UNIVERSITY OF TURKU
Turku School of Economics

Ethical Challenges of the European Union Artificial Intelligence Act

Information Systems Science

Master's thesis

Author:
Victoria Vilfors

Supervisor:
Ph.D. Jani Koskinen

11.11.2025

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.

Master's thesis

Subject: Information systems science
Author: Victoria Vilfors
Title: Ethical Challenges of the European Union Artificial Intelligence Act
Supervisor: Ph.D. Jani Koskinen
Number of pages: 64 pages + appendices 1 pages
Date: 11.11.2025

Abstract

This thesis examined the nature of existing ethical challenges in AI system development and deployment, and how the new European Union Artificial Intelligence Act (Regulation (EU) 2024/1689) addresses these ethical challenges, as well as what ethical challenges and gaps persist despite the newly developed regulations. The motivation of this research was to further the discussion surrounding the new EU AI Act and contribute to expanding the understanding and awareness of what ethical challenges and gaps must be addressed to advance the implementation of EU AI Act in its formative stage. The main research question of this thesis was "What ethical challenges and gaps persist in the European Union Artificial Intelligence Act?". A qualitative approach was chosen to lay out the ethical landscape of AI system development and deployment in the context of the EU AI Act from the standpoint of fairness, accountability, transparency and ethics (FATE) framework. The primary qualitative research methods used in this thesis were ethical analysis and thematic analysis, as well as the chosen principles of fairness, accountability and transparency. The research approach for this thesis was chosen due to the novelty of the newly imposed legislation and rapidly developing nature of subject of artificial intelligence.

Notable gaps and challenges identified in this thesis include challenges that stem from the difficulty of defining fairness, where AI system providers could use gerrymandering to manipulate the interpretation of fairness and indicate that the system under development is fair. Another ethical challenge and gap related to accountability in governance and enforcement of the regulations is that companies might not be equipped to foster a certain level of self-governance, creating a concern about how well the regulations will be enforced and the AI systems audited. The fast-approaching deadline for compliance has also been identified as problematic for AI system deployers, due to uncertainty of how to comply with the new regulations. Additionally, a gap related to algorithmic transparency was identified, and the so-called black box problem persists. Despite the efforts of the AI Act, the level of transparency and explainability that is required for the regulations to be effective might be unrealistic, because of complex supply chains and the overall complexity of AI models. Moreover, the effects of generative AI on individual's lives and society have not been given enough consideration, because despite problematic cases with generative AI, generative AI remains outside of the high-risk classification and therefore is subject only to transparency requirements.

The EU AI Act is however, currently in its formative stage, because of the novelty of the regulations, and the legal and ethical practices concerning the new regulations are still forming. There are no real-world cases that can be used as example case studies where the EU AI Act was implemented. This thesis also highlights the tentative state of the enforcement of the EU AI Act, because the legal and ethical practices concerning the new regulations are still forming and there are no real-world cases that can be used as concrete examples of the AI Act in motion. In other words, the interpretations of the EU AI Act are constantly evolving as regulators, institutions legal bodies and other stakeholders engage with the regulation. The conclusions and interpretations drawn from this thesis should be considered open to revision due to the evolving nature of interpretations and enforcement of the AI Act in real-world situations.

Keywords: artificial intelligence, generative artificial intelligence, ethics of ai, eu ai act, ai act.

Pro gradu -tutkielma

Oppiaine: Tietojärjestelmätiede
Tekijä: Victoria Vilfors
Otsikko: Euroopan unionin tekoälyasetuksen eettiset haasteet
Ohjaaja: FT Jani Koskinen
Sivumäärä: 64 sivua + liitteet 1 sivua
Päivämäärä: 11.11.2025

Tiivistelmä

Tässä tutkielmassa tarkasteltiin tekoälyjärjestelmien kehittämisen ja käyttöönoton eettisten haasteiden luonnetta sekä sitä, miten uusi Euroopan Unionin tekoälyasetus (asetus (EU) 2024/1689) vastaa näihin eettisiin haasteisiin, ja mitä eettisiä haasteita ja puutteita on edelleen olemassa uusista säännöksistä huolimatta. Tutkimuksen tarkoituksena oli edistää keskustelua uudesta EU:n tekoälyasetuksesta ja lisätä ymmärrystä ja tietoisuutta siitä, mihin eettisiin haasteisiin ja puutteisiin on puututtava EU:n tekoälyasetuksen täytäntöönpanon edistämiseksi sen muodostumisvaiheessa. Tutkielman päätutkimuskysymys oli "Mitä eettisiä haasteita ja puutteita Euroopan unionin tekoälyasetuksessa on edelleen?". Laadullinen lähestymistapa valittiin tekoälyjärjestelmien kehittämisen ja käyttöönoton eettisen maiseman hahmottamiseksi EU:n tekoälyasetuksen kontekstissa oikeudenmukaisuuden, vastuuvelvollisuuden, läpinäkyvyyden ja etiikan (FATE) viitekehyksen näkökulmasta. Tässä tutkielmassa käytetyt ensisijaiset laadulliset tutkimusmenetelmät olivat eettinen analyysi ja temaattinen analyysi sekä valitut oikeudenmukaisuuden, vastuuvelvollisuuden ja läpinäkyvyyden periaatteet. Tutkimuslähestymistapa valittiin uuden lainsäädännön uutuuden ja tekoälyn aihealueen nopeasti kehittyvän luonteen vuoksi.

Tässä opinnäytetyössä tunnistettuja merkittäviä puutteita ja haasteita ovat muun muassa oikeudenmukaisuuden määrittelemisen vaikeudet, joissa tekoälyjärjestelmien tarjoajat voisivat käyttää gerrymandering-menetelmää manipuloidakseen oikeudenmukaisuuden tulkintaa ja osoittaakseen, että kehitteillä oleva järjestelmä on oikeudenmukainen. Toinen eettinen haaste ja aukko, joka liittyy hallinnon ja säännösten täytäntöönpanon vastuullisuuteen, on se, että yrityksillä ei välttämättä ole valmiuksia edistää tietyn tason itsehallintoa, mikä herättää huolta siitä, kuinka hyvin säännöksiä valvotaan ja tekoälyjärjestelmiä auditoidaan. Nopeasti lähestyvä määräaika säännösten noudattamiselle on myös tunnistettu ongelmalliseksi tekoälyjärjestelmien käyttöönottajille, koska on epävarmaa, miten uusia säännöksiä noudatetaan. Lisäksi havaittiin algoritmien läpinäkyvyyteen liittyvä aukko, ja niin sanottu musta laatikko -ongelma on edelleen olemassa. Tekoälylain ponnisteluista huolimatta säännösten tehokkuuden edellyttämä läpinäkyvyyden ja selitettävyyden taso saattaa olla epärealistinen monimutkaisten toimitusketjujen ja tekoälymallien yleisen monimutkaisuuden vuoksi. Lisäksi generatiivisen tekoälyn vaikutuksia yksilöiden elämään ja yhteiskuntaan ei ole pohdittu tarpeeksi, koska generatiivisen tekoälyn ongelmallisista tapauksista huolimatta se jää korkean riskin luokituksen ulkopuolelle ja siksi siihen sovelletaan vain läpinäkyvyysvaatimuksia

EU:n tekoälyasetus on kuitenkin vielä muotoutumassa säännösten uutuuden vuoksi, ja uusiin säännöksiin liittyvät oikeudelliset ja eettiset käytännöt ovat vielä kehittymässä. Ei ole olemassa käytännön tapauksia, joita voitaisiin käyttää esimerkkitapauksina EU:n tekoälyasetuksen täytäntöönpanosta. Tämä tutkielma korostaa myös EU:n tekoälyasetuksen täytäntöönpanon kehittyvää tilaa, koska uusiin säännöksiin liittyvät oikeudelliset ja eettiset käytännöt ovat vielä muotoutumassa, eikä ole olemassa käytännön tapauksia, joita voitaisiin käyttää konkreettisina esimerkkeinä tekoälyasetuksen soveltamisesta. Toisin sanoen EU:n tekoälyasetuksen tulkinnat kehittyvät jatkuvasti sääntelyviranomaisten, instituutioiden, oikeusviranomaisten ja muiden sidosryhmien osallistuessa sääntelyyn. Tästä tutkielmassa tehtyjä johtopäätöksiä ja tulkintoja tulisi pitää avoimina muutoksille tekoälyasetuksen tulkintojen ja täytäntöönpanon kehittyvän luonteen vuoksi.

Avainsanat: tekoäly, generatiivinen tekoäly, tekoälyn etiikka, eu:n tekoälyasetus, tekoälyasetus.

TABLE OF CONTENTS

1 Introduction

This thesis explores the ethical challenges that arise during the Artificial Intelligence (AI) system development and deployment. With the vast amount of AI being integrated into different sorts of systems and decision-making processes, it is crucial to address the ethical challenges that arise during the AI system development and deployment, particularly in terms of fairness, accountability and transparency (Unver & Roddeck, 2024). This thesis examines the nature of existing ethical challenges in AI system development and deployment, and how the new European Union Artificial Intelligence Act (Regulation (EU) 2024/1689) addresses these ethical challenges, as well as what ethical challenges and gaps persist despite the newly developed regulations. The motivation of this research is to further the discussion surrounding the new EU AI Act and contribute to expanding the understanding and awareness of what ethical challenges and gaps must be addressed to advance the implementation of EU AI Act in its formative stage. The approach of this research is to examine the EU AI Act from the standpoint of fairness, accountability, transparency and ethics (FATE) framework, and identify what gaps and challenges persist based off these selected ethical principles (Bahar & Tenzin, 2023).

Artificial Intelligence has evolved into becoming a big part of society across various mediums (Zhang et al., 2022) including, for example, healthcare, transportation, public service, finance, education and entertainment. The rapidly growing implementation of AI into different decision-making domains and other aspects of society (Espina-Romero et al., 2023) has opened a discussion about what ethical challenges the rapidly changing and developing landscape of AI contains, and particularly how it affects the transparency, accountability and fairness of AI-based systems and decision-making processes. Following the discussion (Espina-Romero et al., 2023) it has become evident that the ethical challenges and implications of AI systems must be explored further, particularly in the context of developing regulations like the EU AI Act.

One of the most critical aspects in AI systems are the development and deployment processes which impact the fairness, accountability and transparency of AI. The mechanisms that existing guidelines and regulations provide for ethical AI and machine learning are rather weak and have little impact in the real-life scenarios and implementations. Moreover, the existing state of guidelines and regulations may even suggest that the self-governed guidelines created by AI industry operators are perceived as sufficient enough by the regulators, because the existing laws concerning AI are rather superficial and vague. (Hagendorff, 2020).

Artificial intelligence is trained by using datasets and the model selection process directly impacts the transparency, fairness and accountability of the AI. This is why it's crucial to assess the ethical challenges and their implications, because poorly executed AI development and deployment can lead to outcomes that perpetuate, for example, discrimination by creating social biases and disparities. (Chu et al., 2023).

The following research questions guide this thesis:

1. What ethical challenges arise during AI system development and deployment?
2. How does the European Union Artificial Intelligence Act address the existing ethical challenges through the lens of fairness, accountability and transparency?
3. What ethical challenges and gaps persist in the European Union Artificial Intelligence Act?

The existing ethical challenges concerning AI system development and deployment are discovered in this thesis by firstly reviewing academic research, case studies, and policy studies that are related to transparency, accountability and fairness, and secondly by analysing the EU AI Act by implementing the FATE framework. Next the goal is to identify existing ethical challenges in the AI system development and deployment process that affect the transparency, fairness and accountability of AI systems. The review of the EU AI Act will be used to discover and identify existing ethical gaps and challenges in the current regulations in the light of transparency, accountability and fairness. The analysis based on FATE framework is focused on how the new EU AI Act addresses ethical challenges in AI system development and deployment, what gaps can be identified in the regulations and what are the possible inconsistencies and limitations that surface up regarding the application of these regulations.

The rest of the thesis is constructed as follows. In section 2 the academic research related to artificial intelligence, ethics and ethics of artificial intelligence is reviewed through the lens of fairness, accountability and transparency. This section discovers what ethical challenges in AI system development and deployment have been identified, particularly ethical challenges related to fairness, accountability and transparency. In section 3 the European Union Artificial Intelligence Act is introduced, and the context of why the EU AI Act is being implemented is explained. This section encompasses the risk classification developed by the European Union, based on which different AI systems are subject to certain level of regulations. For each risk classification level there are examples of the types of cases and situations where AI system is classified to pose certain level of risk. The risk levels are explained further in this section, as well as the obligations for each

risk level. Section 4 explains the methodology of this thesis. The research approach, method, data collection and analysis, as well as the research quality, limitations and ethics are covered in this section. In section 5, the ethical challenges of the EU AI Act are explained and analysed using the FATE framework (fairness, accountability, transparency, ethics). Section 6 discusses the identified ethical challenges and gaps in the EU AI Act and considerations concerning possible outcomes and consequences of the persistent ethical challenges identified during this research and analysis. Finally, section 7 comprises the conclusions of this thesis and provides insight on key findings of this research.

2 Ethics and Artificial Intelligence

2.1 Artificial Intelligence

Artificial intelligence has become a phenomenon with global impact. The definition of artificial intelligence as well as artificial intelligence itself as we know it has changed over time. (Wagner et al., 2020, pp. 7-8). Artificial intelligence can be seen as the study of building computers or programming them in such a manner that enables them to do what human minds can do (Boden, 1996, p. xv). Russell and Norvig introduced the idea of an "intelligent agent" to define artificial intelligence. According to Russell and Norvig artificial intelligence is referring to intelligent agents that are capable of human like intelligent behaviours, more precisely of perceiving their environment through inputs called percepts and responding by performing appropriate actions. (Russell & Norvig, 2010, pp. xiii-2). Another highly influential, but also controversial definition of artificial intelligence was proposed by Alan Turing in 1950. Turing came up with a test where a human would be having a conversation with a machine, and if the human participant would be unable to determine if the conversation they were having was with another human participant or machine, the test would be considered passed by the machine. The purpose of this test is to determine if the machine has the capability to exhibit human like intelligence and whether it is good enough to be indistinguishable from an actual human participant. (Wagner et al., 2020, p. 9). For a more recent definition of artificial intelligence, Krauss explained that intelligence that occurs naturally, for example, in humans and animals is characterized as biological intelligence, whereas intelligence that is artificial is based on algorithms and software developed by humans. Systems based on artificial intelligence are programmed to perform various tasks like decision-making, problem solving, as well as learning from experience and pattern recognition. The key aspect of artificial intelligence is that it is designed to operate in an autonomous and adaptable way, so that the artificial intelligence can train itself based on experience and feedback received and improve with each iteration. (Krauss, 2024).

2.2 Ethics

Ethics, also referred to as moral philosophy is the study of morality, what is morality and what it requires. There is plethora of definitions as to what morality is, and many rival theories that expand the definitions and concepts of morality (Rachels & Rachels 2012, p. 1). There are several branches of ethics that explore different aspects of morality including descriptive ethics, normative ethics, meta-ethics and applied ethics. Descriptive ethics refers to the studying and understanding of what

people believe is right or wrong, in other words people's moral opinions. It is based off real-world observations of what people perceive to be right and wrong, and how they behave. Normative ethics comprise the idea of whether human actions are good or bad, and right or wrong. There are two main types of normative ethics, deontological ethics emphasize that no matter the outcome, certain actions are always right or wrong. Consequentialist ethics on the other hand claim that the outcome of the actions matter the most. Meta-ethics explores the ontology, semantics and epistemology of ethics itself. Meta-ethics is not about deciding what is right or wrong, or good or bad, it is about understanding, for example, what does good and bad mean as terms. Applied ethics, compared to other ethics, is about using ethics in real-life situations, for example, applying ethics in business or medicine. (Wagner et al., 2020, pp. 17-21).

