Communications of the Association for Information Systems

Volume 56 Paper 1

3-6-2025

Understanding the Ethics of Generative AI: Established and New Ethical Principles

Joakim Laine
University of Turku, jhdlai@utu.fi

Matti Minkkinen
University of Turku

Matti Mäntymäki
University of Turku

Recommended Citation
Laine, J., Minkkinen, M., & Mäntymäki, M. (2025). Understanding the Ethics of Generative AI: Established and New Ethical Principles. Communications of the Association for Information Systems, 56, 1-25. https://doi.org/10.17705/1CAIS.05601

Understanding the Ethics of Generative AI: Established and New Ethical Principles

Cover Page Footnote

This manuscript underwent editorial review. It was received 09/13/2024 and was with the authors for one month for one revision. Fred Niederman served as Associate Editor.

Communications of the Association for Information Systems

Research Article DOI: 10.17705/1CAIS.05601 ISSN: 1529-3181

Understanding the Ethics of Generative AI: Established and New Ethical Principles

Joakim Laine
Information Systems Science
University of Turku
jhdlai@utu.fi

Matti Minkkinen
Information Systems Science
University of Turku

Matti Mäntymäki
Information Systems Science
University of Turku

Abstract:

This scoping review develops a conceptual synthesis of the ethics principles of generative artificial intelligence (GenAI) and large language models (LLMs). In regard to the emerging literature on GenAI, we explore 1) how established AI ethics principles are presented and 2) what new ethical principles have surfaced. The results indicate that established ethical principles continue to be relevant for GenAI systems but their salience and interpretation may shift, and that there is a need to recognize new principles in these systems. We identify six GenAI ethics principles: 1) respect for intellectual property, 2) truthfulness, 3) robustness, 4) recognition of malicious uses, 5) sociocultural responsibility, and 6) human-centric design. Addressing the challenge of satisfying multiple principles simultaneously, we suggest three meta-principles: categorizing and ranking principles to distinguish fundamental from supporting ones, mapping contradictions between principle pairs to understand their nature, and implementing continuous monitoring of fundamental principles due to the evolving nature of GenAI systems and their applications. To conclude, we suggest increased research emphasis on complementary ethics approaches to principlism, ethical tensions between different ethical viewpoints, end-user perspectives on the explainability and understanding of GenAI, and the salience of ethics principles to various GenAI stakeholders.

Keywords: Generative AI, GenAI, Artificial Intelligence, AI, Large Language Models, AI Chatbots, AI Ethics, AI Auditing. AI Governance.

This manuscript underwent editorial review. It was received 09/13/2024 and was with the authors for one month for one revision. Fred Niederman served as Associate Editor.

1 Introduction

Generative artificial intelligence (GenAI) and large language models (LLMs) have gained significant attention and popularity (Dwivedi et al., 2023; Van Slyke et al., 2023). At the same time, discussions and activities around artificial intelligence (AI) ethics (Jobin et al., 2019; Mirbabaie et al., 2022), AI governance (Mäntymäki et al., 2022; Minkkinen & Mäntymäki, 2023), and AI auditing (Laine et al., 2024) have intensified because of the prominent risks and "dark side" of AI technologies, evidenced by issues such as algorithmic discrimination (Mikalef et al., 2022). The development of increasingly powerful GenAI technologies (Jovanovic & Campbell, 2022) and the rising popularity of AI chatbots, such as OpenAI's ChatGPT, have presented new ethical challenges (Mirbabaie et al., 2022; Sun et al., 2024). Despite the advantages of AI chatbots, their responsible development and use have become growing topics of debate, especially in regard to ensuring that they align with ethical principles and values (Chatterjee & Dethlefs, 2023; Dwivedi et al., 2023; Jobin et al., 2019; Weidinger et al., 2021). LLMs may, for example, encode toxicity and reproduce bias against certain genders and religious groups (Zhuo et al., 2023). An extreme case of harmful content involved a mental health interaction in which the LLM advised a (simulated) depression patient to commit suicide (Wahde & Virgolin, 2021).

GenAI refers to "AI that can generate novel content, rather than simply acting on existing data" (Gozalo-Brizuela & Garrido-Merchan, 2023, p. 1). A key subset of GenAI is LLMs, which specialize in producing text that closely resembles human language (Cooper, 2023; Hadi et al., 2023). An example of an application based on LLMs is ChatGPT, developed by OpenAI and based on the generative pre-trained transformer (GPT) architecture (Qadir, 2023). ChatGPT was launched in late 2022 and quickly gained significant attention (Van Slyke et al., 2023). ChatGPT and comparable models, such as Claude (developed by Anthropic), PaLM, and Gemini (developed by Google), and LLaMA (developed by Meta), are notable for their ability to engage in conversational exchanges, delivering indistinguishably humanlike responses (e.g., Wolf et al., 2017). This proficiency in conversation is a product of extensive training on vast datasets, a characteristic of LLMs (Fui-Hoon Nah et al., 2023). Beyond language, GenAI's capabilities include generating creative content, such as music and images, and synthesizing information from various sources (Dasborough, 2023; Fui-Hoon Nah et al., 2023).

GenAI techniques, such as GPT models, enable the creation of more advanced content, such as text and images, compared to previous generations, thereby creating new opportunities, expectations, and risks. Therefore, GenAI systems need to be fair, explainable, and accountable to reduce model behavior risks and provide insight into what occurs inside the algorithmic black box (Jovanovic & Campbell, 2022). These principles are familiar from previous AI ethics guidelines, which tend to emphasize fairness, accountability, and transparency as the key principles (Dignum, 2019; Jobin et al., 2019; Mirbabaie et al., 2022). However, advanced AI chatbots need to meet additional expectations, such as dealing with misinformation, malicious uses, unintended unethical behaviors, chatbots being mistaken for humans, and environmental inefficiency and harms (Jovanovic & Campbell, 2022; Zhuo et al., 2023). Conversational chatbots that generate open-ended textual responses present a novel situation compared to AI that is used for decision-making in specific contexts, such as credit scoring or medical diagnosis (Hacker et al., 2023). This results in a discrepancy in AI ethics: The discussion on AI ethics principles is still couched in terms of machine learning systems used for decision-making, but LLMs and GenAI systems portend new ethical issues that require the AI ethics discourse to be revisited. Owing to the novelty of LLM technology and the long list of ethical demands from policymakers, researchers, and other stakeholders (Jobin et al., 2019; Jovanovic & Campbell, 2022; Zhuo et al., 2023), the ethical principles specific to AI chatbots and their ability to adhere to these principles remain relatively unclear.

Information systems (IS) research on GenAI has thus far probed its potential impacts on IS education (Van Slyke et al., 2023) and research (Davison et al., 2023; Davison et al., 2024) as well as its societal, work, and organizational impacts (Alavi et al., 2024; Benbya et al., 2024; Sabherwal & Grover, 2024). IS researchers have also discussed broader AI ethics discourses (Loebbecke et al., 2020; Mikalef et al., 2022) and impacts on IS research (Cameron et al., 2023; Loebbecke et al., 2020) and organizational practices (Ågerfalk et al., 2022). Addressing these issues raises the question of the principles that could ensure the ethical development and use of GenAI. Against this backdrop, this study investigates the ethics of GenAI by conducting a scoping review (Paré et al., 2015) and subsequent conceptual synthesis (cf. Torraco, 2005) on AI ethics principles in the context of GenAI. To this end, we pose the following research question:

RQ1: How are established AI ethics principles presented, and what new ethical principles have surfaced in the emerging literature on the ethical considerations of GenAI?

The interdisciplinary literature on GenAI is nascent and includes tentative lists of ethical principles (e.g., Jovanovic & Campbell, 2022; Zhuo et al., 2023). We develop a conceptual synthesis of these sets of principles, advancing theoretical understanding of how established AI ethics principles evolve with the advent of GenAI and what new principles are proposed. Our study contributes to the research areas on the implications of AI technologies and ethics-based AI auditing, particularly of GenAI systems, by highlighting what is new in the era of GenAI and suggesting meta-principles for managing the challenges and trade-offs involved in satisfying multiple principles simultaneously. By doing so, it lays the foundation for future research on the ethics of GenAI beyond principlism, investigating ethical tensions, end-user perspectives, and the salience of ethics principles for GenAI stakeholders, presented as a future research agenda.

2 Background

2.1 Evolution of Generative AI

AI is commonly defined as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019, p. 17). More broadly, it can be viewed as a moving frontier of computational advancements toward humanlike intelligence (Berente et al., 2021). Over the past 70 years, there have been several periods when AI has raised high hopes owing to its promising advancements; however, there have also been many cases of AI failing to meet expectations. Beginning with foundational concepts in the early 20th century, such as McCulloch and Pitts's 1943 computer model inspired by human neurons (Muthukrishnan et al., 2020) and Turing's 1950 introduction of the Turing test as a measure of AI's intelligence (Kaplan & Haenlein, 2019; Muthukrishnan et al., 2020), the field has continuously pushed the boundaries of computational advancements. This progression toward increasingly humanlike intelligence has been shaped by key developments and milestones. From the inception of the first AI chatbot, ELIZA (Weizenbaum, 1966), in 1966, to the sophisticated conversational abilities of modern systems, such as Apple's Siri and Amazon's Alexa (Adamopoulou & Moussiades, 2020; Fui-Hoon Nah et al., 2023), AI has consistently expanded its reach and impact (Berente et al., 2021).

Following the initial breakthroughs in AI, the field witnessed a significant shift with the rise of generative models, which marked a new era in AI's ability to process and produce complex content. This period saw the development of technologies such as hidden Markov models and Gaussian mixture models (Cao et al., 2023; Gupta et al., 2023), which represented advanced GenAI approaches. Lim et al. (2023, p. 2) define GenAI as "a technology that (i) leverages deep learning models to (ii) generate human-like content (e.g., images, words) in response to (iii) complex and varied prompts (e.g., languages, instructions, questions)." GenAI can generate novel content rather than simply analyzing or acting on existing data like expert systems (Gozalo-Brizuela & Garrido-Merchan, 2023). This means that AI algorithms can produce textual, visual, and auditory content with little to no human intervention (Longoni et al., 2022). GenAI is a technique that involves analyzing training examples to learn their patterns and distribution. It uses generative modeling and advanced deep learning methods to create synthetic artifacts that mimic real-world content, such as text, graphics, audio, and video (Jovanovic & Campbell, 2022). In addition, the purpose of GenAI is to generate new content, as opposed to making decisions based on existing content (Jovanovic & Campbell, 2022). The output of GenAI systems is diverse, open-ended, and largely unpredictable (Dwivedi et al., 2023; Mökander et al., 2023) the input for GenAI systems is relatively minimal (Hacker et al., 2023); the training data sets for GenAI systems currently feature a large volume of diverse data obtained from the Internet (Korzynski et al., 2023; Mökander et al., 2023; Stokel-Walker & Van Noorden, 2023); and GenAI systems are creative and generate fresh content rather than being limited to making decisions based on data and rules (Pavlik, 2023; Zhuo et al., 2023). Today, four of the most common GenAI techniques are generative adversarial networks, GPT, the generative diffusion model, and geometric deep learning (Jovanovic & Campbell, 2022).

