced in the electricity sector
d) None - AI in the electricity sector is fully autonomous with no need for human oversight

e) Uncertain - I am unsure of the current level of human oversight in AI implementation within the electricity sector
f) Not applicable - I am not familiar with the use of AI in the electricity sector

22) Given the potential risks and benefits of AI in the electricity sector, how much control do you think humans should have in decision-making processes? (Select all that apply)

a) Humans should have full control at all autonomy levels
b) A higher level of human control is needed for critical decisions
c) Human oversight is important for ethical and safety considerations
d) Humans should have final veto power over AI decisions
e) A balance between human oversight and AI decision-making is necessary
f) Limited human involvement is acceptable, as long as safety measures are in place
g) Human oversight should gradually decrease at higher levels of autonomy
h) Complete automation with no human involvement is preferred

23) What factors do you think should be considered when determining the appropriate level of human oversight for AI in the electricity sector? (Select all that apply)

a) Type of AI technology being used: Depending on the type of AI technology being utilized, the level of human oversight may vary
b) Complexity of tasks performed by AI: The complexity of tasks performed by AI can impact the level of human oversight necessary
c) Potential impact on safety and security: Consideration should be given to the potential impact of AI on safety and security in the electricity sector

d) Risk level associated with the AI system: The risk level associated with the AI system should be evaluated when determining the appropriate level of human oversight
e) Level of decision-making authority of AI: The level of decision-making authority given to AI can affect the required amount of human oversight
f) Potential for errors or malfunctions: The potential for errors or malfunctions should be considered when determining the necessary level of human oversight for AI
g) Adequacy of training and testing of AI: The extent to which AI has been trained and tested may impact the level of human oversight needed
h) Legal and regulatory requirements: Adherence to legal and regulatory requirements may influence the level of human oversight required for AI in the electricity sector
i) Ethical considerations: Ethical considerations regarding the use of AI should be taken into account when determining human oversight
j) Feedback and monitoring capabilities: The level of feedback and monitoring capabilities of AI may determine the level of human oversight needed to ensure accountability

24) What are the potential consequences of relying heavily on AI at different levels of autonomy within the electricity sector? (Select all that apply)

a) Reduced human error and improved efficiency at higher levels of AI autonomy
b) Increased risk of system failures and blackouts at lower levels of human oversight
c) Potential job loss for employees who are replaced by AI
d) Higher costs for consumers due to implementation and maintenance of AI technology
e) Greater dependence on technology, leading to vulnerability to cyber-attacks
f) Improved decision-making and problem-solving capabilities at higher levels of AI maturity
g) Lack of accountability and transparency in decision-making processes
h) Potential for bias and discrimination in AI decision-making
i) Improved safety measures and risk assessment at higher levels of AI autonomy
j) Neglect of important human considerations and ethical concerns in AI development

25) What are the main challenges associated with implementing an oversight and compliance auditing framework for AI management in the electricity sector? (Select all that apply)

a) Lack of standardization and guidelines for AI management in the electricity sector
b) Limited availability and high cost of skilled personnel for carrying out audits
c) Resistance from stakeholders to adopt new AI oversight and compliance measures
d) Inadequate data sharing between different entities in the electricity sector
e) Privacy concerns and potential ethical issues related to AI use
f) Ensuring compatibility of the AI framework with existing systems and processes

g) Constantly evolving technology making it challenging to keep up with compliance requirements
h) Difficulty in quantifying the benefits and ROI of implementing AI oversight and compliance measures

26) What do you consider to be the most important benefits of having an oversight and compliance auditing framework for AI management in the electricity sector? (Select all that apply)

a) Enhanced safety and reliability of AI-powered systems in the electricity sector
b) Increased transparency and accountability in decision-making processes
c) Mitigation of potential risks and ethical concerns associated with AI use
d) Cost savings through early detection and prevention of AI failures or errors
e) Improved data governance and protection as AI systems handle sensitive information
f) Facilitation of regulatory compliance and adherence to industry standards
g) Promotion of fair and non-discriminatory use of AI in the electricity sector
h) Identification of areas for optimization and efficiency improvements through auditing
i) Strengthening of consumer trust and confidence in AI-powered services
j) Effective management of potential biases and unintended consequences of AI implementations

27) What potential risks or drawbacks do you see in implementing an oversight and compliance auditing framework for AI management in the electricity sector? (Select all that apply)

a) High cost of implementation
b) Difficulty in hiring qualified auditors
c) Resistance to change from current practices
d) Lack of clear guidelines or standards
e) Data privacy concerns
f) Complexity of integrating AI with existing systems
g) Lack of understanding or knowledge of AI technology
h) Compliance burden for smaller companies
i) Slow adoption and implementation process

28) Do you believe that integrating Artificial Intelligence lifecycle oversight and compliance protocols into existing governance processes would be beneficial for the electricity sector? (Select 1)

a) Yes, it would increase efficiency, accountability, transparency and trust
b) No, it would create unnecessary red tape and delays, while hindering innovation and progress
c) Potentially, but it would need to be carefully managed and integrated against specific protocols
d) I'm not sure, I would need more information
e) Not necessarily, there may be other methods for ensuring compliance

29) Would you support the inclusion of Artificial Intelligence lifecycle oversight in the existing governance processes, even if it may involve additional costs? (Select 1)

a) Yes, it could be beneficial if implemented properly
b) Depends on the potential benefits
c) Not sure, need more information
d) Only if the costs are reasonable
e) It may be necessary for the future of the electricity sector
f) I trust existing governance processes and do not see the need for additional oversight
g) I am open to considering it as long as it does not significantly impact costs
h) No, AI needs to be governed independent to other processes

APPENDIX B

INFORMED CONSENT

Consent Form for participation in research

Research project title: Oversight methodologies for Artificial Intelligence implementation in the energy sector.

