Preprints.org

Ethical Concerns of Generative AI and Mitigation Strategies: A Systematic Mapping Study

Yutan Huang , Chetan Arora [*] , Wen Cheng Huong , Tanjila Kanij , Anuradha Madulgalla , John Grundy

Article

Ethical Concerns of Generative AI and Mitigation Strategies: A Systematic Mapping Study

Yutan Huang [1], Chetan Arora [1,*], Wen Cheng Huong [1], Tanjila Kanij [2], Anuradha Madulgalla [2] and John Grundy [1]

1   Faculty of Information Technology, Monash University, Clayton, Victoria, Australia
2   School of Science, Computing and Engineering Technologies, Swinburne University, Hawthorn, Victoria, Australia
3   School of Information Technology, Deakin University, Geelong, Victoria, Australia
*   Correspondence: chetan.arora@monash.edu

Abstract: Generative AI technologies, particularly Large Language Models (LLMs), have transformed numerous domains by enhancing convenience and efficiency in information retrieval, content generation, and decision-making processes. However, deploying LLMs also presents diverse ethical challenges, and their mitigation strategies remain complex and domain-dependent. This paper aims to identify and categorize the key ethical concerns associated with using LLMs, examine existing mitigation strategies, and assess the outstanding challenges in implementing these strategies across various domains. We conducted a systematic mapping study, reviewing 39 studies that discuss ethical concerns and mitigation strategies related to LLMs. We analyzed these ethical concerns using five ethical dimensions that we extracted based on various existing guidelines, frameworks, and an analysis of the mitigation strategies and implementation challenges. Our findings reveal that ethical concerns in LLMs are multi-dimensional and context-dependent. While proposed mitigation strategies address some of these concerns, significant challenges still remain. Our results highlight that ethical issues often hinder the practical implementation of the mitigation strategies, particularly in high-stake areas like healthcare and public governance; existing frameworks often lack adaptability, failing to accommodate evolving societal expectations and diverse contexts.

Keywords: generative AI; AI ethics; Large Language Models (LLMs) ; systematic mapping study

---

1. Introduction

The evolution of Generative AI, particularly Large Language Models (LLMs), has seen remarkable advancements since 2020 with the introduction of models like ChatGPT and Bard. LLMs have revolutionized tasks, such as writing assistance, code generation, and customer support automation, by leveraging vast amounts of data to generate coherent and contextually relevant natural language (NL) responses [1,2]. As a subset of Generative AI—systems designed to create new content—LLMs go beyond traditional AI techniques, which focus primarily on analyzing existing data. LLMs, in contrast, are capable of generating text, images, and music that mimic human creativity [3]. This capability is powered by advancements in neural network architectures, especially transformers, which enable LLMs to learn the nuances of human language and produce semantically accurate content [4]. LLMs, such as OpenAI's GPT-4 and Google's Gemini, utilize transformer architecture to understand and generate human-like text. GPT-4, for instance, employs a transformer-decoder architecture with billions of parameters, allowing it to generate detailed and contextually relevant text across various topics [5].

LLMs have been used to address numerous challenges, offering innovative solutions across sectors such as healthcare, education, and finance. It has the potential to bring efficiency and effectiveness into these areas [6]. However, as a rapidly emerging technology, LLMs currently operate with limited regulations or oversight, which raises significant ethical concerns [7]. Without mature and robust

guidelines, these potent tools risk being misused or improperly applied. For instance, there have been cases where AI-generated content has perpetuated biases or disseminated misinformation, underscoring the critical need for comprehensive ethical guidelines [8–11]. Although ethics is a broad concept that varies across contexts and fields, understanding the ethics around LLMs is becoming increasingly important for guiding the responsible use and development of these technologies.

To effectively mitigate these risks and harness the full potential of LLMs, it is crucial to understand the existing ethical concerns and strategies proposed to address these issues. To this end, we present our systematic mapping study (SMS) aimed at identifying primary studies that have investigated the various key ethical challenges associated with LLMs usage and by analyzing these studies to provide insights into developing a more comprehensive ethical framework to guide their responsible use. We observed that all the articles provided their understanding of the ethical concerns and categorized these concerns into specific ethical dimensions. The ethical dimensions were categorized based on existing ethical guidelines and regulatory frameworks. These ethical dimensions refer to topics that group similar ethical issues arising within the field, e.g., privacy and bias. We also found that while all the articles proposed different strategies to address these ethical concerns, most were conceptual ideas with a noticeable lack of evaluation regarding their effectiveness in resolving ethical issues using generative AI.

We followed Kitchenham et al.'s guidelines [12] for performing systematic reviews and Peterson et al.'s guidelines [13] for conducting systematic mapping studies. We wanted to answer the following research questions (RQs):

RQ1. What are the ethical dimensions defined in the use of generative AI across various fields? In RQ1, we wanted to identify key ethical dimensions associated with deploying and creating large language models (LLMs) by analysing relevant literature. These ethical dimensions were then mapped against existing ethical guidelines and regulatory frameworks to assess whether they are acknowledged within these standards.

RQ2. What strategies are used or proposed to address the ethical concerns of using generative AI across different fields? Having identified specific ethical dimensions and issues in RQ1, we then examined mitigation strategies mentioned in the analyzed primary studies used to address these ethical concerns across different dimensions.

RQ3. What are the challenges when implementing the strategies? We reviewed any reported challenges associated with implementing the identified mitigation strategies, detailing specific limitations and obstacles encountered in practice.

The main contributions of this mapping study include:

- We identified a list of 39 primary studies that focus on the ethical concerns of using generative AI. All the studies identified various ethical concerns related to specific ethical dimensions and provided strategies to address these concerns.
- Of these 39 papers, 13 conducted some form of empirical evaluation to either investigate or validate a proposed strategy. The remaining 26 papers are conceptual papers that propose strategies but do not conduct empirical evaluations, or in some cases, empirical evaluation is not feasible.
- We identified the most prominent ethical dimensions where ethical concerns are concentrated and the ethical dimensions that require more focused attention.
- We identified a set of key research directions needed to address ethical issues with using LLMs in practice.

The rest of the paper is structured as follows: Section 2 provides the background and related work on ethical concerns of the use of generative AI. Section 3 details our methodology for conducting a literature search and selection process on ethical considerations of generative AI. Section 4 reports the results from the selected primary studies. Section 5 discusses key results and summaries. Section 6 addresses threats to validity. Section 7 concludes our study.

2. Background

2.1. Terminology

Ethics is the discipline (or philosophy) that deals with conduct and questions of morality, exploring what is right and wrong, as well as key principles that govern human behaviors [14]. This exploration includes normative ethics, which seeks to establish general principles for how people should act; applied ethics, which addresses concrete ethical issues in various fields; and metaethics, which involves analyzing the nature of moral judgments and the underlying assumptions of ethical theories [15,16].

AI ethics is a multidisciplinary field that examines the ethical, legal, and social implications of artificial intelligence (AI) systems [17]. It involves developing frameworks, principles, and guidelines to ensure that AI technologies are designed, implemented, and used to align with societal values, human rights, and ethical standards [18]. AI ethics is heavily grounded in normative ethics, as it seeks to establish principles and guidelines for how AI should be designed and used. Normative questions in AI ethics include what constitutes fair AI, ensuring AI respects human rights, and developers' and users' responsibility in mitigating AI's potential harms [19]. AI ethics is also primarily a form of applied ethics, as it directly deals with practical ethical issues in real-world contexts, including the ethical deployment of AI in areas like healthcare and law enforcement, which concerns specific scenarios and case studies [20]. Metaethical questions are also addressed in the context of AI ethics, including the debate over whether AI systems can be considered moral agents and what ethical responsibilities should be assigned to them [21]. In summary, AI ethics extends ethics into the AI field, addressing both theoretical and practical challenges AI technologies pose.

Ethical dimensions are the ethical considerations that arise when developing, deploying, and utilizing LLMs based on existing ethical guidelines and regulatory frameworks. When using LLMs various ethical dimensions may need to be addressed. **Mitigation strategies** are action-oriented, specific, and actionable approaches designed or proposed by the authors to address and resolve ethical issues directly; they can be evaluated. **Recommendations** are guidance-oriented, broader suggestions or guidelines proposed by authors that aim to influence future actions or policies regarding generative AI ethics; they are not usually subject to empirical testing or short-term evaluation. In our analysis of primary studies in this paper, we identify whether the strategies mentioned in the studies have been evaluated in practice. A strategy is **evaluated** if it has undergone some form of empirical testing, validation, or assessment to determine its effectiveness. If the strategy is theoretical or conceptual, based on logical reasoning, best practices, or expert opinion without verification of actual impact, it is **not evaluated**.

2.2. Evolution and Advancements of Generative AI and LLMs

AI ethics have been discussed since the 1950s, and the discussion related to LLMs began in 2018, with various scholars and institutions proposing frameworks to address the ethical implications of AI technologies [22]. However, it was not until after 2020 that authorities began implementing concrete regulatory and policy frameworks to govern these rapidly emerging and adopted AI technologies. The increasing complexity and capability of LLMs, as well as the incidents of data breach and misuse, underscored the urgent need for robust ethical guidelines and regulatory oversight [23]. For instance, the mass data leakage from Chinese applications in 2020 highlighted significant privacy concerns, raising a global call for stronger data protection measures against [24]. Below, we review the evolution of generative AI and LLMs, the ethical concerns they raise, and the regulatory responses from major jurisdictions and industry leaders.

Initially, AI research focused on rule-based systems and basic machine learning algorithms, but the development of Generative Adversarial Networks (GANs) marked a breakthrough [25]. GANs

enabled AI to create realistic images, videos, and text, which marked the foundation of Generative AI and made the space for further advancements. The introduction of the Transformer model in 2017 revolutionized Natural Language Processing (NLP), leading to the creation of powerful LLM models like GPT, which advanced the generative AI field to a large extent [26].

2.2.1. OpenAI's GPT Series

These models leverage large amounts of data and to perform complex language processing tasks with proficiency. GPT-3, introduced in 2020, comprised 175 billion parameters, was well known for its size and scope, has been used across a wide range of applications, from creative writing to code generation and translation, question answering with fine-tuning, demonstrating its learning capabilities where the model can learn from a few examples of a given task [27].

The successor of GPT-3, GPT-4, advances these capabilities by incorporating more parameters to improve its accuracy, fluency and versatility in tasks. It is evident in areas requiring context-sensitive information, such as legal document drafting, medical diagnosis support, and scientific research summarization [28]. There are several other LLMs. However, the GPT series of LLMs is the most widely known; hence, we only mention them, and specific LLMs are beyond the scope of this paper.

2.3. Ethical Implications of LLMs

The ethical concerns surrounding LLMs have been extensively discussed in the literature, with bias and privacy being primarily discussed. Bender et al. [29] highlighted the risks of bias and misinformation inherent in LLM models as LLMs trained on large, diverse datasets from the internet often reflect and amplify societal biases, potentially leading to harmful stereotypes and discrimination. This issue is compounded by the "black box" nature of LLMs, where the decision-making processes are not easily interpretable and opaque, making LLMs difficult to identify and correct biases that occurred [30]. Additionally, the ability of LLMs to generate realistic text can be misused for creating misinformation and conducting social engineering attacks, raising significant ethical concerns.

On the other hand, the extensive data requirements for training LLMs raise serious privacy issues. In 2017, Shokri et al. [31] demonstrated that AI models could be vulnerable to membership inference attacks, where an attacker can determine whether a specific individual's data was included in the training set. This has underscored the need for robust privacy-preserving techniques in developing and deploying LLMs. In 2021, Facebook experienced a significant data breach involving several popular applications with embedded AI components. The incident exposed the personal information of over 500 million users. The database containing user details such as real names, usernames, gender, location and phone numbers was exposed and offered for sale [32]. Since LLMs are trained on datasets and user-generated content from social media platforms like Facebook, such breaches raise significant concerns about privacy and data security in AI training processes, particularly problematic if the leaked data includes personal identifiers or proprietary details.

2.4. Regulatory and Policy Frameworks

The ethical and privacy concerns associated with LLMs have prompted various regulatory responses worldwide. The European Union's proposed Artificial Intelligence Act aims to establish a comprehensive legal framework to ensure the ethical use of AI technologies, focusing on transparency, accountability, and human oversight. This legislation addresses the need for AI systems to disclose AI-generated content and provide explanations for AI-driven decisions, particularly in high-risk applications such as healthcare and law enforcement [33]. In the United States, the Federal Trade Commission (FTC) has issued guidelines emphasizing fairness, transparency, and accountability in AI systems [34].

Additionally, the National Institute of Standards and Technology (NIST) is developing a framework for AI risk management to provide organizations with practical tools for managing AI-related risks [35]. China's Interim Measures for the Management of Generative AI Services, introduced in 2024, require clear labeling of AI-generated content and enforce strict data collection and usage

guidelines to protect user privacy [36]. These regulatory efforts aim to balance innovation with ethical considerations, ensuring the responsible development and deployment of Generative AI technologies. Furthermore, industry-specific guidelines such as those by IEEE provide a framework for ethical AI design, highlighting principles like transparency, accountability, and data privacy [37].

