UNIVERSITY OF TARTU
Faculty of Social Sciences
School of Economics and Business Administration
Innovation and Technology Management Curriculum

Daria Korobenko

Towards Privacy- and Security-Aware Framework for AI Ethics

Master's Thesis (20 ECTS)

Supervisor(s): Anastasija Nikiforova, PhD
Rajesh Sharma, PhD

Tartu 2024

Towards Privacy- and Security-Aware Framework for AI Ethics

Abstract:
As artificial intelligence (AI) advances, bringing forth more benefits, so does the general unease surrounding the privacy and security of AI-enabled systems. The question of how the development of AI can be controlled on both jurisdictional and organizational levels arises in today's discourse more often than ever. While on the governmental level the European Parliament and Council reached a political agreement on EU AI Act, the world's first comprehensive AI law, organizations still find it challenging to adapt to the fast-evolving AI landscape and lack a universal tool for evaluating privacy and security aspects of their AI models. In response to this, a systematic literature review (SLR) was conducted spanning from 2020 to 2023, aimed at establishing a unified definition of key concepts in AI Ethics, with a particular emphasis on the domains of privacy and security. By synthesizing relevant knowledge from SLR, this study provides a privacy- and security-aware framework for AI Ethics, which is designed to aid a range of stakeholders, including organizations, academic institutions, and governmental bodies, in both developing and critically assessing AI systems. In addition, the study unravels the key issues and challenges concerning privacy and security of AI, and outlines avenues for future research.

Keywords: Artificial Intelligence (AI), AI Ethics, Privacy, Security, Systematic Literature Review, Framework

CERCS: 176

AI Eetika Privaatsus- ja Turva Teadliku Raamistiku Poole

Lühikokkuvõte:

Kuna tehisintellekt (AI) areneb tuues endaga kaasa üha rohkem kasu, suureneb ka üldine rahutus, AI-võimeliste süsteemide privaatsuse ja turvalisuse pärast. Küsimus, kuidas saab tehisintellekti arengut kontrollida nii jurisdiktsioonilisel kui ka organisatsioonilisel tasandil, kerkib tänases diskursuses sagedamini kui kunagi varem. Kuigi valitsustasandil jõudsid Euroopa Parlament ja Nõukogu poliitilisele kokkuleppele ELi tehisintellekti määruse osas, mis on maailma esimene kõikehõlmav AI seadus, on organisatsioonidel endiselt keeruline kohaneda kiiresti areneva tehisintellekti maastikuga ning neil puudub universaalne vahend oma tehisintellekti mudelite privaatsuse ja turvalisuse aspektide hindamiseks. Vastuseks sellele viidi läbi süstemaatiline kirjanduse läbivaatamine (SLR), mis hõlmas aastaid 2020 kuni 2023 ja mille eesmärk oli luua ühtne definitsioon tehisintellekti eetika põhimõistetele, pöörates erilist tähelepanu privaatsuse ja turvalisuse valdkondadele. Käesolevas uuringus sünteesitakse asjakohased teadmised, mis on saadud SLRi käigus, ning esitatakse tehisintellekti eetika privaatsust ja turvalisust arvestav raamistik, mille eesmärk on aidata mitmesuguseid sidusrühmi, sealhulgas organisatsioone, akadeemilisi asutusi ja valitsusasutusi, nii tehisintellekti süsteemide arendamisel kui ka kriitilisel hindamisel. Lisaks sellele tuuakse uuringus välja peamised probleemid ja väljakutsed seoses tehisintellekti privaatsuse ja turvalisusega ning visandatakse tulevaste uuringute võimalused.

Võtmesõnad: Tehisintellekt (AI), AI Eetika, Privaatsus, Turvalisus, Süstemaatiline Kirjanduse Ülevaade, Raamistik

CERCS: 176

Contents

1 Introduction

Artificial intelligence (AI) is becoming increasingly prevalent in today's society, from virtual assistants like Siri and Alexa to autonomous vehicles and facial recognition systems. A growing number of individuals began to embrace AI and incorporate it into their daily routines, propelling AI into the mainstream and prompting major industry players, including Microsoft, Google, and Meta, to release their first publicly accessible generative AI applications [35].

As Artificial Intelligence continues to advance, it has brought numerous concerns about the collection, processing, and storage of personal data [50]. AI technologies typically require extensive datasets to refine their algorithms and enhance performance. This often involves personal data, including sensitive details like medical records and social security numbers. One of the primary concerns surrounding AI is the potential for unauthorized access to the personal data and security breaches. With the vast amount of data being collected and processed, there is an increasing threat of it becoming vulnerable to security attacks. In particular, cybercrimes impact the security of 80% of global businesses, underscoring the profound consequences that mishandled personal data can unleash [5]. Another concern is the application of AI in surveillance, which triggers debates over privacy rights and the potential to abuse these technologies [20].

The problem becomes more evident with an increasing number of adverse incidents within the realm of AI. Notable instances include Facebook's data privacy scandal in 2018 when it was revealed that Cambridge Analytica, a political consulting firm, harvested personal data from millions of Facebook users without their consent [11]. Similarly, Google's mishandling of medical data in 2019 through Project Nightingale, where sensitive patient information was accessed without proper authorization [41], further spotlighted the criticality of privacy and security in the AI landscape. These cases only scratch the surface, hinting at a broader spectrum of concerns potentially yet undisclosed to the public.

Researchers around the world have dedicated extensive efforts to investigating these inquiries, with discussions taking their root in the early 2000s [36]. But despite the ongoing dialogue, no consensus was reached about the key terms and concepts around AI Ethics [44], and there is currently no well-defined framework for evaluating the privacy and security of AI models. In a landmark move, the European Union Parliament, after months of extensive negotiations, reached an agreement on the first legal framework dedicated to AI - EU AI Act [6]. The introduction of this framework has ignited widespread scientific discourse, critically examining its quality and applicability [24].

In the light of these developments and the recognized gap in the field, this study aims to conduct a comprehensive analysis of the recent AI Ethics literature through a Systematic Literature Review (SLR) to uncover key themes and knowledge gaps in privacy and security domains. Subsequently, we intend to propose a privacy- and security-aware framework for designing and assessing AI models.

The following research questions are developed to achieve the given core objectives:

- **RQ1:** What is the current state of the art of AI ethics academic literature focused on privacy and security perspectives?
- **RQ2:** What are the existing frameworks for assessing the adherence of an AI model to the AI ethics principles of privacy and security?
- **RQ3:** What aspects should be considered to ensure the development and deployment of privacy- and security-aware AI model?

The remainder content of the paper is structured as follows: Section 2 presents the background of the study, and the research methodology is reported in Section 3. The SLR results and analysis are discussed in Section 4. The privacy- and security-aware framework for AI Ethics is introduced in Section 5. Finally, Section 6 provides an overview of threats to the validity of the study and existing limitations, and Section 7 concludes the findings with future directions.

2 Theoretical Background

Artificial Intelligence (AI) ethics has evolved into a pivotal field of study, gaining prominence as AI technologies become increasingly intertwined with our daily lives. The intricate nature of defining AI ethics is reflected in the myriad of contradicting and overlapping interpretations, underscoring the complexity of ethical considerations in AI development and deployment.

While some studies concentrate on the morality of AI creators, emphasizing the ethical responsibilities of those who design and develop AI systems [3], others define AI Ethics through the lens of intricate interconnections with other domains, such as information ethics, computer ethics, and robot ethics [21]. These studies illuminate the multifaceted nature of ethical concerns in the digital and automated world, showcasing how these fields overlap and influence each other. Predominantly, however, the studies in this realm gravitate towards defining AI Ethics through a set of key principles [7], as ethical principles provide a structured approach to understanding and navigating the ethical complexities inherent in AI technology and its applications. The study by Jobin et al. is considered to be pivotal in this regard [1, 48], as it is providing, to date, the most comprehensive analysis of ethical AI guidelines. Based on the analysis of 84 ethical AI guidelines, 11 central principles were identified: "transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity" [28].

Privacy and security principles, often conceptually linked together in scientific literature [28], stand out among the major AI ethics principles due to their fundamental role in safeguarding individual rights and societal well-being [29]. Privacy ensures the protection of personal data from unauthorized access and use, while security addresses the integrity and robustness of AI systems, protecting against unauthorized access and mitigating risks of malicious interference. The major public concerns in this regard are centered around security vulnerabilities in AI systems controlling autonomous vehicles and healthcare data breaches [10], showcasing the broader implications of neglecting security measures. Hacks and manipulations in these systems not only pose risks to individual safety but also threaten public well-being. This further emphasizes that addressing privacy and security is not just an ethical imperative but essential for fostering public trust and ensuring the responsible development and deployment of AI technologies. While privacy and security are crucial in the context of AI Ethics, no practical framework to tackle this issues was identified in the existing scientific literature.

Several general AI ethics frameworks have been proposed, demonstrating shared aspirations to establish the common standard that will ensure the responsible and ethical use of AI. Frameworks like the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems and Ethics Guidelines for Trustworthy AI [17] have made commendable efforts to provide guidelines for ethical AI development. However, there exists a discernible gap in the depth to which privacy and security considerations are integrated

into these frameworks. They often offer high-level guidance, leaving room for interpretation and potential oversight in implementing concrete measures to safeguard privacy and enhance security in AI systems.

Regulatory bodies, notably the European Union, have recognized the imperative to address the ethical and societal impact of AI, thus proposing EU AI Act - the first regulation on artificial intelligence [18]. The framework follows a risk-based approach to classify AI systems into 4 potential risk levels: "Unacceptable risk, High risk, Limited risk, Minimal or no risk". The framework aims to define clear requirements and obligations specific to each risk level for AI systems and its creators. While commendable strides have been made in formulating it, existing framework has a number of limitations and gaps [24], further emphasizing the critical need for a holistic approach that integrates privacy and security considerations.