Research investigator: Cedric Alwyn Worthmann

Phone Number: +1 345 936 3419

Mail: cedric@ssbm.ch

Introduction

You are being invited to participate in a research study about the safe and sustainable introduction and use of Artificial Intelligence within the energy/electricity sector. You have been identified as a subject matter expert in electricity, electricity regulation and Artificial Intelligence system development through engagements with relevant industry organizations and associations.

Before you decide whether you wish to participate, it is important for you to understand why the research is being conducted and what it will involve. Please take the time to review the information below to make an informed decision to participate.

Purpose of the research study

The information from this study will be used to guide the framing of my dissertation outcomes in partial fulfilment of the requirements for the Degree of Doctor of Business Administration from the Swiss School of Business and Management Geneva. This study aims to undertake a regulatory, governance and ethical impact assessment of Artificial Intelligence in the energy sector as it evolves and matures. Your participation will provide valuable insights to guide the development of a compliance framework for overseeing the implementation of Artificial Intelligence technologies as a final deliverable of this dissertation.

What Does Participation Involve

If you agree to participate, you will be asked to take part in an online questionnaire, which will take approximately 20 minutes. The questionnaire is aimed at identifying the current knowledge and understanding of the current Artificial Intelligence regulatory and compliance oversight landscape in the energy/electricity sector and to identify any gaps.

Risks and Benefits

There are no risks associated with your participation in this study as no personal or organization specific information is being requested. There are no immediate benefits to participants, but your contribution will add to the knowledge base to structure a detailed Artificial Intelligence oversight or compliance audit framework, which will assist the broader industry in establishing policies, procedures and structures to safely and sustainably implement Artificial Intelligence tools and systems.

Confidentiality

Your responses will be confidential. The records of this study will be kept private by storing and analysing in an environment with controlled access. All individual participants identifying information will not be loaded into any online data analysis portals and will not be reported. The data will be summarized and reported in aggregate format in the dissertation, any publications or presentations resulting from this research.

Voluntary Participation

Your participation in this study is entirely voluntary. It is up to you to decide whether or not to take part in this study. As this study is being conducted online, your completion of the study is taken as consent. You are free to withdraw consent at any time, without giving reason and with no adverse effects via the investigator.

Consent

- I have read and understood the information provided above.
- I have had the opportunity to ask questions, and all my questions have been answered to my satisfaction.
- I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason and without consequence.
- I don't expect to receive any benefit or payment for my participation.
- I agree to take part in this study.

REFERENCES

Agrawal, A., Gans, J. and Goldfarb, A. (2017), "What to expect from artificial intelligence", MIT Sloan Management Review Cambridge, MA, USA.

AI Verify Foundation. (2023), "AI Verify", July, available at: https://aiverifyfoundation.sg/what-is-ai-verify/.

Ala-Pietilä, P. and Smuha, N.A. (2021), "A framework for global cooperation on artificial intelligence and its governance", *Reflections on Artificial Intelligence for Humanity*, Springer, pp. 237–265.

Anderson, M. and Fort, K. (2022), "Human where? A new scale defining human involvement in technology communities from an ethical standpoint", *International Review of Information Ethics*, Vol. 31 No. 1.

Arévalo, P. and Jurado, F. (2024), "Impact of artificial intelligence on the planning and operation of distributed energy systems in smart grids", *Energies*, MDPI, Vol. 17 No. 17, p. 4501.

Ayling, J. and Chapman, A. (2022), "Putting AI ethics to work: are the tools fit for purpose?", *AI and Ethics*, Springer, Vol. 2 No. 3, pp. 405–429.

Baker-Brunnbauer, J. (2021), "TAII framework for trustworthy AI systems", *Baker-Brunnbauer, J.(2021). TAII Framework for Trustworthy AI Systems. ROBONOMICS: The Journal of the Automated Economy*, Vol. 2, p. 17.

Bankins, S. and Formosa, P. (2023), "The ethical implications of artificial intelligence (AI) for meaningful work", *Journal of Business Ethics*, Springer, pp. 1–16.

Barr Advisory. (2024), "ISO42001 White paper - Everything You Need to Know About ISO 42001", Barr Advisory, 24 June.

Beck, J. and Burri, T. (2024), "From 'human control' in international law to 'human oversight' in the new EU act on artificial intelligence", *Research Handbook on*

Meaningful Human Control of Artificial Intelligence Systems, Edward Elgar Publishing, pp. 104–130.

Benraouane, S.A. (2024), *AI Management System Certification According to the ISO/IEC 42001 Standard: How to Audit, Certify, and Build Responsible AI Systems*, CRC Press.

Berente, N., Gu, B., Recker, J. and Santhanam, R. (2021), "Managing artificial intelligence.", *MIS Quarterly*, Vol. 45 No. 3.

Berger, R. (2018), "Artificial intelligence: A smart move for utilities", *Roland Berger*.

Bhuvan, S. (2023), "A STUDY ON GOVERNANCE FRAMEWORK FOR AI AND ML SYSTEMS", *ShodhKosh: Journal of Visual and Performing Arts*, Vol. 4, doi: 10.29121/shodhkosh.v4.i2.2023.1923.

Bignami, F. (2022), "Artificial intelligence accountability of public administration", *The American Journal of Comparative Law*, Oxford University Press UK, Vol. 70 No. Supplement_1, pp. i312–i346.

