UNIVERSITY OF TURKU
Turku School of Economics

Effective human oversight in Artificial Intelligence

Information Systems Science,
Bachelor's thesis

Author:
Aliisa Jauhiainen

Supervisor:
PhD Kai Kimppa

11.12.2025
Turku

Student’s statement regarding the use of Artificial Intelligence (AI) for preparing and/or writing this thesis:

☐ **I have not used any AI-based tools.**

☒ **I have used AI-based tools**. Their use is documented in the Appendix. The AI tools were used in a way that complies with academic integrity guidelines.

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

Bachelor's thesis
Subject: Information Systems Science
Author: Aliisa Jauhiainen
Title: Effective human oversight in Artificial Intelligence
Supervisor: PhD Kai Kimppa
Number of pages: 35 pages + appendices 2 pages
Date: 11.12.2025

Abstract

Artificial Intelligence is one of the most transformative forces of the twenty-first century. How its diverse benefits can be harnessed while mitigating its various risks, is becoming increasingly important. In response, in July of 2024, the European Commission published the EU Artificial Intelligence Act, with the intention of regulating the development and deployment of AI within the European Union. The overarching goal of the Act is to promote human-centric and trustworthy AI by safeguarding the health, safety and fundamental rights of individuals through a risk-based framework. The Act introduces human oversight as one of the core measures of protection in regulating high-risk AI systems.

This thesis explores the role of deployers in implementing oversight and provides insights into how holding overseers accountable can enable effective human oversight. The thesis begins by providing an overview of the EU AI Act, including its definition of oversight and the requirements set for it. Oversight is also examined through relevant ethical guidelines. After providing a definition for accountability and presenting how AI enables its unfair allocation, the thesis determines what constitutes effective oversight, demonstrates how allocating accountability to overseers can enable it in practice and defines the role of deployers in its implementation.

The findings indicate that deployers play a critical role in oversight implementation in practice and that effective oversight requires adherence with legal requirements, ethical principles and fulfilment of oversight goals. Moreover, the findings reveal that accountability enhances human task performance, cognitive complexity, sense of morality and information accumulation and processing, consequently, enabling effective human oversight over AI systems.

This thesis contributes to the ongoing global discourse on Artificial Intelligence regulation, focusing on the EU AI Act and its imminent implementation. It links accountability and human oversight, aims to bridge the gap between regulation and practical implementation and generally contributes to the discussion around human-centric AI. Ultimately, the thesis provides insights for deployers of high-risk AI systems on the importance of effective human oversight in promoting responsible Artificial Intelligence and safeguarding human autonomy and fundamental values.

Keywords: Artificial Intelligence, EU Artificial Intelligence Act, Effective human oversight, Accountability, Artificial Intelligence governance

Kandidaatintutkielma

Oppiaine: Tietojärjestelmätiede
Tekijä: Aliisa Jauhiainen
Otsikko: Tehokas ihmisen suorittama valvonta tekoälylle
Ohjaaja: FT Kai Kimppa
Sivumäärä: 35 sivua + liitteet 2 sivua
Päivämäärä: 11.12.2025

Tiivistelmä

Tekoäly on yksi 2000-luvun merkittävimmistä muutosvoimista. Sen optimaalinen hyödyntäminen, sekä siihen liittyvien riskien minimointi on keskeinen ja ajankohtainen teema. EU:n tekoälyasetus pyrkii säätelemään tekoälyn kehittämistä ja käyttöönottoa Unionin alueella ja edistämään ihmiskeskeistä ja luotettavaa tekoälyä riskiperusteisen sääntelyn kautta. Asetus esittää ihmisen suorittaman valvonnan keskeisenä suojakeinona korkean riskin tekoälyjärjestelmiä vastaan.

Tässä tutkielmassa tarkastellaan käyttöönottajien roolia valvonnan toteuttamisessa sekä sitä, miten valvojina toimivien henkilöiden pitäminen vastuullisena voi mahdollistaa tehokkaan ihmisen suorittaman valvonnan. Tutkimuksessa käydään läpi tekoälyasetuksen määritelmä ihmisen suorittamasta valvonnasta ja sen vaatimuksista, analysoidaan vastuuta, sekä sen epäreilua kohdentumista tekoälyjärjestelmissä. Lopuksi esitellään, mitä tehokas valvonta käytännössä tarkoittaa, miten se voidaan mahdollistaa vastuun kautta ja mikä on käyttöönottajien rooli sen tukemisessa.

Tutkielman tulokset osoittavat, että käyttöönottajilla on keskeinen rooli ihmisen suorittaman valvonnan operatiivisessa toteuttamisessa. Tutkielma tarjoaa määritelmän, jonka mukaan tehokas valvonta edellyttää lainsäädännön vaatimusten, eettisten periaatteiden ja valvonnan tavoitteiden toteutumista. Tulokset osoittavat, että vastuu parantaa ihmisen kognitiivista suorituskykyä, tilannetajua, moraalista vastuuntuntoa ja tiedon käsittelyä, mikä mahdollistaa tehokkaan valvonnan onnistumisen.

Tutkimus osallistuu kansainväliseen keskusteluun tekoälyn sääntelystä ja ihmiskeskeisestä tekoälystä, yhdistäen vastuun ja ihmisen suorittaman valvonnan käsitteet sekä tarjoten käyttöönottajille havaintoja siitä, miten tehokas ihmisen suorittama valvonta edistää vastuullista tekoälyn käyttöä ja turvaa ihmisen autonomiaa ja perusarvoja.

Avainsanat: Tekoäly, EU:n tekoälyasetus, tehokas ihmisen suorittama valvonta, vastuu, tekoälyn hallinto, tekoälyn johtaminen

TABLE OF CONTENTS

FIGURES

1 Introduction

Artificial Intelligence (AI) has become one of the most transformative forces of the twenty-first century. It is reshaping human lives by augmenting our knowledge, expanding our capabilities (Bhatti et al., 2023) and altering our interactions (Floridi et al., 2018). At the societal level, AI is transforming the world around us by reshaping societies (Arora et al., 2023; Bhatti et al., 2023; Floridi et al., 2018; Floridi & Cowls, 2019), upturning industries (Bhatti et al., 2023), advancing various technologies and their synergies (Arora et al., 2023) and even influencing the very environments we exist in (Floridi et al., 2018). These shifts signify a new era of improved existence, where efficiency is enhanced (Bhatti et al., 2023) and innovation exponential (Arora et al., 2023) – an era, that has even been described as the fourth industrial revolution (see e.g. IBM, 2021.; Marr, 2018.; McKinsey & Company, 2022).

However, the rapid development of AI has also sparked concerns regarding the potential risks and negative consequences (see e.g. Arora et al., 2023; Floridi et al., 2018; Mikalef et al., 2022; Wörsdörfer, 2024) as well as the broader societal and human impact of AI (see e.g. Enqvist, 2023; Floridi et al., 2018; Salim et al., 2024). The risks include a lack of transparency and accountability (Cheong, 2024), loss of human autonomy (Koulu, 2020a), risks to life, physical integrity (Botero Arcila, 2024; Corrêa et al., 2025) and property, as well as threats to fundamental rights and erosion of privacy, human dignity and equality (Botero Arcila, 2024). Additional societal concerns include misinformation and targeted disinformation, as well as algorithmic bias and marginalization and the threats they pose to diversity and inclusion (Arora et al., 2023). Such harms often arise from unintended consequences and are the result of good intentions gone awry, but can also be the result of inadvertent overuse or wilful misuse of AI (Floridi et al., 2018).

Finding a balance between leveraging the opportunities offered by AI, while also ensuring that humans retain meaningful control, has thefore become a central societal challenge (Floridi et al., 2018) and in recent years, several frameworks, guidelines and principles for responsible, ethical and human-centric AI have been introduced (see e.g. European Comission, 2019; OECD, 2024). While these frameworks have provided valuable and necessray guidance, their impact has often remained confined to aspirational principles, rather than enforceable action. As noted by Westerstrand (2025), AI ethics often place excessive emphasis on abstract principles withour implementing practical mechanisms that are needed to ensure operational alignment with these values. Moreover, ithout the force of law, ethical principles cannot guarantee compliance or prevent harmful uses of AI (Robles Carrillo, 2020).

Prompted by the lack of regulation and in response to the growing mistrust of AI systems and the potential threats they pose, in July of 2024, the European Council formally published the EU Artificial Intelligence Act (AIA), which entered into force on the 1st of August 2024 (Artificial Intelligence Act, 2024). The AIA aims to regulate the development and deployment of AI in the European Union (EU) and to mitigate the potential risks associated with its use (Salim et al., 2024). Its predominant goal is ensuring that AI systems are secure, trustworthy and ethical and are used in a way that increases societal wellbeing whilst respecting fundamental human rights (Salim et al., 2024; Wörsdörfer, 2024).