The urgency of this topic is highlighted by the rapid advancement of AI technologies and their pervasive integration into various sectors. The ethical implications and societal consequences of AI demand a proactive approach, necessitating comprehensive frameworks that not only uphold ethical principles but also embed security and privacy considerations. Therefore, this study seeks to propose Privacy- and Security-Aware Framework for AI Ethics, by offering a more granular and explicit integration of privacy and security principles into the broader ethical landscape of AI.

In summary, this enhanced background section emphasizes the vital importance of the AI ethics field and articulates why privacy and security considerations are indispensable. The potential risks associated with AI technologies underscore the urgency of integrating robust security and privacy dimensions into ethical frameworks, ensuring responsible AI development and deployment in an evolving technological landscape.

3 Research Methodology

To garner more profound understanding about privacy and security domains in AI Ethics, this paper adopted a systematic literature review (SLR) approach as defined by Kitchenham [34]. Unlike narrative literature reviews, the SLR approach is an evidence-based approach that involves employing a transparent, rigorous, and replicable literature searching and reviewing process. Our review consists of the following steps: 1) study identification, 2) study selection, 3) study relevance and quality assessment, 4) data extraction and 5) data synthesis.

3.1 Study Identification

The primary goal of this initial phase is to identify as many studies related to the research questions as possible using an unbiased search strategy.

To accomplish this objective, the Scopus database was chosen due to its wide range of peer-reviewed sources across different disciplines. As of now, both Scopus and Web of Science are the most renowned scientific databases available, each offering extensive collections of academic resources. However, Scopus is recognized as the largest indexer of global research output, with more than 7000 publishers in 105 countries [39]. As the most expansive interdisciplinary abstract and citation database, Scopus provides a more comprehensive resource for an in-depth literature review.

The set of search keywords (see Table 1) was defined to attain the objective set earlier, which is to uncover the key themes and knowledge gaps in privacy and security domains of AI Ethics. The search process was executed through multiple iterations, with the final search conducted on January 1, 2024.

Table 1. Search terms used for the literature review.

Database	**Search terms in title/abstract/keywords**
Scopus	("AI ethics" OR "artificial intelligence ethics") AND ("privacy" OR "security")

3.2 Study Selection

In the second step, the selection of studies, we defined the exclusion and inclusion criteria (Table 2).

Since AI is evolving at a fast pace, we limited our selection to peer-reviewed studies published in the last three years (i.e., between January 1, 2020, and December 31, 2023) and written in the English language. The defined exclusion criteria filter the search string findings and remove irrelevant, not accessible, redundant, and low-quality studies.

Table 2. Inclusion and exclusion criteria.

Number	Inclusion criteria
In1	Studies published in 2020-2023.
In2	Studies written in English language.
In3	Peer-reviewed studies.
Number	**Exclusion criteria**
Ex1	Studies that fall under the categories of books, book chapters, and conference reviews.
Ex2	Studies that are not available in full text.
Ex3	Studies that are not focused on AI Ethics, specifically privacy or security.

3.3 Study Relevance and Quality Assessment

Step three of our systematic literature review was to assess the relevance and quality of the selected studies. First, for each of the identified studies, we read the title and abstract to filter the studies that are not focused on AI Ethics, specifically privacy or security. Following that, the relevance and quality of the remained studies were evaluated by reviewing the full text. Studies not available in full text were also excluded at this point.

The relevance assessment was based on a set of specific criteria centered around the role and focus of AI Ethics. Primarily, AI Ethics needed to be a substantial or major component of the study, as reflected in its research questions, objectives, and overall thematic focus. Additionally, privacy or security topics had to be covered in the study. Any study that did not at least partly focus on these aspects or treated them as peripheral issues was excluded in this phase.

Building upon this foundational screening, we further scrutinized each study using the following quality dimensions [51]:

- The objectives of the study are clearly stated, and data collection methods are adequately described. References support important statements in the paper.
- The design of the study is appropriate for the research objectives. The study's research questions are answered or the research objective is attained.
- The study's research approach is described in sufficient detail.

Employing the PRISMA approach allowed us to systematically structure and visually map out the entire process of study screening and selection, as shown in Figure 1. The

final selection led us to only 73 studies directly addressing the privacy and security aspects of AI.

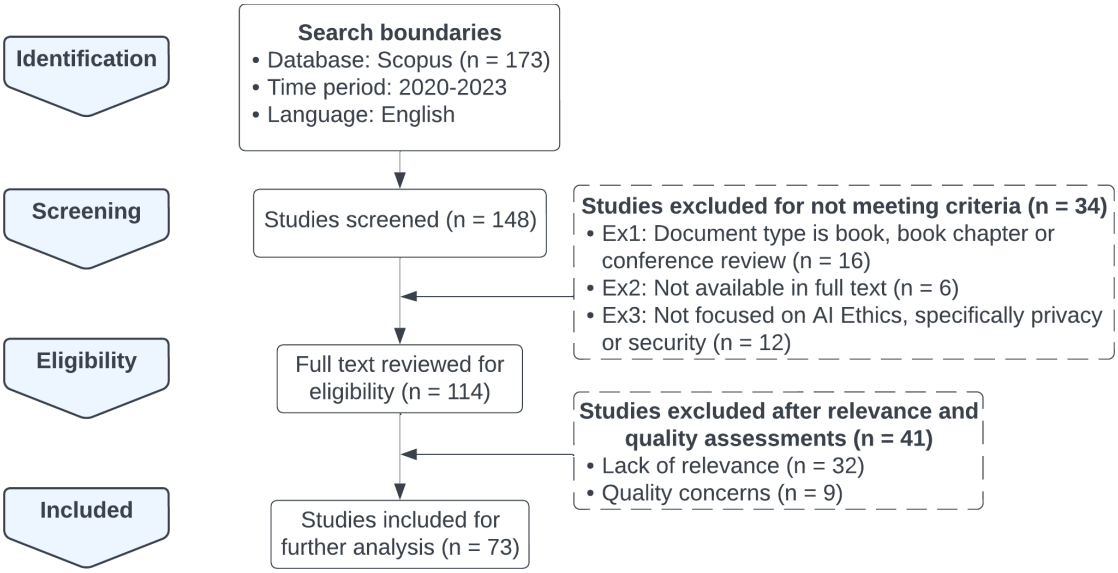

Figure 1. Study selection, assessment, and inclusion (presented using the PRISMA flow diagram).

3.4 Data Extraction

For the data extraction phase of our literature review, a spreadsheet was created to record the metadata for each of the selected studies. Table 3 depicts the metadata we collected about the 73 selected studies, including descriptive information, approach-related information, and AI-related information. To ensure alignment with our objectives, these metadata categories were derived from the research questions.

Table 3. Overview of information collected about each of the selected studies.

Category	Metadata	Description
Descriptive	#	What is the study number, assigned in an Excel worksheet?
information	Title	What is the study title?
	Author(s)	Who is/are the study's author(s)?
	Year	In which year was this study published?

	Country	Which country is/are the author(s) affiliated with?
	Document Type	What is the type of this study (e.g. journal article)?
	Publication	Where was this study published (e.g. name of the journal/conference)?
	DOI / Website	What is the study's DOI? If no DOI is available, through which website can this study be found?
	Keywords	What are the keywords in this study?
	Citation	What is the citation count for this study?
	Database	Which database was used to find this study?
Approach-related information	Objective	What is the research objective / main question?
	Research method	Which method was used to collect data in the study?
	Approach	Does the study use qualitative, quantitative or mixed approach?
	Availability of underlying research data	Does the study contain a reference to the public availability of the underlying research data (or explains why this data is not openly shared)?
	Relevance	Does the study's research focus align with the objectives of this review?
	Quality concerns	Are there any quality concerns (e.g. limited information about the research methods used)?
AI-related information	Primary domain	What is the primary domain/field in this study (e.g. healthcare)?
	Secondary domain	What is the secondary domain/field in this study (e.g. education)?
	Technology	Which technology is this article primarily focused on (e.g. machine learning)?
	AI ethics	How AI ethics is defined in the article?
	Ethical principles	What ethical principles are mentioned in the article (e.g. transparency)?
	Emphasis on privacy or security	Does the study place a stronger emphasis on privacy or security, or both are considered equally?
	Framework	Does the study incorporate any specific framework, and if so, which one?
	Framework application phase	What is the application phase of the framework (if applicable)?
	Key findings	What are the key findings in this study?

3.5 Data Synthesis

The data synthesis phase, marking the final step of our systematic literature review, required a thorough analysis of the metadata extracted earlier. In Section 4, we leveraged Python programming language (mainly Matplotlib library) to efficiently summarize and visualize our findings, highlighting key data patterns and trends. Building on these insights, a privacy- and security-aware framework for AI Ethics is then developed (presented in Section 5). This framework is a direct product of the conducted systematic literature review, offering a nuanced approach to understanding and addressing privacy and security concerns in AI Ethics.

4 Results from Systematic Literature Review

This section describes the results drawn from the analysis of 73 selected research articles that concern privacy and security domains of AI Ethics. The data underlying this study is publicly available in Zenodo, DOI: https://doi.org/10.5281/zenodo.10451282; thereby supporting the open science movement and ensuring replicability of the study.

4.1 Descriptive Analysis

Descriptive Analysis provides the general overview on the examined studies and includes the examination of chronological publication trends, analysis of the publication types, overview of the geographical landscape, citation and keyword analysis.

4.1.1 Chronological Publication Trends

The period between 2020 and 2023 marks a significant upward trajectory in AI Ethics research, with the number of publications in Scopus rising from 4 in 2020 to a peak of 32 by 2023, as depicted in Figure 2. This trend signifies the expanding influence of the field and the heightened attention from scholars on the ethical considerations surrounding AI technology.