Birhane, A., Steed, R., Ojewale, V., Vecchione, B. and Raji, I.D. (2024), "AI auditing: The broken bus on the road to AI accountability", presented at the 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), IEEE, pp. 612–643.

Brown, J.E. and Albert, E.J. (2023), "Beyond the IUDex threshold: human oversight as the conscience of machine learning", *Colorado Technology Law Journal*, Vol. 22.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., *et al.* (2018), "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation", *arXiv Preprint arXiv:1802.07228*.

Bufe, C. (2024), "ISO 42001 Artificial Intelligence - Management system", 31 May.

Burton, S.L. (2019), “Grasping the cyber-world: Artificial intelligence and human capital meet to inform leadership”, *International Journal of Economics, Commerce and Management*, Vol. 7 No. 12, pp. 707–759.

Büthe, T., Djeffal, C., Lütge, C., Maasen, S. and Ingersleben-Seip, N. von. (2022), “Governing AI–attempting to herd cats? Introduction to the special issue on the Governance of Artificial Intelligence”, *Journal of European Public Policy*, Taylor & Francis, Vol. 29 No. 11, pp. 1721–1752.

Caner, S. and Bhatti, F. (2020), “A conceptual framework on defining businesses strategy for artificial intelligence”, *Contemporary Management Research*, Vol. 16 No. 3, pp. 175–206.

Christen, M., Burri, T., Kandul, S. and Vörös, P. (2023), “Who is controlling whom? Reframing ‘meaningful human control’ of AI systems in security”, *Ethics and Information Technology*, Springer, Vol. 25 No. 1, p. 10.

Clarke, R. (2019), “Regulatory alternatives for AI”, *Computer Law & Security Review*, Elsevier, Vol. 35 No. 4, pp. 398–409.

Clavell, G. (2023), *AI Auditing - Checklist for AI Auditing*, European Data Protection Board.

Coglianese, C. and Crum, C.R. (2024), “Taking Training Seriously: Human Guidance and Management-Based Regulation of Artificial Intelligence”, *arXiv Preprint arXiv:2402.08466*.

Costanza-Chock, S., Raji, I.D. and Buolamwini, J. (2022), “Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem”, presented at the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1571–1583.

Cubric, M. (2020), “Drivers, barriers and social considerations for AI adoption in business and management: A tertiary study”, *Technology in Society*, Elsevier, Vol. 62, p. 101257.

Dafoe, A. (2018), “AI governance: a research agenda”, *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK*, Vol. 1442, p. 1443.

Daly, A., Hagendorff, T., Li, H., Mann, M., Marda, V., Wagner, B. and Wang, W.W. (2020), “AI, governance and ethics: global perspectives”, *University of Hong Kong Faculty of Law Research Paper*, No. 2020/051.

Davidovic, J. (2023), “On the purpose of meaningful human control of AI”, *Frontiers in Big Data*, Frontiers, Vol. 5, p. 1017677.

De Silva, D. and Alahakoon, D. (2022), “An artificial intelligence life cycle: From conception to production”, *Patterns*, Elsevier, Vol. 3 No. 6.

van Diggelen, J., Boshuijzen-van Burken, C. and Abbass, H. (2024), “Team Design Patterns for Meaningful Human Control in Responsible Military Artificial Intelligence”, presented at the International Conference on Bridging the Gap between AI and Reality, Springer, pp. 40–54.

Dignam, A. (2020), “Artificial intelligence, tech corporate governance and the public interest regulatory response”, *Cambridge Journal of Regions, Economy and Society*, Oxford University Press UK, Vol. 13 No. 1, pp. 37–54.

Djeffal, C., Siewert, M.B. and Wurster, S. (2022), “Role of the state and responsibility in governing artificial intelligence: A comparative analysis of AI strategies”, *Journal of European Public Policy*, Taylor & Francis, Vol. 29 No. 11, pp. 1799–1821.

Dotan, R., Blili-Hamelin, B., Madhavan, R., Matthews, J. and Scarpino, J. (2024), "Evolving AI Risk Management: A Maturity Model based on the NIST AI Risk Management Framework", *arXiv Preprint arXiv:2401.15229*.

Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., *et al.* (2021), "Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy", *International Journal of Information Management*, Elsevier, Vol. 57, p. 101994.

Enqvist, L. (2023), "'Human oversight'in the EU artificial intelligence act: what, when and by whom?", *Law, Innovation and Technology*, Taylor & Francis, Vol. 15 No. 2, pp. 508–535.

European Commission. (2024), "AI Act, Shaping Europe's digital future", 8 August, available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (accessed 17 September 2024).

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., *et al.* (2021), "Governing AI safety through independent audits", *Nature Machine Intelligence*, Nature Publishing Group UK London, Vol. 3 No. 7, pp. 566–571.

Fanni, R., Steinkogler, V.E., Zampedri, G. and Pierson, J. (2023), "Enhancing human agency through redress in Artificial Intelligence Systems", *AI & Society*, Springer, Vol. 38 No. 2, pp. 537–547.

Fatima, H., Jan, S., Khan, A.K., Javed, S. and Rashid, M. (2024), "EFFECT OF ARTIFICIAL INTELLIGENCE ON THE HUMAN WORKFORCE", *International Journal of Contemporary Issues in Social Sciences*, Vol. 3 No. 1, pp. 1197–1203.

Fenwick, A. and Molnar, G. (2022), "The importance of humanizing AI: using a behavioral lens to bridge the gaps between humans and machines", *Discover Artificial Intelligence*, Springer, Vol. 2 No. 1, p. 14.