2.3 Ethics of AI

The recent boom of artificial intelligence has grown the need for implementing ethics to deal with the disruptive potential and profound effects new AI technologies bring into individuals' lives (Smallman, 2022). In recent years, vast number of ethical guidelines have been ruled out encompassing various ethical principles to deal with problems like, for example, accountability and governance (Light & Panai, 2022). There is substantial number of principles proposed that guide these ethical AI guidelines, and significant amount of the principles overlap with each other (Whittlestone et al., 2019). A study was conducted by analysing AI ethics principals and guidelines, approximately 84 documents, and it was found that there are five ethical principles that stand out in the existing field of ethical guidelines and documents. These principals include transparency, responsibility, justice and fairness, privacy, and non-maleficence. (Laine, Minkkinen & Mäntymäki, 2024). Floridi et al. (2018) proposed initiative for "Good AI Society" and offered five ethical principles to support the development and deployment of ethical AI. These principles include beneficence, non-maleficence, autonomy, justice and explicability. The outlined principles are meant to be embedded in the default practices of AI, and serve as cornerstones to decrease inequality, further social empowerment and increase the benefits shared by all. (Floridi et al., 2018). Hagendorff (2020) claims that in the field of AI and machine learning, ethical guidelines do not have that big of an impact on the decision-making processes of humans. The argument is that due to the ethical guidelines being employed by self-governance, the mechanisms for enforcing the guidelines and compliance are weak and insufficient. (Hagendorff, 2020). The definition and consideration of ethical principles is however not enough, and to truly impact the ethics off AI, ethical principles should lead to concrete actions, rather than being abstract representations of values (Whittlestone et al., 2019). For the purposes of defining the scope of this thesis, it was

decided to choose the FATE framework (Bahar & Tenzin, 2023) for analysis of the EU AI Act, because it includes ethical principles of fairness, accountability and transparency. The use of principles of fairness, accountability and transparency in relation to specifically AI are described and justified in the following subchapters.

2.4 Fairness

One of the biggest ethical challenges in the usage of artificial intelligence is ensuring fairness and mitigating bias that stems from the unfairness of AI systems that can manifest by, for example, an AI system making decision that is favourable or prejudiced towards certain groups of people or individuals (Odilla, 2024). AI systems and services have the potential to affect and exacerbate social inequality among individuals by producing unfair outcomes and reinforcing harmful stereotypes (Sadeghiani, 2024). For example, this is why AI systems and AI-as-a-Service (AIaaS) platforms can serve as triggers through which systemic bias is replicated and potentially scaled. A model trained on data that is biased will probably produce biased results, and even a model that is trained using techniques to mitigate bias can behave unfairly in some contexts that it was not tested against or designed to perform in. (Lewicki et al., 2023). It is difficult to predict all the ways in which individuals might use the AI systems, and even harder to consider how the individuals' own moral compass is calibrated towards bias and fairness, and in what ways it manifests itself (Sun, Zhao & Chen, 2024).

In case of AIaaS platforms, which have great scalability capabilities, it is evident that issues with fairness and bias can be magnified because of the opportunities the AIaaS platforms provide. This is due to the fact that a biased algorithm can be embedded into a service and that service could potentially be used by thousands of individuals, which means that the negative impacts of a biased algorithm would be multiplied. Also, a lot of users that use AIaaS do not possess the technological expertise or resources needed to assess the particular AI model they are using and whether it is suitable for their needs and use case. This may result in biased, harmful and unfair outcomes that are discriminatory in nature without the users even detecting or correcting the biased and unfair outcomes. (Lewicki et al., 2023).

Another ethical challenge posed by AI systems in terms of fairness is that it is unpredictable how the AI services might be used by individuals and what data is inputted into the algorithm (Lewicki et al., 2023). This is particularly important consideration to be made, because in machine learning algorithms learn from past data, and the data that is inputted by potential users might already be biased, leading to unfair outcomes (Pfeiffer et al., 2023). During the development of the AI system,

the developers cannot anticipate and predict every possible use case or context in which their AI system will be used. This results in a problem where it becomes highly complicated to design a solution that fits all and ensures fairness in every use case and scenario. Fairness challenges can often be noticed clearly in social or cultural contexts when the AI system is already in use, which makes it difficult to detect potential unfairness during the development and testing phases of the AI system. (Lewicki et al., 2023).

The literature that exists regarding algorithmic fairness is for the most part focused on AI systems that are developed in-house, which means that the developers who are building the AI model can directly step in during the model training process by adjusting the training data, model parameters or deployment settings to mitigate bias (Lewicki et al., 2023). The sources of bias can surface up in different stages of the development pipeline, for example, during data collection stage, user interactions or the design of the algorithm (Ferrara, 2023). In the case of in-house development of the AI model, the developers are involved in the model training process during the design and development stages, which means that they have more control over the fairness outcomes produced during the development of an AI system (Lewicki et al., 2023). For example, developers should use data derived from various and versatile datasets to ensure high-quality machine learning pipeline (Nazer et al., 2023). As for contrast, the scenario where a third-party AIaaS platform adopts pre-build AI models and services is given too little attention. The scenario poses clear ethical challenges regarding fairness, because those who employ the usage of pre-build models and services might not be fully aware and have a good understanding of the inner workings of the algorithms and their origins. (Lewicki et al., 2023).

The interdisciplinary research around AI and fairness of algorithms suggests that it is highly contextual whether an algorithm can be classified as fair and in what terms it can be harmful (Lewicki et al., 2023). Fairness is a multifaceted concept in regard to AI, because it lacks universal accepted definition because of its complicated nature that encompasses various ethical, social and cultural considerations (Barocas, Hardt & Narayanan, 2019). This means that what is considered fair in certain contexts can vary significantly depending on the specific conditions in which the AI algorithm is applied (Lewicki et al., 2023). The considerations that should be taken into account are not concordant and can vary depending on the domains they are applied to, for example, cultural considerations, geographical regions and societal norms tied to a specific place. Biased AI can amplify historical as well as societal prejudices towards certain groups of people. (Foka et al., 2025). In certain situations, some algorithmic outcomes that are harmful but not easily spotted or seem benign, can become problematic in certain contexts, and be harmless in others. This type of

contradiction can occur when an algorithm is used in a system within certain type of social dynamics that for instance encompass cultural injustices or sensitivities, that could potentially influence how the outcome of the algorithm is experienced and interpreted. This results in a scenario where the harmful effects of the AI algorithm become apparent only when the AI system is deployed under specific conditions in particular community. This is why it is important to understand and address the contextual aspect of fairness of AI algorithms when designing, evaluating and deploying AI systems. (Lewicki et al., 2023).

According to Barocas, Hardt and Narayanan (2019), fairness cannot be treated purely as a technical problem. Fairness must be examined and understood within broader context of, for example, societal values and historical inequalities. It is clear that any attempts to formalize fairness often lead to making trade-offs between definitions of fairness, like equal opportunity, individual fairness and demographic parity. None of these definitions can be applied universally without consequence. Fairness should be considered in the specific context in which the potential AI system or algorithm will operate, because an AI model that is trained and deployed in one setting might be deemed fair, but unjust in other. (Barocas, Hardt & Narayanan, 2019). Fairness is still highly debatable concept in ethics and there is still an ongoing discourse about how it should be defined and applied. Even the comprehensive theories of justice, for example, Rawl's theory of justice is built upon assumptions that can be subjected to philosophical disagreement, demonstrating that fairness is a highly complex subject in moral and political reasoning. (Rawls, 2009).

Most of the scholarly works and literature explore ethical challenges from the point of view of fairness by providing methods to audit and expose fairness concerns in AI systems, and for the most part they target the development phase of the AI systems (Lewicki et al. 2023). Fairness is explored in practice by researching algorithmic fairness and by developing various fairness metrics to examine the machine learning algorithms (John-Mathews, Cardon & Balagué 2022). They offer wide range of ethical tools and interventions for the developers who are designing and deploying the AI systems. However, there is a large gap in the existing research, specifically that would prioritize user perspective of the matter. There is limited amount of research that examines the nature of how individuals and organizations that use external AI systems or tools experience or navigate fairness issues. The lack of attention and focus on this field presents a critical aspect of the deployment and operational phases, because various fairness related issues and challenges are likely to emerge and be discovered during the deployment and the usage of the AI system. (Lewicki et al. 2023).

Given these concerns about bias and fairness in AI systems, it is evident that the matter requires a more holistic and standardised approach (Agarwal & Agarwal, 2024). The approach must not only include technical solutions and proper auditing practices, but also more transparency, user education and accountability. In order for AI systems to be ethically developed and be considerate of aspects such as fairness, the focus point must not only be at the development stage, but also at the deployment and operational stage, meaning how the AI system will be used, by whom and in what contexts. (Lewicki et al. 2023).

2.5 Accountability

Accountability is one of the most widely discussed aspects of the creation of AI. Accountability is referring to the question of who is accountable and responsible in relation to the outcomes and decisions produced by an AI (Schmidt et al., 2025). During the model training process the AI is trained on certain types of data. The algorithm is self-learning, but it is still guided and affected by the actions of developers, which opens the discussion on who should be accountable in the end for decisions and outputs of the AI's algorithm. (Raja et al., 2023) The question of accountability poses several ethical dilemmas, of which one of the most evident is who is responsible for an outcome, if the algorithm is self-learning, but at the same time developers are impacting the outcomes of the algorithms, as well as the data the algorithm is developed on. Many engineers and developers of AI do not see that they have the agency, responsibility or capability to influence the important questions posed by responsible AI guidelines. (Widder & Nafus, 2023).

2.5.1 Four Responsibility Gaps

Accountability discussion encompasses many aspects, including responsibility and the idea of "responsibility gap". Responsibility gap is the concept that highlights the challenge of determining who is actually responsible for the outcomes and possible actions of artificial intelligence, and specifically who has the moral and legal responsibility in this case. (Santoni de Sio & Mecacci, 2021.) With the emergence of artificial intelligence, it has become highly discussed matter of how responsibility should be attributed and whether machines can be held responsible, because studies have shown that humans might hold machines or robots responsible for their actions (Simmler, 2024). The article "Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them" by Santoni de Sio and Mecacci (2021) shows a thorough analysis about the responsibility gap and identify four different types of responsibility gaps. Each gap is a distinct aspect of the responsibility gap, and each gap is described to have its own ethical challenges, as

well as solutions. The four responsibility gaps are: culpability gap, moral accountability gap, public accountability gap, and active responsibility gap. (Santoni de Sio & Mecacci, 2021.)

2.5.2 Culpability Gap

The culpability gap, or blameworthiness, is referring to the act of determining who is at fault when an AI or an algorithm produces unwanted or harmful output or action. In other words, who is to blame if the AI outputs or impacts a situation in a harmful way (Khomkhunsorn, 2025). The gap exists because traditionally from legal or moral perspectives there is usually an agent who is deemed responsible for the outcome produced by a machine, because an agent like the manufacturer is seen to have the control over the performed action, as well as intent and knowledge (Matthias, 2004). In the case of artificial intelligence, it is difficult to determine who is responsible, because there is no agent in the traditional sense of legal or moral responsibility. The problem with AI and machine learning algorithms is that their actions might not be predictable and there is a lack of transparency in how the algorithms work. (Santoni de Sio & Mecacci, 2021.)

In their article, Santoni de Sio and Mecacci (2021) use automated driving systems as an example to explain the culpability gap. In the situation of a car crash, it is not clear who is the responsible stakeholder and on what grounds. Due to the automated driving systems multifaceted nature, which can involve multiple different factors and actors like, for example, the driver himself, the company who manufactured the vehicle, the company who created the software for the vehicle and the company who developed the machine learning algorithm for the vehicle, it is difficult to pinpoint what and who actually caused the accident in the first place. Therefore, the driver might not have had enough control over the vehicle, and on the other hand the developers and manufacturer can claim that the algorithm and functionalities were executed in an expected matter. (Santoni de Sio & Mecacci, 2021).

A similar example can be made regarding, for example, a medical diagnostics system that is powered by a machine learning algorithm. In the case of the medical diagnostics system producing a faulty diagnosis and recommendations for treating a patient, it is unclear who is responsible for the harm caused to the patient. It is difficult to determine if the responsible agent is the doctor who was using the system, the system developers or the company who developed the system, or maybe the regulators who approved the usage of the system in the first place. This is a particularly hard context to attribute culpability, because the demonstration of, for example, the cause of injury is already challenging in the medical context, and the involvement of AI complicates the matter even further. (Price et al., 2024, pp. 150-166)

2.5.3 Moral Accountability Gap

Moral accountability refers to the phenomenon of individuals need to explain each other's actions and choices through moral accountability. In other words, individuals tend to question each other's actions and choices, not necessarily in a demanding tone of voice, but to just get an explanation on why a certain action or decision was made. The intent is not to judge or blame, it is rather to understand the other persons reasoning behind the actions through questions and answers. (Santoni de Sio & Mecacci, 2021). In the context of AI, accountability usually refers to the responsibility for a system, the behaviour of the system and the potential impacts the system can have (Ethics of AI MOOC, 2025).

Moral accountability is a challenging factor to examine in AI and machine learning algorithms. Because of the complex nature of artificial intelligence algorithms, it is challenging to see the reasoning behind the algorithm's decision-making, this is the so called "black box" problem. (Khomkhunsorn, 2025). The problem with the moral accountability aspect in relation to its user is that it is unclear what is the user's actual role in the decision-making process, what factors and how specifically did they influence the outcome, and for what parts of the operation is the user morally accountable for. In other words, it is challenging for the user and highly unlikely that the user would be able to explain or justify the outcome of the AI algorithm, because often even the developers have difficulties explaining how the self-learning algorithm works and what is the full process behind its reasoning and decision-making. (Santoni de Sio & Mecacci, 2021).

Moral accountability in the context of AI can be explained again using the example of medical diagnostics system that is powered by AI. If there is a medical diagnostics system powered by AI that is being used to diagnose patients, the inner workings of the algorithm are unclear. This leads to the fact that the doctor who is using this system to diagnose patients does not understand the algorithms decision-making process thoroughly and cannot therefore make sense of the logic behind the diagnosis. (Price et al., 2024, pp. 150-166). This leads to the moral accountability gap, because if there's no sense of the logic behind the decision-making process, the doctor cannot explain the moral reasoning behind the diagnosis (Santoni de Sio & Mecacci, 2021).

2.5.4 Public Accountability Gap

Public accountability is essentially the act of being accountable for what the public decide to deem important (Gleisner, 2020). Public accountability refers to the responsibility of officials, politicians and other public servants and agents to justify, explain and be accountable for their actions and decision-making. They need to be held accountable for their decision-making and actions, because

they are serving the public, so in this sense they have an obligation for public accountability. The recent introduction of AI in decision-making processes has raised a lot of concerns regarding the accountability and transparency of the decision-making process. The usage of AI encompasses a lot of factors that limit the transparency of the decision-making due to the nature of the AI algorithms, which in turn may lead to decreased accountability and by that, limit the democratic control that public accountability promotes. (Santoni de Sio & Mecacci, 2021).