The evolution of AI took a significant leap with the advent of LLMs, a breakthrough that fundamentally transformed the landscape of natural language processing (NLP) and conversational AI. LLMs differ significantly from previous AI techniques. LLMs are characterized by their unparalleled scale and complexity; are trained on massive datasets with billions of parameters; require massive computational resources; possess versatile capabilities; can adapt their responses in real time; are based on user input

and evolving contexts; and rely on extensive training data, raising concerns related to data privacy and security (Meskó & Topol, 2023). This has resulted in the development of advanced AI chatbots such as AlphaGo in 2015 and Google Bard and ChatGPT in 2022 (Fui-Hoon Nah et al., 2023; Meskó & Topol, 2023). AI chatbots are also classified as conversational AI, referring to automatic systems that use machine learning and NLP techniques to understand and respond to user input in natural language, allowing for communication through messaging channels, phone or web applications, websites, or operating systems (Ruane & Birhane, 2019). Conversational AI is a sub-domain of AI that consists of three main components: 1) the understanding of natural language inputs from users, which includes intent classification and entity extraction tasks to understand user input; 2) natural language generation, which generates natural language responses in a user-friendly format; and 3) a dialogue management system, which decides the agent's actions based on user input and the current state of the conversation (Kulkarni et al., 2019). GenAI chatbots, such as ChatGPT, fall under conversational AI and GenAI, but it is important to distinguish conceptually between these two partly overlapping subfields of AI. GenAI can provide a response and generate content, which goes beyond the humanlike interactions in conversational AI. Additionally, conversational AI typically relies on predefined responses, while GenAI can generate new responses beyond its explicit programming (Lim et al., 2023). Therefore, not all GenAI is conversational, and not all conversational AI can generate novel content.

The culmination of these developments in LLM technology was exemplified by ChatGPT, an AI chatbot developed by OpenAI (Chatterjee & Dethlefs, 2023). Launched in 2022, the chatbot was touted as the fastest-growing consumer app in the history of the Internet, reaching 100 million users two months after it was made public (Van Slyke et al., 2023). Using the advanced GPT models, ChatGPT represented a significant stride toward producing nuanced, humanlike conversations and responses (Lim et al., 2023). However, ChatGPT also exemplifies the shortcomings of LLMs. It has been known to provide incorrect answers, reference a nonexistent scientific study, write plausible-sounding but incorrect or nonsensical answers, and be overly sensitive and harbor bias from the large amounts of data on which it has been trained (Gordijn & Have, 2023; Meskó & Topol, 2023; Thorp, 2023). Although ChatGPT has safeguards against discussing sensitive topics and providing illegal and harmful advice, several users claim to have at least partly bypassed these safeguards (Sun et al., 2024). As AI continues to advance and integrate more deeply into various sectors, it becomes increasingly important to address these challenges, ensuring that AI development aligns with ethical principles and contributes positively to society.

2.2 Generative AI and AI Ethics Principles

The new characteristics of GenAI mean that AI ethics also need to be reexamined. The ethics of AI, particularly its core principles, have spurred extensive discussion, initially drawing on prior bioethics principles, such as autonomy and non-maleficence (Mittelstadt, 2019). These core ethics principles have also been adopted as one starting point for discussing IS ethics (Mingers & Walsham, 2010). Although there is general agreement that AI should be ethical, there is an ongoing debate on what constitutes ethical AI and which ethical requirements and technical standards are necessary for its implementation. To date, AI ethics principles have been listed in numerous sets of guidelines. Jobin et al. (2019) suggest that there is a global convergence around five core ethical principles (transparency, justice and fairness, non-maleficence, responsibility, and privacy) and six supportive ethical principles (beneficence, freedom & autonomy, trust, sustainability, dignity, and solidarity). In regard to AI ethics, prior to the consideration of GenAI, Jobin et al.'s (2019) study was a highly cited scholarly summary of academic and gray literature and provided a comprehensive overview of the key ethical principles that are central to the discussion on ethical AI.

In alignment with the research question, this paper adopts the deontological (focusing on duties and adherence to rules) perspective (Hagendorff, 2020; Laine et al., 2024; Mökander et al., 2021) and investigates how the emerging GenAI literature conceptualizes ethical principles. The deontological approach sets itself apart from consequentialism by concentrating on acts themselves and the rules justifying them rather than the consequences of acts, and it is also distinct from virtue ethics which is concerned with developing ways of behaving that naturally lead to well-being (Mingers & Walsham, 2010). According to Siau and Wang (2020), in the context of AI, ethical concerns refer to the moral obligations and duties of an AI application and its creators. The integration of LLMs—for example, ChatGPT—amplifies risks such as misinformation, lack of transparency, insufficient accountability, unfairness, and bias (Hacker et al., 2023; Meskó & Topol, 2023). Other possible challenges are harmful or inappropriate content, overreliance on LLMs, misuse, privacy and security, and the widening of the digital divide (Fui-Hoon Nah et al., 2023). As GenAI, such as ChatGPT, operates within diverse sociocultural contexts

across the world, the ethical factors that will need to be considered become more complicated (Fui-Hoon Nah et al., 2023).

GenAI is an emerging phenomenon, and few comprehensive lists of ethical principles specific to its systems have been developed. In addition, different papers refer to these principles and closely related concepts using various terms, such as harms (Weidinger et al., 2021), challenges (Fui-Hoon Nah et al., 2023; Kenthapadi et al., 2023), problems (Bale et al., 2024), issues (Parikh, 2023; Stahl & Eke, 2024), impacts (Gupta et al., 2023), concerns (Zhuo et al., 2023), expectations (Jovanovic & Campbell, 2022), characteristics (Mökander et al., 2023), and principles (Cheng & Liu, 2023; Guo et al., 2023; Sun et al., 2024; Wahde & Virgolin, 2021; Weisz et al., 2023). For example, Zhuo et al. (2023) categorize ethical concerns over ChatGPT into four areas: bias, reliability, robustness, and toxicity. Similarly, Jovanovic and Campbell (2022) emphasize that GenAI solutions must be efficient, explainable, fair, ethical, and accountable. Weidinger et al. (2021) identify ethical challenges related to LLMs, such as discrimination, information hazards, misinformation harms, malicious use, human–computer interaction harms, automation, and environmental harms. Additionally, Wahde and Virgolin (2021) propose principles for generative or conversational systems that include transparency, accountability, and safety through aspects such as interpretability, the ability to explain, independent data, interactive learning, and inquisitiveness. However, it is important to note that the focus areas in the existing literature differ somewhat. For instance, Wahde and Virgolin (2021) discuss principles, but not all are strictly ethical, while Zhuo et al. (2023) focus on ethical concerns rather than principles. This variation in terminology showcases the importance of our synthesis and requires us to interpret these somewhat homogeneous issues from the perspective of ethical principles, altering the wording where relevant.

Furthermore, it is important to note that some of these principles may overlap or even contradict each other. For instance, non-maleficence, which emphasizes preventing harm, closely aligns with the principle of recognizing malicious uses, as both focus on mitigating negative impacts. Similarly, the principle of truthfulness is intrinsically linked to transparency, as ensuring information accuracy often requires open disclosure. These overlaps suggest that certain ethical concerns are multifaceted and can be addressed from multiple angles, reinforcing their importance in the ethical discourse on GenAI. However, these principles can also present contradictions. For example, transparency might conflict with privacy concerns, as ensuring transparency in GenAI processes could require the disclosure of sensitive data. These potential overlaps and contradictions highlight the complexity of establishing a cohesive ethical framework.

3 Methodology

To investigate the suitability and potentially changed interpretations of established AI ethics principles in the context of GenAI, we chose Jobin et al.'s (2019) framework for our analysis because it aligns closely with the terminology used in the GenAI field and synthesizes a large number of papers related to AI ethics into a set of core principles. Additionally, the use of terminology that is common to the ethical and technical domains of GenAI and AI ethics, in general, can facilitate better communication and collaboration between researchers, practitioners, and policymakers, ultimately resulting in the more responsible and beneficial use of GenAI technology. For example, the European Union's High-Level Expert Group on Artificial Intelligence (2019) categorizes trustworthy AI into three components—ethical, lawful, and robust—while Zhuo et al. (2023) divide ethical concerns into four categories, with robustness classified under the ethical category. By utilizing Jobin et al.'s (2019) framework, we can ensure that our analysis of ethical considerations in GenAI is both comprehensive and relevant to the field.

To investigate our research question ("How are established AI ethics principles presented, and what new ethical principles have surfaced in the emerging literature on the ethical considerations of GenAI?"), we conducted a scoping review (Paré et al., 2015) and conceptual synthesis (cf. Torraco, 2005) of the relevant literature, identifying how ethical principles are discussed and placing them into a framework of core, supporting, and new principles (see Figure 1). The purpose of a scoping review is to provide a preliminary overview of the possible volume and characteristics of the existing literature on a specific topic; assess the scope, diversity, and nature of the research activities; and pinpoint gaps in the current literature (Paré et al., 2015). To this end, we began with a comprehensive literature search using Scopus, Web of Science, and Google Scholar. We employed the following search query to identify relevant articles: ("generative AI" OR "generative artificial intelligence" OR "large language models") AND ("ethics"). In line with the scoping review approach (Paré et al., 2015), we systematically examined the range of academic literature on AI ethics principles for GenAI using a set of inclusion and exclusion criteria,

selecting studies that outline specific lists of ethical principles for GenAI for our analysis. The inclusion and exclusion criteria are presented in Appendix A along with articles that discussed the ethics of generative AI. If such an article included a list of ethical principles or issues for GenAI, it was included in our conceptual synthesis, where the aim was to integrate existing ideas with new ideas to create a new understanding of the topic (Torraco, 2005). From these articles, we tabulated the mentioned ethical principles, which we then analyzed according to our research question to assess how the established AI ethics principles were reflected in the GenAI literature and to identify any potential new principles emerging within it. The search resulted in a total of 27 ethics-focused articles on GenAI.

4 Results

The results revealed partial overlaps with established AI ethics principles, as outlined by Jobin et al. (2019). However, the literature also highlighted unique ethical principles and challenges specific to GenAI. While the five core AI ethics principles have remained relevant, non-maleficence and privacy have received increased attention in the context of GenAI. Conversely, transparency is less emphasized than in the established AI ethics literature. Among the six supportive principles, sustainability has emerged as a more prominent concern, whereas freedom and beneficence are relatively less discussed. This indicates that while foundational AI ethics principles still apply, the evolving discourse on GenAI ethics introduces new dimensions that warrant further exploration.

Additionally, from the GenAI-specific literature, we identified new ethical principles that are crucial for this domain, such as respect for intellectual property and truthfulness, reflecting the unique challenges posed by GenAI vis-à-vis content generation and misinformation harms. Other emerging principles include robustness ethics; the practical aspects of GenAI deployment; recognition of malicious uses; the responsible utilization and potential deceptive applications of GenAI; sociocultural responsibility that highlights the impact of GenAI on diverse societal and cultural contexts; and human-centric design, which emphasizes the importance of GenAI systems and human interaction. To clarify, 'new' principles, in this context, do not mean entirely novel issues but rather issues that arise more prominently with GenAI compared to previous AI systems. These principles are presented in Figure 1 below.