Ferretti, T. (2022), "An institutionalist approach to AI ethics: justifying the priority of government regulation over self-regulation", *Moral Philosophy and Politics*, De Gruyter, Vol. 9 No. 2, pp. 239–265.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. and Srikumar, M. (2020), "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI", *Berkman Klein Center Research Publication*, No. 2020–1.

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J. and Wen, Y. (2022), "CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act", *Available at SSRN 4064091*.

Florkowski, M., Hayashi, H., Matsuda, S., Moribe, H., Moriwaki, N., Jones, D. and Pauska, J. (2024), "Digitally Boosted Resilience: Digitalization to Enhance Resilience of Electric Power System", *IEEE Power and Energy Magazine*, IEEE, Vol. 22 No. 2, pp. 100–109.

Franki, V., Majnarić, D. and Višković, A. (2023), "A comprehensive review of Artificial Intelligence (AI) companies in the power sector", *Energies*, MDPI, Vol. 16 No. 3, p. 1077.

GAO. (2021), *Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities*, US Government Accountability Office.

Gasser, U. and Almeida, V.A. (2017), "A layered model for AI governance", *IEEE Internet Computing*, IEEE, Vol. 21 No. 6, pp. 58–62.

Green, B. (2022), "The flaws of policies requiring human oversight of government algorithms", *Computer Law & Security Review*, Elsevier, Vol. 45, p. 105681.

Green, B. and Kak, A. (2021), “The false comfort of human oversight as an antidote to AI harm”, Slate.

Gudala, L., Shaik, M., Venkataramanan, S. and Sadhu, A.K.R. (2019), “Leveraging Artificial Intelligence for Enhanced Threat Detection, Response, and Anomaly Identification in Resource-Constrained IoT Networks”, *Distributed Learning and Broad Applications in Scientific Research,* Vol. 5, pp. 23–54.

Gutierrez, C.I. and Marchant, G.E. (2021), “A global perspective of soft law programs for the governance of artificial intelligence”, *Available at SSRN 3855171*.

Haakman, M., Cruz, L., Huijgens, H. and van Deursen, A. (2021), “AI lifecycle models need to be revised: An exploratory study in Fintech”, *Empirical Software Engineering,* Springer, Vol. 26, pp. 1–29.

Hadzovic, S., Becirspahic, L. and Mrdovic, S. (2024), “It’s time for artificial intelligence governance”, *Internet of Things,* Elsevier, Vol. 27, p. 101292.

Hartmann, D., de Pereira, J.R.L., Streitbörger, C. and Berendt, B. (2024), “Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society”, *AI and Ethics*, Springer, pp. 1–22.

Hassan, N. (2024), “Data Poisoning (AI Poisoning)”, May, available at: https://www.techtarget.com/searchenterpriseai/definition/data-poisoning-AI-poisoning (accessed 28 August 2024).

Hickman, E. and Petrin, M. (2021), “Trustworthy AI and corporate governance: the EU’s ethics guidelines for trustworthy artificial intelligence from a company law perspective”, *European Business Organization Law Review,* Springer, Vol. 22, pp. 593–625.

Hickok, M. (2021), “Lessons learned from AI ethics principles for future actions”, *AI and Ethics*, Springer, Vol. 1 No. 1, pp. 41–47.

Huang, C., Zhang, Z., Mao, B. and Yao, X. (2022), "An overview of artificial intelligence ethics", *IEEE Transactions on Artificial Intelligence*, IEEE.

ICO, U. (2020), "Guidance on the AI auditing framework: Draft guidance for consultation", Information Commissioner's Office.

ICO, U. (2022), "A Guide to ICO Audit - Artificial Intelligence (AI) Audits", Information Commissioners Office, UK, 18 November.

ICO, U. (2023), "Guidance on AI and data protection", 15 March, available at: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ (accessed 17 September 2024).

IEA. (2017), "Digitalisation and Energy", *Https://Www.Iea.Org/Reports/Digitalisation-and-Energy*, Vol. License: CC BY 4.0, p. 15.

IIA. (2023), *IIA AI Auditing Framework*, Chartered Institute of Internal Auditors.

Institute of Internal Auditors. (2017), "The IIA's Artificial Intelligence Auditing Framework - Practical Applications, Part A", December.

ISACA. (2018), "Auditing Artificial Intelligence", *Https://Store.Isaca.Org/s/Store#/Store/Browse/Detail/a2S4w000004KoGpEAK*, p. 12.

ISACA. (2024), "Unlocking AI's Potential: How COBIT Can Guide Your Business Transformation", 2 May, available at: https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2024/unlocking-ais-potential-how-cobit-can-guide-your-business-transformation (accessed 21 September 2024).

Isensee, C., Griese, K.M. and Teuteberg, F. (2021), "Sustainable artificial intelligence: A corporate culture perspective.", *In Sustainability Management Forum| NachhaltigkeitsManagementForum*, Vol. 29 No. 3–4, pp. 217–230.

ISO. (2023), “ISO/IEC 42001:2023 Information Technology - Artificial Intelligence Management System”, Standard, ISO, December, doi: 03.100.70, 35.020.

Janaćković, G., Vasović, D. and Vasović, B. (2024), “ARTIFICIAL INTELLIGENCE STANDARDISATION EFFORTS”, *ENGINEERING MANAGEMENT AND COMPETITIVENESS (EMC 2024)*, p. 250.

Jarrahi, M.H., Askay, D., Eshraghi, A. and Smith, P. (2023), “Artificial intelligence and knowledge management: A partnership between human and AI”, *Business Horizons*, Elsevier, Vol. 66 No. 1, pp. 87–99.