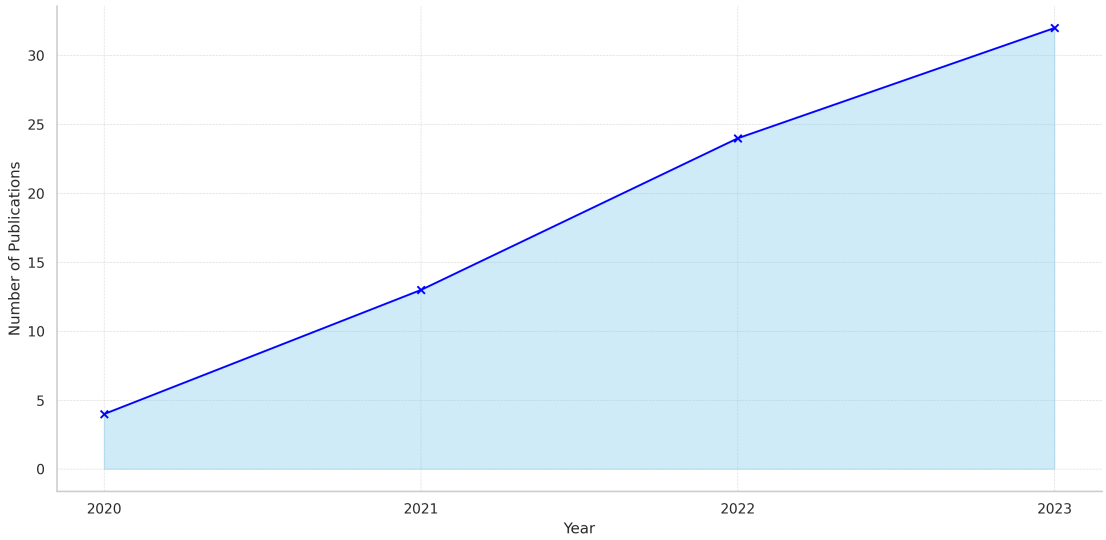

Figure 2. Publication trends over time.

4.1.2 Publication Type Distribution

The dataset underscores a focused trajectory in scholarly output, predominantly featuring journal articles and conference papers, as illustrated in Figure 3. Journal articles, accounting for approximately 67.12% of the studies, are the prevalent publication type, showing the field's strong emphasis on rigorous peer-reviewed research. Conference papers account for roughly 32.88%, reflecting the field's vibrant exchange of ideas and findings in dynamic, collaborative forums.

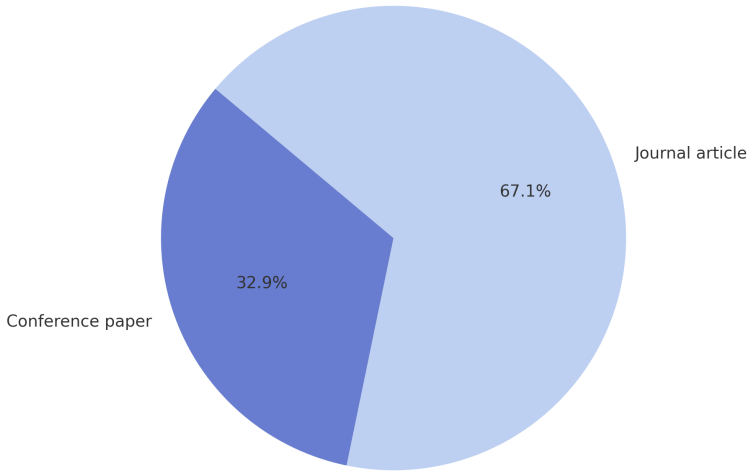

Figure 3. Publication type distribution.

It should be noted that while our initial dataset included a broader spectrum of publication types, such as books, book chapters, and conference reviews, these were methodically filtered out during the screening process. The exclusion of these publication types does not detract from their value but rather serves to sharpen the focus of our current inquiry.

4.1.3 Geographical Landscape

The data reveals a clear picture of the global research landscape in AI Ethics field, highlighting specific countries (Figure 4) as key contributors in the scientific arena.

The United States secures a leading position in our dataset, contributing 15.09% of the total research publications. This substantial share underscores the strong academic and research infrastructure in the United States. Australia and Germany, each with a notable contribution of 10.38%, reflect their considerable involvement in AI Ethics research. This is especially noteworthy for Australia, which maintains a high output despite its smaller population size compared to the United States and Germany. Canada

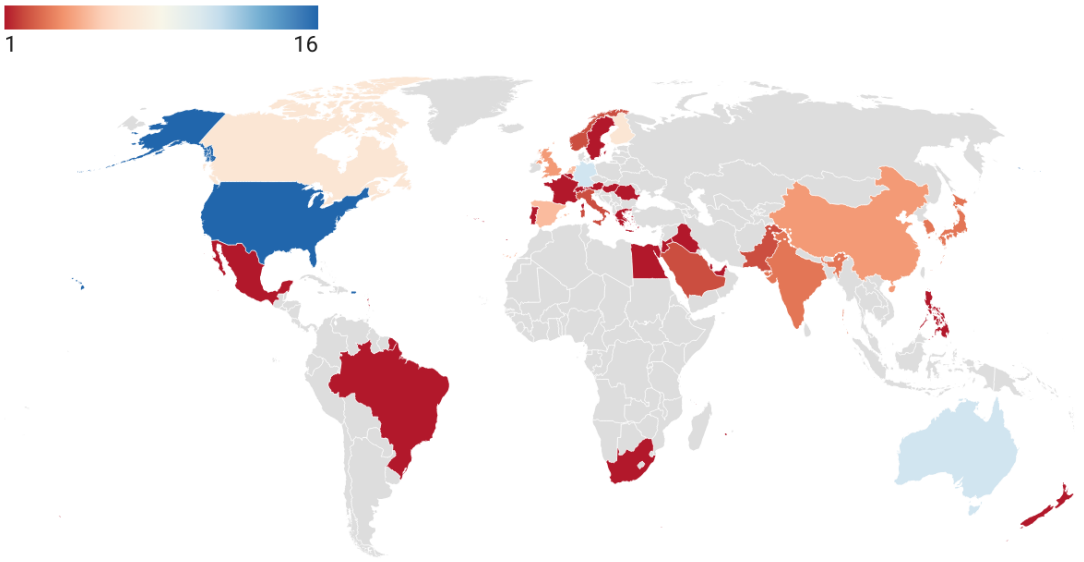

Figure 4. Geographical distribution of publications.

and Finland, contributing 6.6% each, also highlight the importance and focus on AI Ethics research within these countries, indicating a strong commitment to advancing in this domain. Other countries like the Netherlands, Spain, China, and the United Kingdom, each offering around 4-5% of the total publications, further demonstrate the global interest in AI Ethics.

From a regional perspective, Europe stands out as the dominant force, accounting for 41.51% of the total research output. This significant percentage suggests a dense concentration of research activities spread across several European countries, marking the continent as a central hub in the field of AI Ethics. North America, led by the United States and supported by Canada, also plays a crucial role, with a combined contribution of 22.64%, reflecting the region's strong influence in shaping global research trends. Asia, represented by countries like China, India, Japan, and the Republic of Korea, contributes 19.81%, showcasing the region's growing momentum in AI Ethics research, driven by technological advancements and a focus on innovation. Oceania, predominantly represented by Australia, makes a significant contribution as well, with 11.32% of the total publications, indicating its active role in the global AI Ethics research landscape. While less frequent, contributions from Africa and South America are essential for adding diversity to the global research perspectives.

In summary, these findings highlight a widespread engagement in AI Ethics research, with notable concentrations in Europe and North America, and growing participation from regions like Asia and Oceania.

4.1.4 Citation Analysis

Citation analysis in Figure 5 presents a skewed distribution with a majority of articles receiving a lower number of citations, while a few have significantly higher citations. This pattern is typical in academic literature, where certain pioneering or breakthrough studies gain substantial recognition. This can be attributed to many studies being relatively recent and not having sufficient time to accumulate citations. Additionally, the zero-citation count for 30.13% of articles may reflect the time lag in the dissemination and integration of new research into the broader academic discourse.

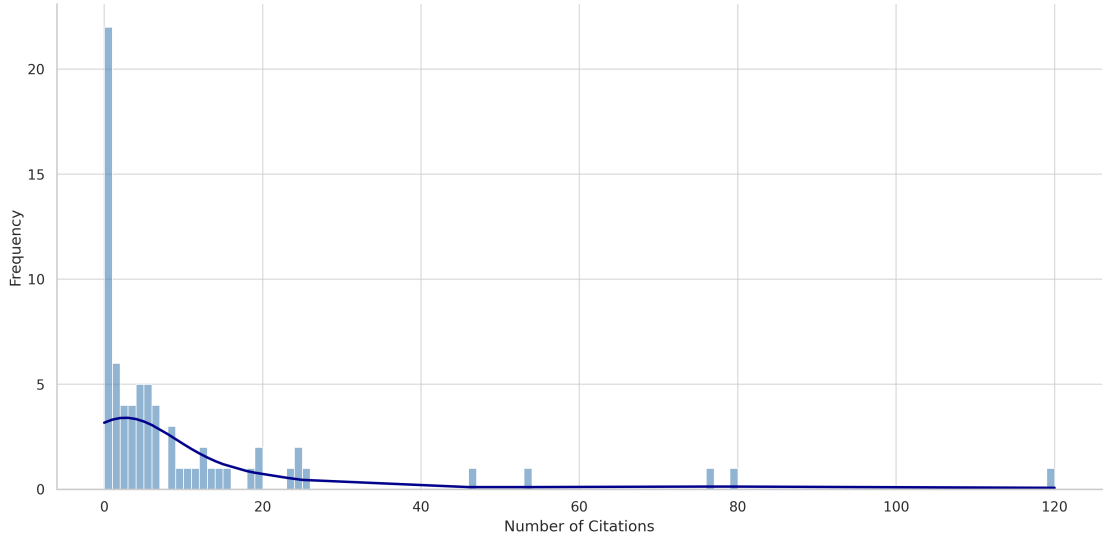

Figure 5. Citation analysis.