Tech companies have also established their own ethical guidelines. For instance, Google's AI principles emphasize fairness, interpretability, privacy, and safety, aiming to prevent misuse of AI technologies [38]. OpenAI has similar guidelines focusing on long-term safety, robustness, and compliance with ethical standards [39]. Despite these regulatory advancements, challenges remain. Overly descriptive regulations might slow down innovation and make compliance burdensome for companies, particularly smaller enterprises with the needs of using LLMs. Additionally, the global nature of AI development necessitates international cooperation to integrate standards and ensure effective enforcement across jurisdictions. The varying regulatory landscapes across different countries can lead to inconsistencies and gaps in ethical standards and protections [22,40].

2.5. Industry and Governmental Guidelines Selected

As mentioned in the previous section, there have been multiple frameworks and guidelines produced to aid in identifying and addressing AI ethics and governance issues. We selected four guidelines/frameworks due to their availability, authority, and significant impact on shaping the ethical use of AI technologies.

The IEEE Ethically Aligned Design Document aims to establish ethical principles for Autonomous and Intelligent Systems (A/IS) that advance human beneficence, prioritize benefits to humanity and the environment, and mitigate risks associated with these technologies. It emphasizes that prioritizing human well-being must not conflict with environmental sustainability [41]. A/IS must also evolve with accountability and transparency to ensure the trustworthiness of socio-technical systems.

National Institute of Standards and Technology (US)'s Artificial Intelligence Risk Management Framework outlines characteristics necessary for building trustworthy AI systems, recognizing both the transformative potential and unique risks of AI. The document emphasizes the importance of balancing various trustworthiness attributes to minimize potential harms [42].

Microsoft Responsible AI Standard aims to operationalize ethical principles across Microsoft's AI development and deployment processes by providing concrete, actionable guidelines [43]. The goals are organized into six key areas, each designed to support responsible and trustworthy AI.

European Union (EU)'s AI Act introduces a comprehensive regulatory framework to manage the development and deployment of AI across the EU, using a risk-based approach to tailor requirements for various AI applications [33].

2.6. Existing Mitigation Strategies for Ethical Concerns

Several mitigation strategies have been proposed and implemented to address the ethical concerns associated with LLMs. One key approach is the development of bias detection and mitigation techniques. Researchers have created tools such as the Large Language Model Bias Index (LLMBI) to quantify and address biases in LLM outputs [44]. These tools use advanced NLP techniques to detect and correct biases related to race, gender, and other sensitive attributes.

Another strategy involves enhancing transparency and interpretability through techniques like SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), which provide insights into how models make decisions and allow users to understand and trust AI outputs [45]. Privacy-preserving techniques such as differential privacy and federated learning are also being employed to protect individual data while training LLMs. Differential privacy introduces noise to the data to prevent the identification of specific individuals, while federated learning allows models to be trained across multiple decentralized devices without sharing raw data, thus enhancing privacy [46].

Collaboration between AI developers, policymakers, and other stakeholders is crucial in creating ethical guidelines and regulatory frameworks that are both effective and adaptable. Industry-wide initiatives, such as the Partnership on AI, bring together diverse perspectives to address ethical challenges and promote best practices in AI development and deployment [47]. However, these strategies also have limitations. Bias detection tools can only address known biases and may miss subtle or emerging issues, potentially leading to ongoing ethical challenges [44]. Transparency techniques like SHAP and LIME can help interpret decisions but do not eliminate the fundamental complexity of LLMs, and their effectiveness relies on the correct application and understanding of these methods by users [45]. Privacy techniques like differential privacy and federated learning require careful implementation to balance privacy with model performance, and there is limited empirical evaluation of their effectiveness in large-scale deployments [46]. Moreover, the current mitigation strategies often lack comprehensive evaluation, and their long-term impacts on AI ethics and performance remain uncertain [47].

Overall, while these mitigation strategies are essential for addressing ethical concerns, they must be continuously refined and adapted to keep pace with the rapid advancements in AI technology. Collaborative efforts and ongoing research are vital to realising the benefits of Generative AI while minimizing potential harms.

2.7. Related Systematic Reviews

The ethical implications of AI systems have increasingly attracted attention in recent years, leading to several systematic reviews and mapping studies. Li et al. (2023) conducted a systematic review focusing on the ethical concerns and related strategies for AI in healthcare, identifying 45 relevant studies. This review highlighted the need for transparency, privacy, and accountability in AI design and implementation, but it was primarily centered on healthcare applications, leaving a gap in understanding how these ethical concerns translate across other domains [48]. Atlam et al. (2024) mapped AI ethics research over the past seven years and identified 127 primary studies. Their findings revealed a concentration of research on fairness, transparency, and accountability while pointing out a lack of empirical evidence supporting AI ethics principles [49]. However, the study was more of a high-level categorization, lacking in-depth analysis of specific methodologies used to address these ethical concerns. A systematic literature review by the International Journal of Data Science and Analytics (2023) identified 66 papers focusing on developing objective metrics to assess the ethical compliance of AI systems [50]. While this review underscored the need for standardized, quantifiable metrics to evaluate AI ethics, it was limited by its exclusion of frameworks that require human intervention, which might be necessary for a comprehensive understanding of ethical AI [50]. Although these studies have provided valuable insights into the current development of AI ethics, their limitations highlight gaps in the literature. There is a need for a more comprehensive analysis that covers various domains of AI application when the focus is on healthcare, high-level categorizations and the exclusion of human-centred frameworks. This systematic mapping study aims to fill these gaps by providing a broader, more detailed exploration of AI ethics across diverse sectors, identifying the strategies used to address ethical concerns, and providing a foundation for the development of comprehensive, cross-domain ethical guidelines and frameworks.

3. Methodology

We conducted a systematic mapping study (SMS) on the rising ethical concerns in using GenAI, particularly LLMs. We conducted our study on the scientific literature, existing frameworks, and guidelines, e.g., industry guidelines and governmental guidelines for AI ethics. Given the rapidly evolving nature of the field, we recognised that a comprehensive understanding of the ethical concerns surrounding LLMs cannot be derived solely from scientific literature. Therefore, we expanded our study to include industry frameworks, governmental guidelines, and other authoritative sources,

ensuring a more holistic and up-to-date perspective on the ethical challenges in this dynamic area. This SMS has been carried out in accordance with the guidelines for systematic mapping studies as outlined by Peterson et al. [13].

Our mapping study was conducted over three stages, i.e., planning, conducting, and reporting the review. The planning phase involved writing a protocol, formulating research questions (RQs) and reviewing our protocol for alignment to the main aim of our study. The protocol included a plan for a search strategy. During the conducting stage, the first author developed search strings that the other authors reviewed. A librarian from our university was also consulted at this stage to ensure alignment with the best practices for conducting systematic reviews. The search strings went through an iterative process, and some that were not specific or only marginally relevant were removed. The first author then led our search process (illustrated in Figure 1) with the search strings in various databases and selected ACM Digital Library, IEEE Xplore, Proquest, Wiley, Web of Science, and Science Direct. The first author searched all selected databases and removed duplicates using Endnote in the initial paper screening. We identified relevant keywords and search strings to exhaust our exploration of existing empirical evidence. We note that based on the guidelines in this stage, we used the search process to identify the relevant scientific literature. We identified the industry and governmental guidelines based on the references in the scientific literature and snowballing. In total, our search process led to 39 scientific studies and six industry and governmental guidelines. The final step involved evaluating, reviewing, and reporting the final document.
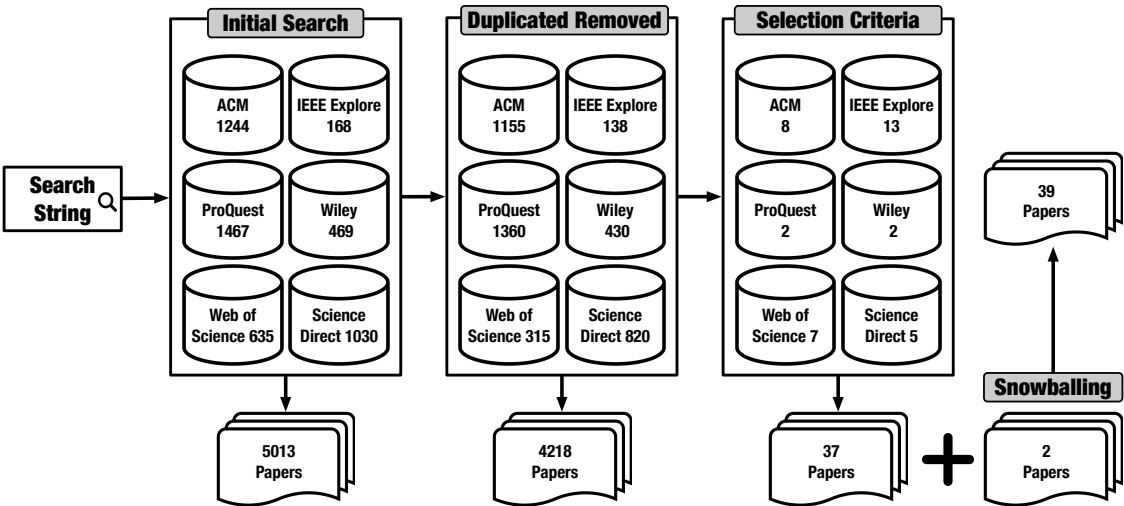

Figure 1. Study Search Process

3.1. Research Questions

We formulated three high-level research questions (RQs):

RQ1 - What are the ethical dimensions defined in the use of generative AI across various fields?

RQ2 - What strategies are used or proposed to address the ethical concerns of using generative AI across various fields?

RQ3 - What are the challenges when implementing the strategies?

We use these RQs to identify studies related to the ethical dimensions of Generative AI usage and the corresponding strategies to address these dimensions. In our study, an ethical dimension refers to specific ethical considerations that arise from using GenAI or LLMs [15,16,18]. These dimensions encompass a range of ethical concerns, principles, and guidelines crucial for the responsible development and use of LLMs.

3.2. Search Strategy

3.2.1. Search String Formulation

We formulated a search string to search online databases using key and alternative terms related to four main areas: large language models, guidelines, development, and ethics. However, we acknowledge that these areas may sometimes overlap with related concepts. Thus, we included a comprehensive set of terms in the search string to capture all relevant literature, shown in Table A1. The search string was tested on six online databases: IEEE Xplore, ACM Digital Library, ProQuest, Web of Science, Wiley Online Library, and Science Direct. These databases were chosen because they are comprehensive and widely recognised sources of high-quality research in computer science, engineering, and technology. The ACM Digital Library and IEEE Xplore are particularly relevant as they contain a vast collection of research papers and conference proceedings focused on artificial intelligence and machine learning. Proquest, Wiley, Web of Science, and Science Direct were included for their extensive coverage of multidisciplinary studies, including regulatory guidelines and ethical considerations in technology. Additionally, we included some of these databases to capture grey literature, e.g., the industry and governmental guidelines. Given that AI ethics is a relatively new field, grey literature (not part of the main 39 papers) provides valuable insights and complements our understanding.

The search string was refined over several iterations to maximise the relevance of the results. For instance, the first author would randomly select a sample of 8-10 papers, review them for relevance check, and further refine the search string.

In constructing the search strategies for various databases, we employed a range of logical connectors to ensure a comprehensive and relevant retrieval of literature on large language models and related issues. Each database required specific adjustments to optimize the search strategy, using various connectors and operators tailored to the database.

In IEEE, Web of Science, Wiley Online Library, and ScienceDirect, the search strategies made use of the "AND" connector to combine key concepts, ensuring that search results included multiple relevant topics such as governance, ethics, transparency, or accountability in relation to large language models (LLMs). The "OR" connector was used to include alternate expressions of similar concepts (e.g.,"ethic" OR "moral", "development" OR "design"). Additionally, symbols like the asterisk (*) were used to capture word variations. For instance, "Large Language Model*" ensures that both "Large Language Model" and "Large Language Models" are retrieved. This strategy helps retrieve documents that mention different term forms, maximizing the search's comprehensiveness. In ACM Digital Library, the search strategy used the "AND" and "OR" connectors to combine multiple concepts and capture alternate terms. "AND" was used to ensure that search results contained all the specified concepts, such as "large language model*" AND "ethics" AND "governance." This ensures that the retrieved documents address all the selected topics together. "OR" was used to include variations or synonyms for a concept, such as "ethics" OR "moral" OR "ethical", which broadens the search to include documents that may use different terms for the same idea.

In ProQuest, we used a more refined approach by employing the "noft" connector, which stands for "not in full text". This operator excludes terms that appear only in the body of the document and not in more critical fields such as the title, abstract, or subject headings. The use of "noft" is particularly useful in ProQuest because this database often contains a large volume of full-text content, including dissertations, reports, and articles that may mention key terms incidentally without focusing on them in depth. By using "noft", we can filter out documents where a term like "governance" is only briefly mentioned in the text but not central to the research focus. For example, using "noft(governance)" ensures that we retrieve documents where "governance" is emphasized in key fields, like the title or abstract, indicating that the topic is a primary focus of the work rather than a tangential mention. This increases the relevance and specificity of the search results, making sure that only documents dealing substantially with "governance" are included. Additionally, the "AND" and "OR" connectors were

also used in ProQuest to combine different thematic elements and provide alternate terms for key concepts, ensuring comprehensive coverage of topics while maintaining the precision of the search.