Javaid, S. (2024), “5 AI Training Steps & Best Practices in 2024”, 17 July, available at: https://research.aimultiple.com/ai-training/ (accessed 25 August 2024).

Johansson, S. and Björkman, I. (2018), “What impact will Artificial Intelligence have on the future leadership role?–A study of leaders’ expectations”.

Jorzik, P., Yigit, A., Kanbach, D.K., Kraus, S. and Dabić, M. (2023), “Artificial Intelligence-Enabled Business Model Innovation: Competencies and Roles of Top Management”, *IEEE Transactions on Engineering Management*, IEEE.

Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J. and Sanchez Martin, J. (2023), “Cybersecurity of Artificial Intelligence in the AI Act”, *Publications Office Eur. Union, Luxembourg, UK, Tech. Rep. JRC134461*.

Kaminski, M.E. (2023), “Regulating the Risks of AI”, *Forthcoming, Boston University Law Review*, Vol. 103.

Kaur, D., Uslu, S., Rittichier, K.J. and Durresi, A. (2022), “Trustworthy artificial intelligence: a review”, *ACM Computing Surveys (CSUR)*, ACM New York, NY, Vol. 55 No. 2, pp. 1–38.

Kazim, E., Denny, D.M.T. and Koshiyama, A. (2021), "AI auditing and impact assessment: according to the UK information commissioner's office", *AI and Ethics*, Springer, Vol. 1, pp. 301–310.

Kazim, E. and Koshiyama, A. (2020), "AI assurance processes", *Available at SSRN 3685087*.

Kelley, A. (2024), "The agency's 2024-2025 Strategic Guidance and National Priorities for Critical Infrastructure highlights continued need to monitor AI's interplay with cybersecurity.", 21 June, available at: https://www.nextgov.com/cybersecurity/2024/06/dhs-highlights-ai-threat-and-asset-critical-infrastructure-new-priority-guidance/397524/ (accessed 11 August 2024).

Kharchenko, V., Fesenko, H. and Illiashenko, O. (2022), "Quality models for artificial intelligence systems: characteristic-based approach, development and application", *Sensors*, MDPI, Vol. 22 No. 13, p. 4865.

KILINÇ, İ. and Aslıhan, Ü. (2020), "Reflections of Artificial Intelligence on C-suite", *Nitel Sosyal Bilimler*, Melih SEVER, Vol. 2 No. 1, pp. 1–18.

Kitsios, F. and Kamariotou, M. (2021), "Artificial intelligence and business strategy towards digital transformation: A research agenda", *Sustainability*, MDPI, Vol. 13 No. 4, p. 2025.

Knowles, B. and Richards, J.T. (2021), "The sanction of authority: Promoting public trust in AI", presented at the Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 262–271.

Konidena, B.K., Malaiyappan, J.N.A. and Tadimarri, A. (2024), "Ethical Considerations in the Development and Deployment of AI Systems", *European Journal of Technology*, Vol. 8 No. 2, pp. 41–53.

Kop, M. (2021), “Establishing a legal-ethical framework for quantum technology”, *Yale Law School, Yale Journal of Law & Technology (YJoLT), The Record*.

Kopeinik, S., Scher, S., Nad, T. and Kowald, D. (2023), *Human Agency and Oversight*, White Paper, SGS Digital Trusts Services GmbH.

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., *et al.* (2024), “Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms”, *Royal Society Open Science*, The Royal Society, Vol. 11 No. 5, p. 230859.

KPMG. (2024), *Decoding the EU AI Act*, KPMG.

de Laat, P.B. (2021), “Companies committed to responsible AI: From principles towards implementation and regulation?”, *Philosophy & Technology*, Springer, Vol. 34, pp. 1135–1193.

Landers, R.N. and Behrend, T.S. (2023), “Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models.”, *American Psychologist*, American Psychological Association, Vol. 78 No. 1, p. 36.

Laplante, P. and Amaba, B. (2021), “Artificial intelligence in critical infrastructure systems”, *Computer*, IEEE, Vol. 54 No. 10, pp. 14–24.

Laroussi, I., Huan, L. and Xiusheng, Z. (2023), “How will the internet of energy (IoE) revolutionize the electricity sector? A techno-economic review”, *Materials Today: Proceedings*, Elsevier, Vol. 72, pp. 3297–3311.

Laux, J., Stephany, F. and Liefgreen, A. (2023), “The Economics of Human Oversight: How Norms and Incentives Affect Costs and Performance of AI Workers”, *arXiv Preprint arXiv:2312.14565*.

Lea, A. (2023), “Why is AI hard to define?”, 24 November, available at: https://www.bcs.org/articles-opinion-and-research/why-is-ai-hard-to-define/ (accessed 11 August 2024).

Leidy, E.N. and Gerstein, D.M. (2024), “Emerging Technology and Risk Analysis: Artificial Intelligence and Critical Infrastructure”, < bound method Organization. get_name_with_acronym of< Organization: RAND ….

Leslie, D., Burr, C., Aitken, M., Cowls, J., Katell, M. and Briggs, M. (2021), “Artificial intelligence, human rights, democracy, and the rule of law: a primer”, *arXiv Preprint arXiv:2104.04147*.

Li, J., Li, M., Wang, X. and Thatcher, J.B. (2021), “Strategic Directions for AI: The Role of CIOs and Boards of Directors.”, *MIS Quarterly*, Vol. 45 No. 3.

