**Joni-Roy Emil Piispanen**

# Current Discourses in Artificial Intelligence Ethics

Master’s thesis of mathematical information technology

June 9, 2023

University of Jyväskylä

Faculty of Information Technology

**Author**: Joni-Roy Emil Piispanen

**Contact information**: joni.r.e.piispanen@student.jyu.fi

**Supervisors:** Ville Vakkuri and Pekka Abrahamsson

**Title:** Current Discourses in Artificial Intelligence Ethics

**Työn nimi:** Tekoälyn Etiikan Nykyiset Diskurssit

**Project:** Master's thesis

**Study line:** Information Technology

**Page count:** 118+29

**Abstract:** This thesis presents the current discourses in artificial intelligence ethics and establishes the relationships between different discourses, topics, and themes. Towards this end relevant literature is collected for analysis using a snowballing approach. The collected data is structured into a map of citation that is analyzed using discourse analysis as the chosen research method. The results obtained in this thesis indicate that the discourses in AI ethics are diverse, containing numerous topics, with both convergent and divergent themes. AI ethics can be regarded as encompassing some overarching focal points, as similar topics and themes with significant overlap are established in the analysis for this thesis. Focusing on discourses revealed the nature and amount of overlap. Conceptualizing the interconnectedness of discourses is one of the achieved results. The citation mapping completed for this thesis revealed that discourses are largely isolated with publications within discourse primarily referencing and citing other publications in their respective clusters.

**Keywords:** Artificial Intelligence, Ethics, Principles, Discourse, Discourse Analysis

**Suomenkielinen tiivistelmä:** Tämä tutkielma käsittelee tekoälyn etiikan nykyisiä diskursseja ja määrittää eri diskurssien, aiheiden ja teemojen väliset yhteydet. Tämän tavoitteen saavuttamiseksi tutkielmassa kerätään olennaista kirjallisuutta analyysiä varten lumipallomenetelmän avulla. Kerätty data strukturoidaan viittausten kartaksi, jota analysoidaan diskurssianalyysin menetelmällä. Tutkielmassa saavutetut tulokset viittaavat siihen, että tekoälyn etiikan diskurssit ovat moninaisia, sisältäen lukuisia aiheita, jotka ovat temaattisesti niin yhteneviä, kuin myös erkanevia. Tekoälyn etiikan voidaan mieltää kattavan joitain yhteisiä keskipisteitä, sillä tutkielman analyysissä aiheiden ja teemojen väliltä löydettiin limittäisyyttä. Keskittymällä diskursseihin voitiin aiheiden päällekkäisyys ja toistuvuus yli diskurssien määrittää. Diskurssien välisten kytkösten käsitteellistäminen on yksi tutkielman saavutetuista tuloksista. Viittausten kartoittaminen paljasti diskurssien eristäytyneisyyden, sillä tutkimusjulkaisut diskurssien sisällä viittasivat pääasiassa toisiin tutkimusjulkaisuihin omien klusteriensa sisällä.

**Avainsanat:** Tekoäly, Etiikka, Periaatteet, Diskurssi, Diskurssi Analyysi

# Glossary

| | |
|---|---|
| AI | Artificial Intelligence |
| ML | Machine Learning |
| AS | Autonomous System |
| A/IS | Autonomous and Intelligent Systems |
| XAI | Explainable Artificial Intelligence |
| AMA | Artificial Moral Agents |
| DA | Discourse Analysis |
| CDA | Critical Discourse Analysis |

# List of Figures

# List of Tables

# Contents

# 1 INTRODUCTION

Artificial Intelligence (AI) has been a part of public discourse for ages, and public perception of AI has been affected by dystopian depictions in science fiction. Academic discourse on AI has changed in the last half century, as the ability to test theoretical possibilities in the form of experimental machines has become feasible (Buchanan 2005). Recent advances in Machine Learning (ML) and AI have however brought renewed interest in AI and especially its impact on society (Paraman and Anamalah 2022). Discourse has shifted from futuristic scenarios to how AI is affecting our daily lives. The impact of AI has been on an unprecedented scale and presented challenges in diverse areas of social life ranging from labor markets (Graetz, Restrepo and Nordström Skans 2022) to health care (World Health Organization 2021).

Mainstream adoption of AI based technologies has transferred long-standing philosophical debates, pertaining to the ethical aspects and considerations in designing AI systems or Autonomous Systems (AS), from academia to the purview of public discourse and scrutiny (Othman 2021). Prominent failures in the governance of AI systems and the ethics of algorithms have resulted in an aura of uncertainty around AI (Brown, Davidovic and Hasan 2021). Highly publicized failures, as seen in the case of Microsoft's chatbot becoming verbally abusive after conversing with users on Twitter (Yampolskiy 2019), and the transformative capabilities of AI systems have the potential to polarize future narratives regarding AI. The ramifications of such media coverage can be seen in public discourse and shape the development and regulation of AI (Ouchchy, Coin and Dubljević 2020). The risks surrounding AI are amplified by a lack of understanding of AI systems and their inner workings (Curtis, Gillespie and Lockey 2022).

## 1.1 Motivation

The research field of AI ethics is multidisciplinary ranging from philosophical topics pertaining to artificial general intelligence and machine consciousness to technological topics relating to the design of ethical autonomous vehicles. This diverse range of topics contains an abundance of ethical considerations, which has resulted in several initiatives from researchers and practitioners alike to combat these topical challenges (Dignum 2018b). The primary outcome of these initiatives has been the issuance and publication of guidelines and frameworks for the ethical design, development, deployment, adoption, and decommissioning of Autonomous and Intelligent Systems (A/IS). This project towards ethical AI has been a combined effort by organizations, governments, companies, and research institutes. Guidelines have been issued by for instance IEEE (2019), ACM (2017), Future of Life Institute (2017), European Commission (2018) and (2021), IBM (2018), with several studies attempting to map the principles proposed in these guidelines, as seen in Larsson (2020), Hagendorff (2020), Ryan and Stahl (2021).

The principles advocated for approximately correspond to those highlighted by Jobin, Ienca and Vayena (2019) in their mapping of principles and guidelines on ethical AI. The researchers found a global convergence on the principles of transparency, justice and fairness, nonmaleficence, responsibility, and privacy in AI ethics guidelines. These results correspond to results attained in later studies, as seen in Ryan and Stahl (2021) and Fahmideh, et al. (2022). These principles are however interpreted in different ways, requiring a considerable amount of ethical analysis and strategies for implementation to support their integration into existing practices (Jobin, Ienca and Vayena 2019).

A principle-based approach, despite its popularity and initial credibility, is concerning in some respects as Mittelstadt (2019) argues, an approach to AI ethics based on principles is problematic, as it suffers from a lack of specificity. Abstract and vague concepts such as beneficence can be interpreted in a myriad of ways. The result of such ambiguity, that is seemingly inherent in guidelines created by AI ethics initiatives, is incoherence (Munn 2022). This property has been built in by design to address the use of AI across varying contexts. From a practitioner's perspective ambiguity presents a challenge, as contested notions of ethical design in guidelines contain inherent disagreements about their normative content (Mittelstadt 2019). Thus, the need for high-level all-encompassing principles and action-guiding recommendations stand in relative opposition to each other.

The problematization of the principle-based approach has revealed inherent deficiencies in tackling AI ethics through principles. This has raised questions about the efficacy of predominantly using principles to guide the development of AI ethics, with its proponents arguing for the merits of high-level principles as guiding forces for the creation of behavioral norms and cultural values (Seger 2022). The realization of the inherent shortcomings of the principle-based approach could serve as a catalyst for the adoption of alternate approaches. Thus, a key issue in this thesis will be determining whether the current discourses in AI ethics have shifted away from principles and a marked influx of interest towards other approaches in recent research can be seen.

## 1.2 Research problem

The amount of literature on AI ethics has increased significantly in recent years (Borenstein, et al. 2021). Due to its ubiquitous nature, AI ethics has attracted interest from a diverse field of researchers with varying backgrounds and methodological approaches. This multidisciplinarity has the potential to obfuscate discourse about the ethics of AI and decrease the intelligibility of the research field. As the predominant principle-based approach to AI ethics has been scrutinized and received substantial criticism, a pathway for alternate approaches has emerged. An overview of the current discourses and issues within AI ethics would potentially alleviate these concerns. Thus, the purpose of this thesis is to research the central discourses in AI ethics and examine whether a shift from establishing principles has occurred. The first research question is:

- What are the current discourses in AI ethics?

The secondary objective for this thesis is to ascertain whether these discourses constitute a coherent field of research with a meaningful amount of interaction or diverge into separate topics of interest. Understanding the current state of research in AI ethics requires an analysis of these discourses for prevalent topics, themes, and the connections between them. Studying the discourses and the interconnectivity between them can reveal whether research has progressed in a meaningful way or whether alternate approaches towards AI ethics are simply reinventing the wheel. Additionally, this will reveal where future research should be focused. Therefore, the second and third research questions are:

- What are the prevalent themes and topics?
- How do they relate to each other?

Through studying these questions, a summary of the current discourses and their relations can be created, which in turn can serve as a compass providing clarity and structure for researchers and other parties interested in the ethics of AI.

## 1.3 Research methods

This thesis adapts the guidelines for systematic literature reviews by Kitchenham and Charters (2007) and guidelines for snowballing in literature studies advocated for by Wohlin (2014) to collect relevant literature for analysis. Snowballing approaches are the preferred procedure to conducting literature reviews, when researching topics with varied terminology (Wohlin 2014). AI ethics being a novel research field with a substantial amount of conceptual ambiguity warrants the use of such an approach. The start set for the snowballing procedure consists of Mittelstadt's (2019) paper as it directly relates to the first research question, while simultaneously being arguably the most impactful critique of a principle-based approach and an influential and cited paper. Additional research literature is collected using the snowballing approach. The collected data is visualized as a map of citations and references. Chapters 3 and 4 elaborate on the research methods used and describe the research process in detail.

The collected literature is analyzed using Discourse Analysis (DA) as a research method with the objective of identifying and studying the main themes and topics in current AI ethics discourses. The research method is applied following the methodology advocated for by Yazdannik, Yousefy and Mohammadi (2017). DA is a qualitative research method primarily used to analyze diverse collections of media in the social sciences and in medical research. DA was chosen as the research method for its inductive approach, which enables the study of a subject matter with varied and contested theoretical perspectives. The objective of this approach is to establish categories of discourses and identify the central themes and topics within the field of AI ethics without them ex ante guiding the research. Previous analyses have primarily been either systematic literature reviews or systematic mapping studies, that have focused on establishing the current state of research in AI ethics. The upshot of a focus on discourses is the discovery of the interconnectedness or separateness of different discourses and topics within AI ethics. Thus, the approach adopted in this thesis can provide novel and complementary results to previous analyses. The identified discourses, topics, and themes are visualized in chapter 5.

## 1.4 Structure of work

Chapter 2 examines the history of AI and discusses common terminology and their definitions. The chapter introduces popular approaches to intelligent systems and discusses the technologies that are commonly regarded as AI. The rest of the thesis will build upon this foundation and the discussion in chapter 2 will provide context for the current issues discussed in later chapters. The chapter progresses into examining the current state of AI ethics research with the objective of grounding this thesis in the larger context of the research field. The chapter will mainly explore the literature of the 21st century, as the central interests of this thesis are the current discourse in AI ethics.

Chapter 3 introduces discourse analysis and its related theoretical foundations. The research method is reviewed in detail by discussing the different schools of thought within discourse analysis and their respective approaches and developments. The research process is documented with critical reflection using relevant literature on methodology. The chapter ends with a discussion on the chosen methodology's relation to discourses in AI ethics.

Chapter 4 presents the research design and data collection methodology for this thesis. The research method used for the literature search is introduced and analyzed. Each phase of the process is documented, and the results of these phases are visualized. Relevant literature on methodology is used to reflect on the process and design choices made during this thesis.

Chapter 5 reviews the results of this thesis. The prominent discourses, topics and themes identified in the collected data for this thesis are examined in detail. The chapter begins with a general overview of identified discourses and their relationships. Subsequently, particular discourses are analyzed with an outline of the articles that are the primary contributors towards the discourse. Each section includes a figure representing the discourse and the articles and their relationships that form the discourse.

Chapter 6 relates the research results to previously obtained results and conclusions drawn in the current literature. The impact and validity of the obtained results are critically examined and reflected on using relevant research literature. The chapter ends with a discussion of the limitations in data collection and analysis performed for this thesis.

Chapter 7 concludes the thesis. The research process and results are summarized with relevant topics for further study being briefly discussed. The chapter provides pathways for complementary and further research, while covering topics that were excluded from this thesis.

# 2 BACKGROUND

This chapter contains a brief review of AI and its history, which serves as a foundation for the discussion of the present-day issues and topics in the ethics of AI. The chapter introduces the established terminology related to intelligent systems and approaches that are commonly regarded as AI. This context will function as the basis that the rest of this thesis builds upon.

## 2.1 Artificial Intelligence

Colloquially AI is used to refer to computer systems and algorithms exhibiting some manner of intelligence and autonomy. There are a multitude of approaches in the research field of AI that attempt to create a foundation for the understanding and creation of intelligent systems (Goertzel and Wang 2007a, 75-76). A prominent approach has been the modelling of human problem-solving and the cognitive processes and mental capabilities related to intelligent behavior, which could be represented in computerized models serving as pathways to the creation of AI (Korteling, et al. 2021). At the other end of the spectrum attempts to formalize intelligence and rationality have also seen prominence (Dobrev 2003). Depending on the approach a different focus is chosen, with some researchers studying internal processes such as reasoning and others external ones such as behavior (Russell, Chang, et al. 2022, 19). Artificially intelligent systems can generally be separate into four distinct categories: systems that think like humans, systems that act like humans, systems that think rationally, systems that act rationally (Kok, et al. 2009). This variety has caused inevitable terminological confusion, with a core issue being the definition of intelligence itself (Dobrev 2003). This thesis will use the previously discussed categorization, which is based on Russell, et al. (2022, 19-22), as the grounding for the discussion on AI.

Intelligence as its commonly understood refers to the mental or intellectual capability displayed by human beings (Bhatnagar, et al. 2018). This perspective intrinsically links intelligence with fidelity to human performance. Viewed in this way, parallels between artificial intelligence and human intelligence can be drawn in a self-evident manner. This notion is exemplified in the Turing test, which stems from Alan Turing's (1950) thought experiment. According to this view a computer programmed in the right way, with the capabilities of natural language processing, knowledge representation, automated reasoning and learning would be considered intelligent, if it could convince a human interrogator by producing indistinguishable responses from a human (Russell, Chang, et al. 2022, 20). Additional requirements of computer vision, speech recognition and object manipulation via robotics have been added by later researchers.

Evaluating the intelligence of an AI based on human task performance benchmarks has provided researchers with a useful metric in determining the breadth and flexibility of these systems (Simon 1995). The approach has however been criticized for its inherent anthropocentrism, which according to its detractors misconstrues AI with human intelligence (Korteling, et al. 2021). They argue that AI should not be the simulation of human intelligence as the architecture and underlying substrate of artificial intelligence differs from biological intelligence.

An alternative approach that has received considerable attention is modelling the cognitive functions of a human mind. Through modelling human cognition researchers can build testable theories of the human mind, unveiling the processes behind intelligence, which in turn can be expressed as computer programs leading to AI (Lieto, et al. 2018). The caveat of these human-centric approaches is that AI is tinted through an anthropomorphized lens and takes on characteristics commonly attributed to humans, which has the inevitable effect of biasing expectations and restricting research (Salles, Evers and Farisco 2020).

At the other end of the spectrum, intelligence can be defined in a formal way through rationality. Rational agents act in a way that achieves the best expected outcome in an environment, striving for optimal action in any situation. In complex environments perfect rationality i.e., taking the optimal action, becomes unfeasible as processing information and taking the appropriate actions requires time (Russell 1997). Thus, a key issue related to optimal action concerns bounded rationality, i.e., rational choice given certain limitations regarding knowledge and computational capacity (Simon 1990, 15-18).

The hitherto discussed approaches are ordinarily categorized under the artificial general intelligence umbrella term (McCarthy 2007). Goertzel (2007b) however argues, that constraining AI to the scale of human performance, might pose overly strict limitations to the future scope and work on general intelligence. Generally intelligent systems or human-level AI is currently an out of reach technology. The technologies that are presently operating in society and referred to as AI fall commonly under the narrow AI umbrella, which encapsulates the more prominent variants of systems exhibiting narrow intelligence and capabilities. These systems perform exceptionally well in specific narrowly defined problem domains such as can be seen in for example chess playing algorithms (Goertzel and Wang 2007a, 36).

The previously discussed perspectives constitute and represent some of the more prominent approaches to AI. These definitions and approaches are however only some of multitudes of conflicting accounts of what should be classified as artificial intelligence (Fast and Horvitz 2017). With the rise of machine learning technologies, especially neural networks, the research field of AI has changed drastically in recent years (Schmidhuber 2015). The approaches discussed serve as representations of what the field of AI contains and exemplify the kinds of technologies referred to in this thesis as AI, even though the subject matter also applies and is pertinent to other kinds of technologies.

## 2.2 AI Ethics

The research field of AI ethics is vast, encompassing a multitude of different technologies and related ethical considerations. AI ethics is rooted in computer ethics and machine ethics, which are concerned with studying the interaction between humans and machines and the operation and creation of artificial intelligence agents, that behave morally (Borenstein, et al. 2021).

AI ethics is multidisciplinary covering both technical and social issues. This abundance of perspectives presents challenges in managing AI related impacts. Thus, AI systems and their effect on society has been heavily scrutinized. Governing AI systems has become a critical issue, as the mechanisms by which AI development is steered towards a socially purposeful and beneficial direction have been uncertain (Djeffal, Siewert and Wurster 2022). Different organizations have addressed the ethical implications of artificial intelligence by publishing policy strategies and frameworks. (Schiff, Biddle and Laas 2021). However, navigating the different roles and responsibilities of governments, industry actors and organizations in governing AI has been a challenge. According to Vica, Voinea and Uszkai (2021), the interests and needs between public and private sector entities are partially incompatibility.

Nevertheless, there has been an overarching effort by companies, organizations, and governments to design, implement, and use AI technologies in a responsible and ethical manner (Trocin, et al. 2021). This project applies to both the organizational level and the systems level of procedures and mechanisms. In this context, responsible AI development is generally regarded as denoting AI that is fair, non-biased, transparent, explainable, secure, safe, privacy-proof, accountable, and to the benefit of humanity (de Laat 2021). The result of these efforts has primarily been research and policy initiatives published by academia, organizations, and companies (Nabavi and Browne 2022). These efforts and initiatives have had a questionable practical impact. As Nababi and Browne (2022) indicate, affecting meaningful change is a challenge for industry actors. Additional challenges pertain to managing stakeholder expectations and involvement in AI development and communicating the risks and benefits of AI appropriately (Liao and Sundar 2022). As Varona and Suárez (2022), argue attaining trustworthiness and fairness requires involvement in the design of AI systems by diverse stakeholders.

The technical challenges involve developing AI systems that adhere to ethical standards. The ethical design of AI technologies can be categorized into three separate focal points, namely ethics by design, ethics in design and ethics for design (Dignum 2018b). Dignum’s structuring of the design on AI ethics technologies is not definitive but provides a useful approach towards making the topic intelligible. Alternative categorizations are discussed by for example Greene, Hoffmann and Stark (2019) or Kazim and Koshiyama (2021).

According to Dignum (2018b), ethics by design comprises issues related to the integration of ethical reasoning capabilities into A/IS through technical or algorithmic means. This includes studying pathways towards implementing ethical decision-making in AI systems, once a set of ethical principles is given. A/IS should be able to consider the varied and conflicting societal, ethical, and moral values of all relevant stakeholders and their respective priorities in multicultural contexts during operation. Taking these considerations into account these systems would have the capability to reason about the ethical aspects of their decisions in an autonomous way. Endowed with such capabilities, these systems could be considered as Artificial Moral Agents (AMA) with an understanding of right and wrong culminating in a system of moral beliefs affecting their decision-making (Dignum, Baldoni, et al. 2018a).

Ethics in design revolves around the engineering methods and related regulation required to support the evaluation and study of the ethical implications of these systems as they are integrated into society and its structures. Ethics for design is concerned with the codes of conduct, standards and certification processes of developers and users, when they are researching, designing, constructing, employing, and managing A/IS. As these systems are constructed, certain values are implicitly embedded within them. The explication of these values in both deliberation and decision processes requires theories and methods to ensure that they are beneficial for society. Therefore, the elicitation of stakeholder values takes on a pivotal role.

## 2.3 Principles for Ethical AI

The design of AI systems raises crucial questions relating to accountability, responsibility, and transparency. The system's ability to explain and justify the mechanisms behind its decisions is crucial when operating in the real world. Ensuring accountability and transparency requires intelligibility and clarity for the different stakeholders when these systems are interacted with (Zicari, et al. 2022). The reproducibility of decisions and evaluation of fallibility in different contexts is of paramount importance when establishing a chain of responsibility in human decision-making. In the case of AI systems further complications arise since machines might be considered less fallible than humans (Giuffrida 2019). The issue is, however, anything but straightforward when dealing with ambiguity of objectives and conflicting moral considerations. Considering the severity of the potential social implications of AI systems and their actions, mechanisms that navigate the chain of responsibility from machines to humans are required (Trocin, et al. 2021).

Among the AI ethics literature, a profusion of guidelines and ethical principles have been proposed, with a tentative convergence upon a set of common issues (Hagendorff 2020).

One of the more prominent among these guidelines is the IEEE (2019) guideline, which articulates high-level principles and imperatives for the design, manufacture, and use of A/IS to ensure the alignment of these systems with societal and ethical values. The general principles proposed in the IEEE (2019) guidelines are those concerning human rights, human well-being, data agency, effectiveness, transparency, accountability, awareness of misuse and creator competence. These principles are meant to guide the development and operation of A/IS to reach beyond the achievement of functional goals and technical problems, promoting trust between the public and automated technology.

The principles set out in AI ethics guidelines approximately correspond to the core principles in principlism, which is a prominent ethical approach in bioethics (Whittlestone, et al. 2019). Principlism designates a framework based on basic human rights culminating in four universal ethical principles, respect for autonomy, nonmaleficence, beneficence, and justice (Beauchamp and Childress 2013).

In the case of AI, respect for human autonomy encompasses navigating the balance between human decision-making and the delegation of tasks to artificial agents. As these systems approach the status of AMAs with ever-increasing intricacy and fidelity, the risks to human autonomy and flourishing escalate (Herzog 2021). Thus, mechanisms that protect or re-establish human control and autonomy are required. This necessitates the intrinsic restriction of machine autonomy and potential to override the decision-making of machines when deemed necessary (Floridi and Cowls 2019). The objective should be to empower human autonomy through the ability of exercising freedom in decision-making and ceding control in the cases where overriding reasons emerge.

The principle of non-maleficence, typically interpreted as do no harm, assumes several roles in relation to AI. Given any degree of autonomy A/IS have the potential to adversely impact society and humanity on a grand scale. Thus, enforcing the principle of non-maleficence requires considering the different capabilities and potential for misuse of A/IS as a technology. Additionally, the principle of non-maleficence can be seen to apply to the developers of these systems. Acknowledging the risks and responsibilities related to the creation of A/IS requires careful deliberation and relevant stakeholder participation as the negative externalities of these systems can be unpredictable.

Beneficence is usually seen as an umbrella term for the promotion of the common good or well-being and benefit of humanity. In AI ethics guidelines beneficence is at times interpreted in a broader sense not only relating to the promotion of human dignity, but also the assurance of sustainability as seen in the European Commission's European Group on Ethics in Science and New Technologies report (2018). Viewed this way, beneficence dictates mitigating the impact of AI on nature and ensuring the environmental protection and continued sustainability of life on earth for future generations.

Artificial intelligence as a technology has the inherent potential to accentuate existing injustice and discrimination (Villegas-Galaviz and Martin 2022). The ability to utilize emerging technologies is fundamentally unequal across society. Thus, ensuring fairness in access and non-discrimination in the widespread deployment of AI requires special deliberation. Mitigating these risks requires the explicit consideration of the principle of fairness and pathways that lead to shared benefit and prosperity from AI (High-Level Expert Group on Artificial Intelligence 2019).

To supplement these four universal ethical principles, some researchers have argued that additional principles such as explicability are needed in the special case of AI as they represent a new dimension of ethical consideration related to computer systems and algorithms (Floridi and Cowls 2019). Explicability as a distinct principle constructs and maintains trust between AI and its users (High-Level Expert Group on Artificial Intelligence 2019). This requires transparency of the internal processes that inform decision-making and communication and explanation of the AI systems purposes and capabilities, when applicable. Explicability should consequently entail the intelligibility and accountability of decision-making processes in AI to the parties directly and indirectly affected (Floridi and Cowls 2019).

As previous studies have concluded, communication plays a pivotal role in building and sustaining ethical practices (Rousi 2022). As Saariluoma, et al. (2018) indicate, trust and understanding of a technology are linked in a fundamental way. The prediction of behavior and performance dictates the user's perception of and confidence in the technology being used. In relation to the ethics of AI, principles of accountability, responsibility, and transparency represent a pragmatic necessity and normative obligation for developers to account for these communicative aspects of AI systems. In the special case of AI systems, explainability and transparency require the explication of the systems inner workings. However, the explainability of decision-making procedures and mechanisms of AI becomes challenging in the case of black box systems, i.e., AI that relies on complex functions learned from large datasets that are therefore opaque and inscrutable (Cortese, et al. 2022). In the case of such systems alternative methods are needed, such as auditability and traceability measures (Cortese, et al. 2022). Therefore, topics relating to Explainable AI (XAI) and the ethical issues regarding the transparency of black box systems have received considerable attention (Arrieta, et al. 2020).

According to Mittelstadt (2019), the convergence of AI ethics around principles of bioethics provides a pertinent backdrop to draw parallels on. Principlism in bioethics establishes pathways towards identifying and conceptualizing ethical challenges and guiding clinical decision-making and health policy by providing a common language. Mittelstadt (2019) argues, that adopting a principle-based approach for AI ethics appears to embed normative considerations into the design and governance of technology. The similarities between AI ethics and bioethics extend only to a certain point. Certain key distinctions between the disciplines suggest that a principle-based approach, that has seen success in bioethics might not be as effective in AI ethics.

Mittelstadt (2019) highlights the apparent lack of a subject in AI ethics, whereas bioethics shares a common goal at a fundamental level, i.e., the health of the patient. This is not the case in AI development, as the interests of private sector companies lie in prioritizing profit and cutting costs at all junctions. This conflict of interests in AI ethics transforms ethical decision-making from a cooperative process into a competitive one, as the purposes of developers and users misalign. This disconnect is traditionally assuaged through fiduciary duties and commitments to uphold public interests in formal professions. AI development is however not considered to be a formal profession and as such doesn’t entail commitment to such value statements. To combat these issues, some researchers have suggested a Hippocratic oath for AI researchers and developers (Smith and Shum 2018, 9). As AI development as a profession progresses and A/IS become ever more ubiquitous, equivalent oaths might become a necessity (Strümke, Slavkovik and Madai 2022).

AI ethics is an emerging discipline with a relatively short history compared to bioethics and as such good practices are not as established. The standards and norms guiding ethical behavior are lacking compared to the accounts of moral obligations that have developed and evolved during the long history of health professions (Mittelstadt 2019). This shortcoming can be perceived in developer understanding and concern regarding ethical issues in A/IS development (Vakkuri, Kemell, et al. 2020b). Remedying this issue would necessitate the incorporation of ethical training in curricula as seen in the training of medical students (Roberts, et al. 2004).