The EU AI Act introduces human oversight as the normative foundation for the relationship between humans and high-risk Artificial Intelligence systems (Beck & Burri, 2024). According to the Act, the purpose of human oversight is to prevent or mitigate risks to health, safety and fundamental rights (European Commission, 2024). Beyond the the aims presented in the regulation, core expectations attributed to human oversight include preserving human autonomy (Beck & Burri, 2024; Enqvist, 2023; Green, 2022; Koulu, 2020a; Methnani et al., 2021), safeguarding human-centric AI (Enqvist, 2023; Sterz et al., 2024) and serving as a critical safeguard against the broader negative impacts of high-risk AI systems (Sterz et al., 2024).

As established, as AI technologies continue to advance and expand across industries, ensuring meaningful and effective human control is becoming increasingly important (Methnani et al., 2021). Yet, the effectiveness of human oversight is widely questioned, with concerns mainly raised regarding its meaningfulness, feasibility and practical effectiveness (Corrêa et al., 2025). Without meaningful implementation, oversight may fail in fulfilling its promises. Furthermore, it risks becoming a symbolic legal measure that provides a false sense of security, enables actors to evade accountability (Green, 2022) and legitimises imperfect systems (Green, 2022; Koulu, 2020b), undermining the very goals of oversight. As (Laux, 2024) notes, how human oversight can be implemented effectively is one of the *"big regulatory questions of our time"*.

A review of the existing literature indicates that research on human oversight – particularly as articulated in the EU AI Act – is still quite limited. Although various studies address the fragility of oversight as a safeguard and the role of system developers in its implemenation, considerably less attention has been devoted to how oversight should be operationalised by deployers, with even the EU Artificial Intelligence Act itself providing minimal guidance. In particular, there is a lack of research on how deployers can contribute to effective human oversight through their role in appointing overseers and their influence in ensuring oversight is effective. This gap highights the

need for a more comprehensive understanding of effective human oversight and the capacities of deployers in facilitating it. This thesis aims to address this gap by examining how holding overseers accountable can enabling effective oversight and what role deployers play in capacitating it. The objective is to provide insights that assist deployers of high-risk AI systems, such as organisations, in meeting the EU AI Act's oversight requirements in an operationally viable and effective manner.

The research questions of this thesis are:

RQ1: How does the EU AI Act define human oversight, and what is the role of deployers in ensuring it?

RQ2: What elements constitute effective human oversight?

RQ3: How can deployers enable effective human oversight by ensuring that overseers are held accountable?

Chapter 2 provides an overview of the EU AI Act, including a detailed discussion of its definition of human oversight, which is further contextualised through relevant AI frameworks and guidelines in the context of the European Union. Chapter 3 examines the concept of accountability, with particular attention to its unfair allocation due to automated systems. Chapter 4 analyses how accountability can enable effective human oversight and explores the role of overseers in implementing it. Finally, in chapter 5, the conclusions, limitations and any future research directions are presented.

While Artificial Intelligence is a technology used worldwide, the scope of this thesis is limited to the European Union and its legal and regulatory context, as the AIA applies only within its legal jurisdiction. Additionally, this thesis is focused on the role of the deployer. It does not provide technical analysis of AI or address technical oversight implementation details. Furthermore, this thesis is based on a literature review, therefore it does not involve primary data collection, but focuses on theoretical and conceptual analysis based on previous literature. Artificial Intelligence has a long and rich history, with research tracing back to the 1950's (see e.g. McCarthy et al., 2006; Turing, 1950) and throughout the years it has gained various definitions (Floridi & Cowls, 2019). Broadly AI can be understood as computer systems that are capable of learning, thinking, deciding and acting like humans (Bhatti et al., 2023). As this thesis does not focus on the technical aspects of AI, a deeper understanding of systems and algorithms is not necessary, and this broad definition can be seen as sufficient.

2 The EU Artificial Intelligence Act and human oversight

The significance of human oversight is widely recognized and thought of as a focal principle for AI development and deployment. Many AI ethics guidelines emphasize oversight as a practical and effective protection against potential harms arising from AI technologies (Koulu, 2020a). Moreover, it is a tempting option for procedural protection, since it can be operationalized relatively easily (Koulu, 2020b). Now, for the first time in history, human oversight over high-risk systems is becoming a legally binding reality through the EU Artificial Intelligence Act.

2.1 Overview of the EU Artificial Intelligence Act

In April of 2021, the European Commission published a proposal titled the *Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence* (European Commission, 2021). Three years later, after extensive negotiations, revisions and political debates, in May of 2024, the European Commission formally adopted the EU AI Act (AIA), – Regulation (EU) 2024/1689, which entered into force on August 1st, 2024 (EU Artificial Intelligence Act, 2024). A historical milestone, the AIA is the first significant government initiative (Wörsdörfer, 2024) and globally the first comprehensive law for AI regulation (Salim et al., 2024; Wörsdörfer, 2024). According to Salim et al., (2024) it is *"set to reshape the landscape of AI development and deployment in Europe."*

The AIA builds on the EU White Paper on Artificial Intelligence (see: European Commission, 2020) and the Ethics Guidelines for Trustworthy AI presented by the High-Level Expert Group on AI (AI HLEG) (see: European Comission, 2019). It is rooted in values set out in Article 2 of the Treaty on European Union (see: European Union, 2016) and in the Charter of Fundamental Rights of the European Union (see: European Union, 2012) – including human dignity, freedom, democracy, equality, the rule of law, human rights and environmental protection. The primary legal basis of the AIA is Article 114 of the Treaty on the Functioning of the European Union, making the Act a part of the EU's Digital Single Market Strategy and New Legislative Framework (Wörsdörfer, 2024), which aim to harmonize rules across the EU, avoid fragmentation and ensure the proper functioning of the internal market. The AIA also complements other initiatives, such as the GDPR and Digital Services Act (Westerstrand, 2025; Wörsdörfer, 2024).

The AIA consists of 12 main titles, 113 articles and 13 annexes (Artificial Intelligence Act, 2024). According to Article 1, the purpose of the EU AI Act is:

> *"To improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation".*

The Act establishes harmonised rules for placing on the market, putting into service and use of Artificial Intelligence systems in the European Union. It prohibits certain AI practices, sets requirements for high-risk AI systems, as well as obligations for their developers and deployers, introduces transparency rules for particular AI systems, regulates general-purpose AI models and provides frameworks for market monitoring, governance, enforcement and innovation support, with particular attention to SMEs and startups. (Artificial Intelligence Act, 2024, Art. 1.)

These objectives are pursued through the implementation of a risk-based framework, set out in chapters II, III and IV, which divides AI systems into four different risk categories: unacceptable risk, high risk, limited risk and minimal risk (Artificial Intelligence Act, 2024). Each level determines how strictly AI should be regulated to ensure that the technology is safe and beneficial. The assessment of risks is based on their potential impact on health, safety and fundamental rights (Salim et al., 2024; Westerstrand, 2025) and, consequently, the aim is to ensure AI systems are used so that they increase societal wellbeing whilst respecting fundamental human rights (Salim et al., 2024; Wörsdörfer, 2024). In line with the level of risk, the AIA prohibits the use of systems deemed to pose unacceptable risks (Chapter II), imposes requirements on high-risk systems (Chapter III), establishes transparency requirements for systems with limited risk (Chapter IV) and provides recommendations and encourages explainability and transparency for those with minimal risk (Chapter X). (Artificial Intelligence Act, 2024.)

The AIA is widely supported and seen as an important first step towards human centric and ethical use of AI (Salim et al., 2024; Wörsdörfer, 2024) while also offering several benefits for organisations, such as single market benefits, securtiy for multinational companies and possibilities to boost an organisations public image and reputation (Wörsdörfer, 2024). Even with its widespread support, the AIA is not without its critics – it is, after all, the third most lobbied EU regulation to date (Westerstrand, 2025). The AIA has been criticised mostly for the lack of implementation guidance and enforcement mechanisms, which would make the transition into compliance clearer. Questions have also been been raised on the account of possible trade offs between ethical principles and corporate interests and the possibility of ethics washing. (Wörsdörfer, 2024.)

Additionally, even if the legally biding nature of the AIA is one of its biggest strenghts, there is a risk of it causing companies to only focus their efforts on compliance, making them drift away from their ethical and moral responsibilities (Westerstrand, 2025). The Act has also faced lobbying due to concerns about a too short compliance deadline (Haeck, 2025).