Lima, G., Grgić-Hlača, N., Jeong, J.K. and Cha, M. (2022), “The conflict between explainable and accountable decision-making algorithms”, presented at the Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 2103–2113.

Lisievici, A. (2024), “EDPB publishes Checklist for AI auditing”, 7 July, available at: https://blog.privacycraft.pro/edpb-publishes-checklist-for-ai-auditing/ (accessed 21 September 2024).

Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D. and Jacquet, A. (2024), “Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering”, *ACM Computing Surveys*, ACM New York, NY, Vol. 56 No. 7, pp. 1–35.

Lyu, W. and Liu, J. (2021), “Artificial Intelligence and emerging digital technologies in the energy sector”, *Applied Energy*, Elsevier, Vol. 303, p. 117615.

Machlev, R., Heistrene, L., Perl, M., Levy, K., Belikov, J., Mannor, S. and Levron, Y. (2022), "Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities", *Energy and AI*, Elsevier, Vol. 9, p. 100169.

Manheim, D., Martin, S., Bailey, M., Samin, M. and Greutzmacher, R. (2024), "The Necessity of AI Audit Standards Boards", *arXiv Preprint arXiv:2404.13060*.

Mannes, A. (2020), "Governance, risk, and artificial intelligence", *Ai Magazine*, Vol. 41 No. 1, pp. 61–69.

Mäntymäki, M., Minkkinen, M., Birkstedt, T. and Viljanen, M. (2022), "Putting AI ethics into practice: The hourglass model of organizational AI governance", *arXiv Preprint arXiv:2206.00335*.

Martin, K., Shilton, K. and Smith, J. ery. (2022), "Business and the ethical implications of technology: Introduction to the symposium", *Business and the Ethical Implications of Technology*, Springer, pp. 1–11.

McKendrick, J. and Thurai, A. (2022), "AI isn't ready to make unsupervised decisions", *Harvard Business Review*, Vol. 15, p. 10.

Michael, K., Abbas, R., Roussos, G., Scornavacca, E. and Fosso-Wamba, S. (2020), "Ethics in AI and autonomous system applications design", *IEEE Transactions on Technology and Society*, IEEE, Vol. 1 No. 3, pp. 114–127.

Michel, A.H. (2023), "Recalibrating assumptions on AI", < bound method Organization. get_name_with_acronym of< Organization: Chatham ….

Mikalef, P., Conboy, K., Lundström, J.E. and Popovič, A. (2022), "Thinking responsibly about responsible AI and 'the dark side'of AI", *European Journal of Information Systems*, Taylor & Francis, Vol. 31 No. 3, pp. 257–268.

Miles, S. (2023), “Enhancing strategic change management practices to stay ahead of external events”, available at: https://www.cefpro.com/enhancing-strategic-change-management-practices-to-stay-ahead-of-external-events/ (accessed 27 September 2024).

Mökander, J., Axente, M., Casolari, F. and Floridi, L. (2022), “Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation”, *Minds and Machines*, Springer, Vol. 32 No. 2, pp. 241–268.

Mökander, J., Morley, J., Taddeo, M. and Floridi, L. (2021), “Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations”, *Science and Engineering Ethics*, Springer, Vol. 27 No. 4, p. 44.

Moldenhauer, L. and Londt, C. (2018), “Leadership, artificial intelligence and the need to redefine future skills development”, presented at the Proceedings of the European Conference on Management, Leadership & Governance, pp. 155–160.

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J. and Floridi, L. (2021), “Ethics as a service: a pragmatic operationalisation of AI ethics”, *Minds and Machines*, Springer, Vol. 31 No. 2, pp. 239–256.

Morris, E., Stamp, K., Halford, A. and Gaura, E. (2022), “The practice of AI and ethics in energy transition futures”, University of Strathclyde.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. and Fernández-Leal, Á. (2023), “Human-in-the-loop machine learning: a state of the art”, *Artificial Intelligence Review*, Springer, Vol. 56 No. 4, pp. 3005–3054.

Munn, L. (2023), “The uselessness of AI ethics”, *AI and Ethics*, Springer, Vol. 3 No. 3, pp. 869–877.

Murray, G., Johnstone, M.N. and Valli, C. (2017), “The convergence of IT and OT in critical infrastructure”.

Musch, S., Borrelli, M. and Kerrigan, C. (2023), “The EU AI Act As Global Artificial Intelligence Regulation”, *Available at SSRN 4549261*.

Nasim, S.F., Ali, M.R. and Kulsoom, U. (2022), “Artificial intelligence incidents & ethics a narrative review”, *International Journal of Technology, Innovation and Management (IJTIM)*, Vol. 2 No. 2, pp. 52–64.

Nazari, Z. and Musilek, P. (2023), “Impact of Digital Transformation on the Energy Sector: A Review”, *Algorithms*, MDPI, Vol. 16 No. 4, p. 211.

Niet, I., van Est, R. and Veraart, F. (2021), “Governing AI in electricity systems: Reflections on the EU artificial intelligence bill”, *Frontiers in Artificial Intelligence*, Frontiers Media SA, Vol. 4, p. 690237.

NIST. (2023), “Artificial Intelligence Risk Management Framework (AI RMF 1.0)”, National Institute of Standards & Technology, USA.

Nothwang, W.D., McCourt, M.J., Robinson, R.M., Burden, S.A. and Curtis, J.W. (2016), “The human should be part of the control loop?”, presented at the 2016 Resilience Week (RWS), IEEE, pp. 214–220.

Novelli, C., Hacker, P., Morley, J., Trondal, J. and Floridi, L. (2024), “A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities”, *AI Board, Scientific Panel, and National Authorities (May 5, 2024)*.

