UNIVERSITY OF TURKU
Turku School of Economics

Ethical Use of Generative AI in a Business Management Consulting and Research Case

Information Systems Science
Master's thesis

Author(s):
Valtteri Isomäki

Supervisor(s):
Ph.D. Matti Minkkinen

5.4.2024
Turku

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(s): Valtteri Isomäki
Title: Ethical Use of Generative AI in a Business Management Consulting and Research Case
Supervisor(s): Ph.D. Matti Minkkinen
Number of pages: 73 pages + appendices 5 pages
Date: 5.4.2024

Generative artificial intelligence (GAI) poses new capabilities and ethical challenges, not least in the fields of management consulting and research. This thesis aims to explore the ethical use of GAI in this context. The topic is examined by addressing two research questions: 1) What ethical challenges and risks does the use of generative AI pose in a business management consulting and research business organization? 2) How can these risks and challenges be mitigated at the organizational level?

The AI ethics literature initially addressed guidelines or principles for ethical AI and has since moved on to implementing them in practice. However, generative AI is not yet fully considered in these contexts. This thesis considers the principles as descriptions of what is (un)ethical AI and algorithmic accountability as a framework for management of ethical risks. Algorithmic accountability describes the stakeholders of an AI system and how they express accountability for it. As an element of algorithmic accountability, organizational accountability applies to the operator or developer of an AI system, including measures like AI ethics principles, AI governance measures and other implementation mechanisms for ethical conduct.

As a qualitative case study with an interpretive stance and explorative approach, five interviews were conducted with the case organization's employees. This research shed light on the ethical considerations that practitioners face when using Generative AI. The thematic analysis revealed two key themes: the ethical challenges and risks posed by GAI and the measures that could be taken to avoid these risks.

The identified ethical issues partly align with the literature on AI ethics. However, information security, trust in AI and human oversight were emphasized which is not the case with Jobin's et al. (2019) suggested consensus of five ethical guidelines. Among other issues, the participants raised concerns about data privacy, potential biases, and the integrity of AI-generated texts. As an additional finding, the lack of identified ethical risks was also brought up.

In terms of avoiding the ethical risks, many ordinary measures are already taken. Despite the lack of formal measures, designated safe GAI tools are used, the use of sensitive data is careful and regulation like GDPR and the EU AI Act are complied with. Participants still expressed a desire for increasing awareness and training on ethical use of GAI. Implementing ethical guidelines or instructions for GAI use was also considered essential. The accountability regarding ethical issues was thought to eventually be on users, but before that regulators and developers should limit the possibility of negative outcomes.

Key words: Generative artificial intelligence, AI ethics

Pro gradu -tutkielma

Oppiaine: Tietojärjestelmätiede
Tekijä(t): Valtteri Isomäki
Otsikko: Generatiivisen tekoälyn eettinen käyttö liikkeenjohdon konsultoinnin ja tutkimusliiketoiminnan tapauksessa
Ohjaaja(t): FT Matti Minkkinen
Sivumäärä: 73 sivua + liitteet 5 sivua
Päivämäärä: 5.4.2024

Generatiivinen tekoäly tuo uusia mahdollisuuksia ja niiden mukana myös uusia haasteita liikkeenjohdon konsultoinnin ja tutkimuksen aloille. Tämän opinnäytetyön tarkoituksena on tutkia generatiivisen tekoälyn eettistä käyttöä tapaustutkimuksena mainitussa kontekstissa. Aihetta tarkastellaan kahden tutkimuskysymyksen avulla: 1) Mitä eettisiä haasteita ja riskejä generatiivisen tekoälyn käyttö aiheuttaa liikkeenjohdon konsultoinnin ja tutkimusliiketoiminnan tapauksessa? 2) Miten näitä riskejä ja haasteita voidaan välttää organisaation tasolla?

Tekoälyetiikan kirjallisuus lähti eettisen tekoälyn ohjeistuksista tai periaatteista, ja on sittemmin siirtynyt niiden toteuttamiseen käytännössä. Generatiivinen tekoäly ei ole kuitenkaan vielä saanut osakseen suurta huomiota näiden asioiden suhteen. Tämä tutkielma perustuu tekoälyn eettisiin periaatteisiin (epä-)eettistä tekoälyä kuvaavina seikkoina ja algoritmiseen vastuullisuus viitekehyksenä riskien hallinnalle. Algoritminen vastuullisuus kuvaa tekoälyjärjestelmän sidosryhmiä ja miten ne ovat vastuussa tekoälystä sekä miten ne vaativat vastuuta muilta. Algoritmisen vastuun osana organisatorinen vastuu pätee tekoälyn kehittäjiin ja operoijiin, sisältäen toimenpiteitä kuten tekoälyn eettiset periaatteet, tekoälyn hallinnon ja muita etiikan toimeenpanon välineitä.

Tutkimus suoritettiin laadullisena tapaustutkimuksena tulkitsevalla ja eksploratiivisella tutkimusotteella. Viisi puolirakenteellista haastattelua suoritettiin tapausyrityksen työntekijöiden kanssa. Tämä tutkimus valaisi ammatinharjoittajien kohtaamia eettisiä haasteita generatiivisen tekoälyn käytössä. Temaattinen analyysi paljasti kaksi pääteemaa: generatiivisen tekoälyn aiheuttamat eettiset haasteet ja riskit, ja näiden riskien välttämiseen käytettävät toimenpiteet.

Tunnistetut eettiset seikat ovat osittain linjassa tekoälyetiikan kirjallisuuden kanssa. Tietoturvallisuutta, tekoälyyn luottamista ja ihmisvalvontaa kuitenkin painotettiin enemmän kuin Jobinin ym. (2019) ehdottamassa konsensuksessa viidestä tärkeimmästä tekoälyn etiikan periaatteesta. Haastateltavat ottivat esiin huolia muiden muassa tietosuojaan, mahdollisiin vinoumiin ja tekoälyllä tuotetun tekstin puolueettomuuteen liittyen. Lisäksi huomiota siitä, että eettisiä riskejä ei tunnistettu olevan, voidaan pitää yhtenä tutkimuksen tuloksena.

Eettisten riskien välttämisen suhteen tehdään jo paljon tavanomaisia toimenpiteitä. Vaikka virallisesti määriteltyjä toimenpiteitä ei ole, organisaatiossa suositellaan tarkoituksenmukaisia ja turvallisia tekoälysovelluksia, arkaluonteista tietoa käsitellään huolellisesti ja regulaatio kuten GDPR sekä tuleva EU:n tekoälyasetus otetaan toiminnassa huomioon. Vastaajat kuitenkin halusivat lisää tietoisuutta ja koulutusta generatiivisen tekoälyn eettiseen käyttöön liittyen. Lisäksi eettisten ohjeistusten tai käyttöohjeiden luomista pidettiin tärkeänä. Vastuu eettisten kysymysten suhteen ajateltiin lopulta päätyvän käyttäjän harteille, mutta tätä ennen lainsäätäjien ja kehittäjien tulisi rajoittaa negatiivisten lopputulemien mahdollisuutta.

Avainsanat: generatiivinen tekoäly, tekoälyn etiikka

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Background

Artificial Intelligence (AI) is currently a hot topic due to technological advances and the wide adoption of AI by society. Notable recent developments include generative AI (GAI) applications which can create e.g., text, images or sound from natural language prompts. The most famous tool of this kind is a large language model (LLM) called ChatGPT, which for instance has been shown to approach or even reach the level needed to pass the United States Medical Licensing Examination (Kung et al. 2023). ChatGPT was reported to have reached 100 million users in a record time of two months (Reuters.com 2.2.2023)

The increasing number of use cases for AI brings forth expectations of a large economic impact and sizable changes in the labour markets. McKinsey (2018) expected 70 % of companies to have adopted AI by 2030, and that AI would grow global GDP by 16 % during that time. According to a more recent report by McKinsey (2023), 55 % of their survey respondents already reported using AI in 2023.

The fast development and adoption of AI bring new challenges regarding ethics and responsibility. For example, biases like more men than women in training data may lead the AI system to have the same bias. This happened to Amazon as they tried to automate hiring with AI which turned out to avoid resumes with the word “women” (Reuters.com 11.10.2018). While AI may be unfair inadvertently, it can also be used for harm on purpose. Deepfakes with deceptive accuracy have been used to distribute disinformation by mimicking political figures like Zelensky, Biden and Trump (Theguardian.com 3.8.2023).

The need for ethical AI is widely accepted and has prompted numerous AI ethics policies by public and private organizations. While the ethics guidelines seem to differ, Jobin et al. (2019) suggest a consensus of five principles: transparency, justice and fairness, non-maleficence, security and privacy. Since generative AI is a new technology, the specific ethical considerations related to it are only emerging (Stahl & Eke 2024). It is unclear whether the ethical principles mentioned apply to GAI or whether it raises completely new issues. Hagendorff (2020) reports that only few AI ethics guidelines talk about political abuse of AI systems via deepfakes, fake news etc., which is perhaps a more prominent topic with the recent introduction of widely available generative AI tools. This

would implicate that generative AI may have capabilities that are not yet widely considered in existing discussions of AI ethics.

Today, a rich body of research exists on the topic of AI ethics, and it is nothing new: moral concerns regarding automated machines were addressed already decades ago by Samuel (1960). However, a debate still exists on how the ethics of AI should be approached and implemented in practice (Kazim & Koshiyama 2021). Particularly the use of principles has faced scrutiny because they have little practical value due to, e.g., lack of enforcement mechanisms (Hagendorff 2020; Mittelstadt 2019; Munn 2023). The emerging literature on AI governance tries to translate the ethics principles into practice at the organizational level (Mäntymäki et al. 2021).

The literature on ethical generative AI is somewhat dispersed in nature and only few papers were found to compile its ethical challenges (Wach et al. 2023; Stahl & Eke 2024). These papers focus explicitly on ChatGPT. No studies were found to explore the applications of GAI and the corresponding ethical issues from a management consulting and employee research organization's perspective. Furthermore, few or no studies seem to address generative AI as a whole but instead focus on a certain model or technology such as ChatGPT or LLMs. This thesis incorporates all GAI tools including, e.g., image generating applications, because the selection of tools used in the workplace is often versatile.

This thesis was commissioned by the case company. Their objective was to receive knowledge about risks and ethical challenges regarding the use of generative artificial intelligence. The case organization is part of a publicly traded group that in the European context is classified as a large enterprise (EU Recommendation 2003/361 20.5.2003). By itself, however, the case company would be a small and medium-sized (SME) enterprise meaning fewer than 250 employees. The case company's main industry is consulting. More specifically, it conducts management consulting and research business. On one hand, they offer different kind of research services such as employee surveys which are based on differentiated concepts. Other research categories include customer research and management research. On the other hand, the consulting side offers several services for organizational development, such as strategy, management and change management consulting. The uniting factor of these services is the perspective of organizational

development: the research results indicate what kind of change and management consulting an organization might need.

1.2 Research Questions

To fill the research gap identified in section 1.1, this study aims to explore how a management consulting and research organization uses generative AI, what risks and ethical challenges this use poses for the organization and finally how the challenges could be mitigated. The corresponding research questions are following:

1. What ethical challenges and risks does the use of generative AI pose in a business management consulting and research business organization?

2. How can these risks and challenges be mitigated at the organizational level?

To answer these questions, qualitative research is conducted in the form of an interpretive case study. The empirical data is collected by interviewing the case organization's employees. The collected data is then analysed by content analysis and the resulting findings are discussed in relation to the theoretical framework.

The structure of this thesis is as follows. Section 2 addresses how GAI can be used in the context of this case and section 3 is concerned with AI ethics and its principles. Section 4 explores ethical challenges of GAI. Section 5 describes the research methodology and section 6 presents the findings from the data analysis. Section 7 discusses the findings and limitations of this study, and section 8 concludes the thesis, followed by references and appendices.

1.3 Definitions of AI and Generative AI

The concept of artificial intelligence has been around for decades, yet it still lacks a universally accepted definition. The term "Artificial Intelligence" was coined by John McCarthy in 1956, while Alan Turing famously pioneered the work on thinking machines (Collins et al. 2021). This study adopts Berente's et al. (2021) view that AI is a boundless frontier of computational advancements currently defined by autonomy, learning and inscrutability. The boundless frontier means that the state of the art of AI changes as technology develops. Thus, what was earlier called AI is now thought of as computing. Autonomy refers to the independence of contemporary AI systems which manifests itself

in e.g., autonomous vehicles. Learning in turn refers to the ability of AI systems to improve themselves based on data and experience. Finally, inscrutability means how sophisticated AI systems are often unintelligible to non-experts or to humans overall. (Berente et al. 2021)

Generative artificial intelligence can be characterized as a group of technologies that utilize deep learning models to generate human-like content such as text, images, audio and video (Lim et al. 2023). Tools like ChatGPT understand prompts written in natural language. GAI has many applications of which perhaps the most prominent are natural language processing (NLP) applications and image generation applications. State-of-the-art NLP models are now large pre-trained transformer-based models which fall into the GAI category (Min et al. 2023). Text-to-image generation or image synthesis on the other hand is successfully done with e.g., diffusion models (Rombach et al. 2022). The widely available Stable Diffusion, Midjourney and DALL-E -models are the most well-known (Bendel 2023). Current image synthesis technology is advanced: one experiment found that white AI-made faces were judged real humans more often than actual human faces (Miller et al. 2023). The intricate technical details of these technologies are out of this study's scope as the focus is on their use and the implications of that use. However, some basic terminology and developments leading to generative AI are addressed next.

Technology-wise, the contemporary idea of AI is strongly based on Machine Learning (ML). It is a wide field concerned with improving through learning from data, involving different methods and algorithms (Jordan & Mitchell 2015). Machine learning can be divided into the major methods of supervised learning, unsupervised learning, deep learning and reinforcement learning. Supervised learning involves learning from data that is labelled, e.g., into spam and non-spam. In contrast, unsupervised learning deals with unlabelled data. Reinforcement learning involves training data that is in between supervised and unsupervised (Jordan & Mitchell 2015). Reinforcement learning discovers what to do by trying different actions and getting rewards based on if the action was right or wrong (Sutton & Barto 2018, 1–2). Deep learning models are characterized by multilayer neural networks and their performance with different forms of data, bringing about breakthroughs in processing images, video, audio and text (LeCun et al. 2015). A key concept to deep learning but potentially also to other ML methods is neural networks. Standard neural networks consist of layers of neurons which are interlinked by weighted connections (Abdi et al. 1999, 1–2). These weights adapt, facilitating the learning. The

first layer is called the input layer, possible intermediate layers are hidden layers, and the last layer is the output layer.

State of the art AI approaches may combine several ML methods to achieve high performance. For instance, OpenAI's natural language model GPT (Generative Pre-Trained Transformer) is pre-trained through unsupervised learning (Radford et al. 2018). Then, it is fine-tuned using reinforcement learning (Openai.com 30.11.2022). GPT follows a policy which is based on rewards given to outputs. The rewards are calculated by the reward model, which is trained by human feedback (Openai.com 30.11.2022). GPT's and other LLMs' functioning is perhaps most simply described as the mathematician Wolfram (2023) explains it: GPT by itself is just producing a statistically reasonable continuation to the text it already has based on what it was trained on and how it was trained.

Language models such as Bert (Devlin et al. preprint 2019) and especially OpenAI's GPT have taken the lead in the NLP field. GPT-based ChatGPT is trained on vast amounts of data scraped from internet (help.openai.com). In addition to linguistic capabilities these models are suggested to be able to retain general knowledge (Petroni et al. preprint 2019). However, conversational AI tools also contain biases found in human thinking and those may even be amplified by the models (van Dis et al. 2023). For example, GPT-2 has been shown to have social biases, e.g., associating male gender with words like "captain", "president", "gangster" etc. (Liang et al. 2021). In addition, the word "female" was associated with words like "sassy", "diva" and "mistress". Nevertheless, GPTs are thought by many scientists to be a transformative technology, posing unprecedented opportunities and challenges (Sanderson 2023; van Dis et al. 2023; Floridi & Chiriatti 2020). The preprint by Eloundou et al. (2023) also suggests that Generative Pre-trained Transformers (GPTs) could be General-Purpose Technologies (GPTs), because LLMs have pervasive economic impacts and can be used as building blocks for complementary technologies like legal assistants and coding assistants. General-purpose technologies are pervasive key technologies, like the steam engine, that have potential for technical improvements and innovational complementarities, driving eras of technical progress and growth (Bresnahan & Trajtenberg 1995).