3.2.2. Automated Search and Filtering

Our search was conducted on databases between April and May 2024. The final search string was applied to the selected database online search engines, and an initial pool of 5013 papers was extracted. The list of papers was downloaded and exported to Endnote. Despite the large number of database results obtained by our search string (5013 papers), many papers were irrelevant or duplicates.

We first removed all duplicates (remaining papers = 4218) and applied a set of inclusion and exclusion criteria to filter out pertinent studies as part of our SMS protocol, shown in Table 1). For this mapping study, we only considered English-language studies that directly discuss the ethical dimensions or concerns in their studies. Given that the topic is relatively new, we included short papers of more than four pages to harness emerging ideas. However, we excluded irrelevant studies such as posters, keynotes, opinion papers, and magazine articles. The selection criteria were applied to all studies to identify the most relevant ones, with discussions among all authors during the study filtering process. This left us with 39 papers remaining.

Table 1. Inclusion and Exclusion Criteria

Criteria ID	Criterion
Inclusion Criteria	
I01	Papers discussing ethical concerns in the use of Generative AI
I02	Full text of the article is available.
I03	Peer-reviewed studies, and sector-specific studies.
I04	Papers written in English language.
Exclusion Criteria	
E01	Papers about GenAI that do not discuss ethical concerns
E02	Papers about ethical concerns that are not in the field of GenAI
E03	Papers that are less than four pages in length.
E04	Conference or workshop papers if an extended journal version of the same paper exists.
E05	Papers with inadequate information to extract relevant data.
E06	Vision papers, books (chapters), posters, discussions, opinions, keynotes, magazine articles, experience, and comparison papers.

4. Results

4.1. Selected Studies

After filtering we selected 39 primary studies on ethical concerns surrounding the use of LLMs, published from 2020 to 2024, as illustrated in Figures 2 and 3. The bar chart in Figure 2 shows the year-wise distribution of the selected papers by their year of publication. Publications were sparse and relatively consistent from 2020 through 2022, with only minor fluctuations. However, 2023 saw a sharp increase, with 26 studies published, reflecting a substantial rise in research interest in ethical concerns in AI. This upward trend continued into 2024, though with a slight decrease, likely due to our research cutoff in June 2024.

Figure 3 displays the distribution of these studies by publication venue. Most studies, 56.4%, were published in journals, followed by 33.3% in conferences. Publications on Arxiv accounted for 7.7%. This breakdown underscores that journals are the primary medium for disseminating research on AI ethics, while conferences also contribute significantly. We included the Arxiv papers as the topic of Ethical AI is relatively new, and Arxiv provides us with studies and research that may not yet have gone through peer review, which allows us to explore emerging ideas, diverse perspectives, and early-stage research that might not yet be available in journals and conferences.
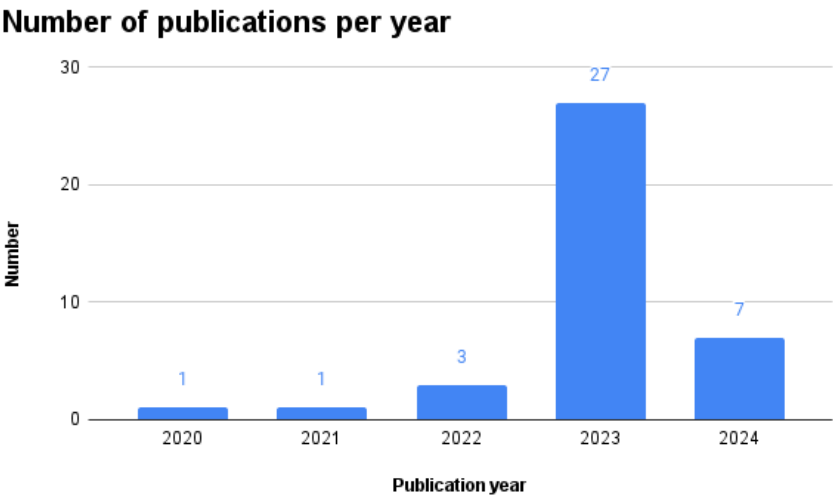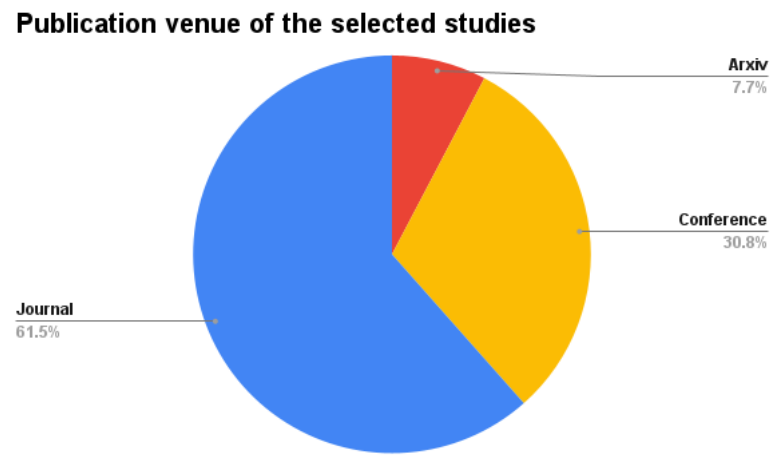
Figure 2. Study Distribution by Year

Figure 3. Study Distribution by publication

Our selected primary studies span multiple application domains. Each domain brings its own set of challenges and ethical considerations, reflecting the diverse applications and potential impacts of LLMs. This section categorizes the studies according to their respective domains mentioned in the primary studies, offering a detailed exploration of how ethical concerns manifest in different contexts.

- *General AI*: Papers categorized under "General AI" address AI's broad concepts and applications without focusing on a specific domain. They discuss relevant ethical challenges across multiple sectors where AI is deployed.

- *Healthcare*: Healthcare is a domain that frequently engages with ethical concerns such as privacy, safety, and bias, especially with the increasing use of AI and LLMs for medical applications. The papers in this domain discuss how AI-driven systems can both enhance and complicate healthcare practices, with patient confidentiality, data security, and algorithmic fairness being recurring themes. This domain is highly sensitive due to the direct impact on human well-being and medical decision-making.

- *Legal*: Papers focusing on the legal domain highlight the intersection of AI with law, exploring issues like accountability, transparency, and bias in legal AI systems. The integration of AI in legal contexts raises concerns about fairness, potential racial biases in predictive policing tools, and the transparency of AI-driven legal decisions.

- *Education*: The educational domain explores the ethical use of AI in learning environments, where concerns include privacy, fairness, and the transparency of AI systems in assessing student performance or providing educational content.

- *Public safety*: This domain involves the use of AI in public safety systems, where the ethical dimensions center on accountability, bias, and safety. AI systems used in policing, emergency response, and public surveillance must be scrutinized for fairness and transparency to avoid unintended harm to communities, especially marginalized groups.

- *Societal Impact*: Papers in this domain examine how LLMs interact with broader societal issues, such as their role in shaping public discourse, policy, and societal norms. Ethical concerns focus on the responsibility of those deploying LLMs to consider their social impact, including how they influence public opinion, access to information, and equality.

- *Cybersecurity*: This domain highlights mainly the privacy concerns specifically related to systems that use AI components or LLMs. Papers here focus on issues such as the protection of personal data, the risks of data leakage, and how AI systems can be designed to respect user privacy, especially focusing on data-hungry models like LLMs.

- *Economics*: Papers in this domain focus on the economic impact and ethical concerns surrounding AI and LLMs, particularly in relation to automation, job displacement, and the ethical use of AI in economic decision-making. The economic dimension of LLMs also brings up issues of accessibility and fairness in how AI systems are distributed and who benefits from their deployment.

The specific papers associated with each domain can be seen in Table 2.

Table 2. Domains mentioned in papers

Domain	Papers
Cybersecurity	[P1, P23, P27, P29, P31, P33]
Education	[P14, P16, P26]
General AI	[P2, P4, P5, P15, P17, P20, P28, P30, P35, P37, P39]
Healthcare	[P1, P9, P10, P11, P13, P21, P22, P24, P32, P34]
Societal Impact	[P2, P25, P38]
Legal	[P3, P6, P8]
Public Safety	[P7, P18, P19, P27]

We identified a diverse range of LLM models being used in the selected primary studies. A significant number of papers used versions of ChatGPT (P1, P7, P11, P12, P14, P16, P17, P18, P19, P24, P25, P27, P28, P34, P35, P37, and P39), highlighting its versatile use across various domains such as healthcare, education, and general conversational interfaces. Conversational Agents (CA), another key area of focus, were discussed in multiple papers (P2, P5, P15, P29, P31, P36). These agents are noted for their applications in interactive and supportive roles, including customer service and educational tools.

General LLMs were featured in several papers (P4, P6, P20, P21, P30, P32, P33, P38). These papers emphasized the broad capabilities of LLMs in understanding and generating human-like text, highlighting their potential across various sectors. Additionally, LLM-based chatbots and virtual assistants were discussed in papers such as P9, P10, and P13, indicating their growing relevance in enhancing user interaction and automating responses.

Models such as GPT-2 (P3) and transformers (P22) were also explored, with RoBERTa being mentioned alongside ChatGPT in P23 for the use of fine-tuning LLMs. Educational Large Language Models (EduLLMs) were specifically addressed in P26, showcasing their application in creating and facilitating educational content. Moreover, newer AI models like Gemini (Bard) and DALL-E were mentioned in the context of their capabilities and potential applications in papers P37 and P39.

Table 3. Generative AI Technologies Used in Primary Studies

LLM Models	Number Mentions
ChatGPT	18
Conversational Agents	6
DALL-E	1
EduLLMs	1
Gemini (Bard)	2
GPT-2	1
General LLMs	8
LLM-based Chatbots	1
LLM Virtual Assistants	1
RoBERTa	1
Transformers	1

4.2. RQ1 Results

We wanted to identify the various ethical issues related to AI that have been discussed across the selected papers. The purpose was to explore both the breadth and depth of these issues and how they align with key ethical dimensions.

In defining the key ethical dimensions for our study, we focused on safety, accountability, bias, transparency, and privacy. These dimensions are central to the international frameworks and guidelines, discussed in Section 2.5.

Safety is a prominent requirement in the EU AI Act, where high-risk AI systems must undergo thorough safety assessments to protect users and ensure system integrity in critical areas like healthcare and transportation. IEEE's Ethically Aligned Design also emphasizes safety through the principle of beneficence, which prioritizes human well-being in A/IS design [33]. Our focus on safety incorporates the need for operational robustness and user protection across different contexts, recognizing it as a foundational element of trustworthy AI.

Accountability also emerged as a crucial dimension, highlighted in both the NIST AI Risk Management Framework and Microsoft Responsible AI Standard, which call for strong accountability measures to hold developers and operators responsible for AI system outcomes. NIST's framework identifies accountability as key to effective risk management, requiring organizations to account for both foreseeable and unforeseen impacts [42]. By selecting accountability, we underscore the importance of defining ethical boundaries and responsibilities within AI operations.

Bias is widely acknowledged as an ethical concern due to the potential for AI systems to propagate discrimination and inequity. This dimension is explicitly covered in frameworks like IEEE's Ethically Aligned Design and the NIST AI RMF, which both stress the importance of reducing algorithmic bias to ensure fair, equitable outcomes [41,42]. Bias mitigation is especially pertinent in high-stakes areas like hiring and law enforcement, where biased AI outcomes can directly and significantly impact the society.

Transparency is a universally emphasized dimension, with all four frameworks underscoring the need for AI systems to be explainable, interpretable, and auditable. The EU AI Act enforces transparency requirements for high-risk applications, demanding that developers provide clear explanations of AI decision-making processes to foster trust [33]. Similarly, NIST lists transparency as fundamental to AI trustworthiness, enabling stakeholders to understand and evaluate AI operations. This dimension is central to building user trust and ensuring regulatory oversight, both necessary for responsible AI deployment [42].

Privacy remains a critical concern in AI ethics, particularly with the influence of GDPR embedded within the EU AI Act, which enforces stringent data protection for AI systems [33]. NIST's framework also highlights privacy-enhanced AI, advocating for systems that protect personal data and uphold users' informational rights [42]. By focusing on privacy, we address data misuse risks and safeguard users' rights to control their information.