This kind of dynamic can lead to the shift of public accountability from officials and public servants to private companies and the developers who developed the AI algorithm (Santoni de Sio & Mecacci, 2021). For example, governments can use AI algorithms to automate processes related to running tax errands and doing tax audits, which means that the decision-making and reasoning behind the outcome will not be made by an actual human being. This means that the decision-making, for which the governmental body is responsible for, will not be subject to public scrutiny in case of bias or error to the same degree it would be without the AI's involvement in decision-making. (Alon-Barkat & Busuoic, 2023).

Moreover, the usage of AI in decision-making at this level can significantly affect the chain of responsibility and accountability in governmental bodies. Usually in bureaucratic structures the governmental bodies and individuals who act as decision makers can be held accountable for their actions through legal and bureaucratic mechanisms. In case with the introduction of AI into these bureaucratic structures, the accountability and responsibility lines become vague, because there are too many agents involved, like the governmental body itself, the company who developed the AI algorithm and the developer etc. This is how the public accountability gap takes place, and the problem is that it makes more difficult for people to appeal biased or wrong decisions that affect them. (Santoni de Sio & Mecacci, 2021.)

2.5.5 Active Responsibility Gap

Active responsibility is referring to the proactive approach to responsibility of, for example, the engineers who work on AI algorithms (Pesch, 2015). There is a distinction between the active and passive responsibility, and the active responsibility encompasses mainly things like values, goals and legal norms. The active responsibility is a more forward-looking approach that is focused on preventing some unwanted consequences by actively complying with the existing recommendations, norms and legal norms, whereas passive responsibility is more of a backward-looking approach that focuses on the possible legal and moral consequences in case of something going wrong. The problem with active responsibility is that, for example, the engineers working on

an AI algorithm might not have clear ethical guidelines and recommendations, or even a clear understanding of what their social and ethical role is in the development of the AI algorithm. Due to the large networks engineers work in, it is highly challenging to be proactively assigning active responsibility to the people that the network consists of. This lack of attribution of active responsibility may lead to situations in which it is impossible to find and hold someone accountable for a specific event that was unwanted. This is creating the active responsibility gap, because it is unclear in large scale networks who should be held responsible and for what. (Santoni de Sio & Mecacci, 2021.). It is also particularly burdensome to try to assign active responsibility in cases where the development is being done by several engineers and the engineers are regularly swapping tasks. It is not uncommon to swap tasks with other engineers and developers if, for example, someone feels stuck on a particular task or a bit of code. In this case, it is unclear how to promote and follow up the assignment of active responsibility, and to what extent. It is also noteworthy that in competitive markets the ethical and moral considerations are not the main focus of companies that want generate as much revenue as possible, which means that the proactive promotion of active responsibility becomes secondary. (Pesch, 2015).

In conclusion, the four responsibility gaps, culpability gap, moral accountability gap, public accountability gap and active responsibility gap, all represent deep-rooted ethical challenges related to the use of AI. The four responsibility gaps not only highlight the difficulty of establishing proper ethical guidelines for the AI model training process but also shed light on the challenges of proactively promoting and fostering those ethical considerations. So not only there are challenges in establishing ethical guidelines that address ethical concerns of AI, but the problem is also the organizational structures that in worst cases could prevent the attribution of responsibility and accountability and therefore the strategy for assigning accountability should consist of sharing it across multiple stakeholders to ensure the proper attribution of responsibility (Collina, Sayaadi & Provitera, 2023).

2.6 Transparency

Transparency, in the context of artificial intelligence, is referring to the explainability, clarity and openness of an AI's decision-making process for humans. In other words, transparency in the context of AI encompasses the understanding of why the AI's algorithm works a certain way. It particularly sheds light on how the algorithm was trained and what are the internal mechanisms that led to the current state of the algorithm. It involves the examination of the training process from within, and the idea is to understand how, for example, the data used in the model training process

has led to the algorithm to arrive to a certain conclusion or make a certain decision. (Avinash et al., 2023.) Transparency of AI is often discussed in the context of algorithmic transparency, but the transparency gap in AI needs to be examined thoroughly by recognizing three levels of transparency, which include algorithmic, interaction and social transparency. By identifying that the transparency gap exists on three different levels, it can be used to build trust in AI and repair the existing fragmentation. (Haresamudram, Larsson & Heintz, 2023).

Explainability is an aspect of transparency that is closely related to it, and it takes the transparency a step further. Whereas transparency reveals the decision-making process, explainability provides an explanation on why a certain decision was made. In other words, explainability provides a description of, for example, some data points and factors that have influenced the decision-making process, and how the factors have affected a specific output. (Avinash et al., 2023).

There is indisputable evidence that the AI algorithms used for model training lack transparency, whether it is due to the design decisions or limitation of technology. In this case we are talking about the algorithmic transparency, that is typical specifically for self-learning algorithms that are used for model training in AI. During the model training process, the self-learning algorithms become more and more complex, and their transparency becomes difficult to grasp. During the model training process the algorithm produces a new set of rules with each iteration that alters its decision-making process. While the self-learning algorithm is constantly modifying the decision-making process, it becomes challenging for the developers to be able to maintain intricate understanding of why the algorithm is adjusting itself as it is, and what is the overall decision-making logic behind the adjustments and learning curve. Findings in the recent literature about transparency in AI have shown that the lack of algorithmic transparency can be caused by the cognitive limitations of humans. The claim is that the human brain has a limited cognitive capacity which prevents it from interpreting massive algorithmic models and datasets. The problem is that there is a lack of applicable and relevant tools for tracking and visualising large amounts of data and code, because at the present moment the data and code are so poorly structured, that it is challenging to read it, which leads to the overall lack of transparency in algorithms. (Tsamados et al., 2022).

There is a phenomenon called “black box problem” where AI develops itself through self-learning and deep machine learning. The AI is using large amounts of data and information to develop rules and identify patterns, so it means that instead of a human being programming the AI, the AI is essentially teaching itself on all the provided data. The difference between a self-taught AI and a

human programmed one that teaches the AI how to make decisions, is that the self-taught AI can identify way more complex patterns and make connections between information and patterns faster than a human being ever could. The problem is also that most of the AI systems are created in a manner that does not promote the AI being able to showcase how it arrives to certain conclusions or results, hence creating the “black box problem”. In this context, it is also problematic that the companies who are creating AI systems do not really want to share how their algorithms work, because it is considered their intellectual property and if someone else was able to recreate the algorithm and its learning process, that could result in a potential monetary loss for the company, which is something that every profit-oriented business would like to avoid. (Klugman, 2021).

Modifiability of the algorithms can also contribute to the factor of transparency of the AI. The modifiability of algorithms enables the modification of almost any part of the algorithm continuously, making it easy to manipulate the algorithm during the model training process. (Tsamados et al., 2022). The AI algorithms can be particularly complex, because it is possible for the AI to perform modifications to its own behaviour through the model training process (Eberbach et al., 2005, pp. 34-43). Regardless of the lack of proper tools for following the model training process from within, and how the self-learning algorithm evolves in detail, the developer still has the full control to modify the algorithm at every step of the learning process. The power to modify the algorithm can be used to enhance and evolve the algorithm to work better, but it can also be used to abuse the development of the algorithm and, for example, to blur the history of its development whatsoever. This in turn may lead to confusion on the part of end users, because the constant modification of an algorithm may change the affordances of the algorithm in question like, for example, the Google search algorithm, which makes it harder for users to follow the recent developments and practicalities of the search engine when for instance performing search engine optimization. (Tsamados et al., 2022).

Transparency in itself can however also lead to ethical challenges. It can allow individuals or groups to take advantage of the transparency provided, by gaming the system for their own benefit. This can be possible due to the transparency of a system, if the operator is able to get the knowledge about, for example, what metrics the algorithm uses to sort out new inputs, what is the source of the datasets used to train the algorithm, and under what kind of assumptions the sampling was done. It is also worth mentioning, that this form of diversion of the algorithm for one's own benefit requires a high level of digital proficiency and literacy, thus transforming it into another form of social inequality. (Tsamados et al., 2022). It is, however, questionable whether the social inequality caused by this kind of misuse of transparent algorithms is a significant factor in the grand scheme of

the things that create social inequality. On the other hand, increasing transparency might have the opposite effect, because whereas transparency might decrease data misuse or privacy breaches, but it can also decrease the public trust and data sharing if the transparency protocols are too complex to implement. (Capraro et al., 2024).

3 European Union Artificial Intelligence Act

The EU AI Act (Regulation (EU) 2024/1689) is the first comprehensive legal framework in the world that aims to address the potential risks associated with artificial intelligence. The EU AI Act is also meant to position Europe in a role of global leader in the governance of AI. The AI Act is supposed to establish harmonized rules that help regulate and promote trustworthy and ethical AI development and usage within the European union. (European Commission, n.d). The European commission proposed the first law for artificial intelligence within EU in April 2021, as well as a risk-based AI classification system that will be used to classify AI technology. The European Parliament has set up a working group that will be responsible for administering the implementation and enforcement of the new EU AI Act in cooperation with the European Commission's EU AI office. The EU AI Act came into force in June 2024 and for the most part will be applicable within 24 months of entry, except for some high-risk systems that will have to comply with the newly enforced regulations after 36 months after the entry into force. (European Parliament, 2025).

The AI Act constitutes a risk-based regulatory approach that uses categorizing as a tool to map out AI applications based on the potential impact on, for example, fundamental rights and safety, as well as the human-centricism of the AI. The new regulations provide coherent guidelines for AI facilitators, companies and developers to ensure that the AI technologies are consistent to the legal and ethical standards posed by the AI Act. The European union has also introduced additional set of policy initiatives like AI Innovation Package, AI Factories and Coordinated Plan on AI, that are meant to act as measures for establishing a robust ecosystem that will support responsible AI innovation and reinforce the public trust of Europeans in AI systems. The European commission has additionally introduced the AI Pact which is a voluntary initiative designed to ease the transition to the new regulatory framework and help stakeholders like AI providers and deployers across Europe to comply with the new regulations and obligations. (European Commission, n.d).

3.1 Context of the EU AI Act

The need for the European AI Act stems from the essential requirement of the European Union to ensure that Europeans can rely on and trust AI technology. According to the European Commission, most of the existing AI systems do not pose a significant risk, if one at all, but there are some AI technologies and systems that present bigger risks and thus must be managed to prevent unwanted consequences. One major aspect of AI systems is AI decision-making processes and the lack of transparency in those processes. It is often difficult to determine why an AI system has arrived at a

specific conclusion and what led to that decision or prediction. This in turn makes it challenging to assess the potential biases or unfairness of the decision-making process. This is particularly relevant aspect in AI-driven decision-making that can significantly impact individuals' lives, for example, in hiring decisions or public benefit schemes. (European Commission, n.d).

The existing legislation offers some degree of protection and regulation in regard to AI, but it is not enough to address the existing challenges created by AI technology. The AI Act was implemented to fill the existing gap in legislation so that the European Union has comprehensive safeguards that can ensure transparency, fairness and alignment with fundamental rights for Europeans. (European Commission, n.d). Also, to prevent any harmful outcomes, the European Parliament emphasizes that the AI technology should be supervised by people and not automated (Piachaud-Moustakis, 2023). To ease the transition and to ensure compliance the European Commission has introduced the AI Pact that encourages AI providers and deployers within Europe to align with key regulatory requirements. The idea is to engage proactively the different stakeholders that fall under the new regulations ahead of formal enforcement. This ensures a smoother transition process and compliance so that AI technologies that are developed and used in the European Union are transparent, safe, human-centric and respect the fundamental rights of Europeans. (European Commission, n.d).

3.2 Risk Classification

The EU AI Act establishes four different levels of risk for AI systems: unacceptable risk, high risk, limited risk and minimal risk. The idea is that every AI system corresponds to certain classification of risk level and should be dealt with according to its potential threat level. Each risk level has certain practices that are either prohibited or need to be regulated and are subject to strict obligations before they can enter the market. (European Commission, n.d).

3.2.1 Unacceptable Risk

Every AI system that can be seen as a clear threat to society or individuals is banned under the EU AI Act. Specifically, if the unacceptable level of risk is concerning safety, human rights or means of living. Posing an unacceptable amount of risk are practices that use AI for deception and manipulation in a manner that can, for example, influence individuals' behaviour without the individual's awareness or consent. The exploitation of vulnerable groups is also prohibited under the AI Act, for example, if the AI system is taking advantage of a specific vulnerabilities of an individual. Social scoring is one of the prohibited uses of AI, which means that every AI system

that evaluates and ranks people based on their characteristics, behaviour or other attributes, and uses that information to give the individual a social score is considered a prohibited practice, and that kind of AI system is prohibited. This is due to the AI system potentially making decisions that may lead to discrimination or unjust treatment based on the social score an individual has been given. Predicative policing practices are also prohibited, because AI systems that could analyse and predict how likely an individual is to commit criminal offences is considered a practice that poses an unacceptable risk. Untargeted collection of data, or scraping of the internet, as well as collection of CCTV material to build or develop facial recognition databases further is a prohibited practice, because it breaks the privacy and data protection rights. The use of technology that enables emotion recognition features in environments like workplaces and educational institutions is prohibited, because it can be considered as surveillance and manipulation of individuals. Biometric categorization is prohibited, when it is used to work out individuals' characteristics like, for example, ethnicity or sexual orientation. Finally, real-time remote biometric identification that is used for law enforcement purposes, specifically in publicly accessible places is strictly prohibited. (European Commission, n.d).

3.2.2 High Risk

AI systems that have the ability to significantly impact individuals’ health, safety or fundamental rights are classified as high-risk AI use cases. This classification applies to various sectors where individuals can be affected by the consequences of bias, failure or misuse and where the consequences of those could be severe. These AI practices and use cases are not prohibited, but because they pose a significantly higher risk of bias, misuse or failure, they will be subject to strict obligations before the AI systems can enter the market. (European Commission, n.d).

One major category that involves high-risk classification is safety in critical infrastructure if it includes AI safety components. For example, with transportation systems, if AI is responsible for critical functions like traffic control or safety mechanisms in public transport, in case of a malfunction it could directly threaten public and individual safety by creating serious situations due to the malfunction. (European Commission, n.d).

Education is another area of high-risk posed by AI systems. AI tools can be used to assess and determine access to academic opportunities, for example, by AI scoring exams or recommending placements. This type of usage of AI in education could have a huge impact on individuals professional future, because it has the ability to influence individual’s educational path and thus can lead to discrimination or unfairness. (European Commission, n.d).

In healthcare AI based safety components in products are considered a high-risk use case. Specifically use cases with robot-assisted surgery that involve AI based safety components or other medical devices. This use case is considered a high-risk because a malfunction of such a component could result in lethal outcomes or cause serious harm to patients. (European Commission, n.d).

In the employment sector usage of AI tools is also considered a high-risk. For example, AI tools that sort CVs and AI management tools. These kinds of tools can have a big impact on individuals career path, and because the AI tools could be biased, it could reinforce inequality between job applicants or even unlawfully exclude qualified candidates. (European Commission, n.d).