This shortcoming is emphasized by the absence of empirically validated methods in translating principles into practice (Mittelstadt 2019). For AI ethics guidelines to have a real-world impact, mechanisms that ensure their appropriate adoption accompanied by practical efforts of testing, application and implementation are needed. This becomes especially important, when considering the time and resource cost of pro-ethical design and the ambiguous subsequent return on investment (Morley, Kinsey, et al. 2021b). Compared to bioethics, AI ethics is wanting for legal and professional accountability mechanisms (Mittelstadt 2019). Ensuring the preservation of ethical principles in self-governance therefore demands complementary mechanisms that uphold professional standards and enable the possibility for redress when deemed necessary.

The practical implementation of ethical guidelines has been a continuing development in AI ethics. As Vakkuri, et al. (2019) indicate, developers are generally not well-informed on ethical matters and frameworks for understanding these issues are scarce. This has led to the lackluster adoption of guidelines on ethical AI by software companies (Vakkuri, Kemell and Kultanen, et al. 2020a). Thus, bridging the gap between principles and practice requires the development of new tools and methods that promote tangible ways of implementing ethics.

Morley, et al. (2020) propose that the fragile consensus on principles among guidelines could serve as a common framework in translating high-level ethical principles into actionable recommendations. In their view, overcoming the gap between principles and practice requires a multidisciplinary approach, which could alleviate the challenges of social complexity, algorithmic uncertainty, and unpredictability inherent in AI and ML (Morley, Floridi, et al. 2020). This multidisciplinary approach has the advantage of accentuating the presence and absence of tools and methods available to researchers and practitioners. In addition, it fosters the development of a common language, which reduces the burden created by the use of different terminology and definitions.

Topics relating to the operationalization of guidelines and the governance of A/IS through different mechanisms have been discussed as potential solutions to bridge the gap between principles and practice (Mökander and Floridi 2022). In particular ethics-based auditing has received attention as a mechanism that facilitates the governance of AI in corporate contexts. Ethical auditing of AI systems and impact assessments have been suggested as pragmatic approaches to mitigate negative impacts of algorithms throughout their entire life cycle (Brown, Davidovic and Hasan 2021). Through the auditing of AI systems for ethical considerations, their alignment to relevant stakeholder metrics and interests can be established resulting in increased trustworthiness of algorithms and systems.

# 3 DISCOURSE ANALYSIS

This chapter introduces discourse analysis, which is the chosen research method used to analyze the collected data. The concept of discourse and its related nomenclature is analyzed through the prism of pre-eminent discoursal traditions. The different traditions and perspectives of discourse analysis are reviewed with a brief historical accounting of the methodology. This chapter additionally grounds the tools and methods of discourse analysis and explains their suitability for this thesis.

## 3.1 Discourse and Related Concepts

Conceptually discourse takes on different meanings depending on the perspective and context of its use. In a general sense discourse is used to refer to any form of language in use. Discourse can be further conceived as denoting structures of language in texts that transcend sentences or clauses (Baker and Ellece 2011, 30-31). This conception of discourse relates itself to different contexts of language in use and is connected to different genres of texts, for instance academic discourse and the language contained within and ascribed to academic texts. In a larger context discourse can be interpreted as encompassing the ideas and social practices that are circulating in society, establishing and enforcing prudent ways of creating meaning (Foucault 1972, 48-49).

Discourse analysis has its roots in social constructionism, which in turn presupposes that understanding reality is a historically and culturally situated process that occurs in social situations (Gergen 1985). Social constructionism embraces a critical viewpoint of knowledge and the attribution of meaning. According to this viewpoint, rather than being reflections of reality our structuring and understanding of phenomena is perpetually open to interpretation and therefore mutable (Parker 1998, 88-89).

The predominant focus of discourse analysis is on the social, political, and semiotic aspects of language and its use. Linguists practicing discourse analysis primarily study language in use and are interested in the structure and function of texts within the linguistic and social context they are deployed in (Eisenhart and Johnstone 2008, 8). Studying the connection between language and social reality and the way in which language acts as reality building and as action with consequence is a critical aspect of discourse analysis (Wood and Kroger 2000, 4). Identifying and studying the existence of parallel and competing systems of meaning within discourses is a key facet of research. This includes discerning the processes that lead to these systems and ascertaining their relationship to the contextuality of meaningful action.

The assumption in discourse analysis is that language has different potential uses and related meanings. Thus, social actors have differing possibilities and resources available when using language and participating in discourse (Holzscheiter 2005). This in turn produces a hierarchy of meaning with some discourses in the marginal and others all together excluded. These hierarchies can be analyzed through deconstruction and reconstruction, which serve a pivotal role in discourse analysis, by providing a means towards analyzing texts for included and excluded language or topics (Janks 2005). A text can be analyzed for its ideational, interpersonal, or textual role and function.

Discourse analysis can shift between a focus on specific texts and a focus on the order of discourse. The latter encompassing the relatively durable social structuring of language which is itself one element of the relatively durable structuring and networking of social practices (Fairclough 2003, 3). A texts functions directly relate to the intertextuality of discourses and the order of discourses, which are partially formed through the polyphony of language, i.e., the quoting and referencing of other texts (Tannen, Hamilton and Schiffrin 2015, 44). Through this notion intertextuality becomes interconnected to both the dialogism and addressivity of a text, thereby connecting it to the broader literary and cultural context it is embedded in. As Bakhtin (1986, 95-96) notes, the construction and style of a text is inevitably affected by its writer and addressee. Thus, a text is shaped and formed through its relationship to the prevailing conventions and the order of discourse, which their structuring constitute (Fairclough 1992, 103).

Through this process new conversational turns are formulated, with prior texts and remarks forming pathways and connections thereby creating a larger category of interconnected texts (Johnstone 2008, 164). As Kristeva (1980, 66) argues, any text can be seen as being constructed of a mosaic of quotations, essentially being the absorption and transformation of other texts. These connections in turn form common understanding across contexts and situations (Agha 2005). To a larger extent they represent systems of thought that represent the world from a particular perspective, providing a framework for organizing meaning, guiding actions, and legitimating positions (Fairclough 1992, 64-65).

Scientific articles utilize this process primarily through quotation and referencing of previous texts, thereby positioning themselves as part of a literature (Bazerman 1994). Thus, a consensus regarding the value and meaning of previous studies is achieved through the constant reformulation of prior literature. As Bazerman (1991) points outs, the eventual product of this process is codified knowledge comprised of a standard set of associations. This intertextual field establishes frames of meaning and is explicitly reconstructed through each novel finding, claim or argument (Bazerman 1991). Therefore, analyzing any scientific discovery detached from the situating intertext entails only a partial analysis, as the relevance of the domains dynamic of established knowledge is abandoned.

Discourse analysis in the case of information systems consists of studying the interaction between technology, individuals, and social entities. Discourses in this context relate to understanding technologies and their uses, as well as how perceptions and interpretations affect technology use (Stahl 2004). Technology is designed and implemented in a societal and historical context with a variety of social considerations impacting its use and development (MacKenzie and Wajcman 1999). Therefore, technology takes on characteristics that are defined by the socially shared structures of meaning associated with them (Latzko-Toth 2016). These characteristics reflect the perceptions of various groups of actors towards a technological artifact and their interpretations of its functions. The perceptions and modes of action in relation to technological artifacts constitute interpretive frames, that govern the attribution of meaning (Bijker and Pinch 1987). As such, the impact and social uses of a technology are shaped through the assumptions, expectations, and knowledge imbued to it by people and their insight into its meaning and purpose (Lindgren and Holmström 2020).

The transformative process of shaping a technology consists of negotiating the meaning and compatibility of a technology with society’s prevailing values and norms. The process, whereby technological artifacts and their users reshape one another, is one of co-construction. As Orlikowski and Scott (2015) argue, such material-discursive practices constantly create and reconfigure the world. AI, as a socio-technical phenomenon is formed by virtue of this co-construction process that negotiates and establishes interpretative frames (Lindgren and Holmström 2020). The social and ethical impacts of AI are generally studied from the perspective of the technologies effect on society. The study of these systems is predominantly carried out by the same scientists that develop them (Rahwan, et al. 2019). This results in a research agenda, that is focused on the technology itself, in favor of studying the social and cultural contexts and the impacts of society on AI systems (Lindgren and Holmström 2020). Thus, the social and discursive aspects of AI are neglected, resulting in a narrow and deterministic perspective.

Acknowledging the role of discourses and their effect on technology functions as the premise for the analysis conducted in this thesis. By analyzing the AI ethics literature through references and quotations, research articles can be situated in their respective intertextual fields and their relationship to other texts and the larger formation of discourses can be elucidated. This process culminates in a map of citations, in which clusters of articles and their respective dynamics can be analyzed. Performing this analysis requires an understanding of discourse analysis on a theoretical and methodological level, which are the respective subjects of the following sections.

## 3.2 Theories of Discourse

Theories of discourse can be categorized according to the three traditions of discourse analysis namely British, French, and German (Remes 2006). The traditions can be differentiated by the specific setting of discourse and perspective each is interested in. The British tradition studies conversations and the discourses that are formed in them. This perspective is exemplified for instance in an individual person's operation while situated in a particular context, such as a teacher's interactions and conversations in a classroom (Mayes 2010). In this frame of reference discourses necessitate attitudes, that are seen as coherent and reasonable in the prevailing situation and culture. Analysis in turn determines the avenues through which discourse affects the person's behavior and speech.

In the French tradition on the other hand, discourse has a fundamental link to culture, which acts as a foundational force for the creation of discourses. This perspective is epitomized in Michel Foucault's notion of discourse, where ideology parallels language in the way it provides a systematic way of thinking about a topic (Foucault 1972, 98). Thusly viewed, discourse governs the mechanisms that make a topic intelligible, and objects of knowledge are constituted. According to Foucault (1972, 37-38), a discourse therefore directs reasoning about a topic and constrains dialogue pertaining to it. Analysis accordingly unveils the discourses that are formed and upheld through prevailing practices, providing a means towards describing the nature of the dominant culture in society. Uncovering the pathways that legitimize these prevailing practices enables the understanding of a person's roles and situatedness in a culture.

The German tradition examines the nature of reality and discourses pertaining to ways of influencing it. The most influential discourse theory in this tradition is Jürgen Habermas's. Central to this theory are the notions of communicative action and communicative rationality. Habermas's theory of rationality abandons the connection of rationality with the possession of particular knowledge in favor of a view, where rationality is exemplified in the acquisition and usage of knowledge by speaking and acting subjects (Habermas 1984, 11).

In Habermas's theory, language takes the role of coordinating action and priming subjects with a practical attitude towards mutual understanding. Habermas concept of communicative action refers to subjects engaging with this kind of stance. Mutual understanding necessitates the establishment of mechanisms that facilitate agreement between rationally motivated subjects. Thus, acceptability conditions, i.e., reasons and justifications that would be considered reasonable and convincing and therefore make a speech act understandable, are essential to Habermas's project. This concept of acceptability in Habermas's account designates a category of validity claims, which include in addition to the purely empirically validated truth condition, also notions of moral rightness, authenticity, sincerity, and aesthetic value (Habermas 1984, 20-22).

Discourse analysis as discussed by Hodges, Kuper and Reeves (2008) is divided into formal linguistic discourse analysis, empirical discourse analysis and critical discourse analysis.

Formal linguistic discourse analysis focuses on the meaning of texts and their forms. The purpose of this kind of analysis is to describe the use of language within the frame set by the material and its writers. Of particular interest are the grammar that is used, the vocabulary, the semantics of the text, its cohesion and dialogicality (Hodges, Kuper and Reeves 2008).

Empirical discourse analysis studies the processes in which social reality is produced and maintained. The objective is to uncover shared meanings and interpretations of phenomena, which includes discovering the ways in which discourses construct and shape our understanding (Hodges, Kuper and Reeves 2008).

Critical discourse analysis recognizes and analyzes the role of language in structuring power relations in society. Critical discourse analysis views ideology as an essential component in determining the ways in which meaning is constructed and conveyed. Contained within this perspective is the assumption that responsibility accompanies the means and opportunities of improving existing conditions and inequalities (Hodges, Kuper and Reeves 2008).

This thesis will use discourse analysis as a research method focusing on empirical discourse analysis utilizing a Habermasian theoretical grounding. The research method is applied following the recommendations by Yazdannik, Yousefy and Mohammadi (2017). Due to the interpretative nature of qualitative research and especially discourse analysis, careful documentation and argumentation about design choices is necessary. According to Yazdannik, Yousefy and Mohammadi (2017), establishing the methodological and interpretative rigor of research employing discourse analysis necessitates, the appropriate formulation and specification of the research questions, such that they fit the research method used. The objective of this thesis is researching prominent discourses within AI ethics and establishing prevalent topics and themes. Thus, discourse analysis is considered an appropriate choice.

Additional recommendations necessitate suitable research questions for the chosen corpus of texts. According to Yazdannik, Yousefy and Mohammadi (2017), the corpus should be representative of the studied research area. The collected data for this thesis is arguably representative of the academic discourse on AI ethics, as primarily research articles and research papers were analyzed. The interpretative paradigm, as well as data-gathering and data analysis methods for this thesis are explicated and documented in a comprehensive manner, ensuring the possibility for readers to follow along.

# 4 DATA COLLECTION

The data collection for this thesis is conducted following phases of the guidelines for literature review by Kitchenham and Charters (2007) and guidelines for snowballing in literature studies by Wohlin (2014). Literature studies are a form of secondary studies that collect and synthesize previously published studies called primary studies. The purpose of literature studies is to identify, evaluate and interpret previous research related to a particular research area or phenomenon of interest (Kitchenham and Charters 2007). Literature studies, including both reviews and maps, present researchers the possibility of synthesizing evidence in a research area and consequently form a joint understanding of its status (Wohlin 2014). The chosen research method belongs to the evidence-based software engineering paradigm (Kitchenham, Dybå and Jørgensen 2004).

The snowballing procedure provides an alternative approach to conducting a literature review. In this thesis the snowballing approach was used to acquire a perspective on the current state of the AI ethics literature. The snowballing approach should be preferred, when researching topics for which regular database searches are unsuited for (Wohlin 2014). AI ethics is a novel and emerging research field, with inconsistent and varying terminology, which complicates the conduction of a systematic literature review. Systematic literature studies should conventionally encompass all or most of the relevant research (Kitchenham and Charters 2007). This however poses practical difficulties, especially in broad and undefined research areas such as AI ethics. Thus, a systematic literature review was considered unfeasible.

The purpose of this thesis was to create a perspective of the research area. The snowballing approach provided an avenue to research and map the interconnectivity of discourses in AI ethics. For this objective a partial adaptation of the literature review guidelines by Kitchenham and Charters (2007) was considered prudent. The data collection was completed during October 2022. The data collection process consisted of several stages, which are depicted below:

1. Planning the Review
    - Identification of the need for a review

    - Commissioning a review
    - Specifying the research question(s)
    - Developing a review protocol
    - Evaluating the review protocol
2. Conducting the Review
    - Identification of research
    - Selection of primary studies
    - Study quality assessment
    - Data extraction and monitoring
    - Data synthesis

## 4.1 Planning the Review

Planning the review consists of identifying the need for a review, defining appropriate research questions and a review protocol. Commissioning a review was not relevant for this thesis as the review was not performed on a commercial basis. Furthermore, evaluating the review protocol was deemed unnecessary, which is the researcher's prerogative according to Kitchenham and Charters (2007). The motivation for this thesis was discerning the current discourses in AI ethics, since the recent notable increase in publications. The chosen research questions reflect this objective. To answer the research questions the snowballing procedure starts with Mittelstadt's (2019) article and continues to identify additional literature using both backward and forward snowballing. The obtained literature is screened using eligibility criteria that determine if a publication is suitable to be included in the review.

The eligibility criteria define characteristics that a candidate publication must exhibit to be considered for inclusion. For a candidate to be included in the literature review it needs to fulfill all the inclusion criteria and none of the exclusion criteria. The eligibility criteria are depicted in Table 1. The literature review conducted in this thesis focuses on research published between 2019 and 2022. The cut-off point was chosen, since the publication of Mittelstadt's (2019) article was considered a landmark. The objective of this thesis was analyzing the prevalent themes and topics in current AI ethics discourse. Thus, only candidates that were published in this time period were considered relevant (I1). Furthermore, English was set as a language requirement for candidates (I2). As the purpose of this thesis was to analyze the literature for prevalent themes and discourses, full access to the publications was set as a requirement (I3). The focus of the analysis was on academic discourse related to the ethics of AI, which resulted in the exclusion of publications that were not either research articles or research papers (E1). Duplicates were removed in the final step (E2). These inclusion and exclusion criteria were applied iteratively to make the screening process more efficient.

Table 1 Eligibility Criteria

| **Inclusion Criteria** | **Exclusion Criteria** |
|---|---|
| [I1] Publication Year: 2019-2022 | [E1] Publication Type |
| [I2] Language: English | [E2] Duplicate |
| [I3] Full Access | |

## 4.2 Conducting the Review

Conducting the review consists of identifying relevant literature through a chosen search strategy. For this thesis the chosen search strategy employs the snowballing procedure. Snowballing as discussed by Wohlin (2014) refers to performing a literature review by using the references or citations of a publication to identify potential additional candidates for inclusion in the review. The procedure can be started once an appropriate start set is identified, which serves as the basis for the procedure. Several approaches can be utilized in forming the start set, for instance choosing a seminal paper in the research field or conducting a search in a general database like Google Scholar (Wohlin 2014). A suitable start set preferably contains publications that have the potential to encompass the area being researched. For this thesis the start set contains one publication, namely Mittelstadt's (2019) article. The article was chosen as it directly pertains to the ethics of AI, while simultaneously being highly cited and influential. For this thesis the citations were studied using Google Scholar, as it contained the most citations out of the Google Scholar, Semantic Scholar and Web of Science databases that were analyzed, as depicted in Table 2.

Table 2 Number of Citations in Each Database

| **Database** | **Number of Citations** |
|---|---:|
| Google Scholar | 384 |
| Semantic Scholar | 341 |
| Web of Science | 152 |

Once a start set has been established, the snowballing procedure can be undertaken. The procedure is depicted in Figure 1. During the snowballing procedure each reference and citation related to a publication is analyzed. Analyzing the references and citations, henceforth referred to as candidates, are called backward and forward snowballing respectively. The snowballing procedure consists of several steps, which are iterated until further candidates are no longer identified. In each iteration the evaluated candidates are screened for eligibility. The publications passing the eligibility criteria are then included in the literature review and further iterations of snowballing.

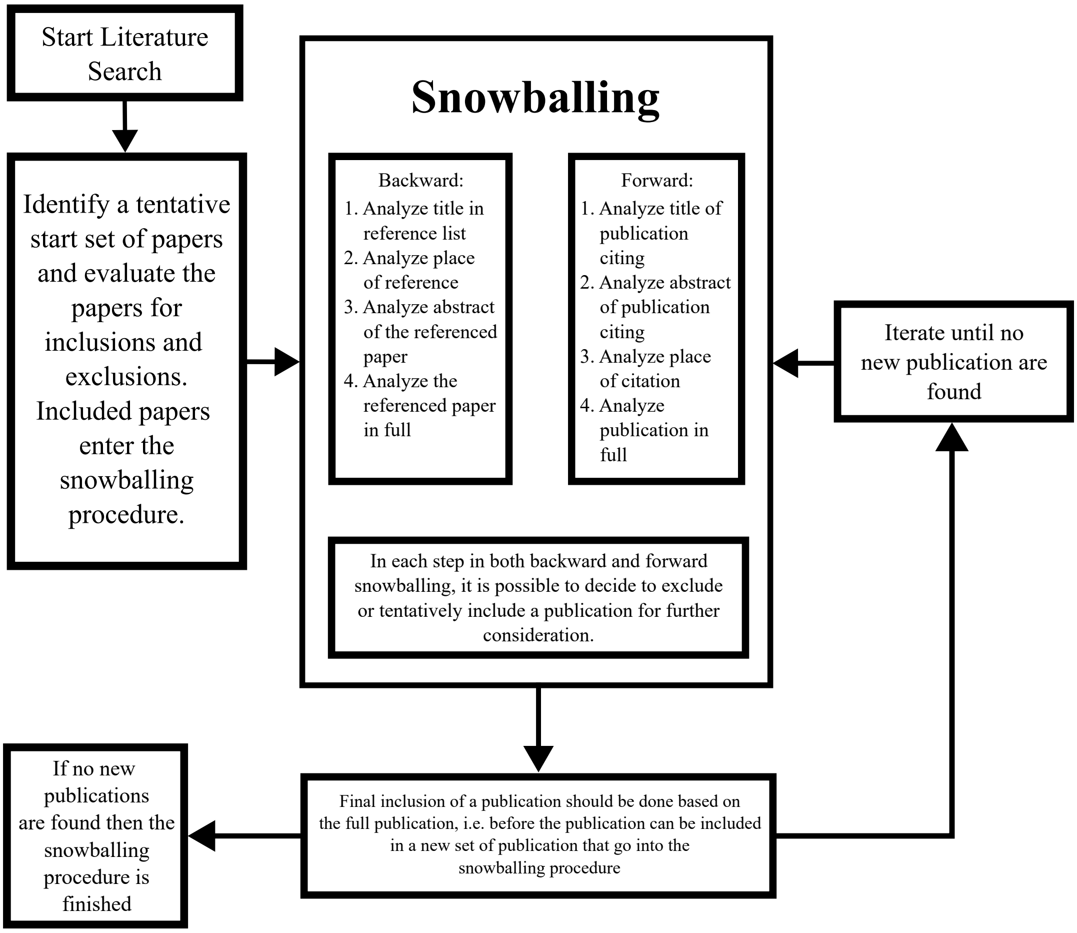

Start Literature Search
Identify a tentative start set of papers and evaluate the papers for inclusions and exclusions. Included papers enter the snowballing procedure.
Snowballing
Backward:
1. Analyze title in reference list
2. Analyze place of reference
3. Analyze abstract of the referenced paper
4. Analyze the referenced paper in full
Forward:
1. Analyze title of publication citing
2. Analyze abstract of publication citing
3. Analyze place of citation
4. Analyze publication in full
In each step in both backward and forward snowballing, it is possible to decide to exclude or tentatively include a publication for further consideration.
Iterate until no new publication are found
Final inclusion of a publication should be done based on the full publication, i.e. before the publication can be included in a new set of publication that go into the snowballing procedure
If no new publications are found then the snowballing procedure is finished

Figure 1 Snowballing Procedure

Backward snowballing refers to examining the reference list of a publication for further publications to include in the review and as candidates for the next iteration of snowballing. Each of these references should be screened using predefined eligibility criteria (Wohlin 2014). The first step in the procedure entails examining the reference list for publications that do not fulfill the inclusion criteria. For the first iteration of backward snowballing the references in Mittelstadt's (2019) article were analyzed for publications published before the cut-off point. After the publications outside of the time period were excluded, the next step was checking the references for publications not written in English. The remaining publications were checked for access in available databases. The accessible publications were screened for publication type, with research articles and research papers being included in the review. The results for the first iteration of backward snowballing are depicted in Table 3 and an in-depth documentation can be found in Appendix A.

Table 3 Excluded Papers in Backward Snowballing

| **Eligibility Criteria** | **Number of Excluded Publications** |
|---|---:|
| [I1] Publication Year: 2019-2022 | 72 |
| [E1] Publication Type | 6 |

Forward snowballing in turn refers to identifying new publications through citations to the original publication being examined. For the first iteration of forward snowballing each citation to Mittelstadt's (2019) article was examined starting with the information provided in Google Scholar. The publications in the database not written in English were excluded. The remaining publications were checked for access in available databases. The accessible publications were screened for publication type, with research articles and research papers being tentatively included in the review. Once all the citations were examined, the tentatively included publications were checked for duplicates. The results for the first iteration of forward snowballing are depicted in Table 4 and an in-depth documentation can be found in Appendix B.

Table 4 Excluded Papers in Forward Snowballing

| **Eligibility Criteria** | **Number of Excluded Publications** |
|---|---:|
| [I2] Language: English | 33 |
| [I3] Full Access | 35 |
| [E1] Publication Type | 34 |
| [E2] Duplicate | 16 |

Appendix A and B contain a detailed documentation of the publications excluded during the search. Table 5 contains the results of both backward and forward snowballing.

Table 5 Publications included

| Backward Snowballing | Forward Snowballing | Total |
|---:|---:|---:|
| 5 | 263 | 268 |

Table 6 contains the results of the screening with each eligibility criteria and its impact.

Table 6 Excluded Papers in Total

| Eligibility Criteria | Number of Excluded Publications |
|---|---:|
| [I1] Publication Year: 2019-2022 | 72 |
| [I2] Language: English | 33 |
| [I3] Full Access | 35 |
| [E1] Research Article or Research Paper | 40 |
| [E2] Duplicate | 16 |

## 4.3 Citation Mapping

The literature included in the review was rendered into a map of citations. The objective of this approach was identifying discourses through citations and references, by clustering connected articles together. The result of this procedure is a map, which provides a perspective of the research area with the prevalent discourses. This approach had the additional upside of revealing the interconnectivity and linkages between different discourses. The citation map was created by first linking the collected literature to the start set, resulting in the first version of the citation map seen in Figure 2.

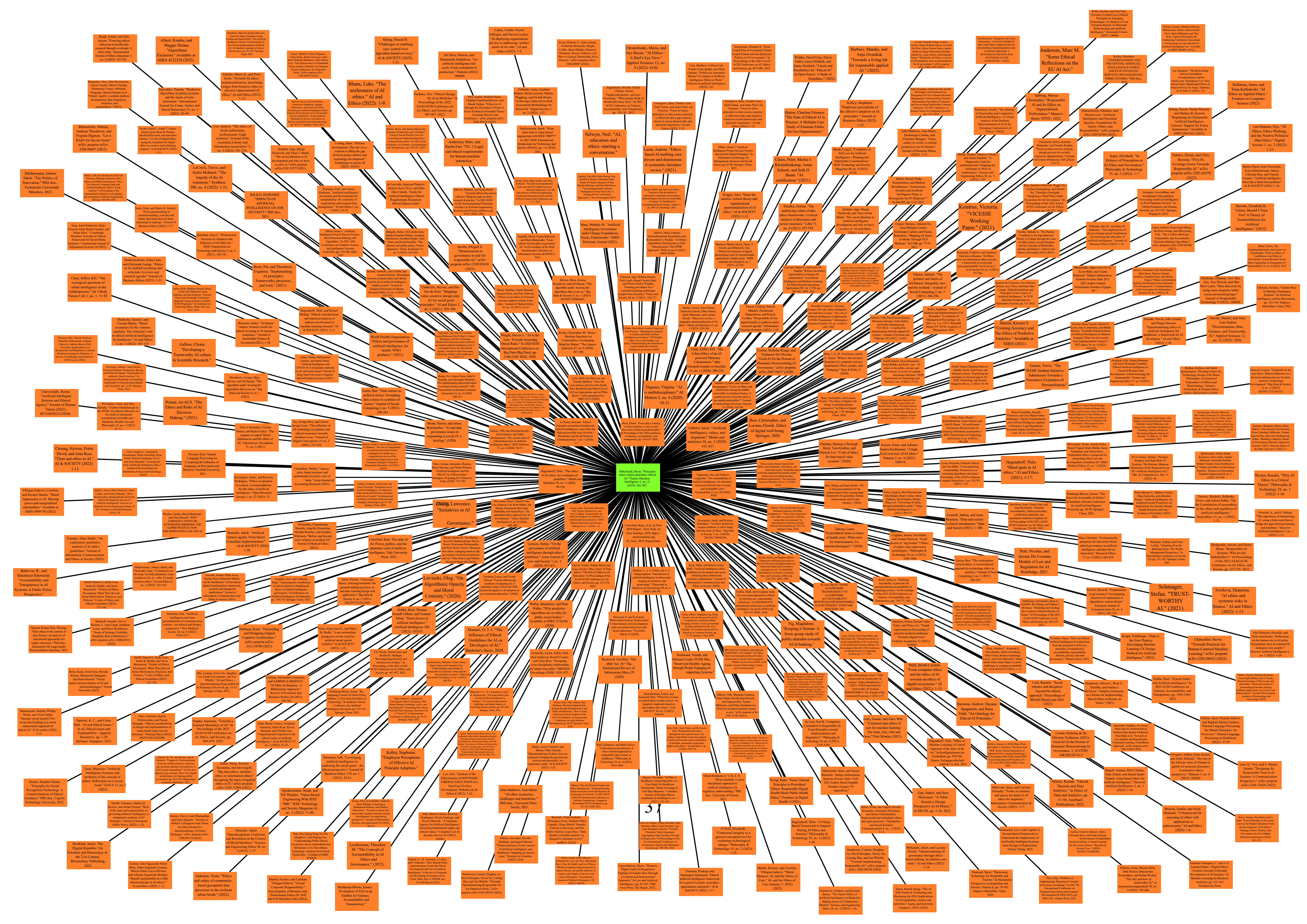
Mittelstadt, Brent. "Principles alone cannot guarantee ethical AI." Nature Machine Intelligence 1, no. 11 (2019): 501-507.
Bietti, Elettra. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy." In Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 210-219. 2020.
Crawford, Kate, et al. AI Now 2019 Report . New York: AI Now Institute, 2019, https://ainowinstitute.org/AI_Now_2019_Report.html
Sargent, Samantha. "AI Bias in Healthcare: Using ImpactPro as a Case Study for Healthcare Practitioners' Duties to Engage in Anti-Bias Measures." Canadian Journal of Bioethics/Revue canadienne de bioéthique 4, no. 1 (2021): 112-116.
Lera-Leri, Roger, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar. "Towards Pluralistic Value Alignment: Aggregating Value Systems Through ℓp-Regression." In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, pp. 780-788. 2022.
Albert, Kendra, and Maggie Delano. "Algorithmic Exclusion." Available at SSRN 4122529 (2022).
Munn, Luke. "The uselessness of AI ethics." AI and Ethics (2022): 1-9.
Selwyn, Neil. "AI, education and ethics–starting a conversation."
Anderson, Marc M. "Some Ethical Reflections on the EU AI Act."
Kontrus, Victoria. "VICESSE Working Paper." (2021).
Schmager, Stefan. "TRUST-WORTHY AI." (2021).
Livinski, Oleg. "On Algorithmic Opacity and Moral Certainty." (2020).
Dignum, Virginia. "AI is multidisciplinary." AI Matters 5, no. 4 (2020): 18-21.