We used these five overarching ethical dimensions to group up studies addressing key LLM usage ethical concerns. We also considered the perspectives of the authors in each paper, incorporating their opinions on which ethical dimension specific issues should belong to. This ensured that our categorization was informed not only by the frameworks and guidelines but also by the authors' interpretations. The analysis involved carefully reviewing each paper and coding the mentioned ethical issues, which were then grouped into themes that fit within these categories. In total, we identified 130 different ethical issues across the studies, which we categorized into 17 distinct themes. Figure 4 shows the key ethical dimensions to theme mappings. Table 4 shows the primary studies addressing each dimension.
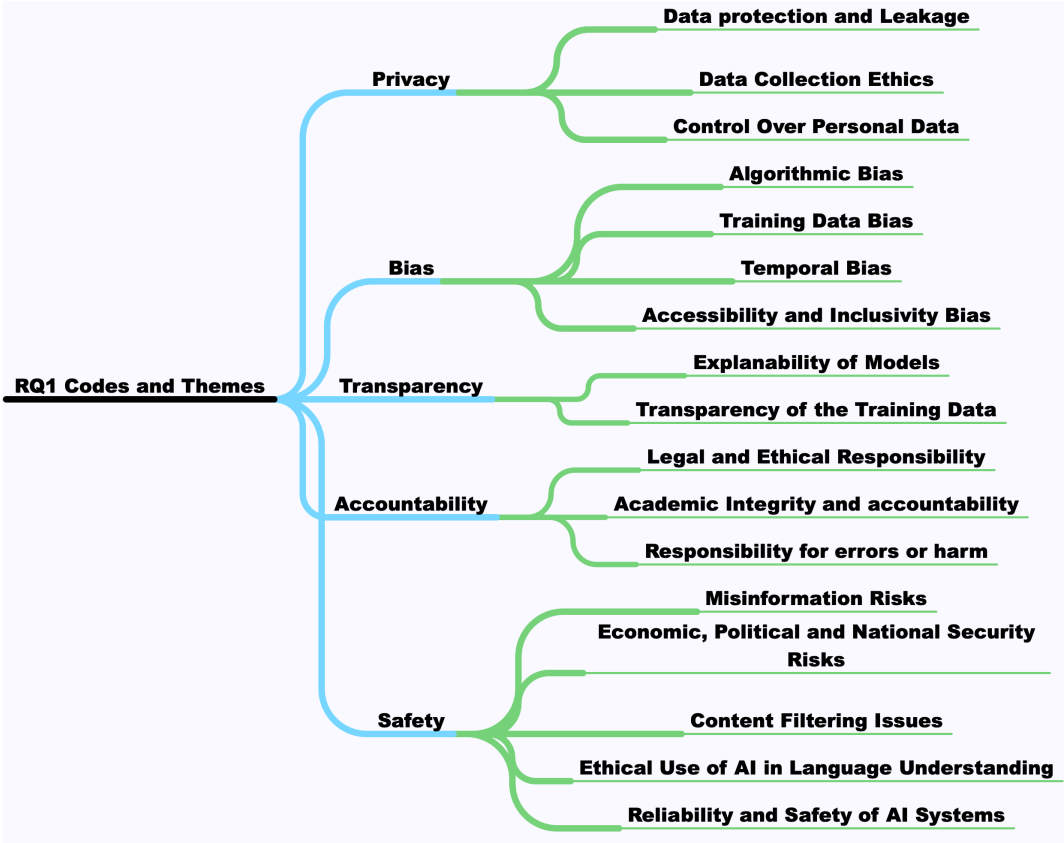

Figure 4. RQ1 codes and themes mapping

Table 4. Ethical Dimensions Identified from papers

Ethical Dimensions	Definition	Paper ID
Safety	Generative AI systems need to be ensured to operate safely and securely, minimizing risks and issues	P2, P6, P7, P8, P9, P10, P12, P13, P14, P18, P19, P21, P22, P25, P27, P28, P30, P32, P34, P38
Privacy	The information and privacy of individuals need to be protected when using generative AI systems	P1, P2, P5, P7, P8, P9, P10, P11, P12, P14, P16, P17, P22, P23, P24, P25, P26, P27, P28, P29,P31, P33, P34, P37, P38, P39
Transparency	The operations and decisions of generative AI systems are open and understandable to stakeholders	P2, P5, P6, P8, P10, P11, P12, P13, P15, P16, P19, P21, P22, P24, P34, P38, P39
Bias	The biases of the system need to be addressed as they may affect the fairness and equity of the AI system	P1, P3, P4, P5, P6, P8, P10, P11, P12, P14, P15, P16, P17, P18, P19, P20, P22, P24, P25, P28, P30, P34, P35, P36, P38
Accountability	AI systems need to be responsible and their actions can be traced and justified	P1, P5, P11, P16, P17, P21, P24, P34, P37, P38

4.2.1. Safety

The theme *"Ethical Use of AI in Language Understanding"* highlights the ethical challenges related to how AI systems, particularly virtual assistants (VAs), process and respond to language, including inappropriate or harmful speech, this theme inclues P10. In P10, a key concern is raised about the need to train VAs with inappropriate language in order for them to recognize and understand it. This brings up ethical questions about whether the VA should be able to understand insults or offensive language and, if so, how it should respond.

The theme *"Misinformation Risk Issues"* identifies the ethical concerns related to AI systems generating and disseminating false or misleading information, which can have serious consequences. This theme was covered in P2, P6, P8, P10, P12, P13, P14, P28, P30, P32, P38. In P12, a significant risk is highlighted with ChatGPT potentially producing sophisticated but hallucinatory fake information, especially in financial outputs, which could be hard to detect and lead to substantial financial losses. In P30, another concern is raised about LLMs generating misinformation, resulting in less informed users and eroding public trust in shared information.

The theme *"Economic, Political, and National Security Risks"* focuses on the dangers posed by AI systems in sensitive areas, particularly regarding the misuse of generative AI like ChatGPT. this theme is covered in P18. In P18, political security risks are noted. AI could inadvertently access and assimilate sensitive information, such as national secrets, trade secrets, or personal data, through user inputs, potentially leading to information leakage. Additionally, military security is threatened by the potential misuse of AI to generate attack codes targeting critical national infrastructure or military facilities. Economic security is also at risk, as tampering with training data could create false financial data and market predictions, undermining economic integrity.

The theme *"Content Filtering Issues"* addresses the ethical concern of LLMs, processing both accurate and harmful content due to inadequate filtering mechanisms. This theme covers P13. In P13, it is noted that LLMs can ingest and incorporate harmful content along with accurate information, which can result in the AI producing outputs that may be inappropriate or damaging.

The theme *"Reliability and Safety of AI Systems"* focuses on the ethical concerns surrounding the accuracy and dependability of AI, particularly in critical situations; this theme is covered in P7, P9, P13, P14, P19, P21, P25, P27, P28, P30, P34. In P19, a significant issue is the provision of oversimplified and erroneous advice on safety matters, with the advice often lacking traceability due to missing cited sources, making fact-checking difficult. This raises concerns about AI's reliability in delivering accurate and trustworthy information in safety-critical contexts. In P21, the issue of trust is discussed, especially

regarding the reliability of LLMs in situations outside their training conditions. This phenomenon, known as an "out-of-distribution shift," can lead to performance failures.

4.2.2. Privacy

The theme *"Control over Personal Data"* emphasizes the ethical importance of protecting individuals' privacy and ensuring they have control over how their personal information is used, particularly in AI-driven systems. As AI increasingly relies on sensitive data, especially in fields like healthcare, safeguarding this data becomes crucial to prevent misuse and ensure privacy compliance, this theme is covered in P1, P14 and P31. In P1, one key concern is the need for robust data protection and transparency in AI's decision-making, especially in healthcare, to avoid bias and ensure fairness. In P14, an example highlights the potential for adding privacy settings to give users control over how their data is collected and shared, reinforcing the need for individual autonomy over personal information.

The theme *"Data Protection and Leakage"* addresses the ethical concerns related to the privacy and security of data in AI systems, particularly regarding the risk of unauthorized access or data breaches, this theme is covered in P1, P2, P5, P7, P8, P9, P10, P11, P12, P14, P16, P17, P22, P23, P24, P25, P26, P27, P28, P29, P30, P33, P34, P37, P38, P39. In P23, the issue is highlighted in the context of AI researchers fine-tuning pre-trained models with private data for various tasks. These models are vulnerable to privacy attacks due to their tendency to memorize training data, known as the "memorization issue," posing risks to sensitive information. In P26, concerns arise around smart education systems, which collect and analyze vast amounts of student data to provide personalized learning.

The theme of "Data Collection Ethics" focuses on the ethical issues surrounding how data is gathered for training LLMs. This theme is covered in P14 and P27. In P27, concerns are raised about scraping information from internet forums and other sources without proper consent or oversight, which raises questions about the ethicality of the data collection process. In P14, another issue is the lack of disclosing the underlying sources used in LLM training, as they are not shared with the public.

4.2.3. Bias

The theme of "Algorithmic Bias" highlights the ethical concerns regarding how AI models can perpetuate social stereotypes and discrimination, particularly when trained on biased data. This theme is covered in P4, P5, P6, P12, P22, P24, P25, P28, P30, P34, P35, P38. In P25, an example is provided where Amazon used AI in human resources, and it was found that women were consistently rated lower than men. This bias arose because the AI analyzed resumes from the past decade, a period where most resumes came from men, indicating that the data was neither large nor diverse enough to mitigate bias. In P30, a similar concern is raised about how LLMs can perpetuate social stereotypes and introduce representational and allocational harms, leading to discrimination.

The theme of "Training Data Bias" addresses the ethical concerns that arise when AI models are trained on datasets that either overrepresent or underrepresent certain demographic groups. This theme is covered in P1, P3, P4, P8, P10, P11, P16, P17, P28, P35. In P35, demographic bias is highlighted as a major issue, where this imbalance in the data can cause models to exhibit biased behaviors toward specific genders, races, ethnicities, or other social groups. In P28, the concern is further emphasized, showing how biased data can perpetuate social stereotypes and lead to unfair discrimination. When models are trained on skewed representations of certain groups, it can result in unjust outcomes, reinforcing existing inequalities.

The theme of "Temporal Bias" focuses on the ethical concerns arising from the time-based limitations of training data used in AI models. This theme is associated with P35. In P35, it is noted that these models are often trained on data from specific time periods or have temporal cutoffs, which can lead to bias when reporting on current events, trends, or opinions. This limitation may result in the model providing outdated or inaccurate information when addressing contemporary topics and an incomplete understanding of historical contexts due to the lack of temporally diverse data. This can affect the relevance and accuracy of AI-generated content in time-sensitive situations.

The theme of "Accessibility and Inclusivity Bias" highlights the ethical concerns related to how AI systems may overlook the needs of individuals with disabilities or those from diverse linguistic backgrounds, this theme is covered in P14, P15, P18, P19, P20, P35. In P14, it is highlighted that accessibility features, such as screen readers, alternative text for images, or video captions, are often inadequate or absent in AI systems. Additionally, limitations in language translation capabilities may exclude non-native speakers or those with diverse linguistic needs.

4.2.4. Accountability

The theme of "Legal and Ethical Responsibility" focuses on the challenges surrounding accountability and moral responsibility when using AI in decision-making processes, especially in critical contexts, this theme is covered in P5, P21, P34, P37, P38. In P5, it is noted that accountability is a pivotal element in AI governance, particularly when delegating tasks such as predictions or decisions to AI systems. However, the definition of accountability in AI remains ambiguous and should be clarified based on factors such as the subject, scope, and context of its application. In P21, the issue of moral responsibility arises in high-stakes decisions, such as medical care, where traditionally, a human is held accountable. When an AI algorithm is involved, it becomes less clear who is responsible for the outcome, especially in adverse decisions. Establishing clear guidelines for when a human professional may "overrule" the AI is essential to maintaining accountability.

The theme of "Academic Integrity and Accountability" highlights the ethical concerns surrounding using AI, particularly ChatGPT, in academic settings, this theme is covered in P11, P16, P17. In P16, the broader academic community has raised concerns about students using ChatGPT to plagiarize assignments, research papers, and other academic work, which undermines academic integrity. In P17, questions arise regarding the proper use of ChatGPT's responses for academic purposes, such as whether AI-generated content should be credited and how accountability is managed if it is misused. The uncertainty extends to legal implications, mainly if AI-generated material is used maliciously or in ways that breach academic or ethical standards.

The theme of "Responsibility for Errors or Harm" involves the ethical concerns related to AI-driven decisions and the risks associated with responsibility. This theme is covered in P1 and P24. In P1, it is highlighted that the availability of AI and automated decision aids can lead to a human tendency to rely too heavily on these technologies, often minimizing cognitive effort. In healthcare, this over-reliance on AI-driven decisions by clinicians can lead to misleading conclusions, potentially endangering patient safety and well-being.

4.2.5. Transparency

The theme of "Explainability of Models" addresses the transparency challenges posed by the complexity of LLMs. In P38, the "black box" nature of LLMs is emphasized, highlighting the difficulty in understanding their decision-making processes due to the vast number of parameters they contain—often in the millions or billions. This theme is covered in P2, P5, P8, P10, P11, P12, P13, P15, P16, P21, P22, P24, P34, P38, and P39. This complexity makes it nearly impossible to explain how these models generate their outputs fully. In P39, the mystery surrounding the emergent capabilities of LLMs is further discussed as researchers remain uncertain about what drives these behaviors.

The theme of "Transparency of the Training Data" focuses on the ethical issues arising from the lack of clarity regarding the data used to train AI models. This theme is covered in P6 and P19. In P6, it is noted that the training data is often treated as a trade secret, and while efforts are underway to reverse engineer what data was used, companies neither confirm nor deny the accuracy of these guesses. This opacity raises concerns about the integrity and representativeness of the training data. In P19, the lack of transparency becomes particularly problematic when ChatGPT provides answers or recommendations, as users are left without a clear understanding of the sources behind its advice. Unlike human experts who can cite specific references, ChatGPT operates on a probabilistic model, making it difficult to trace or explain the origins of its responses.