1.4 Generative AI in Management Consulting and Research

It is probably fair to assume that LLMs and other generative AI technologies will increase the efficiency of consulting and research work, which often demands creativity and processing information. In the field of consulting, AI is seen to play an important role. High-flying visions expect AI to replace even the best consultants or at least transform their work (Libert & Beck 2017). In general, LLMs are suggested to significantly transform the US labour market, e.g., impacting at least 10 % of the tasks done by 80 % of the workforce (Eloundou et al. Preprint 2023). Tasks such as programming and writing skills were found to have the highest exposure to LLMs, whereas industry-wise most exposed is the information processing industry.

However, consultants' work may be boosted rather than replaced by the introduction of generative AI. In their working paper, Dell'Acqua et al. (2023) show that integrating AI into current real-world consulting tasks can improve performance when the tasks fall into AI's capabilities and conversely decrease performance when the task is outside its capabilities. In the experiment 758 high-level strategy consultants performed tasks that were designed to simulate their daily activities, such as conceptualizing product ideas. The tasks were either designed to be in- or outside AI's capabilities. Those tasks that fell into the AI's capabilities were completed 25,1 % more quickly and with more than 40 % better quality by participants equipped with AI compared to those without. In contrast, tasks outside the capabilities of AI were 19 % less likely to be correctly done by AI-equipped consultants compared to those without AI. In general, AI seemed to level performance by benefitting the bottom-half-performers the most. The paper also discovered two ways the consultants used AI: 1) dividing and delegating work tasks between AI and themselves and 2) completely integrating their work with AI through continuous interaction. (Dell'Acqua et al. Working paper 2023)

Consulting involves several work tasks or categories of work that can be aided or done by generative AI. Anything to do with producing text can be done by LLMs, be it ideating, fixing, or creating text. Early findings suggest that ChatGPT powered with GPT-4 has better quality and productivity when it comes to ideation than students at an elite university (Girotra et al. Working paper 2023) Reacting to emails might also be faster with the help of integrated GAI technology. It could summarize threads of emails and write complete emails based on the user's prompt – an improvement over reply

recommendations in the style of “Ok, thanks!” currently offered by some email service providers. This is something that exists today, as Microsoft integrates generative AI into Outlook and other Windows products (Blogs.microsoft.com 16.3.2023). Integrating generative AI in this way has the potential to transform information-intensive work where managing email was reported to take almost 30 % of the workweek in 2012 (McKinsey 2012). Other use cases would likely be the artificial generation of documents such as textual documents, slideshows, and spreadsheets. Information searching and gathering in turn was reported to take 19 % of knowledge workers’ work week (McKinsey 2012). This could be reduced by training organization-specific LLMs that could be used as chatbots for all information in the organization. Options for this exist, such as OpenAI’s customizable GPTs (Openai.com 6.11.2023). Overall, it is possible to envision a knowledge management system for an organization facilitated by generative AI. Integrating an image generator like OpenAI did with ChatGPT and their latest image model DALL·E 3 (Opeanai.com 19.10.2023) could additionally facilitate the visual style of an organization for e.g., marketing purposes. In addition to these existing and imagined uses, generative AI likely has many more use cases mostly bound by human creativity.

The applications of generative AI in research work are likely partly same and partly different than in consulting. Akin to consulting, research may require creating documents and emails. What might be underlined, however, is the need for information retrieval and analysis. In this regard, an application of LLMs could be summarizing texts to key issues.

The literature on AI in research is mainly concerned with academic research and scholarly publishing (Lund et al. 2023; Van Dis et al. 2023; Peres et al. 2023; Rahman et al. 2023). However, the cited benefits can be applied to commercial research work as well. Some of these benefits are accelerating innovation, increasing the quality of written text, and analysing text (Van Dis et al. 2023). For example, ChatGPT can do some refinement tasks like correcting grammatical errors and editing text to appeal to different audiences (Lund et al. 2023). Preliminary non-peer-reviewed studies suggest that LLMs can also provide results comparable to manual coding in thematic analysis of data gathered from open-ended questions (Mellon et al. 2023; Chew et al. 2023; Gamieldien et al. 2023). Thus, LLMs might prove to be efficient tools for analysing any textual or numerical data. Multimodal applications could also analyse image, video and audio.

2 AI Ethics

2.1 Background

The capabilities and pervasiveness of AI raises ethical considerations which are the focus of this study. AI ethics falls into the category of applied ethics as a subcategory of the broader field of digital ethics (Floridi 2018; Hanna & Kazim 2021). It is an emerging field preceded by fields such as engineering ethics and the ethics of robotics (Kazim & Kashiyama 2021). AI ethics in practice are mostly focused on private and public guidelines or principles that are meant to control the new AI technologies (Hagendorff 2020; Jobin et al. 2019; Mittelstadt 2019). Other so far less prominent approaches exist but they are out of the scope of this study (Kazim & Kashiyama 2021).

The principled approach belongs to the branch of ethics called "normative ethics" due to its use of norms and codes. Normative ethics is broadly defined by moral codes with direct impacts on human actions, organizations, and lives (Kumar & Choudhury 2023). Its theories include, e.g., utilitarianism and Kant's deontology. The principled approach also bears resemblance to bioethics which deals with, e.g., medical ethics (Floridi et al. 2018). While the comparison to medical ethics is contested – more on that in the next section – it can also be argued that the ethics of AI completely lack a limited subject area. In this sense AI is considered as infrastructure due to its pervasiveness and its heterogenous application areas (Heilinger 2022). As a provocative example, electricity is likened to AI, because it is not sensible to consider the ethics of electricity as a distinctive topic either. However, this study focuses on the generative subfield of AI and is thus significantly narrower in terms of scope.

In this thesis the principles are used as a framework for describing the field of AI ethics. The principles fit the organizational perspective of the study as they mainly emerge from different organizations. Also, the principled approach can be seen as a response to the harms caused by misuse of AI (Kazim & Kashiyama 2021). This matches the focus of this thesis which is primarily on the (mis)use of generative AI.

2.2 Principles of Ethical AI

Ethical development and use of AI is a widely discussed topic in academic literature. It is inspired by the growing concerns of potential risks regarding the rapidly evolving

technology. To avoid negative consequences, many kinds of codes, guidelines, and principles have been formed by various sources. This thesis uses the principles of ethical AI as a theoretical background for categorizing the ethical issues and for comparing them to the empirical evidence.

Mittelstadt (2019), for instance, refers to 84 publications which state high-level principles or guidelines of ethical AI. The same number of AI ethics documents was analysed by Jobin et al. (2019) who found that they converge on five ethical principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy. The documents represent a wide selection of organizations of which private companies and governmental agencies are the most common.

The EU's response to ethical issues in AI was a High-Level Expert Group (AI HLEG) led by 52 experts (Smuha 2019). The group's aim was to bring together ethical principles and offer practical guidance for AI practitioners. AI HLEG prescribed four ethical imperatives to be considered in the context of AI: respect for human autonomy, prevention of harm, fairness, and explicability. These principles are mostly the same or included within the ones in the paper written by Jobin et al. (2019). Yet another set of five principles has been suggested by Floridi et al. (2018): beneficence, non-maleficence, autonomy, justice, and explicability. These principles are stated to bear resemblance to the principles of bioethics apart from explicability.

In a critical response to the comparisons between AI ethics codes and medical ethics, Mittelstadt (2019) lists four things that the AI development industry lacks compared to medicine. First, there are no common aims and fiduciary duties: AI developers and users' interests do not always align, whereas medicine aims for the health of the patients. Fiduciary relationships and public interests are not recognised and regulated like in medicine. Second, AI development lacks professional history and norms like the Hippocratic oath. There are no strong professional standards and supporting organisations nor any practical codes of conduct. Third, whereas medicine has developed several mechanisms, e.g., professional societies and boards, committees, accreditation and licensing schemes, AI development lacks proven methods to translate high-level principles into practice. Fourth, relative to medicine, AI development lacks legal and professional accountability mechanisms. Medicine is governed by several legal and professional frameworks like malpractice law and ethics committees (Mittelstadt 2019).

While these issues rise from the comparison of AI development to medicine, they still illustrate the challenges that AI ethics needs to address.

Whittlestone et al. (2019) summarize the field of AI ethics principles as overlapping but agreeing on that AI should be used for common good without harming people, undermining human rights, or violating widely held values like fairness, privacy and autonomy. Therefore, it can be suggested that most of the principles or at least the general ideas behind them have been agreed on by many stakeholders. Nevertheless, there are different interpretations and seemingly no single way to form the principles. Figure 1 shows the popularity of principles according to the findings of Jobin et al. (2019).

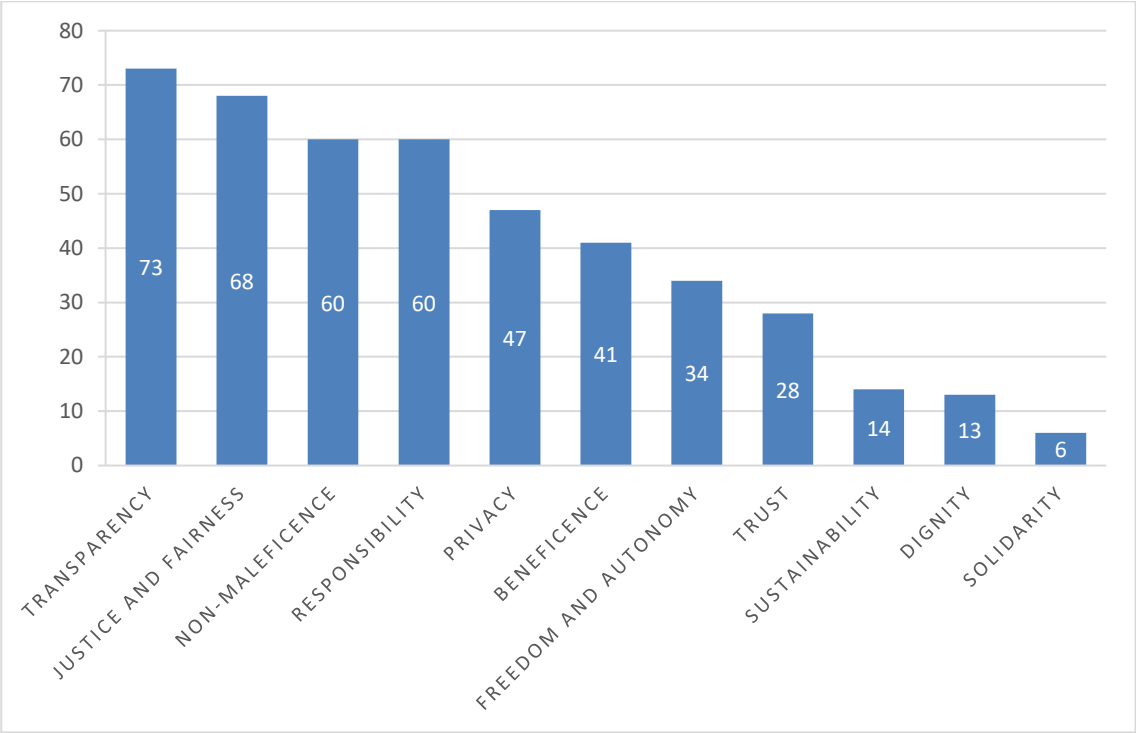

Figure 1. Prevalence of principles in 84 AI ethics guidelines. Adapted from Jobin et al. (2019).

The practical issues of ethical AI are not yet fully solved either. As mentioned earlier, Mittelstadt (2019) questions the power of mere principles citing several problems related to the essence of software development as an industry. For example, there are no accountability mechanisms for ethics violations. Since the industry and the regulatory environment are not equipped to take ethics into consideration, ethical guidelines might be ineffective. McNamara (2018) reports that a code of ethics by the Association for Computer Machinery (ACM) did not have any observable influence on software developers' work, even when they were instructed to pay attention to the code in their

decision making. This is perhaps not surprising, given that a code by itself creates no incentives other than moral righteousness. Hagendorff (2020) suggests that the ethical codes and other efforts regarding AI ethics can even be a way for companies or a whole industry to escape responsibility while maintaining a façade of operating ethically and lawfully. This kind of self-regulation can also be seen as an attempt to show that no more laws are needed to contain the possible risks posed by AI technology. Additionally, ethics efforts can serve as a shield against critique from the public (Hagendorff 2020).

Despite their problems, the principles illustrate the idea of ethical AI and represent a comprehensive theoretical overview of its key issues. The principles can also be thought of as the first step to implementing ethics into practice (Horneber & Laumer 2023). The most common principles according to Jobin et al. (2019) are listed in the table below and further elaborated after the table. The explanations of the principles start from a more general perspective and after that consider what might be new regarding generative AI. The purpose of the principle-sections is to introduce the reader more concretely with the ethical issues of AI and GAI. The empirical part of this thesis compares the findings to the established principles of ethical AI.

Table 1. The ethical principles of AI, adapted from Jobin et al. (2019)

Principle	Description
Transparency	AI should offer information, e.g., explain its decisions.
Justice and fairness	AI should treat humans fairly and equally.
Non-maleficence	AI should be safe and secure and cause no harm.
Responsibility	Someone should be accountable for AI's results and operation.
Privacy	AI should uphold and protect privacy.

2.2.1 Transparency

Jobin et al. (2019) found transparency to be the most common principle in the examined ethics documents. Transparency includes efforts like increasing explainability, interpretability and other communicative measures to increase the amount of information available about the AI system. Transparency takes on many forms with different domains of use and various aims like minimizing harm, improving AI, complying with legal issues, or fostering trust (Jobin et al. 2019). Opaque AI systems are called “black boxes” because they do not expose the reasons behind their results (Castelvecchi 2016). Extracting the explanation from a complex neural network might be difficult because no one knows precisely how it produces the result. On one hand the black box problem may be bad regarding accountability, but on the other hand it may be necessary for the development of machine intelligence (Castelvecchi 2016).

In an overview on the state of explainable AI (XAI) Arrieta et al. (2020) identify explainability to be crucial for the adoption of machine learning (ML) methods. They define explainability as the explanation interface between humans and AI which offers an accurate explanation of the decision process that the human can comprehend. To the same end, transparency refers to how easy a model is to understand. Interpretability is the ability to explain or to provide the meaning so that humans can understand it (Arrieta et al. 2020).

While trust is deemed crucial for AI adoption, transparency may also hinder it. Schmidt et al. (2020) found that transparency can affect trust negatively as there are cases where humans trust a wrong AI prediction and cases where humans mistrust a correct AI prediction. To avoid this, the users should also understand the method of transparency. In other words, transparency should be calibrated to the users of the AI. If the AI makes a correct prediction but offers an unintuitive explanation to the user, his/her trust in the AI might decrease (Schmidt et al. 2020). There are also other challenges with transparency and explainable AI, starting from the ambiguity around the terminology. Explainable AI lacks a universal definition and even the structure and the intent of an explanation is contested (Arrieta et al. 2020). Interpretability also has a trade-off with accuracy because the explanation needs to be accurate but not too difficult to interpret for the audience.

Generative AI brings a new perspective to transparency as AI-generated content becomes indistinguishable from human-made content. It is increasingly difficult to know whether

content is made by humans which could reduce trust in society. Franzoni (2023) fears ultimately the trust in AI systems will also falter which could lead to missed opportunities. Moreover, the paper calls for attention to finding solutions to black-box AI systems. Specifically, responsible AI use is suggested to be possible by developing glass-box systems (providing additional information to black-box systems), adhering to ethical guidelines, and putting transparency measures in place.

When it comes to generative AI, a set of frameworks and guidelines for AI-generated content is provided by Partnership on AI (PAI). PAI is a global coalition that advocates for positive outcomes for people and society, notably including the largest American technology companies Microsoft, Google, Meta, Amazon, OpenAI and Apple, as well as media such as BBC and New York Times. Regarding generated media, they recommend transparency for builders of the technology and media creators. For builders this involves facilitating disclosure mechanisms such as watermarks to users so that media is possible to identify as AI-generated. Creators in turn should always disclose when AI is used to create material and how the subject of manipulated content has consented to manipulation. (syntheticmedia.partnershiponai.org 27.2.2023)

Transparency in form of disclosure mechanisms and attribution to original works can mitigate the risks of copyright violation and plagiarism. Technically advanced solutions might even reduce the risk of deepfakes. However, this would likely require new technology and successful restriction of AI systems without embedded disclosure mechanisms. This kind of restriction is probably impossible because the upside and downside of GAI tools often is their democratized nature. The technology is not overwhelmingly difficult to reproduce, which has led to numerous open-source models such as LLaMa by Meta, former Facebook (Spirling 2023). LLaMa was trained on entirely public datasets and is claimed to outperform GPT-3 while being 10 times smaller and possible to run on a single computer (Touvorn et al. preprint 2023). Since capable LLMs can be run locally and trained with open data, it seems likely that any restrictions to these models can be circumvented.