While certain provisions will apply six months after the Act's entry into force, the EU AI Act is generally set to become fully applicable 24 months after entry into force. An exception is made for specified high-risk system requirements, which will apply 36 months after entry into force. (Artificial Intelligence Act, 2024, Art. 113.) However, in November 2025, under pressure from the United States government, big tech companies like Meta and industry lobby groups, an amendment proposal was presented by the European Commission to postpone parts of the Act's implementation by one year. This proposal is a part of a wider digital simplification package and will still need approval from EU countries and by the European Parliament before becoming applicable. (Haeck, 2025.) At the time of writing, the final implementation timeline for the Act remains uncertain.

2.2 Human oversight in the Artificial Intelligence Act

Chapter III of the Artificial Intelligence Act defines high-risk AI systems (section 1) and their specific requirements (section 2). According to the Act, high-risk systems include, but are not limited to, those intended to be used in biometric identification, critical infrastructure, employment and worker management, law enforcement, administration of justice and democratic processes as well as migration, asylum and border control management. (For complete legal classifications, see: Article 6, Article 7 and the related Annex I and Annex III, Artificial Intelligence Act, (2024)). In short, to comply with requirements set for high-risk systems developers and deployers of high-risk AI systems must implement a continuous risk management system (article 9), adopt data governance and management practices (article 10), maintain technical documentation (article 11) and ensure traceability of system actions (article 12). They are also required to meet transparency obligations (article 13), provide effective human oversight (article 14) and guarantee that systems are accurate, robust and secure (article 15). Provisions for notifying authorities and standardisation norms are also laid out. (Artificial Intelligence Act, 2024.)

Instead of using the term *human control*, which has previously been centric to the discussion around responsible AI, the EU AI Act opts for a softer (but not necessarily weaker) notion of *human oversight* (Beck & Burri, 2024). Article 14 states that the goal of human oversight is to prevent and minimize risks to health, safety and fundamental rights consequent of the use of high-risk AI systems in their intended way or a reasonably foreseeable way of misuse. Requirements imposed

for human oversight include the following: systems need to be designed in a way that allows natural persons to oversee them effectively. The oversight measures need to match the risks, level of autonomy and context of the system. The measures need to be identified and be built into the system by the developer, in a way that they are implementable by the deployer. Additionally, the overseer of the system needs to understand the systems capabilities and limitations, be capable of detecting and addressing issues, not be over-reliant on the system and be able to understand, interpret, critically reflect and override the output of the system or even stop it entirely. For certain systems (defined in 1(a) of annex III), such as remote biometric identification systems, any action or decision the system makes must be verified by at least two individuals, with the necessary competence, training and authority. (Artificial Intelligence Act, 2024, Art. 14.)

These requirements are further developed in Article 26, which lists obligations for deployers of high-risk systems. According to the Article, human oversight must be assigned by deployers to natural persons who have *'the necessary competence, training and authority, as well as the necessary support'.* In addition to this, deployers must implement the oversight measures provided and defined by the developer in practice, according to the instructions provided by the developers and by taking the necessary measures to organize internal processes. (Artificial Intelligence Act, 2024, Art. 26.)

2.3 Human oversight in other frameworks in the European Union

2.3.1 Ethics guidelines for Artificial Intelligence

In April of 2019, before the Artificial Intelligence Act, EU's High-Level Expert Group on AI (AI HLEG), an independent expert group set up by the European Commission in June 2018, presented the Ethics guidelines for trustworthy AI. According to the guidelines, to be trustworthy, AI should be lawful, ethical and robust. All three requirements must be met, since each component is necessary, but not sufficient in itself for achieving trustworthy AI. The framework focuses more on providing guidance on the second and third components, fostering and enabling ethical and robust AI, and aims to go beyond just a list of ethical principles, by providing operational guidance for sociotechnical systems. (European Comission, 2019.)

In chapter II, the guidelines determine 7 key requirements that AI systems should meet, the first of which is human agency and oversight. The others are (II) technical robustness and safety, (III) privacy and data governance, (IV) transparency, (V) diversity, non-discrimination and fairness, (VI) societal and environmental well-being and (VII) accountability. According to the guidelines, AI systems should empower humans while fostering their fundamental rights. Systems should support

human autonomy and decision making, by acting as enablers to a democratic and equitable society, by supporting user agency, fostering fundamental rights and allowing for human oversight. Proper oversight governance mechanisms need to be ensured, either through human-in-the-loop (HITL), human-on-the-loop (HOTL) or human-in-command (HIC) approaches. In HITL the human operator has the ability to intervene in every decision cycle of the AI system. In HOTL the human operator has the capability to intervene during the design and monitoring of the system. In HIC the human operator oversees the overall system and its external impact and has the ability to decide when and how the system is used or not used. The less human oversight, the more extensive testing and stricter governance is needed. (European Comission, 2019.)

Even though legally non-binding, the importance of these guidelines to the design of coherent, trustworthy and human-centric AI is highlighted in article 27 of the AIA. The application of these guidelines is encouraged to all stakeholders, when designing and using AI and when creating codes of conduct in line with the AIA. (European Commission, 2024.)

2.3.2 OECD AI principles

The Organization for Economic Co-operation and Development (OECD) adopted the *Recommendation on Artificial Intelligence* in 2019, making it the first intergovernmental standard on AI. The recommendation was later updated in 2024, to ensure that it was up to date for such a fast-paced industry. The Recommendation contains five AI principles, which are widely known as the OECD AI principles. These principles emphasise that AI systems should respect human rights and democratic values and be designed to allow for appropriate human oversight and determination. They aim to provide guidance for policymakers and AI actors and form a foundation for international cooperation and interoperability. As of December 2025, 38 OECD member countries, including the United States, 10 non-member countries and the European Union have all adhered to the OECD Recommendation. (OECD, 2025.) China, who is a considerable actor in the Artificial Intelligence development and regulation, has not adhered to these principles.

The five principles are (I) inclusive growth, sustainable development and well-being, (II) respect for the rule of law, human rights and democratic values, (III) transparency and explainability, (IV) robustness, security and safety and (V) accountability. The subject of human oversight is introduced in section 1.2, under respect for the rule of law, human rights and democratic values, including fairness and privacy. It states that AI actors should implement mechanisms and safeguards, like human agency and human oversight, to address possible risks arising from use outside the intended

purposes as well as intentional and unintentional misuse. Human oversight is not defined further. (OECD, 2025,)

Even though these recommendations are not explicitly mentioned or highlighted in the AIA nor legally binding, they are still applicable in the EU, since the EU has agreed to adhere to them and therefore worth mentioning.

2.4 Effective oversight

As can be concluded, the AIA divides the responsibility of implementing human oversight between the developer of AI systems, responsible of technically ensuring that the system can be overseen by a natural person, and the deployer, responsible for the organisational implementation through human operators (Artificial Intelligence Act, 2024; Constantino, 2025; Corrêa et al., 2025; Laux & Ruschemeier, 2025), who are arguably also responsible for the oversight (Corrêa et al., 2025). The technical task of identifying the oversight measures is therefore entirely placed on the developer, while the deployer is the executor of human oversight in practice. The roles are clear; however, the AIA falls short on providing ample information on what would make human oversight effective (Laux, 2024).

For something to be effective, it must achieve its goals and produce the desired result. If the findings presented in chapters 2.2 and 2.3 are combined, it can be determined that the requirements that fall under the deployers jurisdiction and are specifically related to the human overseers include the following: Human oversight must be assigned to competent natural persons who are supported by the necessary training and ability to authorise oversight. In addition to this, deployers must implement the oversight measures provided and defined by the developer in practice, according to the instructions of the developers and by taking the necessary measures to organize internal processes. Moreover, as the deployer is responsible of assigning oversight tasks, must the deployer ensure that the appointed overseer of the system properly understands the system’s capabilities and limitations, is able to detect anomalies, capable of addressing issues and aware of the tendency to be over-reliant on the system and, naturally, aims to avoid this. In addition, overseers must be capable of to understanding, interpreting and critically reflecting the systems output and when necessary, be able to decide to not utilize the output or stop the entire system from running. It can be hypothesised that, although not explicitly demanded in the text, meeting the requirements outlined above necessitates that deployers ensure overseers are provided with optimal working conditions and are not subject to organisational constraints, time pressures or excessive interface complexity, since all of these are factors that undermine their capacity to effectively fulfil their responsibilities.