Figure 2 Citation Mapping Phase 1

The second phase of mapping entailed linking articles through direct citation within the set of included articles in the review. This phase involved going through all the citations and references of the articles included in the review and documenting the instances of citation or reference to another article within the set of articles included in the review. Direct citation was chosen as the preferred approach, as articles outside of the set were not considered, rendering alternative methods impractical. Once all the instances of citation and reference were documented, these links were included in the map, resulting in the second version of the citation map seen in Figure 3.

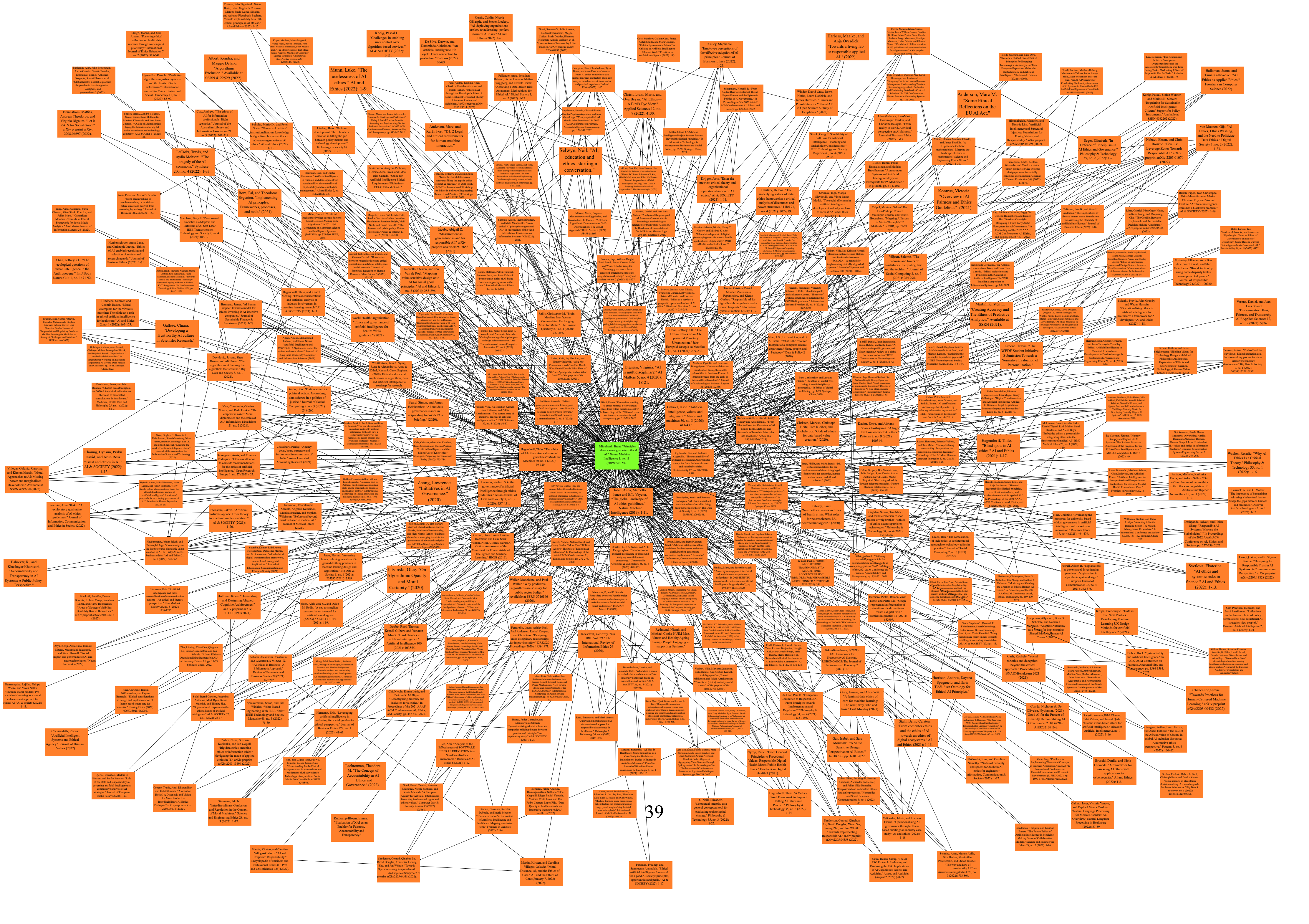

Mittelstadt, Brent. "Principles alone cannot guarantee ethical AI." Nature Machine Intelligence 1, no. 11 (2019): 501-507.
Jobin, Anna, Marcello Ienca and Effy Vayena. "The global landscape of AI ethics guidelines." Nature Machine Intelligence (2019): 1-11.
Gabriel, Iason. "Artificial intelligence, values, and alignment." Minds and machines 30, no. 3 (2020): 411-437.
Bietti, Elettra. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy." In Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 210-219. 2020.
Yigitcanlar, Tan, and Federico Cugurullo. "The sustainability of artificial intelligence: An urbanistic viewpoint from the lens of smart and sustainable cities." Sustainability 12, no. 20 (2020): 8548.
Amann, Julia, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective." BMC Medical Informatics and Decision Making 20, no. 1 (2020): 1-9.
Rességuier, Anaïs, and Rowena Rodrigues. "AI ethics should not remain toothless! A call to bring back the teeth of ethics." Big Data & Society 7, no. 2 (2020): 2053951720942541.
Dignum, Virginia. "AI is multidisciplinary." AI Matters 5, no. 4 (2020): 18-21.
Chan, Jeffrey KH. "The Urban Ethics of an AI-powered Planetary Urbanization." Jahr: Europski časopis za bioetiku 11, no. 1 (2020): 209-232.
Kontrus, Victoria. "Overview of AI Fairness and Ethics Guidelines" (2021).
Müller, Dennis, Maurice Chiodo, and James Franklin. "A Hippocratic Oath for mathematicians? Mapping the landscape of ethics in mathematics." Science and Engineering Ethics 28, no. 5 (2022): 1-30.
Raji, Inioluwa Deborah, Peggy Xu, Colleen Honigsberg, and Daniel Ho. "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance." In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 557-571. 2022.
Litvinski, Oleg. "On Algorithmic Opacity and Moral Certainty." (2020).
Hirsch, Dennis D., Tim Bartley, Aravind Chandrasekaran, Davon Norris, Srinivasan Parthasarathy, and Piers Norris Turner. "Business data ethics: emerging trends in the governance of advanced analytics and AI." Ohio State Legal Studies Research Paper 628 (2020).
Constantinescu, Mihaela, Cristina Voinea, Radu Uszkai, and Constantin Vică. "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context." Ethics and Information Technology 23, no. 4 (2021): 803-814.
Dobbe, Roel, Thomas Krendl Gilbert, and Yonatan Mintz. "Hard choices in artificial intelligence." Artificial Intelligence 300 (2021): 103555.
Boza, Pal, and Theodoros Evgeniou. "Implementing AI principles: Frameworks, processes, and tools." (2021).
Hermann, Erik, and Gunter Hermann. "Artificial intelligence in research and development for sustainability: the centrality of explicability and research data management." AI and Ethics 2, no. 1 (2022): 29-33.
Miller, Gloria J. "Artificial Intelligence Project Success Factors: Moral Decision-Making with Algorithms." In 2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 379-390. IEEE, 2021.
Munn, Luke. "The uselessness of AI ethics." AI and Ethics (2022): 1-9.
Selwyn, Neil. "AI, education and ethics–starting a conversation."
Anderson, Marc M. "Some Ethical Reflections on the EU AI Act."
Zhang, Lawrence. "Initiatives in AI Governance." (2020).
Hagendorff, Thilo. "Blind spots in AI ethics." AI and Ethics (2021): 1-17.
Harbers, Maaike, and Anja Overdiek. "Towards a living lab for responsible applied AI." (2022).
Lechterman, Theodore M. "The Concept of Accountability in AI Ethics and Governance." (2022).

Figure 3 Citation Mapping Phase 2

Once the direct linkages between articles were established, clusters with connected articles were formed. This resulted in several clusters of differing sizes, with some clusters being linked through interconnected articles. The formation of clusters was primarily impacted by the number of direct citations with bibliographic coupling serving a secondary role. The product of this process is the third and final version of the citation map seen in Figure 4.

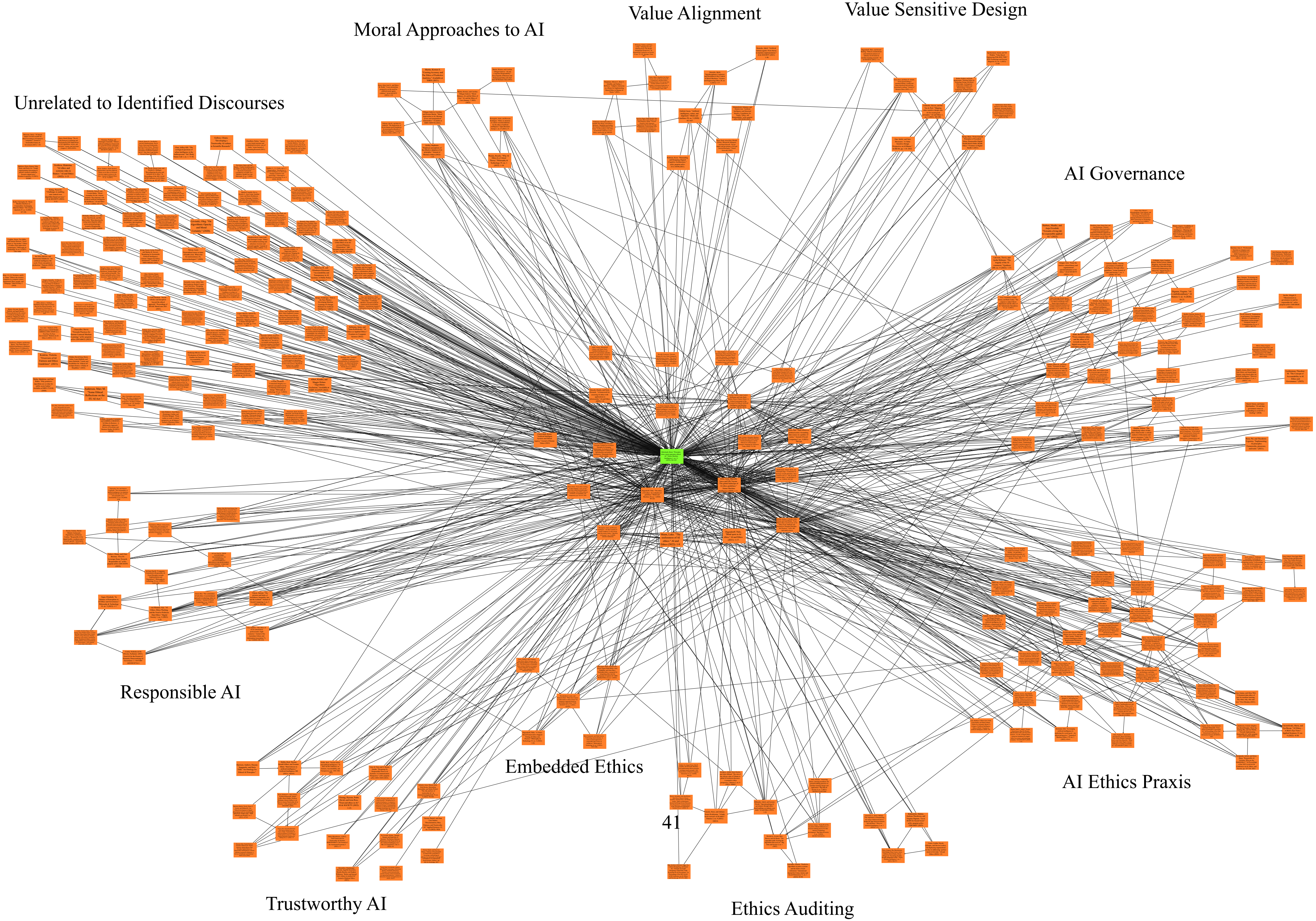

Moral Approaches to AI
Value Alignment
Value Sensitive Design
Unrelated to Identified Discourses
AI Governance
Responsible AI
Embedded Ethics
AI Ethics Praxis
Trustworthy AI
Ethics Auditing

Figure 4 Citation Mapping Phase 3

Important to note is that this process was interpretative by design, as the mapping was done by hand and an alternative clustering could justifiably have been created. The mapping was done by hand for the sake of transparency and repeatability. The use of automatic tools was considered and dismissed, since the documentation of the progression and design choices was considered critical. This approach was considered applicable, as the clusters, articles, and their connections were analyzed and documented in detail.

# 5 RESULTS

This chapter examines the pre-eminent discourses, topics and themes identified in the collected data. The chapter begins with a general overview of identified discourses and their relationships in section 5.1. Sections 5.2 through 5.11 examine a particular discourse and outline the articles that are the primary contributors towards the discourse as identified in the collected literature. Each section includes a figure representing the discourse. The figures contain the articles and citations forming the discourse. Citations and references to articles outside the discourse are omitted for clarity. The sections are constructed by providing a brief general overview of the themes and topics related to the discourse. The themes and topics are grounded by presenting the key articles forming the discourse, as identified in the collected data. Salient parts of the analyzed articles, that either indicate the relationships to other articles, or contribute to the article in a noteworthy way, are directly quoted. These direct quotations are marked by indentation and a citation at the end of the quoted paragraph. The direct quotations are quoted without editing.

## 5.1 AI Ethics Discourses

AI ethics has become a global topic of discussion containing varied themes with vast implications and impacts. Figure 4 and Figure 5 indicate the multifaceted nature of AI ethics containing a multitude of discourses, topics, and themes. Ten central discourses were identified in the collected literature encompassing thirty topics and fifty-eight themes, which are depicted in Figure 5. Additionally, after the completion of the analysis and clustering, a number of articles were identified, that didn’t relate to any of the identified discourses or form additional discourses. These articles are clustered together and labeled as unrelated to identified discourses in Figure 4 for clarity.

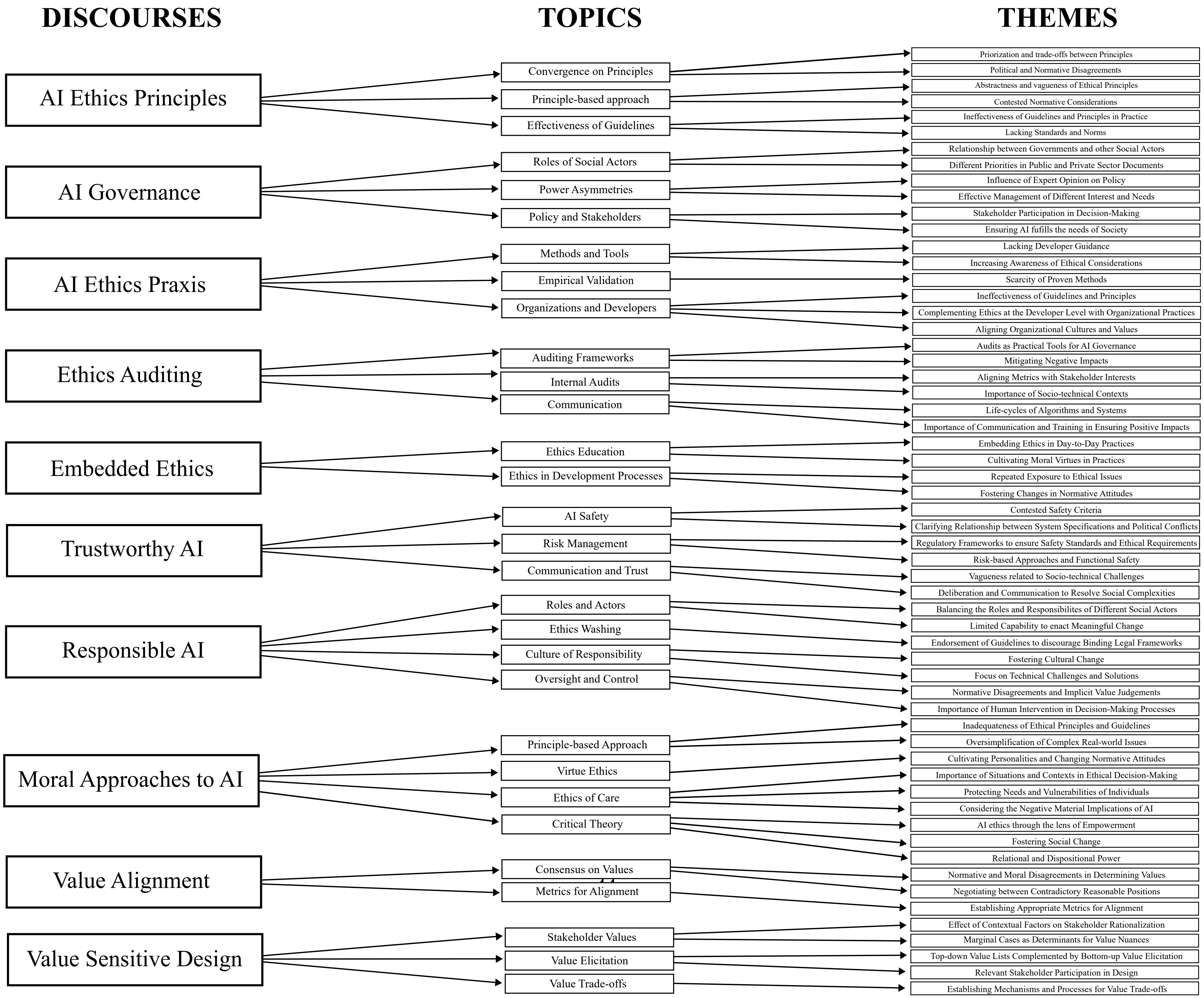

DISCOURSES
TOPICS
THEMES
AI Ethics Principles
Convergence on Principles
Principle-based approach
Effectiveness of Guidelines
AI Governance
Roles of Social Actors
Power Asymmetries
Policy and Stakeholders
AI Ethics Praxis
Methods and Tools
Empirical Validation
Organizations and Developers
Ethics Auditing
Auditing Frameworks
Internal Audits
Communication
Embedded Ethics
Ethics Education
Ethics in Development Processes
Trustworthy AI
AI Safety
Risk Management
Communication and Trust
Responsible AI
Roles and Actors
Ethics Washing
Culture of Responsibility
Oversight and Control
Moral Approaches to AI
Principle-based Approach
Virtue Ethics
Ethics of Care
Critical Theory
Value Alignment
Consensus on Values
Metrics for Alignment
Value Sensitive Design
Stakeholder Values
Value Elicitation
Value Trade-offs
Priorization and trade-offs between Principles
Political and Normative Disagreements
Abstractness and vagueness of Ethical Principles
Contested Normative Considerations
Ineffectiveness of Guidelines and Principles in Practice
Lacking Standards and Norms
Relationship between Governments and other Social Actors
Different Priorities in Public and Private Sector Documents
Influence of Expert Opinion on Policy
Effective Management of Different Interest and Needs
Stakeholder Participation in Decision-Making
Ensuring AI fufills the needs of Society
Lacking Developer Guidance
Increasing Awareness of Ethical Considerations
Scarcity of Proven Methods
Ineffectiveness of Guidelines and Principles
Complementing Ethics at the Developer Level with Organizational Practices
Aligning Organizational Cultures and Values
Audits as Practical Tools for AI Governance
Mitigating Negative Impacts
Aligning Metrics with Stakeholder Interests
Importance of Socio-technical Contexts
Life-cycles of Algorithms and Systems
Importance of Communication and Training in Ensuring Positive Impacts
Embedding Ethics in Day-to-Day Practices
Cultivating Moral Virtues in Practices
Repeated Exposure to Ethical Issues
Fostering Changes in Normative Attitudes
Contested Safety Criteria
Clarifying Relationship between System Specifications and Political Conflicts
Regulatory Frameworks to ensure Safety Standards and Ethical Requirements
Risk-based Approaches and Functional Safety
Vagueness related to Socio-technical Challenges
Deliberation and Communication to Resolve Social Complexities
Balancing the Roles and Responsibilites of Different Social Actors
Limited Capability to enact Meaningful Change
Endorsement of Guidelines to discourage Binding Legal Frameworks
Fostering Cultural Change
Focus on Technical Challenges and Solutions
Normative Disagreements and Implicit Value Judgements
Importance of Human Intervention in Decision-Making Processes
Inadequateness of Ethical Principles and Guidelines
Oversimplification of Complex Real-world Issues
Cultivating Personalities and Changing Normative Attitudes
Importance of Situations and Contexts in Ethical Decision-Making
Protecting Needs and Vulnerabilities of Individuals
Considering the Negative Material Implications of AI
AI ethics through the lens of Empowerment
Fostering Social Change
Relational and Dispositional Power
Normative and Moral Disagreements in Determining Values
Negotiating between Contradictory Reasonable Positions
Establishing Appropriate Metrics for Alignment
Effect of Contextual Factors on Stakeholder Rationalization
Marginal Cases as Determinants for Value Nuances
Top-down Value Lists Complemented by Bottom-up Value Elicitation
Relevant Stakeholder Participation in Design
Establishing Mechanisms and Processes for Value Trade-offs

Figure 5 AI Ethics Discourses, Topics and Themes

The primary discourses identified in the collected data relate to: *AI ethics principles, AI governance, AI ethics praxis, ethics auditing, embedded ethics, trustworthy AI, Responsible AI, moral approaches to AI, AI value alignment, and value sensitive design*.

The overarching substructure underlying most discourses identified in the collected literature concerns acknowledging or critiquing an approach towards AI ethics based on ethical principles. The principle-based approach has been the predominant approach towards ethical AI (Villegas-Galaviz and Martin 2022a). Thus, it has received its fair share of acknowledgement and criticism, with subsequent approaches positioning themselves in support or opposed to principles as guiding mechanisms for ethical AI (Munn 2022). This dichotomy can be seen permeating throughout the discourses in the collected data.

Discourse pertaining to *AI ethics principles* concerns examining and evaluating different guidelines and establishing a consensus on agreed upon principles. Additionally, the effectiveness of guidelines and principles in guiding AI development towards ethical practices is a primary topic of discussion. The discourse contains both articles in support of guidelines and principles as guiding mechanisms for ethical AI, as well as articles arguing for their uselessness in affecting meaningful change. The *AI ethics principles* discourse is examined in detail in section 5.2.

*AI governance*, as discussed in the collected data, focuses on the different roles of social actors in the effective governance of AI. Additionally, *AI governance* comprises the mechanisms, such as regulatory and technical standards, in guiding AI development towards a socially purposeful and beneficial direction. The *AI governance* discourse is concerned with the emancipation of different stakeholders, acknowledging prevailing power asymmetries, and promoting opportunities for different social actors. In the collected data the discourse on *AI governance* is interconnected with multiple other discourses related thematically to the different roles and responsibilities in ethical AI development. The scope of governance encompasses both the regulatory policies, as well as technical standards established by organizations, thereby having far-reaching implications towards other discourses. The importance of *AI governance* can be seen in the substantial number of articles contained in the discourse. The *AI governance* discourse is examined in detail in section 5.3.

Producing actionable tools and methods for ethical AI has been challenging (Aastha, et al. 2022). Adopting ethical principles and guidelines in design practice has proved to be difficult for developers, as the necessary tools for translating these innumerable abstract notions of ethicality into practical instructions have been lacking (Ryan and Stahl 2021). This rift between principles and practice has caused a shift towards the production of concrete tools for AI practitioners (Georgieva, et al. 2022). This movement towards practical tools can also be seen in the data collected for this thesis, as the discourse pertaining to *AI ethics praxis* contains the largest number of articles. Contained within this discourse are several tools designed to incorporate ethical principles into design processes, as well as research aimed at empirically validating the effectiveness of these tools. Additionally, discourse related to organizational practices and aligning organizational cultures and values is a central topic of interest. The *AI ethics praxis* discourse is examined in detail in section 5.4.

In the collected literature, the discourse related to *AI ethics praxis* is connected to several other discourses. This interdiscursivity is most prominently seen in the discourse related to *ethics auditing*. *Ethics auditing*, as discussed in the gathered data, concerns assessing and evaluating AI systems and algorithms for ethical concerns, especially negative impacts. Audits are proposed as a practical mechanism and actionable tool in evaluating AI systems. In the literature audits are discussed as both an internal tool for stakeholders within organization, as well as a tool for external stakeholders in assessing AI systems. The discourse pertaining to auditing contains several proposed frameworks, in addition to debate about the roles of different stakeholders in auditing. Additionally, discourse about the role of communication in the effective adoption of auditing is a primary issue. The *ethics auditing* discourse is examined in detail in section 5.5.

The discourse related to *embedded ethics* provides a different perspective towards operationalizing AI ethics, serving as a complementary approach with respect to developer tools and ethical auditing. As discussed in the collected literature, *embedded ethics* refers to both the incorporation of ethics into technology and data science education, as well as the integration of ethics into the design and operation of AI systems. In the latter sense, *embedded ethics* resembles the aspirations seen in the developer-oriented praxis literature. Thus, certain overlap in themes and interdiscursivity can be noticed between the *AI ethics praxis* and *embedded ethics* discourses. In comparison to praxis and auditing literatures, *embedded ethics* comprises of only a handful of articles, representing the smallest identified discourse. The *embedded ethics* discourse is examined in detail in section 5.6.