> **RQ1 Key Takeaways**
>
> The RQ1 findings reveal critical ethical concerns across safety, transparency, accountability, privacy, and bias. Safety issues, such as misinformation risks and reliability, highlight the need for secure, trustworthy AI systems in high-stakes areas. Transparency challenges, including model explainability and unclear training data, reflect ongoing difficulties in understanding LLM decisions. Accountability concerns raise questions about defining responsibility in AI-driven decisions. Privacy issues emphasize user control over personal data in AI applications, while bias concerns rooted in algorithmic and training data show how LLMs risk reinforcing societal inequalities.
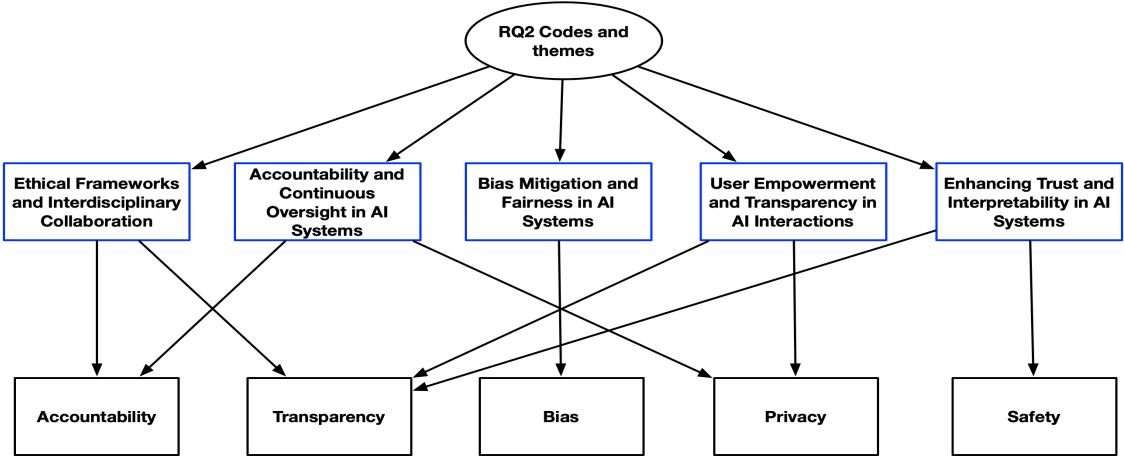
4.3. RQ2 Results

Figure 5. RQ2 Mapping

To address the second research question, "What strategies are used to address the ethical dimensions?", we conducted a thorough review of the studies and extracted the mitigation strategies and recommendations presented in each paper. A mitigation strategy refers to a concrete approach or action that is proposed or implemented to directly address or reduce a specific risk, issue, or challenge. In contrast, a recommendation is a suggestion or guidance proposed by the authors for future actions, often highlighting potential solutions or best practices without immediate implementation. These strategies were categorized accordingly, and we further explored their evaluation status. If a mitigation strategy or recommendation was deemed "evaluated," it indicates that the authors conducted experiments or implemented the solution within a specific context and obtained results. On the other hand, if it was categorized as "not evaluated," it means the authors either did not verify the solution or the solution could not be assessed in the short term. This process allowed us to systematically identify and classify the approaches proposed to address the ethical dimensions of AI use. The mitigation strategies and recommendations were evaluated only in 5/39 (13.2%) studies. After identifying the evaluation status, we conducted a thematic analysis and categorized various mitigation strategies or recommendations, reflecting the proposed or practical solutions presented.

- *Ethical Frameworks and Interdisciplinary Collaboration*: Establishing ethical guidelines for LLMs requires input from diverse disciplines to address the complex ethical challenges they present, this theme is covered in P1, P2, P4, P6, P16, P18, P24, P25, P30, P35, P37, P38. In P1, the recommendation emphasizes the need to establish ethical guidelines and regulatory frameworks for AI in healthcare, advocating for interdisciplinary collaboration to ensure that ethical standards are comprehensive and adaptable to advancements in technology. This collaboration is essential to create frameworks that are both effective and applicable across various domains. In P2, a Machine Ethics approach is proposed, suggesting that ethical standards and reasoning should be

directly embedded within AI systems. This approach would enable AI to make ethically informed decisions autonomously, integrating ethical principles into the core of AI functionality. In P6, there is a focus on the legal dimensions of AI, suggesting that the training phase of AI may be covered under fair use; however, clearer guidelines are needed to inform system users. These examples illustrate the necessity of establishing ethical frameworks that are continuously refined through interdisciplinary cooperation, ensuring that LLMs operate ethically and are aligned with societal expectations.

- *Bias Mitigation and Fairness in AI systems*: Addressing biases and promoting fairness in LLMs requires a multifaceted approach, with strategies implemented at various stages of the AI development process, this theme is covered in P1, P2, P3, P4, P6, P8, P16, P19, P25, P28, P35. In P3, a mitigation strategy is employed using a fill-in-the-blank method in a GPT-2 model to ensure that context is carefully considered during predictions, focusing on binary racial decisions between "White" and "Black." This method involves examining racial bias by masking racial references within the context, prompting the model to assign probabilities and observe potential biases. In P4, a combination of techniques is used to handle biases systematically: Pre-processing methods transform input data to reduce biases before training, while in-processing techniques modify learning algorithms to eliminate discrimination during training. Additionally, post-processing methods are applied to adjust the model's output after training, particularly when retraining is not an option, treating the model as a black box. These strategies showcase a comprehensive effort to mitigate biases at different levels of the AI development pipeline, aiming to create fairer and more equitable AI systems.

- *Enhancing Trust and Interpretability in AI Systems*: Building trust in LLMs relies heavily on making AI systems transparent and understandable to users, especially in critical fields like healthcare, this theme include P1, P7, P9, P15, P20, P24, P30. In P1, a recommendation is made to focus on improving the interpretability of AI algorithms so that healthcare professionals can clearly understand and trust the decisions generated by these systems. This involves making the decision-making processes of AI transparent, allowing professionals to verify and rely on AI outputs confidently. In P7, a mitigation strategy involves using leading questions to guide ChatGPT when it fails to generate valid responses, enabling users to refine the system's outputs until they are accurate iteratively. This interaction fosters trust by giving users greater control over the AI's response quality. In P9, the HCMPI method is recommended to reduce data dimensions, focusing on extracting only relevant K-dimensional information for healthcare chatbot systems. This reduction makes the AI's reasoning clearer and more interpretable, allowing users to follow the underlying logic without getting overwhelmed by excessive data. These strategies aim to enhance AI systems by making them more interactive and interpretable.

- *User Empowerment and Transparency in AI Interactions*: Empowering users and ensuring transparency in AI systems are critical for fostering ethical interactions and trust, this theme is covered in P1, P2, P5, P8, P10, P11, P12, P13, P16, P17, P24, P28, P29, P30, P31, P33. In P2, a recommendation is made to emphasize critical reflection throughout conversational AI's design and development phases. This includes giving users more control over their interactions with AI agents and being transparent about the AI's non-human nature and limitations. Such transparency allows users to understand the AI's capabilities and limitations, enabling more informed decision-making. In P8, a mitigation strategy involves techniques like logit output verification and proactive detection of hallucinations. These strategies are paired with participatory design, where users actively shape AI systems, ensuring that the AI's responses remain accurate and meaningful. In P30, further mitigation strategies include functionality audits to assess whether LLM applications meet their intended goals and impact audits to evaluate how AI affects users, specific groups, and the broader environment. These strategies prioritize user involvement and clarity, fostering a transparent and user-centered AI ecosystem.

- *Accountability and Continuous Oversight in AI Systems*: Ensuring that AI systems are accountable and continuously monitored is essential to maintain ethical standards and protect users, this theme is covered in P5, P7, P9, P10, P11, P12, P13, P20, P24, P25, P30, P31, P32, P35. In P24, a recommendation is made for AI tools like ChatGPT to be evaluated by regulators, specifically in healthcare settings, to ensure safety, efficacy, and reliable performance. This highlights the importance of oversight from regulatory bodies to ensure that AI applications adhere to established standards. In P9, a mitigation strategy involves the Healthcare Chatbot-based Zero Knowledge Proof (HCZKP) method, which enables the use of data without making it visible. This approach reduces the need for extensive data collection, safeguarding privacy, and ensuring ethical data handling. Additionally, the strategy recommends adopting data minimization principles by collecting only necessary essential data and decentralizing patient data during feedback training. These strategies emphasize the need for ongoing oversight and accountability in how AI systems manage and utilize data, particularly in sensitive environments.

> **RQ2 key Takeaways**
>
> The findings from RQ2 identify various mitigation strategies across the five ethical dimensions of bias, transparency, accountability, privacy, and safety. Strategies include bias mitigation techniques to promote fairness, enhancing interpretability to build transparency and user trust, and fostering accountability through regulatory oversight. Privacy is safeguarded through data minimization practices, while safety is reinforced by interdisciplinary frameworks that adapt ethical standards to evolving AI technologies.
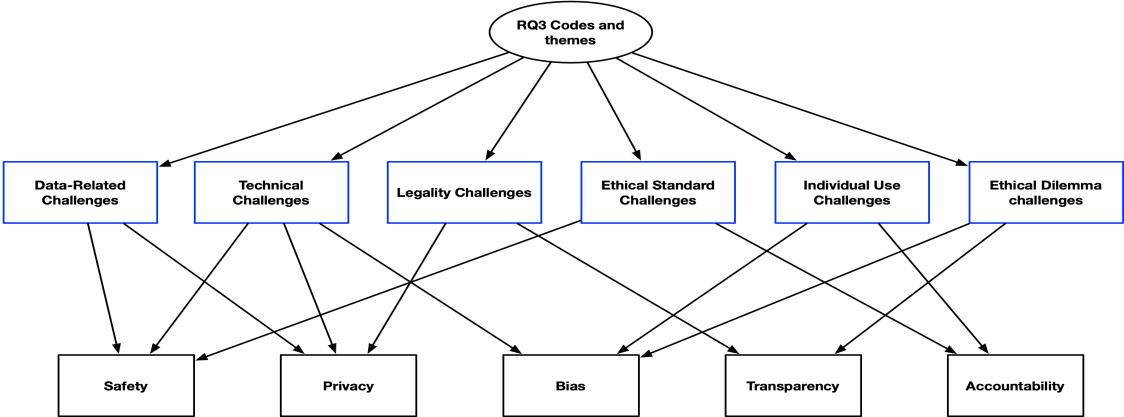
4.4. RQ3 Results

Figure 6. RQ3 codes and themes mapping

While numerous mitigation strategies have been proposed to address the ethical concerns surrounding the use of LLMs, many studies indicate that significant challenges persist even after these strategies are implemented. Authors in several papers have explicitly acknowledged that the mitigation efforts, while promising, often fall short due to various technical, legal, and governance barriers. These challenges highlight the complexity of achieving effective ethical governance in LLM applications. To better understand these issues, we have categorized the identified challenges into key themes, each reflecting the unique difficulties encountered during the practical implementation of these mitigation strategies.

- *Technical Challenges*: Technical challenges in implementing mitigation strategies for LLMs often revolve around the complexity of integrating advanced techniques within diverse and sensitive contexts, this theme is covered in P1, P2, P4, P7, P10, P11, P12, P14, P20, P21, P22, P25, P28, P29, P30, P32, P33. In P1, the use of Natural Language Processing (NLP) on Electronic Health Records (EHR) data highlights multiple challenges, such as accurately identifying clinical entities,

ensuring privacy protection, handling spelling errors, and managing the de-identification of sensitive information. These challenges are compounded by the lack of labeled data, the difficulty of detecting negations, and the complexities of deciphering numerous medical abbreviations. In P2, Gonen and Goldberg (2019) critique the current debiasing methods applied to word vectors. While these techniques produce high scores on self-defined metrics, they often result in only superficially hiding biases rather than genuinely eliminating them. P25 emphasizes that Red Teaming, a method used to identify security vulnerabilities, must be a continuous process and not a one-time solution. It requires a persistent commitment to security throughout the development and operation of AI systems, with careful consideration of both legal and ethical implications. These examples demonstrate the multifaceted technical hurdles that complicate the practical application of ethical mitigation strategies in LLMs.

- *Ethical Standard Challenges*: Implementing ethical standards in the context of LLMs is another challenging task, particularly due to the dynamic and diverse regulatory and cultural environments in which these technologies operate, this theme inclues P2, P8, P10, P30, P32, P34. In P10, the proposed recommendations are based on the analysis of the current regulatory landscape and anticipated regulations by the European Commission. However, as regulations continue to evolve, these recommendations may need to be revised to remain applicable and relevant. This highlights the importance of staying updated with the latest legal requirements in each jurisdiction to ensure ongoing compliance. P34 underscores the significance of incorporating a user-centered approach in cybersecurity, recognizing that addressing human factors can enhance the effectiveness of these measures. Yet, as cybersecurity threats evolve, there is a need for continuous development of strategies to stay ahead of emerging challenges. In P8, it is noted that a one-size-fits-all approach to ethical standards is unlikely to be effective, as LLMs span diverse cultural and global contexts. The absence of standardized or universally regulated frameworks for LLMs adds to the complexity, requiring adaptable and context-sensitive solutions. These examples illustrate the challenges of creating consistent and effective ethical standards in a rapidly changing and diverse landscape.