When applying the recommendations set out for human oversight in the AI HLEG ethics guidelines and OECD AI principles, it can be determined that overseers must be empowered by deployers to protect fundamental rights, human autonomy and decision making, whilst fostering democratic and equitable values. Consequently, it can be determined that deployers are also responsible of ensuring that overseers are aware of their position as safeguards against the risks associated with the AI systems under their supervision and of their societal, moral and ethical implications. To fulfil these responsibilities effectively, deployers must support the overseer's professional agency and authority and ensure they can act without external pressure or constraints. Ultimately, it can be argued that only by meeting the above-mentioned legal requirements and fulfilling the ethical principles, can oversight be effective.

3 Accountability and AI governance

AI and the challenges posed by it need to be approached from an interdisciplinary perspective and only a comprehensive understanding of all necessary perspectives can facilitate responsible Artificial Intelligence (Robles Carrillo, 2020). Accountability is one of those perspectives. Accountability is a multifaceted phenomenon that can be studied from various theoretical perspectives and through various methods (Bovens et al., 2014). In governance, accountability holds a strong promise of ensuring fair and equitable outcomes and is of great importance in practice, as it helps to ensure the legitimacy of the governing processes. In political discourse and documents, accountability is much utilized, because it promises transparency and promotes the idea of trustworthiness, especially among external stakeholders. (Bovens, 2007.) Despite this, accountability has been described as an evocative political word, that is used to "*patch up a rambling argument, to evoke an image of trustworthiness, fidelity and justice or to hold critics at bay".* Its elusive nature is further reinforced by the fact that its meaning varies between individuals. (Bovens, 2007.)

In Artificial Intelligence, accountability and making sure it is not just used for empty promises, takes on an entirely new scope. AI systems may act on their own accord, without human control or intervention (Santoni de Sio & van den hoven, 2018) and make decisions that vary in fairness and consequence. Therefore, ensuring fair accountability allocation and mitigation of unacceptable risks becomes difficult, yet more important than ever (Santoni de Sio & van den hoven, 2018).

3.1 Artificial Intelligence governance

AI principles are continually evolving, as every new emerging AI technology and advancement introduces a new set of conditions that must be considered (Papagiannidis et al., 2025). Consequently, responsible AI requires compliance with a wide range of standards through the whole lifecycle of AI applications (Robles Carrillo, 2020). Concurrently, the use and deployment of Artificial Intelligence should align with an organisation's overarching strategies, objectives and values while also complying with legal requirements and upholding ethical standards (Mäntymäki et al., 2022). In practice, this means that AI technology should be leveraged to its maximum value, while simultaneously minimising risks and preventing unintended consequences (Mäntymäki et al., 2022; Papagiannidis et al., 2025). The complexity of these demands emphasises that increased focus should be applied to establishing processes, mechanisms and structures that meet these goals (Papagiannidis et al., 2023). Ultimately, these multifaceted challenges and obligations can only be effectively addressed through comprehensive Artificial Intelligence governance.

AI governance is not an isolated process – it occurs in a complex organisational landscape, as a part of the overall governance of an organisation (Mäntymäki et al., 2022), which, broadly defined, refers to the ways in which corporations are managed, operated, regulated and financed (Cihon et al., 2021). Several frameworks, guidelines and principles with varying theoretical backgrounds for the responsible governance of AI have been introduced in recent years (Papagiannidis et al., 2025; Westerstrand, 2025). For example, Papagiannidis et al. (2025) define responsible AI governance as "*a set of practices for developing, deploying, and monitoring AI applications in a safe, trustworthy, and ethical manner that ensures appropriate functionality of AI over the entire lifecycle*". Similarly, Mäntymäki et al. (2022) define AI governance as "*a system of rules, practices, processes, and technological tools that are employed to ensure an organization's use of AI technologies aligns with the organization's strategies, objectives, and values; fulfils legal requirements; and meets principles of ethical AI followed by the organization*". Overall, the slightly varying definitions provide procedural guidance for regulatory, economic and organisational boundaries for responsible AI, while emphasising the importance of moral, social and ethical responsibility.

Although there is still a lack of empirical evidence, succesful AI governance can be tied to multiple benefits for businesses and organisations. Succesfull AI governance enhances an organisation's sense of external legitimacy as a reliable and trustworthy actor and decreases internal threat perception, while improving synergy and cooperation, resulting in higher employee job satisfaction, productivity and overall well-being. (Papagiannidis et al., 2025.) Responsible use of AI can also lead to higher profitability, greater customer retention and satisfaction as well as better strategic alliances and partnerships (Enholm et al., 2022). Ultimately, comprehensive AI governance enables deployers to implement the legal requirements imposed in the AIA, such as human oversight, in practice.

3.2 Accountability

Historically, accountability in socio-technical systems has been relatively straightforward to assign. Responsibility of the consequences for the operation of machines has been traditionally ascribed to the operator of the machine, if it operates as specified by the manufacturer or, respectively, to the manufacturer, if the machine does not work as prescribed. This approach is based on the notion that control is a necessary condition of accountability. (Matthias, 2004.) However, as autonomous systems are increasingly integrated into complex socio-technological systems, who actually is accountable has become more difficult to determine as overseers are loosing their control over the systems (see e.g Elish, 2019; Matthias, 2004). As AI systems themselves cannot be legally held

accountable, since they do not have legal personhood (Corrêa et al., 2025), and the humans assigned to oversee the autonomuous systems may not truly understand or control them (Elish, 2019; Matthias, 2004), accurately locating who is accountable has become increasingly important (Elish, 2019).

In his study of accountability in European governance, Bovens (2007) defines accountability in a narrow sense, as a specific social relation – *"a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences*". This relationship is highlighted in figure 1. The actor can be an individual or an organisation, the forum a (superior) person, a journalist (or another external contributor) or an agency that has the possibility to debate and judge the actions of the actor and assign suitable, positive or negative consequences. This can be divided into three stages: First, the actor is obliged to inform the forum directly about their conduct, by providing documentation, explanations and justifications. Second, the forum must be able to interrogate the actor and question the adequacy of the provided documentation and the legitimacy of the conduct. Third, the forum can pass judgement on the conduct of the actor and assign consequences depending on the nature of the action. (Bovens, 2007.)

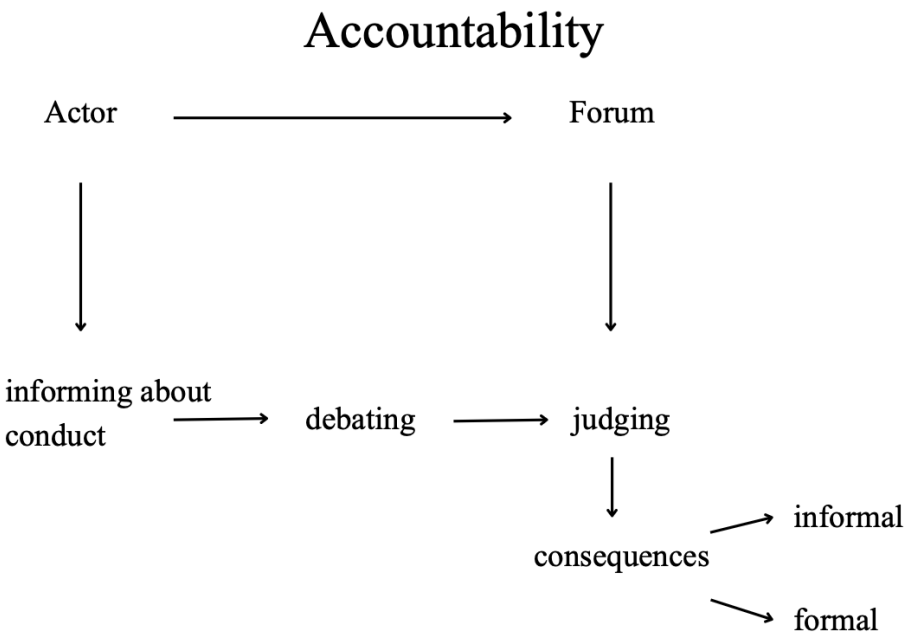

Figure 1. Elements of an accountability relationship (As in Bovens, 2007)

Bovens, (2007) divides accountability into different types based on the nature of the forum, nature of the actor, nature of the conduct and nature of the obligation. The aim of this divide is to help answer four important questions: to *whom*, by *who*, *what* about and *why* accountability should be rendered. Based on the nature of the actor *(by who),* Bovens (2007) identifies four accountability types, one for corporate accountability – in which the organisation can be held accountable for the collective outcome because of its independent legal status – and three focused on individual actors. The first is hierarchical accountability, in which accountability starts from the top and works its way down the chain of command. The actor at the top of the chain is externally and internally accountable for all those below them and middle managers act as both actors and forums, being accountable *to* their superiors (actor) and holding account *for* their inferiors (forum). The second is collective accountability, in which any member of the organisation can be held accountable for the conduct of the organisation as a whole. The third is individual accountability, in which each individual is liable based on their personal contribution, not their position. These strategies are susceptible to accountability inadequacies, either as deficits (lack of accountability arrangements) or excesses (dysfunctional accumulation of accountability mechanisms). (Bovens, 2007.)