Trustworthiness, safety, and responsibility are seen as critical issues in AI ethics discourses, as substantiated by the collected data. The overarching point of concern in these discourses is the need for stakeholder participation in deliberation and decision-making processes, in addition to the explication of varied normative implications and value judgements included in the design and implementation of AI systems. The critical points of contestation comprise the different roles of social actors in risk management, as well as the relative importance of regulation, policy, and guidelines in managing risks and impacts of AI systems. The difficulties, as discussed in the collected literature, are compounded by socio-technical challenges stemming from the wide-reaching impacts and interconnected nature of AI systems. Thus, the discourses related to *trustworthy AI* and *responsible AI* share common themes and issues with the *AI governance* discourse. The *trustworthy AI* discourse specifically focuses on AI safety risks in complex social contexts and the role of communication in building trust. The *trustworthy AI* discourse is examined in detail in section 5.7. The *responsible AI* discourse is focused on establishing consensus on contested normative attitudes and navigating value disagreements. The discourse on *responsible AI* is focused on establishing common ground and shared goals, despite containing varied topics. The discourse on *responsible AI* is examined in detail in section 5.8.

Ethical principles and guidelines have been criticized as insufficient guiding mechanisms for ethical AI (Mittelstadt 2019). Thus, the principle-based approach has received its fair share of criticism. In the discourse on *moral approaches to AI*, as seen in the collected literature, alternative approaches towards ethical AI are proposed and discussed. The prominent proposals advocate for either a virtue-oriented approach, or an ethics of care approach, as these can represent divergent ethical facets, yet can be complementary to a principle-based approach. Exemplified especially in the ethics of care approach is the recommendation of stakeholder participation and consideration of stakeholder values (Villegas-Galaviz and Martin 2022b). A similar necessity is propounded in the collected literature on *trustworthy AI* and *responsible AI.* The discourse on *moral approaches to AI* is examined in detail in section 5.9.

The discourses regarding *value alignment* and *value sensitive design* in the collected literature form distinct discourses, that explore similar topics in considering conflicting value systems and value trade-offs. From the alignment perspective designing AI systems that adhere to human values is portrayed as a complicated process, requiring the establishment of consensus building frameworks and moral exemplars (Lera-Leri, et al. 2022). The *value sensitive design* discourse approaches values from a more pragmatic perspective. From this perspective, the elicitation of relevant stakeholder values and translating these into design requirements are regarded as pivotal issues, as argued by Gan and Moussawi (2022). Despite being thematically closely related, the discourses as seen in the collected data form distinct clusters and are not connected through references or citations. Therefore, the discourses are separated and analyzed in their respective sections. The AI *value alignment* discourse is examined in detail in section 5.10. The *value sensitive design* discourse is examined in detail in section 5.11.

## 5.2 AI Ethics Principles

There has been a proliferation of principles and guidelines for ethical AI in the past decade. The *AI ethics principles* discourse is concerned with the ethical development, deployment, and use of artificial intelligence, that respects human values, minimizes harm, and promotes human well-being. This interest in principles and guidelines can be seen in the collected data, which indicates that the discourse is prominent and ongoing. The visualization of the principles discourse, as seen in the collected data, is depicted in Figure 6. As Figure 4 and Figure 5 indicate, the discourse on principles forms the substructure upon which the other discourses are built upon, containing the largest amount of references and citations, and a considerable amount of articles. The visualization, depicted in Figure 6, shows the interconnectedness and strong referentiality of the *AI ethics principles* discourse.

In the collected literature, the *AI ethics principles* discourse contains articles examining ethics principles and guidelines. However, a critical undertone can be seen permeating through part of the *AI ethics principles* discourse, seen especially in the recent articles related to AI ethics principles. This critical stance towards ethical principles and guidelines as guiding mechanisms towards ethical AI is exemplified in the critique by Munn (2022).

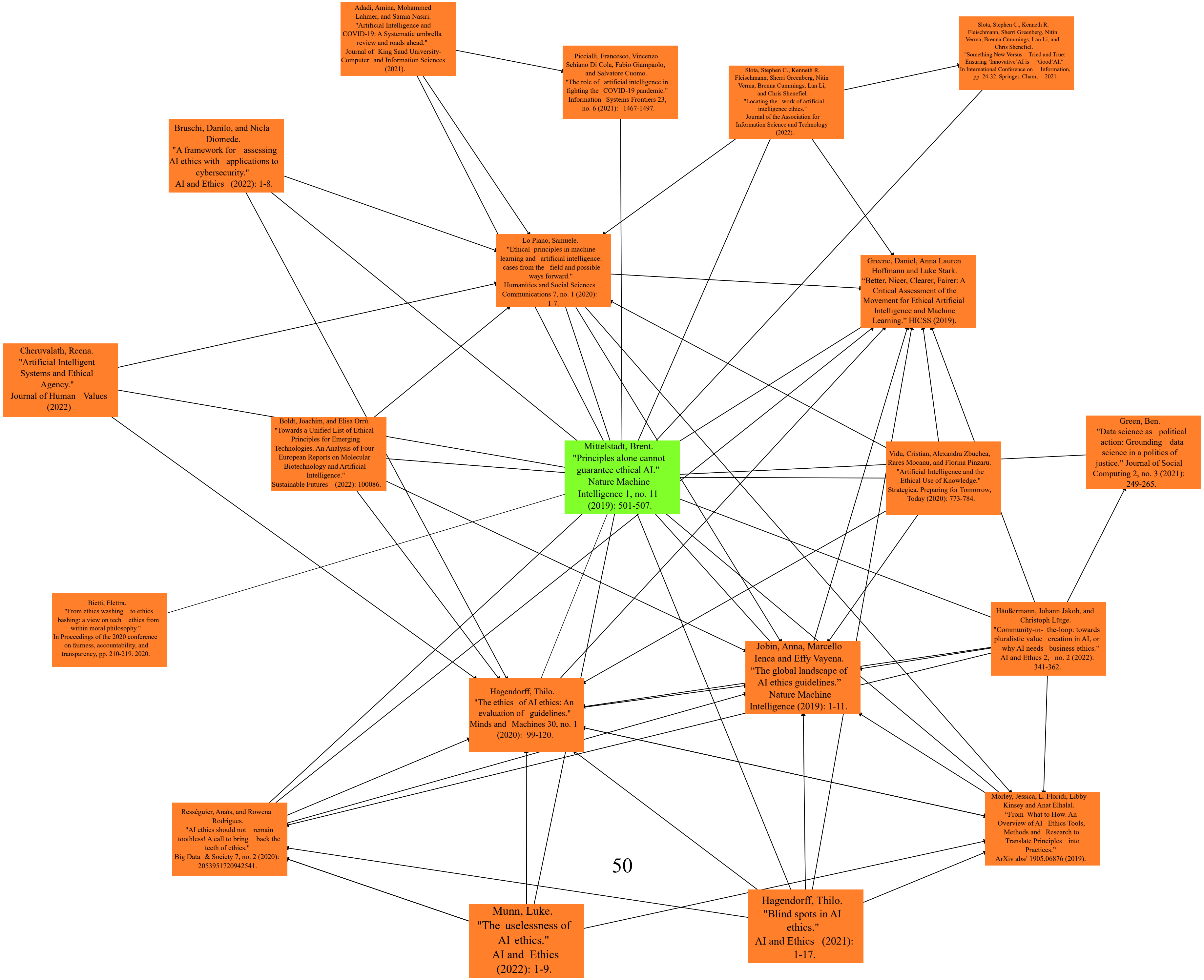

Adadi, Amina, Mohammed Lahmer, and Samia Nasiri. "Artificial Intelligence and COVID-19: A Systematic umbrella review and roads ahead." Journal of King Saud University-Computer and Information Sciences (2021).
Piccialli, Francesco, Vincenzo Schiano Di Cola, Fabio Giampaolo, and Salvatore Cuomo. "The role of artificial intelligence in fighting the COVID-19 pandemic." Information Systems Frontiers 23, no. 6 (2021): 1467-1497.
Slota, Stephen C., Kenneth R. Fleischmann, Sherri Greenberg, Nitin Verma, Brenna Cummings, Lan Li, and Chris Shenefiel. "Something New Versus Tried and True: Ensuring 'Innovative'AI is 'Good'AI." In International Conference on Information, pp. 24-32. Springer, Cham, 2021.
Slota, Stephen C., Kenneth R. Fleischmann, Sherri Greenberg, Nitin Verma, Brenna Cummings, Lan Li, and Chris Shenefiel. "Locating the work of artificial intelligence ethics." Journal of the Association for Information Science and Technology (2022).
Bruschi, Danilo, and Nicla Diomede. "A framework for assessing AI ethics with applications to cybersecurity." AI and Ethics (2022): 1-8.
Lo Piano, Samuele. "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward." Humanities and Social Sciences Communications 7, no. 1 (2020): 1-7.
Greene, Daniel, Anna Lauren Hoffmann and Luke Stark. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." HICSS (2019).
Cheruvalath, Reena. "Artificial Intelligent Systems and Ethical Agency." Journal of Human Values (2022)
Boldt, Joachim, and Elisa Orrù. "Towards a Unified List of Ethical Principles for Emerging Technologies. An Analysis of Four European Reports on Molecular Biotechnology and Artificial Intelligence." Sustainable Futures (2022): 100086.
Mittelstadt, Brent. "Principles alone cannot guarantee ethical AI." Nature Machine Intelligence 1, no. 11 (2019): 501-507.
Vidu, Cristian, Alexandra Zbuchea, Rares Mocanu, and Florina Pinzaru. "Artificial Intelligence and the Ethical Use of Knowledge." Strategica. Preparing for Tomorrow, Today (2020): 773-784.
Green, Ben. "Data science as political action: Grounding data science in a politics of justice." Journal of Social Computing 2, no. 3 (2021): 249-265.
Bietti, Elettra. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy." In Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 210-219. 2020.
Häußermann, Johann Jakob, and Christoph Lütge. "Community-in- the-loop: towards pluralistic value creation in AI, or —why AI needs business ethics." AI and Ethics 2, no. 2 (2022): 341-362.
Jobin, Anna, Marcello Ienca and Effy Vayena. "The global landscape of AI ethics guidelines." Nature Machine Intelligence (2019): 1-11.
Hagendorff, Thilo. "The ethics of AI ethics: An evaluation of guidelines." Minds and Machines 30, no. 1 (2020): 99-120.
Morley, Jessica, L. Floridi, Libby Kinsey and Anat Elhalal. "From What to How. An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices." ArXiv abs/ 1905.06876 (2019).
Rességuier, Anaïs, and Rowena Rodrigues. "AI ethics should not remain toothless! A call to bring back the teeth of ethics." Big Data & Society 7, no. 2 (2020): 2053951720942541.
Munn, Luke. "The uselessness of AI ethics." AI and Ethics (2022): 1-9.
Hagendorff, Thilo. "Blind spots in AI ethics." AI and Ethics (2021): 1-17.

Figure 6 AI Ethics Principles Discourse

The discourse on *AI ethics principles* is built on the objective of ensuring ethical AI development and governance. At the forefront of this project, private companies, research institutions, and public sector organizations have set out different visions for the future of AI. The discourse on *AI ethics principles*, as seen in the collected data, is largely grounded in the mapping and analysis of ethical AI principles and guidelines by Jobin, Ienca and Vayena (2019). According to Jobin, Ienca and Vayena (2019) emerging within the corpus of guidelines and principles is a convergence around five ethical principles: transparency, justice and fairness, non-maleficence, responsibility, and privacy. Hagendorff (2022b) however argues, that the consensus seen in the different guidelines and codes of ethics is due to their referential and derivative nature, where previous guidelines guide the formation and composition of later guidelines, thereby resulting in the creation of mere echoes. This discussion on the effectiveness of principles and guidelines in ensuring ethical AI development is seen throughout the topics in the discourse on *AI ethics principles*.

Mittelstadt (2019) notes, contained within this seeming consensus on principles are a numerous amount of political and normative disagreements. This sentiment is shared by Jobin, Ienca and Vayena (2019).

> Nonetheless, our thematic analysis reveals substantive divergences in relation to four major factors: (i) how ethical principles are interpreted, (ii) why they are deemed important, (iii) what issue, domain or actors they pertain to, and (iv) how they should be implemented. Furthermore, unclarity remains as to which ethical principles should be prioritized, how conflicts between ethical principles should be resolved, who should enforce ethical oversight on AI and how researchers and institutions can comply with the resulting guidelines. (Jobin, Ienca and Vayena 2019)

Mittelstadt (2019) argues, that a principle-based approach to AI ethics is problematic, as it is too abstract to be action-guiding. From an AI practitioner's perspective ambiguous notions of ethical design present a challenge (Mittelstadt 2019). Furthermore, the effectiveness of reading ethical guidelines and its effect towards the decision-making of software developers has been criticized (Hagendorff 2020). As Munn (2022) argues, ethical principles are generally contested verging on incoherence, making them difficult to apply. Mittelstadt (2019) concurs, noting that adopting a principle-based approach for AI ethics appears to embed contested normative considerations into the design and governance of AI. Thus, ethical principles fail to mitigate the negative impacts of AI in any meaningful sense. According to Munn (2022), this also applies to the project of operationalization, as the translation of complex social concepts to technical requirements is non-trivial.

Mittelstadt (2019) further highlights the conflicts of interest in AI ethics, as the purposes of developers and users misalign, thereby transforming ethical decision-making into a competitive process, instead of a cooperative one. This has resulted in both ethics-washing and ethics-bashing, as described by Bietti (2020).

> Weaponized in support of deregulation, self-regulation or hands-off governance, "ethics" is increasingly identified with technology companies' self-regulatory efforts and with shallow appearances of ethical behavior. So-called "ethics washing" by tech companies is on the rise, prompting criticism and scrutiny from scholars and the tech community at large. In parallel to the growth of ethics washing, its condemnation has led to a tendency to engage in "ethics bashing." This consists in the trivialization of ethics and moral philosophy now understood as discrete tools or pre-formed social structures such as ethics boards, self-governance schemes or stakeholder groups. (Bietti 2020)

These attempts at demarcating AI ethics are exacerbated by the lack of fiduciary duties and commitments to uphold public interest in AI development (Mittelstadt 2019). As Mittelstadt (2019) highlights, AI ethics currently lacks standards and norms guiding ethical behavior. Thus, the principle-based approach dominating AI ethics is prone to manipulation by industry actors attempting to avert regulation (Rességuier and Rodrigues 2020). These issues are compounded by the currently prominent endeavor towards technological solutionism, where the ethics in AI are addressed in terms of technical and design expertise (Greene, Hoffmann and Stark 2019). These concerns are shared by Häußermann and Lütge (2022), who additionally highlight the risks of overlooking important ethical considerations, by adhering to the tendency towards technical fixes in AI ethics.

Green (2021a) on the other hand criticizes the focus on vague moral principles for their lack of applicability in practice.

> First, technology ethics principles are abstract and lack mechanisms to ensure that engineers follow ethical principles. Second, technology ethics has a myopic focus on individual engineers and on

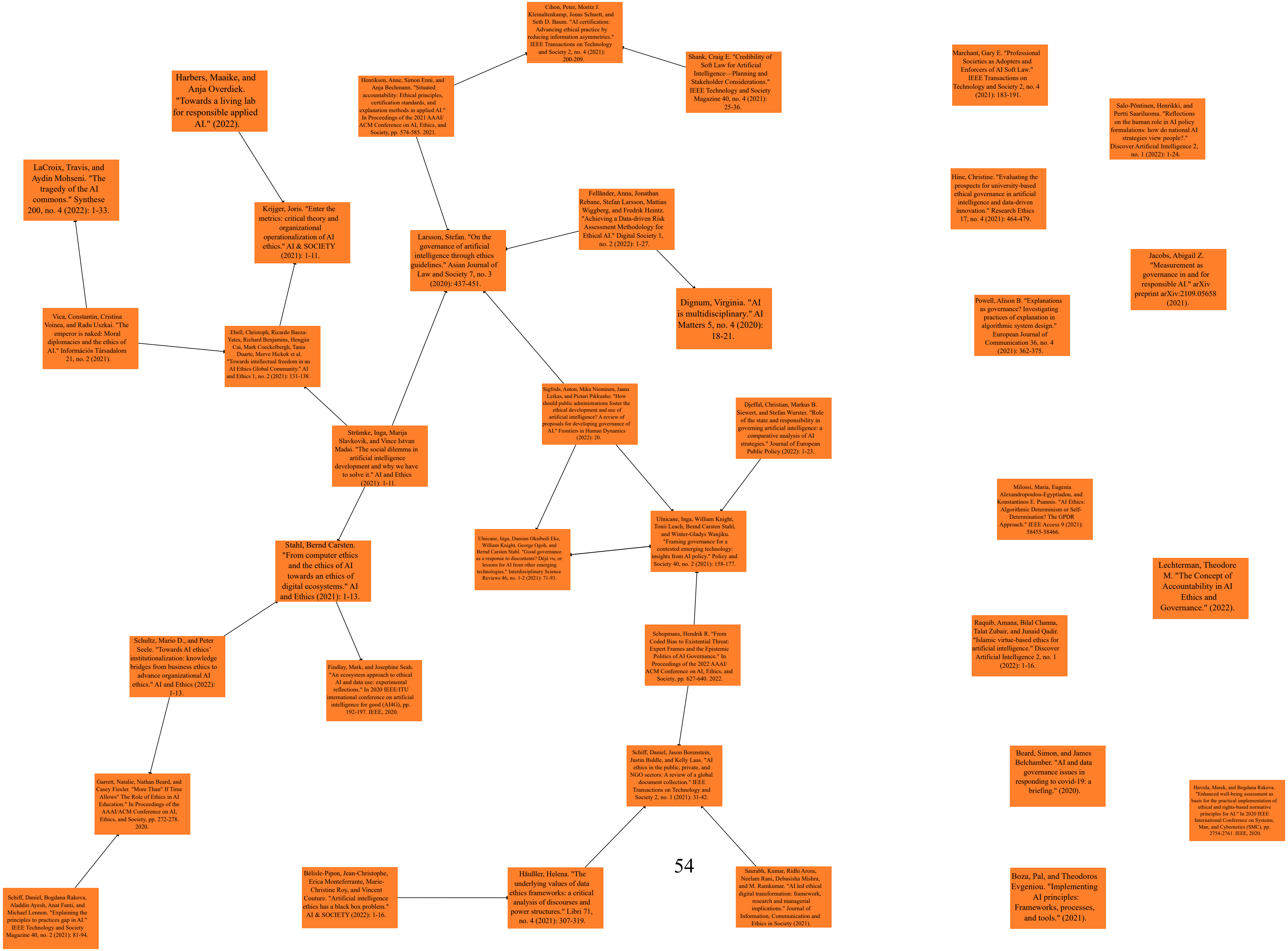
> technology design, overlooking the structural sources of technological harms. Third, technology ethics is subsumed into corporate logics and practices rather than substantively altering behavior. (Green, Data science as political action: Grounding data science in a politics of justice 2021a)

Green (2021a) advocates for a focus on engagement with competing perspectives, values, and goals. The gap between principles and practice is considerable. Unless mechanisms that facilitate practical guidance are developed, principles by themselves risk accentuating the costs of ethical mistakes, thereby outweighing the benefits of ethical successes (Morley, Floridi, et al. 2020). Such failures can lead to the undermining of public trust and acceptance of AI systems, resulting in reduced adoption and missed opportunities (Morley, Floridi, et al. 2020).

## 5.3 AI Governance

*AI governance* is a wide-reaching multidimensional issue, as the dispersity in the visualization of the governance discourse, depicted in Figure 7, indicates. Additionally, this multidimensionality can be seen in the number of articles and represented perspectives seen in the discourse. As seen in Figure 7, the discourse on *AI governance* is fragmented, encompassing topics from organizational roles in *AI governance* to empowering a larger share of stakeholders in managing the impact of AI technologies. A central connecting theme in the discourse surrounding *AI governance* is focused on the roles of different social actors in the effective governance of AI.

Navigating the different roles and responsibilities of governments, industry actors and organizations in *AI governance* is a challenging issue. The multifaceted nature of different paradigms and approaches towards effective governance is exemplified by the shallow referentiality and interconnectedness of the *AI governance* discourse. Thus, certain competitiveness and incompatibility in interests and needs between public and private sector entities needs to be overcome for the sake of reaching effective *AI governance* (Vica, Voinea and Uszkai 2021).



Cihon, Peter, Moritz J. Kleinaltenkamp, Jonas Schuett, and Seth D. Baum. "AI certification: Advancing ethical practice by reducing information asymmetries." IEEE Transactions on Technology and Society 2, no. 4 (2021): 200-209.

Shank, Craig E. "Credibility of Soft Law for Artificial Intelligence—Planning and Stakeholder Considerations." IEEE Technology and Society Magazine 40, no. 4 (2021): 25-36.

Henriksen, Anne, Simon Enni, and Anja Bechmann. "Situated accountability: Ethical principles, certification standards, and explanation methods in applied AI." In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 574-585. 2021.

Harbers, Maaike, and Anja Overdiek. "Towards a living lab for responsible applied AI." (2022).

Marchant, Gary E. "Professional Societies as Adopters and Enforcers of AI Soft Law." IEEE Transactions on Technology and Society 2, no. 4 (2021): 183-191.

Salo-Pöntinen, Henrikki, and Pertti Saariluoma. "Reflections on the human role in AI policy formulations: how do national AI strategies view people?." Discover Artificial Intelligence 2, no. 1 (2022): 1-24.

LaCroix, Travis, and Aydin Mohseni. "The tragedy of the AI commons." Synthese 200, no. 4 (2022): 1-33.

Krijger, Joris. "Enter the metrics: critical theory and organizational operationalization of AI ethics." AI & SOCIETY (2021): 1-11.

Larsson, Stefan. "On the governance of artificial intelligence through ethics guidelines." Asian Journal of Law and Society 7, no. 3 (2020): 437-451.

Felländer, Anna, Jonathan Rebane, Stefan Larsson, Mattias Wiggberg, and Fredrik Heintz. "Achieving a Data-driven Risk Assessment Methodology for Ethical AI." Digital Society 1, no. 2 (2022): 1-27.

Hine, Christine. "Evaluating the prospects for university-based ethical governance in artificial intelligence and data-driven innovation." Research Ethics 17, no. 4 (2021): 464-479.

Jacobs, Abigail Z. "Measurement as governance in and for responsible AI." arXiv preprint arXiv:2109.05658 (2021).

Dignum, Virginia. "AI is multidisciplinary." AI Matters 5, no. 4 (2020): 18-21.

Powell, Alison B. "Explanations as governance? Investigating practices of explanation in algorithmic system design." European Journal of Communication 36, no. 4 (2021): 362-375.

Vica, Constantin, Cristina Voinea, and Radu Uszkai. "The emperor is naked: Moral diplomacies and the ethics of AI." Információs Társadalom 21, no. 2 (2021).

Ebell, Christoph, Ricardo Baeza-Yates, Richard Benjamins, Hengjin Cai, Mark Coeckelbergh, Tania Duarte, Merve Hickok et al. "Towards intellectual freedom in an AI Ethics Global Community." AI and Ethics 1, no. 2 (2021): 131-138.

Sigfrids, Anton, Mika Nieminen, Jaana Leikas, and Pietari Pikkuaho. "How should public administrations foster the ethical development and use of artificial intelligence? A review of proposals for developing governance of AI." Frontiers in Human Dynamics (2022): 20.

Djeffal, Christian, Markus B. Siewert, and Stefan Wurster. "Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies." Journal of European Public Policy (2022): 1-23.

Strümke, Inga, Marija Slavkovik, and Vince Istvan Madai. "The social dilemma in artificial intelligence development and why we have to solve it." AI and Ethics (2021): 1-11.

Milossi, Maria, Eugenia Alexandropoulou-Egyptiadou, and Konstantinos E. Psannis. "AI Ethics: Algorithmic Determinism or Self-Determination? The GPDR Approach." IEEE Access 9 (2021): 58455-58466.

Ulnicane, Inga, William Knight, Tonii Leach, Bernd Carsten Stahl, and Winter-Gladys Wanjiku. "Framing governance for a contested emerging technology: insights from AI policy." Policy and Society 40, no. 2 (2021): 158-177.

Ulnicane, Inga, Damian Okaibedi Eke, William Knight, George Ogoh, and Bernd Carsten Stahl. "Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies." Interdisciplinary Science Reviews 46, no. 1-2 (2021): 71-93.

Stahl, Bernd Carsten. "From computer ethics and the ethics of AI towards an ethics of digital ecosystems." AI and Ethics (2021): 1-13.

Lechterman, Theodore M. "The Concept of Accountability in AI Ethics and Governance." (2022).

Raquib, Amana, Bilal Channa, Talat Zubair, and Junaid Qadir. "Islamic virtue-based ethics for artificial intelligence." Discover Artificial Intelligence 2, no. 1 (2022): 1-16.

Schopmans, Hendrik R. "From Coded Bias to Existential Threat: Expert Frames and the Epistemic Politics of AI Governance." In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 627-640. 2022.

Schultz, Mario D., and Peter Seele. "Towards AI ethics' institutionalization: knowledge bridges from business ethics to advance organizational AI ethics." AI and Ethics (2022): 1-13.

Findlay, Mark, and Josephine Seah. "An ecosystem approach to ethical AI and data use: experimental reflections." In 2020 IEEE/ITU international conference on artificial intelligence for good (AI4G), pp. 192-197. IEEE, 2020.

Schiff, Daniel, Jason Borenstein, Justin Biddle, and Kelly Laas. "AI ethics in the public, private, and NGO sectors: A review of a global document collection." IEEE Transactions on Technology and Society 2, no. 1 (2021): 31-42.

Beard, Simon, and James Belchamber. "AI and data governance issues in responding to covid-19: a briefing." (2020).

Garrett, Natalie, Nathan Beard, and Casey Fiesler. "More Than" If Time Allows" The Role of Ethics in AI Education." In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 272-278. 2020.

Havrda, Marek, and Bogdana Rakova. "Enhanced well-being assessment as basis for the practical implementation of ethical and rights-based normative principles for AI." In 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2754-2761. IEEE, 2020.

Bélisle-Pipon, Jean-Christophe, Erica Monteferrante, Marie-Christine Roy, and Vincent Couture. "Artificial intelligence ethics has a black box problem." AI & SOCIETY (2022): 1-16.

Häußler, Helena. "The underlying values of data ethics frameworks: a critical analysis of discourses and power structures." Libri 71, no. 4 (2021): 307-319.

Saurabh, Kumar, Ridhi Arora, Neelam Rani, Debasisha Mishra, and M. Ramkumar. "AI led ethical digital transformation: framework, research and managerial implications." Journal of Information, Communication and Ethics in Society (2021).

Boza, Pal, and Theodoros Evgeniou. "Implementing AI principles: Frameworks, processes, and tools." (2021).

Schiff, Daniel, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. "Explaining the principles to practices gap in AI." IEEE Technology and Society Magazine 40, no. 2 (2021): 81-94.

Figure 7 AI Governance Discourse

Governance of AI is a frequently mentioned issue in AI policy documents and guidelines as a mechanism that mitigates risks and facilitates benefits related to AI technologies. As Ulnicane, et al. (2021b) highlight, governance as depicted in relation to AI takes on several meanings, being either integrated into the functions of governments, or alternatively represented as something similar to ethics, that simultaneously enforces it.

> Elements of governance which emphasize the importance of interaction between government and a broad range of societal actors in decision-making can be found in AI policy documents which suggest multi-stakeholder approaches, inclusion and dialogue to ensure that AI is developed and used according to interests and needs of the society (Ulnicane et al 2020). (Ulnicane, Eke, et al. 2021b)

The discourse around *AI governance* in ethics guidelines has converged on a set of topics and themes. As Ulnicane, et al. (2021a) suggest, this convergence shouldn’t be taken for granted as it might hide disagreements between different social actors.