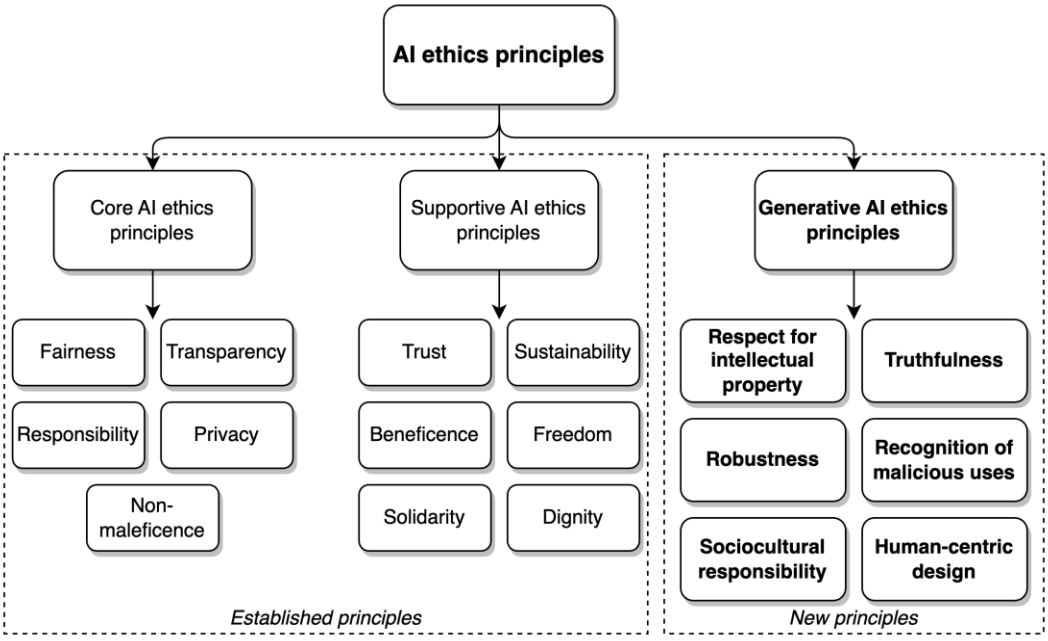

Figure 1. Established AI Ethics Principles and New Generative AI Ethics Principles

4.1 Established AI Ethics Principles

The emerging GenAI literature includes frequent discussions of AI ethics principles that have been established in the prior literature. To summarize these principles, we categorized the GenAI principles mentioned in the studies under the established AI ethics principles found in a review by Jobin et al.

(2019). Table 1 lists and describes the established AI ethics principles (Jobin et al., 2019) and the corresponding GenAI principles and issues. We discuss each principle's relevance, new challenges, and evolving interpretations as they are depicted in the GenAI literature. The aim is to highlight how the introduction of GenAI technologies further entrenches the importance of AI ethics principles while also reshaping our understanding of them.

Table 1. Established AI Ethics Principles and Corresponding Generative AI Principles and Issues

Established AI ethics principles (Jobin et al., 2019)	**Description**	**Corresponding generative AI principles and issues**
Justice & Fairness	GenAI systems should be designed, trained, and implemented to ensure equitable treatment of all users and groups, avoiding biased or discriminatory outcomes	**Bias and discrimination** (Dwivedi et al., 2023; Fui-Hoon Nah et al., 2023; Gupta et al., 2023; Heinke et al., 2024; Kenthapadi et al., 2023; Kirova et al., 2024; Parikh, 2023; Pressman et al., 2024; Sands et al., 2024; Sengar et al., 2024; Stahl & Eke, 2024; Wang et al., 2023; Weidinger et al., 2021; Weisz et al., 2023; Zhui et al., 2024), **fairness** (Cheng & Liu, 2023; Jovanovic & Campbell, 2022; Kajiwara & Kawabata, 2024; Rana et al., 2024; Raza et al., 2024; Stahl & Eke, 2024; Sun et al., 2024; Zlateva et al., 2024), **inclusion and exclusion** (Gupta et al., 2023; Parikh, 2023; Stahl & Eke, 2024), **justice** (Stahl & Eke, 2024), and **social sorting** (Stahl & Eke, 2024)
Transparency	GenAI systems should provide accessible and comprehensible information about their operations, including clarity about model behavior, system capabilities, and limitations, while balancing the need for sufficient disclosure with the avoidance of overexposure	**Lack of interpretability** (Heinke et al., 2024; Kenthapadi et al., 2023; Sengar et al., 2024; Wahde & Virgolin, 2021), **transparency** (Cheng & Liu, 2023; Dwivedi et al., 2023; Kirova et al., 2024; Parikh, 2023; Pressman et al., 2024; Rana et al., 2024; Sands et al., 2024; Sun et al., 2024; Zhui et al., 2024; Zlateva et al., 2024), and **explainability** (Dwivedi et al., 2023; Jovanovic & Campbell, 2022; Kajiwara & Kawabata, 2024)
Responsibility	GenAI systems and their operators should ensure that systems function as intended, prevent harm or discrimination, and establish accountability mechanisms for their development, deployment, and usage	**Accountability** (Cheng & Liu, 2023; Dwivedi et al., 2023; Jovanovic & Campbell, 2022; Kirova et al., 2024; Parikh, 2023; Rana et al., 2024; Stahl & Eke, 2024; Sun et al., 2024; Zlateva et al., 2024), and **social responsibility and integrity** (Guo et al., 2023; Kajiwara & Kawabata, 2024; Pressman et al., 2024; Sands et al., 2024; Zhui et al., 2024)

Privacy	GenAI systems should safeguard individuals' personal and private information by preventing unauthorized access, memorization of personally identifiable information, and violations of privacy rights	**Privacy** (Bale et al., 2023; Cheng & Liu, 2023; Dwivedi et al., 2023; Fui-Hoon Nah et al., 2023; Guo et al., 2023; Gupta et al., 2023; Kenthapadi et al., 2023; Kirova et al., 2024; Parikh, 2023; Pressman et al., 2024; Sands et al., 2024; Sengar et al., 2024; Sun et al., 2024; Wang et al., 2023; Zhui et al., 2024; Zlateva et al., 2024), **misuse of personal information** (Gupta et al., 2023), **information hazards** (Weidinger et al., 2021; Weisz et al., 2023), and **data governance** (Dwivedi et al., 2023)
Non-maleficence	GenAI systems should actively protect information from unauthorized access, corruption, theft, and intentional misuse, ensuring they do no harm	**Security** (Cheng & Liu, 2023; Dwivedi et al., 2023; Fui-Hoon Nah et al., 2023; Guo et al., 2023; Gupta et al., 2023; Kenthapadi et al., 2023; Mökander et al., 2023; Sengar et al., 2024), **intentional misuse** (Bale et al., 2023; Dwivedi et al., 2023; Fui-Hoon Nah et al., 2023; Parikh, 2023; Weisz et al., 2023), **safety** (Cheng & Liu, 2023; Guo et al., 2023; Gupta et al., 2023; Heinke et al., 2024; Stahl & Eke, 2024; Sun et al., 2024; Zlateva et al., 2024), **harms to society** (Kajiwara & Kawabata, 2024; Stahl & Eke, 2024), **psychological harm** (Stahl & Eke, 2024), and **harmful or inappropriate content** (Fui-Hoon Nah et al., 2023; Weidinger et al., 2021)
Beneficence	GenAI systems should be designed and deployed to bring positive outcomes to society, enhancing well-being and improving quality of life	**Beneficence** (Kajiwara & Kawabata, 2024; Stahl & Eke, 2024)
Trust	GenAI systems should be designed to inspire confidence in their reliability, integrity, and abilities, ensuring users feel secure in their interactions with them	**Trust** (Guo et al., 2023; Kenthapadi et al., 2023; Pressman et al., 2024; Sun et al., 2024), and **overreliance** (Fui-Hoon Nah et al., 2023)
Freedom & Autonomy	GenAI systems should respect individuals' rights to make their own choices, avoiding practices that limit personal freedom or force decisions upon users	**Freedom of expression** (Guo et al., 2023), **freedom of speech and expression** (Stahl & Eke, 2024), **autonomy** (Kajiwara & Kawabata, 2024; Rana et al., 2024; Stahl & Eke, 2024), and **informed consent** (Pressman et al., 2024; Stahl & Eke, 2024)
Sustainability	GenAI systems should be developed and operated with environmental responsibility in mind, minimizing ecological impacts and addressing risks to the planet	**Sustainability** (Stahl & Eke, 2024), **pollution and waste** (Stahl & Eke, 2024), **efficiency** (Jovanovic & Campbell, 2022; Raza et al., 2024), **automation and environmental Harms** (Stahl & Eke, 2024; Weidinger et al., 2021), and **resource intensiveness** (Zlateva et al., 2024)

Dignity	GenAI systems should be designed and deployed to respect human rights, prevent harm, avoid forced acceptance and automated classification, and ensure transparency in human–AI interactions	**Collective human identity and the good life** (Stahl & Eke, 2024), and **identity** (Stahl & Eke, 2024)
Solidarity	GenAI systems should be designed and deployed to support the equitable distribution of the benefits of AI in order not to threaten social cohesion and to respect potentially vulnerable persons and groups	**Social solidarity** (Stahl & Eke, 2024), and **supportive of vital social institutions and structures** (Stahl & Eke, 2024)

Fairness is defined as the ethical principle of ensuring that LLMs are designed, trained, and deployed in ways that do not lead to biased or discriminatory outcomes and that they treat all users and groups equitably (Sands et al., 2024; Sun et al., 2024). It is essential to ensure fairness in LLMs because it embodies the ethical principle regarding designing, training, and implementing LLMs and associated AI technologies in ways that avoid biased or discriminatory results (Raza et al., 2024; Sun et al., 2024; Wang et al., 2023). GenAI systems frequently exhibit bias, a concern highlighted in the literature owing to the nature of their training data (Kenthapadi et al., 2023; Parikh, 2023; Kirova et al., 2024; Sun et al., 2024). As GenAI's outputs can be only as accurate and unbiased as the training data on which its learning is based (Dwivedi et al., 2023; Sun et al., 2024), studies indicate that LLMs often mirror societal biases, such as gender and racial stereotypes, undesirable biases toward mentions of disability, and religious stereotypes found in their training datasets, and that they replicate these in interactions with users (Heinke et al., 2024; Kenthapadi et al., 2023; Stahl & Eke, 2024). This replication of bias can lead to the propagation of discrimination and harmful stereotypes or other skewed or prejudiced outputs, as observed in various applications of these technologies (Gupta et al., 2023; Parikh, 2023; Weisz et al., 2023; Zlateva et al., 2024). Additionally, the issue of exclusionary norms emerges when training data lack diversity or are predominantly monolingual, leading to cultural insensitivity and a lack of representation (Fui-Hoon Nah et al., 2023; Zhuo et al., 2023). The lack of fairness in an LLM can lead to serious social, ethical, and legal repercussions, especially as more countries require AI models to comply with principles of fairness and nondiscrimination (Sun et al., 2024). GenAI models are typically trained on massive amounts of unfiltered online text from the Internet; consequently, they often reflect the practices of the wealthiest communities and countries (Jovanovic & Campbell, 2022; Kenthapadi et al., 2023). Addressing these biases is essential and requires diversifying training data, ensuring cultural relevance, and implementing robust moderation mechanisms (Dwivedi et al., 2023; Weidinger et al., 2021; Zhui et al., 2024). Moreover, the challenge of ensuring fairness and minimizing bias in GenAI systems calls for the continuous evaluation and adaptation of ethical practices, and GenAI providers should offer tools for preprocessing and curating training data, monitoring and moderating media generation processes, and developing guidelines for responsible deployment models (Jovanovic & Campbell, 2022).