- *Ethical Dilemma Challenges*: Ethical dilemmas in the implementation of LLMs often arise from conflicting values and considerations, especially when navigating fairness, transparency, and privacy. In P3, the issue of learned associations within AI models highlights a significant ethical dilemma, this theme is covered in P3, P5, P8, P9, P10. For example, in P3, language models trained on data that frequently describe suspects and criminals as "black males" may reinforce stereotypes, raising concerns about legal equality. This brings up the ethical question of whether such associations, which exist in the training data, should be allowed in AI systems deployed for legal purposes, where fairness and equality are paramount. In P5, another ethical dilemma is identified in balancing transparency and protecting non-public information. While transparency is crucial for understanding an AI system's behavior and decision-making process, companies often prioritize commercial secrecy to protect proprietary technology and business models. These conflicts illustrate the complexities of ethical decision-making in AI, where satisfying one ethical principle may compromise another, necessitating careful consideration and context-specific solutions.

- *Individual Use Challenges*: Implementing mitigation strategies for LLMs encounters specific challenges related to how individuals use and interpret AI-generated content, this theme is covered in P13, P25, P27, P30. In P13, even with established guidelines and recommendations, reports have surfaced in the healthcare and wellness sectors where individuals consult LLMs like ChatGPT for health-related matters and take its advice without proper scrutiny. Despite mitigation efforts, the widespread sharing of such AI-generated content on social media demonstrates the difficulty in ensuring that users critically assess AI advice. In P25, engaging multiple stakeholders introduces further complications, even when measures are in place to involve diverse perspectives. Some stakeholders may use their involvement to advance personal agendas, reinforce existing biases, or misuse sensitive data, undermining transparency and ethical intentions. These challenges

illustrate the complexities in managing individual behavior and ensuring responsible AI use, despite mitigation strategies aimed at fostering ethical and informed engagement with LLMs.

- *Legality Challenges*: Legal challenges in implementing mitigation strategies for LLMs often involve navigating the complexities of intellectual property, censorship, and transparency, this theme is covered in P7, P8, P30. In P7, one significant concern is the potential leakage of business secrets and proprietary information when users interact with LLMs like ChatGPT. If proprietary code is inadvertently shared during AI interactions, it may become part of the chatbot's knowledge base, raising issues of copyright infringement and the preservation of business confidentiality. This poses a risk for organizations that depend on protecting sensitive information. In P8, censorship within LLMs, while intended to prevent harmful outputs, introduces legal dilemmas. There is often no clear or objective standard for determining what content is harmful, which can lead to the suppression of free speech or creative expression. Overly restrictive censorship may also hinder important debates, while a lack of transparency around censorship policies can create distrust in the AI system. These examples underscore the legal complexities of implementing effective mitigation strategies, where balancing ethical considerations with regulatory compliance is a persistent challenge.

- *Data-related Challenges*: Data-related challenges are a critical factor in the implementation of mitigation strategies for LLMs, as they affect the quality, availability, and reliability of the datasets used in AI development, this theme is covered in P11, P12, P28, P32. In P28, concerns are raised about the future limitations of data collection and usage in machine learning. Research indicates that high-quality language data could be exhausted by 2026, with lower-quality data potentially running out by 2060. This forecast suggests that the limited availability of suitable datasets may constrain the future development and improvement of LLMs, affecting their ability to perform effectively and ethically. In P32, the quality of data is further questioned, as a substantial portion of source material comes from preprint servers that lack rigorous peer review. This reliance on unverified data can limit the generalizability and reliability of LLMs, particularly when data is drawn from diverse and variable contexts. These examples highlight the significant data challenges faced when attempting to implement mitigation strategies, where the quality, scope, and future availability of data play a crucial role in shaping the effectiveness of LLM interventions.

> **RQ3 Key Takeaways**
>
> The RQ3 findings reveal significant challenges in applying mitigation strategies for LLMs, spanning technical, ethical, individual, legal, and data-related obstacles. Technical challenges arise from the complexity of integrating mitigation techniques in diverse and sensitive contexts, while ethical standards are difficult to maintain across varied regulatory landscapes and cultural settings. Ethical dilemmas further complicate implementation, as strategies often conflict with values like fairness and transparency. Individual use challenges highlight the difficulty of ensuring users critically engage with AI-generated content, even with guidelines in place. Legal barriers around intellectual property, censorship, and transparency introduce further complexity, and data-related issues, including data quality and availability, present constraints for sustainable LLM development.

5. Discussion

Our systematic mapping study reveals that the ethical concerns surrounding the deployment and usage of LLMs are multifaceted and have evolved significantly over recent years. This complexity is reflected in the increasing volume of research and the diverse range of mitigation strategies aimed at addressing these ethical dimensions. In this section, we analyze the prominent ethical issues identified, including safety, transparency, accountability, privacy, and bias. These dimensions, discussed across

the selected studies, underscore a need for robust ethical frameworks and regulatory oversight to responsibly guide LLM deployment. Additionally, we discuss the challenges encountered in implementing these strategies, highlighting gaps that require further research and practical solutions to support ethical AI deployment.

5.1. Contextual Significance of Ethical Dimensions Accross Different Domains

In examining the RQ1 data, we noticed that the significance of each ethical dimension—safety, privacy, accountability, bias, and transparency—varies across domains. This suggests that the ethical concerns identified span a range of relevance depending on the application context. For instance, safety is particularly emphasized in public safety applications within the education sector, where autonomous systems must prioritize safeguarding individuals, especially in student-focused scenarios [51]. Conversely, safety appears less frequently in general AI applications where the primary concerns shift towards bias and transparency, suggesting that safety might hold less immediate significance in these broader AI use cases [52].

Privacy emerges as a critical issue in healthcare, where data sensitivity is high, particularly for applications in AI privacy frameworks and general use of LLMs [17]. This contrasts with economics-focused applications, where privacy remains essential but is balanced with other ethical considerations like transparency in auditing contexts. In healthcare, protecting user data aligns with legal mandates like HIPAA, making privacy a non-negotiable dimension, while in economic applications, transparency takes precedence due to regulatory oversight needs [53].

Bias shows a strong presence in legal and economic contexts, specifically regarding predictive models that could exacerbate racial or socioeconomic disparities. Bias is not as prominently discussed in more general AI applications, where transparency and interpretability may be prioritized over fairness, especially in domains where the risks associated with bias are not as pronounced [54].

Transparency is crucial across most domains but is particularly highlighted in the use of LLMs for public-facing applications, such as in education and healthcare, where trustworthiness directly impacts user experience and compliance. For example, in economic applications, transparency is essential in auditing AI outputs, ensuring that stakeholders can trust and verify model decisions [55].

Accountability, while universally relevant, is most emphasized in contexts with high legal or public stakes, such as healthcare and government AI applications, where the ability to trace and attribute decisions is essential for compliance and ethical standards [56]. In more experimental AI applications, accountability is less prioritized, with a stronger focus instead on developing fundamental functionalities [57].

5.2. The need for Continuous Evaluation and Iterative Improvements

The ethical concerns surrounding the use of Large Language Models (LLMs), including bias, privacy, and transparency, are complex and continuously evolving. These issues are compounded by the dynamic nature of LLM deployments, which means that ethical strategies cannot remain static but must adapt to emergent challenges. Although frameworks such as the EU AI Act and the National Institute of Standards and Technology (NIST) emphasize the importance of responsive, adaptable frameworks, the real-world application of this principle remains limited. Most current ethical approaches primarily focus on initial implementation, with less attention to iterative evaluation, leaving gaps in sustained effectiveness. For instance, Nigam Shah et al. highlight the critical need for ongoing assessment in AI prediction algorithms to ensure long-term reliability and ethical alignment [58].

An additional concern is the lack of empirical evaluations of these strategies. Despite the recognized need for robust testing of bias mitigation, privacy safeguards, and transparency practices, relatively few systematic studies assess these methods in practical, high-stakes environments [50]. This lack of empirical evidence means many strategies are based on theoretical or conceptual frameworks rather than validated, real-world outcomes. Consequently, organizations may find it difficult to anticipate how well these ethical measures will hold up over time or under evolving conditions.

The shifting landscape of AI ethics, driven by advances in data science and changing societal norms, also creates challenges for maintaining ethical standards. Strategies designed for earlier LLMs may not be effective for models trained on more diverse or sensitive datasets, where biases may arise unexpectedly, or privacy risks could be amplified. As AI continues to develop, the frameworks that govern its deployment must keep pace, ensuring that ethical practices evolve alongside technological progress [59]. The limitations of "checkbox" ethics, as noted by Sara Kijewski et al., further highlight the importance of deeper, more adaptable standards that extend beyond initial implementation [60].

Frameworks like Microsoft's AI Standard and IEEE's guidelines advocate for continuous accountability, reinforcing the notion that ethical oversight should be an ongoing process rather than a one-time checkpoint. These frameworks aim to create adaptable, accountable structures that can evolve in response to new insights and societal changes [61]. However, implementing such adaptable strategies is challenging and requires a commitment to monitoring, reviewing, and refining ethical measures throughout the lifecycle of an AI system. The role of regular auditing, as highlighted by Li and Goel, is particularly essential in supporting ongoing ethical evaluation and accountability [62].

> **Recommendation**
>
> We recommend that mitigation approaches focus on addressing emerging biases, security vulnerabilities, and aligning with legal standards, particularly those related to evolving datasets. Ethical strategies should not remain static but should adapt over time. It is essential to prioritize accountability and support the development of flexible and evolving frameworks and guidelines.

5.3. Achieving Cross-Framework Consistency

Implementing ethical standards in Artificial Intelligence (AI) is often complicated by the lack of harmonized guidelines across various frameworks and domains. For instance, the European Union's AI Act provides explicit regulatory requirements for privacy and bias mitigation, whereas frameworks from organizations like the NIST and IEEE offer guidelines that are often voluntary and context-specific [63]. This discrepancy can lead to inconsistencies, especially when Large Language Models (LLMs) are deployed across regions with differing regulatory landscapes.

The absence of standardized global ethical principles poses significant challenges for international organisations. Without a unified foundation, companies may struggle to ensure consistency in ethical AI practices, leading to potential ethical lapses and legal complications [64]. This fragmentation is evident in the varying approaches to AI ethics across different jurisdictions, as highlighted by Hagendorff, who notes that the proliferation of AI ethics guidelines has led to a "bewildering variety" of recommendations, often lacking coherence and practical applicability [54].

Moreover, the dynamic nature of AI technologies necessitates adaptable ethical frameworks that can evolve with technological advancements. However, the current lack of cross-framework consistency hampers the development of such adaptable guidelines. Vakkuri and Kemell emphasize the importance of implementing AI ethics in practice through empirical evaluation, yet they acknowledge the challenges posed by the absence of standardized ethical frameworks [65].

> **Recommendation**
>
> We recommend establishing a set of standardized global ethical principles, providing a unified foundation on which different frameworks can develop tailored guidance for specific domains. This coordinated approach would support companies operating across jurisdictions by ensuring consistency in ethical AI practices while allowing flexibility for domain-specific requirements.

5.4. The lack of Scalability of Mitigations

Implementing ethical mitigation strategies for LLMs presents significant scalability challenges, particularly when these models are deployed across diverse and high-volume contexts [66]. Ensuring consistent application of bias detection, privacy protections, and transparency measures across various domains is a complex endeavor. Bias mitigation, for instance, often requires specialized expertise and substantial resources, which may not be accessible to all organizations, leading to uneven ethical standards across AI applications [67].

Scalability is especially critical in high-risk sectors such as healthcare and finance, where ethical lapses can have serious consequences. The EU AI Act and NIST guidelines recognize the importance of scalable solutions in these contexts. However, translating high-level guidelines into practical, domain-specific applications remains a challenge, as detailed by Fjeld et al., who examine the limitations of current AI frameworks in meeting diverse operational requirements across sectors [68].

One pressing issue is the need for tools and frameworks to adapt to the varied operational scales at which LLMs are deployed. Binns et al. highlight this challenge, noting that most bias mitigation strategies lack the flexibility to be scaled up effectively, often requiring case-specific customization that hampers broader application [69]. Furthermore, the resource-intensive nature of comprehensive ethical compliance poses additional barriers for smaller organizations, which may lack the funding to implement scalable solutions [70,71].

> **Recommendation**
>
> We recommend building scalability into the core guidelines and frameworks for large-scale LLMs, ensuring they support the creation of ethical AI solutions that are accessible and inclusive.

5.5. Engaging with End Users and Interdisciplinary Collaboration

The ethical deployment of LLMs necessitates active engagement with end users and robust interdisciplinary collaboration. Historically, AI development has often prioritized technical performance over societal impact, leading to unintended consequences such as bias and reduced trust among users [72]. Involving end users, especially those from marginalized communities disproportionately affected by AI biases, is crucial for aligning LLMs with societal values and human rights [73].