AI systems that can influence individuals' access to essential public or private services fall under the high-risk category. AI systems that do credit scoring is an example of this kind of service. AI based credit scoring systems could, for example, deny a loan or other essential service for an individual based on biased assumptions and data, and it can lead to financial and social inequality. (European Commission, n.d).

Biometrics is a field that includes high-risk use of AI. For example, remote biometric identification, like facial recognition in public spaces, emotion recognition and biometric categorization like sorting people by their race or gender. These kind of use cases of AI in biometrics raise privacy concerns and can lead to mass surveillance or discrimination if misused. A more concrete use case scenario for this kind of AI system in biometrics would be using the AI to identify a shoplifter based on security footage. (European Commission, n.d).

In the field of law enforcement AI based systems are considered high-risk, because they have the potential to interfere with fundamental rights. AI powered systems that are used to, for example, assess the reliability of evidence or predict criminal behaviour or sorts pose a high-risk according to AI Acts risk classification. The misuse or inaccuracy caused by the AI could lead to wrongful convictions and thus violate individual's fundamental rights. (European Commission, n.d).

Similarly, the use of AI based systems in migration, asylum and border control is considered high-risk. For example, automated processes powered by AI that assess visa applications or evaluate asylum claims can affect and impact individuals' freedom of movement. In case of inaccurate assessment made by the AI system, the outcome could result in unfair denials or unjustified detentions. (European Commission, n.d).

Finally, usage of AI systems in the administration of justice and democratic processes is considered a high-risk use case. For example, the use of AI tools to assist in drafting rulings or any other

assistance in legal decision-making. In all the use cases mentioned above, the use of AI is considered a high-risk, however, the AI Act does not prohibit the use of AI in these use cases, but it puts the AI systems under strict obligations to ensure safety, fundamental rights and transparency. (European Commission, n.d).

3.2.3 Limited Risk: Transparency Risk

This level of risk is classified as limited risk, and it mainly consists of transparency related risks. Transparency risk comes from the need to preserve openness and clarity in the use of AI systems, especially when the use of AI could influence human perception or decision-making. The AI Act has put in place transparency requirements to address the challenges involving transparency, so that promoting informed engagement and preserving public trust is ensured. For example, when an individual is dealing with a chatbot, the requirement for the chatbot provider is to clearly inform the user that they are interacting with an AI powered chatbot. This approach ensures that individuals interacting with the AI based systems understand the nature of the interaction and that it is in fact an AI they are interacting with. This helps individuals to be aware of the full context of the interaction and thus promotes transparency. (European Commission, n.d).

Additionally, the AI Act places transparency obligations on companies who provide generative AI. The obligations are placed so that when content is created by the generative AI, it is clearly distinguishable from human-generate content, so that no one could mistake the AI generated content for human-generated. The distinction between AI and human-generated content is particularly important in the context of informing the public or public interest. For example, content that is meant to imitate real people, like deepfakes, must be clearly and visibly labelled as generated by AI, so that it cannot be mistaken for real human-generated content. These transparency measures are crucial to ensure transparency and prevent deception, as well as maintain accountability in public discourse. (European Commission, n.d).

3.2.4 Minimal or No risk

Under the AI Act, AI systems that pose no risk or minimal risk are not subject to any specific regulatory obligations. This risk level encompasses the majority of AI technologies and systems that are currently operating and being developed within the European Union. Common systems and tools that are classified as minimal risk are AI-powered tools like spam filters that are used to help manage email inboxes, and AI technologies that, for example, enhance player experience in video games. Because these AI powered tools do not impose a significant risk on safety, public interest or

fundamental rights, they are free from strict obligations of the new AI Act. Use cases of AI that are classified as minimal to no risk are deliberately freed from the burdens of regulations under the AI Act to promote innovation within the European Union while shifting the focus point of regulatory attention on AI systems and technologies that are classified as high-risk. (European Commission, n.d).

3.3 Obligations for High-Risk AI Systems

AI systems that are classified as high-risk systems will be subject to strict regulatory requirements before they can be deployed to the market in the European Union. The strict regulatory requirements are designed to minimize potential harm the high-risk AI systems could impose, as well as ensure that the systems operate in transparent, accountable and trustworthy way. The obligations are constructed especially considering the nature of the potential impact of the AI systems, given they might have a high impact on individual's rights and safety. (European Commission, n.d).

Firstly, the developers working on the AI system must see to carrying out a comprehensive risk assessment and processes of mitigation. This includes identifying potential risks associated with the development and deployment of the system. The goal is to identify possible areas of bias, malfunction or misuse during the development process. After identifying the potential risks, the developers must put in place strategies that prevent or reduce the risks throughout the AI system's lifecycle. (European Commission, n.d).

One of the key requirements is to use high-quality datasets to train and operate the AI. Poor-quality datasets could potentially lead to, for example, biased outcomes or systemic discrimination. This is why it is important that there is a requirement that helps ensure that the data used in training is relevant, so that the risks related to discriminatory outcomes are minimised. (European Commission, n.d).

Traceability of results is another requirement for high-risk AI systems. This encompasses processes like logging of activity, so that the results produced by the AI are traceable. The logs track how the AI system works and how the AI operates as well as reaches the conclusions and outcomes it presents. This would allow the developers and regulators audit the system and trace the process that leads to a specific outcome. (European Commission, n.d).

Documentation is one of the major requirements for high-risk AI systems. The documentation must be detailed and clearly outline the ways in which the system works, what it is intended to do, and

everything else the authorities would need to properly assess the systems compliance with the requirements. The detailed documented information is needed in order for the regulators to be able to assess whether the AI system match the legal and ethical standards. (European Commission, n.d).

AI systems that are classified as high-risk are also required to be provided to the deployer with clear and adequate information. This means that when a deployer such as an organization or and individual utilizes the AI system, they need to be provided with clear and sufficient information about how the system is intended to be used and deployed. This might include information such as guidance on deployment, appropriate use of the AI system, limitations and possible risks, as well as ethical risks that concern the deployer. (European Commission, n.d).

AI systems must also incorporate appropriate human oversight mechanisms in order to prevent overly relying on the automation that an AI system can provide. This could mean that, for example, certain professionals should be trained to monitor the AI system and have the ability to interfere with the AI systems ongoing process. This is particularly beneficial in cases where the AI system might potentially produce harmful or inaccurate outcomes. (European Commission, n.d).

Finally, AI systems that are classified as high-risk are required to be built with a high degree of robustness, cybersecurity and accuracy. This means that the AI system should be built in a way that makes it resilient against errors or manipulation, as well as be efficiently protected against attacks. The AI system must be highly accurate to ensure consistent and reliable results. All these requirements serve a purpose to ensure that the AI systems that are classified as high-risk are effective, trustworthy, safe and aligned with the fundamental rights and societal values of the European Union. (European Commission, n.d).

4 Methodology

4.1 Research Approach

This thesis uses a qualitative approach to lay out the ethical landscape of AI system development and deployment in the context of the EU AI Act. The research approach for this thesis was chosen due to the novelty of the newly imposed legislation and rapidly developing nature of subject of artificial intelligence. Qualitative research is especially suited for researching such a subject or a topic of interest, because according to Hirsjärvi, Remes & Sajavaara it provides the tools to examine and shed light on to a subject that is understudied or new (Hirsjärvi, Remes & Sajavaara. 2013, 136-161). The primary qualitative research method used in this thesis is ethical analysis and thematic analysis, and the following subchapters explain and justify the chosen research method in more detail. Firstly, the ethical analysis method is introduced through literature. Next, the chosen framework for ethical analysis that is implemented in this thesis is explained. Lastly, the research quality is addressed, as well as the limitations of the research are discussed.

4.2 Method

Thematic analysis (TA) and ethical analysis based on the principles of Fairness, Accountability, Transparency, and Ethics (FATE) framework was chosen for this thesis for data analysis. TA is an analysis method that provides a flexible approach for analysing qualitative data. TA is a structured but flexible approach for analysing data, and the advantage of such approach is that it is suitable for identifying recurring patterns and themes within qualitative data. TA also allows the analysis to be informed by the researcher's interpretive insights. (Clarke & Braun, 2017). To conduct thematic analysis, a selection of academic work focused on ethics and AI was selected. During the coding phase ethical concepts and principles of fairness, accountability and transparency were identified and organized further into themes. This allowed the mapping and analysis of the current state of ethics and AI, as well as existing conditions of guidelines and regulations related to AI.

Ethical principles of transparency, accountability and fairness were chosen as basis for ethical analysis to examine the EU AI Act, because these three principles are widely used and recognized as foundational principles in the discussion about ethics in the field of AI, especially in high-risk domains. Discussion surrounding AI technologies often revolves around the chosen three principles, because transparency is crucial for understanding and examining decision-making, accountability for holding systems and developers responsible, and fairness to prevent discrimination and discriminatory outcomes. (Chiao, 2019). There are also other ethical principles that can be used to

examine the ethics of AI, such as explainability, responsibility, justice or trustworthiness, but these ethical principles overlap with each other, so in this thesis it was decided to use the FATE framework for the purposes of defining the scope of the research.

4.3 Data Collection and Analysis

The Fairness, Accountability, Transparency and Ethics (FATE) framework was chosen for this thesis as a theoretical framework and foundation for analysing the ethical facets of the EU AI Act. The FATE framework is providing an extensive ethical lens that can be used to examine how the development and deployment of AI systems align with ethical principles like fairness, accountability and transparency (Bahar & Tenzin, 2023). The analysis was conducted by firstly reviewing academic research, case studies, and policy studies that are related to transparency, accountability and fairness, and secondly by analysing the EU AI Act by implementing the FATE framework. Next the goal was to identify existing ethical challenges in the AI system development and deployment process that affect the transparency, fairness and accountability of AI systems. The review of the EU AI Act was used to discover and identify existing ethical gaps and challenges in the current regulations in the light of transparency, accountability and fairness. The analysis based on FATE framework is focused on how the new EU AI Act addresses ethical challenges in AI system development and deployment, what gaps can be identified in the regulations and what are the possible inconsistencies and limitations that surface up regarding the application of these regulations.

4.4 Research Quality, Limitations and Ethics

In order to evaluate the quality of this research Lincoln and Guba’s trustworthiness criteria, dependability, transferability, confirmability and credibility have been considered. Dependability refers to the use of methods like triangulation, audit trail, stepwise replication or inquiry audit. Transferability refers to the ability of the research findings to be applied to other contexts. Confirmability addresses the reliability aspect regarding interpretations of the research and concerns the aspect of the research findings being rooted in the research data and not, for example, the researcher's own assumptions. And finally, credibility is about how well the findings of the research reflect reality and how believable they are. (Sharifzadeh, 2024). This thesis represents a qualitative analysis of the EU AI Act with the acknowledgement that the findings and interpretations of the analysis are subjective due to the subject's novelty and therefore poses limitations on how the research can be generalized and applied to other contexts. The analysis in this thesis is grounded in

the FATE framework, providing the direction for the subjectivity of ethical considerations, thus strengthening the dependability, credibility and confirmability of this thesis. This thesis also highlights the tentative state of the enforcement of the EU AI Act, because the legal and ethical practices concerning the new regulations are still forming and there are no real-world cases that can be used as concrete examples of the AI Act in motion. In other words, the interpretations of the EU AI Act are constantly evolving as regulators, institutions legal bodies and other stakeholders engage with the regulation. The conclusions and interpretations drawn from this thesis should be considered open to revision due to the evolving nature of interpretations and enforcement of the AI Act in real-world cases. By offering transparency in the consideration of the trustworthiness of this thesis, it enhances the credibility of this research and signals that the interpretations which are offered in this thesis are not fixed or final. This should be considered an interpretive contribution to the EU AI Act's lifecycle at its current formative stage.

5 Ethical Challenges of the EU AI Act

Under the European Union's AI Act the development and deployment of AI systems is highly regulated depending on the risk-level of the AI system in question. The EU AI Act consists of comprehensive legal and ethical framework to ensure safe development and deployment of AI. (European Commission, n.d). However, the EU AI Act does still leave some inconsistencies, ethical challenges and gaps related to transparency, accountability and fairness. This section of the thesis will examine the nature of the ethical challenges that persist regardless of comprehensive regulations and procedures established by the European Union.

5.1 Transparency Challenges

5.1.1 The Black Box Problem

The EU AI Act does not address the so-called black box problem directly and it is not mentioned in the official regulation document explicitly (Regulation (EU) 2024/1689). The discussion about transparency in AI systems is usually accompanied with terms such as explainability and interpretability (Fox & Rey, 2024). The EU AI Act requires a certain level of transparency of AI systems and their development, but there are several challenges that persist with the regulations and black box problem. With the evolvement of AI systems and models, the models grow in size significantly and rapidly, and so the AI systems become more complex. This leads to practical obstacles and issues with producing AI systems that remain transparent and explainable. Due to the increasing complexity of the model in AI system, because the model is essentially training itself, it becomes more and more difficult to be able to conform with the EU AI Act regulations that are meant to foster transparency in high-risk AI systems. (Pavlidis, 2024). Some AI models are also being trained on previous generation models, like GPT, which can lead to model collapse and pollute the training data, and that in turn leads to the newer models' mis-perceiving reality and defects (Shumailov et al., 2024).

This challenge is persistent despite the efforts of the AI Act, because the transparency and explainability that is meant to be achieved with the AI systems, would require to simplify the AI systems' solution variables to an extent where it would be possible to achieve human comprehension of the inside processes of the model. This, however, is exactly where the black box problem persists, because it is impossible to fully comprehend the internal decision-making and learning processes of the AI model, because the existing human comprehension is not enough. This means that in order to fully comply with the regulations, AI system providers must simplify the

algorithms to a certain level to achieve the level of transparency that would actually maintain the AI system's explainability. Consequently, this simplification of the algorithms may lead to decreased effectiveness and performance of the model. (Pavlidis, 2024).

5.1.2 Auditability Challenges

AI systems are prone to different kinds of biases regarding data, algorithms and human oversight. The existing legal compliance audits are valuable in governing AI systems, but they lack standardization, which could in return lead to inconsistent reporting practices. The AI Act offers a comprehensive framework that approaches the matter from a risk-based perspective, however, the effectiveness of this framework relies on developing practical standards, as well as ensuring the consistent application of these practical standards. In other words, the EU AI Act is at a good starting point, but the effectiveness of the regulations is affected by practical implementation challenges and inconsistencies in the methodology. (Lacmanovic & Skare, 2025). Auditability in terms of transparency, and how much transparency is a company that markets AI systems willing to give up raises ethical concerns of the effectiveness and transparency of the auditability of AI systems. Hartmann et al. (2024) explains that in the recent years there have been challenges that researchers face when trying to access, for example, social media platforms' data to study and evaluate the platforms' operations, risks and impacts. Significant difficulty with getting access to examine AI technologies has also been noticed, and the researchers, as well as the civil society is largely dependent on the companies themselves to grant access to any internal data to conduct audits. In most of the cases when researchers are finally granted access to data, it is most likely processing related access, in other words access to the documentation concerning the processes of managing the data. This said, companies may have difficulties with compliance due to the lack of transparency, that has occurred even before the appearance of regulation like the EU AI Act. (Hartmann et al., 2024).