> However, existence of similar themes and suggestions in AI discourse does not immediately mean that there is total convergence and differences across countries and organizations have disappeared. (Ulnicane, Knight, et al. 2021a)

Schiff, Biddle and Laas (2021) indicate, that differences between participatory processes in public sector and private sectors actors are prominent. Public sector organizations usually utilize task forces, hearings, or similar mechanisms to broaden inclusion of expertise (Schiff, Biddle and Laas 2021). Additionally, the perspective of private sector organizations appears to be focused and narrower in scope, which the researchers conjecture leads to a more superficial approach to ethical issues.

> Public sector and NGO documents are predominantly driven by inclusive and participatory processes, whereas private sector documents are largely not. Do these differences in participation lead to divergences in ethical priorities? Our results suggest that they do, given the distinctly narrower ethical scope of the private sector and correlations between participation and ethical breadth and depth (see Appendix Fig. 7 in the supplementary material). (Schiff, Biddle and Laas 2021)

The discourse on *AI governance*, as seen in the collected data, is marked with a critical undertone. According to Bélisle-Pipon, et al. (2022), the development and governance of AI is affected by a variety of stakeholders with a marked power asymmetry. As Häußler (2021) notes, a continuing struggle between various actors exists with disparity in meaningful inclusion in debate and discourse. Schopmans (2022) argues, that experts in the field of AI have considerable influence on the constitution of *AI governance* and policy interventions and the effects of this misalignment in influence are undetermined.

> While these studies have critically deconstructed dominant discourses and uncovered the sociotechnical imaginaries reproduced in AI policies, their focus, too, has been on political responses to a phenomenon that is largely seen as given. Less attention has been paid to the politics of AI expertise—the competitive processes in which different epistemic actors have constructed AI as an object that is problematic, and thus in need of governance, in the first place. (Schopmans 2022)

These shortcomings are emphasized by Sigfrids, et al. (2022), who argue that the currently available tools for governing AI, such as regulatory and technical standards, are not sufficient in steering AI in a socially purposeful and beneficial direction. Sigfrids, et al. (2022) highlight the approach advocated for by Ulnicane, et al. (2021a), to incorporate a wider category of stakeholders and communities in *AI governance*.

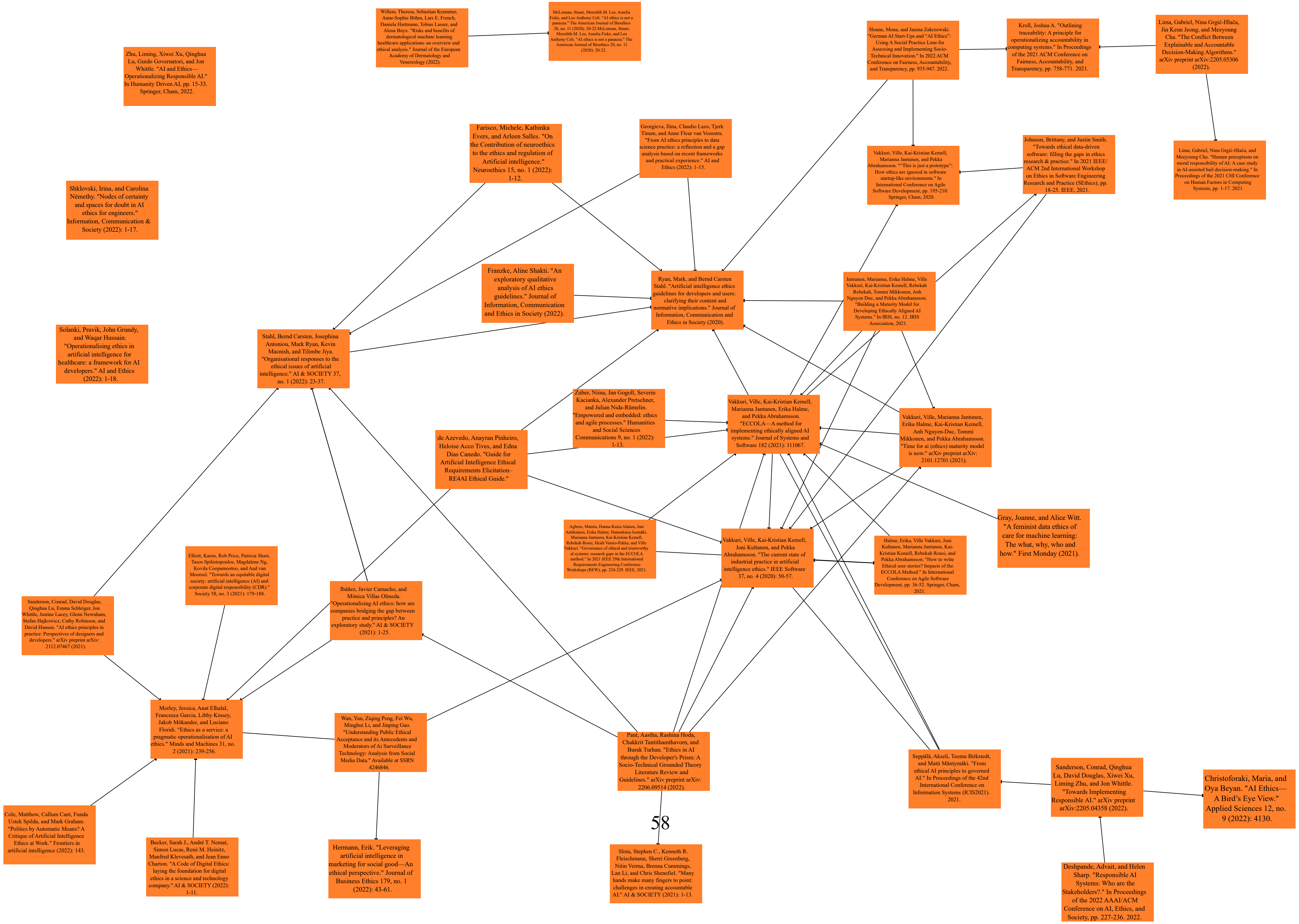
> For example, researchers have suggested that a new regulatory agency should be developed to support the operationalization of good governance and ethical principles, the assessment of ethical issues and social impacts should be an indispensable part of AI development, and governance should utilize more people-centered and inclusive policy-making (e.g., Floridi et al., 2018; de Almeida et al., 2021; Ireni-Saban and Sherman, 2021; Stahl, 2021; Taeihagh, 2021; Ulnicane et al., 2021). (Sigfrids, et al. 2022)

In the discourse on *AI governance,* the importance of participation by different social actors and stakeholders in articulating goals and principles is generally emphasized. The call for flexible and adaptive governance in the form of decentralized decision-making in a bottom-up fashion, internal and external expertise, and continuous adaptation to uncertainty take on key roles in governing AI (Sigfrids, et al. 2022). States and especially policymakers serve a crucial role in this process by addressing societal challenges through policy (Ulnicane, Eke, et al. 2021b). In this context states mediate between the different interests and needs of stakeholders, while managing risks and supporting inclusion. As Djeffal, et al. (2022) argue, depending on the situation states have to manage between serving either a proactive or a passive role.

> First, governments can play a proactive role in the development of AI technologies (strong state intervention), or they take a more passive stance by stepping back and giving private actors and/or the markets as much leeway as possible in the governance of AI (weak state intervention). Second, governments can concentrate on regulating potential risks of AI technologies (enclosure-and-control approach), or they can prioritise the deployment of AI and see their role primarily in promoting its development (stimulation approach). (Djeffal, Siewert and Wurster 2022)

## 5.4 AI Ethics Praxis

The research data collected for this thesis contained discourse around awareness of ethical issues during AI system development, including roles and responsibilities in implementing ethics from the AI developer perspective. The *AI ethics praxis* discourse is depicted in Figure 8. As the visualization indicates, the discourse is partially fragmented, revolving around a few central publications that other articles reference. An overarching topic connecting the central publications seen in the discourse, is their attempt at providing methodological impetus, thereby resulting in the production of actionable tools, methods, and frameworks that progress the state of the art. Based on the collected data, the praxis discourse can be regarded as one of the more prominent discourses, containing a sizable number of articles and references. As seen in Figure 8, the praxis discourse is continuing with some of the more recent articles focusing on empirically validating the proposed tools, methods, and frameworks. Additionally, a prevalent topic in the *AI ethics praxis* discourse pertains to organizational responses in addressing the ethical issues related to AI in practice.



Willem, Theresa, Sebastian Krammer, Anne-Sophie Böhm, Lars E. French, Daniela Hartmann, Tobias Lasser, and Alena Buyx. "Risks and benefits of dermatological machine learning healthcare applications–an overview and ethical analysis." Journal of the European Academy of Dermatology and Venereology (2022).

McLennan, Stuart, Meredith M. Lee, Amelia Fiske, and Leo Anthony Celi. "AI ethics is not a panacea." The American Journal of Bioethics 20, no. 11 (2020): 20-22.McLennan, Stuart, Meredith M. Lee, Amelia Fiske, and Leo Anthony Celi. "AI ethics is not a panacea." The American Journal of Bioethics 20, no. 11 (2020): 20-22.

Sloane, Mona, and Janina Zakrzewski. "German AI Start-Ups and “AI Ethics”: Using A Social Practice Lens for Assessing and Implementing Socio-Technical Innovation." In 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 935-947. 2022.

Kroll, Joshua A. "Outlining traceability: A principle for operationalizing accountability in computing systems." In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 758-771. 2021.

Lima, Gabriel, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. "The Conflict Between Explainable and Accountable Decision-Making Algorithms." arXiv preprint arXiv:2205.05306 (2022).

Zhu, Liming, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. "AI and Ethics—Operationalizing Responsible AI." In Humanity Driven AI, pp. 15-33. Springer, Cham, 2022.

Farisco, Michele, Kathinka Evers, and Arleen Salles. "On the Contribution of neuroethics to the ethics and regulation of Artificial intelligence." Neuroethics 15, no. 1 (2022): 1-12.

Georgieva, Ilina, Claudio Lazo, Tjerk Timan, and Anne Fleur van Veenstra. "From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience." AI and Ethics (2022): 1-15.

Vakkuri, Ville, Kai-Kristian Kemell, Marianna Jantunen, and Pekka Abrahamsson. "“This is just a prototype”: How ethics are ignored in software startup-like environments." In International Conference on Agile Software Development, pp. 195-210. Springer, Cham, 2020.

Johnson, Brittany, and Justin Smith. "Towards ethical data-driven software: filling the gaps in ethics research & practice." In 2021 IEEE/ ACM 2nd International Workshop on Ethics in Software Engineering Research and Practice (SEthics), pp. 18-25. IEEE, 2021.

Lima, Gabriel, Nina Grgić-Hlača, and Meeyoung Cha. "Human perceptions on moral responsibility of AI: A case study in AI-assisted bail decision-making." In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-17. 2021.

Shklovski, Irina, and Carolina Némethy. "Nodes of certainty and spaces for doubt in AI ethics for engineers." Information, Communication & Society (2022): 1-17.

Franzke, Aline Shakti. "An exploratory qualitative analysis of AI ethics guidelines." Journal of Information, Communication and Ethics in Society (2022).

Ryan, Mark, and Bernd Carsten Stahl. "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications." Journal of Information, Communication and Ethics in Society (2020).

Jantunen, Marianna, Erika Halme, Ville Vakkuri, Kai-Kristian Kemell, Rebekah Rebekah, Tommi Mikkonen, Anh Nguyen Duc, and Pekka Abrahamsson. "Building a Maturity Model for Developing Ethically Aligned AI Systems." In IRIS, no. 12. IRIS Association, 2021.

Solanki, Pravik, John Grundy, and Waqar Hussain. "Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers." AI and Ethics (2022): 1-18.

Stahl, Bernd Carsten, Josephina Antoniou, Mark Ryan, Kevin Macnish, and Tilimbe Jiya. "Organisational responses to the ethical issues of artificial intelligence." AI & SOCIETY 37, no. 1 (2022): 23-37.

Zuber, Niina, Jan Gogoll, Severin Kacianka, Alexander Pretschner, and Julian Nida-Rümelin. "Empowered and embedded: ethics and agile processes." Humanities and Social Sciences Communications 9, no. 1 (2022): 1-13.

Vakkuri, Ville, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. "ECCOLA—A method for implementing ethically aligned AI systems." Journal of Systems and Software 182 (2021): 111067.

Vakkuri, Ville, Marianna Jantunen, Erika Halme, Kai-Kristian Kemell, Anh Nguyen-Duc, Tommi Mikkonen, and Pekka Abrahamsson. "Time for ai (ethics) maturity model is now." arXiv preprint arXiv: 2101.12701 (2021).

de Azevedo, Anayran Pinheiro, Heloise Acco Tives, and Edna Dias Canedo. "Guide for Artificial Intelligence Ethical Requirements Elicitation–RE4AI Ethical Guide."

Gray, Joanne, and Alice Witt. "A feminist data ethics of care for machine learning: The what, why, who and how." First Monday (2021).

Agbese, Mamia, Hanna-Kaisa Alanen, Jani Antikainen, Erika Halme, Hannakaisa Isomäki, Marianna Jantunen, Kai-Kristian Kemell, Rebekah Rousi, Heidi Vainio-Pekka, and Ville Vakkuri. "Governance of ethical and trustworthy al systems: research gaps in the ECCOLA method." In 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), pp. 224-229. IEEE, 2021.

Vakkuri, Ville, Kai-Kristian Kemell, Joni Kultanen, and Pekka Abrahamsson. "The current state of industrial practice in artificial intelligence ethics." IEEE Software 37, no. 4 (2020): 50-57.

Halme, Erika, Ville Vakkuri, Joni Kultanen, Marianna Jantunen, Kai-Kristian Kemell, Rebekah Rousi, and Pekka Abrahamsson. "How to write Ethical user stories? Impacts of the ECCOLA Method." In International Conference on Agile Software Development, pp. 36-52. Springer, Cham, 2021.

Elliott, Karen, Rob Price, Patricia Shaw, Tasos Spiliotopoulos, Magdalene Ng, Kovila Coopamootoo, and Aad van Moorsel. "Towards an equitable digital society: artificial intelligence (AI) and corporate digital responsibility (CDR)." Society 58, no. 3 (2021): 179-188.

Sanderson, Conrad, David Douglas, Qinghua Lu, Emma Schleiger, Jon Whittle, Justine Lacey, Glenn Newnham, Stefan Hajkowicz, Cathy Robinson, and David Hansen. "AI ethics principles in practice: Perspectives of designers and developers." arXiv preprint arXiv: 2112.07467 (2021).

Ibáñez, Javier Camacho, and Mónica Villas Olmeda. "Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study." AI & SOCIETY (2021): 1-25.

Morley, Jessica, Anat Elhalal, Francesca Garcia, Libby Kinsey, Jakob Mökander, and Luciano Floridi. "Ethics as a service: a pragmatic operationalisation of AI ethics." Minds and Machines 31, no. 2 (2021): 239-256.

Wan, Yan, Ziqing Peng, Fei Wu, Minghui Li, and Jinping Gao. "Understanding Public Ethical Acceptance and its Antecedents and Moderators of Ai Surveillance Technology: Analysis from Social Media Data." Available at SSRN 4246846.

Pant, Aastha, Rashina Hoda, Chakkrit Tantithamthavorn, and Burak Turhan. "Ethics in AI through the Developer's Prism: A Socio-Technical Grounded Theory Literature Review and Guidelines." arXiv preprint arXiv: 2206.09514 (2022).

Seppälä, Akseli, Teemu Birkstedt, and Matti Mäntymäki. "From ethical AI principles to governed AI." In Proceedings of the 42nd International Conference on Information Systems (ICIS2021). 2021.

Sanderson, Conrad, Qinghua Lu, David Douglas, Xiwei Xu, Liming Zhu, and Jon Whittle. "Towards Implementing Responsible AI." arXiv preprint arXiv:2205.04358 (2022).

Christoforaki, Maria, and Oya Beyan. "AI Ethics—A Bird’s Eye View." Applied Sciences 12, no. 9 (2022): 4130.

Cole, Matthew, Callum Cant, Funda Ustek Spilda, and Mark Graham. "Politics by Automatic Means? A Critique of Artificial Intelligence Ethics at Work." Frontiers in artificial intelligence (2022): 143.

Becker, Sarah J., André T. Nemat, Simon Lucas, René M. Heinitz, Manfred Klevesath, and Jean Enno Charton. "A Code of Digital Ethics: laying the foundation for digital ethics in a science and technology company." AI & SOCIETY (2022): 1-11.

Hermann, Erik. "Leveraging artificial intelligence in marketing for social good—An ethical perspective." Journal of Business Ethics 179, no. 1 (2022): 43-61.

Slota, Stephen C., Kenneth R. Fleischmann, Sherri Greenberg, Nitin Verma, Brenna Cummings, Lan Li, and Chris Shenefiel. "Many hands make many fingers to point: challenges in creating accountable AI." AI & SOCIETY (2021): 1-13.

Deshpande, Advait, and Helen Sharp. "Responsible AI Systems: Who are the Stakeholders?." In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 227-236. 2022.

Figure 8 AI Ethics Praxis Discourse

Central issues in *AI ethics praxis* regard communication challenges in implementing ethics and practical tools to bridge the gap between principles and practice. Different strategies towards enhancing ethical consideration were proposed in the literature. Among these, speculation of socio-ethical impacts of systems and group discussion around raising of awareness about ethical issues related to AI were prominent, as discussed by Aastha, et al. (2022).

AI ethics suffers from a lack of empirically proven methods to translate ethical principles into actionable tools. The various AI ethics guidelines have not had a marked impact on industry practices (Vakkuri, Kemell and Kultanen, et al. 2020a). The principles propagated in guidelines are nevertheless relevant (Vakkuri, Jantunen, et al. 2021b). According to Ryan and Stahl (2021), the normative content within the different guidelines needs to be distilled and articulated, since there exist substantial differences contained within the seeming overlap of main issues and themes. Ryan and Stahl (2021) highlight, that different topics are emphasized in different guidelines, articulated with differing tones, described in varying lengths, created for different audiences at different levels of technicality. This has resulted in interest towards researching and testing practical methods that enable developers and practitioners to implement ethical guidelines in practice.

In the research data collected for this thesis, the discourse around practical methods focused on a few prominent tools. At the center of this discourse is ECCOLA, which is designed for developers to implement ethics in AI development (Vakkuri, Kemell, et al. 2021a).

> ECCOLA (Fig. 1) is intended to provide developers an actionable tool for implementing AI ethics. To utilize the various AI ethics guidelines in practice, the organization seeking to do so has to somehow make them practical first. ECCOLA, on the other hand, is intended to be practical as is, and ready to be incorporated into any existing method. (Vakkuri, Kemell, et al. 2021a)

ECCOLA is manifested in the form of a deck of cards (Vakkuri, Kemell, et al. 2021a). These cards are thematical categorized according to AI ethics topics and principles as seen in prominent AI ethics guidelines, such as in the IEEE Ethically Aligned Design guidelines (2019) and EU ethics guidelines for trustworthy AI (2019). ECCOLA is designed with the objectives of helping create awareness around AI ethics issues and their importance, making a modular method applicable in varied software engineering contexts, and embedding ethics in agile development processes. As Vakkuri, et al. (2021a) indicate, ECCOLA is intended to facilitate ethical thinking in AI development. Following ECCOLA enables the documentation of trade-offs in ethical considerations during the development of AI. Thus, the researchers argue that the trustworthiness of systems can be ascertained. ECCOLA has been further developed and researched, as can be seen in the collected data for this thesis. The discourse around actionable tools, such as ECCOLA, incorporates topics related to *AI governance* and development methods for *trustworthy AI* systems, as seen in Agbese, et al. (2021). As they argue, ECCOLA could be supplemented by representing aspects related to data governance and information governance in a more comprehensive way.

> The analysis revealed that IG has the least representation in ECCOLA. The identified governance practices exist in ECCOLA but to varying degrees. Corporate governance practices are represented in all the cards, but data governance and IG practices are not completely represented in all the cards. This indicates that ECCOLA can be improved. (Agbese, et al. 2021)

According to Zuber, et al. (2022) the previously discussed approach is lacking in some respects.

> While Vakkuri et al. embed ethics into processes using familiar methods of software engineering it remains unclear what guides such an ethical deliberation besides sheer luck of looking at the right card at the right moment. Therefore, it is of great importance to combine such a software engineering approach with ethical frameworks that support the identification of relevant normative aspects. (Zuber, et al. 2022)

Thus, it is argued that the cultivation of evaluative rationality needs to be facilitated and the procedure of normative reason must be exercised by developers (Zuber, et al. 2022). According to Zuber, et al. (2022), principles, values, and laws provide orientation and guidance in this process, even when they are underdetermined.

ECCOLA, while being a prominent method in the discourse related to the implementation of AI ethics, is not the only such method. An alternative approach found in the collected data is the RE4AI Ethical Guide (Siqueira de Cerqueira, et al. 2022). The approach resembles

ECCOLA in its use of a deck of cards, but according to the researchers differs in some key facets.

> This Guide differ from the ECCOLA method by Vakkuri et al. [4] in many aspects, while the latter is presented only as a deck of cards in Portable Document Format, our guide is developed as a web-based system (using HTML, CSS and JS), allowing interactivity in card selection through filters and comparisons between multiple cards. Furthermore, the addition of tools suggestion in the content of the cards, as well as extensive supporting material (how to use, principles, tools, trade-offs). (Siqueira de Cerqueira, et al. 2022)
>
> We also found the need for the inclusion of traditional software engineering practices, such as requirements elicitation, for the context of Artificial Intelligence, in addition to the characteristics of a Guide to implement ethics in AI [10]: broad, operationalisable, flexible, iterative, guided and participatory. (Siqueira de Cerqueira, et al. 2022)

The ethical principles used in these approaches are based on different guidelines and incorporate different principles, but according to de Cerqueira, et al. (2022) differ only in name, not content. Thus, the researchers were able to map the cards used in ECCOLA to principles and further standardize them, which served as the starting point for their guide. The researchers argue that RE4AI as a practical tool has several benefits.

> Our findings suggest that the RE4AI Ethical Guide is perceived to be of great interest by participants, receiving an overall positive evaluation. The Guide, by operationalising ethical principles, can help mitigate challenges present in the literature, such as: lack of tools to implement AI ethics at the project level [26], [27]; lack of tools that assist software development teams as a whole [4]; with practicality and usability offering help to be used in practice [26]; as well as the lack of tools that do not focus mostly on explicability [26]. (Siqueira de Cerqueira, et al. 2022)

The discourse around actionable tools, as seen in the collected data, additionally focused on evaluating the merits of methods such as ECCOLA. As Halme, et al. (2021) indicate, results are generally positive, even though further improvements can be made.

> As the summarizing PECs [Primary Empirical Contributions] in the above table show, the ECCOLA method [10] seemed to improve user stories in various ways. However, PEC2 also highlights an interesting observation in that ECCOLA did not make the user stories notably more focused on the themes of the ECCOLA cards in question. Moreover, the ECCOLA cards used in this study contained typical SE themes such as system security and privacy & data. (Halme, et al. 2021)

> Even if overall, ECCOLA produced positive results in this study, the contents of the cards may need adjusting based on PEC2. (Halme, et al. 2021)

Despite advancements, gaps in research especially related to ethics in data-driven software development still exist. According to Johnson and Smith (2021), ethical software development practices at the organizational level and across different domains is needed to supplement research conducted at the developer level. Johnson and Smith (2021) argue, that this will result in increased scalability and generalizability of findings. A similar necessity for further research on organizational practices is mentioned by Stahl, et al. (2022), who call for clarity in defining responsibilities related to managing issues and measures for ethical AI.

In addition to the aforementioned methods, maturity models were discussed in the collected literature as a means to incorporate principles into a more practical form (Vakkuri, Jantunen, et al. 2021b). Maturity models have seen extensive use in industry and could therefore provide a pathway for standardizing ethical AI development practices (Vakkuri, Jantunen, et al. 2021b). An AI ethics maturity model could facilitate the incorporation of ethical considerations at the developer level and organizational level (Jantunen, et al. 2021).

> Moreover, and perhaps on a more practical note concerning the implementation, uptake and potential success of such a model, there is the requirement to understand and align the cultures, values, decisions and actions of developers and organizations as a whole towards more common understandings of international ethical practice (Vakkuri, Kemell, Kultanen, et al., 2019; Weller, 2017). (Jantunen, et al. 2021)

Sanderson, et al. (2022) note, that implementing *AI ethics principles* requires a framework, which system design and development are only a component of. They highlight the categorization of Seppälä, et al. (2021), which postulates governance, AI design and development, competence and knowledge development, and stakeholder communication as the required sets of practices to implement principles in practice. Similarly, Morley, et al. (2021a), indicate the combination of law, ethical governance policies, practices, and procedures, with contextual discursive and procedural support in medical ethics, as a reasonable point of comparison, which pro-ethical design in AI ethics could be reflected with and built upon.

## 5.5 Ethics Auditing

Governance of AI systems has become a widespread issue, with a direct link to public trust. As a solution to concerns related to potential impacts of AI systems, researchers have suggested several mechanisms to mitigate harms (Kazim and Koshiyama 2021). In the collected literature the prominent proposals center around ethical auditing of AI systems and impact assessments, as pragmatic approaches. The *ethics auditing* discourse is depicted in Figure 9. As the visualization indicates, two distinct, yet thematically connected clusters related to ethical auditing were identified in the collected data. *Ethics auditing*, as discussed in the collected data, encompasses both technical aspects related to mitigating negative impacts of algorithms throughout their entire life cycle and social aspects related to alignment of relevant metrics with stakeholder interests and the effect of communication on trust. The discourse on ethical audits is still evolving, with several articles in the collected data focusing on frameworks and proposals for auditing. As the collected data indicates, empirical validation and case studies are still needed to further substantiate the merits and shortcomings of different auditing mechanisms.