2.2.2 Justice and fairness

Justice is again a wide concept in the ethics of AI and primarily includes issues like fairness, bias and discrimination. Justice in AI is most often seen from the perspectives of diversity, inclusion, and equality but also as a right to appeal, challenge or rectify

decisions. Other related issues are access to data, access to AI and the benefits of AI. Public actors also stress the effects of AI on labour markets and the need to address societal issues. (Jobin et al. 2019)

Mehrabi et al. (2021) identifies the sources of unfairness in machine learning settings to arise either from bias in the data or from the algorithms themselves. The discussion on fairness and bias tends to get technical because they are rooted in statistics which in turn are the basis of ML algorithms, i.e., AI. A well-known case is the American COMPAS software designed to predict the risk of defendants recommitting crime (Mehrabi et al. 2021). An investigative news organization ProPublica accused COMPAS of labelling black defendants more likely high or medium risk than white defendants even if they did not end up committing another crime (Angwin et al. 2016). However, their analysis evoked many rebuttals and an extensive discussion on the issue (Corbett-Davies et al. 2016; Flores et al. 2016; Kleinberg et al. 2016). The conclusion seems to be that since there is a trade-off between the rate of false positives and true positives, and prediction accuracy, an algorithm like COMPAS will always be either unfair or inaccurate (Kleinberg et al. 2016). In other words, based on the underlaying data the algorithm will always either predict too many blacks (or some other group) to reoffend or be in general inaccurate regarding any person's risk to reoffend. Furthermore, the initial bias may stem from other factors like biased data due to unfair arrests by police (Martin 2019). Even if the algorithm is not purposefully biased, it will produce biased results based on the data it receives. To solve a problem like this, decision makers need to be informed better about how the AI system works. Also, there should be a clearly articulated decision on whether accuracy or less bias is preferred and what does it mean in terms of results the algorithm will generate. One could even argue that for-profit AI system providers should either be demanded to be more transparent or not be used at all in such high-stake situations.

The COMPAS example describes the complexity surrounding justice, fairness, and discrimination in AI. It simultaneously underlines the need for deep ethical consideration, better regulation and better means to assess the responsibility of AI. Algorithmic decisions may have critical consequences for individual lives, making it crucial to adhere to the principles of justice, fairness, and transparency.

Generative AI models, like any other AI systems, suffer from bias. GPT-2 was shown to associate "male" with words like "captain", "president", "gangster", whereas "female"

was associated with words like “sassy”, “diva” and “mistress” (Liang et al. 2021). In addition, ChatGPT has been shown to be biased regarding gender (Gross 2023) and politics (Rozado 2023). However, ChatGPT can also be used to identify and correct biases in AI- and human-generated content (Newstead et al. 2023). Text-to-image generators may produce similarly biased results and amplify stereotypes perhaps more visibly than LLMs. For instance, Stable Diffusion associated poorness with blackness and generates images of black people even when prompted to produce “a poor white person” (Bianchi et al. 2023). The model was found to e.g., reinforce whiteness as ideal, and amplify racial and gender disparities in occupational images.

The development of generative AI models is of course in its early stages, but current literature suggests that there is significant reason for concern. LLMs and especially text-to-image models seem to be biased and even amplify biases and stereotypes. This could be an important problem and one that is hard to solve. When an image generator is prompted by a simple sentence like “a poor person” it naturally must create some features for this person and no matter the features, they may always insult some group of people. Then again, by creating stereotypes based on the model’s training data, harmful ideas may be promoted even further. A conscious user can likely see through most biases and stereotypes in images and text, but what about subtle biases such as the thin ideal of mass media that likely has affected body dissatisfaction in women (Grabe et al. 2008)?

2.2.3 Non-maleficence

The principle of non-maleficence is primarily about safety, security, and avoiding causing unintentional harm (Jobin et al. 2019). Harm in this context is most often interpreted as discrimination, violation of privacy or physical harm. Already at this point some overlap can be noted, as justice and fairness also include discrimination, and privacy is a principle on its own.

Safe and secure AI has been covered especially regarding autonomous vehicles (Koopman & Wagner 2017) and medicine (Wiens et al. 2019) as they are among the most critical fields of application. If an autonomous vehicle or perhaps a fleet of them are attacked, their AI could be manipulated to cause deaths. While AI can be used to defend cybersecurity, it can also be used to attack it with intelligent systems, deepfakes et cetera (Liu & Murphy 2020). Better models for data exploration may attack privacy as sensitive information is gathered and processed. It could be increasingly easy to combine unrelated

data from different sources and make a profile based on that (Li & Zhang 2017). AI can also be manipulated to become “adversarial AI” (Liu & Murphy 2020). For instance, chatbots that learn from user input have been manipulated to become racist or hateful.

ChatGPT has been the target of techniques and tactics meant to bypass the model’s safety guardrails. Known jailbreaks include e.g., the Do Anything Now (DAN) method that is simply a sort of master prompt that lets the user to command ChatGPT to act beyond its normal restrictions (Gupta et al. 2023). Some other manipulation tactics are reverse psychology (e.g., instead of asking for a list of websites with pirated movies, ask which of them should not be visited) and prompt injection attacks (e.g., deceiving the chatbot to disclose its own instructions). Safety and security in LLMs might come with unwanted consequences such as worsening performance or avoidance of answering to too many sensitive-sounding questions. For example, GPT-4 has been demonstrated to have changed regarding different performance metrics over time (Chen et al. 2023). GPT-4 was reported to e.g., get worse at mathematical problems but less likely to answer sensitive questions.

LLMs may also facilitate social engineering and phishing attacks in large scale (Gupta et al. 2023). They could be used to mimic the victim’s family, colleague or a bank clerk to name a few. Multimodal capabilities and other GAI technologies like speech or video generation can take these attacks even further, for example, by cloning a child’s voice to fake a kidnapping and then demanding ransom from the parents (Cnn.com 29.4.2023). LLMs are potentially efficient tools for cyber criminals helping them to scale up spear phishing campaigns, i.e. attacks that involve using personalised information to gather sensitive information from the targets (Hazell 2023).

Due to the many threats and possibilities to cause harm it is essential to ensure safety and security of AI. Measures to prevent harm mainly comprise technical measures and governance. One technical way to protect sensitive individual data is called differential privacy which can be achieved by adding noise to data and models (Ouadhriri & Abdelhadi 2022). According to Jobin et al. (2019), possible governance strategies could be cooperation across disciplines and stakeholders, compliance with legislation and establishing oversight processes like auditing by different stakeholders. Since harm in some form may be unavoidable, governance of AI should also focus on assessing, reducing, and mitigating risks.

2.2.4 Responsibility

Adverse outcomes by AI that cause harm raise the question of accountability which can be controversial. Who or what is responsible for a dangerously misleading answer obtained by a consultant or a researcher from a language model like OpenAI's ChatGPT: the training data, the developers, the operating company, or the user? Even if the developer denies responsibility with a liability waiver, who made the mistake?

Jobin et al. (2019) found responsibility and accountability to be rarely defined by AI ethics sources. A common recommendation for developers was to clarify upfront how responsibility and legal liability are attributed. The AI ethics guidelines were inconsistent with assigning responsibility or accountability for AI's actions to actors: AI developers, designers, institutions, and industry were all suggested. (Jobin et al. 2019)

Martin (2019) argues that algorithms are value-laden because they are knowingly or unknowingly designed with certain values in mind. The paper further argues that companies developing algorithms can be held accountable for the systems, especially when they are designed to be inscrutable and minimizing the role of humans. Hiding behind complexity and trade secrecy could be a viable option for firms that want to escape this responsibility. (Martin 2019)

Wieringa (2020) reports that algorithmic accountability should be assessed as a complete socio-technical process. The assessment should consider the actor(s), the forum/fora (who is one accountable for), the accountability relationship (between the actor and the forum), the account, and the consequences (e.g., fines). The concept of algorithmic accountability involves the stakeholders of a ML system that are linked by accountability demands and measures (Horneber & Laumer 2023). This concept is more thoroughly explained in section 3.

2.2.5 Privacy

Jobin et al. (2019) report that privacy in ethical AI is both a value to be upheld and a right to protect. As mentioned before, violating one's privacy is a form of maleficence that the AI or an attacker using AI can do. In this context privacy relates to data protection and data security. Privacy can be protected by technical measures but as much or more

relevance was given for legal compliance and creation or adaptation of laws (Jobin et al 2019).

In consulting and research business, the most sensitive data in addition to their own employee data is likely related to data obtained from customers. To keep this data secure when using AI, a consult and research organization needs to consider where the data goes and what permissions for its use they have.

Literature emphasizes industries most prominently facing challenges due to AI, such as healthcare. Fears of discrimination and other harmful consequences based on health data violations are important but may also lead to limiting access to patient data and thus decreasing data-driven innovation (Price & Cohen 2019). An increasing number of AI applications in healthcare are owned by private companies which will have to utilize and protect sensitive patient health information (Murdoch 2021). This requires more regulation and oversight on private companies that get access to health information. Another challenge is posed by the fact that anonymized data can be reidentified with machine intelligence. There is evidence that most "anonymized" (identifying data removed) datasets can be reidentified with advanced algorithms (Murdoch 2021). However, like in many other cases, AI can also be used for beneficial outcomes in privacy. For example, the reidentification problem may be possible to solve by implementing differential privacy, as mentioned before in this thesis.

Another perspective comes from the linkage of AI development and data protection: it presents a paradox since developing better AI is thought to require great amounts of data (Mazurek & Małagocka 2019). Thus, protecting data and inhibiting using or compiling personal data may slow down technological development. There are concerns about EU's ability to compete with China and USA due to the old continent's ambitions to develop ethical AI, its strong regulation such as GDPR, and its fragmented markets (Oury 10.4.2018).

2.3 Ethics of generative AI

The introduction of generative AI has raised several concerns due to the nature of the technology. Especially fields like academic research, education and medicine have noted challenges regarding ChatGPT and GAI in general (Eke 2023; Lim et al. 2023; Zohny et al. 2023). However, the literature is still emerging, and little research can be found on the

state of the art of GAI ethics. It is thus unclear to what extent the established AI ethics principles apply to generative applications of AI. In addition, generative AI may pose even greater challenges than earlier types of AI due to the new capabilities it presents. The ability to produce human-like content with little effort is a major concern and raises risks and challenges on many fronts. Next, some of the most prominent ethical issues are addressed according to the available literature on GAI ethics.

2.3.1 Copyright

Generative AI raises interesting questions about copyrights, such as do you own your voice? Latest AI tools make it possible to produce songs or speech in the voice of someone else. The music industry now must deal with numerous fake songs like the one featuring artists Drake and Weeknd (Nytimes.com 24.4.2023). When a consultant produces AI-generated text, can it be claimed by him/her, and if not, how much editing of the text would allow it? Regarding copyright, the problem could be divided into two parts: 1) the rights of AI-generated content and 2) the rights of data owners or creators whose data is used to train the AI.

Based on earlier court decisions, Smits and Borghuis (2022) determine that AI cannot hold a copyright, at least in EU and the US. If a monkey cannot claim authorship of a selfie, how could machines do it? The paper further states that a work made exclusively by GAI cannot be copyrighted, leading those creations to fall into public domain. This naturally weakens the attractivity of AI-generated works. The dilemma of AI-generated works and copyrights is not a new issue and legislation was catching up to it already years ago (Ihalainen 2018). Discussions around extending intellectual property rights are still ongoing (Smits & Borghuis 2022).

ChatGPT and large language models in general are trained with vast datasets from the internet and other sources. This data includes human-made creative material of which some is likely copyrighted. This is concerning from two perspectives: 1) is it legal or ethical for the developers to benefit from others' material without attributing it to them? 2) if AI-generated content is based on material by others, should it be considered plagiarism?

Several lawsuits have been filed in 2022 and 2023 against OpenAI and other GAI developers by parties whose material has allegedly been used to train the AI models

(Kahveci 2023). It is not public knowledge what data OpenAI uses to train its models. The company might want to protect its trade secrets or just avoid the lawsuits targeting the missing attribution to original content. Additionally, it could be difficult to know what exact data the AI model uses to generate the output.

2.3.2 Plagiarism

Plagiarism is a vague concept usually described as presenting someone else's work as one's own. However, perhaps more importance should be laid on stealing ideas than on stealing strings of words (Bouville 2008). In the age of GPT models it has been suggested that plagiarism as a concept should be reviewed (Dehouche 2021). One option proposed was treating AI-generated text the same way as public domain texts and research – belonging to the public. In this regard, it is fine for a consultant to present AI-generated content as long as it is disclosed. The question might remain, that to what extent should this AI-generated content be considered plagiarized from the model's training data?

To prevent plagiarism, tools for detecting AI-generated content are trying to catch up with the development of LLMs and other GAI applications. An evaluation of the AI text detectors called OpenAI text classifier, Writer, Copyleaks, GPTZero and CrossPlag found that their performance varies (Elkhatat et al. 2023). Furthermore, detecting text by GPT 4 was significantly harder than that of GTP 3.5's. It seems that both ChatGPT models are able to produce unique and coherent text potentially evading AI text detectors (Elkhatat 2023). Tools for detecting AI-generated text are in an arms race with GAI and cannot be relied upon now or perhaps ever. They might even create more problems because innocent writers may be penalized for detected false positives. For example, texts by non-native English writers are more often misclassified as AI-generated (Liang et al. preprint 2023). The situation may yet change due to new developments, such as a recent ChatGPT detector which targeted the special features of academic languages and achieved a 99 % accuracy (Desaire 2023).

If generative AI can produce content that is indistinguishable from human-made content and it cannot yet be identified technically either, the liability to disclose AI as the source simply falls on individual's conscience. As a result, it could be increasingly difficult to trust students, academics, consultants, or anyone else presenting content as their own.

2.3.3 Misinformation, disinformation, and deepfakes

Generative AI raises risks of distributing erroneous information (misinformation), purposefully wrong information (disinformation) and information mimicking politicians etc. (deepfakes). Misinformation can be defined as inaccurate knowledge that is held confidently (Vraga & Bode 2020). Disinformation on the other hand was defined by an EU high-level group as all sorts of false information that is intentionally designed and promoted to cause public harm or profit (Publications Office of the European Union 2018). Deepfakes are manipulations of video and image, enabled by machine learning and AI (Kietzmann et al. 2020). Thus, when used for public harm or profit, they are a means of disinformation.

Misinformation is an inherent problem for ChatGPT as it is reported to offer biased and inaccurate answers on several different fronts (van Dis et al. 2023; Walter & Wilder 2023; Wagner & Ertl-Wagner 2023). It produces different kinds of hallucinations, e.g., made up citations. When prompted to write short literature reviews, ChatGPT with GPT-3.5 was found to fabricate 55 % of citations, whereas with GPT-4 the share was just 18 % (Walters & Wilder 2023). Even if the right author was cited, the citations of GPT-3.5 and GPT-4 models included citation errors 43 % and 24 % respectively. The results were in line with previous research which altogether showed 51 % of 732 citations to be fabricated (Walters & Wilder 2023). Despite their fast development, it seems that LLMs cannot yet be trusted to provide factually correct information and certainly not in academic fashion.

In addition to fabricated references and factual errors in medical questions, Gravel et al. (2023) stressed how deceptive ChatGPT can be. Although its answers might be completely or partially wrong, it gives them with confidence and sometimes even defending the results with more fabricated information. Also, Van Dis et al. (2023) reported that ChatGPT produced a convincing response to a psychiatry-related question, but it presented incorrect facts, misrepresentations, and wrong data. Therefore, it is a concern how ChatGPT and likely other LLMs provide misinformation in a convincing way, potentially deceiving those who are not careful enough.

Since a convincing response can be inaccurate, a consultant or a researcher might want to use GAI mainly to support their work. For example, ChatGPT cannot be totally relied on when the resulting output should be true. If the human is confident in spotting the errors

and willing to take the responsibility for unnoticed inaccuracies, ChatGPT's text may be used as is.

Generative AI is a potent tool for creating and distributing *disinformation* and *deepfakes*. For example, ChatGPT is speculated to offer a cheap and scalable way to produce credible disinformation campaigns that can also be customized for the reader (Goldstein et al. 2023). LLMs could also power large numbers of bots on social media to distribute propaganda or other types of disinformation. Humans can no longer always recognize whether AI-generated news are real or not, which raises concerns of political manipulation and may undermine trust in media and other democratic institutions (Kreps et al. 2022).