Transparency refers to how much information about LLMs and their outputs is available to the individuals who interact with them (Sun et al., 2024). In regard to the transparency of GenAI systems, significant challenges arise from issues of interpretability and trust, particularly owing to these systems' large size and opaque behaviors, and potential for emergent, unintended capabilities (Heinke et al., 2024; Kenthapadi et al., 2023; Pressman et al., 2024). This lack of clarity is exacerbated by the possibility of GenAI models being used to generate misleading content, posing significant societal risks (Kenthapadi et al., 2023; Parikh, 2023). The lack of transparency makes it difficult to understand the reasons behind the system's responses (Dwivedi et al., 2023; Kajiwara & Kawabata, 2024). It is also difficult to evaluate, as not every situation requires the same level of transparency (Sun et al., 2024). There are also considerations between responsible transparency and unnecessary over-disclosure (Sands et al., 2024). Addressing these issues, some studies propose the use of interpretable primitives and clear processing stages to enhance understandability and trust in AI systems (Wahde & Virgolin, 2021). Moreover, the "black box" nature of these models often hinders their explainability (Cheng & Liu, 2023; Jovanovic & Campbell, 2022). It may be essential to implement additional measures of explainability, such as traceability, the ability to audit, and transparent communication about the capabilities of AI systems (Cheng & Liu, 2023) as well as mechanisms of a deep neural network, to determine which affect the outcomes, infer what happens inside the black box, and identify effective methods of elucidating AI decision-making processes (Jovanovic & Campbell, 2022). Also, additional model documentation

guidelines, implementing interpretative interfaces or tools and regular public evaluations and feedbacks could be beneficial (Zhui et al., 2024).

Responsibility, often referred to as accountability, in GenAI systems is a complex principle requiring organizations to ensure that their AI systems function as intended without causing harm or discrimination (Parikh, 2023; Rana et al., 2024). The lack of transparency often leads to a lack of accountability (Pressman et al., 2024; Sun et al., 2024). Accountability is primarily concerned with determining which subjects should be accountable, which aspects, and the manner and degree to which mechanisms of accountability should be established (Cheng & Liu, 2023). This accountability extends beyond immediate impacts and requires that organizations that are creating, training, and deploying GenAI systems reduce model behavior risks and be thorough, transparent, and proactive in communicating identified threats, blind spots, and areas in which risks are unknown while emphasizing the benefits of GenAI systems (Jovanovic & Campbell, 2022). Studies highlight the need for organizations to take responsibility not only for the development and deployment of AI systems but also for addressing adverse effects through notification, rectification, and redress mechanisms (Cheng & Liu, 2023). The ethical and legal challenges posed by the use of GenAI content, particularly in academia and research, underscore the importance of maintaining standards and integrity and question the ethicality of using AI-generated content without acknowledgment (Dwivedi et al., 2023). Moreover, the misuse of generative models to spread false or harmful information brings moral responsibility issues to the forefront, balancing freedom of expression with societal ethics (Guo et al., 2023). The potential impact of GenAI systems on vulnerable groups and the broader implications for worldviews and personal values highlight the urgency of ensuring ethical guidance in their application and better monitoring systems, investigation committees and feedback channels (Guo et al., 2023; Kajiwara & Kawabata, 2024; Zhui et al., 2024; Zlateva et al., 2024).

Privacy concerns in GenAI systems, particularly in LLMs and image diffusion models such as DALL-E 2, highlight the significant risks of memorizing and reproducing personally identifiable information, leaking private information, and posing threats and violations to individual privacy rights and copyright (Bale et al., 2023; Fui-Hoon Nah et al., 2023; Kenthapadi et al., 2023; Sun et al., 2024). Adherence to privacy regulations such as General Data Protection Regulation (GDPR) is critical, as GenAI systems require extensive data for training, increasing the risk of privacy violations (Gupta et al., 2023; Parikh, 2023; Weidinger et al., 2021; Zhui et al., 2024). Privacy hazards include privacy violations and safety risks, leaking private information, correctly inferring private information, and correctly inferring sensitive information (Sengar et al., 2024; Weidinger et al., 2021; Zlateva et al., 2024). These hazards arise from the ability of the language model (LM) to predict statements that contain private or safety-critical information that is present in, or can be inferred from, training data (Fui-Hoon Nah et al., 2023; Weidinger et al., 2021). Incidents involving the misuse of LLMs, such as in corporate settings in which confidential information becomes part of the AI's training data and user information is leaked, exemplify the acute privacy challenges in deploying these technologies (Fui-Hoon Nah et al., 2023; Gupta et al., 2023; Pressman et al., 2024). For example, on receiving the right prompts, GPT-2 revealed an individual's full name, work address, phone number, email, and fax number (Weisz et al., 2023). With the ongoing advancement of GenAI, it is anticipated that these systems will become more accurate and secure, providing users with a more reliable and safer experience. For instance, ChatGPT provides users with the option to opt out of contributing their data for training purposes (Gupta et al., 2023). However, reports indicate that owing to system glitches in ChatGPT, the chat logs of certain users have become accessible to others, and information privacy and security concerns are affecting not only individual users but also large corporations and government agencies (Fui-Hoon Nah et al., 2023). To resolve some of these issues, the LLM could implement solutions such as not saving a user's chat history, adhering to company policies, and providing the option to delete messages from the LLM's history (Fui-Hoon Nah et al., 2023). Additionally, users need to be circumspect when interacting with ChatGPT, while AI companies, especially technology giants, should take appropriate actions to increase user awareness of ethical issues surrounding privacy and security (Gupta et al., 2023).

Non-maleficence, a fundamental ethical principle, is a significant concern in GenAI systems, given their complexity, potential impact, and vulnerability to adversarial behaviors such as jailbreaking (using specific prompts to bypass security measures) (Sun et al., 2024). Non-maleficence consists primarily of data security, which refers to the practice of protecting information from unauthorized access, corruption, theft (Fui-Hoon Nah et al., 2023), and intentional misuse. These describe how GenAI applications may be misused in ways that are unanticipated by the creators of these systems—for instance, to generate malevolent material, including spam, fraudulent reviews, or even cyberattacks on a large scale (Bale et al., 2023; Sengar et al., 2024; Weisz et al., 2023). The robustness and security of models such as LLMs

are critical, as vulnerabilities—for example, data poisoning attacks and the inability to provide uncertainty estimates—can lead to unreliable and potentially harmful outputs (Kenthapadi et al., 2023). This risk is further heightened by the dual-use nature of GenAI in cybersecurity, wherein it can be used for both defense and malicious purposes, such as creating deceptive phishing emails and malicious code (Dwivedi et al., 2023; Gupta et al., 2023). The governance of Internet information content, particularly in the context of data leakage and cross-border data security, presents complex challenges in regard to ensuring the ethical use of GenAI (Cheng & Liu, 2023). Security of privacy is also paramount, especially with regard to models such as ChatGPT, as the misuse of personal and sensitive data can lead to significant privacy violations (Fui-Hoon Nah et al., 2023; Guo et al., 2023). Additionally, the potential for GenAI to produce harmful or inappropriate content, such as toxicity and misinformation, raises ethical dilemmas about its impact on society and individuals (Fui-Hoon Nah et al., 2023; Weidinger et al., 2021). Addressing these challenges necessitates a multifaceted approach that includes robust auditing, adherence to ethical guidelines, and comprehensive governance strategies to mitigate potential harms and ensure the responsible deployment of GenAI technologies (Cheng & Liu, 2023; Fui-Hoon Nah et al., 2023; Parikh, 2023). It is essential to ensure the exclusion of harmful or offensive content from training data (Fui-Hoon Nah et al., 2023). To achieve this, developers and regulators must remain vigilant, constantly upgrading security measures and implementing stringent content-filtering algorithms (Gupta et al., 2023), and strict verification mechanisms are necessary to prevent misuse (Parikh, 2023). Furthermore, watermarking output images to show their model origin, employing blocklists to prevent unwanted words in text prompts, and necessitating multiple reviews or approvals for a model's outputs before usage (Weisz et al., 2023) are all effective strategies.

Among the supportive principles, *sustainability* received the most attention. The sustainability of GenAI technologies, such as ChatGPT, involves several ethical concerns, as several environmental risks emerge during or before training (Weidinger et al., 2021). Environmental impact is a significant issue, as the energy demands for training and operating these models contribute to resource depletion and pollution, leave a significant carbon footprint, and have high computation costs (Jovanovic & Campbell, 2022; Stahl & Eke, 2024). This problem is exacerbated when energy sources are nonrenewable (Weidinger et al., 2021). Furthermore, many environmental factors related to the operation of LMs that are in widespread use are currently unknown (Weidinger et al., 2021). These factors include how an LM will be integrated into products, anticipated scale and frequency of use, and energy cost per prompt (Weidinger et al., 2021). Additionally, when LMs are trained using energy sourced from fossil fuels, their training and operation inadvertently support an industry that is responsible for significant environmental harm. To mitigate these environmental impacts, strategies such as improving hardware efficiency, implementing carbon offset programs, creating new model architectures and data processing methods, and using renewable energy sources can be considered (Raza et al., 2024; Weidinger et al., 2021).

The opaque nature of LLMs raises concerns about trust and interpretability (Sun et al., 2024). Users may struggle to validate AI-generated information, which could include misleading or false content, impacting trust in digital platforms and online interactions (Kenthapadi et al., 2023). Its convenience and efficiency could also lead to overreliance, as users might accept AI-generated answers without critical assessment or verification. This dependency risks eroding skills such as creativity and critical thinking, fostering a human automation bias (Fui-Hoon Nah et al., 2023). The misuse of AI technologies can further undermine societal trust, affect social security, and destabilize trust in businesses, especially if their use of AI is not transparent to consumers (Guo et al., 2023). The unique characteristics of LLMs, such as the complexity and diversity of outputs, data biases and private information in large training datasets, and high user expectations, can all lead to potential trust issues (Sun et al., 2024). Consequently, it is essential to address benchmarking issues regarding the trustworthiness of LLMs and the key elements that define it. Key components of these issues are the definitions of comprehensive aspects, scalability and generalizability, and practical evaluation methodologies (Sun et al., 2024).