4.1.5 Keyword Analysis

A word cloud visualization of keywords is employed to present the predominant themes emerging from the dataset, as depicted in Figure 6.

A word cloud offers an intuitive and visually engaging means to understand the frequency of keywords within a text corpus. In this graphical representation, the size of each word or phrase is directly proportional to its frequency of occurrence. Therefore, more prominent words in the word cloud, excluding those utilized in our search strategy, such as "Machine learning" and "Machine ethics", "Trustworthy AI", "AI canvas" and "Responsible AI", "Data governance" and "AI governance", "AI Act" and "GDPR", and various AI Ethics principles, indicate their higher prevalence in the dataset. The dominance of these particular words underscores their significance in the context of our research.

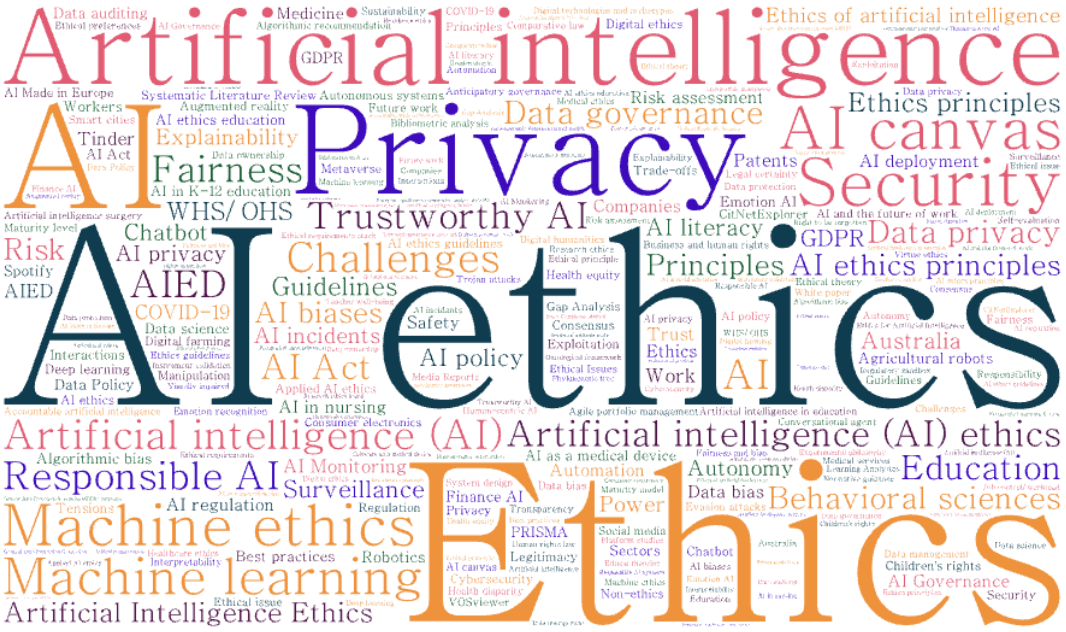

Figure 6. Word cloud based on keywords.

4.2 Approach Analysis

The selected studies utilize a variety of research methods, with literature review being by far the most dominant research approach (n = 40) (Figure 7). The array of other methods employed in the examined studies includes document analysis, surveys, interviews, case studies, expert panels, comparative analysis, workshops, action research, and review of existing initiatives.

83.56% of the studies in our sample utilize qualitative method, and the remainder of the studies employ quantitative or mixed methods (combining qualitative and quantitative approaches) (Figure 8).

Only 12 studies have openly made the underlying research data available, despite the growing trend in openly sharing the underlying research data as a positive open science practice, which increases transparency and trust and allows scrutiny of the findings [51]. 3 studies mention that research data can be made available upon request from the author. The unavailability of underlying research data could be attributed to the authors' concerns regarding the data privacy.

4.3 Quality Analysis

In addition to general inclusion and exclusion criteria, it is generally considered important to assess the quality of the studies. For 43 out of the 73 studies, the research

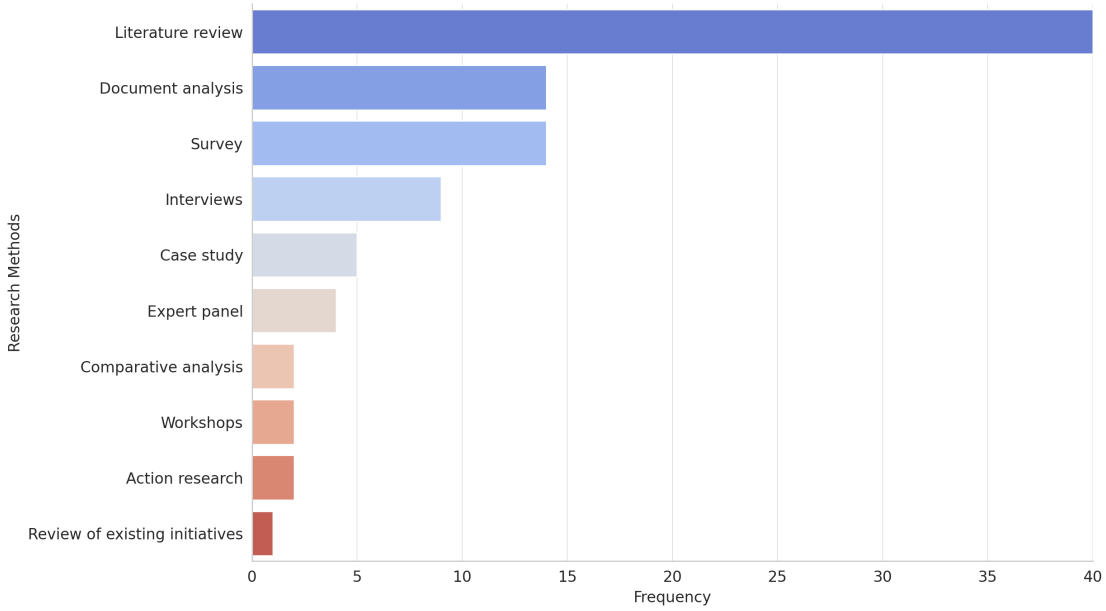

Figure 7. Research methods utilized in publications.

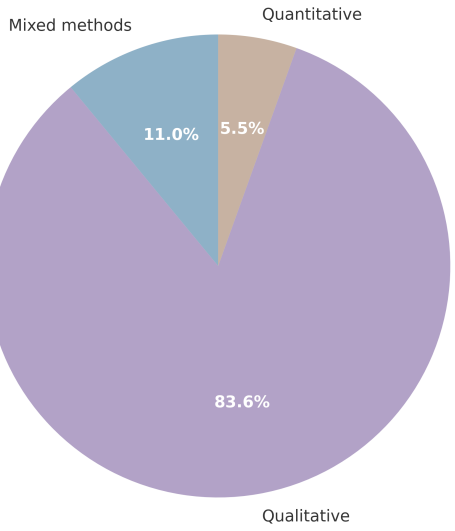

Figure 8. Research approaches utilized in publications.

design was appropriate. For 30 studies, we had minor quality concerns – for instance, missing information about the literature review methodology, like the total number of search results for each database explored or the quality assessment mechanisms of the reviewed studies. Studies for which we had significant quality concerns (n = 9) had already been removed during the full study assessment.

4.4 Content Analysis

4.4.1 AI Ethics Definition and Principles

Out of the 73 studies that were reviewed, only 7 included a formal definition of AI Ethics, as indicated in Table 4. While several of these studies acknowledged the absence of a scientific consensus on the definition of AI Ethics [44], the predominant approach was to define AI Ethics in terms of its principles. For instance, Ashok et al. [7] described AI Ethics by highlighting its core principles of responsibility, transparency, auditability, incorruptibility, predictability, and the morality of machines.

Table 4. AI Ethics Definitions.

Definition	Reference
"deals with the issues raised in developing, deploying, and using AI systems and involves the moral behavior of humans in the design and development of AI"	[2]
"a human-centered perspective which focuses on the morality of humans who deal with the AI systems, including developers, manufacturers, operators, consumers, etc"	[3]
"the moral behavior of humans as well as the moral behavior of AI agents in the process of designing, constructing, using, and handling AI beings"	[9]
“a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies”	[14]
"an emerging and interdisciplinary field concerned with addressing ethical issues of AI. AI ethics involves the ethics of AI, which studies the ethical theories, guidelines, policies, principles, rules, and regulations related to AI, and the ethical AI, that is, the AI that can uphold ethical norms and behaves ethically."	[25]
"a subfield of applied ethics focusing on the ethical issues raised in the development, deployment, and use of AI"	[26]
"specifies the moral obligations and duties of an AI and its creators"	[47]

Subsequently, we identified the key principles of AI Ethics prevalent in the studies (Figure 9). The discourse is heavily dominated by principles of transparency, privacy, and fairness, whereas the principle of sustainability, among others, garners considerably

less scholarly attention. The principles of beneficence and non-maleficence, while being significantly covered in the literature, might be less emphasized due to their broader definitions and the challenges associated with the practical implementation.