Interdisciplinary collaboration is equally vital in the ethical implementation of LLMs. Bringing together technical experts, legal professionals, ethicists, and domain specialists facilitates the integration of ethical considerations into technical processes [74]. Such collaborations have been initiated by organizations like the European Union and IEEE, setting a precedent for effective cooperation [75].

However, challenges persist in achieving meaningful engagement and collaboration. Power imbalances, differing priorities, and communication barriers can hinder effective participation [76]. Addressing these challenges requires deliberate efforts to create inclusive environments where diverse perspectives are valued and integrated into the AI development process [77].

> **Recommendation**
>
> We encourage LLM developers to work closely with marginalized communities, often the most affected by AI biases or overlooked during development. Building relationships through open and inclusive conversations can foster trust, improve understanding, and ensure that LLM systems are designed and used in ways that respect societal values and protect human rights.

6. Threats to Validity

In this section, we discuss the threats of validity in our study, the steps we have taken to mitigate them

Internal Validity: One of the key threats to internal validity in a systematic mapping study (SMS) is selection bias, which can arise from subjective interpretation during the study selection process. To mitigate this, we employed multiple strategies. We identified a set of keywords considered relevant to the ethical use of LLMs, tested them, and consulted a university librarian to refine the keywords. We conducted searches across six databases to ensure a broader variety of studies could be included. Pilot tests were performed and validated by all authors to ensure the reliability of the results. Additionally, forward snowballing was performed to capture studies that may not have been initially identified.

Construct Validity: Construct validity in our study refers to the extent to which the selected studies are relevant and appropriate to our research goals. To address this, we selected papers that directly aligned with our research questions (RQs) and excluded papers that focused solely on LLMs without discussing the ethical issues associated with their use or deployment. We also held meetings and discussions to establish the inclusion and exclusion criteria for the selected papers. The timeframe for our SMS was set between 2023 and July 2024, so any study published after July 2024 would not be reflected in our results.

Conclusion Validity: One of the major threats to conclusion validity in systematic mapping studies is bias in data extraction. In our study, the data extraction process was guided by our research questions (RQs), ensuring that the selected data was directly relevant to our study objectives. To mitigate potential bias, we used Google Forms to facilitate the coding process, enabling us to categorize data systematically through thematic analysis. This approach allowed us to group the findings based on predefined codes and themes derived from existing literature. As new themes emerged during coding, they were incorporated as needed. The coding was collaboratively performed by the first, second, and third authors to minimize individual bias. We held regular meetings to discuss and refine the data extraction and analysis processes, ensuring agreement on selecting relevant data and how it should be presented.

7. Conclusion

In conclusion, this study offers a comprehensive examination of ethical concerns surrounding LLMs, categorized into five primary dimensions—safety, transparency, accountability, privacy, and bias—developed from a synthesis of selected guidelines and existing literature. Within these dimensions, numerous ethical themes emerge, highlighting complex issues such as the potential for LLMs to propagate misinformation, the opacity in model decision-making, challenges in establishing accountability, risks to user privacy, and the perpetuation of social biases. Our findings reveal that, although numerous mitigation strategies have been proposed to address these dimensions, substantial implementation challenges remain due to technical complexities, evolving regulatory frameworks, ethical dilemmas, and user awareness and data quality issues. These obstacles highlight the need for adaptable, interdisciplinary ethical frameworks/guidelines capable of evolving alongside AI advancements. By elucidating these ethical dimensions and the practical difficulties in deploying mitigation strategies, this study contributes valuable insights to the discourse on responsible AI, informing future research, policy, and best practices to align LLM development with ethical standards better.

Appendix A. Selected Primary Studies

P1: Sathe, N., Deodhe, V., Sharma, Y., & Shinde, A. (2023, December). A Comprehensive Review of AI in Healthcare: Exploring Neural Networks in Medical Imaging, LLM-Based Interactive Response Systems, NLP-Based EHR Systems, Ethics, and Beyond. In 2023 International Conference on Advanced Computing & Communication Technologies (ICACCTech) (pp. 633-640). IEEE.

P2: Ruane, E., Birhane, A., & Ventresque, A. (2019). Conversational AI: Social and Ethical Considerations. AICS, 2563, 104-115.

P3: Malic, V. Q., Kumari, A., & Liu, X. (2023, December). Racial skew in fine-tuned legal AI language models. In 2023 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 245-252). IEEE.

P4: Bansal, R. (2022). A survey on bias and fairness in natural language processing. arXiv preprint arXiv:2204.09591.

P5: Bang, J., Lee, B. T., & Park, P. (2023, August). Examination of ethical principles for llm-based recommendations in conversational ai. In 2023 International Conference on Platform Technology and Service (PlatCon) (pp. 109-113). IEEE.

P6: Lofstead, J. (2023, September). Economic, Societal, Legal, and Ethical Considerations for Large Language Models. In 2023 Fifth International Conference on Transdisciplinary AI (TransAI) (pp. 155-162). IEEE.

P7: Khoury, R., Avila, A. R., Brunelle, J., & Camara, B. M. (2023, October). How secure is code generated by chatgpt?. In 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 2445-2451). IEEE.

P8: Jiao, J., Afroogh, S., Xu, Y., & Phillips, C. (2024). Navigating llm ethics: Advancements, challenges, and future directions. arXiv preprint arXiv:2406.18841.

P9: Cai, Z., Chang, X., & Li, P. (2023, November). HCPP: A Data-Oriented Framework to Preserve Privacy during Interactions with Healthcare Chatbot. In 2023 IEEE International Performance, Computing, and Communications Conference (IPCCC) (pp. 283-290). IEEE.

P10: Piñeiro-Martín, A., García-Mateo, C., Docío-Fernández, L., & Lopez-Perez, M. D. C. (2023). Ethical challenges in the development of virtual assistants powered by large language models. Electronics, 12(14), 3170.

P11: Parray, A. A., Inam, Z. M., Ramonfaur, D., Haider, S. S., Mistry, S. K., & Pandya, A. K. (2023). ChatGPT and global public health: applications, challenges, ethical considerations and mitigation strategies.

P12: Khan, M. S., & Umer, H. (2024). ChatGPT in finance: Applications, challenges, and solutions. Heliyon, 10(2).

P13: Harrer, S. (2023). Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine, 90.

P14: Chauncey, S. A., & McKenna, H. P. (2023). A framework and exemplars for ethical and responsible use of AI Chatbot technology to support teaching and learning. Computers and Education: Artificial Intelligence, 5, 100182.

P15: Ansarullah, S. I., Kirmani, M. M., Alshmrany, S., & Firdous, A. (2024). Ethical issues around artificial intelligence. In A Biologist s Guide to Artificial Intelligence (pp. 301-314). Academic Press.

P16: Patton, D. U., Landau, A. Y., & Mathiyazhagan, S. (2023). ChatGPT for social work science: Ethical challenges and opportunities. Journal of the Society for Social Work and Research, 14(3), 553-562.

P17: Wu, X., Duan, R., & Ni, J. (2024). Unveiling security, privacy, and ethical concerns of ChatGPT. Journal of Information and Intelligence, 2(2), 102-115.

P18: 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.

P19: Oviedo-Trespalacios, O., Peden, A. E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod, J. E., ... & Reniers, G. (2023). The risks of using ChatGPT to obtain common safety-related information and advice. Safety science, 167, 106244.

P20: Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. New directions for evaluation, 2023(178-179), 33-46.

P21: MacIntyre, M. R., Cockerill, R. G., Mirza, O. F., & Appel, J. M. (2023). Ethical considerations for the use of artificial intelligence in medical decision-making capacity assessments. Psychiatry research, 328, 115466.

P22: Grote, T., & Berens, P. (2024). A paradigm shift?—On the ethics of medical large language models. Bioethics, 38(5), 383-390.

P23: Behnia, R., Ebrahimi, M. R., Pacheco, J., & Padmanabhan, B. (2022, November). Ew-tune: A framework for privately fine-tuning large language models with differential privacy. In 2022 IEEE International Conference on Data Mining Workshops (ICDMW) (pp. 560-566). IEEE.

P24: Wang, C., Liu, S., Yang, H., Guo, J., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. Journal of Medical Internet Research, 25, e48009.

P25: Mitsunaga, T. (2023, October). Heuristic Analysis for Security, Privacy and Bias of Text Generative AI: GhatGPT-3.5 case as of June 2023. In 2023 IEEE International Conference on Computing (ICOCO) (pp. 301-305). IEEE.

P26: Gan, W., Qi, Z., Wu, J., & Lin, J. C. W. (2023, December). Large language models in education: Vision and opportunities. In 2023 IEEE international conference on big data (BigData) (pp. 4776-4785). IEEE.

P27: Kshetri, N. (2023). Cybercrime and privacy threats of large language models. IT Professional, 25(3), 9-13.

P28: Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity. arXiv preprint arXiv:2301.12867.

P29: Zhang, X., Xu, H., Ba, Z., Wang, Z., Hong, Y., Liu, J., ... & Ren, K. (2024). Privacyasst: Safeguarding user privacy in tool-using large language model agents. IEEE Transactions on Dependable and Secure Computing.

P30: Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). Auditing large language models: a three-layered approach. AI and Ethics, 1-31.

P31: Curzon, J., Kosa, T. A., Akalu, R., & El-Khatib, K. (2021). Privacy and artificial intelligence. IEEE Transactions on Artificial Intelligence, 2(2), 96-108.

P32: Bano, M., Hoda, R., Zowghi, D., & Treude, C. (2024). Large language models for qualitative research in software engineering: exploring opportunities and challenges. Automated Software Engineering, 31(1), 8.

P33: Khoje, M. (2024, February). Navigating Data Privacy and Analytics: The Role of Large Language Models in Masking conversational data in data platforms. In 2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC) (pp. 1-5). IEEE.

P34: Jeyaraman, M., Balaji, S., Jeyaraman, N., & Yadav, S. (2023). Unraveling the ethical enigma: artificial intelligence in healthcare. Cureus, 15(8).

P35: Ferrara, E. (2023). Should chatgpt be biased? challenges and risks of bias in large language models. arXiv preprint arXiv:2304.03738.

P36: Akinci D'Antonoli, T., Stanzione, A., Bluethgen, C., Vernuccio, F., Ugga, L., Klontzas, M. E., ... & Koçak, B. (2023). Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. Diagnostic and Interventional Radiology, Epub-ahead.

P37: Belzner, L., Gabor, T., & Wirsing, M. (2023, October). Large language model assisted software engineering: prospects, challenges, and a case study. In International Conference on Bridging the Gap between AI and Reality (pp. 355-374). Cham: Springer Nature Switzerland.

P38: Liyanage, U. P., & Ranaweera, N. D. (2023). Ethical considerations and potential risks in the deployment of large language models in diverse societal contexts. Journal of Computational Social Dynamics, 8(11), 15-25.

P39: Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative ai. Educational Measurement: Issues and Practice, 43(2), 16-29.

Appendix B. Database Search Strings

Table A1. Database Search Strings

Database	Search Strings
IEEE	((("Large Language Model*" OR LLMs) AND (guideline* OR standard* OR framework* OR compliance OR principles OR practices OR governance OR impact OR oversight OR algorithmic OR policy OR policies) AND (development OR deployment OR use OR design OR implementation) AND (ethics OR ethical OR moral OR bias OR fairness OR transparency OR accountability OR privacy OR security OR sustainability OR responsible OR trustworthiness OR equit* OR inclus* OR diversity OR legal OR rights OR cultural))
ACM	((([All: "large language model*"] OR [All: llms]) AND ([All: guideline*] OR [All: standard*] OR [All: framework*] OR [All: compliance] OR [All: principles] OR [All: practices] OR [All: governance] OR [All: impact] OR [All: oversight] OR [All: algorithmic] OR [All: policy] OR [All: policies] AND ([All: development] OR [All: use] OR [All: design] OR [All: implementation]) AND ([All: ethics] OR [All: ethical] OR [All: moral] OR [All: bias] OR [All: fairness] OR [All: transparency] OR [All: accountability] OR [All: privacy] OR [All: security] OR [All: sustainability] OR [All: responsible] OR [All: trustworthiness] OR [All: equit*] OR [All: inclus*] OR [All: diversity] OR [All: legal] OR [All: rights] OR [All: cultural]))
ProQuest	((noft("Large Language Model*") OR noft(LLMs)) AND (noft(guideline*) OR noft(standard*) OR noft(framework*) OR noft(compliance) OR noft(principles) OR noft(practices) OR noft(governance) OR noft(impact) OR noft(oversight) OR noft(algorithmic) OR noft(policy) OR noft(policies)) AND (noft(development) OR noft(deployment) OR noft(use) OR noft(design) OR noft(implementation)) AND (noft(ethics) OR noft(ethical) OR noft(moral) OR noft(bias) OR noft(fairness) OR noft(transparency) OR noft(accountability) OR noft(privacy) OR noft(security) OR noft(sustainability) OR noft(responsible) OR noft(trustworthiness) OR noft(equit*) OR noft(inclus*) OR noft(diversity) OR noft(legal) OR noft(rights) OR noft(cultural)))
Web of Science	(("Large Language Model*" OR llms) AND (guideline* OR standard* OR framework* OR compliance OR principles OR practices OR governance OR impact OR oversight OR algorithmic OR policy OR policies) AND (development OR deployment OR use OR design OR implementation) AND (ethics OR ethical OR moral OR bias OR fairness OR transparency OR accountability OR privacy OR security OR sustainability OR responsible OR trustworthiness OR equit* OR inclus* OR diversity OR legal OR rights OR cultural))
Wiley Online Library	(("Large Language Model*" OR llms) AND (guideline* OR standard* OR framework* OR compliance OR principles OR practices OR governance OR impact OR oversight OR algorithmic OR policy OR policies) AND (development OR deployment OR use OR design OR implementation) AND (ethics OR ethical OR moral OR bias OR fairness OR transparency OR accountability OR privacy OR security OR sustainability OR responsible OR trustworthiness OR equit* OR inclus* OR diversity OR legal OR rights OR cultural))
Science Direct	("Large Language Model" OR "LLMs") AND ("guideline" OR "standard") AND ("development" OR "design") AND ("ethics" OR "moral")

References

1. M. Alawida, S. Mejri, A. Mehmood, B. Chikhaoui, O. Isaac Abiodun, A comprehensive study of chatgpt: advancements, limitations, and ethical considerations in natural language processing and cybersecurity, Information 14 (8) (2023) 462.