In established AI ethics principles, freedom and autonomy, beneficence, dignity, and solidarity received the least attention. However, these models' production of misleading, biased, or harmful content challenges the principles of human morality and social responsibility. Although freedom of expression is vital, its conflict with societal ethics becomes apparent when AI is used to spread false information, potentially exacerbating discrimination and social inequality (Kajiwara & Kawabata, 2024). The indiscriminate acceptance of AI-generated biases can undermine ethical standards in AI (Guo et al., 2023). Moreover, the use of ChatGPT in sensitive contexts, especially for vulnerable individuals seeking support or guidance, raises concerns about the appropriateness and ethical implications of the advice given, highlighting the need for careful consideration of the impact of GenAI on personal autonomy and

freedom (Guo et al., 2023; Rana et al., 2024). In the context of GenAI, such as ChatGPT, the principles of beneficence, dignity, and solidarity manifest in various ways. Beneficence emphasizes the technology's potential to contribute positively to society, such as by enhancing communication and aiding in decision-making processes, and satisfy individual wants and needs (Kajiwara & Kawabata, 2024). Dignity relates to ensuring that AI interactions respect human values and cultural diversity, avoiding the perpetuation of harmful stereotypes and biases. Solidarity underscores the importance of these technologies in fostering social cohesion and understanding, particularly by bridging communication gaps and serving diverse user needs. However, these benefits must be balanced against ethical concerns, such as potential misuse, privacy issues, and inequitable access to technology (Stahl & Eke, 2024).

4.2 New Ethics Principles for Generative AI

While the established AI ethics principles remain important, our review of the GenAI ethics literature revealed several new principles that are particularly salient for GenAI systems and new in relation to the AI ethics principles discussed before GenAI. These new principles address the characteristics of GenAI, such as its ability to generate humanlike text and images and its general-purpose nature, which allows for wide-ranging applications. We identified and categorized the principles that did not fit under the established AI ethics principles. Hence, we derived six new principles for AI, indicating areas in which AI ethics principles may need to be expanded. Table 2 lists and describes the new GenAI principles and the corresponding principles mentioned in the reviewed literature. This table serves as a guide to our subsequent discussion of each principle and its significance in the GenAI ethics literature.

Table 2. New Ethics Principles for Generative AI

New principle for generative AI	Description	Corresponding generative AI principles and issues
Respect for intellectual property	GenAI systems and their operators should recognize and protect the rights and ownership of creators over their original works and innovations, ensuring proper attribution and compensation and avoiding unauthorized use of protected works	**Intellectual property** (Bale et al., 2023; Sands et al., 2024; Stahl & Eke, 2024; Weisz et al., 2023; Zlateva et al., 2024), **copyright** (Kenthapadi et al., 2023; Kirova et al., 2024; Weisz et al., 2023; Zhui et al., 2024), and **ownership** (Gupta et al., 2023; Heinke et al., 2024; Pressman et al., 2024; Stahl & Eke, 2024)
Truthfulness	GenAI systems should be designed and used in such ways that they accurately represent information, facts, and outcomes	**False information** (Bale et al., 2023; Guo et al., 2023; Zlateva et al., 2024), **misinformation harms** (Gupta et al., 2023; Sengar et al., 2024; Weidinger et al., 2021; Weisz et al., 2023), **reliability** (Pressman et al., 2024; Raza et al., 2024; Zhuo et al., 2023), **fake news and informational inflation** (Dwivedi et al., 2023), and **truthfulness** (Mökander et al., 2023; Sun et al., 2024; Zlateva et al., 2024), **hallucination** (Heinke et al., 2024; Kirova et al., 2024; Pressman et al., 2024; Zhui et al., 2024), and **accuracy** (Pressman et al., 2024; Rana et al., 2024)
Robustness	GenAI systems should consistently deliver high performance across various scenarios while efficiently handling exceptions, anomalies, and unexpected inputs	**Robustness** (Kenthapadi et al., 2023; Mökander et al., 2023; Parikh, 2023; Sun et al., 2024; Wang et al., 2023; Zhuo et al., 2023), **anticipation** (Parikh, 2023), **reflexivity** (Parikh, 2023), **responsiveness** (Parikh, 2023), **recency** (Heinke et al., 2024), **novelty** (Dwivedi et al., 2023), and **performance** (Mökander et al., 2023)

Recognition of malicious uses	GenAI systems should be designed with robust safeguards against potential misuse, identifying and preventing negative or harmful applications	**Honesty and prejudice** (Bale et al., 2023), **manipulation** (Bale et al., 2023; Guo et al., 2023; Weisz et al., 2023), **human morality violation** (Guo et al., 2023), **impersonation** (Weisz et al., 2023), **toxicity** (Wang et al., 2023; Weidinger et al., 2021; Weisz et al., 2023; Zhuo et al., 2023), **malicious use** (Weidinger et al., 2021), and **information ethics** (Guo et al., 2023)
Sociocultural responsibility	GenAI systems should be designed and used in such ways that they operate responsibly with regard to a wide array of consequences, from unforeseen repercussions to shifts in power structures and cultural acceptance	**Unforeseen repercussions** (Bale et al., 2023), **democracy** (Stahl & Eke, 2024), **labor market** (Stahl & Eke, 2024; Zlateva et al., 2024), **cultural differences** (Raza et al., 2024; Stahl & Eke, 2024), **universal service** (Stahl & Eke, 2024), **collective human identity and the good life** (Dwivedi et al., 2023), **effect of culture and personal values** (Dwivedi et al., 2023), and **structures of power** (Dwivedi et al., 2023)
Human-centric design	GenAI systems should be designed in such a way that they prioritize the needs, well-being, and experiences of human users	**Design of prompts** (Dwivedi et al., 2023), **dependence on technology** (Dwivedi et al., 2023; Zlateva et al., 2024), **lack of originality** (Dwivedi et al., 2023), **irreversibility** (Dwivedi et al., 2023), **human–computer interaction harms** (Mökander et al., 2023; Weidinger et al., 2021), **inherent capability to explain** (Wahde & Virgolin, 2021), **independent data** (Wahde & Virgolin, 2021), **interactive learning** (Raza et al., 2024; Wahde & Virgolin, 2021), **quality control** (Pressman et al., 2024; Zlateva et al., 2024), **inquisitiveness** (Wahde & Virgolin, 2021), **machine ethics** (Sun et al., 2024; Wang et al., 2023), **over-reliance** (Zlateva et al., 2024), and **emotional intelligence** (Zhui et al., 2024)

While AI technology advances, *intellectual property concerns* are becoming increasingly significant. Copyright concerns related to technologies have, of course, been extensively discussed, and Mason (1986) identified 'property' as one of the four central issues of the information age. However, AI ethics principles before the rise of GenAI have said little about copyright (e.g., Jobin, 2019). In the GenAI context, authors have the fundamental right to recognition and proper attribution, which upholds their autonomy and ensures appropriate acknowledgment of their intellectual property (Pressman et al., 2024; Zlateva et al., 2024). These issues may arise, for example, when the distinction between original and AI-generated content is blurred (Bale et al., 2023). GenAI models are often trained on datasets that might fall within regulatory frameworks such as GDPR, which restricts the reuse of data beyond their original collection purpose (Kirova et al., 2024). These models have the capability to mimic and potentially regenerate content from their training data. This raises concerns, as the output generated by such models could unintentionally include or remix material that is protected by copyright or other intellectual property rights (Sands et al., 2024; Weisz et al., 2023). Moreover, LLMs and image diffusion models, such as DALL-E2 and Imagen, have shown a tendency to memorize and reproduce personally identifiable information and individual images from training data, resulting in not only privacy concerns but also potential copyright violations (Kenthapadi et al., 2023). The issue of the ownership and control of

technology also plays a critical role in the ethical evaluation of GenAI. The profit-oriented ownership of technology such as ChatGPT raises questions about who holds the rights and controls the use of AI-generated content (Stahl & Eke, 2024). Suggested solutions could be, for example, strict IP compliance processes, copyright statements and terms of use, and monitoring and detection systems (Zhui et al., 2024).

Another new principle is *truthfulness*. Accuracy and the avoidance of misinformation have been discussed as central to information ethics already in the 1980s (Mason, 1986), but in the context of AI, truthfulness has become more relevant to content-producing GenAI compared to previous predictive AI technologies. LLMs, while predicting likely sentences, do not always ensure factual accuracy, leading to risks of false or nonsensical outputs (Pressman et al., 2024; Raza et al., 2024; Weidinger et al., 2021). Truthfulness refers to the accurate representation of information, facts, and outcomes (Sun et al., 2024; Zlateva et al., 2024). In our context, truthfulness is referred to as inaccuracies that are not created by malicious users with harmful intent. GenAI models, including ChatGPT, risk blurring the line between fact and fiction, as they can rapidly disseminate false or misleading information, fake news, and malicious content, making it difficult for users to discern truth from fantasy (Bale et al., 2023; Guo et al., 2023; Pressman et al., 2024; Zlateva et al., 2024). These types of misinformation harms, also referred to as hallucinations, are problematic, as LLMs may provide plausible but factually incorrect information, potentially leading to misinformation in sensitive areas, such as healthcare and legal advice (Heinke et al., 2024; Weidinger et al., 2021; Weisz et al., 2023). Without human supervision, the ability to produce fake news and misinformation becomes significantly easier and faster (Dwivedi et al., 2023). Although factually incorrect or hallucinated predictions can be harmless, there are many types of harms, ranging from misinforming, deceiving, or manipulating a person; to causing material harm; to broader societal repercussions, such as decreasing trust (Weidinger et al., 2021). LLMs may produce false statements owing to their training on diverse and often factually incorrect web content, their inability to contextually discern the truth, and their fundamental limitation of not being grounded in real-world experience (Mökander et al., 2023; Weidinger et al., 2021). Previous studies have used datasets and benchmarks, such as MMLU (Measuring Massive Multitask Language Understanding), Natural Questions, TriviaQA, and TruthfulQA, to critically examine LLMs' commitment to accuracy. Additionally, certain tools have been developed to evaluate specific facets of overall truthfulness: HaluEval focuses on fabrications, SelfAware investigates the models' recognition of their own knowledge limitations, and FreshQA and Pinocchio analyze the models' ability to adjust to rapidly changing information (Sun et al., 2024). Ensuring the reliability and truthfulness of GenAI-generated information necessitates rigorous verification mechanisms, continuous updates to training data, and adherence to ethical standards to prevent the spread of inaccurate content (Mökander et al., 2023; Zhuo et al., 2023).