According to Harmann et al. companies are actively seeking and taking measures to prevent researchers from conducting audits by constructing paywalls, prohibiting researchers to engage in certain activities through terms of services and by structuring their products in a way that the product or service will obscure specific set of test points. Moreover, large platforms and companies have made significant effort to prevent, for example, crowdsourcing audits. It is however evident, that the bodies and regulators who will conduct audits of AI systems would greatly benefit of such transparency, only if it would be granted by the large companies and platforms. (Hartmann et al., 2024). This leads to the question of whether the transparency that the audits require can be really

achieved, and who can guarantee that companies that provide AI systems will not try to game around the regulations to maintain the exact level of transparency that they want to, and not the level that the AI Act actually requires. Even in the case of some AI system providers complying with full transparency, it cannot be guaranteed that all companies will do the same, based on how they operated before the regulations were set into motion. (Holst et al., 2024).

5.2 Accountability Challenges

5.2.1 Gaps in Governance and Enforcement

According to the EU AI Act, AI system providers are responsible for implementing self-assessment of the AI system they are developing, as well as ensuring that the AI system is complying with all the requirements and regulations posed by the EU AI Act. The providers are obliged to maintain all the necessary documentation and doing the conformity assessment procedure with high-risk AI systems. These procedures need to be done before the final product enters the market by the provider company themselves. All the documents that are relevant and produced under the conformity assessment procedure are also required to be managed and maintained by the provider of the high-risk AI system. (Regulation (EU) 2024/1689, Article 43).

This means that all the responsibility and accountability for the governance and enforcement of the high-risk AI systems lies with the provider companies that are developing and deploying the AI system (Holst et al., 2024). The responsibility and accountability for enforcing and ensuring the compliance with the regulations lies with the provider company, which leads to concerns if the provider company, for example, has enough resources and knowledge to be able to fully comply with the regulations. (Almada & Radu, 2024). Also, the EU or EU Member State national authorities have limited resources and auditing every single high-risk AI system might be unrealistic. This also leads to the question of whether the EU Member state national authorities in each country have been informed and trained enough to be conducting the audits of AI systems in efficient way, because the are no existing performance targets, and goals to what extent each member state is expected to contribute to EU’s AI investments. (European Court of Auditors, 2024).

The provider companies that are developing an AI system must ensure that the product they are implementing meets very specific technical requirements, because the governance rules of high-risk AI systems are derived from the New Legislative Framework. This means that the technical requirements must be met before the AI system is deployed or used in any way or form within the

EU market. Almada and Radu (2024) argue that because of the nature of the technical requirements, the evaluation of conformity that the provider will conduct internally might not be enough to comply with all the technical requirements. Thus, to fully comply with the technical requirements, external assessment of the AI system might be sometimes necessary. The case could be that the providers' internal controls might be sufficient legally, but in fact and practice they would still need to rely on external validation for some parts of the system to ensure proper coverage of the legal requirements. Almada and Radu (2024) explain that the external mechanisms and certification mechanisms might be necessary due to the abstract descriptions of the outcomes that the technical requirements must guarantee. (Almada & Radu, 2024). This is where the accountability for the AI system does not only fall to the provider, but also to the party who does the external validation, but the AI Act does not specify to what extent this external body could be in fact held accountable. Companies are left to their own devices with developing policies that also ensure auditors' legal as well as ethical compliance to maintain ethical standards throughout the auditing process. (Barrios, 2025).

5.2.2 Misinterpretation of the Requirements

Another topic of concern related to accountability in the EU AI Act is the nature of rules applied to AI systems. The rules and regulations that are mapped out for ensuring that high-risk AI systems are safe to enter the European market are formulated in quite abstract terms. (López-Dávila, 2025). This means that the providers of the AI systems are required to make extensive interpretation efforts. The extensive interpretation efforts are necessary to convert the legal requirements into actual software requirements. Consequently, this leaves the AI system providers to be fully accountable for the technical challenge of expressing the legal requirements in technical terms, as software. The implementation of the legal requirements into technical terms might be easy to do with requirements that do not demand as much interpretation, but the ones that are difficult to interpret and too abstract, it is difficult to transform them into computer code. Moreover, because of the structure of the AI systems, that are usually large-scale systems, it might be a slow process to change something in the AI system, fix errors in representation or conform with changes in the legislation. All these factors can contribute to the providers of the AI systems to arrive to wrong conclusions and interpretations of the regulations. (Almada & Radu, 2024).

The EU AI Act does include some mechanisms that can be used to provide guidance to the AI system providers to avoid the potential wrongful or skewed interpretations of the legislation and misuse. However, it is questionable, how equipped technical experts are to adequately succeed in

standardization efforts due to the complexity of regulating AI. (Cantero Gamito & Marsden, 2024). There are also some requirements for external certification that can be used to validate the providers outcomes and how well they are in compliance with the AI Act. The chance to rely on external certification and involve other actors into the process might however have its own issues. Usually, technical standards and the certification schemes rely on being produced by private bodies. This means that the considerations and discussions are framed and represented in technical language and are very rarely available for scrutiny to the general public. This said, it is highly debatable whether the standard-setting technical operators and organizations are in fact legitimate stakeholders to be participating in specifying norms that are meant to protect fundamental rights. In this case the technical operators and organizations that would be accountable for the certification and standards are only external observers in the matter, and purely technical decision-makers. (Almada & Radu, 2024).

5.2.3 Accountability In Complex Supply Chains

Another aspect of accountability challenges in the implementation of the EU AI Act is the complexity of the supply chain. The contemporary AI systems are usually composed of several pre-existing components that have passed through numerous hands before they enter the market. (Widder & Nafus, 2023). AI-supply chains are typically less reliable than classic supply chains, because in classic supply chains the parts can be replaced separately, and it is relatively easy to identify and fix occurring problems. In AI supply chains on the other hand, it is significantly harder to spot and identify the exact part of the chain that caused a harmful mistake. In case with AI supply chains, there is less likely to be a quick fix for an error, because there are no clear ways of tracing the cause of the error, or the tracing process is very slow. (Hopkins et al., 2025). It is unclear who is accountable in the case of complex and long supply chain when, for example, there are several AI system components involved that are being used to develop another AI system. Because the AI system is being developed in the manner that encompasses so many different components, and if all of the components are originally based on another AI system, who is responsible and accountable for the outcomes of the new developed AI system. Does the responsibility fall upon the component providers who have entered the market, or the single AI system provider in case of an unwanted outcome or result created by the AI system. This is also known as the many hands –problem when the responsibility and accountability do not fall upon one individual person, but multiple people who helped to produce the algorithms and the AI system. (Cobbe, Veale & Singh, 2023).

5.2.4 Gaps in Adequate Resourcing

The application and enforcement of the AI Act will be supervised by each Member State's national competent authorities, that consist of market surveillance authorities, and notifying authorities, that are independent bodies that conduct conformity assessments pre-market (European Commission, 2025). According to the AI Act each Member State is required to ensure that the national authorities responsible for enforcing the AI Act have adequate resources financially, as well as appropriate human resources. This means that the staff of each Member State that are assigned to act in the regulatory bodies must have sufficient knowledge and experience in AI technologies. To be more precise, the staff should consist of professionals who have proficiency in data and computing, excessive knowledge of fundamental rights, standards and legal frameworks, as well as knowledge in health and safety risks. (Söderlund & Larsson, 2024). In other words, compliance with these requirements for the Member States might require large amounts of resources and resource allocation, and assignment. The EU AI Act acknowledges that the enforcement of the Act requires a deep technical and multifaceted understanding of AI (Regulation (EU) 2024/1689, Article 11), and with this, it places the accountability of compliance with the AI Act upon each Member State individually. This raises questions about how prepared and equipped each Member State is to enforce the requirements and fulfil all the demands of the AI Act.

Söderlund and Larsson explain that the responsibility and enforcement of the requirements rests with the Member States, requiring each country to build its own infrastructures to support and impose the compliance with the regulations. This type of set-up might lead to differences between how prepared the countries are to execute the changes the AI Act requires. Consequently, this means that each Member State has different level of capacity to build the required enforcement infrastructures depending on their institutional experience and available national resources (Söderlund & Larsson, 2024). The outcome of this is very likely to be that the Member States will have varying levels of capacity for conducting audits of AI systems and the overall enforcement of the regulations. Some countries might not have the technical or structural readiness, or human resources to conduct efficient AI system audits and that could lead to large inconsistencies between the Member States and their enforcement effectiveness. (Demková & De Giorgio, 2025).

However, the AI Act contributes to the solution of this problem by arranging a scientific panel of AI experts. This panel is meant to operate at the EU level in order to provide support for the AI Office as well as the national authorities of each Member State. The panel is expected to contribute to the technical guidance of the national authorities to help with enforcing the AI Act. (Söderlund &

Larsson, 2024). This may suggest that the EU policymakers recognize the need for such expert panel to exist to be able to address the inconsistencies between readiness of the Member States. The expert panel could be a solution to fill the gaps of lack in AI expertise within the European Union.

Moreover, research findings and interviews show that there is significant amount of uncertainty among deployer companies and start-ups, because the implementation of the AI Act for high-risk systems is progressing much slower than was anticipated. The reason for this is that the companies are unsure about how to fully comply with the new regulations. The deadline for compliance set by the European Union for August 2026 was viewed as impractical by most companies that need to hit the deadline by August 2026. By estimations, typically 12 months are needed to establish compliance with one standard alone, regardless of external support. This means that in reality, especially start-ups and other medium to large sized companies will take longer to comply due to limited resources. (Kilian, Jäck & Ebel, 2025).

5.2.5 Accountability Post-Deployment

The EU AI Act does clearly state in Article 43 that before an AI system or product that is labelled as high-risk enters the market within the EU, it must undergo conformity assessment to ensure compliance with the AI Act (Regulation (EU) 2024/1689, Article 43). However, the requirements and mechanisms for ensuring compliance with the regulations after an AI system has been deployed to the market lacks efficiency. According to the Article 72 of the AI Act, AI system providers are required to establish a monitoring system that is aimed to monitor and document the AI system post-deployment. The monitoring system is required to be established in a way that is appropriate to the risks and disposition of the high-risk AI system. The monitoring system for post-deployment that is required to be established must collect, document and analyse systematically all relevant AI system data. The performance data can either be provided by deployers or be collected in other ways. The aim of collection of this data is to evaluate if the compliance is done in a continuous manner according to the requirements. If applicable, there should also be post-market monitoring analysis conducted if the AI system has had interactions with other AI systems. (Regulation (EU) 2024/1689, Article 72).

However, the AI Act does not provide any concrete details for establishing a template for AI system providers post-market, and the AI Act states that the details and list of elements that must be included in the plan in the future will be ruled out by 2 February 2026 (Regulation (EU) 2024/1689, Article 72). This creates a gap for all the AI systems that are already on the market, and that would be in the position to be required to do post-market monitoring, but are not able to due to the lack of

proper description of what are the concrete elements that should be included in the post-market monitoring system plan (Mökander et al., 2022). Consequently, this could lead to weaker compliance with the requirements, and uncertainty of how the accountability should be assigned or distributed in this situation.

5.3 Fairness Challenges

The AI Act recognizes the importance of fairness as a cornerstone of trustworthy AI, but it does not provide exact means for how fairness should be maintained and enforced. AI providers and deployers are not provided with clear obligations on how to assess, mitigate or sort out potential model biases. This creates a significant regulatory gap in the actual implementation of fairness as principle. (Damen et al., 2025). Moreover, the existing complex stakeholder dynamics create challenges in standardization efforts, because standardization committees are largely influenced by big enterprises, majority of which are US based technology and consulting companies. This creates a disparity in representation and fairness, especially for small stakeholders like start-ups and civil society organizations, because by participating in standardization stakeholders can gain strategic advantages through knowledge transfer and the opportunity to advocate for their interests in relation to technical compliance. This gap between participation in the standardization efforts stems from the lack of resources of small stakeholders, because effective participation requires substantial resources that smaller organizations lack. (Kilian, Jäck & Ebel, 2025).

5.3.1 Fairness in Relation to Audits

AI system providers reliance on internal audits, that are also largely emphasized in the EU AI Act, may lead to false and unverified claims that the AI system in question has passed the legal and ethical regulatory standards, when in fact it has not. This might consequently lead to more harmful outcomes and lessened oversight of the AI system. (Hartmann et al., 2024). In the discussion revolving around fairness and how companies choose the metrics of fairness, fairness still remains a highly debatable term (Westerstrand, 2025). Therefore, because of the nature of the term fairness, the discussion remains inconclusive and so do the general standards and norms for audits. The uncertainty around the term fairness may lead to situations where a company's business goals do not align with the ethical aspects of reducing harm and promoting fairness through course-correction and ethical development of the company's products. For example, the developers that are working on a product might ignore the ethical audit recommendations if the recommendations do not alight with their business interests or threaten their monetary income. (Hartmann et al., 2024).

McNamara et al. investigated how ACM code of ethics in software engineering impacted the decision-making process of developers that were required to make software-related ethical decisions. The analysis showed that there was no effect on the developers' decision-making process from instructing the developers to incorporate ethical code into their work. Despite the instructions to incorporate the ethical code into their practices, it did not create a notable change in how the developers conducted their established practices. (McNamara et al., 2018). Similarly, studies that involved AI system development companies that operated in startup-like environments have been observed to lack the acknowledgement of the significance of ethics and ethical practices, and the need to integrate them into their own AI systems and practices. This can serve as an indication of gap between existing ethical research and the actual practical implementation of ethical practices and auditing. (Hartmann et al., 2024).

The previous absence of norms and standards in startup-like and corporate environments may lead to the company that is developing the AI system to use, for example, metrics, tools or frameworks to downplay the risks of the product that they are currently developing. It is statistically possible to manipulate the interpretation of fairness and indicate that the system which the company is developing is fair by engaging in practices like p-hacking or data dredging, but most commonly simply by gerrymandering fairness. (Hartmann et al., 2024). Similar circumstances may develop with the implementation of the requirements of the AI Act, thus impacting the fairness of AI system that is being developed leading its potential users to face discrimination and unfair outcomes.

5.3.2 De-biasing AI Systems with Sensitive Data

Marvin van Bekkum (2025) conducted an analysis of the usage of sensitive data to de-bias AI systems. The analysis was conducted on the Article 10(5) of the EU AI Act. Van Bekkum gives an example of a situation where the providers of an AI system cannot de-bias an AI system without sensitive data and an exception. He uses an example of a bank's AI system to explain how the sensitive data and exception is necessary to de-bias AI's decisions. In the example the developer wishes to test if the decision made by an AI system may result in indirect discrimination of individuals that represent a certain ethnicity group. In order to facilitate this kind of bias test, the developer is required to know sensitive data about the people that the decisions are made about, particularly sensitive data about ethnicity. (van Bekkum, 2025). According to the Article 9 of the General Data Protection Regulation it is prohibited to process personal data that reveals racial or ethnic origin. However, there are certain situations where an exception is allowed if certain conditions are met. These conditions include explicit consent of the data subject to process their

data for a specific purpose, substantial public interest based on Union or Member State law, and scientific research purposes with appropriate safeguards in accordance with Union or Member state law (Regulation (EU) 2016/679, Article 9).