Based on the nature of the obligation (*why),* which can be either formal or informal, there are three forms of accountability: vertical, diagonal and horizontal. In vertical accountability the forum wields power over the actor, like in most hierarchical relationships, such as in organisations (e.g. managers overseeing teams). Horizontal accountability is less formal and more moral in nature, like the accountability an organisation (actor) might have for external stakeholders or civil society (forum). Diagonal accountability is accountability "*in the shadow of a hierarchy"*, so not through a direct hierarchical relationship but informally under a forum's power, like accountability to external regulators. A single accountability relationship can be classified in various categories. Although accountability may contain various diverse aspects, in its most concise definition, accountability can be understood as "*the obligation to explain and justify conduct"*. (Bovens, 2007.)

3.3 Unfair distribution of accountability

As established, AI brings new challenges to accountability and its fair allocation. These challenges can be illuminated through Matthias' (2004) and Elish's (2019) theories of accountability in socio-technical contexts. Matthias (2004) introduces the concept of responsibility gaps, which were first introduced in relation to "learning automata", but have been more recently discussed in relation to AI (see e.g. Santoni de Sio & Mecacci, 2021). Responsibility gaps are deficits of accountability that emerge as machines become increasingly autonomous and humans no longer hold control over

them. These gaps cannot be bridged through traditional means of responsibility ascription, as it is possible that the responsible actor no longer possesses sufficient control over a machine's actions to legitimately assume responsibility for them. (Matthias, 2004.) This is closely related to the relationship between an AI system and a human overseer, who, under the AIA, is expected to possess the necessary knowledge and competence to supervise the system. However, since the AIA offers very limited guidance or specific requirements for carrying out such oversight, ensuring that human overseers are actually in control remains challenging.

Elish (2019) approaches this issue through introducing the concept of a moral crumple zone, where accountability is deflected off automated systems towards human operators, who possess only limited knowledge, capacity or control over the system. The integrity of a system is preserved at the expense of the nearest human operator who absorbs the blame, inadvertently taking advantage of the operator to fill the gaps in accountability. The technology itself is maintained as faultless, while the human operator becomes the apparent weak point of the system. (Elish, 2019.) This issue is, again, closely related to the relationship between an AI system and a human overseer.

In addition to these more extreme cases, there are accountability deficits that arise from the complexity of the entire causal chain of an AI system, not merely from the system's and actor's relation itself. Identifying who is accountable becomes exceedingly difficult, due to problems coined as the problem of many hands and the problem of many eyes. The problem of many hands manifests because of the complexity of an AI system through its lifecycle, from early-stage design and training to ongoing supervision and implementation. It becomes difficult to unravel who has contributed in what way, to what part of the system and, consequently, who should be held accountable if the system does not work as intended. The problem of many eyes on the other hand refers to the problem of multiple accountability forums that hold an actor accountable and all demand different types of justification. This is particularly an issue for organisations and public bodies, potentially leading to accountability excesses and over complication of accountability networks and relationships. (Bovens, 2007.) Accountability gaps can also form purely from human disagreements, if participants disagree on if they share an accountability relationship, what its terms are and if they fail to uphold them (Lechterman, 2024).

In conclusion, it can be determined that in the context of Artificial Intelligence, accountability is easily eroded, overcomplicated, allocated unfairly or not allocated at all. Together these concepts highlight the need for effective human oversight, as these issues are mainly the result of the loss of

human control or incompetent understanding of Artificial Intelligence systems. Both of which are merits that effective oversight aims to mitigate.

4 The role of the deployer and fulfilling the promises of oversight

4.1 Objectives and promises of human oversight

According to the AIA, the main aim of human oversight is to prevent or minimise the risks to health, safety and fundamental rights (European Commission, 2024). In addition to this a core expectation attributed to human oversight is the protection of human autonomy (Beck & Burri, 2024; Enqvist, 2023; Green, 2022; Koulu, 2020a; Methnani et al., 2021) and human-centric AI (Sterz et al., 2024). Human autonomy, though not explicitly defined in the AIA or any related EU AI regulation or framework, is strongly associated with the right to dignity and liberty. To ensure that AI systems are developed, deployed and used in a trustworthy manner, protecting human autonomy is a fundamental ethical principle that must be respected (European Comission, 2019). It is an integral part of most AI ethics guidelines and regulations and the EU's overall policy towards Artificial Intelligence (see e.g.; European Comission, 2019; Artificial Intelligence Act, 2024). Human-centric AI, on the other hand, refers to the design, development and deployment of AI systems in a way that places human beings at their core, by embedding human values and ethical considerations throughout their structure. Its aim is to meet human needs, safeguard individual rights, like human autonomy, and enhance overall human well-being. (Enqvist, 2023.)

Purely from an operational perspective, the aim of human oversight is to ensure the presence of human judgment in decisions that significantly affect individuals. Essentially the human is playing the role of quality control, acting as protection against harmful, mistaken or biased algorithmic decisions. (Green, 2022.) This demonstrates that human judgement is highly valued and that we attribute the ability to produce legitimacy and justifications for decisions to humans, a task we think fundamentally impossible for machines (Koulu, 2020b). All of this is based on the idea that humans are able to better make complex decisions and deductions (Enqvist, 2023) and on society's tendency to attribute the fears and risks associated with AI and automated systems to the technology itself, not humans (Koulu, 2020a).

It is evident, that human oversight holds significant promise and bears the heavy burden of navigating a multitude of interrelated expectations. Its effectiveness in practice is, therefore, a question of utmost significance. As argued in chapter 2.4, to be thought of as effective, human oversight must meet its legal requirements whilst fulfilling relevant ethical principles. When the above-mentioned objectives are added, it can be concluded that to be effective, human oversight must, in addition, deliver on its promises.

4.2 Holding overseers accountable

Accountability is recognized as essential for ensuring meaningful human control (Methnani et al., 2021) of which human oversight is a critical component (Beck & Burri, 2024). Therefore, it can be argued that holding overseers accountable can ensure effective human oversight. The deployer acts as the forum, assigning accountability to the overseer who acts as the actor with the obligation to explain and justify conduct.

According to the legal requirements for oversight, discussed in chapter 2.2, the appointed human overseer must properly understand the systems capabilities and limitations, be able to detect anomalies and capable of addressing potential issues. Research indicates that accountability can enhance cognitive complexity, improve human judgement and increase situational understanding. Specifically, accountability can comply decision makers to seek out additional, more comprehensive information and lead to them engaging in more complex information processing. (Skitka et al., 2000.) When accountability is likened to motivation to perform well, research shows that it leads to people initiating and intensifying goal-directed behaviour and maintaining epistemic access which can be defined as *"sufficient knowledge of their decision situation*" (Sterz et al., 2024). Additionally, Skitka et al. (2000) highlight that accountability is particularly beneficial in contexts where high situational awareness is associated with better decision making. This closely aligns with human oversight positions. Taken together, these results suggest that accountability leads to the fulfilment of the aforementioned requirement by fostering a deeper understanding of the system's capabilities, limitations and operational context, due to an increase in cognitive involvedness, improved willingness to seek out additional information and increased situational awareness. Consequently, this will lead to the overseer being more equipped to monitor the system in an informed and competent manner.

Human overseers are also required to remain aware of the tendency to be over-reliant on the system, otherwise known as automation bias, and aim to avoid this. Automation bias occurs when actors ascribe greater power and authority to automated systems over other resources or even their own capabilities. This leads to decisions that are not based on a thorough information analysis but are strongly biased by the system's generated advice. Automation bias can lead to errors of omission and commission. An error of omission is when an actor fails to respond to a system mistake. An error of commission, on the other hand, is when the actor follows the advice of the system, even if that advice is incorrect. (Parasuraman & Manzey, 2010.) Automation bias occurs, mainly, because humans tend to choose the path of least effort, both in cognitive effort and in practice, automation is

trusted as superior in capability and because accountability is diffused between the actor and the system, leading to the actor to believing themselves less accountable. All of this leads to the actor reducing their own effort in the monitoring task and in analysing all available information, thus into automation bias. Automation bias is closely linked with automation complacency, in which the actor might fail to miss the incorrect working of the system due to an unjustified assumption that the system is working as it should. (Parasuraman & Manzey, 2010.)