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., *et al.* (2023), “Six human-centered artificial intelligence grand challenges”, *International Journal of Human–Computer Interaction*, Taylor & Francis, Vol. 39 No. 3, pp. 391–437.

Paul, J. and Criado, A.R. (2020), "The art of writing literature review: What do we know and what do we need to know?", *International Business Review*, Elsevier, Vol. 29 No. 4, p. 101717.

PDPC and IMDA. (2020), *Model Artificial Intelligence Governance Framework 2nd Edition*, Infocomm Media Development Authority/Personal Data Protection Commission Singapore.

Peifer, Y., Jeske, T. and Hille, S. (2022), "Artificial intelligence and its impact on leaders and leadership", *Procedia Computer Science*, Elsevier, Vol. 200, pp. 1024–1030.

Perry, B. and Uuk, R. (2019), "AI governance and the policymaking process: key considerations for reducing AI risk", *Big Data and Cognitive Computing*, MDPI, Vol. 3 No. 2, p. 26.

Priya, B., Sharma, V., Awotunde, J.B. and Adeniyi, A.E. (n.d.). "Artificial Intelligence in Industry 5.0: Transforming Manufacturing through Machine Learning and Robotics in Collaborative Age", *Computational Intelligence in Industry 4.0 and 5.0 Applications*, Auerbach Publications, pp. 61–100.

Pugliese, R., Regondi, S. and Marini, R. (2021), "Machine learning-based approach: Global trends, research directions, and regulatory standpoints", *Data Science and Management*, Elsevier, Vol. 4, pp. 19–29.

Raji, I.D., Xu, P., Honigsberg, C. and Ho, D. (2022), "Outsider oversight: Designing a third party audit ecosystem for ai governance", presented at the Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 557–571.

Roberts, H., Cowls, J., Hine, E., Mazzi, F., Tsamados, A., Taddeo, M. and Floridi, L. (2021), "Achieving a 'Good AI Society': Comparing the Aims and Progress of the EU and the US", *Science and Engineering Ethics*, Springer, Vol. 27, pp. 1–25.

Roberts, H., Zhang, J., Bariach, B., Cowls, J., Gilburt, B., Juneja, P., Tsamados, A., *et al.* (2022), "Artificial intelligence in support of the circular economy: ethical considerations and a path forward", *AI & SOCIETY*, Springer, pp. 1–14.

Roski, J., Maier, E.J., Vigilante, K., Kane, E.A. and Matheny, M.E. (2021), "Enhancing trust in AI through industry self-governance", *Journal of the American Medical Informatics Association*, Oxford University Press, Vol. 28 No. 7, pp. 1582–1590.

Ryan, M. and Stahl, B.C. (2020), "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications", *Journal of Information, Communication and Ethics in Society*, Emerald Publishing Limited, Vol. 19 No. 1, pp. 61–86.

Scannell, B., Moore, L. and Hayes, R. (2024), "The Time To (AI) Act Is Now: A Practical Guide To Emotion Recognition Systems Under The AI Act", 25 July, available at: https://www.mondaq.com/ireland/new-technology/1497172/the-time-to-ai-act-is-now-a-practical-guide-to-regulatory-sandboxes-under-the-ai-act (accessed 26 September 2024).

Schiff, D., Biddle, J., Borenstein, J. and Laas, K. (2020), "What's next for ai ethics, policy, and governance? a global overview", presented at the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 153–158.

Schultz, M.D. and Seele, P. (2023), "Towards AI ethics' institutionalization: knowledge bridges from business ethics to advance organizational AI ethics", *AI and Ethics*, Springer, Vol. 3 No. 1, pp. 99–111.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A. and Hall, P. (2022), "Towards a standard for identifying and managing bias in artificial intelligence", *NIST Special Publication*, Vol. 1270 No. 10.6028.

Scott, R. (2024), “Navigating the Landscape: Ethics and Generative Ai in the Age of HYPERREAL”, 5 February, available at: https://www.hyperreal.io/post/ethical-considerations-in-generative-ai-hyperreal-takes-a-stand (accessed 25 September 2024).

Shadman, S.A.K. (2023), “The Impact of Incorporating Artificial Intelligence on Leadership from the Perspective of Leadership Experts”, Alliant International University.

Sharkov, G., Todorova, C. and Varbanov, P. (2021), “Strategies, policies, and standards in the eu towards a roadmap for robust and trustworthy ai certification”, *Information & Security*, ProCon Ltd., Vol. 50 No. 1, pp. 11–22.

Sheikh, H., Prins, C. and Schrijvers, E. (2023), “Artificial Intelligence: Definition and Background”, in Sheikh, H., Prins, C. and Schrijvers, E. (Eds.), *Mission AI: The New System Technology*, Springer International Publishing, Cham, pp. 15–41, doi: 10.1007/978-3-031-21448-6_2.

Simbeck, K. (2024), “They shall be fair, transparent, and robust: auditing learning analytics systems”, *AI and Ethics*, Springer, Vol. 4 No. 2, pp. 555–571.

Slate, D.D., Parisot, A., Min, L., Panciatici, P. and Van Hentenryck, P. (2024), “Adoption of Artificial Intelligence by Electric Utilities”, *Energy LJ*, HeinOnline, Vol. 45, p. 1.

Smuha, N.A. (2021), “From a ‘race to AI’to a ‘race to AI regulation’: regulatory competition for artificial intelligence”, *Law, Innovation and Technology*, Taylor & Francis, Vol. 13 No. 1, pp. 57–84.