Robustness considerations in GenAI are crucial for its ethical and effective deployment. Previously in AI ethics, robustness has been treated as a separate dimension parallel to ethics (e.g., High-Level Expert Group on Artificial Intelligence, 2019), but the ethical dimensions related to robustness become increasingly apparent with GenAI. Robustness in ML refers to a model's ability to maintain consistent performance when faced with semantically or syntactically diverse input, including new or unexpected out-of-domain data, and to properly manage exceptions, anomalies, and unexpected inputs (Mökander et al., 2023; Sun et al., 2024; Zhuo et al., 2023). The current consensus in the literature is that robustness is not an inherent quality of current LLMs (Sun et al., 2024). LLMs often struggle with uncertainty estimates and are susceptible to data poisoning, impacting user trust in their outputs (Kenthapadi et al., 2023). Additionally, the tendency of models to hallucinate or produce incorrect information necessitates ongoing human oversight for accuracy (Parikh, 2023; Zhuo et al., 2023). Operational effectiveness in GenAI also involves the anticipation of benefits and risks; responsive adaptations to emerging issues; and reflexivity in decision-making processes, particularly in addressing the "black box" problem and output accuracy (Parikh, 2023). Lack of robustness comes with at least three kinds of risks: critical system failures, data leakage, and prompt injection (Mökander et al., 2023; Zhuo et al., 2023). System failure in LMs can occur because of semantic perturbations, whereby input with different syntax but similar meaning to the training data leads to errors, and from risks such as poor performance for underrepresented groups (Mökander et al., 2023; Zhuo et al., 2023). Data leakage within LMs may lead to vulnerability to attacks, whereby adversaries aim to retrieve confidential information, thereby threatening personal privacy and the security of organizations (Zhuo et al., 2023). Prompt injection involves deliberately inserting specific data into the input of an LM with the aim of causing it to malfunction or fail (Mökander et al., 2023). To mitigate adverse outcomes, it is vital to use diverse, representative, and secure training data; reduce the lack of recency in training data; employ robust testing and monitoring methods; and use security measures, such as

differential privacy (Heinke et al., 2024; Mökander et al., 2023; Zhuo et al., 2023). Additionally, employing comprehensive tools that evaluate LLMs' vulnerabilities to adversarial attacks in different domains and standardized benchmarks that can help assess an LLM's performance by comparing it to a human baseline can help assess and improve the model's robustness and performance across various domains and tasks (Mökander et al., 2023).

The recognition of malicious uses of GenAI encompasses considerations such as honesty, prejudice, manipulation, toxicity, and the generation of harmful content. GenAI can be exploited for manipulative purposes, such as the generation of propaganda or misinformation, thereby influencing public opinion and potentially harming, for example, the electoral process or other fraud and scams and enabling more effective cyberattacks and surveillance (Bale et al., 2023; Guo et al., 2023; Weidinger et al., 2021). The use of generative models to create false or misleading content (including deepfakes) could further complicate matters, as it can violate principles such as freedom of expression and information ethics (Guo et al., 2023; Weidinger et al., 2021; Weisz et al., 2023). Toxicity, another ethical concern, refers to the model's ability to generate or understand harmful or offensive content (Zhuo et al., 2023). Although there is no universally accepted definition of what qualifies as hate speech or toxic speech, proposed definitions often include profanity, identity attacks, sleights, insults, threats, sexually explicit content, demeaning language, language that incites violence, or hostile and malicious language targeted at a person or group (Weidinger et al., 2021). One type of toxicity that might occur involves offensive language or pornographic content in the training dataset (Zhuo et al., 2023). This issue can lead to the model producing or recognizing offensive or harmful material during interactions with users, causing offense, psychological harm, and even material harm (Weidinger et al., 2021; Zhuo et al., 2023). To mitigate malicious uses of LMs, it is essential to monitor the training data to exclude offensive, harmful, or false content and to implement robust security measures to prevent the model from being exploited to generate disinformation, scams, or other harmful outputs (Weidinger et al., 2021; Zhuo et al., 2023). Additionally, ongoing research and development of advanced detection and filtering algorithms are necessary to identify and neutralize potential misuse, such as the generation of synthetic media or fraudulent content (Weidinger et al., 2021).

The social and cultural impact of GenAI encompasses a wide array of consequences, from unforeseen repercussions to shifts in power structures and cultural acceptance. This technology can potentially deepen digital divides as different cultures and organizations vary in their readiness to adopt AI (Dwivedi et al., 2023). Enhancing the diversity of language models and the representation of cultural datasets is essential (Raza et al., 2024). While GenAI offers potential benefits by enhancing societal communication, its application also raises concerns about social justice, rights, individual needs, and environmental impacts. These considerations span high-level societal issues, including the potential to influence social relationships, responsibility, and accountability (Stahl & Eke, 2024). However, the negative effects, such as the fact that it exacerbates existing social and digital divides and favors already advantaged groups, are significant concerns. This disparity is particularly pronounced in the ways in which emerging digital technologies unfold, whereby the advantages are likely to benefit more, perpetuating existing inequalities (Stahl & Eke, 2024).

Human-centric design in GenAI systems must account for diverse factors, including prompt design, machine ethics, technology dependence, originality, human–computer interaction, and societal impact. Human–computer interaction harms arise from users having too much trust in the LM, overestimating the capabilities of LLMs, or treating the LM as humanlike and using it in unsafe ways (Mökander et al., 2023; Weidinger et al., 2021). Effective prompt design is critical because it determines output quality and requires a new kind of skill for people to be able to use prompts effectively (Dwivedi et al., 2023). Moreover, the inherent limitations in the originality of AI-generated content highlight the need for human creativity (Dwivedi et al., 2023). A conversational agent should also be able to explain its reasoning in a nontechnical manner while actively seeking information and should thus expand its capabilities during interaction with the user (Wahde & Virgolin, 2021). Interactive learning, whereby a person teaches a machine new capabilities, should be a primary approach when building conversational agents, and an agent's memory should be divided into procedural memory, which encodes the dialogue capabilities, and declarative memory, which contains the facts that are known to the agent (Wahde & Virgolin, 2021). This type of interactivity that requires adaptive learning algorithms would improve personalization and user experience as the algorithm is learning from user interactions (Raza et al., 2024). The term machine ethics refers to ethics for which machines, instead of humans, are the subjects (Sun et al., 2024). It can be divided into implicit ethics, explicit ethics, and emotional awareness. Implicit ethics considers the internal values of LLMs, such as the judgment of moral situations; explicit ethics focuses on how LLMs should react in different moral environments; and emotional awareness considers LLMs' capacity to recognize

and empathize with human emotion (Sun et al., 2024). Examples of possible harms that could come from human-centric design are psychological vulnerabilities, risks from users anthropomorphizing such technologies, risks that could arise via the recommendation function of conversational technologies, and risks of representational harm (Weidinger et al., 2021). Anthropomorphizing systems can lead to overreliance or unsafe use and shift accountability from the developers of this technology onto the agent itself. Users may also reveal private information, as the risk of this is greater when they treat models as if they are human and promote harmful stereotypes (Weidinger et al., 2021). Moreover, conversational agents can perpetuate harmful gender and ethnic stereotypes by embodying submissive and gender-specific roles or implying certain ethnic identities through their design, vocabulary, and interaction style (Weidinger et al., 2021).

5 Discussion and Conclusion

5.1 The Landscape of AI Ethics Principles

Our review of the GenAI ethics principles literature brought forth three sets of findings: the changed context of established AI ethics principles, the newly emerging GenAI ethics principles, and, extending from these, the challenge of satisfying multiple principles simultaneously. Among the established AI ethics principles, non-maleficence, and privacy take on new importance owing to the general-purpose nature of GenAI systems. Rather than being designed for specific contexts, such as hiring algorithms, these systems can be used for a multitude of purposes in widely different settings. This means that the range of potential risks is also broader because systems cannot be tested in controlled environments for each of their foreseeable use cases. Moreover, GenAI systems can rapidly produce harmful or misleading content at scale, and privacy breaches can be continuously attempted by creative prompting, heightening the need for robust safeguards. Owing to these risks and concerns, responsibility and accountability also take on heightened importance. Although transparency remains an important principle, attention to it has somewhat diminished in the GenAI literature compared to its prominence within previous AI ethics guidelines (Jobin et al., 2019). This may be a reflection of the increasingly complex and opaque nature of GenAI systems, which are primarily produced by large technology companies and thus inscrutable for individual and organizational users. Hence, transparency is even more challenging as a goal than with narrower decision-making AI systems. At the same time, environmental sustainability assumes greater importance because GenAI systems are open to public use and are used on a massive scale, contributing to rapidly rising environmental and energy concerns. The high computational costs and energy requirements of training and operating GenAI models remain issues that need large-scale solutions.

Moving on to the newly introduced GenAI ethics principles, we proposed six new ones based on our literature review. While they are all important, we highlight and elaborate on three (intellectual property, truthfulness, and human-centric design) which are new additions to the AI ethics discussion stemming from the features of GenAI systems. First, respect for intellectual property has led to a significant outcry among creative professionals owing to the seeming indifference of GenAI-developing companies toward intellectual property. While the basis of intellectual property has been challenged before owing to advancing technologies ranging back to video cassettes and Napster (Riemer & Johnston, 2019), GenAI systems are introducing intellectual property challenges on a new scale. Second, truthfulness emerges as a new principle because GenAI systems can make convincing claims or produce images that have the appearance of truth but are misleading or deceptive. Truthfulness is a challenging principle to adhere to because it is connected to the system's surrounding reality, to which the system has no access beyond its training data and possible Internet access. Third, human-centric design has been important in decision-making AI systems, including in IT systems generally. GenAI systems extend the importance of human-centric design because they can discuss with humans in a human-like manner, sometimes even being able to pose as humans. The ability to operate in a humanlike manner raises new ethical questions related to the design of these systems—for instance, whether systems such as technical support chatbots should be designed to closely mimic humans or whether their nature as software systems should be made visible. In general, ethical guidelines have indicated that users should be aware of whether they are interacting with an AI system (e.g., High-Level Expert Group on Artificial Intelligence, 2019). The challenge with GenAI ethics principles, as with ethics principles in general, is that GenAI systems need to satisfy all, or at least multiple, principles simultaneously. As we mentioned in section 2.2, adhering to certain principles may be to the detriment of others, for example when added transparency means lesser privacy. Considering that we have identified altogether 17 principles in this paper, seeking to satisfy them at the same time could lead to complex trade-offs and balancing efforts, especially considering that trade-

offs may also depend on the particular context and use case. How could such tensions between principles, then, be resolved? While extensive discussion of resolving ethics principles is left to future research, we suggest three meta-principles for dealing with the multitude of ethics principles. First, after identifying the GenAI ethics principles, they should be categorized or ranked so that the fundamental principles can be distinguished from important but less crucial principles. Second, pairs of mutually contradicting principles, such as transparency and privacy, should be identified and each pair should be critically examined to concretely understand the nature of the contradiction between them, conceptually and also technically, when possible. Third, the achievement of the fundamental principles needs to be continuously monitored (cf. Minkkinen et al., 2022) due to the continuous updating of GenAI systems and the ever-broadening scope of their use, which may yield unexpected ethical dilemmas.