Parasuraman & Manzey (2010) determine that susceptibility to automation bias is dependent on how accountable actors perceive themselves. Additionally, a study by Skitka et al. (2000) demonstrates that accountability can reduce the tendency to make errors of omission and commission, which, are the two forms of automation bias (Parasuraman & Manzey, 2010). Therefore, it can be determined that accountability can reduce over reliance on AI systems, therefore fulfilling the requirement efficiently. It is also important to note, that in addition to mitigating automation bias, accountability can be beneficial in alleviating accountability gaps (Wagner et al., 2023) that are, as established in chapter 3.3., a considerable risk associated with automatous systems such as AI. This constitutes an additional benefit of accountability in enhancing the overall deployment of responsible AI.

In addition to the requirements discussed above, human overseers must be capable of correctly interpreting the system's outputs, deciding not to use the system when needed, overriding or reversing its results and intervening in its operation, including stopping it entirely when necessary. While the system must evidently technically allow for such measures, accountability can enable these functions in an operational capacity. As established earlier, accountability can enhance cognitive ability and encourage overseers to seek out broader relevant information (Skitka et al., 2000; Sterz et al., 2024), both of which are essential traits in accurately interpreting system outputs and making informed intervention decisions. Research further shows that accountability can improve task performance, especially in monitoring and tracking tasks (Skitka et al. 2000), motivate actors to maintain high self-control and lead them to uphold intentions that are in line with their assigned responsibilities (Sterz et al., 2024). Moreover, accountability creates pressure to provide well-founded justifications for decisions (Skitka et al., 2000) and similarly, as Schillemans (2022) finds, accountability raises perceived stakes for actors, leading to higher quality of judgement, by compelling them to allocate more time and effort into tasks. Although this can be seen as a negative, if the aim of a system is to be as effective as possible, it aligns well with the objectives of human oversight, which emphasise deliberation and critical engagement rather than speed. Combined these findings establish that accountability enhances human oversight by improving task

performance, encouraging more intentional actions, motivating better self-control and compelling clearer justifications for actions. These qualities directly support the overseer's ability to correctly interpret a system's outputs, increasing preparedness to question, override or stop the system, when necessary, ultimately ensuring that oversight is effective.

In addition to meaningfully fulfilling legal requirements, to be effective, oversight must also uphold ethical values and meet its promised goals. As established, the overarching aim of human oversight includes protecting the health and safety of individuals, safeguarding fundamental rights and preserving human autonomy and human-centricity in Artificial Intelligence. Considering that high-risk AI systems may threaten all these values, it is essential to ensure that human overseers comprehend the moral, ethical and societal consequences of their actions. This aligns with the recommendations of the ethical guidelines and principles for AI, discussed further in chapter 2.3, which demand that that overseers must protect fundamental rights, human autonomy and decision making, whilst fostering democratic and equitable values. Overseers should be aware of their position as safeguards against the risks associated with the AI systems under their supervision and of their societal, moral and ethical implications.

Felt accountability is fundamental in fostering such comprehension and enabling responsible oversight. As Pesch (2015) argues, if people are held accountable for their actions, they are more likely to recognize, accept and fulfil their moral and social responsibilities. Similarly Bovens et al. (2014) contend that when actors understand the consequences of their actions and the that way in which they might be held accountable, both during the task and in its eventual outcomes, as opposed to merely following procedures, even when those procedures might be flawed or insufficient, they are more likely to pay meaningful attention to said task. This underscores the importance of accountability in motivating moral engagement and thereby strengthening the quality of human oversight and its effectiveness in practice.

In conclusion, it can be determined that overseers understanding their accountability and perceiving themselves held accountable can enable oversight to meet its legal obligations, endorse essential ethical principles and ensure that oversight fulfils its promises, therefore ensuring its effectiveness.

4.3 Operationalising oversight

In addition to fulfilling the requirements discussed in the previous chapter that are more related to the human overseer's capabilities, deployers are also responsible of ensuring that overseers are

granted appropriate authority and of implementing the oversight measures in practice by taking the necessary measures to organize internal processes. This requires comprehensive AI governance.

According to Weill (2009) for information technology (IT) governance to be successful, organisations need to allocate specific decision rights and accountabilities. Governance must be implemented through clear mechanisms, like individual roles, committees or teams. These mechanisms must clarify how each decision will be made and who is held accountable. (Weill, 2009.) This same logic extends directly to AI governance and the aforementioned compliance requirement. Defined mechanisms, distinct decision rights and accountability structures are a necessity in enabling and operationalising effective human oversight. It is important to note that, as established previously, AI systems themselves cannot be held accountable due to their lack of legal personhood. Thus, accountability still occurs between human actors, with AI only as an intermediary (Nabben, 2024). Therefore, assigning the necessary roles and authority as solely between human actors is relevant.

Finally, in addition to internal restructuring, deployers are responsible of ensuring that the overseers receive the necessary training and, ultimately, are capable of executing their tasks competently. The limitations of human capabilities to control and provide oversight over complex technological systems are widely acknowledged (Beck & Burri, 2024; Green, 2022; Holzinger et al., 2025; Koulu, 2020a, 2020b; Laux, 2024; Parasuraman & Manzey, 2010) further enforcing need of necessary training. Lack of sufficient oversight can result from various human errors: from boredom to alert fatigue (Elish, 2019; Koulu, 2020b) and decision biases, like automation bias and automation complacency, discussed in more detail in chapter 4.2. Additionally, there is a risk of algorithmic aversion, which occurs when actors prefer human advice over algorithmic predictions, even when the algorithm has been shown to be more accurate (Dietvorst et al., 2015). Consequently, is it as simple as human oversight being inherently flawed if it is built on the assumption of an idealised human overseer (Koulu, 2020b). Therefore, implementing effective oversight requires not only organisational structural clarity, as established above, but also appropriate system implementation and interaction design in a way, that does not impose unrealistic expectations on human operators (Koulu, 2020b), takes into account human cognitive limits (Koulu, 2020b; Laux, 2024) and trains overseers accordingly.

5 Conculsions

This thesis set out to examine what constitutes effective human oversight in Artificial Intelligence, how overseers being held accountable can enable it and what is the role of system deployers in its practical implementation. The fist question of this thesis asked how the EU AI Act defines human oversight, and what is the role of deployers in ensuring it. The found answer was that the Act defines human oversight as a safeguard against high-risk AI systems and the many risks they pose – ranging from threats to health, safety, and fundamental rights, to the protection of human autonomy and of human-centric AI.

The role of the deployer is to execute the practical implementation of human oversight, by ensuring that necessary organisational measures are taken to enable oversight in practice and assigning oversight to natural persons who possess the necessary competence, training and authority. Consequently, as deployers are responsible of assigning competent overseers, they are responsible of ensuring that the overseers themselves meet the imposed requirements, which essentially demand that overseers are able to critically interpret system outputs, detect anomalies, avoid over-reliance on automated systems, and be capable of intervening when necessary. This responsibility is highlighted in figure 2, with the added responsibility of holding overseers accountable, which, as discussed in more detail below, is essential in ensuring effective human oversight and can only be executed by the deployers.

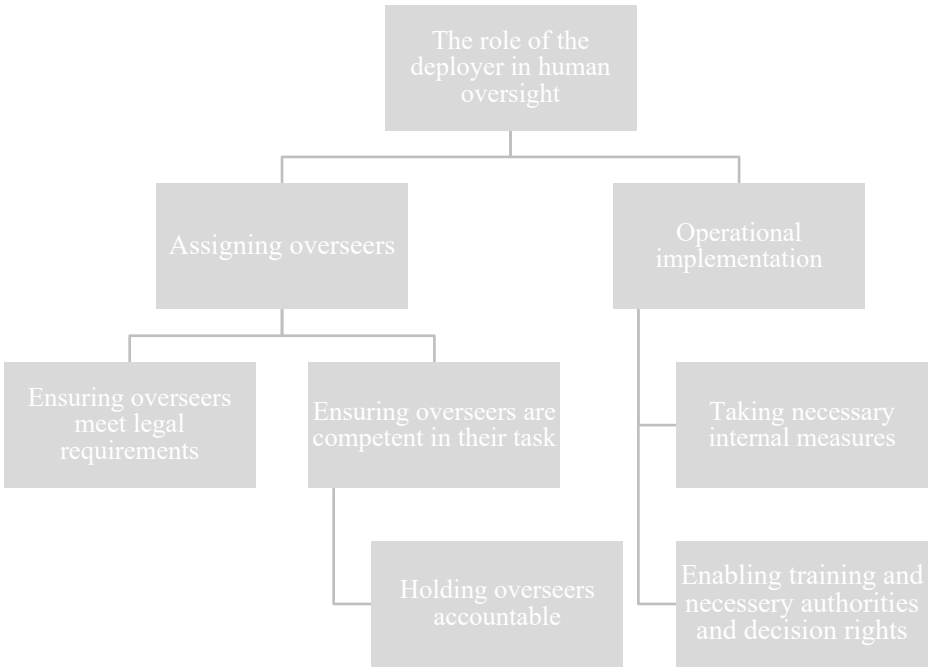

Figure 2. The role of the deployer in human oversight

The second question explored was what constitutes effective human oversight. The findings indicate that oversight is effective when it meets the legal requirements of the EU AI Act, adheres to core ethical principles and fulfils on its ultimate objectives and goals. Only by fulfilling all the three conditions, highlighted in figure 3, can oversight be considered effective.