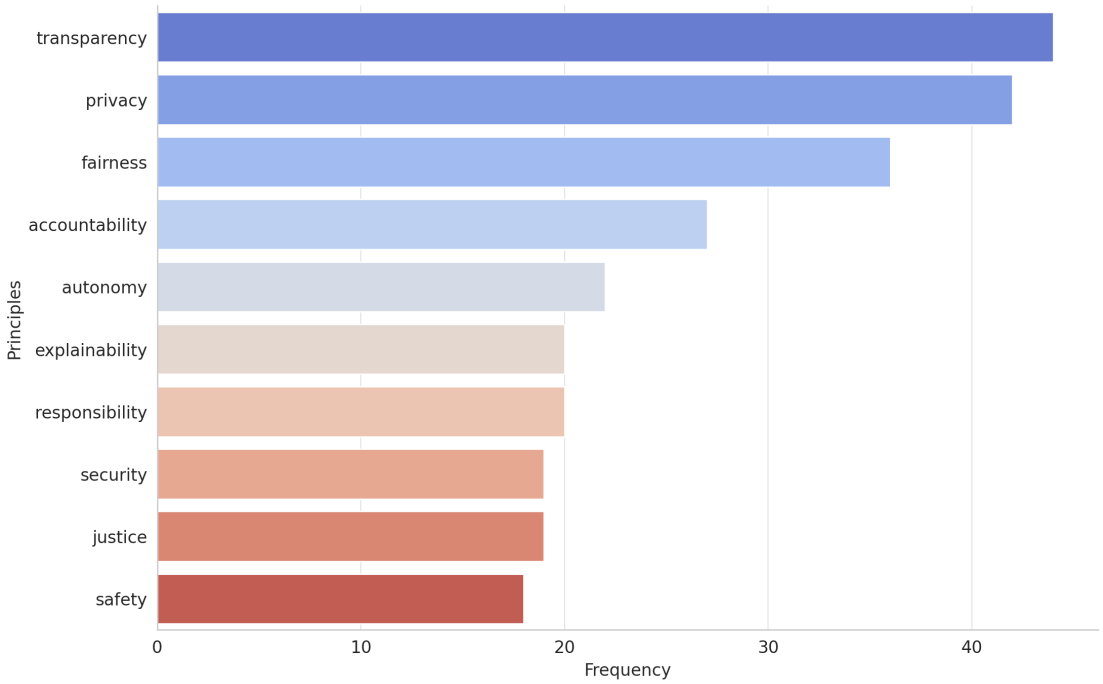

Figure 9. Ethical principles identified in publications.

By mapping together the principles that are interrelated and often conceptually linked in the literature, we identified 5 most frequently cited ethical principles, which are transparency and explainability, privacy and security, fairness and justice, responsibility and accountability, and freedom and autonomy. A brief description for each of these principles is provided in Table 5. The arrangement of these principles in the table follows a top-down order, reflecting their popularity in the current scientific discourse.

Table 5. AI Ethics Principles.

Ethical principle	**Description**
Transparency and explainability	AI systems should be designed to provide clear insight into their processes and decisions, ensuring that users and regulators can understand AI actions and outputs.
Privacy and security	AI systems should safeguard personal data, respecting privacy rights and maintaining robust defenses against unauthorized access and breaches.

Fairness and justice	AI systems should be inclusive and operate without bias, offering equal opportunity and treatment while ensuring just outcomes for all individuals and groups.
Responsibility and accountability	There should be a clear assignment of responsibility for AI behavior, with mechanisms in place to hold the creators accountable for the system's impact.
Freedom and autonomy	AI systems should enhance human decision-making without constraining individual freedoms, allowing for personal autonomy and self-determination.

After reviewing the existing literature, we define AI Ethics as *an interdisciplinary field focused on the moral conduct of both humans and AI agents involved in the development, deployment, and use of AI technologies*. Central to AI Ethics are the principles of transparency and explainability, privacy and security, fairness and justice, as well as freedom and autonomy.

4.4.2 Classification of Research Studies

In our investigation into the interdisciplinary nature of AI Ethics, we conducted a domain analysis for the selected studies to map out the various fields concerned with AI Ethics.

Each study was meticulously reviewed to identify its primary domain — the main field of research or application. The primary domain is the central focus of a study, where AI ethics is most prominently addressed or scrutinized. For instance, if a study primarily addresses the data privacy of patients, *Healthcare* would be designated as its primary domain.

Furthermore, we also sought to identify any secondary domains present within each study. A secondary domain is an auxiliary field that provides additional context, background, or implications for the primary discussion of AI ethics. It supports the primary domain but is not the main focus of the ethical inquiry. For example, a study primarily focused on *Healthcare* might also touch upon the current governance initiatives aimed at addressing privacy issues, thus assigning *Law and Governance* as a secondary domain.

The domain analysis was visualized in a network graph (Figure 10), where nodes represent domains and edges signify the conceptual links between these domains and AI Ethics. The size of the node corresponds to the number of studies within the domain.

General domain, represented by the highest number of studies in our dataset, covers a wide range of themes, with the central one being to provide a comprehensive view of the AI Ethics landscape [14]. Significant topics include the exploration of ethical principles [30] and their trade-offs [42], such as transparency versus privacy, and the

Figure 10. Classification of publications by domains.

formulation of guidelines to navigate these complex ethical terrains [16]. Additionally, this domain encompasses discussions on the broader risks and challenges posed by AI technologies [43].

Among the specific fields analyzed, *Business* domain emerged as the most extensively studied. The primary focus in this domain was on bridging the gap between theoretical concepts and practical applications [26]. Key discussions revolve around how ethical principles can be operationalized in real-world scenarios. This includes exploring the implications of AI in the workplace, particularly in relation to employee safety [12, 13]. A significant topic in this domain is the use of AI in hiring processes [49], examining

both the potential benefits and the ethical concerns, such as bias and fairness. Studies in this domain essentially shed light on the challenges and opportunities of integrating AI ethically in a corporate environment, which seems to be the main concern in the current scientific literature.

In the realm of *Education & Research*, our analysis indicates a growing emphasis on integrating AI Ethics into academic curricula. The studies in this domain delve into the rationale and methodologies for incorporating AI Ethics into high school [33] and university programs [23]. Another salient theme within this domain is the use of AI for surveillance and online proctoring in educational settings [15, 20]. These studies critically assess the advantages and drawbacks of AI-enabled applications, exploring both the effectiveness of these technologies in maintaining academic integrity and their impact on student privacy and perceptions. *Education & Research* domain, therefore, encompasses not only the content of AI Ethics education but also the ethical dimensions of AI applications in the educational process and research practice itself.

Healthcare domain focuses on the ethical dimensions surrounding the development and use of AI in a medical context. A primary theme here involves evaluating the risks and challenges associated with AI-enabled medical devices [22] and products [40]. Data protection and patient privacy emerge as prevalent topics [10], reflecting the heightened sensitivity of personal health information in the age of AI. Additionally, the use and implications of conversational AI, such as chatbots, in healthcare are explored [4]. Collectively, the *Healthcare* domain addresses the multifaceted ethical issues at the intersection of AI technology and patient care, with an emphasis on personal data protection.

Law & Governance domain, often encountered as a secondary topic, is mostly centered around data privacy under the GDPR framework [8, 37]. A substantial portion of the discussion also focuses on the existing regulatory initiatives of different countries and the necessity for robust governance to address the ethical challenges posed by AI[16]. The studies in this domain critically examine the gaps in the regulatory frameworks and advocate for more comprehensive policies. *Law & Governance* domain underscores the importance of legal structures in safeguarding ethical standards in AI.

In *Society* domain the focal point is predominantly on public opinion towards AI. Studies in this domain explore societal perspectives on AI and how the public perceives the importance of various ethical principles [32]. The main focus is on the general attitudes towards AI [27], including concerns and expectations, and public reactions to the potential misuse of AI [45]. By capturing the societal sentiment towards AI, the studies in this domain shed light on the ethical dimensions of AI as viewed through the lens of the broader public. This perspective is vital for understanding the societal impact of AI and for guiding responsible AI development that aligns with public values and concerns.

Computer Science domain, the last significantly covered field in the reviewed liter-

ature, delves into the ethical considerations specific to various AI technologies. Key themes include the ethics of particular technologies like emotion recognition systems [38], where the moral implications of technology use and its impact on privacy and autonomy are examined. Additionally, the domain encompasses the study of ethical design principles, especially in the development of AI-driven tools such as chatbots [9], focusing on how these technologies can be designed to uphold ethical standards. Cybersecurity also emerges as a critical topic, particularly in the context of adversarial attacks [31], where the integrity and resilience of AI systems against malicious interventions are scrutinized. This domain, therefore, intersects the technical aspects of AI with ethical deliberation, emphasizing the need for ethical mindfulness in the technological design and deployment of AI.

Domains like *Agriculture*, *Biology*, *Finance*, and *Urban Environment* are underrepresented in our dataset, possibly due to their specific niche applications or the nascent stage of AI Ethics in these areas. These fields might be overshadowed by sectors where AI's impact is more immediate and visible to society, leading to a disproportionate focus in current research. In addition, *Philosophy & Religion* domain, while being fundamental in framing broader ethical questions, seems to be less significant than more practical concerns in other domains.

4.4.3 AI Ethics Frameworks

In our comprehensive analysis of 73 studies, a mere fraction - just over 13 percent - focused primarily on the development of frameworks for AI Ethics.

Frameworks discovered within *General* domain exhibit a predominantly theoretical focus. For instance, Ashok et al. developed an ontological framework aimed at conceptualizing the ethical impacts of AI across physical, cognitive, information, and governance domains [7]. This framework provides a high-level theoretical overview rather than actionable steps for ethical AI implementation. Similarly, Sharma et al. proposed a framework for evaluating ethics in AI centered around 7 abstract checkpoints, which, while insightful, does not translate easily into practical measures [46].

Within *Healthcare* domain, a single framework was identified, which explores the risks associated with AI-enabled chatbots in the context of medical ethics [19]. The study's framework, although adeptly maps ethical principles to potential risks, stops short of offering a structured risk management approach.