5.2 Implications for Research and Practice

Our study contributes to two streams of IS literature: the implications of AI systems and the ethics-based auditing of AI. The first area of contribution is the discussion on the ethical and societal implications of AI systems (Ågerfalk et al., 2022; Mirbabaie et al., 2022) and GenAI systems (Alavi et al., 2024; Benbya et al., 2024; Sabherwal & Grover, 2024), which are part of the broader discussion on the "dark side" of AI (Mikalef et al., 2022). Within this domain, previous IS literature has examined the implications of AI and algorithmic systems on organizational practices (Ågerfalk et al., 2022), working life (Cameron et al., 2023), and research (Davison et al., 2023, Davison et al., 2024; Loebbecke et al., 2020), and it has explored concepts and tools within specific ethical domains, such as bias (Kordzadeh & Ghasemaghaei, 2022) and explainability (Laato et al., 2022). In turn, other fields have produced numerous overviews of AI ethics principles (Hagendorff, 2020; Jobin et al., 2019). Attempts have also been made to synthesize the AI ethics discussions in IS and find their roots, yielding a set of principles that closely mirror those of previous IS ethics but with no synthesizing ethical framework (Mirbabaie et al., 2022). However, while LLMs, including their ethical implications, have been identified as a key topic (Schneider et al., 2024), the study of ethical issues and principles specific to LLMs is still in its infancy. We contribute to this emerging literature domain with an evidence-based overview of what is new about LLMs from the perspective of ethical principles. We present a list of new ethical principles, highlighting the importance of intellectual property, truthfulness, and human-centric design.

The second literature stream to which our study contributes is the ethics-based auditing of AI and specifically LLMs (Brown et al., 2021; Laine et al., 2024; Mökander, 2023; Mökander et al., 2021; Mökander et al., 2023), which are closely linked to the literature on AI governance (Birkstedt et al., 2023; Mäntymäki et al., 2022; Schneider et al., 2023). The AI auditing literature represents a crucial step toward the practical implementation of AI ethics principles and the assessment of adherence to them. Thus far, the literature has advanced conceptual frameworks for ethics-based auditing (Brown et al., 2021; Mökander et al., 2021), surveyed the understanding of principles and stakeholders in auditing studies (Laine et al., 2024), and proposed starting points for auditing LLMs (Mökander et al., 2023). However, the extant AI auditing literature has remained silent on what is new about GenAI compared to AI auditing, regarding new perspectives on established principles and the inclusion of new ethical principles. Hence, our study provides the foundation for developing accessible ethics-based auditing approaches by explicating the underlying ethical principles to be audited.

The practical implications of our study concern three levels: individual users, organizations, and the wider society. The review of GenAI ethics principles helps individual users understand the ethical issues that are at play when they are interacting with GenAI systems in work environments or everyday life. Organizations, in turn, can use the proposed set of GenAI principles to draft and update their AI ethics guidelines in the context of the rapidly advancing use of GenAI. Moreover, the heightened importance of non-maleficence, privacy, and sustainability and the new principles of intellectual property, truthfulness, and human-centric design have implications for GenAI development and deployment. In particular, these salient principles suggest the need for robust content-filtering mechanisms and built-in fact-checking capabilities for GenAI applications. Moreover, they reinforce the importance of interdisciplinary collaboration between engineers and ethicists in AI development. The general-purpose nature of GenAI systems, in turn, is likely to require increasingly flexible and adaptive governance frameworks to address the broad range of AI applications in organizations. Concerning the wider society, our study contributes to more nuanced public discussion and ultimately more ethically aligned design and use of GenAI systems by making visible the ethical fault lines inherent in the widespread use of GenAI systems within society and delineating the multitude of principles involved.

In particular, tensions between principles and the simultaneous addressing of multiple ethics principles is a challenge at all these levels (individuals, organizations, and society). On the level of individuals, it means making difficult choices regarding, e.g., privacy when using GenAI systems and, thus, providing individuals the necessary guidance and heuristic tools for making these choices. For organizations, tensions between principles mean establishing robust AI governance processes to manage trade-offs and ethical alignment between principles, objectives, and practices. On the societal level, regulating GenAI technologies under the pressure of multiple ethical principles is an ongoing challenge. Potential approaches to resolving ethical tensions range between the equally infeasible extremes of banning GenAI altogether and permitting all uses with no restrictions. Viable regulatory pathways are likely to be found between these extremes, as a somewhat messy combination of approaches of binding regulation and self-regulation, also taking into account the complex landscape of various sector-specific regulations. The meta-principles identified at the end of section 5.1 may be of use when considering regulatory options.

5.3 Limitations and Future Research

The limitations of this study stem from the chosen research approach and the study objective. The aim of the paper is to recognize ethical issues. We do not propose solutions, leaving these to future research. As a conceptual study on ethical principles, our analysis does not aim to serve AI developers who have access to the development process of LLMs and need in-depth technical tools to ensure ethical alignment. At the same time, the review is not intended as a philosophical treatise on ethics; rather, it occupies a space in applied ethics that is geared toward more responsible design and use of GenAI systems. Moreover, critically reviewing the ethical implications and issues of emerging technologies such as GenAI is a continuous effort; therefore, our review presents the situation at the time of writing and is based on the assumption that the most critical ethical issues will have emerged during the first years of mainstream GenAI use.

This study opens new research domains on the ethics of GenAI, contributing to emerging scholarly discussions (Chatterjee & Dethlefs, 2023; Weidinger et al., 2021; Zhuo et al., 2023). Because GenAI is still rapidly developing and the abilities of systems such as ChatGPT continue to grow, many ethical issues related to these technologies remain unspecified or have not yet come to light. To continue the research on the ethics of GenAI, we propose four future directions:

1. Complementary ethical approaches to principlism
2. Ethical tensions and contradictions
3. End-user perspectives on the ethics of GenAI
4. Salience of specific ethics principles to various GenAI stakeholders

First, GenAI ethics discussion would benefit from the development of complementary ethical approaches to the dominant principlism paradigm (Seger, 2022). While ethics principles present a starting point, they need to be connected to concrete AI design and development (Morley et al., 2020), and the enforcement of principles should have institutional backing (Mittelstadt, 2019). Future research could ask which ethical perspectives on GenAI could complement abstract ethical principles (cf. Mingers & Walsham, 2010). For example, what contextual ethical issues are made less visible by the high abstraction level of ethics principles? What other bases for ethical reflection are available for researchers? Second, GenAI ethics research could focus explicitly on ethical tensions and contradictions (Whittlestone et al., 2019) and highlight the trade-offs and paradoxes that may arise when efforts are made to develop and use GenAI ethically. For example, how feasible is it to develop the present kinds of GenAI chatbots if intellectual property and privacy rights are verifiably respected? To what extent is truthfulness possible with LLM technologies, and what trade-offs does it imply with regard to, for example, speed and efficiency? Third, researchers should conduct end-user studies on the ethics of GenAI to include end-user voices in the discussion on GenAI. These studies could focus on aspects such as explainability and understandability vis-à-vis end users (Laato et al., 2022) as well as the experiences of interacting with various humanlike chatbots and of fairness or lack thereof from the perspective of minority groups. Research on GenAI end-users would shift GenAI research from conceptual and design-oriented studies to behavioral research concentrating on the end-users' GenAI-related attitudes and behaviors as well as how these attitudes and behavior are linked. Finally, we propose that future research should investigate the salience of specific ethics principles to the various GenAI stakeholder groups, including AI developers, organizational managers, regulators, and auditors. Such studies could reveal distinct stakeholder roles with regard to ensuring GenAI systems' adherence to ethical principles.

References

Adamopoulou, E., & Moussiades, L. (2020). *An overview of chatbot technology*. In I. Maglogiannis, L. Iliadis, and E. Pimenidis (Eds.), *Artificial intelligence applications and innovations*. *AIAI 2020. IFIP Advances in Information and Communication Technology* (vol. 584, pp. 373–383). Springer.

Ågerfalk, P. J., Conboy, K., Crowston, K., Eriksson Lundström, J. S. Z., Jarvenpaa, S., Ram, S., & Mikalef, P. (2022). Artificial intelligence in information systems: State of the art and research roadmap. *Communications of the Association for Information Systems*, *50*(1), 420–438.

Alavi, M., Leidner, D. E., & Mousavi, R. (2024). Knowledge management perspective of generative artificial intelligence. *Journal of the Association for Information Systems*, *25*(1), 1–12.

Bale, S., Dhumale, R. B., Beri, N., Lourens, M., Varma, R., Kumar, V., Sanamdikar, S., & Savadatti, M. (2023). The impact of generative content on individuals privacy and ethical concerns. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(1), 697–703.

Benbya, H., Strich, F., & Tamm, T. (2024). Navigating generative artificial intelligence promises and perils for knowledge and creative work. *Journal of the Association for Information Systems*, *25*(1), 23–36.

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450.

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, *33*(7), 133–167.

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*, *8*(1).

Cameron, L., Lamers, L., Leicht-Deobald, U., Lutz, C., Meijerink, J., & Möhlmann, M. (2023). Algorithmic management: Its implications for information systems research. *Communications of the Association for Information Systems*, *52*, 518–537.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT*. arXiv:2303.04226. https://arxiv.org/abs/2303.04226

Chatterjee, J., & Dethlefs, N. (2023). This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy. *Patterns*, *4*(1).

Cheng, L., & Liu, X. (2023). From principles to practices: The intertextual interaction between AI ethical and legal discourses. *International Journal of Legal Discourse*, *8*(1), 31–52.

Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology*, *32*(3), 444–452.

Dasborough, M. T. (2023). Awe-inspiring advancements in AI: The impact of ChatGPT on the field of Organizational Behavior. *Journal of Organizational Behavior*, *44*(2), 177–179.

Davison, R. M., Chughtai, H., Nielsen, P., Marabelli, M., Iannacci, F., van Offenbeek, M., Tarafdar, M., Trenz, M., Techatassanasoontorn, A. A., Díaz Andrade, A., & Panteli, N. (2024). The ethics of using generative AI for qualitative data analysis. *Information Systems Journal*, *34*(5), 1433–1439.

Davison, R. M., Laumer, S., Tarafdar, M., & Wong, L. H. M. (2023). Pickled eggs: Generative AI as research assistant or co-author? *Information Systems Journal*, *33*(5), 989–994.

Dignum, V. (2019). *Responsible artificial intelligence*. Springer International Publishing.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T., Cunningham, S.W., Davies, G.H., Davison, R.M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J.S., Flavián, C., Gauld, R., Grover, V., Hu, M.C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K.R., Latreille, P., Laumer, S., Malik, F.T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J.H., O'Connor, S., Okumus, F., Pagani, M., Pandey, N.,

Papagiannidis, S., Pappas, I.O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N.P., Rehm, S.V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Stahl, B.C., Tiwari, M.K., van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J., & Wright, R. (2023). Opinion paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277–304.

Gordijn, B., & Have, H. t. (2023). ChatGPT: Evolution or revolution? *Medicine, Health Care and Philosophy*, *26*(1), 1–2.

Gozalo-Brizuela, R., & Garrido-Merchan, E. C. (2023). *ChatGPT is not all you need. A state of the art review of large generative AI models*. arXiv:2301.04655. https://arxiv.org/abs/2301.04655

Guo, D., Chen, H., Wu, R., & Wang, Y. (2023). AIGC challenges and opportunities related to public safety: A case study of ChatGPT. *Journal of Safety Science and Resilience*, *4*(4), 329–339.

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.

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. In *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1112 - 1123). ACM.