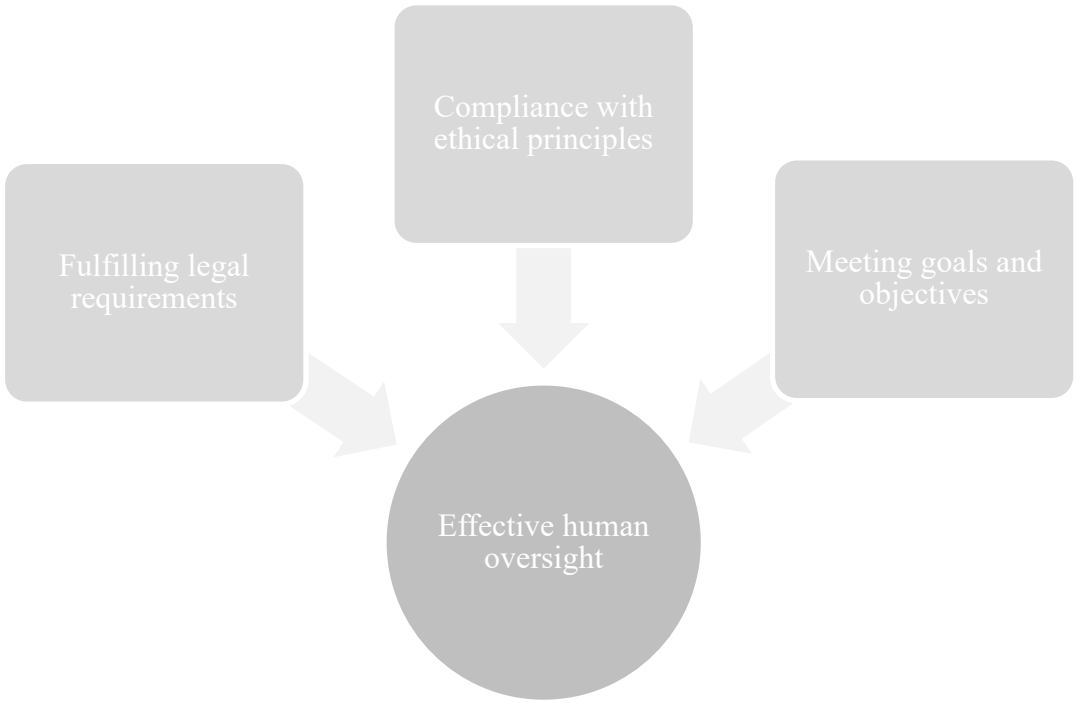

Figure 3. Requirements of effective human oversight

The third research question addressed how deployers can enable effective human oversight by ensuring that overseers are held accountable. Accrding to the findings, accountability is the relationship between an actor and a forum, in which the actor has the obligation to explain and justify conduct. This relation was found to be a key mechanism for effective oversight, as accountability enhances human cognitive complexity, increases situational awareness, reduces the tendency to over-rely on automated systems and improves task performance, overall assisting in overseers being competent in their tasks and meeting both legal and ethical requirements, that are, as established, necessary for effective oversight. Additionally, accountability also encourages moral engagement and social responsibility, ensuring that human oversight does more than simply exist on paper – it fulfils its promises and therefore fills all the requirements of effective human oversight.

In conclusion, the EU AI Act is globally the first AI regulation and, ultimately, it is recongnised as an important step towards human centric and ethical use of AI. The AIA, in line with ethical frameworks, introduces human oversight as the protector of human-centric Artificial Intelligence.

However, it falls short on guiding its implementation and fails to provide guidance on what effective oversight would look like in practice. Without meaningful implementation, oversight may fail in fulfilling its promises by becoming a symbolic legal measure that legitimises imperfect systems, consequently undermining its intentions.

Therefore, the meaningful implementation of oversight, as well as ensuring its effectiveness, is a pressing and highly consequential current issue. As the deployers make up one half of the parties responsible of implementing human oversight, their role is exceptionally important. Deployers bear the responsibility of holding overseers accountable and for ensuring that oversight is operationalised as intended, thereby assuming ultimate responsibility for the effectiveness of human oversight. This relation is highighted in figure 4. In conclusion, to put it simply, the effectiveness of oversight depends on overseers being held accountable, but its realisation ultimately rests on deployers fulfilling their responsibilities.

Figure 4. The relation between deployers and effective human oversight

This thesis contributes to the understanding of how oversight can be operationalized in practice, offering insights for deployers of high-risk AI systems, such as organizations, who bear a legal and ethical responsibility for the AI they implement. Yet, this is only the beginning. The benefits of accountability described here remain largely theoretical, with much of the supporting evidence drawn from related research domains. The research on this area is in its infancy, and much more future research, especially empirical studies, are needed to illuminate efficient oversight and to provide concrete guidance for deploying it oversight in practice.

References

Arora, A., Barrett, M., Lee, E. L., Oborn, E., & Prince, K. (2023). Risk and the future of AI: Algorithmic bias, data colonialism, and marginalization. *Information and Organization*, *33*, 100478. https://doi.org/10.1016/j.infoandorg.2023.100478

Beck, J., & Burri, T. (2024). From "human control" in international law to "human oversight" in the new EU act on artificial intelligence. In *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*. Edward Elgar Publishing. https://doi.org/10.4337/9781802204131.00014

Bhatti, D. S., Tariq, N., Ali, Z., & Raza, U. A. (2023). AI's challenge to ethics and law: Privacy, bias, and beyond. *2023 25th International Multitopic Conference (INMIC)*, 1–7. https://doi.org/10.1109/INMIC60434.2023.10466146

Botero Arcila, B. (2024). AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight? *Computer Law & Security Review*, *54*, 106012. https://doi.org/10.1016/j.clsr.2024.106012

Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework. *European Law Journal*, *13*(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Bovens, M., Goodin, R. E., Schillemans, T., Patil, S. V., Vieider, F., & Tetlock, P. E. (2014). Process Versus Outcome Accountability. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford Handbook of Public Accountability*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0002

Cheong, B. C. (2024). *Transparency and Accountability in AI Systems: Safeguarding Wellbeing in the Age of Algorithmic Decision-Making* (SSRN Scholarly Paper No. 4961260). Social Science Research Network. https://doi.org/10.3389/fhumd.2024.1421273

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate Governance of Artificial Intelligence in the Public Interest. *Information*, *12*(7), 275. https://doi.org/10.3390/info12070275

Constantino, J. (2025). Accountable AI: It Takes Two to Tango. In R. Gsenger M. MSc & M.-T. Sekwenz (Eds.), *Digital Decade: How the EU Shapes Digitalisation Research* (1st ed., pp. 95–114). Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783748943990-95

Corrêa, A. M., Garsia, S., & Elbi, A. (2025). Better together? Human oversight as means to achieve fairness in the European AI Act governance. *Cambridge Forum on AI: Law and Governance*, *1*, e29. https://doi.org/10.1017/cfl.2025.10010F

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Elish, M. C. (2019). Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial Intelligence and Business Value: A Literature Review. *Information Systems Frontiers*, *24*(5), 1709–1734. https://doi.org/10.1007/s10796-021-10186-w

Enqvist, L. (2023). 'Human oversight' in the EU artificial intelligence act: What, when and by whom? *Law, Innovation and Technology*, *15*(2), 508–535. https://doi.org/10.1080/17579961.2023.2245683

European Commission. (2019). *Ethics guidelines for trustworthy AI*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/346720

European Commission. (2020). *White Paper on Artificial Intelligence: A European approach to excellence and trust*. Publications Office of the European Union. https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_en?filename=commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. (2021). *Proposal for a regulation laying down harmonised rules on artificial intelligence (AI Act) (COM 2021 206 final).* EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

European Parliament & Council of the European Union. (2024). *Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union, L 168*, 1–202. http://data.europa.eu/eli/reg/2024/1689/oj