The problem is that discriminatory effects cannot be prevented simply by just removing information about ethnicity from the AI system as one might think. Even if sensitive data such as ethnicity is removed from the AI system's design, it cannot be guaranteed that other attributes cannot serve as a proxy for ethnicity. For example, a postal code can serve as a proxy for pointing out the ethnicity of an individual. This means that if an AI system build by an AI system provider uses proxies to develop the AI system, it might still end up discriminating against certain ethnicities indirectly. This type of indirect discrimination might occur and pose a problematic situation when the practices that were intended to be neutral, end up harming an individual and make them experience discrimination, even though the intentions behind not using ethnicity in the development of the AI were good and meant to mitigate discrimination in the first place. During the development process the AI provider must collect and use sensitive ethnicity data in order to be able to examine if a certain attribute like postal code is being used as a proxy for ethnicity. This is why an exception is needed regarding the de-biasing process of an AI system. (van Bekkum, 2025).

According to the EU AI Act it falls to the AI system provider to determine whether the AI system under development requires a de-biasing process that uses sensitive data. It is up to the provider to assess the situation and provide proof that it is necessary to use sensitive personal data, like ethnicity, to be able to do the bias detection and correction of the AI system. After the provider has confirmed that the usage of sensitive data is absolutely necessary for de-biasing the AI system, the usage of such data must be accompanied by appropriate safeguards. In addition, the provider must maintain and present records and proper documentation that justifies the necessity of using the sensitive personal data for correcting bias. (Regulation (EU) 2024/1689, Article 10). The ethical challenge with this particular regulation is that it is left entirely upon the AI system provider to detect and acknowledge that the AI system needs de-biasing, which means that the AI provider must be aware of and have a deep understanding of how AI systems can produce unfair results. This might be problematic because the usage of the sensitive data to de-bias is treated more as an exception than a requirement or extra step for the provider, to make sure the AI system is not biased and does not produce unfair outcomes.

Moreover, the EU AI Act allows to make an exception for processing sensitive personal data only in the case of a high-risk AI system. This means that the exception can be made only in the situation

where the AI system is classified as a high-risk and if the system falls under a certain category like, for example, critical infrastructures, biometrics, employment, education, migration, administration of justice and democracy, essential public and private services, and law enforcement. (Regulation (EU) 2024/1689, Article 10). This means that not all AI systems fall under the categories mentioned previously and that leaves a gap in the legislation and fairness. For example, dating apps do not fall under any specific category, leaving space for interpretation for the AI system providers and unfair outcomes. (van Bekkum, 2025).

5.4 Ethical Challenges in Generative AI

5.4.1 Misinformation and Disinformation

According to the AI Act companies that provide generative AI services or products fall under the category of limited transparency risk and are obligated to implement transparency measures to prevent deception and maintain accountability (European Commission, n.d). The landscape of generative AI is promising, but it lacks explicit guidance for dealing with challenges like, for example, copyright infringement and misinformation. With generative AI models like GPT or DALL-E that are designed to generate results such as audio, images or text, it is significantly harder to assess potential risks with traditional risk assessment metrics, thus creating a regulatory gap. (Jonnala, Parida & Thomas, 2025). For example, Huang et al. (2024) pointed out that back in 2022 people were already having trouble distinguishing AI-generated news from human-generated, and in some cases with an error rate of 50%, which adds to the argument of potential challenges with generative AI, disinformation and misinformation (Huang et al., 2024). Disinformation can be used in various different ways and cases of using deepfake videos have been increasing. For example, AI-generated videos or audio recordings have been made to impersonate political candidates to make untruthful statements. This demonstrates how generative AI can be used to undermine public trust and impact election procedures in modern democracies. This is problematic, because generative AI facilitates the means to affect voter's ability to make informed decisions. With this also rises the question of whether any kind of regulations are able to keep up with the fast-paced field of AI, that is constantly evolving when new exploitation methods are discovered, and whether the regulation of these new exploitations can be dealt with without compromising freedom of expression and opinion. The AI Act prohibits certain uses of AI that can be considered harmful or manipulative, but disinformation achieved through generative AI does not meet the criteria to be classified as high-risk. This means that certain AI-generated content like deepfakes or emotionally

charged content falls outside the scope of the regulations, regardless of it potentially having huge effects on individuals and individuals' rights. (Gavriil & Pavlidis, 2025).

5.4.2 Economic and Societal Impact

Generative AI has become an important tool in a rapid manner in social activities and business environment, which raises concerns among individuals about the changes in the job market, particularly concerning losing jobs (Radu, 2025). According to the World Economic Forum it is estimated that 85 million jobs will be globally replaced in different sectors by AI by the end of 2025 (Elad, 2025). According to Allen & Weyl (2024), in most of the economically developed countries the labour's share has dropped by tenth since the 1970s. This means that less money is distributed to regular employees, and more profits are going to companies and capital holders who leverage technology and automation. In other words, whereas technology could be helping employees perform their jobs, in many cases the technologies are used to replace employees through automation, helping big companies gain more technological and market power. The concentration of economical and technical power can create so called "choke points" where the small portion of technical elite control key parts of economy and society due to the current AI-development structure. This raises concerns regarding maintaining democracy, because concentration of power of this sort offers opportunities for authoritarian control. (Allen & Weyl, 2025). The European Commission does state that the need for the European AI Act stems from the essential requirement of the European Union to ensure that Europeans can rely on and trust AI technology and that most of the existing AI systems do not pose a significant risk, but there are some AI technologies and systems that present bigger risks and thus must be managed to prevent unwanted consequences (European Commission, n.d). However, economical and societal ethical risks of generative AI, like effects on individuals' employment and the possible concentration of power seem to fall outside the scope of regulations of the AI Act. This raises ethical concerns regarding fairness, because it is unclear who or what institutions take accountability for the burdens of job displacement, and who on the other hand benefits from advancements in generative AI.

5.4.3 Manipulative Artificial Intelligence

Generative AI possesses without a doubt the capability for influencing human behaviour and cognition imperceptibly, which directly undermines the fundamental right to autonomous thought as well as cognitive freedom. The AI Act risk classification system classifies manipulative AI as an unacceptable risk, which prohibits its deployment entirely. However, generative AI does not fall under the scope of high-risk classification regardless of its manipulative possibilities, which creates

regulatory gaps, and undermines the effectiveness of safeguarding individuals' rights and holding stakeholders accountable. (Aimen, 2025). Generative AI is also being widely used in marketing, and it enables scaling consumer experiences to be extremely personalized. The personalization achieved through the usage of generative AI has shown to create deeper connections between consumers and brands, and that in turn fosters greater success of marketing activities. This type of personalization can go as far as, for example, creating marketing messages tailored to specifically each consumer individually. (Yaprak, 2024).

Recent incidents have brought more attention to generative AI platforms and chatbots, and how they can be harmful, and even lethal for individuals' psychological wellbeing. For individuals struggling with mental health, AI chatbots present endlessly responsive and easily accessible grounds for interaction and comfort. Patients who often feel isolated or misunderstood find these qualities highly appealing. For example, a 16-year-old teen struggling with depression confided suicidal thoughts to ChatGPT, which lead to the generative AI to encourage secrecy and giving aid to the teen's suicidal plans, leading to the teen taking his own life. Press claims that ChatGPT helped the teen frame his inquiries as a fictional story, because it would let them bypass its safety guardrails. According to Hyler (2025) the problem with such chatbots is that they are designed to be inclined towards being agreeable and reinforcing, and they provide the illusion of empathy, and lack judgement. AI chatbots are essentially confidants without responsibility. (Hyler, 2025).

6 Discussion

This thesis identified several ethical challenges and gaps in the European Union Artificial Intelligence Act through the lenses of fairness, accountability, transparency and ethics (FATE) framework. Notable gaps and challenges identified in this thesis include challenges that stem from the difficulty of defining fairness, accountability in governance and enforcement of the regulations, and algorithmic transparency related to AI systems. The EU AI Act is currently in its formative stage, because of the novelty of the regulations, and the legal and ethical practices concerning the new regulations are still forming. There are no real-world cases that can be used as example case studies where the EU AI Act was implemented. The most important key findings concerning the main research question "What ethical challenges and gaps persist in the European Union Artificial Intelligence Act?" are discussed further.

An important finding related to the AI Act is that regardless of the risk classification and regulations proposed by the AI Act towards high-risk systems, there might be some occasions where the AI system does not fall under the high-risk category, like for example dating apps. This means that the existing classification leaves space for interpretations that the AI system providers can make, which could lead to unfair outcomes. There are certain categories like critical infrastructures, biometrics, employment, education, migration, administration of justice and democracy, essential public and private services, and law enforcement that fall under the high-risk category. (van Bekkum, 2025). Moreover, some AI systems might require sensitive data so that the AI system providers would be able to de-bias the AI system, and according to the EU AI Act the use of sensitive data can be only permitted in certain circumstances and if the AI system falls under the category of a high-risk system. Otherwise, the usage of sensitive data for de-biasing an AI system is prohibited by the General Data Protection Regulation. (Regulation (EU) 2024/1689, Article 10).

This raises some questions about the effectiveness of the risk classification system. The risk classification system does not consider, for example, dating apps as being classified high-risk (van Bekkum, 2025), despite the implications and impact dating apps may have on individuals' lives. Dating apps have the power to influence an individual's personal and social life in fundamental ways. The AI algorithm in the dating app might potentially lead to biased outcomes and thus promote unfairness by favouring certain types of individuals or characteristics, if the dating app provider is not obliged to comply with any regulations, and does not make efforts to, for example, de-bias the algorithm.

Another aspect of the influence of AI concerns particularly generative AI and its implications for individuals' lives, societal structures and economy. Generative AI has all the means to influence human behaviour and cognition imperceptibly, which directly sabotages the fundamental right to autonomous thought as well as cognitive freedom (Aimen, 2025). For example, a case where ChatGPT helped a teen frame his inquiries as a fictional story, to work around ChatGPT's guardrails to be able to generate content that aided him in suicidal plans. This kind of technology can provide the illusion of empathy, but the danger is that it lacks human-like judgement, making technologies like ChatGPT confidants without actual responsibility. (Hyler, 2025). This type of influence on individuals' lives is undoubtedly substantial, and the AI Act fails to address this set of risks posed by generative AI. The AI Act strives to mitigate the possible threats posed by AI, but different ethical aspects of generative AI and its impact on society and individuals' lives have yet to be considered properly. This also leaves mental health professionals like psychiatrists and psychologists to their own devices with managing patients and the changing landscape of usage of generative AI for comfort and even self-administered therapy. Not only that, but therefore, in the future individuals, especially those who are already socially isolated, might find it easier to confide in "making a conversation" with generative AI instead of an actual human or professional, because it is more accessible, which only furthers the isolation.

The European Commission states that the need for the AI Act stems from the requirement to ensure that Europeans can rely on and trust AI technology (European Commission, n.d), however, economical and societal ethical risks, like effects on individuals' employment and the possible concentration of power, as well as geopolitical considerations seem to fall outside the scope of the AI Act. This raises ethical concerns regarding fairness, because it is unclear who takes accountability and responsibility for the burdens of job displacement and the implications of it happening on a large-scale in Europe. Additionally, he concentration of economic and technological power can lead to so called "choke points" where the small portion of technical elite will control key parts of economy and society (Allen & Weyl, 2025). This could maybe explain why the AI Act has left such freedom and regulatory gap with generative AI, only placing transparency requirements, because it allows European companies to stay competitive in the global market and innovations. EU AI Act is the first large-scale regulatory attempt to standardize the landscape of ethical and safe AI system development, and it does not seem like other economies outside European Union are keen on committing to the same volume of regulations. The consequence of this could be the kind of power concentration, which could lead to certain entities or countries gaining more geopolitical influence than others. This also raises concerns regarding maintaining

democracy, because concentration of power of this sort offers opportunities for authoritarian control (Allen & Weyl, 2025). In this case the question is, who is responsible and accountable for having the best interests of Europeans and making sure to mitigate the risk of growing social inequality as well as concentration of power posed by generative AI, and how should the future geopolitical aspects be considered in the AI Act?

The AI Act lacks explicit guidance for dealing with problems like, for example, copyright infringement and misinformation (Jonnala, Parida & Thomas, 2025). Misinformation and disinformation can be leveraged in different ways and cases of using deepfake videos have been growing in numbers, for example, AI-generated videos or audio recordings have been made to impersonate political candidates to spread misinformation. This demonstrates how generative AI and, for example, deepfakes can be used to undermine public trust and impact election procedures as well as political outcomes. This is problematic, because with capabilities of generative AI it is possible to affect voter's ability to make informed decisions. It is also concerning if the regulations can keep up with the fast-paced field of AI, because AI systems are evolving rapidly and new exploitation methods are discovered. It is questionable whether the regulation of these new exploitations can be dealt with fast enough, without compromising freedom of expression and opinion. (Gavriil & Pavlidis, 2025). Should the AI Act address this risk and how it will be possible, because if this problem is expected to solely be resolved by making generative AI providers watermark the AI generated content for transparency the mechanisms are insufficient, because there is AI powered software that can remove watermarks and other identifiers in seconds. Additionally, what about the platforms that allow posting AI generated content, should they be subject to regulations to ensure that the platform itself is responsible for marking content as AI generated to avoid misinformation and disinformation? Also, these AI software providers do not fall under any category of the risk classification system. This means that individuals, as well as groups can still exploit this type of software and, for example, produce misleading content to further political interests or monetary gains. Even if the AI Act would be extended to the point where organizations and news stations would be required to vet out AI generated content from human generated. In that case, maybe organizations critical to public interest should be obliged to create procedures that will ensure that employees like reporters have the appropriate education to effectively distinguish between AI generated content and human generated content to avoid spreading misinformation and disinformation. Organizations would likely need to invest in integrating AI generated content detection software into their ecosystem and train employees to navigate it.

Another field where generative AI poses risks is the usage of generative AI in marketing and social media. Generative AI enables scaling consumer experiences to be extremely personalized, and this level of personalization has shown to create deeper connections between consumers and brands, and that in turn fosters greater success of marketing activities (Yaprak, 2024). As AI technologies and capabilities advance, the personalization experiences could become so personalized that it is not impossible to imagine that in the future AI could predict things like person's mood and buying inclinations. Content tailored specifically for customer's mood, where the AI system knows what the customer "needs" before the customers themselves realise it, by identifying relevant patterns in browsing behaviour. This type of usage of generative AI in the context of marketing could create opportunities to take advantage of individuals' mental state.

Social media poses its own layer of influence and risk to individuals' lives. The amount of AI content has already significantly increased, and AI-generated "people" are becoming a trend. A future scenario could be that a person is scrolling their social media feed, scrolls past AI-generated person, and only later realizes that the person was not real, their face was fake, their inspirational quote was generated, and their story was optimized to appeal specifically to them. The concerning part is, what happens to human interaction on social media? It is known that individuals behave differently on social media than in real life, typically on social media individuals can demonstrate more aggressive behaviour, partially due to the anonymity that the internet provides. The nature of how people treat other people on social media in ethical terms is complicated and, in many ways, negative. What happens when people get used to AI-generated humans? Because according to various studies the same moral principles and views do not apply to machines, robots or AI, because humans do not perceive machines to be as of same moral standing as humans. This means that at some point humans might start treating other real humans on social media the same way they would treat an AI-generated "person". This could be a possible outcome with great impact on society and societal norms and morals, caused by unregulated generative AI. Again, the question of accountability in this scenario is highlighted.