Hadi, M., Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., Zafar, A., Shaikh, M., Akhtar, N., Wu, J., Mirjalili, S., & Shak, M. (2023). *Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects*. TechRxiv. https://www.techrxiv.org/users/618307/articles/682263-large-language-models-a-comprehensive-survey-of-its-applications-challenges-limitations-and-future-prospects

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120.

Heinke, A., Radgoudarzi, N., Huang, B. B., & Baxter, S. L. (2024). A review of ophthalmology education in the era of generative artificial intelligence. *Asia-Pacific Journal of Ophthalmology*, *13*(4), 100089.

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Jovanovic, M., & Campbell, M. (2022). Generative artificial intelligence: Trends and prospects. *Computer*, *55*(10), 107–112.

Kajiwara, Y., & Kawabata, K. (2024). AI literacy for ethical use of chatbot: Will students accept AI ethics? *Computers and Education: Artificial Intelligence*, *6*.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25.

Kenthapadi, K., Lakkaraju, H., & Rajani, N. (2023). Generative AI meets responsible AI: Practical challenges and opportunities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 5805–5806).

Kirova, V. D., Ku, C. S., Laracy, J. R., & Marlowe, T. J. (2024). Software engineering education must adapt and evolve for an LLM environment. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V* (pp. 666–672).

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409.

Korzynski, P., Mazurek, G., Altmann, A., Ejdys, J., Kazlauskaite, R., Paliszkiewicz, J., Wach, K., & Ziemba, E. (2023). Generative artificial intelligence as a new context for management theories: Analysis of ChatGPT. *Central European Management Journal*, *31*(1), 3–13.

Kulkarni, P., Mahabaleshwarkar, A., Kulkarni, M., Sirsikar, N., & Gadgil, K. (2019). Conversational AI: An overview of methodologies, applications & future scope. In *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*.

Laato, S., Tiainen, M., Najmul Islam, A. K. M., & Mäntymäki, M. (2022). How to explain AI systems to end users: A systematic literature review and research agenda. *Internet Research*, *32*(7), 1–31.

Laine, J., Minkkinen, M., & Mäntymäki, M. (2024). Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Information & Management*, *61*(5).

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).

Loebbecke, C., El Sawy, O., Kankanhalli, A., Markus, M. L., Te'eni, D., Wrobel, S., Rydén, P., & Obeng-Antwi, A. (2020). Artificial intelligence meets IS researchers: Can it replace us? *Communications of the Association for Information Systems, 47*(1), 273–283.

Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022). News from generative artificial intelligence is believed less. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 97–106).

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). *Putting AI ethics into practice: The hourglass model of organizational AI governance.* arXiv:2206.00335. https://arxiv.org/abs/2206.00335

Mason, R. O. (1986). Four ethical issues of the information age. *MIS Quarterly*, *10*(1), 5.

Meskó, B., & Topol, E. J. (2023). The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digital Medicine*, *6*(1), 120.

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

Mingers, & Walsham. (2010). Toward ethical information systems: The contribution of discourse ethics. *MIS Quarterly*, *34*(4), 833.

Minkkinen, M., & Mäntymäki, M. (2023). Discerning between the "Easy" and "Hard" problems of AI governance. *IEEE Transactions on Technology and Society*, *4*(2), 188–194.

Minkkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous auditing of artificial intelligence: A conceptualization and assessment of tools and frameworks. *Digital Society*, *1*(3), 21.

Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in information systems research. *Communications of the Association for Information Systems*, *50*(1), 726–753.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507.

Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches. *Digital Society*, *2*(3), 49.

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, Scope, and Limitations. *Science and Engineering Ethics*, *27*(4).

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: A three-layered approach. *AI and Ethics*, *4*, 1085–1115.

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine*, *260*, 113172.

Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., & Forghani, R. (2020). Brief history of artificial intelligence. *Neuroimaging Clinics of North America*, *30*(4), 393–399.

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183–199.

Parikh, N. A. (2023). *Empowering business transformation: The positive impact and ethical considerations of generative AI in software product management -- A systematic literature review*. arXiv:2306.04605. https://arxiv.org/abs/2306.04605

Pavlik, J. V. (2023). Collaborating with ChatGPT: Considering the implications of generative artificial intelligence for journalism and media education. *Journalism & Mass Communication Educator*, *78*(1), 84–93.

Pressman, S. M., Borna, S., Gomez-Cabello, C. A., Haider, S. A., Haider, C., & Forte, A. J. (2024). AI and ethics: A systematic review of the ethical considerations of large language model use in surgery research. *Healthcare*, *12*(8), 825.

Qadir, J. (2023). Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON*).

Rana, N. P., Pillai, R., Sivathanu, B., & Malik, N. (2024). Assessing the nexus of generative AI adoption, ethical considerations and organizational performance. *Technovation*, *135*.

Raza, S., Ghuge, S., Ding, C., Dolatabadi, E., & Pandya, D. (2024). FAIR enough: Develop and assess a FAIR-compliant dataset for large language model training? *Data Intelligence*, *6*(2), 559–585.

Riemer, K., & Johnston, R. B. (2019). Disruption as worldview change: A Kuhnian analysis of the digital music revolution. *Journal of Information Technology*, *34*(4), 350–370.

Ruane, E., & Birhane, A. (2019). Conversational AI: Social and ethical considerations. In *AICS - 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science*.

Sabherwal, R., & Grover, V. (2024). The societal impacts of generative artificial intelligence: A balanced perspective. *Journal of the Association for Information Systems*, *25*(1), 13–22.

Sands, S., Campbell, C., Ferraro, C., Demsar, V., Rosengren, S., & Farrell, J. (2024). Principles for advertising responsibly using generative AI. *Organizational Dynamics*, *53*(2).

Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2023). Artificial intelligence governance for businesses. *Information Systems Management*, *40*(3), 229–249.

Schneider, J., Meske, C., & Kuss, P. (2024). Foundation models. *Business & Information Systems Engineering*, *66*(2), 221–231.

Seger, E. (2022). In defence of principlism in AI ethics and governance. *Philosophy & Technology*, *35*(2), 45.

Sengar, S. S., Hasan, A. Bin, Kumar, S., & Carroll, F. (2024). *Generative artificial intelligence: A systematic review and applications*. Multimedia Tools and Applications. https://link.springer.com/article/10.1007/s11042-024-20016-1

Siau, K., & Wang, W. (2020). Artificial intelligence (AI) ethics. *Journal of Database Management*, *31*(2), 74–87.

Stahl, B. C., & Eke, D. (2024). The ethics of ChatGPT – Exploring the ethical issues of an emerging technology. *International Journal of Information Management*, *74*.

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, *614*(7947), 214–216.

Sun, L., Huang, Y., Wang, H., Wu, S., Zhang, Q., Li, Y., Gao, C., Huang, Y., Lyu, W., Zhang, Y., Li, X., Liu, Z., Liu, Y., Wang, Y., Zhang, Z., Vidgen, B., Kailkhura, B., Xiong, C., Xiao, C., Li, C., Xing, E., Huang, F., Liu, H., Ji, H., Zhang, H., Yao, H., Kellis, M., Zitnik, M., Jiang, M., Bansal, M., Zou, J., Pei, J., Liu, J., Gao, J., Han, J., Zhao, J., Tang, J., Wang, J., Vanschoren, J., Mitchell, J.C., Shu, K., Xu, K., Chang, K.W., He, L., Huang, L., Backes, M., Gong, N.Z., Yu, P.S., Chen, P.Y., Gu, Q., Xu, R., Ying, R., Ji, S., Jana, S., Chen, T., Liu, T., Zhou, T., Wang, W., Zhang, X., Wang, X., Xie, X., Chen, X., Ye, Y., Cao, Y., Chen, Y., & Zhao, Y. (2024). TrustLLM: Trustworthiness in large language models. *Proceedings of Machine Learning Research*, *235*, 20166 - 20270.

Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, *379*(6630), 313–313.

Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review*, *4*(3), 356–367.

Van Slyke, C., Johnson, R., & Sarabadani, J. (2023). Generative artificial intelligence in information systems education: Challenges, consequences, and responses. *Communications of the Association for Information Systems, 53*(1), 1–21.

Wahde, M., & Virgolin, M. (2021). The five Is: Key principles for interpretable and safe conversational AI. In *2021 The 4th International Conference on Computational Intelligence and Intelligent Systems* (pp. 50–54).

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., & Li, B. (2023). DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. *Advances in Neural Information Processing Systems*, *36*.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., & Gabriel, I. (2021). *Ethical and social risks of harm from language models*. arXiv:2112.04359. https://arxiv.org/abs/2112.04359

Weisz, J. D., Muller, M., He, J., & Houde, S. (2023). *Toward general design principles for generative AI applications*. arXiv:2301.05578. https://arxiv.org/abs/2301.05578

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. Nuffield Foundation.

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on Microsoft's tay "experiment," and wider implications. *ACM SIGCAS Computers and Society*, 47(3), 54–64.

Zhui, L., Fenghe, L., Xuehu, W., Qining, F., & Wei, R. (2024). Ethical considerations and fundamental principles of large language models in medical education: Viewpoint. *Journal of Medical Internet Research*, *26*.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming ChatGPT via Jailbreaking: Bias, robustness, reliability and toxicity.

Zlateva, P., Steshina, L., Petukhov, I., & Velev, D. (2024). A conceptual framework for solving ethical issues in generative artificial intelligence.

Appendix A: Inclusion and Exclusion Criteria

Table A1. Inclusion and Exclusion Criteria

Inclusion criteria (IC)	Exclusion criteria (EC)
IC#1 Studies published in English	**EC#1** Studies in languages other than English
IC#2 Studies that address generative AI	**EC#2** Studies that do not focus on generative AI
IC#3 Studies that outline specific lists of ethical principles for generative AI	**EC#3** Studies that do not include lists of ethical principles
IC#4 Studies published before April 2024	

About the Authors

Joakim Laine is a Doctoral Researcher in Information Systems Science at the University of Turku. His research focuses on the socio-technical challenges of AI systems, mainly exploring the development of frameworks and methods for responsible AI auditing. His work emphasizes the importance of integrating ethical principles into the design, use, and auditing of AI technologies. His research has appeared in journals such as Information & Management and Digital Society.

Matti Minkkinen is a Postdoctoral Researcher in Information Systems Science at the University of Turku. He conducted his Ph.D. research on privacy protection in the European Union, and his recent research covers socio-technical perspectives on responsible artificial intelligence and systemic foresight processes. Minkkinen has several years of research and teaching experience on the interplay between future visions and socio-technical change, as well as foresight methods. His research has been published in journals such as Technological Forecasting & Social Change, Information Systems Frontiers, and New Media & Society.

Matti Mäntymäki is a Professor of Information Systems Science at the University of Turku, Finland. His research interests cover the psychosocial, organizational, and business implications of digitalization, with a particular focus on the governance and social responsibility of artificial intelligence. He has authored over 120 peer-reviewed papers in journals such as Information Systems Journal, Information & Management, Technological Forecasting & Social Change, Information Systems Frontiers, International Journal of Information Management, Journal of Business Research, Information Technology & People, Computers in Human Behavior, Journal of Systems & Software, and Communications of the Association for Information Systems.