A more applied approach is evident in *Business* domain, with several frameworks designed to integrate ethical considerations into the corporate landscape. These include methodologies for assessing workplace health and safety in the context of AI [12, 13]. Agbese et al. propose a novel way of implementing ethical requirements in organizations by utilizing the Agile portfolio management framework [2]. It is worth noting, however, that this framework primarily concentrates on the strategic and planning aspects of companies, emphasizing the integration of ethical requirements rather than specific

ethical aspects that should be considered. Meanwhile, another notable study in this domain introduces an AI maturity model for small and medium sized companies (also known as SME) [44]. While this model proves to be practical and has undergone validation by real organizations, its main function lies in providing an initial assessment of companies, offering limited insights and strategies for improvement.

ECCOLA method by Vakkuri et al. represents a beacon of practical application within our dataset, employing a deck of cards to facilitate ethical considerations in AI systems [48]. It is "a sprint-by-sprint process designed to facilitate ethical thinking in AI and autonomous systems development" that demonstrates a tangible approach to embedding ethical thinking within organizational processes. ECCOLA comprises a set of 21 cards, split into 8 distinct themes. These themes include "analyze, transparency, safety & security, fairness, data, agency & oversight, wellbeing, and accountability", with each theme encompassing 1 to 6 cards. The employed methodology is covering all key ethical principles and providing organizations with a practical tool for implementing ethically aligned AI systems.

When reviewing AI Ethics frameworks across diverse fields, from General domain to specific sectors like Healthcare and Business, a pattern becomes evident from looking beyond the unique context of each study. Despite the varied applications and focus areas, there is a consistent emergence of four dimensions, specifically Data, Technology, Process, and People. These dimensions reoccur in frameworks regardless of their specific field of application which might signify a universal approach within AI Ethics.

Our review, aimed at uncovering frameworks with a strong focus on AI privacy and security, identified only one study (ECCOLA methodology) that partially addressed our inquiry. The scarcity of frameworks that effectively meld ethical theory with actionable practice highlights a noticeable gap in the existing scientific literature. Despite the heightened public and scholarly concern regarding privacy and security in AI systems, among the studies scrutinized, none provided a framework dedicated exclusively to the practical evaluation of privacy or security aspects in AI models.

5 Towards Privacy- and Security-Aware Framework for AI Ethics

Crafting a privacy- and security-aware framework for AI Ethics requires a multi-faceted approach. Through the systematic literature review, we have managed to identify four critical dimensions - Data, Technology, People, and Process - that serve as the pillars of our framework, guiding the ethical assessment of AI systems in organizational environments.

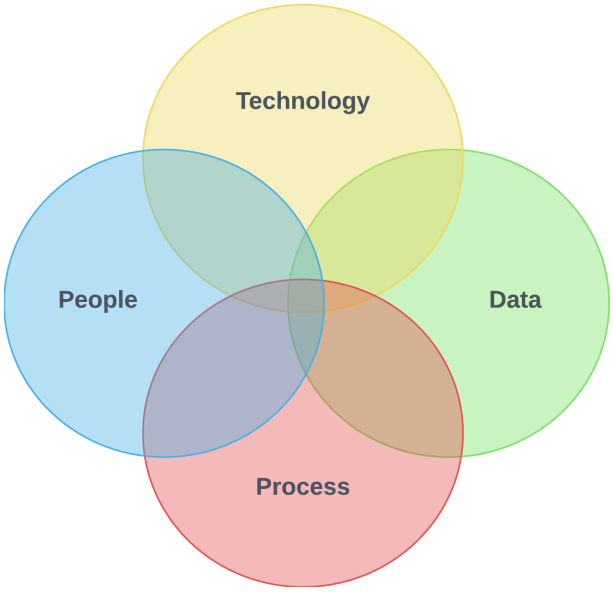

Figure 11. Four pillars of AI systems.

Data dimension is foundational, recognizing that data is the substrate from which AI derives its functionality and intelligence. Ethical considerations in data governance, such as consent, anonymization, and data rights, are paramount, particularly as they relate to privacy and the potential for misuse.

> ***Recommendation 1:*** *Organizations should mandate rigorous scrutiny of data handling practices to mitigate risks and protect individual rights.*

Technology dimension addresses the tangible components of AI systems. The security mechanisms and technical safeguards are not merely adjuncts but are intrinsic to the ethical fabric of the technology. This necessitates an ongoing commitment to adapt and refine these measures, in line with the latest technological advancements.

> ***Recommendation 2:*** *Organizations should adopt a 'security-by-design' approach and continuously align with security standards, ensuring that security mechanisms are foundational elements in the development of AI.*

The human element, encapsulated in **People dimension**, brings to the fore the societal and individual factors impacting AI systems. This dimension emphasizes the importance of secure authentication and permission management. Additionally, it advocates for active engagement with a diverse array of stakeholders, ensuring that AI practices are informed by a broad spectrum of ethical perspectives and experiences.

> ***Recommendation 3:*** *Organizations should establish robust protocols for authentication and permission management.*

> ***Recommendation 4:*** *Organizarions should foster awarness of AI privacy and security concerns and nurture the practical skills necessary to address these challenges effectively.*

Process dimension is crucial in materializing ethical principles into tangible actions. It involves instituting robust processes for ethical oversight, from compliance checks to ethical auditing and incident response strategies, ensuring that AI systems adhere to established standards throughout their lifecycle. This dimension captures the operationalization of ethical principles through established procedures and includes scrutinizing the extent to which these processes are ingrained in the organization's culture and AI system management.

> ***Recommendation 5:*** *Organizations should establish and rigorously enforce comprehensive governance processes, encompassing ethical oversight, compliance checks, and incident response strategies.*

For our privacy- and security-aware framework we synthesized 4 dimensions derived from the SLR – Data, Technology, People, and Process – into an integrated structure, as illustrated in Table 6, enabling a comprehensive and grounded approach to AI ethics.

Each dimension is guided by a set of specific questions to encompass the overarching themes of privacy and security within AI systems. Data, People, and Process dimensions are each guided by three questions, while Technology dimension is informed by two questions. These questions are deliberately wide-ranging, designed to provoke thorough contemplation and exploration of ethical practices. They were formulated based on the SLR-established domain knowledge, ensuring that they are deeply rooted in the current understanding and challenges of AI ethics.

To enhance the framework's utility, we elaborated a suite of best practices corresponding to each question, offering organizations targeted guidance for ethical assessment. Data and People dimensions incorporate 10 best practices each, while Technology and Process dimensions are outlined with 8 best practices. This methodical compilation of best practices serves as a concrete reference point for organizations seeking to evaluate and improve their AI systems.

Table 6. Privacy- and Security-Aware Framework for AI Ethics.

Dimension	**Question**	**Best Practice**	**Risk Level**
Data	How do you classify your data?	Identify the data	High
		Define data protection controls	High
		Define data lifecycle management	Low
	How do you manage data privacy?	Implement consent management	High
		Utilize anonymization techniques	High
		Define data rights and ownership policies	Medium
	How do you protect your data?	Implement secure key management	High
		Implement encryption	High
		Implement access control	Medium
		Utilize tools to prevent direct access to data	Low
Technology	How do you protect your resources?	Perform vulnerability management	High
		Control network traffic at all layers	High
		Reduce attack surface	High
		Validate software integrity	Medium
	How do you secure your applications?	Perform regular penetration testing	High
		Assess security of pipelines	High
		Deploy software programmatically	High
		Centralize services for packages and dependencies	Medium

People	How do you manage authentication?	Utilize strong sign-in mechanisms	High
		Utilize temporary credentials	High
		Audit and rotate credentials periodically	High
		Store and use secrets securely	High
	How do you manage access control?	Determine access requirements	High
		Follow the principle of least privilege	High
		Manage access based on lifecycle	Low
		Share resources securely	Medium
	How do you ensure stakeholder engagement?	Facilitate regular interactive privacy and security trainings	High
		Implement stakeholder feedback mechanisms	Medium
Process	How do you ensure regulatory compliance?	Implement regular compliance audits	High
		Perform regulatory impact assessments	High
	How do you detect potential incidents?	Configure logging	High
		Analyze metrics and logs centrally	High
		Automate monitoring of critical events	Medium
	How do you manage incidents?	Identify key personnel and resources	High
		Create incident management plans	High
		Facilitate learning from incidents	Medium

Embracing the complexity of ethical considerations, we adopted a risk-based approach utilized in EU AI Act. This is a nuanced approach proposed by the European Commission to regulate AI systems by categorizing them into four levels: *Unacceptable Risk*, *High Risk*, *Limited Risk*, and *Minimal or No Risk*. AI systems associated with unacceptable risk pose direct threats to public safety and fundamental rights, meriting their outright prohibition. This is followed by high-risk AI systems, employed in areas such as critical infrastructure and healthcare, where the potential for significant impacts on safety and rights necessitates rigorous regulatory adherence. Limited-risk AI systems, like chatbots, require explicit transparency measures, ensuring user awareness of AI interaction. The spectrum concludes with minimal or no risk AI systems, like most AI-enabled video games or spam filters, which are subject to minimal regulatory oversight [18].

In our framework, we've mirrored this risk-based approach, focusing on High, Limited, and Minimal Risk categories, aligning our best practices in a manner that reflects the level of potential impact, defined as *High*, *Medium*, or *Low*. This provides us with the possibility to critically evaluate and manage ethical risks of AI. Our framework strategically excluded unacceptable risks, as systems falling under this classification are not only prohibited by the AI Act but also represent a class of AI development that fundamentally contradicts ethical principles and standards.

The proposed framework indicates the prevalence of practices potentially leading to high-risk impact across all dimensions. This includes practices such as implementing consent management in Data dimension and performing vulnerability management within Technology dimension. People dimension contains the highest count of practices with high-risk impact (n = 7), while the other three dimensions have equal numbers (n = 6). This underscores the importance of human-centric considerations in AI. Practices with medium-risk impact are equally distributed across dimensions (n = 2), and low-risk impact is present only in Data and People dimensions.