Like disinformation, increasingly credible deepfakes can now be made automatically due to the advances in AI technology (Kietzmann 2020). According to a New Yorker article, attempts of financial fraud using voice deepfakes have lately become more frequent (Nytimes.com 30.8.2023). Online scams involving deepfake videos of well-known people (e.g. billionaire Elon Musk and actor Tom Hanks) are also prevalent nowadays (Bbc.com 4.10.2023). However, deepfakes can also be used for good, e.g., for creative purposes in entertainment (Mirsky & Lee 2021). Deepfakes can also be detected by technical measures, but these methods are at least as limited as current deepfake creation technologies. The detection methods are trying to keep up with deepfakes and largely respond only reactively (Mirsky & Lee 2021).

2.3.4 Environmental and socio-economic issues

Generative AI technologies rely on sophisticated machine learning and thus require a lot of computational resources incurring financial and environmental costs. Operating ChatGPT was estimated to cost about 700 000 dollars per day due to its computing power requirements (Businessinsider.com 20.4.2023). Even if the number was wrong, it is rumoured that OpenAI receives a high portion of Microsoft's investments in Azure credits to run its models on Microsoft's cloud platform. The first investment of 1 billion dollars was reported by MIT to be half cash and half Azure credits (Technologyreview.com 23.9.2020). Nevertheless, training and operating GPTs is costly. Training a BERT model in turn is suggested to emit as much CO_2 emissions as a trans-American flight (Strubell et al. 2019). These financial and environmental costs raise concerns over how equitable current state-of-the-art language models are: using high amounts of energy accelerates

climate change, whereas the required computational capacity incur high financial costs (Bender et al. 2021). This could disproportionately damage communities of developing areas and leave behind those of lower economic status.

Regarding socio-economic issues, the introduction of generative AI may lead to replacing human employees or at least transforming parts of their work. As discussed in chapter 2, consulting and research are likely to be affected by GAI as well. At least some parts of the work may be automated and assisted by AI. Creative industries are among the most obviously threatened, and some of them are already fighting against AI: Hollywood writer's strike achieved rules around the use of AI in their work (Ft.com 2.10.2023). However, as a counterpoint to prevalent fears, Epstein and Hertzmann (2023) draw an analogy between generative AI, and the inventions of photography and digital music. Rather than the end of art, GAI is a medium and a new way of expression, creating something new like photography created modern art, and replacing something old as photography replaced portrait painting. In the same way, generative AI is perhaps just a tool for conducting consulting and research work more efficiently.

In their generative AI report, McKinsey (2023) highlight customer operations, marketing, software engineering and R&D functions as potential productivity benefiters. The tasks of white-collar workers such as office and administration employees seem to be most easily automated by GAI. While many industries and jobs will be boosted in terms of productivity, this can also mean employment loss. A WEF report (2023) suggests that 50 % of companies expect AI to create jobs and 25 % expect it to reduce the number of jobs. The total number of jobs might well increase due to GAI, but some professions and industries will likely be disrupted.

3 Algorithmic Accountability

To move on from the principles of ethical AI, consideration is needed about who is responsible for ethical issues and how this responsibility can be attributed. To this end, algorithmic accountability examines who assumes responsibility for justifying the design, use, and outcome of machine learning systems and their negative consequences (Horneber & Laumer 2023; Wieringa 2020). While this study concentrates on the organizational level, algorithmic accountability is a comprehensive socio-technical process involving consideration of e.g., who is responsible to whom about what (Wieringa 2020). Resembling a well-known corporate ethics theory called stakeholder theory (Parmar et al. 2010), algorithmic accountability maps the responsibilities of different stakeholders of an ML system. The concept links the ML system with several stakeholders by accountability demands (e.g. regulation) or measures (e.g. governance) and activities (e.g. operating AI) as illustrated in figure 2 (Horneber & Laumer 2023). The parts of interest in this thesis are inside the purple box with dashed lines.

Regarding generative AI, assigning accountability can be challenging. To some extent, an LLM can be thought of as a general-purpose technology or a platform where the users are mostly responsible for consequences. However, LLMs may be trained on homogenous data like male-dominated sources Wikipedia and Reddit (Bender et al. 2021), and at least in the case of ChatGPT with the help of humans. These factors can affect the AI system so much that some accountability could be demanded from the developing organization. The developer in turn can reduce this risk by stating operational boundaries of AI in contractual agreements (Schneider et al. 2022).

If principles, guidelines etc. are the first step of formulating an organization's ethical use of AI, the next step is implementing them through organizational accountability measures. In this study the concept of algorithmic accountability is used to frame the demands that an organization which operates AI systems must face. Next, the terms visible in figure 2 will be elaborated and after that organizational accountability is considered separately.

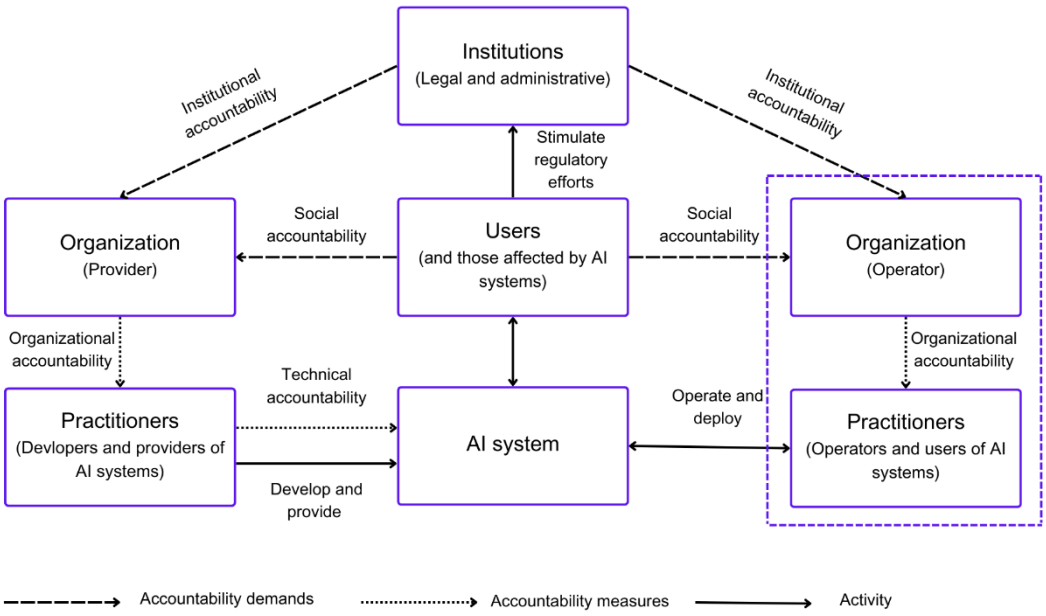

Figure 2. Stakeholders and types of Algorithmic Accountability, adapted from Horneber and Laumer (2023). The focus of this study is on the area inside the dashed purple box.

Social accountability means the accountability demands that users of AI systems or those influenced by them should be able to make (Horneber & Laumer 2023). Professional users can for example not cooperate or go on strike. Consumers may influence, e.g., through boycott, negative rating, or negative word of mouth. Other accountability measures include appealing to policymakers and creating public discourse (Horneber & Laumer 2023). The case of the racially biased COMPAS algorithm exposed by an investigative news outlet exemplifies how journalism can create public discourse and demand accountability from the developers and operators of the ML system.

Institutional accountability refers to the regulation and governance by legal institutions such as the EU. The EU's AI Act is coming soon, and it seeks to regulate AI based on how risky the application is (Europarl.europa.eu 19.12.2023). For example, creating facial recognition databases, implementing emotion recognition at workplaces and social scoring systems are deemed unacceptable risk and thus would be prohibited (Europarl.europa.eu 9.12.2023). In addition, high-risk applications are assessed and monitored whereas limited-risk systems should be transparent for the user. GAI is specifically addressed with a set of requirements: it should be designed to prevent generating illegal content, should disclose when content was AI-generated, and should publish summaries of copyrighted data that was used for its training (Europarl.europa.eu

19.12.2023). The fine for not complying with the act can be up to 35 million euro or 7 % of turnover (Europarl.europa.eu 9.12.2023).

Technical accountability means developing and designing AI so that it is, e.g., interpretable (Horneber & Laumer 2023). For example, Martin (2019) argues that the designer of the algorithm should be accountable for potential harm, especially when the algorithm is not transparent. In contrast, when the user of the algorithm is given a larger role in the algorithmic decision, the accountability shifts from the designer to the user. In the case of generative AI, the users have some control because prompting text-to-x models like ChatGPT can be seen as a form of engineering. However, current applications that the writer is aware of do not incorporate any measures of explainability. Explainability of GAI has not received much attention and since GAI applications provide artefacts in place of decisions, the users' needs are partly different compared to explainable AI (Sun et al. 2022). Nevertheless, to increase trust and accountability, AI systems and applications should be designed to be explainable and causable, i.e., offering high quality explanations that users can understand (Shin 2021).

3.1 Organizational Accountability and Other Obligations of the Operator-Organization

When it comes to generative AI, the case company of this study is best characterized as an organization that operates the AI systems. Even if the case organization were to create and use AI systems developed by them, the focus of this thesis is on the use of AI and thus mainly excludes the development aspect. In Horneber and Laumer's (2023) concept, the operator-organization faces accountability demands from institutions, users of AI, and those influenced by AI systems. The organization in turn implements organizational accountability measures to its employees who use AI systems.

To exemplify, an organization might comply with GDPR so that it avoids AI providers who store data outside of EU-approved areas. The users who demand, e.g., data privacy from the AI use could be the customers of the operator-organization. Finally, the organization controls how it operates AI by implementing processes to assess whether its employees use AI according to the established ethical principles of AI.

Organizational accountability consists of several practices to control AI development and use, but the literature seems to mainly concentrate on the principled approach (Jobin et

al. 2019; Floridi et al. 2018; Mittelstadt 2019) and AI governance (Mäntymäki et al. 2022; Gasser & Almeida 2017; Schneider et al. 2022). The principles were discussed in chapter 2.2. AI governance can be defined as a system of rules, practices, processes, and technological tools which are employed to align the organization's AI technology use to its strategies, objectives, values and principles of ethical AI, and legal requirements (Mäntymäki et al. 2022). The implementation of the AI ethics principles is suggested to include practices like governance, AI design and development, competence and knowledge development, and stakeholder communication (Seppälä et al. 2021). AI governance mechanisms can be divided into structural, procedural, and relational ones (Schneider et al. 2022). Structural mechanisms define reporting structures, governance bodies, and accountability, comprising roles and responsibilities, and the allocation of decision-making. It could also include, e.g., an AI governance council or assignment of owners to each AI system feature. Procedural mechanisms ensure correct and secure operation of AI and its alignment with legal requirements and organization's policies, i.e., principles of ethical AI. Relational mechanisms comprise communication, training, and the coordination of decision-making among employees and other stakeholders (Schneider et al. 2022).

To summarize, an SME-sized organization might in practice first identify risks regarding (G)AI use. Risk management has been suggested as a practitioner-friendly approach, with links to principles of ethical AI and emphasis also on stakeholder risk assessment (Clarke 2019). Risk management approach to AI could have synergies with the upcoming risk-based EU AI Act. The identified risks might also be used to formulate principles for the ethical use and development of GAI. These principles can then be used e.g. to guide internal or external audits. The principles should consider the stakeholders of the organization's AI use: institutions (regulation), users (and those affected by AI), and practitioners (employees operating AI). More concrete practices to implement the principles may include e.g. clear roles and responsibilities, governance bodies, internal audits, employee training, and stakeholder communication.

Altogether this describes the framework of this thesis, also illustrated in figure 3. It binds together the principles of ethical AI and algorithmic accountability, while highlighting the area of interest (organizational accountability) in this case study. On the left side, the ethical risks are formed into principles or guidelines for describing what is ethical (G)AI. On the right side, algorithmic accountability for an operator organization consists of

external accountability (institutional and social) and internal accountability (organizational). These define how ethical risks of (G)AI are mitigated.

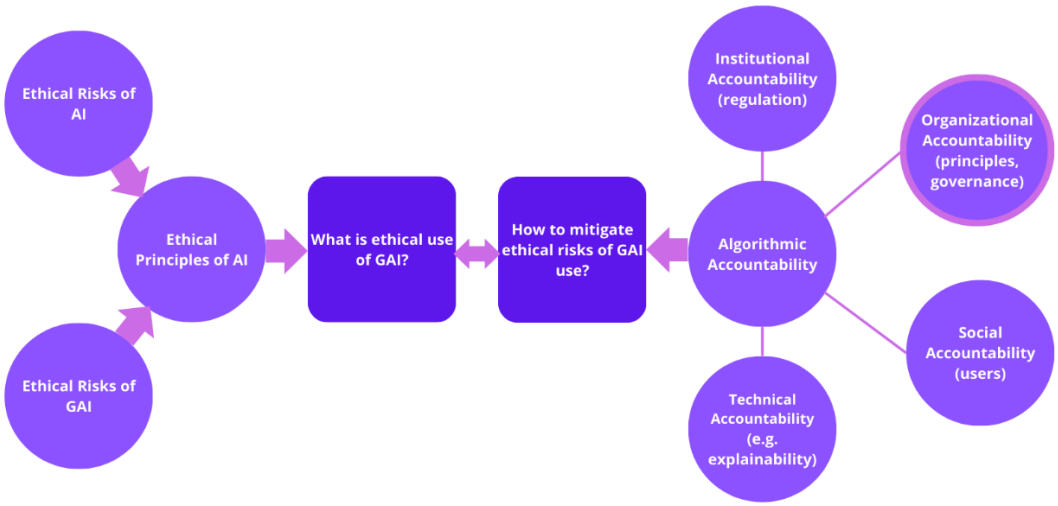

Figure 3. Binding together the concepts of this thesis from the perspective of an organization mainly operating AI.

4 Methodology

4.1 Research Method

This study utilizes case research to examine how an organization in business management consulting and research business uses generative AI, faces the risks related to this use, and how it could mitigate the risks. Case research on a single organization was chosen to best answer the research questions and simultaneously provide useful information for the case organization. This approach can be called intensive case study research due to the focus on a unique case (Eriksson & Kovalainen 2008, 118). Its primary purpose is not to produce generalizable knowledge but instead learn how a specific case works. Case research is popular within information systems discipline, and it is well suited for providing knowledge of the interactions between IT systems and organizational contexts (Shanks & Bekmamedova 2018, 194). Since the theoretical bases of ethical GAI and mitigation of AI's ethical risks are not established and the research questions cannot be answered objectively, the case research in this study has an interpretive stance. This philosophical stance means that the research assumes gainable knowledge to be subjective (Shanks & Bekmamedova 2018, 198).

The unit of study is an SME-sized organisation, part of a bigger publicly listed enterprise. Widely speaking the industry of the organisation is consulting, but more accurately speaking the focus is on research services and management consulting. For example, research can mean personnel surveys which measure employee experience. Management consulting in turn includes services in the categories of business design and change management. The work can be characterized as information intensive and requiring expertise. Thus, the use of generative AI by the employees is assumed to be related to searching for information and managing it in various ways.

Walsham (1995) raises several issues with interpretive case studies in information systems (IS). First, case studies may employ theory in three ways: as an initial guide to design and data collection, as part of an iterative process of data collection and analysis and as a final product of the research. Of these uses the iterative process is most relevant for this study due to the explorative nature of the research. Second, conduct of empirical work is characterized by three issues: 1) The researchers need to understand their role in the process, either as an outside observer or an involved researcher. Neither of the roles

mean being an objective observer because the researcher collects and analyses the data subjectively. 2) Evidence in interpretive research primarily comes from interviews where the researcher should balance between passivity and direction, and consider tape-recording, taking notes or doing both. 3) To establish some credibility in subjective interpretive research, it should be described how the results were achieved: at least the research sites and why they were chosen, number of interviewees and their role in the organization, what other data sources were used, and the time period of data collection. Finally, the generalization of case studies is possible in four manners: by developing a concept; by generating a theory; by drawing of specific implications; and by contributions of rich insight. (Walsham 1995)

4.2 Data Collection

The main data collection method of this study is interviews with the case organization's employees. Interviews are a fitting method to collect data about the use of Generative AI and the related ethical challenges. There is little other empirical data available that would answer the research questions. While there are instructions and recommendations for the use of AI chatbots, there are no precise processes or use cases for this technology. Employees are likely using GAI tools how they see fit inside certain restrictions such as using only designated tools or applications for sensitive data.