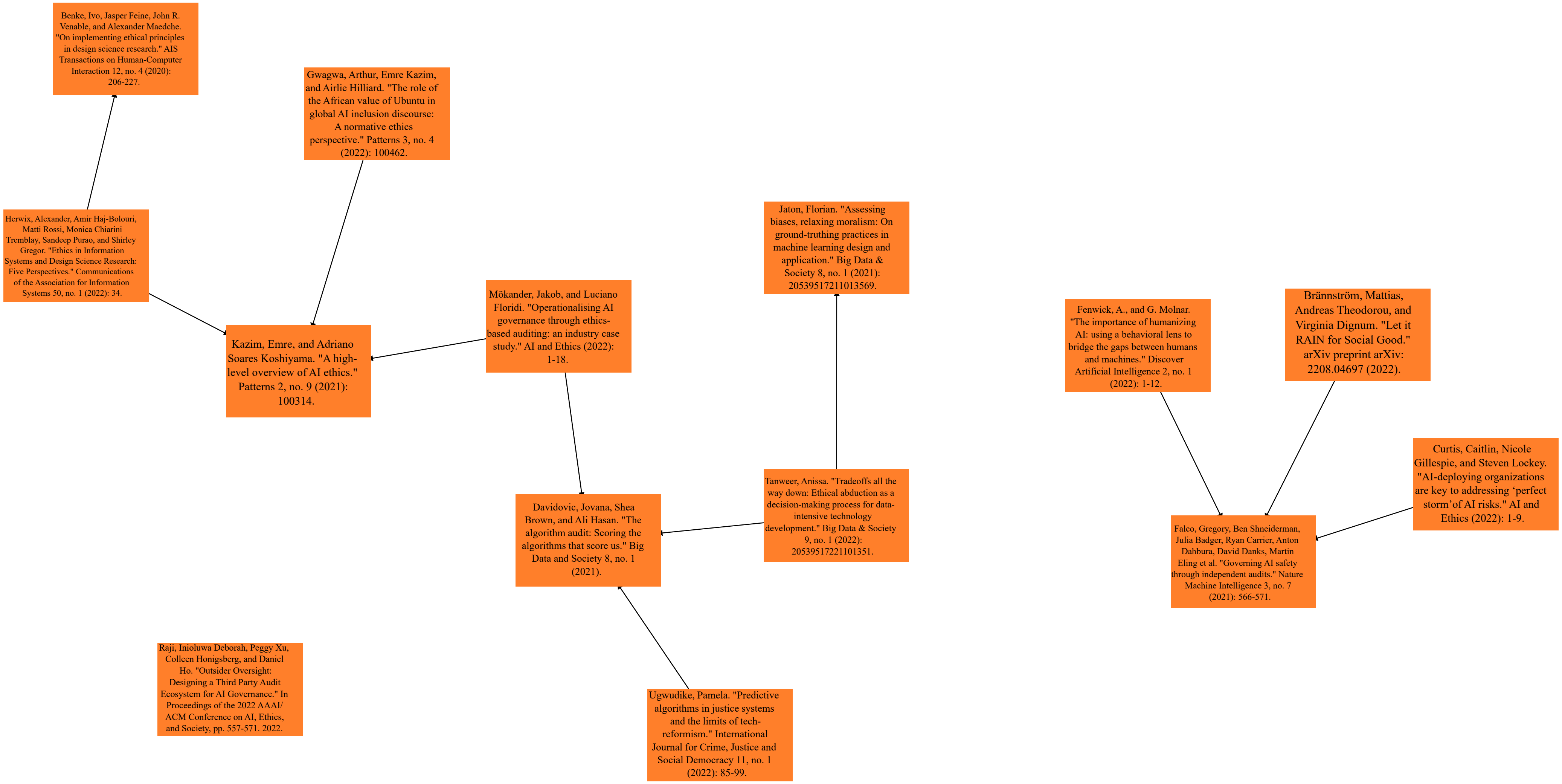

Benke, Ivo, Jasper Feine, John R. Venable, and Alexander Maedche. "On implementing ethical principles in design science research." AIS Transactions on Human-Computer Interaction 12, no. 4 (2020): 206-227.
Gwagwa, Arthur, Emre Kazim, and Airlie Hilliard. "The role of the African value of Ubuntu in global AI inclusion discourse: A normative ethics perspective." Patterns 3, no. 4 (2022): 100462.
Jaton, Florian. "Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application." Big Data & Society 8, no. 1 (2021): 20539517211013569.
Herwix, Alexander, Amir Haj-Bolouri, Matti Rossi, Monica Chiarini Tremblay, Sandeep Purao, and Shirley Gregor. "Ethics in Information Systems and Design Science Research: Five Perspectives." Communications of the Association for Information Systems 50, no. 1 (2022): 34.
Mökander, Jakob, and Luciano Floridi. "Operationalising AI governance through ethics-based auditing: an industry case study." AI and Ethics (2022): 1-18.
Fenwick, A., and G. Molnar. "The importance of humanizing AI: using a behavioral lens to bridge the gaps between humans and machines." Discover Artificial Intelligence 2, no. 1 (2022): 1-12.
Brännström, Mattias, Andreas Theodorou, and Virginia Dignum. "Let it RAIN for Social Good." arXiv preprint arXiv: 2208.04697 (2022).
Kazim, Emre, and Adriano Soares Koshiyama. "A high-level overview of AI ethics." Patterns 2, no. 9 (2021): 100314.
Curtis, Caitlin, Nicole Gillespie, and Steven Lockey. "AI-deploying organizations are key to addressing 'perfect storm'of AI risks." AI and Ethics (2022): 1-9.
Tanweer, Anissa. "Tradeoffs all the way down: Ethical abduction as a decision-making process for data-intensive technology development." Big Data & Society 9, no. 1 (2022): 20539517221101351.
Davidovic, Jovana, Shea Brown, and Ali Hasan. "The algorithm audit: Scoring the algorithms that score us." Big Data and Society 8, no. 1 (2021).
Falco, Gregory, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling et al. "Governing AI safety through independent audits." Nature Machine Intelligence 3, no. 7 (2021): 566-571.
Raji, Inioluwa Deborah, Peggy Xu, Colleen Honigsberg, and Daniel Ho. "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance." In Proceedings of the 2022 AAAI/ ACM Conference on AI, Ethics, and Society, pp. 557-571. 2022.
Ugwudike, Pamela. "Predictive algorithms in justice systems and the limits of tech-reformism." International Journal for Crime, Justice and Social Democracy 11, no. 1 (2022): 85-99.

Figure 9 Ethics Auditing Discourse

Audits, as discussed in the literature, typically involve evaluating algorithms or systems for negative impacts towards certain groups in particular contexts (Brown, Davidovic and Hasan 2021). Evaluation is generally concerned with assessing potential biases in technologies and unfair treatment of groups and individuals. Negative impacts have emerged as a pragmatic aggregation metric for evaluation, as they are immediately related to risk management and as such of primary interest for regulators (Brown, Davidovic and Hasan 2021). The proposals in the *ethics auditing* discourse potentially meet some of the challenges emphasized in the *AI ethics praxis* and *AI governance* discourses. However, based on the references and citations seen in Figure 4, the discourses are currently insubstantially connected.

The discourse in the collected data centered around several auditing frameworks generally designed to assist regulators and system designers in meeting legal standards and policy guidelines, mitigating potential risks, and enabling ethical assessments of systems. A key factor in ethical auditing is assessing the socio-technical context, that an algorithm is deployed in. This feature is often overlooked in ethical audits, detracting from the validity of attained ethical assessments (Brown, Davidovic and Hasan 2021). Brown, Davidovic and Hasan (2021) propose, that auditing should encompass evaluating the processes related to the development of the algorithm, preparing the data for the training algorithm, delivering an algorithm to its primary user, in addition to evaluating the setting within which the algorithm is used. The audit tool proposed by Brown, Davidovic and Hasan (2021), consists of identifying relevant metrics for algorithms, evaluating the algorithms performance in respect to these metrics and identifying the impact of these metrics towards relevant stakeholder interests.

The collected literature additionally contained discourse related to independent audits. Falco, et al. (2021), set out to establish an audit framework for practical *AI governance* in their research paper.

> This paper outlines a regulatory mechanism to achieve assurance at scale: the Independent Audit of AI Systems (IAAIS - pronounced "eyes"). The proposed audit framework could embody the authors'

> proposed “AAA” governance principles: 1) Prospective Assessments before highly automated systems are implemented 2) Audit trail to analyze failures and help assess accountability 3) System Adherence to jurisdictional requirements. (Falco, et al. 2021).

Governing AI systems through a principle-based approach has been criticized for its lack of practical tools and methods. As Curtis, et al. (2022) state, independent audits as discussed by Falco, et al. (2021), provide an actionable and enforceable tool for the governance of AI systems.

> Algorithmic Impact Assessment Tools are being encouraged and used in several jurisdictions [55], and independent audits have been also proposed as a mechanism of AI governance that is actionable and enforceable [56]. (Curtis, Gillespie and Lockey 2022).

Bridging the gap between abstract high-level principles and particular application is key in establishing socially beneficial AI. This requires co-operation by different stakeholders as Brännström, Theodorou, and Dignum (2022) argue.

> Further exacerbating the problem is that the sociotechnical domain typically consist of not a single actor, the AI developer, but an interplay between developers, procurers, customers and users [7, 8]. It is within this socio-technical multi-actor sphere where the effects of AI on society develop [7, 8, 6]. (Brännström, Theodorou and Dignum 2022)

> A continual chain encapsulating all levels of the socio-technical landscape is needed to ensure relevance to both actual applications and the larger society [8, 7]. (Brännström, Theodorou and Dignum 2022)

Brännström, Theodorou, and Dignum refer to governance mechanisms as seen in Falco et al. (2021), as potential solutions to these issues.

> Cross-application of local organisational policies and national and international guidelines allows procurers to set their own terms and requirements on their suppliers, enabling each layer of the chain to take responsibility [7, 8]. (Brännström, Theodorou and Dignum 2022)

Such an approach enables a shift from abstract notions of ethical standards to discussions focused on topical and relevant real-world issues. As Fenwick and Molnar (2022) argue, audit trails as proposed by Falco, et al. (2021), provide tools for the explainable and responsible operationalization of AI.

> To guide the evolution of AI operationalization in an explainable and responsible manner, various (micro-level) types of mechanisms need to be in place. These mechanisms i.e., audit trails (e.g. [54]), interpretability (e.g. [55]), and algorithmic design choices (e.g., [56]) can guide AI development and deployment into the future. (Fenwick and Molnar 2022)

> Other mechanisms which can be used to ensure more human centricity, accountability, and safety in application are audit trails (e.g. [54]), Responsible AI governance (e.g., [86]), and data bias and proportionality checks (e.g., [87]) among others. (Fenwick and Molnar 2022)

Raji, et al. (2022) indicate, the assumption in many AI policy proposals is, that auditing is a process for meeting compliance standards designed for internal stakeholders in organizations. Thus, their format and content are generally oriented towards operationalization by internal stakeholders and their concerns. Raji, et al. (2022) argue, that internal audits are similar to algorithmic impact assessments in this regard but lack comparable stakeholder involvement, articulation of trade-offs and decision-making. Thus, internal auditing in isolation fails to capture the wider range of harms faced by stakeholders (Raji, et al. 2022). This sentiment is shared by Ugwudike (2022), arguing for holistic audits, that ensure algorithms take into consideration structural outcomes, not being solely optimized for technical imperatives, such as validity and accuracy. Additionally, it is argued, that external accountability necessitates the need for independent third-party ethical audits in conjunction with internal auditing (Ugwudike 2022).

According to Mökander and Floridi (2022b), the feasibility and effectiveness of different auditing procedures in ethics-based auditing, require further substantiation by empirical research. In their case study of AstraZeneca's ethics-based AI audit, they highlight some of the practical issues organizations face when conducting ethical audits.

> While previous literature concerning EBA [Ethics-Based Auditing] has focussed on proposing or analysing evaluation metrics or visualisation techniques, our findings suggest that the main difficulties large multinational organisations face when conducting EBA mirror classical governance challenges. These include ensuring harmonised standards across decentralised organisations, demarcating the scope of the audit, driving internal communication and change management, and measuring actual outcomes. (Mökander and Floridi 2022b)

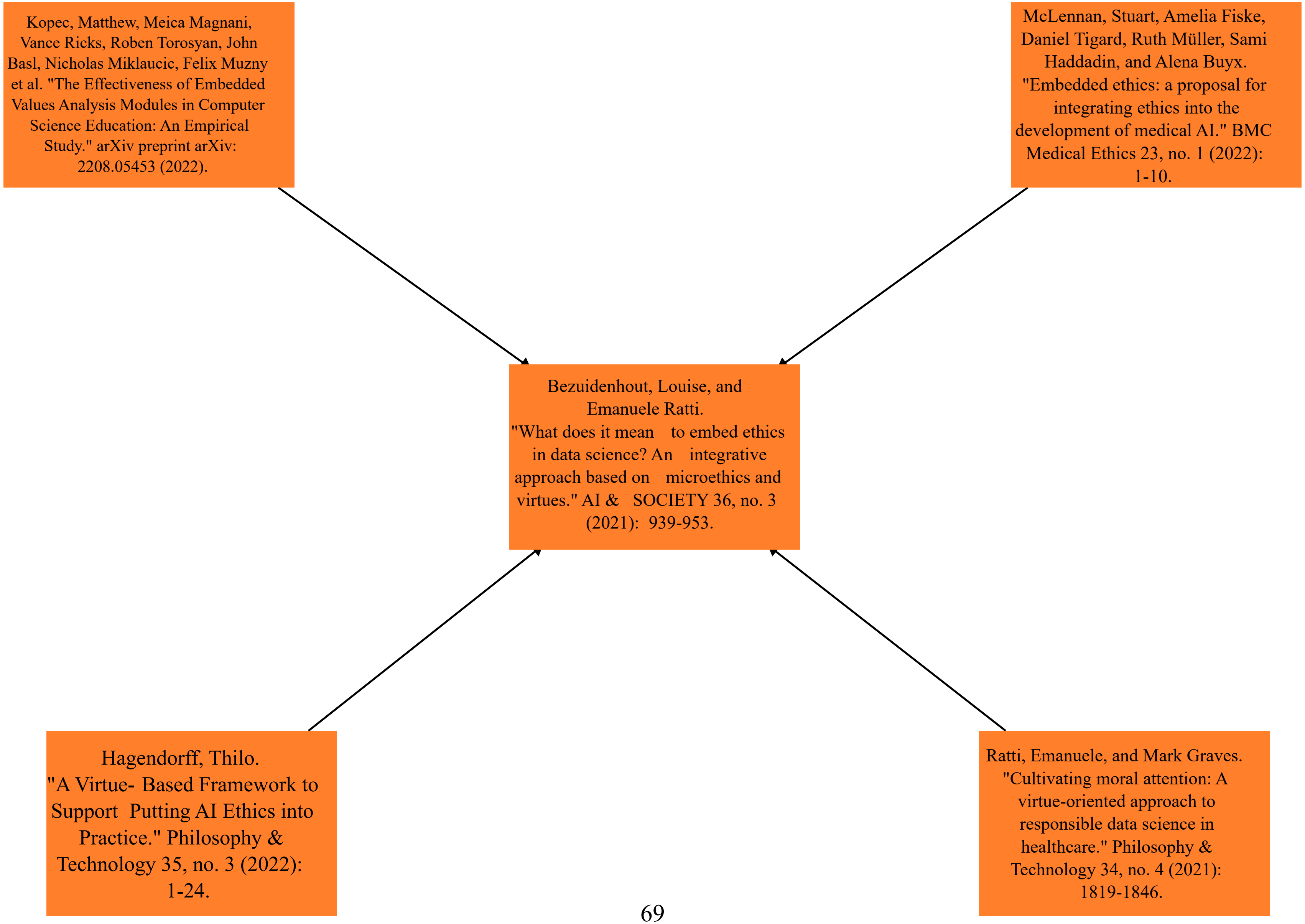
Mökander and Floridi (2022b) highlight the importance of procedural regularity and transparency in ethics-based auditing, setting traceable documentation as a vital objective.

Additionally, for ethics-based auditing to have a marked impact, the material scope of governance of AI must be accepted throughout the organization. Thus, internal communication and training efforts serve a key role in anchoring proposed policy with stakeholders and are essential for operationalizing corporate governance of AI. Towards this end Mökander and Floridi (2022b) suggest three constructive measures: support by senior executives in communication efforts of governance of AI, anchoring of ethics-based auditing procedures with employee's daily tasks, and communication related to the relevance of ethics-based auditing procedures.

## 5.6 Embedded Ethics

The *embedded ethics* discourse is depicted in Figure 10. As the visualization indicates, the discourse on *embedded ethics* in the collected data is relatively small containing only a few articles. The *embedded ethics* discourse, as seen in the collected data, consists of two distinct topics with some degree of thematic overlap. Firstly, the discourse revolves around embedding ethics in technology education. In the collected data the discourse around ethics education in technology contexts focused on methods to make theoretical proposals operational. The central proposals discussed in the collected data focused on a virtue-oriented approach to ethics education.

Secondly, the discourse revolves around embedding ethics in development processes and AI systems. The overarching themes connecting the two distinct topics center on changing normative attitudes, cultivating moral virtues, and fostering attention to moral contexts. By combining ethics education with repeated exposure to the ethical facets of AI, wider attention to ethical considerations in development processes and broader awareness of socio-technical issues and contextual factors can be cultivated.



Kopec, Matthew, Meica Magnani, Vance Ricks, Roben Torosyan, John Basl, Nicholas Miklaucic, Felix Muzny et al. "The Effectiveness of Embedded Values Analysis Modules in Computer Science Education: An Empirical Study." arXiv preprint arXiv: 2208.05453 (2022).

McLennan, Stuart, Amelia Fiske, Daniel Tigard, Ruth Müller, Sami Haddadin, and Alena Buyx. "Embedded ethics: a proposal for integrating ethics into the development of medical AI." BMC Medical Ethics 23, no. 1 (2022): 1-10.

Bezuidenhout, Louise, and Emanuele Ratti. "What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues." AI & SOCIETY 36, no. 3 (2021): 939-953.

Hagendorff, Thilo. "A Virtue- Based Framework to Support Putting AI Ethics into Practice." Philosophy & Technology 35, no. 3 (2022): 1-24.

Ratti, Emanuele, and Mark Graves. "Cultivating moral attention: A virtue-oriented approach to responsible data science in healthcare." Philosophy & Technology 34, no. 4 (2021): 1819-1846.

Figure 10 Embedded Ethics Discourse

Embedding ethics in data science practice by illustrating the emergence of ethical issues in day-to-day decision-making practices by data scientists has been considered as a practical way to make data ethics approachable (Bezuidenhout and Ratti 2021). Bezuidenhout and Ratti (2021) argue, that a virtue theory framework would be suitable in cultivating moral virtues and promoting agency in the pursuit of responsible action. Hagendorff (2022a) proposes, the encouragement and practice of specific virtues by emulating behavior considered representational of that specific virtue, as exemplified in patterns, narratives, and social models. This process would eventually include consideration of contextual factors and deliberation of the relative importance of conflicting virtues. A virtue ethical approach to data science approaches morality through a practical perspective, by going beyond compliance with principles. Being ethical and promoting morally responsible data science involves using virtues in decision-making and identifying the ethically relevant features and impacts of routine activities (Ratti and Graves 2021).

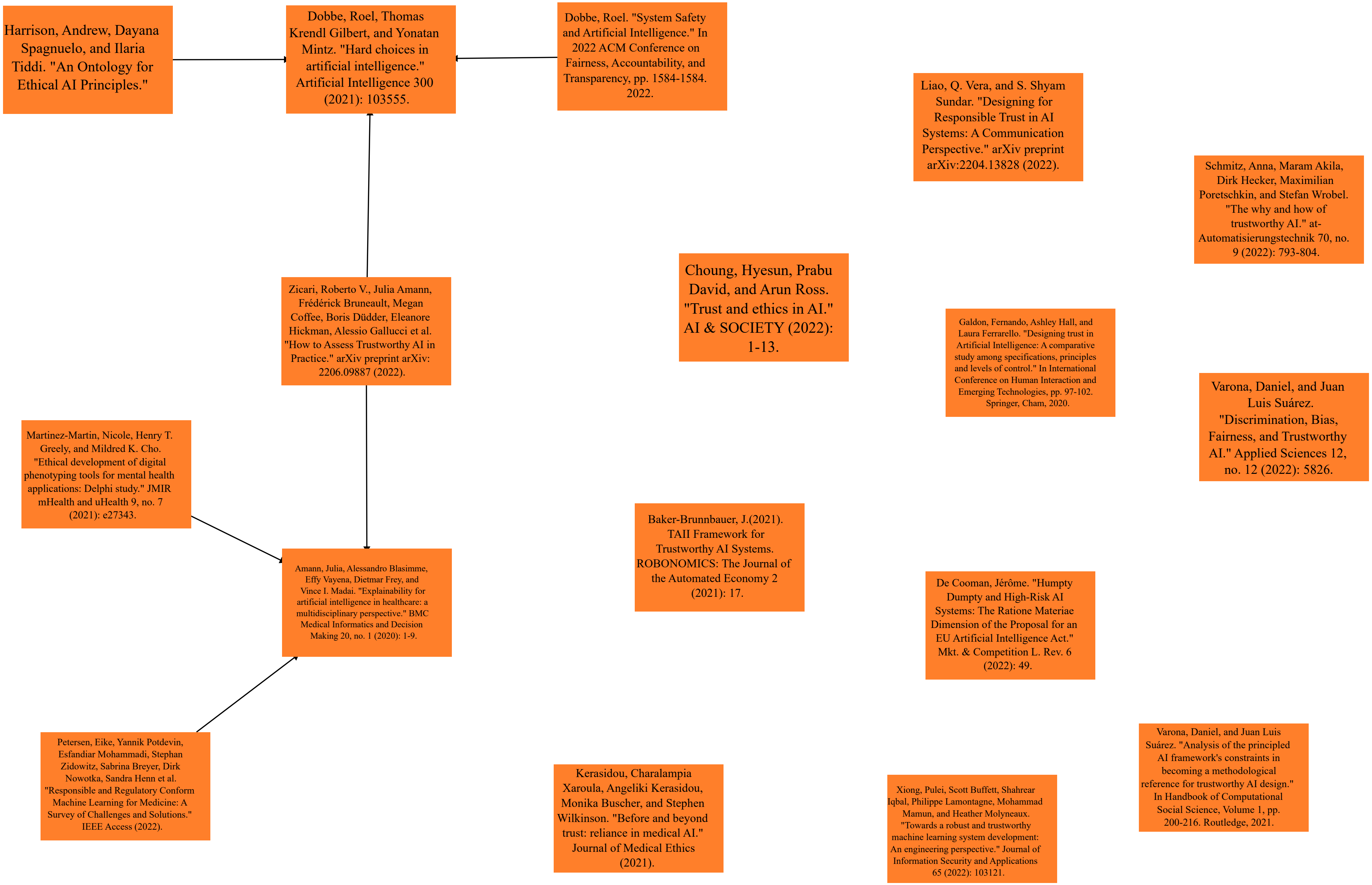
Embedding ethics courses into technology education has been a developing endeavor. Kopec, et al. (2022) emphasize, the existing lack of space to include technology ethics in education. This issue has been met by embedding ethics modules into a range of pre-existing courses, consequently reinforcing students' growth through repeated exposure to ethical issues. Kopec, et al. (2022) argue, that this kind of approach can result in notable changes in normative attitudes.

> And although student self-reports about the positive impacts of the modules are clearly fallible, the fact that students overall seemed to believe the modules had the impact we had hoped for in terms of noticing, caring about, and knowing how to better navigate ethical dilemmas is at least some evidence that they are effective on that front as well. (Kopec, et al. 2022)

In addition to the discussion about effective ways to incorporate ethics into education, the discourse around *embedded ethics* includes integrating ethics into development processes in a broader sense. From this perspective *embedded ethics* is viewed as encompassing ethically and socially responsible AI systems, that benefit society. Embedding ethics in development processes can serve a complementary role to educating AI developers and engineers, as well as enforcing the benefits of legislative and regulatory measures (McLennan, et al. 2022). Such an approach turns ethics into a collaborative and interdisciplinary enterprise, in which ethical issues can be addressed through an iterative and ongoing process (McLennan, et al. 2022). This regularity of exchanges on ethical considerations should be emphasized according to McLennan, et al. (2022), as irregular and disorganized consideration would undermine the authenticity, ethical awareness, and critical reasoning capacities of developers in ethical matters.

## 5.7 Trustworthy AI

The discourse on *trustworthy AI* is depicted in Figure 11. As the visualization indicates, the discourse on *trustworthy AI* is scattered and diverse, encompassing both technical and social aspects. On the technical side, discourse revolving around AI safety is a key issue. As Dobbe, Gilbert and Mintz (2021) indicate, assessing safety risks in complex social contexts remains difficult due to unclear and contested criteria. On the social side, issues pertaining to stakeholder participation and the role of communication in building trust can be seen as crucial issues. According to Varona and Suárez (2022), stakeholder involvement in the design and development of AI systems should extend across the life cycle and include diverse perspectives. Kerasidou, et al. (2022) argue, that adherence to ethical principles without complementary legal frameworks, isn't sufficient to ensure public trust. The dichotomy of technical and social topics contained within the collected data results in a fragmented discourse with most articles being relatively disconnected from each other thematically.



Harrison, Andrew, Dayana Spagnuelo, and Ilaria Tiddi. "An Ontology for Ethical AI Principles."

Dobbe, Roel, Thomas Krendl Gilbert, and Yonatan Mintz. "Hard choices in artificial intelligence." Artificial Intelligence 300 (2021): 103555.

Dobbe, Roel. "System Safety and Artificial Intelligence." In 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1584-1584. 2022.

Liao, Q. Vera, and S. Shyam Sundar. "Designing for Responsible Trust in AI Systems: A Communication Perspective." arXiv preprint arXiv:2204.13828 (2022).

Schmitz, Anna, Maram Akila, Dirk Hecker, Maximilian Poretschkin, and Stefan Wrobel. "The why and how of trustworthy AI." at-Automatisierungstechnik 70, no. 9 (2022): 793-804.

Zicari, Roberto V., Julia Amann, Frédérick Bruneault, Megan Coffee, Boris Düdder, Eleanore Hickman, Alessio Gallucci et al. "How to Assess Trustworthy AI in Practice." arXiv preprint arXiv: 2206.09887 (2022).

Choung, Hyesun, Prabu David, and Arun Ross. "Trust and ethics in AI." AI & SOCIETY (2022): 1-13.

Galdon, Fernando, Ashley Hall, and Laura Ferrarello. "Designing trust in Artificial Intelligence: A comparative study among specifications, principles and levels of control." In International Conference on Human Interaction and Emerging Technologies, pp. 97-102. Springer, Cham, 2020.

Varona, Daniel, and Juan Luis Suárez. "Discrimination, Bias, Fairness, and Trustworthy AI." Applied Sciences 12, no. 12 (2022): 5826.

Martinez-Martin, Nicole, Henry T. Greely, and Mildred K. Cho. "Ethical development of digital phenotyping tools for mental health applications: Delphi study." JMIR mHealth and uHealth 9, no. 7 (2021): e27343.

Baker-Brunnbauer, J.(2021). TAII Framework for Trustworthy AI Systems. ROBONOMICS: The Journal of the Automated Economy 2 (2021): 17.

Amann, Julia, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I. Madai. "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective." BMC Medical Informatics and Decision Making 20, no. 1 (2020): 1-9.

De Cooman, Jérôme. "Humpty Dumpty and High-Risk AI Systems: The Ratione Materiae Dimension of the Proposal for an EU Artificial Intelligence Act." Mkt. & Competition L. Rev. 6 (2022): 49.

Varona, Daniel, and Juan Luis Suárez. "Analysis of the principled AI framework's constraints in becoming a methodological reference for trustworthy AI design." In Handbook of Computational Social Science, Volume 1, pp. 200-216. Routledge, 2021.

Petersen, Eike, Yannik Potdevin, Esfandiar Mohammadi, Stephan Zidowitz, Sabrina Breyer, Dirk Nowotka, Sandra Henn et al. "Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions." IEEE Access (2022).

Kerasidou, Charalampia Xaroula, Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. "Before and beyond trust: reliance in medical AI." Journal of Medical Ethics (2021).

Xiong, Pulei, Scott Buffett, Shahrear Iqbal, Philippe Lamontagne, Mohammad Mamun, and Heather Molyneaux. "Towards a robust and trustworthy machine learning system development: An engineering perspective." Journal of Information Security and Applications 65 (2022): 103121.

Figure 11 Trustworthy AI Discourse

AI safety criteria remain contested as systems are integrated into complex social contexts and critical social domains (Dobbe, Gilbert and Mintz 2021). *Trustworthy AI*, as depicted in guidelines, is outlined through abstract requirements and recommendations leading to open-ended interpretation regarding measures and methods that are considered applicable and appropriate (Schmitz, et al. 2022). Discourse relating to AI safety is characterized by a certain vagueness inherent to socio-technical challenges and contexts these systems are deployed in. Dobbe, Gilbert and Mintz (2021) argue, that deliberation should not only be the procedure by which AI safety is ensured, but also be its goal. According to these researchers, deliberation could alleviate vagueness related to normative considerations associated with the social context systems operate in.

Through deliberation a consensus on roles and responsibilities related to system operation can be reached, with a pre-established process and set of requirements. Dobbe, Gilbert and Mintz (2021) additionally stress the importance of sufficient perspectives and distribution of decision-making during this process.