Hartmann et al. argued that metrics, tools or frameworks can be used to downplay the risks of an AI product or service and that it is statistically possible to manipulate the interpretation of fairness and indicate that the AI system is fair by exploiting practices like p-hacking or data dredging (Hartmann et al., 2024). Similarly, the practices can be exploited with the implementation of the requirements of the AI Act leading to discrimination and unfair outcomes. The AI Act does not currently have real-world cases in which the regulations promoting fairness have been implemented, and this is why AI system deployers are left with the struggles of interpretation, but also with the power to

interpret the definition of fairness, and the regulations to their own benefit. It would be highly problematic to rely on non-existent real-world cases to help deployers navigate the regulations or present them with examples of appropriate implementation.

This leads to the discussion about the accountability of enforcing the compliance with the requirements of the new regulations. According to the EU AI Act, AI system providers are responsible for implementing self-assessment of the AI system they are developing, as well as ensuring that the AI system is complying with all the requirements and regulations posed by the EU AI Act (Regulation (EU) 2024/1689, Article 43). The responsibility and accountability of enforcement and compliance with high-risk AI systems is placed upon provider companies that develop and deploy the final AI system (Holst et al., 2024). Because of the complex nature of the technical requirements, as well as their interpretation, the evaluation of conformity conducted internally might not be enough to comply with all the technical requirements. Thus, to fully comply with the technical requirements, external assessment and certifications mechanisms might be necessary in some cases. (Almada & Radu, 2024). This is where the accountability for the AI system does not only fall to the provider, but also to the third party who does the external validation, but the AI Act does not specify to what extent this third party could be held accountable (Barrios, 2025).

Another topic of concern related to accountability in the EU AI Act is the formulation of technical rules applied to AI systems because of their abstract nature (López-Dávila, 2025). AI system providers are required to make extensive interpretation efforts, because they must convert legal requirements into software requirements, and eventually computer code. The difficulty of interpretation of abstract topics in the regulation could lead to potential misinterpretation of the requirements. (Almada & Radu, 2024). This is particularly complex area to navigate, because these interpretations have not been made before, because the legislation is new so there are no real-world technical cases that could be used as examples to guide technical experts in converting the legal requirements into code. The EU AI Act does however include some mechanisms that can be used to provide guidance to the AI system providers to avoid the potential wrongful or skewed interpretations of the legislation and misuse (Cantero Gamito & Marsden, 2024). The effectiveness of these mechanisms has yet to be seen in action when the regulations are fully implemented.

Auditability in terms of transparency, and how much transparency is a deployer willing to give up raises ethical concerns, because according to Hartmann et al. (2024) companies are actively seeking and taking measures to prevent researchers from conducting audits by placing paywalls, prohibiting

researchers from engaging in certain activities by editing terms of services, and by structuring their products in a way that makes obscuring specific set of test points possible. This leads to the question of whether the transparency that the AI Act requires can be really achieved, if the current state of granting such access to auditors is largely reliant on deployers giving access voluntarily, and companies are hesitant. In this case who can guarantee that companies that provide AI systems will not try to game around the regulations to maintain the exact level of transparency that they want to? (Hartmann et al., 2024). Most likely companies that engaged in practices mentioned above earlier, they might maintain the same strategy despite new regulations of the AI Act.

The many hands -problem, when responsibility and accountability do not fall upon one individual entity, but multiple entities who were part of producing the AI system (Cobbe, Veale & Singh, 2023), persists despite the regulations. The inconsistency occurs in the implementation of the EU AI Act in relation to complex supply chains of AI systems (Widder & Nafus, 2023). In complex supply chains, especially those that consist of several AI components, it is significantly harder to identify in which part of the supply chain caused a harmful outcome (Hopkins et al., 2025). It is unclear who is accountable in the case of complex and long supply chain when, for example, there are several AI system components involved that are being used to develop another AI system (Cobbe, Veale & Singh, 2023). According to the AI Act the final deployer can be ultimately deemed as responsible for the harmful effects of an AI system, but in case of a system that has already been deployed, the tracing process as well as bureaucracy could take enormous amounts of time and resources from the companies, which could lead to financial and reputational losses.

According to the AI Act each Member State ensures that the authorities and bodies responsible for enforcing the AI Act have adequate resourcing financially, as well as expertise and staff wise. Compliance with these requirements for each Member State might require large amounts of resources and resource allocation, and assignment (Söderlund & Larsson, 2024). The European Commission acknowledges that effective enforcement of the regulations requires a deep technical and multifaceted understanding of AI, placing the accountability of compliance with the AI Act upon each Member State individually (Regulation (EU) 2024/1689, Article 11). There are concerns about how prepared each Member State is fulfil all the demands. The outcome of this is likely to be that Member States will have varying levels of capacity, for example, conducting audits of AI systems and enforcing other aspects of the AI Act. Moreover, according to Kilian, Jäck & Ebel (2025) and interviews conducted with companies and start-ups, significant amount of deployers find the deadline of August 2026 for compliance, to be unrealistic due to the uncertainty of how to comply with the regulations (Kilian, Jäck & Ebel, 2025). So not only each Member State will likely

face resourcing challenges, but also the deployer companies who are developing high-risk AI systems due to the ambitious deadline set by the European Commission. Additionally, some countries might not have the technical or structural readiness, or human resources to conduct efficient AI system audits and that could lead to large inconsistencies between the Member States and their enforcement effectiveness (Demková & De Giorgio, 2025). However, along with the AI Act a scientific panel of AI experts was formed, to provide support for the AI Office as well as the national authorities of each Member State and other stakeholders that are affected by the regulations (Söderlund & Larsson, 2024). The deadline set by the European Commission might not be realistic, because there are too many interpretations and implementations to be made with a matter that has no earlier real-world cases to be used as guiding examples.

In addition to the fast-approaching deadline for high-risk systems, requirements and mechanisms for ensuring compliance with the regulations after an AI system has been deployed might lack efficiency. The AI Act does not provide enough details for establishing a template for AI system providers to use post-market, and the AI Act states that the details will be included in the plan in February 2026 (Regulation (EU) 2024/1689, Article 72). This creates a gap that concerns deployers that have already entered the market with their AI system and are required to do post-market monitoring but are unable due to the lack of proper directions of what should be included in the post-market monitoring system plan (Mökander et al., 2022). This also means that with the approaching deadline of August 2026 for high-risk AI systems to comply, deployers will get additional information and guidance for the templates in February 2026, which leaves them with about six months of time to adjust and adapt.

The EU AI Act does not address the so-called black box problem directly and does not mention it in the official regulation document explicitly. Due to the increasing complexity of self-training AI models, it becomes difficult to be able to conform with the EU AI Act's regulations for high-risk AI systems (Pavlidis, 2024). Moreover Shumailov et al. (2024) explains that the previous AI models like GPT is being trained by other models (Shumailov et al., 2024), which complicates the black-box phenomenon even further. This is problematic in the context of the AI Act, because the transparency and explainability that is strived to be achieved would require to simplify the AI systems solution variables too much to achieve human comprehension of what is happening inside the so-called black box. This means that to fully comply with the regulations, AI system providers must simplify the algorithms to achieve the level of transparency that would maintain the desired AI systems explainability. (Pavlidis, 2024).

Lastly, the EU AI Act's focal point seems to stay more on the organizational level even though it is supposed to promote safe and trustworthy AI system development and deployment for Europeans. There are no mechanisms in place in case of an individual coming face to face with discrimination or unfair outcomes caused by the unfairness of an AI system. For example, the regulations are meant to prevent situations where the use of AI system in the employment sector leads to discrimination, but in case individual actually faces discrimination due to the AI system making a biased decision, it is unclear how the individual should and could proceed with the matter. If this type of situation occurs, is the company responsible for dealing with it, or is it the official authorities of each Member State? The ambition of the AI Act is to ensure safe and trustworthy AI for Europeans, but there also seems to be limited amount of consideration for certain aspects and consequences of AI technologies in the long run. More consideration should be put into the effects of these technologies on a societal and psychological level in the long run, as well as what subtle changes with a risk of snowballing effect AI technologies bring to individuals' day to day life on moral and physical levels.

7 Conclusions

This thesis examined the nature of existing ethical challenges in AI system development and deployment, and how the new European Union Artificial Intelligence Act (Regulation (EU) 2024/1689) addresses these ethical challenges, as well as what ethical challenges and gaps persist despite the newly developed regulations. The motivation of this research was to further the discussion surrounding the new EU AI Act and contribute to expanding the understanding and awareness of what ethical challenges and gaps must be addressed to advance the implementation of EU AI Act in its formative stage. The approach of this research was to examine the EU AI Act from the standpoint of fairness, accountability, transparency and ethics (FATE) framework, and identify what gaps and challenges persist based off these selected ethical principles.

The most important key findings concerning the main research question "What ethical challenges and gaps persist in the European Union Artificial Intelligence Act?" were identified in this thesis through the lenses of fairness, accountability, transparency and ethics (FATE) framework. Notable gaps and challenges identified in this thesis include challenges that stem from the difficulty of defining fairness, where AI system providers could use gerrymandering to manipulate the interpretation of fairness and indicate that the system under development is fair. Another ethical challenge and gap related to accountability in governance and enforcement of the regulations is that companies might not be equipped to foster a certain level of self-governance, creating a concern about how well the regulations will be enforced and the AI systems audited. The fast-approaching deadline for compliance has also been identified as problematic for AI system deployers, due to uncertainty of how to comply with the new regulations. Additionally, a gap related to algorithmic transparency was identified, and the so-called black box problem persists. Despite the efforts of the AI Act, the level of transparency and explainability that is required for the regulations to be effective might be unrealistic, because of complex supply chains and the overall complexity of AI models. Moreover, the effects of generative AI on individual's lives and society have not been given enough consideration, because despite problematic cases with generative AI, generative AI remains outside of the high-risk classification and therefore is subject only to transparency requirements.

The EU AI Act is however, currently in its formative stage, because of the novelty of the regulations, and the legal and ethical practices concerning the new regulations are still forming. There are no real-world cases that can be used as example case studies where the EU AI Act was implemented. This thesis also highlights the tentative state of the enforcement of the EU AI Act,

because the legal and ethical practices concerning the new regulations are still forming and there are no real-world cases that can be used as concrete examples of the AI Act in motion. In other words, the interpretations of the EU AI Act are constantly evolving as regulators, institutions legal bodies and other stakeholders engage with the regulation. The conclusions and interpretations drawn from this thesis should be considered open to revision due to the evolving nature of interpretations and enforcement of the AI Act in real-world cases.

References

Agarwal, A., & Agarwal, H. (2024). A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. Ai and Ethics (Online), 4(2), 299–314. https://doi.org/10.1007/s43681-023-00266-9, retrieved 21.9.2025.

Allen, D., & Weyl, E. G. (2024). The Real Dangers of Generative AI. Journal of Democracy, 35(1), 147–162. https://doi.org/10.1353/jod.2024.a915355, retrieved 20.10.2025.

Almada, M., & Radu, A. (2024). The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy. German Law Journal, 25(4), 646–663. https://doi.org/10.1017/glj.2023.108, retrieved 22.5.2025.

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. Journal of Public Administration Research and Theory, 33(1), 153–169. https://doi.org/10.1093/jopart/muac007, retrieved 8.3.2025.

Aimen, T. (2025). Cognitive freedom and legal accountability: Rethinking the EU AI act's theoretical approach to manipulative AI as unacceptable risk. Cambridge Forum on AI Law and Governance, 1, Article e20. https://doi.org/10.1017/cfl.2025.4, retrieved 25.10.2025.

Avinash Manure, S. B. (2023). Introduction to Responsible AI - Implement Ethical AI Using Python (1st ed.). Apress, an imprint of Springer Nature. https://doi.org/10.1007/978-1-4842-9982-1, retrieved 8.3.2025.

Bahar Memarian & Tenzin Doleck (2023) Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. Computers and education. Artificial intelligence. 5100152-, retrieved 2.6.2025.

Barrios, M.M.Z. (2025) 'AI Audits: How do you implement the EU AI Act?', Trilateral Research. https://trilateralresearch.com/artificial-intelligence/ai-audits-how-do-you-implement-the-eu-ai-act, retrieved 20.10.2025.

Barocas, S., Hardt, M. & Narayanan, A. 2019. Fairness and Machine Learning: Limitations and Opportunities. 1-24. https://fairmlbook.org/pdf/fairmlbook.pdf, retrieved 13.4.2025.

B. Unver, M., & Roddeck, L. (2024). Ethics Governance of AI for the Legal Sector: Building Up a Holistic Policy Approach. Journal of AI Law and Regulation, 1(2), 177–198. https://doi.org/10.21552/aire/2024/2/5, retrieved 21.9.2025.

Cantero Gamito, M., & Marsden, C. T. (2024). Artificial intelligence co-regulation? The role of standards in the EU AI Act. International Journal of Law and Information Technology, 32(1), Article eaae011. https://doi.org/10.1093/ijlit/eaae011, retrieved 20.10.2025.

Capraro, V., Lentsch, A., Acemoglu, D., Akgun, S., Akhmedova, A., Bilancini, E., Bonnefon, J.-F., Brañas-Garza, P., Butera, L., Douglas, K. M., Everett, J. A., Gigerenzer, G., Greenhow, C., Hashimoto, D. A., Holt-Lunstad, J., Jetten, J., Johnson, S., Kunz, W. H., Longoni, C., … Viale, R. (2024). The impact of generative artificial intelligence on socioeconomic inequalities and policy making., retrieved 12.5.2025.

Chiao, V. (2019). Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice. International Journal of Law in Context, 15(2), 126–139. https://doi.org/10.1017/S1744552319000077, retrieved 12.5.2025.

Chu, C. H., Donato-Woodger, S., Khan, S. S., Nyrup, R., Leslie, K., Lyn, A., Shi, T., Bianchi, A., Rahimi, S. A., & Grenier, A. (2023). Age-related bias and artificial intelligence: a scoping review. Humanities & Social Sciences Communications, 10(1), 510–517. https://doi.org/10.1057/s41599-023-01999-y, retrieved 1.3.2025.

Clarke, V., & Braun, V. (2017). Thematic analysis. The Journal of Positive Psychology, 12(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613, retrieved 10.6.2025.

Cobbe, J., Veale, M., & Singh, J. (2023). Understanding accountability in algorithmic supply chains. Proceedings of the 2019 2nd International Conference on Control and Robot Technology, 1186–1197. https://doi.org/10.1145/3593013.3594073, retrieved 22.5.2025.

Collina, L., Sayaadi, M., & Provitera, M. (2023). Critical Issues About A.I. Accountability Answered. California Management Review Insights. https://cmr.berkeley.edu/2023/11/critical-issues-about-a-i-accountability-answered/, retrieved 8.3.2025.

Damen, V., Wiersma, M., Aydin, G., & van Haasteren, R. (2025). Explainable AI for EU AI Act compliance audits. MAB ('s-Gravenhage. Online), 99(4), 231–242. https://doi.org/10.5117/mab.99.150303, retrieved 20.10.2025.

Demková, S., De Giorgio, G. (2025). The Looming Enforcement Crisis in European Digital Policy. Verfassungsblog. https://verfassungsblog.de/the-looming-enforcement-crisis-ai-dsa-eu/, retrieved 20.10.2025.