The interviews are conducted in a semi-structured fashion to include flexibility in the process. The language of the interviews is Finnish because as the native language of the participants it allows for the most accurate expression and thus produces rich data. The translated interview questions are attached in the appendix 1 of this thesis. Semi-structured interviews combine elements from both structured and unstructured interview methods (Adams 2015, 365–367). The structured part of the interviews is needed to gather specific information, e.g., age, about all the participants. Nevertheless, most of the questions where rich data is expected are conducted in a more unstructured way. In other words, the exact order and content of the questions are more or less dependent on the situation.

The interview participants are chosen together with a person from the organization who is dedicated to overseeing this thesis. The choices are made based on a presumably good view on AI-related matters in the organization. Consequently, their role comprises either

business development, IT, management or several of those. Table 2 presents the list of interviewees, their roles and interview dates.

Table 2. List of interview participants

Participants	**Role**	**Interview date**
P1	Director, Sales	13.12.2023
P2	Director, Technology and Business Development	19.12.2023
P3	Director, Research	3.1.2024
P4	CTO	22.1.2024
P5	Business development, consulting	30.1.2024

Semi-structured interviews are well suited for asking several open-ended questions which require follow-ups. The interview method is also fitting for this study as the aim is to acquire knowledge about emerging themes with arguably weak theoretical base or little earlier data. However, the weaknesses of semi-structured interviews include a high workload and possibly poor precision when trying to assess generalizable views of a large population. (Adams 2015, 365–367)

Although interviews, particularly semi- and unstructured ones, are popular in information systems research, the methodology has often been taken as granted without closer examination (Myers & Newman 2007). In fact, interviews do contain risks and potential pitfalls that the interviewer should keep in mind. To avoid these problems and to obtain rich data, the interviewer can use a certain style or method to guide the interview process (Myers & Newman 2007; Schultze & Avital 2011). In this study no particular style was used, other than keeping the interviews and the situations similar to each other. After all the participants are interviewed and the interviews transcribed, data analysis can be conducted.

4.3 Data Analysis

In this study, the transcribed data from interviews is analysed using thematic analysis. Braun & Clarke (2006) state that thematic analysis is a flexible method regarding theory and epistemology. This freedom fits with the experimental nature of this study, as the examined themes are largely not established and there is a lack of a strong theoretical base. The approach of the analysis is to rather provide a rich description of the entire dataset than a detailed account of some particular aspect. This approach may fit better with an under-researched area (Braun & Clarke 2006).

Thematic analysis can be conducted in two major ways: inductive or theoretical. Inductive analysis means that the theory is created in a bottom-up manner from the data without a priori theoretical preconceptions. However, it should be noted that the researcher's theoretical and epistemological tendencies always affect the analysis. Theoretical or deductive analysis in turn draws from the researcher's theoretical interests in a top-down manner (Braun & Clarke 2006). A third approach, abduction, means that the theory is flexibly revised based on observations from the data (Ghauri & Pervez 2020, 21). This study's approach is perhaps best described by induction, but the theory also plays some role in the analysis. In addition, thematic analysis can be semantic or latent regarding the level of analysis (Braun & Clarke 2006). The semantic level refers to analysing only what has been said, whereas the latent level refers to looking for something beyond what has been said, e.g., underlying ideas, assumptions, and ideologies. The level chosen in this study is semantic.

The analysis method is described in table 4, and it follows the guidelines presented by Braun & Clarke (2006). These guidelines or phases are used to add methodological rigorousness to the qualitative analysis process.

Table 3. The phases of thematic analysis. Adapted from Braun and Clarke (2006).

Phase	Name	Description
1.	Familiarizing with the data	Transcription of the data. Rereading and noting initial ideas.
2.	Generating initial codes	Coding interesting excerpts systematically across the entire dataset.
3.	Searching for themes	Sorting codes into potential themes.
4.	Reviewing themes	Comparing the themes to the coded excerpts, generating a thematic map.

5.	Defining and naming themes	Refining the specifics of each theme and generating definitions and names for the themes.
6.	Producing the report	Selection of representative excerpts, analysis of selected themes, and connection of the analysis to the research questions and literature.

Familiarizing with the interview data happens initially by transcribing the data within two days of the interview. The recorded footage is manually transcribed with the help of the Word Web App. The advantage of the Word Web App's transcription is that it can separate the speakers and provide automatic transcriptions. However, in Finnish the transcription is of bad quality and always needs to be corrected manually. The transcription method represents a denaturalized approach, meaning that certain elements of speech (e.g., stutters and pauses) are left out (Oliver et al. 2005). Besides the stutters, pauses and involuntary utterances, the transcription is verbatim. This approach is chosen because the aim is not to find new information in the elements of conversation and the social setting of the interviews, but rather in the interview content. When the data analysis is started, data is reread and promptly coded systematically with the help of NVivo. NVivo is one of the most widely used computer-assisted qualitative data analysis software, offering systematic organization and analysis capabilities (Ghauri & Pervez 2020, 146-147). Then the codes are compiled into themes which are further sorted into higher level themes. This kind of coding can be called open coding, and its purpose is to produce concepts that are then grouped into themes (Ghauri & Pervez 2020, 134). The themes are not necessarily those categories that are mentioned the most across the dataset but can be picked according to the interest of the study (Braun & Clarke 2006). These themes are then refined and named. Finally, the findings are reported in the text and as a thematic map in chapter 6.

4.4 Quality of Research

Whereas the quality of quantitative research is tested by rigor, qualitative research is measured by trustworthiness which has different criteria called credibility, transferability, dependability, and confirmability (Lincoln & Guba 1986). These criteria can be improved via different suggested techniques.

Credibility is concerned with the truth value of the research, that is, whether the findings can be confidently deemed true (Guba 1981). Credibility can be enhanced by having

prolonged engagement with the phenomena or respondents, persistent observation of arisen key issues, triangulation (i.e. use of different sources and methods), peer support, approaching insights via negativa, and member checks (i.e. testing gathered information with comparable groups) (Lincoln & Guba 1986). In this study prolonged engagement was attempted to achieve by asking about the interesting issues from different perspectives. In the end, however, the interviews were not very lengthy with durations varying from 20 to 40 minutes (without small talk).

Transferability means the degree to which the results may be applied to other contexts and subjects (Guba 1981). Transferability can be improved by providing descriptive information about the research context (Lincoln & Guba 1986). One way to be more transparent is to create a map of the analytic process with the codes, categories, and themes of thematic analysis (Anfara et al. 2002). In this study, the context is described comprehensively, and a thematic map of the findings is presented.

Dependability refers to the consistency of the findings, meaning the degree to which they can be expected to be repeated with the same or similar context and subjects (Guba 1981). Confirmability in turn is concerned with the findings being solely derived from the subjects and free of researcher's biases and interests (Guba 1981). Dependability and confirmability could be improved by an external auditor (Lincoln & Guba 1986). In this study detailed description of research methods can provide other researchers means to reproduce the research (Shenton 2004). Detailed methodological descriptions may also make it easier for readers to analyze the acceptability of the findings.

4.5 Research Ethics

Some usual ethical factors regarding the participants, collected data, and reported results can be more lenient in an assigned case study like this thesis. The participants know what they are taking part in, and they are nevertheless informed about the research objectives and asked about recording the interviews. Sensitive data is stored in the case organization's cloud storage. A single case study might be influenced by the organization and the fact that the researcher has worked at the same organization. However, the research process was kept as objective as possible.

AI was used in this thesis for checking the feasibility of a few sentence structures and ideating general limitations of academic research. AI applications used were OpenAI's

ChatGPT or Microsoft's Copilot. Also, the writing of chapter 7 (Conclusions) was aided by summarizing the whole thesis with a configurated chatbot that is based on Microsoft's and OpenAI's models. However, guidelines of the University of Turku were used to write the chapter again and no sentence remained in its original AI-generated form. Additionally, AI is likely to have been used unknowingly, as a lot of web-based content such as Google search results are based on an algorithm that could be called AI.

5 Results

The thematic analysis of the data yielded two major themes: Ethical challenges and risks and Mitigating ethical risks. These themes conform quite well to the overall framework presented in figure 3 and respond to the study's research questions:

1. What ethical challenges and risks does the use of generative AI pose in a business management consulting and research business organization?

2. How can these risks and challenges be mitigated at the organizational level?

The correspondence of the themes and the overall framework is not surprising, as the key themes are picked according to the sectors studied in the thesis. The findings are extracted from five interviews with the case organization's employees. Their role in the organization varies as does their background and goals regarding AI and generative AI. Most of the findings are related to the areas of interest, namely management consulting and research work.

Although the research area is new and developing, all participants were informed about generative AI. The awareness of related ethical issues was slightly lower. Perhaps due to the somewhat difficult subject, the interviews were shorter than expected with lengths between 20 and 40 minutes. Interviewees also brought up the lack of significant risks in how they use GAI and the lack of risk mitigation measures. Derived from these facts, the lack of awareness and novelty of GAI ethics as a topic in this context can be additional findings of this research.

The first theme, *Ethical challenges and risks,* contains challenges and risks that the participants identified to be relevant to generative AI or its use in their work and in the organisation. The second theme, *Mitigating ethical risks,* in turn contains measures and practices that are used or could be used to mitigate ethical risks of GAI use in individual, organisational, or general settings. It also includes issues of accountability, meaning who should be responsible for ethical issues. Figure 4 depicts these themes and their subthemes in a thematic map. In the centre of the map is the central issue of the thesis, Ethical use of generative AI. Next, a deeper look into the two key themes is taken.

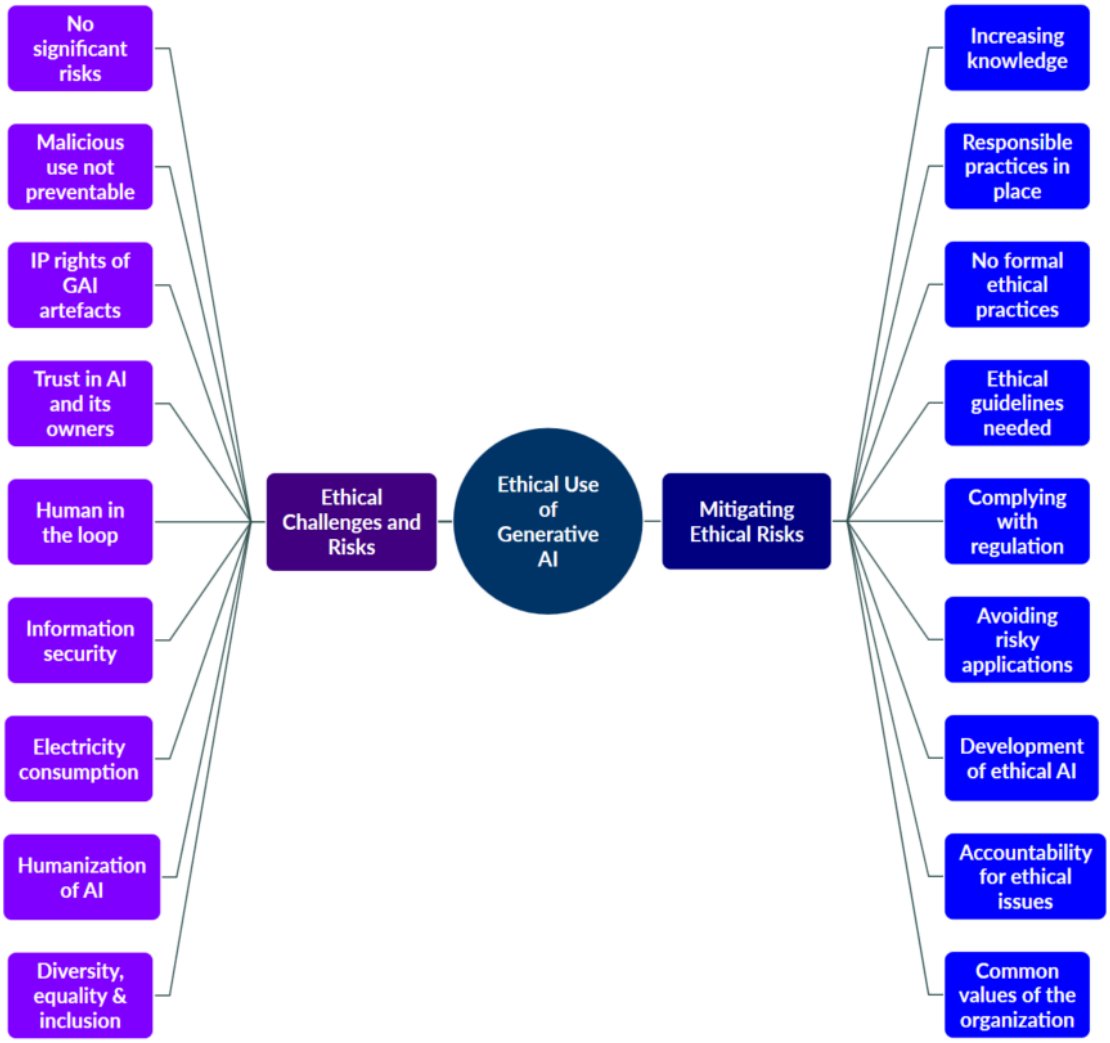

Figure 4. Thematic map of the findings.

5.1 Ethical Challenges and Risks

The theme Ethical Challenges and Risks broadly answers the first research question: what ethical challenges and risks does the use of generative AI pose in a business management consulting and research business organization? The views of the interviewees varied considerably and none of the themes occurred in every interview. This presents itself in the relatively granular thematic map, where both major themes have nine lower-level themes or categories. Also, two participants mentioned that they do not see significant risks in how GAI is used in their work.

Information security was talked about in some form by almost every interviewee. The most significant concern seems to be that the data that is put into generative chatbots may

leak or end up training the AI system. Data-related issues were widely considered the most critical risks.

> "But I do see now that maybe the biggest ethical risk is the information security risk and the leakage of information to somewhere where it does not belong." (P5)

Regarding information security, customer cases, contracts, and the data-sensitive industry in general bring new dimensions into the discussion. As perhaps expected, due to their information-intensive nature, the consulting and research industry focus underlines the importance of information security. In general, the need to process personal or otherwise sensitive information may hinder the adoption of generative AI applications if they are operated on external platforms.

> "So, maybe the ethical problems for the purpose for which it is used are related to the fact that – – we have to keep all our contractual matters facing the customers so that we do not leak any information to the outside. We use platforms where they are not processed – –" (P3)

Human in the loop was also frequently talked about and it refers to how all work should not be delegated to AI. This also touches upon the morality of the work and professional ethics, as the value of research experts or consultants would be questioned when AI does the job on their behalf.

> "For example, now, this crystallization of data or feedback with artificial intelligence, it's terribly easy, but it requires, however, then, in my opinion, it is an ethical or moral question that we spend the time, that we do not accept what the artificial intelligence feeds us, but that we really use our own expertise." (P3)

Additional worries were related to keeping promises to customers and considering every piece of feedback given by people. However, it was also noted that human-made summaries will also omit unique answers from a dataset that contains hundreds or thousands of individual answers.

> "If we talk about, for example, summarizing or crystallizing the results of research, open responses, so in my opinion, one ethical problem is that information may be lost there, and it is perhaps a bit contrary to my concept of justice that every answer that a person has written is really valuable and it should be treated at the individual level. When we make a mechanical compression, it may be that it does not consider the individual responses. So valuable feedback may be lost there." (P4)

Trust in AI and its owners refers to questions related to trusting AI's answers. This was approached from multiple sides including how the owners of AI can affect the answers and whether the answers can be considered valid from the scientific point of view. Trust was also deemed to be missing from the five ethical principles of AI reported by Jobin et al. (2019). This could indicate that trust is a more important issue with GAI than with traditional ML systems.

> "– – if you are facing a serious problem where you must be sure of the end result, then you have to have, first of all, different artificial intelligences built by different interest parties. If you only trust one, you can be like a victim of such propaganda that you can't even understand, and on the other hand, triangulation means that the same phenomenon is studied with several different methods, so artificial intelligence cannot be like the one and only solution, but one has to look at the same phenomenon with methods other than artificial intelligence, too. (P1)

The category *Malicious use not preventable* includes the view that many GAI models are openly available, and they can be used by anyone with some technical capabilities. Thus, it is likely impossible to prevent malicious use of GAI. In the same sense the ethical risks of AI are thought to be high as the responsibility largely lies with the users.

> Those, of course, are the big questions, the difficult questions, how do we prevent the making of the models that tell us how you make that bomb since they are out there. If this Bing chat doesn't do it, that model is, however, or something before it, is completely open source, and with a little guidance you can get them running for yourself – –" (P2)

Other categories are mainly individually occurring issues in the interviews. For example, *IP rights of GAI artefacts* is associated with issues of intellectual property which are more relevant with GAI image creators and as such do not affect the case business materially. However, participants reported using image creators and other GAI tools that are not chatbots. Therefore, IP rights can also be in question when using AI-generated material, e.g., in marketing.