Stahl, B.C. (2022), “Responsible innovation ecosystems: Ethical implications of the application of the ecosystem concept to artificial intelligence”, *International Journal of Information Management*, Elsevier, Vol. 62, p. 102441.

Stahl, B.C. and Stahl, B.C. (2021), "Ethical issues of AI", *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, Springer, pp. 35–53.

Statista Research Department. (2024), "Electric utilities in the U.S. - statistics & facts", 1 July.

StatPlan Energy Ltd. (2020), "Electricity Utility Customer Analysis Ed 1 2020", Database, StatPlan Energy Ltd, February.

Stix, C. (2021a), "Actionable principles for artificial intelligence policy: three pathways", *Science and Engineering Ethics*, Springer, Vol. 27 No. 1, p. 15.

Stix, C. (2021b), "The ghost of AI governance past, present and future: AI governance in the European Union", *arXiv Preprint arXiv:2107.14099*.

Stuurman, K. and Lachaud, E. (2022), "Regulating AI. A label to complete the proposed Act on Artificial Intelligence", *Computer Law & Security Review*, Elsevier, Vol. 44, p. 105657.

Światowiec-Szczepańska, J. and Stępień, B. (2022), "Drivers of digitalization in the energy sector—The managerial perspective from the catching up economy", *Energies*, MDPI, Vol. 15 No. 4, p. 1437.

Synergist Technology. (2024), "How to audit Artificial Intelligence using COBIT 2019", 4 March, available at: https://synergist.technology/blogs/ai-kbase/how-to-audit-artificial-intelligence-using-cobit-2019 (accessed 21 September 2024).

Taeihagh, A. (2021), "Governance of artificial intelligence", *Policy and Society*, Oxford University Press, Vol. 40 No. 2, pp. 137–157.

Talbot, J. (2018), "Risk BowTie Method", 25 May, available at: https://www.juliantalbot.com/post/risk-bow-tie-method (accessed 16 September 2024).

Thelisson, E. and Verma, H. (2024), "Conformity assessment under the EU AI act general approach", *AI and Ethics*, Springer, Vol. 4 No. 1, pp. 113–121.

Thomas, C., Roberts, H., Mökander, J., Tsamados, A., Taddeo, M. and Floridi, L. (2024), "The case for a broader approach to AI assurance: addressing 'hidden' harms in the development of artificial intelligence", *AI & SOCIETY*, Springer, pp. 1–16.

Torrance, A.W. and Tomlinson, B. (2023), "Governance of the AI, by the AI, and for the AI", *Miss. LJ*, HeinOnline, Vol. 93, p. 107.

Ulnicane, I., Eke, D.O., Knight, W., Ogoh, G. and Stahl, B.C. (2021), "Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies", *Interdisciplinary Science Reviews*, Taylor & Francis, Vol. 46 No. 1–2, pp. 71–93.

Walter, J. (2023), *Human Oversight Done Right: The AI Act Should Use Humans to Monitor AI Only When Effective*, ZEW policy brief.

Walter, Y. (2024), "Managing the race to the moon: Global policy and governance in Artificial Intelligence regulation—A contemporary overview and an analysis of socioeconomic consequences", *Discover Artificial Intelligence*, Springer, Vol. 4 No. 1, p. 14.

Walz, A. and Firth-Butterfield, K. (2019), "Implementing ethics into artificial intelligence: a contribution, from a legal perspective, to the development of an AI governance regime", *Duke L. & Tech. Rev.*, HeinOnline, Vol. 18, p. 176.

Williams, R., Cloete, R., Cobbe, J., Cottrill, C., Edwards, P., Markovic, M., Naja, I., *et al.* (2022), "From transparency to accountability of intelligent systems: Moving beyond aspirations", *Data & Policy*, Cambridge University Press, Vol. 4, p. e7.

Wilson, C. and Van Der Velden, M. (2022), “Sustainable AI: An integrated model to guide public sector decision-making”, *Technology in Society*, Elsevier, Vol. 68, p. 101926.

Winfield, A.F. and Jirotka, M. (2018), “Ethical governance is essential to building trust in robotics and artificial intelligence systems”, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, The Royal Society Publishing, Vol. 376 No. 2133, p. 20180085.

Winfield, A.F., Michael, K., Pitt, J. and Evers, V. (2019), “Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]”, *Proceedings of the IEEE*, IEEE, Vol. 107 No. 3, pp. 509–517.

World Health Organization. (2021), “Ethics and governance of artificial intelligence for health: WHO guidance”, World Health Organization.

Xia, B., Lu, Q., Zhu, L. and Xing, Z. (2024), “An ai system evaluation framework for advancing ai safety: Terminology, taxonomy, lifecycle mapping”, presented at the Proceedings of the 1st ACM International Conference on AI-Powered Software, pp. 74–78.

Xue, L. and Pang, Z. (2022), “Ethical governance of artificial intelligence: An integrated analytical framework”, *Journal of Digital Economy*, Elsevier, Vol. 1 No. 1, pp. 44–52.

Yigit, Y., Ferrag, M.A., Sarker, I.H., Maglaras, L.A., Chrysoulas, C., Moradpoor, N. and Janicke, H. (2024), “Critical infrastructure protection: Generative ai, challenges, and opportunities”, *arXiv Preprint arXiv:2405.04874*.

Zimmer, M.P., Minkkinen, M. and Mäntymäki, M. (2022), “Responsible Artificial Intelligence Systems Critical considerations for business model design”, *Scandinavian Journal of Information Systems*, Vol. 34 No. 2, p. 4.