2. C. Arora, J. Grundy, M. Abdelrazek, Advancing requirements engineering through generative ai: Assessing the role of llms, in: Generative AI for Effective Software Development, Springer, 2024, pp. 129–148.

3. A. A. Linkon, M. Shaima, M. S. U. Sarker, N. Nabi, M. N. U. Rana, S. K. Ghosh, M. A. Rahman, H. Esa, F. R. Chowdhury, et al., Advancements and applications of generative artificial intelligence and large language models on business management: A comprehensive review, Journal of Computer Science and Technology Studies 6 (1) (2024) 225–232.

4. D. H. Hagos, R. Battle, D. B. Rawat, Recent advances in generative ai and large language models: Current status, challenges, and perspectives, IEEE Transactions on Artificial Intelligence (2024).

5. S. Bengesi, H. El-Sayed, M. K. Sarker, Y. Houkpati, J. Irungu, T. Oladunni, Advancements in generative ai: A comprehensive review of gans, gpt, autoencoders, diffusion model, and transformers., IEEE Access (2024).

6. S. Reddy, Generative ai in healthcare: an implementation science informed translational path on application, integration and governance, Implementation Science 19 (1) (2024) 27.

7. Y. J. P. Bautista, C. Theran, R. Aló, Ethical considerations of generative ai: A survey exploring the role of decision makers in the loop, in: Proceedings of the AAAI Symposium Series, Vol. 3, 2024, pp. 391–398.

8. N. Bontridder, Y. Poullet, The role of artificial intelligence in disinformation, Data & Policy 3 (2021) e32.

9. F. Germani, G. Spitale, N. Biller-Andorno, The dual nature of ai in information dissemination: Ethical considerations, JMIR AI 3 (2024) e53505–.

10. T. W. Sanchez, M. Brenman, X. Ye, The ethical concerns of artificial intelligence in urban planning, Journal of the American Planning Association (2024) 1–14.

11. A. Rezaei Nasab, M. Dashti, M. Shahin, M. Zahedi, H. Khalajzadeh, C. Arora, P. Liang, Fairness concerns in app reviews: A study on ai-based mobile apps, ACM Transactions on Software Engineering and Methodology (2024).

12. B. Kitchenham, L. Madeyski, D. Budgen, Segress: Software engineering guidelines for reporting secondary studies, IEEE Transactions on Software Engineering 49 (3) (2022) 1273–1298.

13. K. Petersen, S. Vakkalanka, L. Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Information and software technology 64 (2015) 1–18.

14. J. Dewey, J. H. Tufts, Ethics, DigiCat, 2022.

15. S. Sivasubramaniam, D. H. Dlabolová, V. Kralikova, Z. R. Khan, Assisting you to advance with ethics in research: an introduction to ethical governance and application procedures, International Journal for Educational Integrity 17 (2021) 1–18.

16. D. C. Poff, Academic ethics and academic integrity, in: Encyclopedia of business and professional ethics, Springer, 2023, pp. 11–16.

17. A. Jobin, M. Ienca, E. Vayena, The global landscape of ai ethics guidelines, Nature machine intelligence 1 (9) (2019) 389–399.

18. S. L. Anderson, M. Anderson, Ai and ethics, AI and Ethics 1 (1) (2021) 27–31.

19. J.-C. Põder, Ai ethics–a review of three recent publications (2021).

20. E. Kazim, A. S. Koshiyama, A high-level overview of ai ethics, Patterns 2 (9) (2021).

21. M. Coeckelbergh, AI ethics, Mit Press, 2020.

22. L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations, Minds and machines 28 (2018) 689–707.

23. R. Eitel-Porter, Beyond the promise: implementing ethical ai, AI and Ethics 1 (1) (2021) 73–80.

24. E. Hickman, M. Petrin, Trustworthy ai and corporate governance: the eu's ethics guidelines for trustworthy artificial intelligence from a company law perspective, European Business Organization Law Review 22 (2021) 593–625.

25. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

26. A. Vaswani, Attention is all you need, Advances in Neural Information Processing Systems (2017).

27. T. B. Brown, Language models are few-shot learners, arXiv preprint arXiv:2005.14165 (2020).

28. V. Alto, Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT3 and GPT4, Packt Publishing Ltd, 2023.

29. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

30. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (5) (2019) 206–215.

31. R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE symposium on security and privacy (SP), IEEE, 2017, pp. 3–18.

32. J. Li, W. Xiao, C. Zhang, Data security crisis in universities: identification of key factors affecting data breach incidents, Humanities and Social Sciences Communications 10 (1) (2023) 1–18.

33. T. Madiega, Artificial intelligence act, European Parliament: European Parliamentary Research Service (2021).

34. E. Jillson, Aiming for truth, fairness, and equity in your company's use of ai, Federal Trade Commission 19 (2021).

35. N. AI, Artificial intelligence risk management framework: Generative artificial intelligence profile (2024).

36. S. Migliorini, China's interim measures on generative ai: Origin, content and significance, Computer Law & Security Review 53 (2024) 105985.

37. I. S. Association, et al., The ieee global initiative on ethics of autonomous and intelligent systems. ieee. org. retrieved march 12, 2021 (2017).

38. S. Pichai, Ai at google: our principles, The Keyword 7 (2018) (2018) 1–3.

39. OpenAI, Openai safety standards, https://openai.com/safety-standards/ (2024).

40. M. C. Buiten, Towards intelligent regulation of artificial intelligence, European Journal of Risk Regulation 10 (1) (2019) 41–59.

41. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2, http://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf (2017).

42. N. AI, Artificial intelligence risk management framework (ai rmf 1.0) (2023).

43. Microsoft, Microsoft responsible ai standard, version 2: General requirements, https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf (2022).

44. A. Oketunji, M. Anas, D. Saina, Large language model (llm) bias index—llmbi, Data & Policy (2023).

45. M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

46. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial intelligence and statistics, PMLR, 2017, pp. 1273–1282.

47. P. on AI, Annual report 2021 (2021).
     URL https://partnershiponai.org/resource/annual-report-2021/

48. F. Li, N. Ruijs, Y. Lu, Ethics & ai: A systematic review on ethical concerns and related strategies for designing with ai in healthcare, Ai 4 (1) (2022) 28–53.

49. E. Atlam, M. Almaliki, A. Alfahaid, I. Gad, G. Elmarhomy, M. Alwateer, A. Ahmed, Slm-aie: A systematic literature map of artificial intelligence ethics (2024).

50. G. Palumbo, D. Carneiro, V. Alves, Objective metrics for ethical ai: a systematic literature review, International Journal of Data Science and Analytics (2024) 1–21.

51. D. Leslie, Understanding artificial intelligence ethics and safety, arXiv preprint arXiv:1906.05684 (2019).

52. L. Floridi, J. Cowls, A unified framework of five principles for ai in society, Machine learning and the city: Applications in architecture and urban design (2022) 535–545.

53. V. Dignum, Responsible artificial intelligence: how to develop and use AI in a responsible way, Vol. 2156, Springer, 2019.

54. T. Hagendorff, The ethics of ai ethics: An evaluation of guidelines, Minds and machines 30 (1) (2020) 99–120.

55. J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices, Science and engineering ethics 26 (4) (2020) 2141–2168.

56. A. Anawati, H. Fleming, M. Mertz, J. Bertrand, J. Dumond, S. Myles, J. Leblanc, B. Ross, D. Lamoureux, D. Patel, et al., Artificial intelligence and social accountability in the canadian health care landscape: A rapid literature review, PLOS Digital Health 3 (9) (2024) e0000597.

57. H. Smith, Clinical ai: opacity, accountability, responsibility and liability, Ai & Society 36 (2) (2021) 535–545.

58. N. H. Shah, M. A. Pfeffer, M. Ghassemi, The need for continuous evaluation of artificial intelligence prediction algorithms, JAMA Network Open 7 (9) (2024) e2433009–e2433009.

59. R. Ortega-Bolaños, J. Bernal-Salcedo, M. Germán Ortiz, J. Galeano Sarmiento, G. A. Ruz, R. Tabares-Soto, Applying the ethics of ai: a systematic review of tools for developing and assessing ai-based systems, Artificial Intelligence Review 57 (5) (2024) 110.

60. S. Kijewski, E. Ronchi, E. Vayena, The rise of checkbox ai ethics: a review, AI and Ethics (2024) 1–10.

61. M. Srikumar, R. Finlay, G. Abuhamad, C. Ashurst, R. Campbell, E. Campbell-Ratcliffe, H. Hongo, S. R. Jordan, J. Lindley, A. Ovadya, et al., Advancing ethics review practices in ai research, Nature Machine Intelligence 4 (12) (2022) 1061–1064.

62. Y. Li, S. Goel, Making it possible for the auditing of ai: A systematic review of ai audits and ai auditability, Information Systems Frontiers (2024) 1–31.

63. E. Prem, From ethical ai frameworks to tools: a review of approaches, AI and Ethics 3 (3) (2023) 699–716.

64. A. A. Khan, M. A. Akbar, M. Fahmideh, P. Liang, M. Waseem, A. Ahmad, M. Niazi, P. Abrahamsson, Ai ethics: an empirical study on the views of practitioners and lawmakers, IEEE Transactions on Computational Social Systems 10 (6) (2023) 2971–2984.

65. V. Vakkuri, K.-K. Kemell, Implementing ai ethics in practice: An empirical evaluation of the resolvedd strategy, in: Software Business: 10th International Conference, ICSOB 2019, Jyväskylä, Finland, November 18–20, 2019, Proceedings 10, Springer, 2019, pp. 260–275.

66. C. Deng, Y. Duan, X. Jin, H. Chang, Y. Tian, H. Liu, H. P. Zou, Y. Jin, Y. Xiao, Y. Wang, et al., Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas, arXiv preprint arXiv:2406.05392 (2024).

67. A. A. Khan, S. Badshah, P. Liang, M. Waseem, B. Khan, A. Ahmad, M. Fahmideh, M. Niazi, M. A. Akbar, Ethics of ai: A systematic literature review of principles and challenges, in: Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering, 2022, pp. 383–392.

68. J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, M. Srikumar, Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai, Berkman Klein Center Research Publication (2020-1) (2020).

69. R. Binns, Fairness in machine learning: Lessons from political philosophy, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 149–159.

70. W. Hoffmann-Riem, Artificial intelligence as a challenge for law and regulation, Regulating artificial intelligence (2020) 1–29.

71. C. Cath, Governing artificial intelligence: ethical, legal and technical opportunities and challenges, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376 (2133) (2018) 20180080.

72. K. Murphy, E. Di Ruggiero, R. Upshur, D. J. Willison, N. Malhotra, J. C. Cai, N. Malhotra, V. Lui, J. Gibson, Artificial intelligence for good health: a scoping review of the ethics literature, BMC medical ethics 22 (2021) 1–17.

73. A. Hagerty, I. Rubinov, Global ai ethics: a review of the social impacts and ethical implications of artificial intelligence, arXiv preprint arXiv:1907.07892 (2019).

74. S. Pink, E. Quilty, J. Grundy, R. Hoda, Trust, artificial intelligence and software practitioners: an interdisciplinary agenda, AI & SOCIETY (2024) 1–14.

75. M. Al-kfairy, D. Mustafa, N. Kshetri, M. Insiew, O. Alfandi, Ethical challenges and solutions of generative ai: An interdisciplinary perspective, in: Informatics, Vol. 11, MDPI, 2024, p. 58.

76. S. Keles, Navigating in the moral landscape: analysing bias and discrimination in ai through philosophical inquiry, AI and Ethics (2023) 1–11.

77. R. Gianni, S. Lehtinen, M. Nieminen, Governance of responsible ai: From ethical guidelines to cooperative policies, Frontiers in Computer Science 4 (2022) 873437.