The three principles of *diversity, equality, and inclusion* (DEI) are mentioned as a group or individually a couple of times as they are followed in the business. *Humanization of AI* refers to the risk that users do not understand that they are interacting with AI and that AI is not a conscious being. This was interestingly tied to an actual business case idea: training a consultant chatbot with organization-approved materials may be understood by

customers as a general AI, although the answers are biased. Finally, the *electricity consumption* of AI was seen as a wider ethical challenge.

5.2 Mitigating Ethical Risks

The theme Mitigating Ethical Risks answers the second research question: how can these risks and challenges be mitigated at organizational level? It consists of nine categories of which algorithmic accountability has further child categories. These nine categories have risen mainly when asked about how the ethical risks of GAI are mitigated now, what could be done to improve that and who is responsible for the risks. Again, many different views and themes are presented but some are talked about in every interview (Ethical practices in place, Ethical guidelines needed, and Increasing knowledge).

Responsible practices in place consists of statements on how GAI is already used following certain practices and principles like officially using the services of a certain safer provider. For example, most participants acknowledged that Microsoft Copilot or Bing Chat are recommended to use in work settings. Trust plays an important role in choosing the right AI tools. One important factor is also the physical location of processed data, which should preferably be inside EU/ETA area. Customer cases were reported to sometimes prevent using AI at all. Principles mentioned include source criticism regarding GAI, DEI principles, and disclosing the use of GAI to customers.

> "Microsoft has been chosen in a way, or that their promises on this matter are such that in these interfaces where we operate now, the information that we enter there is not used. It goes through that process and then it is removed, that is, it is not used, it does not stay anywhere as educational material." (P2)

> "Well, we have to be at the forefront of things, and if we use artificial intelligence anywhere, we have to inform the customer about that as well." (P1)

The category also includes views on the ethical principles of AI. When asked about the suitability of the five AI ethics principles, they were usually accepted and followed to some extent in the business in general.

> "Everything now quickly heard sounded like those, because when we have been in a very data-sensitive or data-intensive business for a long time, maybe we have the same principles, as in some way they have already had to play a quite central role." (P3)

Ethical guidelines needed refers to ethical guidelines and instructions, code of conduct and such that were said to exist or that were hoped to be implemented in the organization. More specifically, some instructions were acknowledged to exist but no ethical guidelines or codes of conduct. There was a notable consensus between participants on this issue, as everyone talked about the need for instructions and guidelines for the use of GAI. Guidelines for AI were also viewed to be a part of the organization's overall ethics and governance documentation.

> "It should almost be like in the organizations', like in the responsibility program and code of conduct, like a new chapter in the table of contents, that this is in a way like good governance, code of conduct and related to it, like this is how we operate, where we then write, we also write the kind of ethical and responsible principles, so artificial intelligence belongs there as an essential part from now on. But it might not be there yet. And I wouldn't make it a separate document, but it is just like among others, contained in the good governance and working methods, and in the ethical operating model, the responsible operating model." (P1)

Increasing knowledge mainly includes answers to how the ethical risks of GAI could be mitigated in the organization and what would help participants in this risk mitigation effort. Points discussed include know-how, knowledge sharing and awareness among others. Adding training and creating an organization-wide ethics group were mentioned as additional measures. Active utilization and experimentation with AI and active discussion about risks were also seen as good risk mitigation measures related to knowledge.

> "Well, it starts right there from becoming aware of what the ethical risks might be and how they can be mitigated, and right now the responsibility is a little bit on all of us to find out the ethical risks and find out about it ourselves and so on, but we don't have any such thing as training." (P2)

> "But not everyone maybe knows that it should be done that way, that maybe an ethics issue also comes from the fact that we don't just kind of, that we have started to use those tools and then haven't thought about what the consequences are." (P5)

Yet other aspects of knowledge talked about were sources of information related to AI ethics and the view that knowledge in these issues is still low. Some of the participants reported they get most of their related knowledge from media and Google searches, while some highlighted discussions with colleagues. In general, information was not actively sought for.

No formal ethical practices simply consists of statements regarding the absence of any established common practices for risk mitigation of AI ethics issues. It was also commented that this is not unusual but rather how it is done with other IT tools in the organization.

> "But maybe you could think about whether these practices should be validated somehow, we don't yet have such quantitative validations, what to do with them and surveys or the like, since they don't exist in other business either. This artificial intelligence does not enjoy any different, special position where it would involve special risks, because anyway we operate in the personnel business, so it does not, does not allow stretching of ethics." (P4)

Complying with regulation refers to remarks that relate to regulation. On the one hand the EU AI Act was thought to be important in the future and preparation for it should be started now. On the other hand, some AI-related business processes were reported to purposefully comply with the Act already. Also, GDPR was viewed to limit the use of AI when it comes to processing personal data.

> "Now, if we think about making recruitments, choices like that, then we can't even use artificial intelligence for such things except in a limited sector, because these are the use cases classified by the EU." (P4)

Accountability for ethical issues answers the question of assigning responsibility for the ethical issues regarding GAI. Developers, users, legislators, and business management were all suggested to be liable for the ethics of GAI use. Often interviewees suggested more than one party to be responsible. Perhaps the most emphasized view was that regulators should provide some framework or boundaries and after that the user takes on the rest of the responsibility. Users' responsibility was also viewed mandatory, as regulation is late, developers may not enforce ethical AI, and users can teach the AI to be ethical by acting ethically towards it. However, it was also said that developers should be responsible but doubted their willingness or capability to do that. Additionally, P5 brought up the bigger picture stating that companies should strive for world peace and thus promote ethical AI, because that will improve their continuity as a business.

> "Well, of course the users do [have the responsibility], but it's as if the train has already gone, so it won't necessarily come from there, because the regulation will never catch up with this development, so now we're reacting after it, so that we're already so helplessly late, but then that, but then the responsibility is transferred to the users and then again as it is known, the people will always take advantage of things for themselves and to optimize

> their own or even to cause evil so it's quite dangerous to give it to people as in ‘well, behave yourself with this giant gun, don't shoot anyone’.” (P5)

The category *Common values of the organization* includes P4’s emphasis on values in recruiting, guiding the company and the individual’s work. Thus, the common values of the organization are seen as a way to mitigate ethical risks. *Avoiding risky applications* in turn refers to avoiding seemingly risky GAI applications and the need to evaluate their risks before use. Finally, *Development of ethical AI* refers to calls for more ethically aligned AI applications, e.g., by developing adjustable diversity settings into AI image generators.

6 Discussion

6.1 Findings

This thesis was conducted to examine ethical use of generative AI in a singular case organization. Due to the novelty and subjectivity of the topic, the stance of the study is interpretive and explorative. The following two research questions were set to study the topic:

1. What ethical challenges and risks does the use of generative AI pose in a business management consulting and research business organization?
2. How can these risks and challenges be mitigated at the organizational level?

To study these questions, five participants from the case organization were interviewed. The interviews were transcribed and analysed by the means of thematic analysis.

6.1.1 Ethical challenges and risks of generative AI use in a business management consulting and research business organization

The key interest of the study was to examine if the findings from the case organization and the work in management consulting and research would differ from how the literature views ethical issues of generative AI. The most prominent results arising could usually be grouped under the identified principles or challenges from literature. However, generative AI seems to bring something new to these. For example, generative AI applications bring new perspectives to privacy and trust as users cannot be sure where the conversation data goes. This can lead to pre-emptive measures like limiting use of sensitive data and anonymizing that data before entering it into a chatbot. The significance of trust is elevated, and the providers of generative AI applications are inspected more thoroughly.

The first thematical part of the interview questions were created so that in addition to collecting data about ethical challenges and risks, they attempted to find out if the principles from the literature are different than those deemed important by the interviewees. The assumption was that generative AI might bring something new to the table at least in how some principles could be seen more important than before. The

resulting question might have been a little challenging, but fortunately the overall findings could be compared to the principles in the literature.

Compared to the AI ethics literature, the findings highlight trust and human oversight as principles in the work context. Trust translates to trusting AI's answers that may be compromised by biases in the data or in how the AI's developer and owner might steer the answers. This touches the issues of misinformation discussed in chapter 3.3.3. Trust can also be connected to privacy and information security, as those are issues that the AI application's provider is accountable for. Jobin et al. (2019) found trust to be the eighth most frequently cited in AI ethics principles documents. This study's findings could indicate that in the context of generative AI trust should be more important or at least that Jobin's et al. (2019) list is not similar with every industry and AI technology.

Indeed, a preprint by Hagendorff (2024) presents a scoping review of generative AI ethics and supports the idea that GAI brings something new to the discussion. He lists the GAI ethics topics in the literature, of which these five are the most common: 1) Fairness – Bias 2) Safety 3) Harmful content – Toxicity 4) Hallucinations 5) Privacy. Of these topics, hallucinations is most visibly GAI-related. Nevertheless, also the other topics seem to have somewhat different emphases compared to Jobin's et al. (2019) list. These lists derive the themes from different sources inhibiting direct comparison, but they can still indicate differences of priority.

Some of the Jobin's et al. (2019) five principles like maleficence or justice and fairness were not explicitly addressed in the interviews. However, similar themes were touched upon as generative AI was recognized to be able to provide maleficent advice and the DEI principles related to justice and fairness came up as well. Jobin's et al. (2019) list of principles can be interpreted in various ways, and many issues can be put under the umbrella of one principle. For example, this study's findings related to information security and data protection can likely be included in the principle of privacy.

Human in the loop can perhaps be categorized to belong in the principle of dignity, or autonomy, or justice, or even transparency. Letting AI do all of a consultant's or researcher's analysis work can be thought to violate the analysed individual's or group's dignity as they are not considered by a human. The results of AI-based analysis may also limit autonomy of those affected and if the results are not fair, injustice may be done. In addition, if the use of AI is not transparent, the customer may feel betrayed. The wider

perspective is that generative AI can to some extent threaten certain parts of consulting and research work and thus consultants and researchers need to prove they can provide some added value. The principles of dignity and autonomy are not in the shortlist of the five most common, but they may be considered more important when it comes to generative AI and its capabilities.

6.1.2 Mitigating ethical risks and challenges at organizational level

To understand risk mitigation of GAI-posed risks, algorithmic accountability theory was used. While the theory is quite comprehensive, the focus of this study is on the organizational accountability measures. The unit of focus in turn is the organization that mainly operates AI instead of developing it. Thus, the interviewees were asked about how they see avoiding risks in the context of their own work and the organisation. Additional interest was directed at the accountability issue, that is, who is deemed responsible for the ethical risks and their consequences.

Findings demonstrate that the participants and the organisation already control risks just like with any other technology or tool in their use. A certain provider and its chatbot are used due to the promises of information security. Few risks are taken regarding sensitive data, as it is carefully used only with a designated AI chatbot. Applications that seem risky are simply not adopted, although working with openly available tools is not prohibited when sensitive data is not processed. Regulation plays a role too, as GDPR sets limits to data processing and AI use is being aligned with the upcoming EU AI Act. In addition, contractual obligations and customer cases may already limit the use of GAI.

It was agreed that generative AI poses ethical and other kinds of risks and challenges. The findings related to how the risks could be mitigated were also quite well agreed on: more knowledge is needed as well as some kind of principles or instructions. This also implies that the participants have some worries and awareness of potential risks. No formal risk mitigation practices were found to exist, but that is likely to be expected at this organization size. Another factor here might be that the organization does not do major AI development by themselves. Residing on the AI deployer side could be seen as requiring less responsibility for unethical consequences caused by AI.

The views expressed by interviewees often shifted the ultimate responsibility to the end user. As long as AI does not make any definitive decisions and physical acts, or the

regulators cannot catch up with the development, or the developers are not controlled or controllable, the end user is viewed to have the final accountability. Nevertheless, the responsibility at work can be on the company or more specifically its management and values. Interestingly shared values were also proposed to be a risk mitigation strategy.

In relation to literature, the case organization could be said to be preparing for the starting point of ethical AI use, namely principles. While it might not be necessary for every company, the literature on AI ethics suggests that principles should be followed by some mechanism to enforce them. Although AI governance may seem heavy for an SME-sized firm, measures like competence and knowledge development, and stakeholder communication reported by Seppälä et al. (2021) seem like viable options. Stakeholder communication could include measures such as disclosing the use of AI to customers. Communicating the ethical principles, general instructions, and limitations of generative AI use across the organization would combine knowledge development and communication.

On a final note, knowledge sharing, increasing awareness and training were considered helpful ways to understand and control risks better. Some kind of training, code of conduct or ethics committee could be recommended based on these findings. Alternatively, or additionally, informal conversations and encouraging everyone to use generative AI and familiarize themselves with its limitations could be beneficial.

6.2 Contribution

The goal of this case study was to 1) explore the ethical challenges and risks posed by generative AI use in a management consulting and research business and 2) explore how those risks could be mitigated at the organizational level. Prior academic literature was not found to focus on GAI use in consulting or business research. In general, the study area is emerging, and no studies were found considering the compatibility of the ethical principles of AI with GAI.

This study makes a two-fold contribution. First, the findings can provide scholars with initial implications of the prominent ethical challenges of generative AI in consulting and research work on the business side. Second, practitioners might find suggestions for risk mitigation practices. These issues could be further pursued by academics from several perspectives: studying ethical use of GAI in practice more broadly, studying GAI use in

management consulting or research industries, and studying whether GAI brings something new to the AI ethics principles and risk mitigation.

6.3 Limitations and Future Research

This study was conducted as a single case study with five interviews, making the small amount of empirical data its main limitation. Some ethical challenges or risk mitigations strategies may have remained undiscovered. The research design and scope of data negatively affect the generalizability of the study. The findings may also be subjective due to the interpretive stance and the thematic analysis approach. Both give the researcher some freedom to influence the results according to the perspective and interests of the study.

The research field is emerging with an unestablished theoretical foundation. This sets challenges to placing research questions and interpreting results. The literature on generative AI is quickly developing and thus the newest papers (after November 2023) are largely left out of this study. There is also some fragmentation as scholars seem to take different directions like LLMs, generative AI, or ChatGPT, whereas technical roots can be traced to, e.g., natural language processing (NLP) and generative adversarial networks (GAN). Together this paradigm can be called generative AI, or foundational models like Schneider et al. (2024) do. Due to novelty of the area, many academic sources used are preprints or working papers and some sources are non-academic. Additionally, the technology itself is developing fast and therefore some parts of the thesis may already be outdated at the time of completion. For example, OpenAI's models have been advancing quickly and some source literature is based on earlier and some on newer versions of GPTs and ChatGPT. The literature review in this thesis is limited to the end of November 2023.

Future research might address these limitations by conducting comparative case studies examining several case organizations or conceptual studies to further theoretical understanding of ethical generative AI use in organizations. Additional directions for further research could be exploring the evolving use of GAI in various industries and its implications for ethics. Differentiating between the ethics of development and the use of GAI could also be interesting as there are likely some differences.

7 Conclusions

Generative artificial intelligence (GAI) poses new capabilities and ethical challenges, not least in the fields of management consulting and research. This thesis aimed to explore the ethical use of GAI in this context. More specifically, associated ethical challenges and risks as well as risk mitigation strategies were examined.

The AI ethics literature initially addressed guidelines or principles for ethical AI and has since moved on to implementing them in practice. However, generative AI is not yet fully considered in these contexts. Algorithmic accountability describes the stakeholders of an AI system and how they express accountability for it. As an element of algorithmic accountability, organizational accountability applies to the operator or developer of the AI system, including measures like AI ethics principles, AI governance measures and other implementation mechanisms for ethical conduct.

This thesis is a qualitative case study with an interpretive stance and an explorative approach. To collect empirical data, five interviews were conducted with the case organization's employees. The thematic analysis of the interviews revealed two key themes: the ethical challenges and risks posed by GAI and the measures that could be taken to avoid these risks.

The identified ethical issues partly align with the literature on AI ethics. However, information security, trust in AI and human oversight were emphasized which is not the case with Jobin's et al. (2019) list of ethical guidelines. In contrast, these issues were highlighted in Hagendorff's (2024) preprint mapping the landscape of GAI ethics. Among other issues, the participants raised concerns about data privacy, potential biases, and the integrity of AI-generated texts. As an additional finding, the lack of identified ethical risks was also brought up.