> Without clarifying this landscape, it will not be possible to evaluate whether particular governance mechanisms at different institutional scales are more or less appropriate for addressing the indeterminacies at stake. (Dobbe, Gilbert and Mintz 2021)

Harrison, et al. (2021) on the other hand propose, that knowledge graphs based on their AI principles ontology could serve as complementary tools for deliberation processes, thereby alleviating difficulties related to differing terminology and contextual factors. An AI principles ontology would provide context and connect ethical terms with prior usages in the knowledge graph, providing valuable information in situations of normative uncertainty. As Zicari, et al. (2022) however point out, tensions between differing requirements and their respective balancing represent a complex issue, that requires further research.

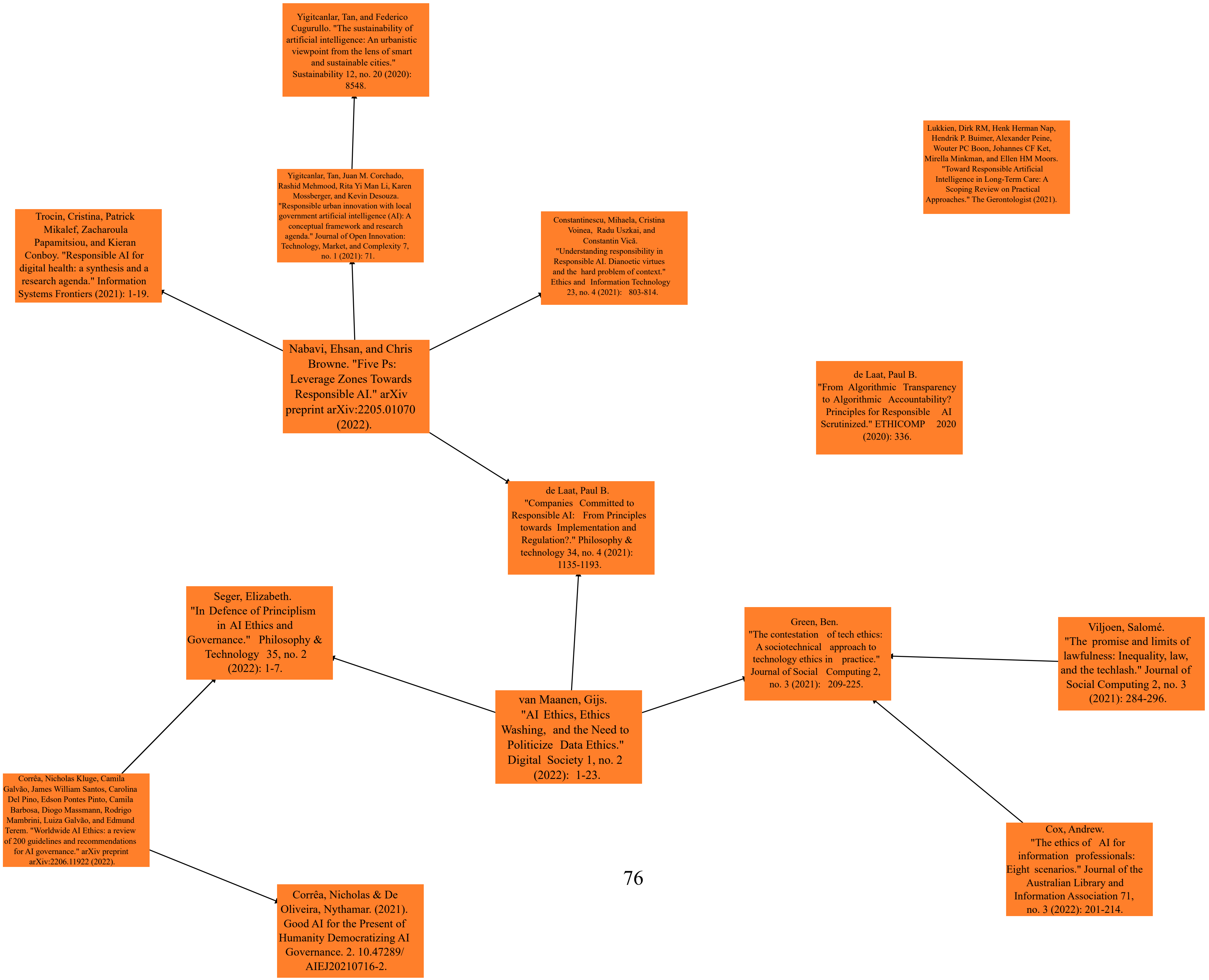
In addition to the previously discussed topics, AI safety discourse in the collected data additionally encompasses discussions of risk management and regulatory frameworks for mitigating risks. As Petersen, et al. (2021) note, technical challenges related to machine learning systems can be viewed as risks within a risk minimization framework. Safety concerns, such as susceptibility to adversarial attacks, correspond to requirements in risk mitigation strategies in pre-existing regulatory frameworks to some degree (Petersen, et al. 2021). These requirements can be seen in for instance the proposed regulatory framework for AI by the EU commission (2021), which emphasizes the importance of effective governance and enforcement of existing law on fundamental rights and safety requirements applicable to AI systems. Thus, regulatory frameworks and legislation play an essential role in ensuring AI systems adhere to safety standards and ethical requirements.

Risk-based approaches are prominent governance mechanisms, that ensure the fulfillment of rigorous safety requirements, by identifying possible risks and their probability of occurrence. In the context of AI systems an additional aspect of concern is functional safety, which denotes the reliable performance of a system and its intended functions (Martinez-Martin, Greely and Cho 2021). Risk-based approaches provide additional safeguards in cases where regulation and standards for evaluation of validity, accuracy, and effectiveness are lacking and vary across contexts. Safety and efficacy concerns related to AI systems, however, require further consideration when placed within larger structures of interconnected systems (Martinez-Martin, Greely and Cho 2021).

According to Varona and Suárez (2022), achieving *trustworthy AI* requires involving all affected stakeholders, throughout the AI system's development life cycle. This necessity for stakeholder participation and the importance of communication can be seen extending throughout the social side of the *trustworthy AI* discourse. The mechanisms by which people judge trustworthiness and the ways in which the trustworthiness of AI should appropriately be communicated, are largely neglected in the *trustworthy AI* discourse. As Liao and Sundar (2022) highlight, principles underlying trustworthy AI are researched prominently, however the ways in which individuals develop trust is largely overlooked. Individual, organizational, and cultural contexts play an important role in understanding how AI is perceived and trust in AI is established (Choung, David and Ross 2022).

## 5.8 Responsible AI

The discourse on *responsible AI* is depicted in Figure 12. As the visualization indicates, the discourse on *responsible AI* is diverse, encompassing topics related to different roles and responsibilities of various social actors in ensuring *responsible AI* development, as well as topics related to establishing consensus on contested normative attitudes and navigating value disagreements towards compromises, affecting positive change and steering AI development towards a culture of responsibility. The discourse on *responsible AI*, as seen in the collected data, is thematically connected in trying to establish common ground and shared goals, despite containing varied topics. The discourse on *responsible AI* is still evolving, with a considerable number of articles, seen in the collected data, being published in recent years.



Yigitcanlar, Tan, and Federico Cugurullo. "The sustainability of artificial intelligence: An urbanistic viewpoint from the lens of smart and sustainable cities." Sustainability 12, no. 20 (2020): 8548.

Lukkien, Dirk RM, Henk Herman Nap, Hendrik P. Buimer, Alexander Peine, Wouter PC Boon, Johannes CF Ket, Mirella Minkman, and Ellen HM Moors. "Toward Responsible Artificial Intelligence in Long-Term Care: A Scoping Review on Practical Approaches." The Gerontologist (2021).

Yigitcanlar, Tan, Juan M. Corchado, Rashid Mehmood, Rita Yi Man Li, Karen Mossberger, and Kevin Desouza. "Responsible urban innovation with local government artificial intelligence (AI): A conceptual framework and research agenda." Journal of Open Innovation: Technology, Market, and Complexity 7, no. 1 (2021): 71.

Trocin, Cristina, Patrick Mikalef, Zacharoula Papamitsiou, and Kieran Conboy. "Responsible AI for digital health: a synthesis and a research agenda." Information Systems Frontiers (2021): 1-19.

Constantinescu, Mihaela, Cristina Voinea, Radu Uszkai, and Constantin Vică. "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context." Ethics and Information Technology 23, no. 4 (2021): 803-814.

Nabavi, Ehsan, and Chris Browne. "Five Ps: Leverage Zones Towards Responsible AI." arXiv preprint arXiv:2205.01070 (2022).

de Laat, Paul B. "From Algorithmic Transparency to Algorithmic Accountability? Principles for Responsible AI Scrutinized." ETHICOMP 2020 (2020): 336.

de Laat, Paul B. "Companies Committed to Responsible AI: From Principles towards Implementation and Regulation?." Philosophy & technology 34, no. 4 (2021): 1135-1193.

Seger, Elizabeth. "In Defence of Principlism in AI Ethics and Governance." Philosophy & Technology 35, no. 2 (2022): 1-7.

Green, Ben. "The contestation of tech ethics: A sociotechnical approach to technology ethics in practice." Journal of Social Computing 2, no. 3 (2021): 209-225.

Viljoen, Salomé. "The promise and limits of lawfulness: Inequality, law, and the techlash." Journal of Social Computing 2, no. 3 (2021): 284-296.

van Maanen, Gijs. "AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics." Digital Society 1, no. 2 (2022): 1-23.

Corrêa, Nicholas Kluge, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, Rodrigo Mambrini, Luiza Galvão, and Edmund Terem. "Worldwide AI Ethics: a review of 200 guidelines and recommendations for AI governance." arXiv preprint arXiv:2206.11922 (2022).

Cox, Andrew. "The ethics of AI for information professionals: Eight scenarios." Journal of the Australian Library and Information Association 71, no. 3 (2022): 201-214.

Corrêa, Nicholas & De Oliveira, Nythamar. (2021). Good AI for the Present of Humanity Democratizing AI Governance. 2. 10.47289/ AIEJ20210716-2.

Figure 12 Responsible AI Discourse

The role of different social actors is contested in the discourse on *Responsible AI*. Responsibility in the context of AI is used to refer to the overarching efforts by companies, organizations, and governments to design, implement, and use AI technologies in an ethical manner (Trocin, et al. 2021). This project applies to both the organizational level and the systems level of procedures and mechanisms. In this sense, responsibility is generally understood as an umbrella term denoting AI that is fair, non-biased, transparent, explainable, secure, safe, privacy-proof, accountable, and to the benefit of humanity (de Laat 2021). As Nabavi and Browne (2022) highlight, the effort towards *responsible AI* has been a combined effort by academia, organizations, and companies through research and policy initiatives. However, the effectiveness of these initiatives and efforts is contested, as Nababi and Browne (2022) indicate, companies have limited capacities to realize meaningful change. The endorsement of ethical AI development often amounts to ticking the right checkboxes (de Laat 2020). As such, future research on *responsible AI* innovation at the level of specific technologies and their local contexts of use is required, otherwise *responsible AI* development risks being confined to the hypothetical, detached from the realities of practice and real impacts (Lukkien, et al. 2021).

Nabavi and Browne (2022) however argue, that the efforts thus far are good first steps and any steps in the right direction are appreciated in moving the industry towards *responsible AI*. The result of these initiatives has mostly been tools and software designed to translate high-level principles into engineering practices as well as standards and frameworks, that support *responsible AI* development (Nabavi and Browne 2022). Several of these frameworks were analyzed in chapter 5.4 on the *AI ethics praxis* discourse. The role of high-level principles and design guidelines in *responsible AI* development has been questioned and criticized extensively (Seger 2022). A common criticism towards adopting ethics principles and guidelines by AI researchers and developers has been labeling it as ethics washing i.e., an attempt to discourage the creation of truly binding legal frameworks, by purporting to the adherence of ethical guidelines, that are generally vague and ill-defined (Constantinescu, et al. 2021). According to Cox (2022), AI ethics codes are generally individualistic in nature, focusing on the design process and placing the burden of decision making on individual technical experts, thereby evading questions pertaining to organizational responsibility and the larger context of decision-making. Thus, ethical review is applied to design decisions, but the business practices driving decision-making are overlooked (Cox 2022). These criticisms correspond to similar challenges seen in the *AI ethics praxis* discourse, where the excessive focus on the developer level instead of organizational practices was criticized.

Despite the criticisms, calls for ethical AI have resulted in initiatives and guidelines, that have spurred considerable debate and brought ethics to the forefront of AI discourse (Viljoen 2020). Seger (2022) argues, that guidelines and principles serve an important role in establishing social norms and influencing AI development culture towards responsibility and beneficence. Extrinsic rules and requirements reinforce the internalization of new values and enable the progression towards policy goals (Seger 2022). Principles in combination with rules, requirements and explicit regulation foster movement towards cultural change, consequently going beyond compliance to minimum ethical standards, preventing ethics washing and ensuring the aims of beneficence and responsibility in AI development (Seger 2022).

The overarching challenges of *responsible AI* have generated a temptation towards technological solutionism, i.e., the mistaken belief that complex social issues can be resolved through technical means (Green 2021b). This belief is built on the premise, that AI will be beneficial once the appropriate ethical conceptions are made computable and implementable (de Laat 2021). This conception of technological solutionism is intertwined with the belief in technological determinism, i.e., technology viewed as determining social outcomes and evolving autonomously (Green 2021b).

As Green (2021b) notes, viewing technology through a socio-technical lens reveals the inherent concealment of normative disagreement and implicit value judgements contained within. Thus, guiding technological progress towards beneficial socio-technical outcomes requires acknowledging the fact, that ethics is not objective or neutral and, that adopting ethics doesn’t in and of itself prompt substantive change (Green 2021b). These challenges correspond to criticisms by Mittelstadt (2019) analyzed in chapter 5.2 on the discourse on *AI ethics principles*.

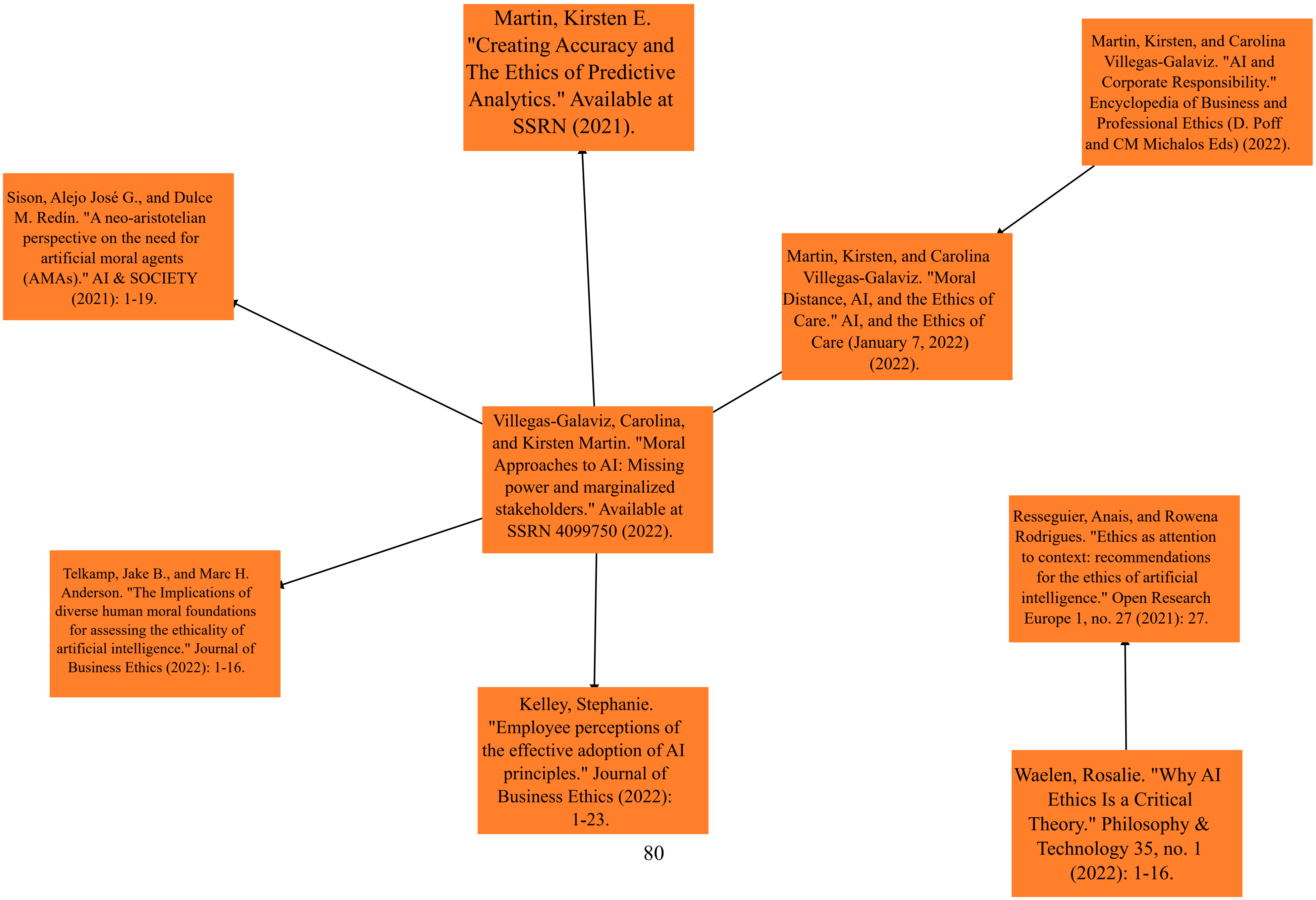
Nevertheless, attempts at meeting the challenges of *responsible AI* through guidelines, brochures, checklists, and software implementations serve as positive first steps (de Laat 2021). As van Maanen (2022) notes, ethical principles and guidelines advance the gradual identification of a conception of good practice. This needs to be complemented with appropriate policy and regulation, as an organization's adherence to ethical norms based on commitments to guidelines isn't sufficient by itself.

> Paul B. De Laat's thorough attempt to diffuse the ethics washing critique based on commercial actor's self-published material illustrates this (de Laat, 2021). As he explains himself, it is not possible to say much about companies' actual adherence to ethical norms and their implementation based on material found online. (van Maanen 2022)

*Responsible AI* development faces additional issues of human oversight and control, as the growth in complexity of technologies coincides with less human intervention in decision-making processes (Constantinescu, et al. 2021). The issues of responsibility are compounded by the involvement of multiple actors in the development, regulation, and use of AI systems, leading to the obscurement of the attribution of responsibility for AI related outcomes (Constantinescu, et al. 2021).

## 5.9 Moral Approaches to AI

The discourse on *moral approaches to AI* is depicted in Figure 13. As the visualization indicates, the discourse on *moral approaches to AI*, as seen in the collected data, is relatively limited. Despite containing only a handful of articles, the perspectives represented in the collected data are diverse, ranging from virtue ethics to ethics of care. The overarching thematic overlap between topics consists in their critique of a principle-based approach towards AI ethics. Additionally, represented in the collected data is a call for consideration of contextual factors and asymmetries in systemic power relations, which are not effectively represented in a principle-based approach.



Martin, Kirsten E. "Creating Accuracy and The Ethics of Predictive Analytics." Available at SSRN (2021).

Martin, Kirsten, and Carolina Villegas-Galaviz. "AI and Corporate Responsibility." Encyclopedia of Business and Professional Ethics (D. Poff and CM Michalos Eds) (2022).

Sison, Alejo José G., and Dulce M. Redín. "A neo-aristotelian perspective on the need for artificial moral agents (AMAs)." AI & SOCIETY (2021): 1-19.

Martin, Kirsten, and Carolina Villegas-Galaviz. "Moral Distance, AI, and the Ethics of Care." AI, and the Ethics of Care (January 7, 2022) (2022).

Villegas-Galaviz, Carolina, and Kirsten Martin. "Moral Approaches to AI: Missing power and marginalized stakeholders." Available at SSRN 4099750 (2022).

Resseguier, Anais, and Rowena Rodrigues. "Ethics as attention to context: recommendations for the ethics of artificial intelligence." Open Research Europe 1, no. 27 (2021): 27.

Telkamp, Jake B., and Marc H. Anderson. "The Implications of diverse human moral foundations for assessing the ethicality of artificial intelligence." Journal of Business Ethics (2022): 1-16.

Kelley, Stephanie. "Employee perceptions of the effective adoption of AI principles." Journal of Business Ethics (2022): 1-23.

Waelen, Rosalie. "Why AI Ethics Is a Critical Theory." Philosophy & Technology 35, no. 1 (2022): 1-16.

Figure 13 Moral Approaches to AI Discourse

AI ethics has been predominantly tackled through a principle-based approach as discussed in chapter 5.2 in the discourse on *AI ethics principles*. Instilled within these principles are the societal ideals and values requisite for AI systems to be considered beneficial and ethical. The principle-based approach to AI ethics envisions principles in a way that is reminiscent of duties in deontological ethics (Villegas-Galaviz and Martin 2022a). In this normative ethical approach, the moral rightness of actions depends on their conformity with an agent's obligations (Villegas-Galaviz and Martin 2022a). From this perspective, guidelines present themselves as opportune mechanisms to establish ethical principles and address relevant duties in novel scenarios related to AI ethics.

> In AI ethics, the first impulse has been towards the search for principles to guide developers and users in unknown terrain. Several moral guidelines on AI have been proposed in the search for moral principles that guide machine ethics. (Villegas-Galaviz and Martin 2022a)

Principles appear to be necessary for ethics, they are however not sufficient in and of themselves to guarantee ethical AI (Villegas-Galaviz and Martin 2022a). This has led some researchers to shift their focus from duties and rules to individuals, by advocating for a virtue ethical approach to AI ethics. The discourse pertaining to virtue ethical approaches identified in the collected data, focused on a neo-Aristotelian perspective as presented by Sison and Redín (2021).

> Most of these applications refer to a neo-Aristotelian approach, where neo indicates the resolved variety of virtue ethics that rejects Aristoteles's views on women and slavery, as well as children, vulnerabilities, and dependence (Sison and Redín, 2021). (Villegas-Galaviz and Martin 2022a)

This depiction of virtue ethics advances the project of changing attitudes, cultivating personalities, strengthening responsibilities, and refraining from actions deemed unethical. As Villegas-Galaviz and Martin (2022a) argue, a virtue ethical approach can in certain instances be complemented with the principles of a deontological account of ethics, as the reliance on goodwill and virtue of character might be insufficient. However, the virtue ethical approach contains inherent flaws, that need to be overcome.

> Nevertheless, the approach applied in isolation may encounter some limitations. First, "conceptions of virtue and human flourishing are never universal. There have always been, and will always be, coherent accounts of the good life that cannot be reduced to or fully reconciled with other" (Vallor, 2017). (Villegas-Galaviz and Martin 2022a)

The unforeseen impacts of AI systems necessitate the establishment of frameworks that resolve issues of responsibility and accountability when mistakes and harms occur. Normative approaches centered on these aspects navigate the allocation of individual responsibility in collective settings (Villegas-Galaviz and Martin 2022a). As technology is inherently value-laden, AI systems can exacerbate issues of fairness and justice (Martin 2021). Moral approaches centering on these ethical aspects of AI face the problem of attempting to achieve consensus on the concept of fairness and justice in differing contexts. Critical theories emphasize such asymmetries in systemic power relations and attempt to contribute to structural changes (Martin 2021).

Ethics of care adopts this perspective, as interdependent relationships and their effect on ethical decision-making is stressed. In this approach, marginalized groups, people's vulnerabilities, and protection against harms are brought to the foreground (Villegas-Galaviz and Martin 2022b). In the context of AI, ethical consideration should be placed on a system's effects on ensuring individuals meet their needs and, that the vulnerabilities of individuals are not exploited. This approach calls attention towards particular situations in favor of the detached perspective prescribed in the principle-based approach (Resseguier and Rodrigues 2021).

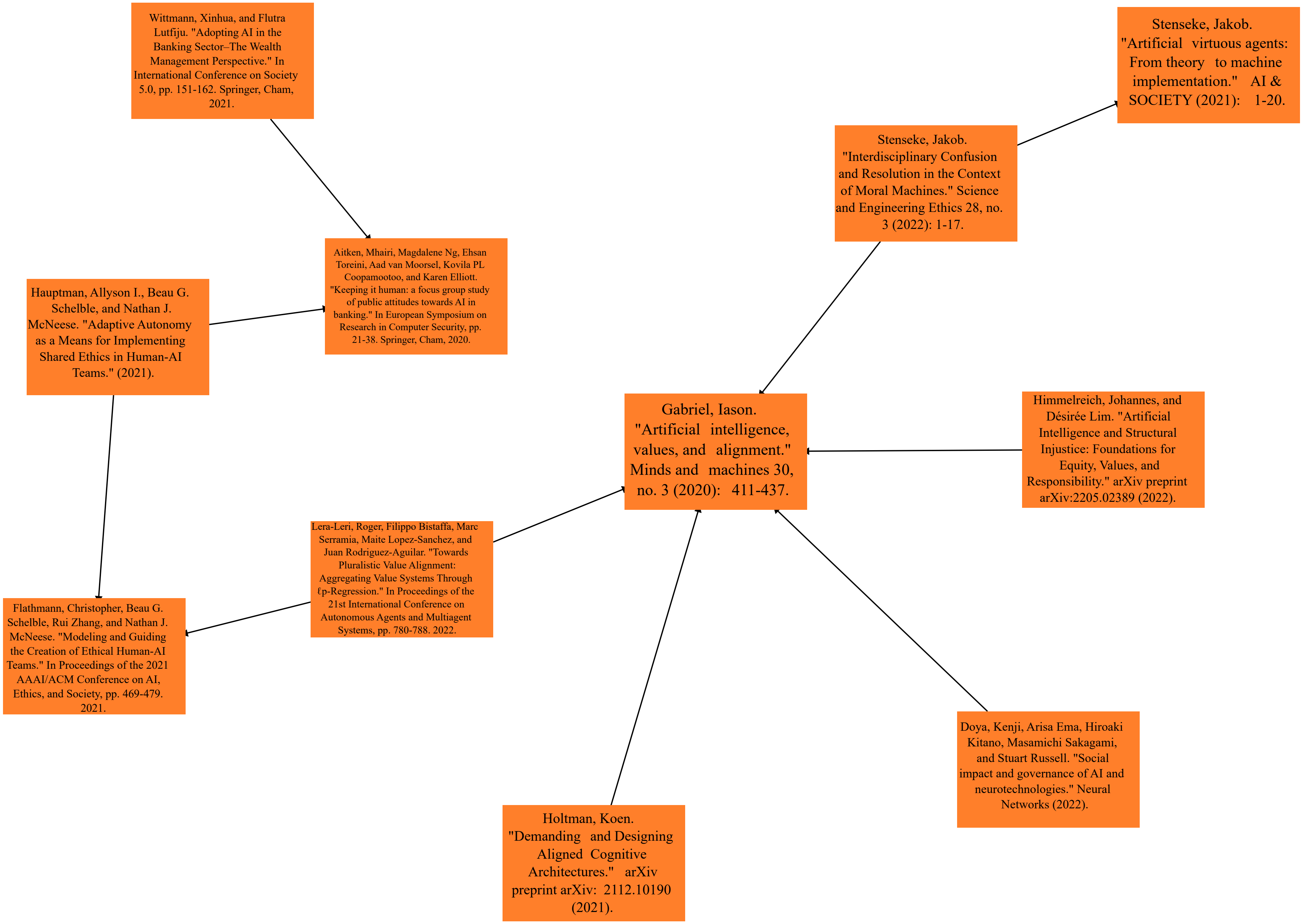
> The ethics of care has been proposed in situations where the interests of the least advantaged stakeholders are not being considered. In other words, where the distance between those making the decisions and those impacted by the decisions is too great and the marginalized stakeholder's interest are not being seen or judged to be legitimate. (Villegas-Galaviz and Martin 2022b)

In this regard, the ethics of care views empathy, emotions, context, relationships, and vulnerabilities as appropriate guides to ethical decision-making (Villegas-Galaviz and Martin 2022b). As Resseguier and Rodrigues (2021) indicate, in an effort to address the negative material implications and impacts of AI, the ethics of care advocates for a shift towards the consideration of concrete practices and the socio-political context and materialities related to AI. This approach advocates for a notion of care, viewed as contextualized responsiveness to others, superseding the mere delineation of rules or calculated consequences (Villegas-Galaviz and Martin 2022b). In this sense care is seen as a relational approach to morality. The crux of this approach is balancing between a focus on the disadvantages of AI and ethical issues unrelated to vulnerabilities and harms (Villegas-Galaviz and Martin 2022b). Thus, an ethics of care approach needs to be complemented by other approaches.