European Union. (2012). *Charter of Fundamental Rights of the European Union*. Official Journal of the European Union, C 326, 391–407. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:12012P/TXT

European Union. (2016). *Consolidated versions of the Treaty on European Union and the Treaty on the Functioning of the European Union (CELEX 12016ME/TXT).* EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12016ME/TXT

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, *45*, 105681. https://doi.org/10.1016/j.clsr.2022.105681

Haeck, P. (2025, November 19). The EU promised to lead on regulating artificial intelligence. Now it's hitting pause. *Politico*. https://www.politico.eu/article/the-eu-wanted-to-lead-on-regulating-ai-now-its-hitting-pause/ Retrieved 05.12.2025

Holzinger, A., Zatloukal, K., & Müller, H. (2025). Is human oversight to AI systems still possible? *New Biotechnology*, *85*, 59–62. https://doi.org/10.1016/j.nbt.2024.12.003

IBM. (2021, August 4). *What is Industry 4.0 and how does it work? https://www.ibm.com/think/topics/industry-4-0* Retrieved 04.12.2025

Koulu, R. (2020a). Human control over automation: EU policy and AI ethics. *European Journal of Legal Studies*, *1*, 9–46. https://doi.org/10.2924/EJLS.2019.019

Koulu, R. (2020b). Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law*, *27*(6), 720–735. https://doi.org/10.1177/1023263X20978649

Laux, J. (2024). Institutionalised distrust and human oversight of artificial intelligence: Towards a democratic design of AI governance under the European Union AI Act. *AI & SOCIETY*, *39*(6), 2853–2866. https://doi.org/10.1007/s00146-023-01777-z

Laux, J., & Ruschemeier, H. (2025). Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI. *European Journal of Risk Regulation*, 1–16. https://doi.org/10.1017/err.2025.10033

Lechterman, T. M. (2022). The Concept of Accountability in AI Ethics and Governance. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford Handbook of AI Governance* (1st ed., pp. 164–182). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.10

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

Marr, B. (2018, Septermber 2). What is Industry 4.0? Here's A Super Easy Explanation For Anyone. *Forbes*. Retrieved December 5, 2025, from

https://www.forbes.com/sites/bernardmarr/2018/09/02/what-is-industry-4-0-heres-a-super-easy-explanation-for-anyone/

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine*, *27*(4), 12. https://doi.org/10.1609/aimag.v27i4.1904

McKinsey & Company. (2022, August 17). *What is industry 4.0 and the Fourth Industrial Revolution?https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-industry-4-0-the-fourth-industrial-revolution-and-4ir#/* Retrieved 04.12.2025

Methnani, L., Aler Tubella, A., Dignum, V., & Theodorou, A. (2021). Let Me Take Over: Variable Autonomy for Meaningful Human Control. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.737072

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, *31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Nabben, K. (2024). AI as a constituted system: Accountability lessons from an LLM experiment. *Data & Policy*, *6*, e57. https://doi.org/10.1017/dap.2024.58

OECD. (2024). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI Governance: Identifying Best Practices and Potential Barriers and Outcomes. *Information Systems Frontiers*, *25*(1), 123–141. https://doi.org/10.1007/s10796-022-10251-y

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, *34*(2), 101885. https://doi.org/10.1016/j.jsis.2024.101885

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Pesch, U. (2015). Engineers and Active Responsibility. *Science and Engineering Ethics*, *21*(4), 925–939. https://doi.org/10.1007/s11948-014-9571-7

Robles Carrillo, M. (2020). Artificial intelligence: From ethics to law. *Telecommunications Policy*, *44*(6), 101937. https://doi.org/10.1016/j.telpol.2020.101937

Salim, S., Jayasudha., J. S. & Soniya, B., (2024). Ensuring Ethical AI: Unpacking the Significance of Risk Analysis Under the European Union's Artificial Intelligence Act, *2024 IEEE Region 10 Symposium (TENSYMP)*, 1-6. https://doi.org/10.1109/TENSYMP61132.2024.10752133

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, *5*. https://doi.org/10.3389/frobt.2018.00015

Schillemans, T. (2022). Accountability and the Quality of Regulatory Judgment Processes. Experimental Research Offering Both Confirmation and Consolation. *Public Performance & Management Review*, *45*(3), 473–498. https://doi.org/10.1080/15309576.2022.2040034

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., & Langer, M. (2024). On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2495–2507. https://doi.org/10.1145/3630106.3659051

Turing, A. M. (1950). Computing machinery and intelligence, *Mind*, *Volume LIX, Issue 236*, 433 – 460. https://doi.org/10.1093/mind/LIX.236.433

Wagner, B., De Gooyert, V., & Veeneman, W. (2023). Sustainable development goals as accountability mechanism? A case study of Dutch infrastructure agencies. *Journal of Responsible Technology*, *14*. https://doi.org/10.1016/j.jrt.2023.100058

Weill, P., Ross, J. W. (2009). *IT savvy: What top executives must know to go from pain to gain*. Harvard Business Press.

Westerstrand, S. (2025). Fairness in AI systems development: EU AI Act compliance and beyond. *Information and Software Technology*, *187*, 107864. https://doi.org/10.1016/j.infsof.2025.107864

Wörsdörfer, M. (2024). Mitigating the adverse effects of AI with the European Union's artificial intelligence act: Hype or hope? *Global Business and Organizational Excellence*, *43*(3), 106–126. https://doi.org/10.1002/joe.22238

Appedix 1: Use of Generative Artificial Intelligence

In the creation of this thesis, I utilized generative Artificial Intelligence for several support tasks. The tools, their purpose and the verification measures are detailed below. I confirm that I have used AI tools with the necessary care and caution, have fully disclosed their use in accordance with university policy and take full responsibility for all content presented in this thesis.

1. Tool: OpenAI's ChatGPT (GPT-4 Version)

- **Stage of Use:** Thesis topic ideation and refining research questions
- **Purpose of Use:** I knew I wanted to research ethical AI and somehow include The EU AI Act. I used ChatGPT to brainstorm initial research avenues and research questions.
    - **Example Prompt (November 8, 2025):** *"What would be current Batchelor's Thesis research avenues that combine the EU AI Act and ethical AI, and fit well within Information Systems Science?"*
- **Verification:** The AI results suggested multiple research ideas and possible research questions for each area, with varying usability and personal interest, such as *"Transparency-by-Design Under the EU AI Act: Technical and Legal Perspectives"*, with the relevant research questions: *"What technical transparency tools (e.g., logging, XAI) best support compliance? How does transparency impact organizational trust and risk management?"* and "*Human Oversight Requirements in the EU AI Act: Practical Implementation in Information Systems*" with the relevant research questions: "*What oversight mechanisms are required under the EU AI Act for high-risk systems? How can these be embedded into system architecture and workflows? How does oversight influence system accuracy, fairness, or user trust?".* I treated these as a starting point and refined my own research topic and research questions based on this idealisation. No text from the AI was used in the thesis itself; it only guided my own research process.
- **Stage of Use:** Rephrasing individual sentences
- **Purpose of Use:** Used in rephrasing individual sentences to improve academic tone and flow.

    - **Example Prompt (December 8, 2025).** *"improve: as well as act as a potential safeguard to counter the negative aspects of high-risk AI applications more broadly."*

- **Verification:** The AI results suggested the following options *"…while also functioning as a potential safeguard to mitigate the wider risks associated with high-risk AI applications"* and *"…as well as serve as a potential safeguard against the broader negative impacts of high-risk AI applications".* I ended up phrasing the sentence as: *"…and serving as a critical safeguard against the broader negative impacts of high-risk AI systems."* In this instance I took inspiration from the AI provided improvements, making the final phrasing myself but some individual sentences I asked ChatGPT to rephrase, were used as suggested.

Stage of Use: Citation formatting

- **Purpose of Use:** I used AI in clarifying how some references, mainly for the legal sources, should be formatted according to APA 7.

- **Example Prompt (November 25, 2025):** *"Is it correct to refer to this in text as (Artificial Intelligence Act, 2024) http://data.europa.eu/eli/reg/2024/1689/oj?"*

- **Verification:** The AI results confirmed that it is in line with APA 7 to refer to The EU AI Act as I thought.

    - **Example Prompt (November 25, 2025): "***Is this correct for APA 7: Bovens, M., Goodin, R. E., Schillemans, T., Patil, S. V., Vieider, F., & Tetlock, P. E. (2014) Process Versus Outcome Accountability. M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), The Oxford Handbook of Public Accountability. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0002"*

- **Verification:** The AI results stated that the reference was mostly correct, but required a few minor tweaks, which I did.