This comprehensive analysis of the dimensions, practices, and associated risk levels highlights the framework's capacity to navigate the complex terrain of AI ethics. It showcases the emphasis on a balanced approach, where both the micro-level intricacies and macro-level perspectives are equally considered, ensuring that the deployment of AI systems is both pragmatic and grounded in a robust understanding of the evolving landscape.

6 Discussion and Limitations

This study has introduced a Privacy- and Security-Aware Framework for AI Ethics, developed as a result of the SLR. The framework, centered around the critical dimensions of Data, Technology, People, and Process, is designed to guide organizations in evaluating and enhancing their AI systems' privacy and security postures. By addressing key questions within each dimension, we propose a series of best practices accompanied by risk assessments to inform the decision-making.

The integration of findings from our SLR has been instrumental in shaping this framework. The SLR underscored the escalating concerns regarding AI's privacy and security aspects and revealed a gap in practical, actionable frameworks addressing these issues. Aligning our framework with these findings equips it with the acumen to effectively address these concerns, offering organizations a clear path to enhance AI ethics compliance and practice.

While the framework aims to be comprehensive, certain limitations must be acknowledged. The SLR that provided the underlying data for this framework has inherent limitations. The exclusive selection of studies from 2020 to 2023 and reliance on a single database may narrow the framework's scope, potentially omitting seminal works and diverse perspectives. The targeted search strategy, focusing on the intersection of privacy, security, and AI Ethics, ensures the depth of the research but could also miss broader ethical concerns.

To date, the framework has not been empirically tested within the organizational environment. As a result, a key aspect of future work involves validation of the proposed theoretical and therefore conceptual framework, which will provide empirical data to refine it further and adapt to diverse organizational contexts. It is also imperative to consider the scalability of the framework, ensuring that it can be effectively utilized by organizations of varying sizes and capacities.

Moreover, the framework assumes a certain level of organizational maturity in privacy and security practices, which may not be present in all organizations. There can also be industry-specific considerations that the framework does not fully encapsulate, necessitating customization in certain sectors. The framework might also benefit from incorporating stakeholders into its structure, as it would further operationalize it.

The rapid pace of technological innovation and the continuous evolution of legal standards mean that the framework will need ongoing updates to remain relevant and maintain compliance.

Acknowledging these constraints is not to undermine the framework's value but to pave the way for its continuous improvement. As AI Ethics is an ever-evolving field, the framework must adapt and expand, incorporating a wider array of sources and perspectives to remain both current and comprehensive.

7 Conclusion

The development and introduction of the Privacy- and Security-Aware Framework for AI Ethics signify a critical advance in the field of AI Ethics. This framework, with its emphasis on the dimensions of Data, Technology, People, and Process, offers a systematic method for organizations to assess and enhance their AI systems, ensuring that privacy and security are not ancillary considerations but central to the ethical deployment of AI technologies.

Our work commenced with a systematic literature review that illuminated pressing concerns in the public and academic discourses regarding privacy and security in AI. This review laid the foundation for a framework that bridges the gap between theoretical ethical concepts and their practical application. By incorporating elements like risk assessments and best practices tailored to each dimension, we have crafted a tool that addresses the urgent need for structured guidance in navigating the ethical complexities of AI.

The practical implications of this framework are far-reaching. It can serve as a blueprint for organizations to evaluate and fortify the ethical integrity of their AI systems, focusing particularly on privacy and security aspects. We encourage organizations to utilize this framework as a guide for ethical decision-making, thereby fostering an environment where AI is developed and utilized with a strong commitment to ethical principles.

The subsequent phase of this research will involve implementing the framework in a variety of organizational settings. This practical application is expected to yield valuable insights, allowing for the refinement of the framework and its adaptation to a wide range of industries and organizational sizes. It is anticipated that this will also uncover new challenges and ethical considerations,

Looking ahead, the framework is positioned for ongoing evolution. We recognize the necessity of adapting to emerging technologies and evolving ethical standards. As the AI landscape continually transforms, so must our approaches to ethical governance. The Privacy- and Security-Aware Framework for AI Ethics marks the beginning of a sustained endeavor in ethical discourse, refinement, and practical implementation. The ultimate goal of our research is to create a paradigm where ethical considerations are seamlessly integrated into the fabric of AI development and deployment that will be made publicly available for a wider audience via its publishing and presenting at the 25th Annual International Conference on Digital Government Research, to which it is currently submitted.

References

[1] C. Adams et al. “Ethical principles for artificial intelligence in K-12 education”. In: *Computers and Education: Artificial Intelligence* 4 (2023). ISSN: 2666920X (ISSN). DOI: https://doi.org/10.1016/j.caeai.2023.100131.

[2] M. Agbese et al. “Implementing AI Ethics: Making Sense of the Ethical Requirements”. In: *ACM Int. Conf. Proc. Ser.* Association for Computing Machinery, 2023, pp. 62–71. ISBN: 979-840070044-6. DOI: https://doi.org/10.1145/3593434.3593453.

[3] K. Ahmad et al. “Developing future human-centered smart cities: Critical analysis of smart city security, Data management, and Ethical challenges”. In: *Computer Science Review* 43 (2022). ISSN: 15740137. DOI: https://doi.org/10.1016/j.cosrev.2021.100452.

[4] R. Ahmad et al. “The benefits and caveats of personality-adaptive conversational agents in mental health care”. In: *Annu. Americas Conf. Inf. Syst., AMCIS*. Association for Information Systems, 2021. ISBN: 978-173363258-4. URL: https://www.researchgate.net/publication/351659439_The_Benefits_and_Caveats_of_Personality-Adaptive_Conversational_Agents_in_Mental_Health_Care.

[5] “AI and Privacy: The privacy concerns surrounding AI, its potential impact on personal data”. In: *The Economic Times* (Apr. 2023). URL: https://economictimes.indiatimes.com/news/how-to/ai-and-privacy-the-privacy-concerns-surrounding-ai-its-potential-impact-on-personal-data/articleshow/99738234.cms.

[6] *Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI | News | European Parliament*. Sept. 12, 2023. URL: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai.

[7] M. Ashok et al. “Ethical framework for Artificial Intelligence and Digital technologies”. In: *International Journal of Information Management* 62 (2022). ISSN: 02684012. DOI: https://doi.org/10.1016/j.ijinfomgt.2021.102433.

[8] D. Banciu and C.E. Cirnu. “AI Ethics and Data Privacy compliance”. In: *Int. Conf. Electron., Comput. Artif. Intell., ECAI*. Institute of Electrical and Electronics Engineers Inc., 2022. ISBN: 978-166549535-6. DOI: https://doi.org/10.1109/ECAI54874.2022.9847510.

[9] J. Bang et al. “Ethical Chatbot Design for Reducing Negative Effects of Biased Data and Unethical Conversations”. In: *Int. Conf. Platf. Technol. Serv., PlatCon - Proc.* Institute of Electrical and Electronics Engineers Inc., 2021. ISBN: 978-166541766-2. DOI: https://doi.org/10.1109/PlatCon53246.2021.9680760.

[10] J. Baric-Parker and E.E. Anderson. “Patient Data-Sharing for AI: Ethical Challenges, Catholic Solutions”. In: *Linacre Quarterly* 87.4 (2020), pp. 471–481. ISSN: 00243639. DOI: https://doi.org/10.1177/0024363920922690.

[11] C. Cadwalladr and E. Graham-Harrison. “Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach”. In: *The Guardian* (Mar. 2018). URL: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

[12] A. Cebulla, Z. Szpak, and G. Knight. “Preparing to work with artificial intelligence: assessing WHS when using AI in the workplace”. In: *International Journal of Workplace Health Management* 16.4 (2023), pp. 294–312. ISSN: 17538351. DOI: https://doi.org/10.1108/IJWHM-09-2022-0141.

[13] A. Cebulla et al. “Applying ethics to AI in the workplace: the design of a scorecard for Australian workplace health and safety”. In: *AI and Society* 38.2 (2023), pp. 919–935. ISSN: 09515666. DOI: https://doi.org/10.1007/s00146-022-01460-9.

[14] M. Christoforaki and O. Beyan. “AI Ethics—A Bird’s Eye View”. In: *Applied Sciences (Switzerland)* 12.9 (2022). ISSN: 20763417. DOI: https://doi.org/10.3390/app12094130.

[15] S. Coghlan, T. Miller, and J. Paterson. “Good Proctor or “Big Brother”? Ethics of Online Exam Supervision Technologies”. In: *Philosophy and Technology* 34.4 (2021), pp. 1581–1606. ISSN: 22105433. DOI: https://doi.org/10.1007/s13347-021-00476-1.

[16] N. Corrêa et al. “Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance”. In: *Patterns* 4.10 (2023). ISSN: 26663899. DOI: https://doi.org/10.1016/j.patter.2023.100857.

[17] *Ethics guidelines for trustworthy AI | Shaping Europe’s digital future*. Apr. 8, 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[18] *EU AI Act: first regulation on artificial intelligence | News | European Parliament*. Aug. 6, 2023. URL: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[19] E. Fournier-Tombs and J. McHardy. “A Medical Ethics Framework for Conversational Artificial Intelligence”. In: *Journal of Medical Internet Research* 25 (2023). ISSN: 14388871. DOI: https://doi.org/10.2196/43068.

[20] B. Han, G. Buchanan, and D. McKay. “Learning in the Panopticon: Examining the Potential Impacts of AI Monitoring on Students”. In: *ACM Int. Conf. Proc. Ser.* Association for Computing Machinery, 2022, pp. 9–21. ISBN: 979-840070024-8. DOI: https://doi.org/10.1145/3572921.3572937.