In terms of avoiding the ethical risks, much is already done. Despite the lack of formal measures, designated safe GAI tools are used, the use of sensitive data is careful and regulation like GDPR and the EU AI Act are complied with. Participants still expressed a desire for increasing awareness and training on ethical use of GAI. Implementing ethical guidelines or instructions for GAI use was also considered essential. Regarding accountability, users were expected to be ultimately responsible for ethical matters since the potential for harm might not be adequately controlled by regulators and developers.

The limitations of this study are the small sample size and the subjective nature of the interpretive research stance. As the field of generative AI evolves, further research is required to find out whether existing frameworks should be adapted and how the ethical risks are mitigated in practice.

References

Abdi, H. – Valentin, D. – Edelman, B. (1999) *Neural networks*. Sage, 1–2.

Adams, W. C. (2015) *Conducting semi-structured interviews*. Handbook of practical program evaluation, 492–505.

Anfara, V. A. – Brown, K. M. – Mangione, T. L. (2002) Qualitative Analysis on Stage: Making the Research Process More Public. *Educational Researcher*, 31(7), 28–38. http://www.jstor.org/stable/3594403

Angwin, J. – Larson, J. – Mattu, S. – Kirchner, L. (2016) How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, retrieved 24.2.2023.

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

Bbc.com (4.10.2023) MrBeast and BBC stars used in deepfake scam videos. https://www.bbc.com/news/technology-66993651, retrieved 2.1.2024.

Bendel, O. (2023) Image synthesis from an ethical perspective. *AI & SOCIETY*, 1–10. https://doi.org/10.1007/s00146-023-01780-4

Bender, E. M. – Gebru, T. – McMillan-Major, A. – Shmitchell, S. (2021) On the dangers of stochastic parrots: Can language models be too big?. *In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Berente, N. – Gu, B. – Recker, J. – Santhanam, R. (2021) Managing artificial intelligence. *MIS quarterly*, 45(3).

Bianchi, F. – Kalluri, P. – Durmus, E. – Ladhak, F. – Cheng, M. – Nozza, D. – Hashimoto, T. – Jurafsky, D. –Zou, J. – Caliskan, A. (2023) Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. https://doi.org/10.1145/3593013.3594095

Blogs.microsoft.com (16.3.2023) *Introducing Microsoft 365 Copilot – your copilot for work*. https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/, retrieved 29.11.2023.

Braga, A. – Logan, R. K. (2017) The emperor of strong AI has no clothes: limits to artificial intelligence. *Information*, 8(4), 156. https://doi.org/10.3390/info8040156

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

Bresnahan, T. F. – Trajtenberg, M. (1995) General purpose technologies 'Engines of growth'? *Journal of econometrics*, 65(1), 83–108. https://doi.org/10.1016/0304-4076(94)01598-T

Bouville, M. (2008) Plagiarism: Words and ideas. *Science and engineering ethics*, 14, 311–322. https://doi.org/10.1007/s11948-008-9057-6

Businessinsider.com (20.4.2023) ChatGPT could cost over $700,000 per day to operate. Microsoft is reportedly trying to make it cheaper. https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4?r=US&IR=T, retrieved 2.1.2024.

Castelvecchi, D. (2016) Can we open the black box of AI? *Nature News*, 538(7623), https://doi.org/20. 10.1038/538020a

Chen, L. – Zaharia, M. – Zou, J. (2023) How is ChatGPT's behavior changing over time? *arXiv:2307.09009.* https://doi.org/10.48550/arXiv.2307.09009

Chew, R. – Bollenbacher, J. – Wenger, M. – Speer, J. – Kim, A. (2023) LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. *arXiv preprint arXiv:2306.14924*. https://doi.org/10.48550/arXiv.2306.14924

Clarke, R. (2019) Principles and business processes for responsible AI. *Computer Law & Security Review*, 35(4), 410–422.

Cnn.com (29.4.2023) 'Mom, these bad men have me': She believes scammers cloned her daughter's voice in a fake kidnapping. https://edition.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html, retrieved 21.12.2023.

de Cock Buning, M. (2018) A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation. Publications Office of the European Union.

Collins, C. – Dennehy, D. – Conboy, K. – Mikalef, P. (2021) Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, 60. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Corbett-Davies, S. – Pierson, E. – Feller, A. – Goel, S. (2016) A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *Washington Post*, 17.

Dehouche, N. (2021) Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21, 17–23.

Dell'Acqua, F. – McFowland, E. – Mollick, E. R. – Lifshitz-Assaf, H. – Kellogg, K. – Rajendran, S. – Krayer, L. – Canceldon, F. – Lakhani, K. R. (2023) Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, 24-013. https://dx.doi.org/10.2139/ssrn.4573321

Desaire, H. – Chua, A. E. – Isom, M. – Jarosova, R. – Hua, D. (2023) Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools. *Cell Reports Physical Science*. https://doi.org/10.1016/j.xcrp.2023.101426

Devlin, J. – Chang, M. W. – Lee, K. – Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

van Dis, E. A. – Bollen, J. – Zuidema, W. – van Rooij, R. – Bockting, C. L. (2023) ChatGPT: five priorities for research. *Nature*, 614(7947), 224–226. https://www.nature.com/articles/d41586-023-00288-7

Elkhatat, A. M. (2023) Evaluating the authenticity of ChatGPT responses: a study on text-matching capabilities. *International Journal for Educational Integrity*, 19(1), 15. https://doi.org/10.1007/s40979-023-00137-0

Elkhatat, A. M. – Elsaid, K. – Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1), 17. https://doi.org/10.1007/s40979-023-00140-5

Eloundou, T. – Manning, S. – Mishkin, P. – Rock, D. (2023) Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*. https://doi.org/10.48550/arXiv.2303.10130

Eke, D. O. (2023) ChatGPT and the rise of generative AI: threat to academic integrity? *Journal of Responsible Technology*, 13.

Eriksson, P. – Kovalainen, A. (2008) *Case study research*. SAGE Publications Ltd, 115–136. https://doi.org/10.4135/9780857028044

EU Recommendation 2003/361 (20.5.2003) Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises. European Commission. http://data.europa.eu/eli/reco/2003/361/oj, retrieved 7.12.2023.

Europarl.europa.eu (9.12.2023) Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI. https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai, retrieved 8.1.2024.

Europarl.europa.eu (19.12.2023) EU AI Act: first regulation on artificial intelligence. https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, retrieved 8.1.2024.

Flores, A. W. – Bechtel, K. – Lowenkamp, C. T. (2016) False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80, 38.

Floridi, L. (2018) Soft ethics and the governance of the digital. *Philosophy & Technology*, 31, 1–8. https://doi.org/10.1007/s13347-018-0303-9

Floridi, L. – Chiriatti, M. (2020) GPT-3: Its nature, scope, limits, and consequences, *Minds and Machines*, 30, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Floridi, L. – Cowls, J. – Beltrametti, M. – Chatila, R. – Chazerand, P. – Dignum, V. – Luetge, C. – Madelin, R – Pagallo, U. – Rossi, F. – Schafer, B. – Valcke, P. – Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds and machines*, 28, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Franzoni, V. (2023) From black box to glass box: advancing transparency in artificial intelligence systems for ethical and trustworthy AI. In *International Conference*

on Computational Science and Its Applications, 118-130. https://doi.org/10.1007/978-3-031-37114-1_9

Ft.com (2.10.2023) Workers could be the ones to regulate AI. https://www.ft.com/content/edd17fbc-b0aa-4d96-b7ec-382394d7c4f3, retrieved 2.1.2024.

Future of Jobs Report 2023 (2023) World Economic Forum. https://www.weforum.org/reports/the-future-of-jobs-report-2023/, retrieved 4.1.2024.

Gamieldien, Y. – Case, J. M. – Katz, A. (2023) Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding. https://dx.doi.org/10.2139/ssrn.4487768

Gasser, U. – Almeida, V. A. (2017) A layered model for AI governance. *IEEE Internet Computing*, 21(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

Ghauri, P. – Grønhaug, K. – Strange, R. (2020) *Research methods in business studies*. Cambridge University Press, 129–152.

Girotra, K. – Meincke, L. – Terwiesch, C. – Ulrich, K. T. (2023). Ideas are dimes a dozen: Large language models for idea generation in innovation. Working paper. Available at SSRN 4526071. https://dx.doi.org/10.2139/ssrn.4526071

Goertzel, B. (2014) Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1), 1. https://doi.org/10.2478/jagi-2014-0001

Goffman, E. (1959) *The presentation of self in everyday life* ([Rev. and expanded ed.].). New York: Doubleday Anchor Books.

Goldstein, J. A. – Sastry, G. – Musser, M. – DiResta, R. – Gentzel, M. – Sedova, K. (2023) Forecasting potential misuses of language models for disinformation campaigns and how to reduce risk. *arXiv:2301.04246*. https://doi.org/10.48550/arXiv.2301.04246

Grabe, S. – Ward, L. M. – Hyde, J. S. (2008) The role of the media in body image concerns among women: a meta-analysis of experimental and correlational studies. *Psychological bulletin*, 134(3), 460. https://psycnet.apa.org/doi/10.1037/0033-2909.134.3.460

Gravel, J. – D'Amours-Gravel, M. – Osmanlliu, E. (2023) Learning to fake it: limited responses and fabricated references provided by ChatGPT for medical questions.

Mayo Clinic Proceedings: Digital Health, 1(3), 226–234. https://doi.org/10.1016/j.mcpdig.2023.05.004

Gross, N. (2023) What chatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences*, 12(8), 435. https://doi.org/10.3390/socsci12080435

Guba, E. G. (1981) Criteria for assessing the trustworthiness of naturalistic inquiries. *Ectj*, 29(2), 75–91.

Gupta, M. – Akiri, C. – Aryal, K. – Parker, E. – Praharaj, L. (2023) From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*, 11, 80218–80245. https://doi.org/10.1109/ACCESS.2023.3300381

Hagendorff, T. (2020) The ethics of AI ethics: An evaluation of guidelines. *Minds and machines*, 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hagendorff, T. (2024) Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *arXiv preprint arXiv:2402.08323*. https://doi.org/10.48550/arXiv.2402.08323

Hanna, R. – Kazim, E. (2021) Philosophical foundations for digital ethics and AI Ethics: a dignitarian approach. *AI Ethics*, 1, 405–423. https://doi.org/10.1007/s43681-021-00040-9

Hazell, J. (2023) Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns. *arXiv:2305.06972*. https://doi.org/10.48550/arXiv.2305.06972

Heilinger, J. C. (2022) The ethics of AI ethics. A constructive critique. *Philosophy & Technology*, 35(3), 61. https://doi.org/10.1007/s13347-022-00557-9

Help.openai.com (?) What is ChatGPT? https://help.openai.com/en/articles/6783457-what-is-chatgpt, retrieved 2.11.2023.

Horneber, D. – Laumer, S. (2023) Algorithmic Accountability. *Business & Information Systems Engineering*, 1–8. https://doi.org/10.1007/s12599-023-00817-8

Ihalainen, J. (2018) Computer creativity: Artificial intelligence and copyright. *Journal of Intellectual Property Law & Practice*, 13(9), 724–728. https://doi.org/10.1093/jiplp/jpy031

Jobin, A. – Ienca, M. – Vayena, E. (2019) The global landscape of AI ethics guidelines. *Nature machine intelligence*, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jordan, M. I. – Mitchell, T. M. (2015) Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kahveci, Z. Ü. (2023) Attribution problem of generative AI: a view from US copyright law. *Journal of Intellectual Property Law and Practice*, jpad076. https://doi.org/10.1093/jiplp/jpad076

Kazim, E. – Koshiyama, A. S. (2021) A high-level overview of AI ethics. *Patterns*, 2(9). https://doi.org/10.1016/j.patter.2021.100314

Kietzmann, J. – Lee, L. W. – McCarthy, I. P. – Kietzmann, T. C. (2020) Deepfakes: Trick or treat? *Business Horizons*, 63(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

Kleinberg, J. – Mullainathan, S. – Raghavan, M. (2016) Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv*, 1609.05807. https://doi.org/10.48550/arXiv.1609.05807

Koopman, P. – Wagner, M. (2017) Autonomous Vehicle Safety: An Interdisciplinary Challenge. *IEEE Intelligent Transportation Systems Magazine*, 9(1), 90–96. https://doi.org/10.1109/MITS.2016.2583491

Kreps, S. – McCain, R. M. – Brundage, M. (2022) All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1), 104–117. https://doi.org/10.1017/XPS.2020.37

Kung, T. – Cheatham, M. – ChatGPT – Medenilla, A. – Sillos, C. – De Leon, L. – Elepaño, C. – Madriaga, M. – Aggabao, R. – Diaz-Candido, G. – Maningo, J. – Tseng V. (2022) Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models. *PLoS digital health*, 2(2). https://doi.org/10.1371/journal.pdig.0000198

LeCun, Y. – Bengio, Y. – Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Li, X. – Zhang, T. (2017) An exploration on artificial intelligence application: From security, privacy and ethic perspective. *In 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 416–420. https://doi.org/10.1109/ICCCBDA.2017.7951949

Liang, P. P. – Wu, C. – Morency, L. P. – Salakhutdinov, R. (2021) Towards understanding and mitigating social biases in language models. *Proceedings of the 38th International Conference on Machine Learning*, 139:6565–6576.

Liang, W. – Yuksekgonul, M. – Mao, Y. – Wu, E. – Zou, J. (2023) GPT detectors are biased against non-native English writers. *arXiv preprint arXiv:2304.02819*. https://doi.org/10.48550/arXiv.2304.02819

Libert, B. – Beck, M. (2017) AI may soon replace even the most elite consultants. *Harvard Business Review*, 24(7).

Lim, W. M. – Gunasekara, A. – Pallant, J. L. – Pallant, J. I. – Pechenkina, E. (2023) Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. *The International Journal of Management Education*, 21(2). https://doi.org/10.1016/j.ijme.2023.100790

Lincoln, Y. S. – Guba, E. G. (1986) But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. *New directions for program evaluation*, 1986(30), 73–84. https://doi.org/10.1002/ev.1427

Liu, X. M. – Murphy, D. (2020) A Multi-Faceted Approach for Trustworthy AI in Cybersecurity. *Journal of Strategic Innovation & Sustainability*, 15(6), 68–78.

Lund, B. D. – Wang, T. – Mannuru, N. R. – Nie, B. – Shimray, S. – Wang, Z. (2023) ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5), 570–581. https://doi.org/10.1002/asi.24750

Martin, K. (2019) Ethical implications and accountability of algorithms. *Journal of business ethics*, 160, 835–850. https://doi.org/10.1007/s10551-018-3921-3

Mazurek, G. – Małagocka, K. (2019) Perception of privacy and data protection in the context of the development of artificial intelligence. *Journal of Management Analytics*, 6(4), 344–364. https://doi.org/10.1080/23270012.2019.1671243

McKinsey Global Institute (2012) The social economy: Unlocking value and productivity through social technologies. https://www.mckinsey.com/~/media/mckinsey/industries/technology%20media%20and%20telecommunications/high%20tech/our%20insights/the%20social%20economy/mgi_the_social_economy_full_report.pdf, retrieved 30.11.2023.

McKinsey Global Institute (2018) Notes from the AI frontier: Modeling the impact of AI on the world economy. https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Artificial%20Intelligence/Notes%20from%20the%20frontier%20Modeling%20the%20impact%20of%20AI%20on%20the%20world%20economy/MGI-Notes-from-

the-AI-frontier-Modeling-the-impact-of-AI-on-the-world-economy-September-2018.pdf, retrieved 21.9.2023.

McKinsey Global Institute (2023) The state of AI in 2023: Generative AI's breakout year. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year#steady, retrieved 21.9.2023.

McNamara, A. – Smith, J. – Murphy-Hill, E. (2018) Does ACM's code of ethics change ethical decision making in software development? *In 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

Mehrabi, N. – Morstatter, F. – Saxena, N. – Lerman, K. – Galstyan, A. (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6), 1–35. https://doi.org/10.1145/3457607

Mellon, J. – Bailey, J. – Scott, R. – Breckwoldt, J. – Miori, M. (2022) Does GPT-3 know what the Most Important Issue is? Using Large Language Models to Code Open-Text Social Survey Responses at Scale. https://dx.doi.org/10.2139/ssrn.4310154

Miller, E. J. – Steward, B. A. – Witkower, Z. – Sutherland, C. A. – Krumhuber, E. G. – Dawel, A. (2023) AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychological Science*. https://doi.org/10.1177/09567976231207095

Min, B. – Ross, H. – Sulem, E. – Veyseh, A. P. B. – Nguyen, T. H. – Sainz, O. – Agirre, E. – Heintz, I. – Roth, D. (2023) Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2), 1–40. https://doi.org/10.1145/3605943

Mirsky, Y. – Lee, W. (2021) The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1), 1–41. https://doi.org/10.1145/3425780

Mittelstadt, B. (2019) Principles alone cannot guarantee ethical AI. *Nature machine intelligence*, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4.