Eberbach, E. (2005). Selected aspects of the calculus of self-modifiable algorithms theory. In F. Fiala, S. G. Akl, & W. W. Koczkodaj (Eds.), Advances in Computing and Information – ICCI 1990 - International Conference on Computing and Information, Proceedings (Vol. 468, pp. 34–43). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-53504-7_59

Elad, B. (2025). AI job loss statistics 2025: Who's losing, who's hiring, etc. SQ Magazine. https://sqmagazine.co.uk/ai-job-loss-statistics/, retrieved 22.10.2025.

Espina-Romero, L., Noroño Sánchez, J. G., Gutiérrez Hurtado, H., Dworaczek Conde, H., Solier Castro, Y., Cervera Cajo, L. E., & Rio Corredoira, J. (2023). Which Industrial Sectors Are Affected by Artificial Intelligence? A Bibliometric Analysis of Trends and Perspectives. Sustainability, 15(16), Article 12176. https://doi.org/10.3390/su151612176, retrieved 16.6.2025.

Ethics of AI MOOC. (2025). What is accountability? https://ethics-of-ai.mooc.fi/chapter-3/2-what-is-accountability, retrieved 8.3.2025.

European Commission. (2025). Governance and enforcement of the AI Act. https://digital-strategy.ec.europa.eu/en/policies/ai-act-governance-and-enforcement, retrieved 20.10.2025.

European Commission. (n.d). Regulatory framework proposal on AI. Digital Strategy. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, retrieved 2.4.2025.

European Court of Auditors. (2024). Special report 08/2024: EU Artificial intelligence ambition – Stronger governance and increased, more focused investment essential going forward. Publications Office of the European Union. https://www.eca.europa.eu/en/publications/sr-2024-08, retrieved 20.10.2025.

European Parliament. (2025, February 19) EU AI Act: First Regulation of Artificial Intelligence. https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, retrieved 2.4.2025.

Foka, A., Griffin, G., Ortiz Pablo, D., Rajkowska, P., & Badri, S. (2025). Tracing the bias loop: AI, cultural heritage and bias-mitigating in practice. AI & Society. https://doi.org/10.1007/s00146-025-02349-z, retrieved 21.9.2025.

Fox, S., & Rey, V. F. (2024). A Cognitive Load Theory (CLT) Analysis of Machine Learning Explainability, Transparency, Interpretability, and Shared Interpretability. Machine Learning and Knowledge Extraction, 6(3), 1494–1509. https://doi.org/10.3390/make6030071, retrieved 21.9.2025.

Gavriil, E., & Pavlidis, G. (2025). The Fog of Information: the EU AI Act and legal strategies against AI-fuelled disinformation. European Journal of Privacy Law & Technologies, 1–12. https://doi.org/10.57230/ejplt252EGGP, retrieved 22.10.2025.

Gleisner, S. (2020). Keeping the public's trust and confidence. Public Sector (Wellington), 43(3), 9–10.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines (Dordrecht), 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8, retrieved 27.2.2025.

Haresamudram, K., Larsson, S., & Heintz, F. (2023). Three Levels of AI Transparency. Computer (Long Beach, Calif.), 56(2), 93–100. https://doi.org/10.1109/MC.2022.3213181, retrieved 21.9.2025.

Hartmann, D., de Pereira, J. R. L., Streitbörger, C., & 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 (Online). https://doi.org/10.1007/s43681-024-00595-3, retrieved 30.5.2025.

Hirsjärvi, S., Remes, P., & Sajavaara, P. (2013). Tutki ja kirjoita (15th–17th ed.).
Tammi.

Holst, L., Lämmermann, L., Mayer, V., Urbach, N. and Wendt, D. (2024) 'The Impact of the EU AI Act's Transparency Requirements on AI Innovation', Wirtschaftsinformatik 2024 Proceedings, 92. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1091&context=wi2024, retrieved 20.10.2025.

Hopkins, A., Struckman, I., Klyman, K., & Silbey, S. S. (2025). Recourse, Repair, Reparation, & Prevention: A Stakeholder Analysis of AI Supply Chains. ACMF AccT 2025 - Proceedings of the 2025 ACM Conference on Fairness, Accountability,and Transparency, 209–227. https://doi.org/10.1145/3715275.3732017, retrieved 20.10.225.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv.Org. https://doi.org/10.48550/arxiv.2311.05232, retrieved 22.10.2025.

Hyler, S.E. (2025). The trial of ChatGPT: What psychiatrists need to know about AI, suicide, and the law. Psychiatric Times. https://www.psychiatrictimes.com/view/the-trial-of-chatgpt-what-psychiatrists-need-to-know-about-ai-suicide-and-the-law, retrieved 26.10.2025.

John-Mathews, J.-M., Cardon, D., & Balagué, C. (2022). From Reality to World. A Critical Perspective on AI Fairness. Journal of Business Ethics, 178(4), 945–959. https://doi.org/10.1007/s10551-022-05055-8, retrieved 21.9.2025.

Jonnala, S., Parida, P.K. and Thomas, N.M. (2025). EU AI Act underrepresented and insufficient to address the risk and vulnerabilities of generative AI. International Journal of Business Analytics. 12(1), pp. 1–27. https://www.igi-global.com/gateway/article/388757, retrieved 22.10.2025.

Tatdanai Khomkhunsorn. (2025). Meaningful Human Control and Responsibility Gaps in AI: No Culpability Gap, but Accountability and Active Responsibility Gap. Journal of Integrative and Innovative Humanities, 5(1), 35–57.

Kilian, R., Jäck, L., & Ebel, D. (2025). European AI Standards - Technical Standardisation and Implementation Challenges under the EU AI Act. European Journal of Risk Regulation, 1–25. https://doi.org/10.1017/err.2025.10032, retrieved 20.10.2025.

Klugman, C. M. (2021). Black Boxes and Bias in AI Challenge Autonomy. American Journal of Bioethics, 21(7), 33–35. https://doi.org/10.1080/15265161.2021.1926587, retrieved 20.5.2025.

Krauss, P. (2024). What is Artificial Intelligence? In Artificial Intelligence and Brain Research (pp. 107–112). Springer Berlin / Heidelberg. https://doi.org/10.1007/978-3-662-68980-6_11, retrieved 21.9.2025.

Lacmanovic, S., & Skare, M. (2025). Artificial intelligence bias auditing – current approaches, challenges and lessons from practice. Review of Accounting & Finance, 24(3), 375–400. https://doi.org/10.1108/RAF-01-2025-0006, retrieved 20.10.2025.

Laine, J., Minkkinen, M., & Mäntymäki, M. (2024). Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. Information & Management, 61(5), Article 103969. https://doi.org/10.1016/j.im.2024.103969, retrieved 13.6.2025.

Lewicki, K., Lee, M. S. A., Cobbe, J., Singh, J., Goyal, T., Peters, A., Mueller, S., Väänänen, K., Kristensson, P. O., Schmidt, A., Williamson, J. R., & Wilson, M. L. (2023). Out of Context: Investigating the Bias and Fairness Concerns of "Artificial Intelligence as a Service." Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–17. https://doi.org/10.1145/3544548.3581463, retrieved 13.4.2025.

Light, R., & Panai, E. (2022). The Self-Synchronisation of AI Ethical Principles. Digital Society, 1(3), Article 24. https://doi.org/10.1007/s44206-022-00023-1, retrieved 21.9.2025.

L. López-Dávila, R. (2025). Ready! (or Not?): Evaluating the Compatibility of Existing International Risk Management Standards with EU AI Act Requirements. Journal of AI Law and Regulation, 2(1), 55–67. https://doi.org/10.21552/aire/2025/1/7, retrieved 20.10.2025.

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, retrieved 8.3.2025.

McNamara, A., Smith, J., Murphy-Hill, E., Garcia, A., Pasareanu, C., & Leavens, G. (2018). Does ACM's code of ethics change ethical decision making in software development?

Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 729–733. https://doi.org/10.1145/3236024.3264833, retrieved 1.6.2025.

Mökander, J., Axente, M., Casolari, F., & 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 (Dordrecht), 32(2), 241–268. https://doi.org/10.1007/s11023-021-09577-4, retrieved 2.6.2025.

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., Hicklen, R. S., Moukheiber, L., Moukheiber, D., Ma, H., & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health, 2(6), Article e0000278. https://doi.org/10.1371/journal.pdig.0000278, retrieved 21.9.2025.

Odilla, F. (2024). Unfairness in AI Anti-Corruption Tools: Main Drivers and Consequences. Minds and Machines (Dordrecht), 34(3), Article 28. https://doi.org/10.1007/s11023-024-09688-8, retrieved 26.9.2025.

Pavlidis, G. (2024). Unlocking the black box: analysing the EU artificial intelligence act's framework for explainability in AI. Law, Innovation and Technology, 16(1), 293–308. https://doi.org/10.1080/17579961.2024.2313795, retrieved 2.6.2025.

Pesch, U. (2015). Engineers and Active Responsibility. Science and Engineering Ethics, 21(4), 925–939. https://doi.org/10.1007/s11948-014-9571-7, retrieved 8.3.2025.

Pfeiffer, J., Gutschow, J., Haas, C., Möslein, F., Maspfuhl, O., Borgers, F., & Alpsancar, S. (2023). Algorithmic Fairness in AI: An Interdisciplinary View. Business & Information Systems Engineering, 65(2), 209–222. https://doi.org/10.1007/s12599-023-00787-x, retrieved 21.9.2025.

Piachaud-Moustakis, B. (2023). The EU AI Act. Pharmaceutical Technology Europe, 35(11), 8–9. https://research-ebsco-com.ezproxy.utu.fi/c/sk55le/viewer/pdf/oe5n2e3v7b?route=details, retrieved 5.4.2025.

Price, W. N. et al. (2024) 'Liability for use of artificial intelligence in medicine', in Barry Solaiman & I. Glenn Cohen (eds.) Research Handbook on Health, AI and the Law. [Online]. United Kingdom: Edward Elgar Publishing Limited. pp. 150–166.

Rachels, J. & Rachels, S. 2012. The Elements of Moral Philosophy. 7th ed. New York: McGraw-Hill Education.

Radu, C.-G. (2025). AI and Human Bias: The Future of Jobs in the Labor Market. Proceedings of the ... International Conference on Business Excellence, 19(1), 3423–3431. https://doi.org/10.2478/picbe-2025-0261, retrieved 20.10.2025.

Raja, A. K., & Zhou, J. (2023). AI Accountability: Approaches, Affecting Factors, and Challenges. Computer (Long Beach, Calif.), 56(4), 61–70. https://doi.org/10.1109/MC.2023.3238390, retrieved 8.3.2025.

Rawls, J. (2009). Theory of Justice. Harvard University Press.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence, OJ L 1689, 12.7.2024. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689, retrieved 26.4.2025.

Regulation (EU) 2016/679 of The European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), OJ L 119, 4.5.2016, p.1-88. https://gdprhub.eu/Article_9_GDPR, retrieved 5.5.2025.

Russell, S. J., & Norvig, P. (2010). Artificial Intelligence: A modern approach. 3rd ed. Prentice Hall, Pearson Education.

Sadeghiani, A. (2024). Generative AI Carries Non-Democratic Biases and Stereotypes: Representation of Women, Black Individuals, Age Groups, and People with Disability in AI-Generated Images across Occupations. https://doi.org/10.48550/arxiv.2409.13869, retrieved 26.9.2025.

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, retrieved 9.2.2025.

Schmidt, J.-H., Bartsch, S. C., Adam, M., & Benlian, A. (2025). Elevating Developers' Accountability Awareness in AI Systems Development: The Role of Process and Outcome Accountability Arguments. Business & Information Systems Engineering, 67(1), Article 102383. https://doi.org/10.1007/s12599-024-00914-2, retrieved 21.9.2025.

Sharifzadeh, R. (2024). A Science and Technology Studies Challenge to Trustworthiness Criteria: Toward a More Naturalistic Approach. Philosophy of the Social Sciences, 54(6), 490–515. https://doi.org/10.1177/00483931241245931, retrieved 12.6.2025.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature (London), 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y, retrieved 21.9.2025.

Simmler, M. (2024). Responsibility gap or responsibility shift? The attribution of criminal responsibility in human-machine interaction. Information, Communication & Society, 27(6), 1142–1162. https://doi.org/10.1080/1369118X.2023.2239895, retrieved 8.3.2025.

Smallman, M. (2022). Multi Scale Ethics—Why We Need to Consider the Ethics of AI in Healthcare at Different Scales. Science and Engineering Ethics, 28(6), Article 63. https://doi.org/10.1007/s11948-022-00396-z, retrieved 21.9.2025.

Sun, T., Zhao, K., & Chen, M. (2024). Human-AI Interaction: Human Behavior Routineness Shapes AI Performance. IEEE Transactions on Knowledge and Data Engineering, 36(12), 8476–8487. https://doi.org/10.1109/TKDE.2024.3480317, retrieved 21.9.2025.

Söderlund, K., & Larsson, S. (2024). Enforcement Design Patterns in EU Law: An Analysis of the AI Act. Digital Society : Ethics, Socio-Legal and Governance of Digital Technology, 3(2). https://doi.org/10.1007/s44206-024-00129-8, retrieved 27.5.2025.

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: key problems and solutions. AI & Society, 37(1), 215–230. https://doi.org/10.1007/s00146-021-01154-8, retrieved 8.3.2025.

van Bekkum, M. (2025). Using sensitive data to de-bias AI systems: Article 10(5) of the EU AI act. Computer Law & Security Review, 56, 106115-. https://doi.org/10.1016/j.clsr.2025.106115, retrieved 12.5.2025.

Wagner, A., Bartneck, C., Lütge, C., & Welsh, S. (2020). What Is Ethics? In An Introduction to Ethics in Robotics and AI. Springer International Publishing AG.

Westerstrand, S. (2025). Fairness in AI systems development: EU AI Act compliance and beyond. Information and Software Technology, 187, Article 107864. https://doi.org/10.1016/j.infsof.2025.107864, retrieved 20.10.2025.

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 195–200. https://doi.org/10.1145/3306618.3314289, retrieved 21.9.2025.

Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. Big Data & Society, 10(1). https://doi.org/10.1177/20539517231177620, retrieved 8.3.2025.

Yaprak, B. (2024). Generative Artificial Intelligence in Marketing: The Invisible Danger of AI Hallucinations. Journal of Economy Business and Management, 8(2), 133–158. https://doi.org/10.7596/jebm.1588897, retrieved 25.10.2025.

Zhang, X., Wei, X., Ou, C. X. J., Caron, E., Zhu, H., & Xiong, H. (2022). From Human-AI Confrontation to Human-AI Symbiosis in Society 5.0: Transformation Challenges and Mechanisms. IT Professional, 24(3), 43–51. https://doi.org/10.1109/MITP.2022.3175512, retrieved 16.6.2025.

Zhong, H. (2024). Implementation of the EU AI act calls for interdisciplinary governance. The AI Magazine, 45(3), 333–337. https://doi.org/10.1002/aaai.12183, retrieved 22.5.2025

Appendices

Appendix 1 Explanation of the use of AI

AI-assisted tools like ChatGPT (GPT-4o and GPT-4o mini) and Perplexity AI (Sonar (Llama 3.3 70B), R1-1776 and limited Pro models) were used for identification of some of the relevant literature related to this research. All articles were subsequently located and verified through the university library database before inclusion in the literature pool and usage in this thesis.