[21] Jeonghye Han. “An Information Ethics Framework Based on ICT Platforms”. In: *Information* 13.9 (Sept. 2022), p. 440. ISSN: 2078-2489. DOI: https://doi.org/10.3390/info13090440. (Visited on 01/04/2024).

[22] Jonathan Herington et al. “Ethical Considerations for Artificial Intelligence in Medical Imaging: Data Collection, Development, and Evaluation”. In: *Journal of Nuclear Medicine* 64.12 (Dec. 1, 2023), pp. 1848–1854. ISSN: 0161-5505, 2159-662X. DOI: https://doi.org/10.2967/jnumed.123.266080.

[23] A. Hingle and A. Johri. “Recognizing Principles of AI Ethics through a Role-Play Case Study on Agriculture”. English. In: *ASEE Annu. Conf. Expos. Conf. Proc.* American Society for Engineering Education, 2023. ISBN: 21535965. URL: https://peer.asee.org/44029.

[24] H. Hirsch-Kreinsen and T. Krokowski. “Trustworthy AI: AI made in Germany and Europe?” In: *AI and Society* (2023). ISSN: 09515666. DOI: https://doi.org/10.1007/s00146-023-01808-9.

[25] C. Huang et al. “An Overview of Artificial Intelligence Ethics”. In: *IEEE Transactions on Artificial Intelligence* 4.4 (2023), pp. 799–819. ISSN: 26914581. DOI: https://doi.org/10.1109/TAI.2022.3194503.

[26] J. Ibáñez and M. Olmeda. “Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study”. In: *AI and Society* 37.4 (2022), pp. 1663–1687. ISSN: 09515666. DOI: https://doi.org/10.1007/s00146-021-01267-0.

[27] Y. Ikkatai et al. “Octagon Measurement: Public Attitudes toward AI Ethics”. In: *International Journal of Human-Computer Interaction* 38.17 (2022), pp. 1589–1606. ISSN: 10447318. DOI: https://doi.org/10.1080/10447318.2021.2009669.

[28] Anna Jobin, Marcello Ienca, and Effy Vayena. “The global landscape of AI ethics guidelines”. In: *Nature Machine Intelligence* 1 (Sept. 2019). DOI: https://doi.org/10.1038/s42256-019-0088-2.

[29] M.K. Kamila and S.S. Jasrotia. "Ethical issues in the development of artificial intelligence: recognizing the risks". English. In: *International Journal of Ethics and Systems* (2023). Publisher: Emerald Publishing. ISSN: 25149369 (ISSN). DOI: 10.1108/IJOES-05-2023-0107. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85164325205&doi=10.1108%2fIJOES-05-2023-0107&partnerID=40&md5=d8b2e62229483f746e2f349df545d5f0.

[30] A. Khan et al. "AI Ethics: An Empirical Study on the Views of Practitioners and Lawmakers". In: *IEEE Transactions on Computational Social Systems* (2023), pp. 1–14. ISSN: 2329924X. DOI: https://doi.org/10.1109/TCSS.2023.3251729.

[31] M. Khosravy et al. "When AI Facilitates Trust Violation: An Ethical Report on Deep Model Inversion Privacy Attack". In: *Proc. - Int. Conf. Comput. Sci. Comput. Intell., CSCI*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 929–935. ISBN: 979-835032028-2. DOI: https://doi.org/10.1109/CSCI58124.2022.00166.

[32] K. Kieslich, B. Keller, and C. Starke. "Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence". In: *Big Data and Society* 9.1 (2022). ISSN: 20539517. DOI: https://doi.org/10.1177/20539517221092956.

[33] Z. Kilhoffer et al. "'How technical do you get? I'm an English teacher': Teaching and Learning Cybersecurity and AI Ethics in High School". In: *Proc. IEEE Symp. Secur. Privacy*. Vol. 2023-May. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 2032–2049. ISBN: 978-166549336-9. DOI: https://doi.org/10.1109/SP46215.2023.10179333.

[34] B. Kitchenham. "Procedures for Performing Systematic Reviews". In: *Keele, UK, Keele Univ.* 33 (Aug. 2004). ISSN: 1353-7776. URL: https://www.researchgate.net/publication/228756057_Procedures_for_Performing_Systematic_Reviews.

[35] Sourabh Kulesh and Nikunj Dalmia. "In the public AI: Big tech giants like Microsoft, Meta, Google experiment with artificial intelligence tools". In: *The Economic Times* (Aug. 20, 2023). ISSN: 0013-0389. URL: https://economictimes.indiatimes.com/magazines/panache/in-the-public-ai-big-tech-giants-like-microsoft-meta-google-experiment-with-artificial-intelligence-tools/articleshow/102869083.cms?from=mdr.

[36] B.M. McLaren. "Extensionally defining principles and cases in ethics: An AI model". In: *Artificial Intelligence* 150.1 (2003), pp. 145–181. ISSN: 0004-3702. DOI: https://doi.org/10.1016/S0004-3702(03)00135-8.

[37] M. Milossi, E. Alexandropoulou-Egyptiadou, and K.E. Psannis. “AI Ethics: Algorithmic Determinism or Self-Determination? The GPDR Approach”. In: *IEEE Access* 9 (2021), pp. 58455–58466. ISSN: 21693536. DOI: https://doi.org/10.1109/ACCESS.2021.3072782.

[38] S.M. Mohammad. “Ethics Sheet for Automatic Emotion Recognition and Sentiment Analysis”. In: *Computational Linguistics* 48.2 (2022), pp. 239–278. ISSN: 08912017. DOI: https://doi.org/10.1162/coli_a_00433.

[39] M. Nishanthi, H. U. C. S. Kumara, and K. W. a. M. Konpola. “Research Conception of Palm Leaf Manuscript Conservation: Bibliometric Analysis of Scopus database”. In: *International Journal of Multidisciplinary Studies* 10.2 (July 15, 2023), pp. 13–28. ISSN: 2465-6380. URL: https://journals.sjp.ac.lk/index.php/ijms/article/view/6477.

[40] S. Pasricha. “AI Ethics in Smart Healthcare”. In: *IEEE Consumer Electronics Magazine* (2022), pp. 1–7. ISSN: 21622248. DOI: https://doi.org/10.1109/MCE.2022.3220001.

[41] E. Pilkington. “Google’s secret cache of medical data includes names and full details of millions – whistleblower”. In: *The Guardian* (Nov. 2019). URL: https://www.theguardian.com/technology/2019/nov/12/google-medical-data-project-nightingale-secret-transfer-us-health-information.

[42] C. Sanderson, D. Douglas, and Q. Lu. “Implementing Responsible AI: Tensions and Trade-Offs Between Ethics Aspects”. English. In: *Proc Int Jt Conf Neural Networks*. Vol. 2023-June. Institute of Electrical and Electronics Engineers Inc., 2023. ISBN: 978-166548867-9. DOI: https://doi.org/10.1109/IJCNN54540.2023.10191274.

[43] C. Sanderson et al. “Towards Implementing Responsible AI”. English. In: *Proc. - IEEE Int. Conf. Big Data, Big Data*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 5076–5081. ISBN: 978-166548045-1. DOI: https://doi.org/10.1109/BigData55660.2022.10021121.

[44] T. Schuster and L. Waidelich. “Maturity of Artificial Intelligence in SMEs: Privacy and Ethics Dimensions”. In: *IFIP Advances in Information and Communication Technology*. Vol. 662. Springer Science and Business Media Deutschland GmbH, 2022, pp. 274–286. ISBN: 978-303114843-9. DOI: https://doi.org/10.1007/978-3-031-14844-6_22.

[45] D.B. Shank and A. Gott. “Exposed by AIs! People Personally Witness Artificial Intelligence Exposing Personal Information and Exposing People to Undesirable Content”. In: *International Journal of Human-Computer Interaction* 36.17 (2020), pp. 1636–1645. ISSN: 10447318. DOI: https://doi.org/10.1080/10447318.2020.1768674.

[46] V. Sharma et al. “Framework for Evaluating Ethics in AI”. In: *Int. Conf. Innov. Data Commun. Technol. Appl., ICIDCA - Proc.* Institute of Electrical and Electronics Engineers Inc., 2023, pp. 307–312. ISBN: 979-835039720-8. DOI: https://doi.org/10.1109/ICIDCA56705.2023.10099747.

[47] K. Siau and W. Wang. “Artificial intelligence (AI) Ethics: Ethics of AI and ethical AI”. In: *Journal of Database Management* 31.2 (2020), pp. 74–87. ISSN: 10638016. DOI: https://doi.org/10.4018/JDM.2020040105.

[48] V. Vakkuri et al. “ECCOLA — A method for implementing ethically aligned AI systems”. In: *Journal of Systems and Software* 182 (2021). ISSN: 01641212. DOI: https://doi.org/10.1016/j.jss.2021.111067.

[49] J. Yam and J.A. Skorburg. “From human resources to human rights: Impact assessments for hiring algorithms”. In: *Ethics and Information Technology* 23.4 (2021), pp. 611–623. ISSN: 13881957. DOI: https://doi.org/10.1007/s10676-021-09599-7.

[50] Y. Zhang et al. “Ethics and privacy of artificial intelligence: Understandings from bibliometrics”. In: *Knowledge-Based Systems* 222 (2021). ISSN: 09507051. DOI: https://doi.org/10.1016/j.knosys.2021.106994.

[51] Anneke Zuiderwijk, Yu-Che Chen, and Fadi Salem. “Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda”. In: *Government Information Quarterly* 38.3 (July 1, 2021), p. 101577. ISSN: 0740-624X. DOI: https://doi.org/10.1016/j.giq.2021.101577.

Appendix

I. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Daria Korobenko**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Towards Privacy- and Security-Aware Framework for AI Ethics,
   *(*title of thesis)

   supervised by Anastasija Nikiforova and Rajesh Sharma.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Daria Korobenko
4/1/2024