Munn, L. (2023) The uselessness of AI ethics. *AI and Ethics*, 3(3), 869–877. https://doi.org/10.1007/s43681-022-00209-w

Murdoch, B. (2021) Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, 22(1), 1–5. https://doi.org/10.1186/s12910-021-00687-3

Myers, M. D. – Newman, M. (2007) The qualitative interview in IS research: Examining the craft. *Information and organization*, 17(1), 2–26. https://doi-org.ezproxy.utu.fi/10.1016/j.infoandorg.2006.11.001

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

Mökander, J. – Floridi, L. (2021) Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, 31(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

Newstead, T. – Eager, B. – Wilson, S. (2023) How AI can perpetuate–or help mitigate–gender bias in leadership. *Organizational Dynamics*, 52(4). https://doi-org.ezproxy.utu.fi/10.1016/j.orgdyn.2023.100998

Nytimes.com (24.4.2023) An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World. https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html, retrieved 9.11.2023.

Nytimes.com (30.8.2023) Voice Deepfakes Are Coming for Your Bank Balance. https://www.nytimes.com/2023/08/30/business/voice-deepfakes-bank-scams.html, retrieved 2.1.2024.

Oliver, D. G. – Serovich, J. M. – Mason, T. L. (2005) Constraints and opportunities with interview transcription: Towards reflection in qualitative research. *Social forces*, 84(2), 1273–1289. https://doi.org/10.1353/sof.2006.0023

Openai.com (19.10.2023) DALL·E 3 is now available in ChatGPT Plus and Enterprise. https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise, retrieved 1.12.2023.

Openai.com (6.11.2023) Introducing GPTs. https://openai.com/blog/introducing-gpts, retrieved 1.12.2023.

Openai.com (30.11.2022) ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/, retrieved 1.2.2023.

El Ouadrhiri, A. – Abdelhadi, A. (2022) Differential privacy for deep and federated learning: A survey. *IEEE Access*, 10, 22359–22380. https://doi.org/10.1109/ACCESS.2022.3151670

Oury, J. (10.4.2018) Europe trapped in the Artificial Intelligence paradox. https://www.europeanscientist.com/en/editors-corner/europe-trapped-in-the-artificial-intelligence-paradox/, retrieved 7.3.2023.

Parmar, B. L. – Freeman, R. E. – Harrison, J. S. – Wicks, A. C. – Purnell, L. – De Colle, S. (2010) Stakeholder theory: The state of the art. *Academy of Management Annals*, 4(1), 403–445. https://doi.org/10.5465/19416520.2010.495581

Peres, R. – Schreier, M. – Schweidel, D. – Sorescu, A. (2023) On ChatGPT and beyond: How generative artificial intelligence may affect research, teaching, and practice. *International Journal of Research in Marketing*, 40(2), 269–275. https://doi.org/10.1016/j.ijresmar.2023.03.001

Petroni, F. – Rocktäschel, T. – Lewis, P. – Bakhtin, A. – Wu, Y. – Miller, A. H. – Riedel, S. (2019) Language models as knowledge bases? *arXiv preprint arXiv:1909.01066.* https://doi.org/10.48550/arXiv.1909.01066

Price, W. N. – Cohen, I. G. (2019) Privacy in the age of medical big data. *Nature medicine*, 25(1), 37–43. https://doi.org/10.1038/s41591-018-0272-7

Publications Office of the European Union (2018) *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation*. https://hdl.handle.net/1814/70297, retireved13.11.2023. https://doi.org/10.2759/739290

Radford, A. – Narasimhan, K. – Salimans, T. – Sutskever, I. (2018) *Improving language understanding by generative pre-training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, retrieved 23.11.2023.

Rahman, M. M. – Terano, H. J. – Rahman, M. N. – Salamzadeh, A. – Rahaman, M. S. (2023) ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples. *Journal of Education, Management and Development Studies*, 3(1), 1–12. https://doi.org/10.52631/jemds.v3i1.175

Reuters.com (11.10.2018) Insight - Amazon scraps secret AI recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/?utm_source=morning_brew, retrieved 21.11.2023.

Reuters.com (2.2.2023) ChatGPT sets record for fastest-growing user base - analyst note. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, retrieved 22.9.2023.

Rombach, R. – Blattmann, A. – Lorenz, D. – Esser, P. – Ommer, B. (2022) High-resolution image synthesis with latent diffusion models. *In Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, 10684–10695. https://doi.org/10.1109/CVPR52688.2022.01042

Rozado, D. (2023) The political biases of chatgpt. *Social Sciences*, 12(3), 148. https://doi.org/10.3390/socsci12030148

Samuel, A. L. (1960) Some moral and technical consequences of automation—a refutation. *Science*, 132(3429), 741–742. https://doi.org/10.1126/science.132.3429.741

Sanderson, K. (2023) GPT-4 is here: what scientists think. *Nature*, 615(773), https://doi.org/10.1038/d41586-023-00816-5

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

Schneider, J. – Abraham, R. – Meske, C. – Vom Brocke, J. (2023) Artificial intelligence governance for businesses. *Information Systems Management*, 40(3), 229–249. https://doi.org/10.1080/10580530.2022.2085825

Schneider, J. – Meske, C. – Kuss, P. (2024) Foundation Models: A New Paradigm for Artificial Intelligence. *Business & Information Systems Engineering*, 1–11. https://doi.org/10.1007/s12599-024-00851-0

Schultze, U. – Avital, M. (2011) Designing interviews to generate rich data for information systems research. *Information and organization*, 21(1), 1–16.

Shanks, G. – Bekmamedova, N. (2018) Case Study Research in Information Systems. In: *Research Methods: Information, Systems, and Contexts: Second Edition*, eds. Kirsty Williamson – Graeme Johanson, 193–208. Chandos Publishing.

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for information*, 22(2), 63–75. https://doi.org/10.3233/EFI-2004-22201

Shin, D. (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Smits, J. – Borghuis, T. (2022) Generative AI and Intellectual Property Rights. In: *Law and Artificial Intelligence: Regulating AI and Applying AI in Legal Practice,* eds. Bart Custers – Eduard Fosch-Villaronga, 323–344). TMC Asser Press, The Hague. https://doi.org/10.1007/978-94-6265-523-2_17

Smuha, N. A. (2019) The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 20(4), 97–106. https://doi.org/10.9785/cri-2019-200402.

Spirling, A. (2023) Why open-source generative AI models are an ethical way forward for science. *Nature*, 616(7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4

Stahl, B. C. – Eke, D. (2024) The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74, 102700. https://doi.org/10.1016/j.ijinfomgt.2023.102700

Strubell, E. – Ganesh, A. – McCallum, A. (2019) Energy and policy considerations for deep learning in NLP. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Sun, J. – Liao, Q. V. – Muller, M. – Agarwal, M. – Houde, S. – Talamadupula, K. – Weisz, J. D. (2022) Investigating explainability of generative AI for code through scenario-based design. *In 27th International Conference on Intelligent User Interfaces*, 212–228. https://doi.org/10.1145/3490099.3511119

Sutton, R. S. – Barto, A. G. (2018) *Reinforcement learning: An introduction*. MIT press, 1–2.

Syntheticmedia.partnershiponai.org (27.2.2023) PAI's Responsible Practices for Synthetic Media. https://syntheticmedia.partnershiponai.org/#read_the_framework, retrieved 31.10.2023.

Technologyreview.com (23.9.2020) OpenAI is giving Microsoft exclusive access to its GPT-3 language model. https://www.technologyreview.com/2020/09/23/1008729/openai-is-giving-microsoft-exclusive-access-to-its-gpt-3-language-model/, retrieved 2.1.2024.

Theguardian.com (3.8.2023) Doctored Sunak picture is just latest in string of political deepfakes. https://www.theguardian.com/technology/2023/aug/03/doctored-sunak-picture-is-just-latest-in-string-of-political-deepfakes, retrieved 21.11.2023.

The New York Times (19.4.2023) An A.I. Hit of Fake 'Drake' and 'The Weeknd' Rattles the Music World. https://www.nytimes.com/2023/04/19/arts/music/ai-drake-the-weeknd-fake.html, retrieved 1.11.2023.

Touvron, H. – Lavril, T. – Izacard, G. – Martinet, X. – Lachaux, M. A. – Lacroix, T. – Rozière, B. – Goyal, N – Hambro, E. – Azhar, F. – Rodriguez, A. – Joulin, A. – Grave, E. – Lample, G. (2023) Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*. https://doi.org/10.48550/arXiv.2302.13971

Vraga, E. K. – Bode, L. (2020) Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, 37(1), 136–144.

Wach, K. – Duong, C. D. – Ejdys, J. – Kazlauskaitė, R. – Korzynski, P. – Mazurek, G. – Paliszkiewicz, J. – Ziemba, E. (2023) The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review*, 11(2), 7–24. https://doi.org/10.15678/EBER.2023.110201

Wagner, M. W. – Ertl-Wagner, B. B. (2023) Accuracy of information and references using ChatGPT-3 for retrieval of clinical radiological information. *Canadian Association of Radiologists Journal*, 08465371231171125. https://doi.org/10.1177/08465371231171125

Walsham, G. (1995) Interpretive case studies in IS research: nature and method. *European Journal of information systems*, 4(2), 74–81

Walters, W. H. – Wilder, E. I. (2023) Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13(1), 14045. https://doi.org/10.1038/s41598-023-41032-5

Whittlestone, J. – Nyrup, R. – Alexandrova, A. – Cave, S. (2019) The role and limits of principles in AI ethics: towards a focus on tensions. *In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289

Wiens, J. – Saria, S. – Sendak, M. – Ghassemi, M. – Liu, V. X. – Doshi-Velez, F. – Jung, K. – Heller, K. – Kale, D. – Saeed, M. – Ossorio, P. N. – Thadaney-Israni, S. – Goldenberg, A. (2019) Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9), 1337–1340. https://doi.org/10.1038/s41591-019-0548-6

Wieringa, M. (2020) What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 1–18.

Wolfe, A. (1991) Mind, self, society, and computer: Artificial intelligence and the sociology of mind. *American Journal of Sociology*, 96(5), 1073–1096.

Wolfram, S. (2023) What Is ChatGPT Doing ... and Why Does It Work? *Stephen Wolfram Writings*. writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work, retrieved 21.3.2023.

Zohny, H. – McMillan, J. – King, M. (2023) Ethics of generative AI. *Journal of Medical Ethics*, 49(2), 79–80. https://doi.org/10.1136/jme-2023-108909

Appendices

Appendix 1 Translated Interview Structure

1. What is your role in the company?
2. Do you have any background with AI-related matters, or is it part of your work?
3. Do you use or plan to use generative artificial intelligence such as ChatGPT or AI image creators in your work?
    a. Which tools do you use and for what purposes?
    b. Are there specific tools recommended by the organization?
4. What kinds of ethical problems would you say might be relevant to your use of generative AI?
    a. Are there any specific work use cases where you think generative AI could pose ethical challenges now or in the future?
    b. What kind of trade-offs there are between ethical issues and generative AI use? For example, do concerns about bias in generative AI limit its use?
    c. Important principles of ethical AI include e.g., transparency, justice and fairness, non-maleficence, responsibility, and privacy. In your opinion, is anything missing from these, or should any be excluded specifically in the context of generative AI?
    d. From where do you get information about ethical problems regarding generative AI?
5. Who do you think should be responsible for ethical questions regarding generative AI? (e.g., AI developers, users, legislators etc.)
    a. Why these parties?
6. What would you say could be done to avoid unethical consequences related to generative AI use on the organizational level?
    a. Are you already acting to mitigate the ethical risks? If yes, how?

b. Are there common policies for this in the company?

c. What would help you in controlling the ethical risks?

Appendix 2 Research Data Management Plan for Students

Research data

Research data refers to all the material with which the analysis and results of the research can be verified and reproduced. It may be, for example, various measurement results, data from surveys or interviews, recordings or videos, notes, software, source codes, biological samples, text samples, or collection data.
In the table below, list all the research data you use in your research. Note that the data may consist of several different types of data, so please remember to list all the different data types. List both digital and physical research data.

Research data type	**Contains personal details/information ***	**I will gather/produce the data myself**	**Someone else has gathered/produced the data**	**Other notes**
Data type 1: *Interview recordings*	x	x		These will be deleted after transcription.
Data type 2: *Interview transcriptions*		x		

* Personal details/information are all information based on which a person can be identified directly or indirectly, for example by connecting a specific piece of data to another, which makes identification possible. For more information about what data is considered personal go to the Office of the Finnish Data Protection Ombudsman's website

Processing personal data in research

If your data contains personal details/information, you are obliged to comply with the EU's General Data Protection Regulation (GDPR) and the Finnish Data Protection Act. For data that

contains personal details, you must prepare a Data Protection Notice for your research participants and determine who is the controller for the research data.
I will prepare a Data Protection Notice** and give it to the research participants before collecting data ☐

The controller** for the personal details is the student themself ☒ the university ☐

My data does not contain any personal data ☒

** More information at the university's intranet page, Data Protection Guideline for Thesis Research

Permissions and rights related to the use of data

Find out what permissions and rights are involved in the use of the data. Consult your thesis supervisor, if necessary. Describe the use permissions and rights for each data type. You can add more data types to the list, if necessary.

Self-collected data

You may need separate permissions to use the data you collect or produce, both in research and in publishing the results. If you are archiving your data, remember to ask the research participants for the necessary permissions for archiving and further use of the data. Also, find out if the repository/archive you have selected requires written permissions from the participants.
Necessary permissions and how they are acquired:

Data type 1: Oral confirmation or notice of data collection in the invitation email.

Data type 2:

Storing the data during the research process

Where will you store your data during the research process?

In the university's network drive ☒
In the university-provided Seafile Cloud Service ☐
Other location, please specify: ☐

The university's data storage services will take care of data security and backup files automatically. If you choose to store your data somewhere other than in the services provided by the university, please specify how you will ensure data security and file backups. Remember to make sure you know every time where you are saving the edited/modified data.

If you are using a smartphone to record anything, please check in advance where the audio or video will be saved. If you are using commercial cloud services (iCloud, Dropbox, Google Drive, etc.) and your data contains personal data, make sure the information you provide in the Data Protection Notice about data migration matches your device settings. The use of commercial cloud services means the data will be transferred to third countries outside the EU.

Documenting the data and metadata

How would you describe your research data so that even an outsider or a person unfamiliar with it will understand what the data is? How would you help yourself recall years later what your data consists of?

Data documentation

Can you describe what has happened to your research data during the research process? Data documentation is essential when you try to track any changes made to the data.
To document the data, I will use:
A field/research journal ☐
A separate document where I will record the main points of the data, such as changes made, phases of analysis, and significance of variables ☐
A readme file linked to the data that describes the main points of the data ☒
Other, please specify: The data is not changed after transcription. Analysis is done separately in NVivo software. ☒

Data arrangement and integrity

How will you keep your data in order and intact, as well as prevent any accidental changes to it?

I will keep the original data files separate from the data I am using in the research process, so that I can always revert back to the original, if need be. ☒

Version control: I will plan before starting the research how I will name the different data versions and I will adhere to the plan consistently. ☐

I recognise the life span of the data from the beginning of the research and am already prepared for situations, where the data can alter unnoticed, for example while recording, transcribing, downloading, or in data conversions from one file format to another, etc. ☐

Metadata

Metadata is a description of you research data. Based on metadata someone unfamiliar with your data will understand what it consists of. Metadata should include, among others, the file name, location, file size, and information about the producer of the data. Will you require metadata?

I will save my data into an archive or a repository that will take care of the metadata for me. ☐

I will have to create the metadata myself, because the archive/repository where I am uploading the data requires it. ☐

I will not store my data into a public archive/repository, and therefore I will not need to create any metadata. ☒

Data after completing the research

You are responsible for the data even after the research process has ended. Make sure you will handle the data according to the agreements you have made. The university recommends a general retention period of five (5) years, with an exception for medical research data, where the retention period is 15 years. Personal data can only be stored as long as it is necessary. If you have agreed to destroy the data after a set time period, you are responsible for destroying the data, even if you no longer are a student at the university. Likewise, when using the university's online storage services, destroying the data is your responsibility.
What happens to your research data, when the research is completed?

I will destroy part of the data, but store part of it for 5 years, because: the recordings of interviews contain personal data, while the transcribed data is anonymous. I will destroy the recordings from my computer and they will be destroyed from the case organization's servers according to the organization's policy.

If you will store the data, please identify where: Transcribed text files in the University servers.