In a similar vein to ethics of care, Waelen (2022) characterizes AI ethics as embodying concerns for human emancipation and empowerment, resembling a critical theory. According to Waelen (2022), reframing transparency, privacy, freedom, and autonomy as important value issues, as a consequence of their relation to empowerment, reveals the inherent critical nature of AI ethics. From this perspective principles of trust, justice, responsibility, and non-maleficence are indispensable in virtue of their role in protecting individuals from the negative consequences of AI (Waelen 2022). Waelen (2022) argues, that this position is further justified, when the aspiration towards social change is acknowledged as an intrinsic quality of AI ethics. Consequently, AI ethics is not merely interested in analyzing or diagnosing society, but actively attempts to change it. This reframing resolves the issue of abstraction commonly attributed to AI ethics, by creating a common language through the perspective of relational or dispositional power, serving as a tangible target for change (Waelen 2022). Furthermore, viewing AI ethics as a critical theory provides a new method of analysis in identifying ethical issues, which addresses criticisms of neglect towards social and political impacts (Waelen 2022). The approach advocated for by Waelen (2022), remedies the commonly emphasized issues in AI ethics of vagueness and abstractness, as well as the disregarded of negative impacts and externalities.

## 5.10 Value Alignment

Aligning AI systems with human values has become a significant issue in AI ethics literature, as the autonomy and capabilities of AI systems have escalated. The discourse on *value alignment* is depicted in Figure 14. As the visualization indicates, the discourse on *value alignment* contains discussion from appropriate values to proposals on frameworks. The discourse on *value alignment*, as seen in the collected data, is two-sided, containing both normative and technical topics. On the normative side, determining the values or principles artificial agents should adhere to is a primary concern, as discussed by Gabriel (2020). On the technical side, issues regarding the formal encoding of values and principles in artificial agents are pivotal, as discussed by Stenseke (2021).



Wittmann, Xinhua, and Flutra Lutfiju. "Adopting AI in the Banking Sector–The Wealth Management Perspective." In International Conference on Society 5.0, pp. 151-162. Springer, Cham, 2021.

Stenseke, Jakob. "Artificial virtuous agents: From theory to machine implementation." AI & SOCIETY (2021): 1-20.

Stenseke, Jakob. "Interdisciplinary Confusion and Resolution in the Context of Moral Machines." Science and Engineering Ethics 28, no. 3 (2022): 1-17.

Aitken, Mhairi, Magdalene Ng, Ehsan Toreini, Aad van Moorsel, Kovila PL Coopamootoo, and Karen Elliott. "Keeping it human: a focus group study of public attitudes towards AI in banking." In European Symposium on Research in Computer Security, pp. 21-38. Springer, Cham, 2020.

Hauptman, Allyson I., Beau G. Schelble, and Nathan J. McNeese. "Adaptive Autonomy as a Means for Implementing Shared Ethics in Human-AI Teams." (2021).

Gabriel, Iason. "Artificial intelligence, values, and alignment." Minds and machines 30, no. 3 (2020): 411-437.

Himmelreich, Johannes, and Désirée Lim. "Artificial Intelligence and Structural Injustice: Foundations for Equity, Values, and Responsibility." arXiv preprint arXiv:2205.02389 (2022).

Lera-Leri, Roger, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar. "Towards Pluralistic Value Alignment: Aggregating Value Systems Through ℓp-Regression." In Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems, pp. 780-788. 2022.

Flathmann, Christopher, Beau G. Schelble, Rui Zhang, and Nathan J. McNeese. "Modeling and Guiding the Creation of Ethical Human-AI Teams." In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 469-479. 2021.

Doya, Kenji, Arisa Ema, Hiroaki Kitano, Masamichi Sakagami, and Stuart Russell. "Social impact and governance of AI and neurotechnologies." Neural Networks (2022).

Holtman, Koen. "Demanding and Designing Aligned Cognitive Architectures." arXiv preprint arXiv: 2112.10190 (2021).

Figure 14 Value Alignment Discourse

Establishing a consensus on AI value alignment is restrained by widespread variation in moral beliefs (Gabriel 2020). Furthermore, the issue is exacerbated by issues pertaining to what kind of values AI should align with and who the appropriate moral authority should be. A popular criticism towards the technical side of the *value alignment* issue relates to the tendency to reduce complex socio-technical issues to optimization problems (Stenseke 2022). As Stenseke (2022) notes, normative frameworks, such as deontology and consequentialism have seen popular demand, as they conveniently correspond to software development practices.

> While deontology conveniently corresponds to conditional statements that drives software programming (e.g., "If X→do Y"), consequentialism's emphasis on quantifiable utility elegantly resonates with reward-functions of reinforcement learning and objective functions in mathematical optimization. (Stenseke 2022)

The limitation of deontological and consequentialist approaches lies in their simplification of values, concepts, and theories to correspond with well-defined computational settings in functional applications (Stenseke 2022). Thus, the broader aspects of moral behavior and cognition are disregarded. Because of these shortcomings Stenseke (2021) argues for a virtue-oriented approach based on moral exemplars towards solving the value alignment problem. As Flathmann, et al. (2021) indicate, designing AI systems that align their values with humans by learning from human behavior could be a constructive approach towards value alignment. These interactions would embody ethical principles in a practical manner and continually orient AI to reflecting on the changes in ethical expectations and constraints (Hauptman, Schelble and McNeese 2021).

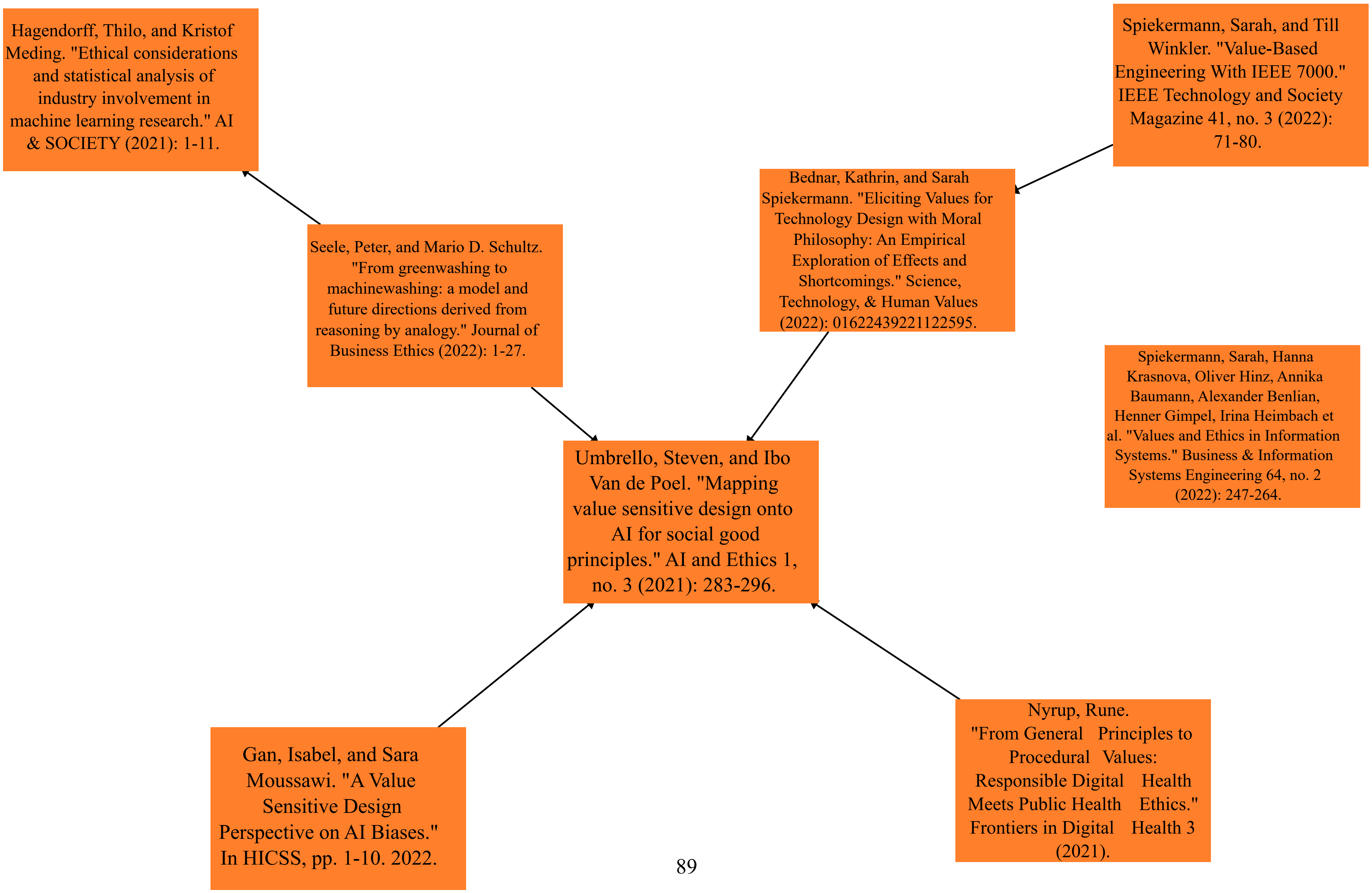
Doya, et al. (2022) argue, that revealed preferences and behavior are not good metrics for evaluation, as it often does not reflect the underlying preferences, since humans are not perfectly rational. Furthermore, Gabriel (2020) argues, that making reliable inferences from behavior is difficult, since at any moment several reward functions for agents exist, that can be considered optimal for the observed behavior. Additionally, such an approach encounters issues in ascertaining preferences for situations seldom encountered, which could nevertheless be morally relevant, such as emergencies. According to Gabriel (2020), the approach also doesn't account for revealed preferences that can be considered harmful or immoral.

Adopting a value-based approach to AI alignment requires elucidating values or principles for agent adherence. This endeavor has commonly been furthered through the establishment of principles of justice endorsed by particular societies. As Gabriel (2020) however notes, conceptions of justice, that are global in nature might be required, as determining AI alignment based on local notions of justice might lead to the endorsement of cultural relativism. Additionally, restricting the scope of AI alignment to local principles of justice, that are considered benevolent, doesn't entail global congruence (Gabriel 2020). Furthermore, advanced AI will evidently become a global technology, rendering the approach based on local notions of justice unfeasible. Thus, Gabriel (2020) advocates the development and design of AI aligned to principles reflecting a global consensus, as seen in for instance the doctrine of universal human rights.

Designing AI systems to align with any set of values poses several difficulties. As Lera-Leri, et al. (2022) state, a variety of reasonable beliefs about values exist, that are nevertheless conflicting and sometimes contradictory. Thus, appropriately aligning AI represents a pluralistic value alignment problem. Himmelreich and Désirée (2022) indicate, that this pluralism could be accommodated, if a set of principles denoting a consensus could be reached. Gabriel (2020) suggests, that the literature on social choice theory could provide mechanisms towards reaching a collective decision. Alternatively, a consensus on ethical principles as established by Jobin, Ienca and Vayena (2019), could serve as a preliminary basis. According to Holtman (2021) alignment in this sense would preferably be guided by a process of informed debate incorporating representatives of all affected human stakeholders towards building a consensus (Holtman 2021).

## 5.11 Value Sensitive Design

The discourse on *value sensitive design* is depicted in Figure 15. As can be seen in Figure 15, the discourse on *value sensitive design* is comparatively small, containing only a handful of articles. The discourse, as seen in the collected data, encompasses both technical and social aspects. On the technical side, Umbrello and Van de Poel (2021) propose, a framework for value sensitive design, that incorporates ethical principles into design as system requirements. On the social side, Gan and Moussawi (2022) argue for a shared understanding of relevant values amongst all stakeholders, as a necessary basis for effective design. The thematic convergence amongst both perspectives is in their endorsement of stakeholder participation in the value discovery process and communication of value trade-offs.



Hagendorff, Thilo, and Kristof Meding. "Ethical considerations and statistical analysis of industry involvement in machine learning research." AI & SOCIETY (2021): 1-11.

Spiekermann, Sarah, and Till Winkler. "Value-Based Engineering With IEEE 7000." IEEE Technology and Society Magazine 41, no. 3 (2022): 71-80.

Bednar, Kathrin, and Sarah Spiekermann. "Eliciting Values for Technology Design with Moral Philosophy: An Empirical Exploration of Effects and Shortcomings." Science, Technology, & Human Values (2022): 01622439221122595.

Seele, Peter, and Mario D. Schultz. "From greenwashing to machinewashing: a model and future directions derived from reasoning by analogy." Journal of Business Ethics (2022): 1-27.

Spiekermann, Sarah, Hanna Krasnova, Oliver Hinz, Annika Baumann, Alexander Benlian, Henner Gimpel, Irina Heimbach et al. "Values and Ethics in Information Systems." Business & Information Systems Engineering 64, no. 2 (2022): 247-264.

Umbrello, Steven, and Ibo Van de Poel. "Mapping value sensitive design onto AI for social good principles." AI and Ethics 1, no. 3 (2021): 283-296.

Gan, Isabel, and Sara Moussawi. "A Value Sensitive Design Perspective on AI Biases." In HICSS, pp. 1-10. 2022.

Nyrup, Rune. "From General Principles to Procedural Values: Responsible Digital Health Meets Public Health Ethics." Frontiers in Digital Health 3 (2021).

Figure 15 Value Sensitive Design Discourse

Value sensitive design is a method for integrating values into technical design, established as a methodology consisting of empirical, conceptual, and technical investigations (Umbrello and van de Poel 2021). These investigations comprise determination of relevant stakeholders, elicitation of stakeholder values, establishing value trade-offs, value issues in current technologies, and value implementation in dew designs (Umbrello and van de Poel 2021). For value sensitive design to be effective, developers and users need a shared understanding of relevant values, in addition to metrics and standards specifically designed for AI technologies shared by all stakeholders (Gan and Moussawi 2022).

Umbrello and van de Poel (2021) propose an approach to value sensitive design in AI contexts, where ethical principles are integrated into value sensitive design as design norms, that enable the derivation of more specific design requirements. Incorporated in this way, ethical principles would have an impact on design and avoid being abstract and inconsequential to developers and practitioners. According to Umbrello and van de Poel (2021), the value sensitive design process should be extended to encompass the AI technologies entire life cycle, so that unintended value consequences could be monitored. This approach could potentially mitigate the negative impacts and externalities of AI systems, that according to Hagendorff (2022b) are commonly disregarded.

According to Umbrello and van de Poel (2021), contextual variables determine how values are rationalized and stakeholders understand them. Thus, eliciting stakeholder values is imperative to ascertaining whether design conforms and fulfills stakeholder expectations both directly and indirectly. Bednar and Spiekermann (2022) emphasize the importance of contextual factors in the value discovery process. By shifting the focus from general values towards marginal cases, contextual value nuances not captured in top-down value list approaches are represented, resulting in the establishment of a more complete spectrum of system requirements (Bednar and Spiekermann 2022). Bednar and Spiekermann (2022) argue, that bottom-up value elicitation processes that include stakeholders can be complemented with top-down approaches incorporating high-level values and principles, as seen in the approach advocate for by Umbrello and van de Poel (2021). According to Bednar and Spiekermann (2022), this synthesis should be accomplished at the conceptual and empirical levels in the value sensitive design tradition.

Nyrup (2021) however notes, that a value sensitive design approach, as described by Umbrello and van de Poel (2021), lacks a method for resolving value trade-offs. This criticism is common to value sensitive design approaches and requires mechanisms or procedures to assist in ascertaining the right kind of procedural values in decision-making. As Nyrup (2021) argues, the accountability for reasonableness (A4R) approach overcomes this issue.

> Proposed by Norman Daniels and James Sabin, the key idea in A4R is to implement decision procedures for reaching compromises which fair-minded people can accept as legitimate, despite their underlying ethical disagreements. This relies on a distinction between ethical rightness and ethical legitimacy. To regard a decision as right is to regard it as the morally correct thing to do in a given situation. To regard it as legitimate is to regard it as appropriately made, i.e., by a decision-maker or procedure whose moral authority to make such decisions should be accepted. (Nyrup 2021)

This approach articulates standards and mechanisms for reaching consensus in deliberation about appropriate value trade-offs and holding decision-makers accountable, thereby empowering stakeholders.

# 6 DISCUSSION

This chapter summarizes the attained results in this thesis and examines similarities and discrepancies to previous meta-analyses with a critical reflection. Additionally, the chapter covers the novel contributions of this thesis to the research field. Section 6.1 presents the theoretical and practical implications of the attained results. Section 6.2 discusses the limitations of the data collection and analysis performed in this thesis.

## 6.1 Implications of Results

The primary theoretical implication of this thesis is the mapping of discourses, topics and themes presented and analyzed in Chapter 5. The discourses in AI ethics are varied, encompassing a multitude of topics, containing both convergent and divergent themes, as Figure 5 presented in section 5.1 indicates. This mapping corresponds to results attained in previous meta-analyses of the AI ethics field of research. However, some notable distinctions are seen in the mapping for this thesis, which result in novel contributions to the research area.

AI ethics is an emerging field of research, that has in recent years received a considerable amount of attention (Borenstein, et al. 2021). Research articles mapping the current state of AI ethics discourses are scarce. Given this lack of previous research, that would provide a benchmark to evaluate the results of this thesis, reflection is challenging. However, fortunately research analyzing some of the prevalent topics and blind spots in AI ethics has been published. The mapping of discourses, topics and themes in this thesis partially corresponds to the categorization of keywords and categories in AI ethics literature by Vakkuri and Abrahamsson (2018). Commonalities include topics pertaining to the different branches of ethics and their application in the context of AI, autonomous agents and moral agency, law and regulation, AI risks, and value alignment.

Some of the blindspots highlighted by Hagendorff (2022b), were observable in the mapping of AI ethics discourses, topics and themes conducted in this thesis. Most notably issues related to negative externalities of AI systems were largely absent. However, as seen particularly in the *ethics auditing* discourse, mitigating negative impacts throughout the life cycle of AI systems is seen as a central issue. Additionally, as seen in the *moral approaches to AI* discourse, calls for greater consideration of the needs of individuals have become increasingly more common. Conversely, the key blindspot in AI ethics of considering the suffering and harms connected to AI systems emphasized by Hagendorff (2022b), was present in multiple discourses analyzed in this thesis. This was especially seen in the discourse on alternative ethical approaches to a principle-based approach, which in the collected data for this thesis comprised of virtue ethics, ethics of care and critical theory. This discourse contained several themes raising concerns related to protecting the needs and vulnerabilities of individuals as well as emphasizing a greater consideration of social change and empowerment. This implies a shift towards more inclusive and wide-ranging discourse and consideration of the needs of individuals and society.

The aforementioned shift can be explained by progress in the discourse on *AI ethics principles*, which has developed significantly in recent years. Mittelstadt's (2019) critique, provided impetus for alternative approaches to emerge and caused a shift towards the operationalization of AI ethics as established in section 5.2. As seen in Munn's (2022) critique, the discourse is prevalent and ethical principles are still debated. However, in the discourse on *moral approaches to AI* analyzed in section 5.9 alternative approaches towards ethical AI are examined. In the collected data the prominent proposals advocate for either a virtue-oriented approach, or an ethics of care approach, since both can represent disparate ethical features, while being complementary to a principle-based approach. The ethics of care approach emphasizes stakeholder needs and consideration of stakeholder values, which corresponds to calls for stakeholder participation seen in the *trustworthy AI* and *responsible AI* discourses. The discourse on *AI ethics principles* is connected to several other discourses in the collected data for this thesis.

The commonly seen ethical principles of trustworthiness and responsibility in AI development were identified in the collected data for this thesis as separate discourses. These match previously attained results by Jobin, Ienca and Vayena (2019), who identified trust and responsibility amongst others as overarching ethical values and principles commonly seen in the AI ethics literature. The separation from the *AI ethics principles* discourse and the formation of distinct and separate discourses for *trustworthy AI* and *responsible AI* implies that the discourse has progressed from a focus on theoretical contemplations and a focus on principles to debating mechanisms and frameworks that facilitate safe and responsible development processes. This assertion is substantiated by the analysis in sections 5.7 and 5.8. However, the mechanisms by which individuals determine trustworthiness and effective processes to communicate the trustworthiness of AI systems, are largely neglected in the *trustworthy AI* and *responsible AI* discourses identified in this thesis. These results match previous findings by Liao and Sundar (2022). Establishing such mechanisms requires further research.

Conversely, while fairness was seen as a central issue in multiple discourses, which is in line with both Jobin, Ienca and Vayena (2019) as well as Hagendorff (2022b), transparency and privacy were not represented in the collected data for this thesis as central issues and were not identified as separate discourses. However, transparency and explainability were represented as important topics in the *ethics auditing* discourse. Thus, while not forming a separate and distinct discourse in the collected data for this thesis, transparency is an important topic in AI ethics.

In comparison to the hitherto discussed, mostly theoretical contributions, the results attained in the following paragraphs are essential to practitioners alike. The *AI ethics praxis* discourse, as indicated by the collected data, has progressed significantly in recent years. The recent publications in the collected data for this thesis focused on empirically validating the proposed tools, methods, and frameworks in the literature. This represents a notable step forward and addresses the central criticisms of being abstract and not actionable that have been issued towards AI ethics as seen in for instance Mittelstadt (2019), as well as Morley, Floridi, et al. (2020). Additionally, discourse pertaining to organizational responses in addressing the ethical issues related to AI in practice were prominent. This represents a step forward and might potentially address the issue of lacking or mixed maturity in tackling AI ethics in practice on an organizational level, previously identified by Vakkuri, et al. (2020a).

However, seen especially in the *responsible AI* discourse is a critical undertone towards the predisposition to technological solutionism and resolving complex social issues through technical means. This criticism is intertwined with a criticism of technological determinism and the view of technology as autonomously evolving. These perspectives inherently conceal implicit value judgements and hide normative disagreements within technology development. Focusing on design processes and placing the burden of decision making on individual technical experts, evades questions of organizational responsibility and the larger context of decision-making. This corresponds to recent criticisms, as seen in for instance Schwartz, et al (2022).

## 6.2 Limitations

The data collection for this thesis was completed using a snowball approach with Mittelstadt's (2019) article serving as the start set. One iteration of both forward and backward snowballing was performed, which resulted in the inclusion of 384 research articles and papers. These were then further pruned using the established eligibility criteria. As the eligibility criteria seen in Table 1 indicate, the data collected for this thesis is restricted and limited in multiple ways. Firstly, research published in languages other than English was excluded. This resulted in the exclusion of a non-negligible amount of research. Secondly, partially related to the first limitation is the representation of different perspectives. The collected literature is overwhelmingly representative of western perspectives and research implemented at academic institutes. Thus, while arguably being representational of academic contexts and discourse related to AI ethics, the results of this thesis might not necessarily apply and extend to discourses and topics prevalent in other contexts. Thirdly, the cutoff point for relevant literature was set as 2019, which further restricted the number of research publications. Fourthly, the researcher conducting the data collection for this thesis was restricted by access to some of the publications that would have provided relevant data for this thesis.

Evidently the research process and methodology applied in this thesis was non-exhaustive and represents a limited overview of the research literature. As such, establishing the reliability and validity of attained results was approached through careful argumentation and documentation of the process that was followed. Qualitative research and especially discourse analysis is interpretive by design. According to Yazdannik, Yousefy and Mohammadi (2017), evaluating the methodological and interpretative rigor of research employing discourse analysis is of paramount importance. Methodological rigor requires that the research question be appropriately formulated and specified, to fit the research method used (Yazdannik, Yousefy and Mohammadi 2017). In this thesis the objective was to research the prominent discourses in AI ethics and establish the prevalent topics and themes. As such, discourse analysis was deemed a prudent methodological approach to researching these issues.

Additionally, the chosen research question must be suitable for the corpus of texts. This corpus should represent the research area studied or alternatively the shortcomings should be clearly explicated in case the representation is lacking in some form (Yazdannik, Yousefy and Mohammadi 2017). As discussed in this chapter the literature collected for this thesis encompassed primarily research articles and papers published in academic contexts. Thus, while containing certain limitations, the collected data is arguably representative of the research area studied in this thesis. Furthermore, a clear description of the interpretative paradigm is required, as well as of the data-gathering and data analysis methods, so that readers can clearly follow along with the process (Yazdannik, Yousefy and Mohammadi 2017). Clear and comprehensive documentation was set as a priority early on in this thesis. Thus, the data collection and analysis steps and their justifications were explicated and documented in detail in chapters 4 and 5.

Interpretative rigor mandates that the findings and their connection to discourse be adequately described using verbatim text in support (Yazdannik, Yousefy and Mohammadi 2017). This was accomplished in chapter 5 through pertinent in-line citations. Additionally, all findings should be explainable and plausible. Verification of these findings should be grounded in previous studies and existing knowledge (Yazdannik, Yousefy and Mohammadi 2017). Chapter 6 grounded the findings in this thesis to previous studies and research, emphasizing results that corroborated previous findings, while simultaneously highlighting novel results. Admittedly, corroboration of the results and findings in this thesis proved to be challenging, since research publications related to meta-analyses and overviews of AI ethics discourses were scarce. However, through careful analysis and reflection similarities and discrepancies to previous research could be established, with several novel findings being established in the discussion section in section 6.1 that were substantiated through analysis in chapter 5.

# 7 CONCLUSION

This chapter concludes the thesis with an overview of the thesis and results, while providing pathways for complementary and further research. Section 7.1 provides an overview of the thesis starting with the research questions and summarizing the attained results. Chapter 7.2 proposes avenues and opportunities for further research, that were noticed during the implementation of this thesis. Additionally, the chapter covers topics that were excluded from this thesis, as they were outside of scope and weren't feasible to research in the allotted time frame for this thesis.

## 7.1 Answers to Research Questions

The primary objective of this thesis was to research the prominent discourses in AI ethics and establish what the prevalent topics and themes are and how they relate to each other. The secondary objective for this thesis was to ascertain whether these discourses constitute a coherent field of research with a meaningful amount of interaction or diverge into separate topics of interest. This entailed studying the discourses and the interconnectivity between them with the aim of revealing whether research has progressed in a meaningful way or whether AI ethics is stuck in a spiral of reinventing the wheel. The research questions for this thesis were:

- What are the current discourses in AI ethics?
- What are the prevalent themes and topics?
- How do they relate to each other?

The answers to the first two research questions are presented in Figure 5. The central discourses identified in the collected data relate to *AI ethics principles*, *AI governance*, *AI ethics praxis*, *ethics auditing*, *embedded ethics*, *trustworthy AI*, *responsible AI*, *moral approaches to AI*, *value alignment*, and *value sensitive design*. As the visualization indicates the research field of AI ethics is vast with multiple overlapping topics and themes. The topics and themes identified in the collected data are inordinately numerous to enumerate and are reasonably intelligibly presented in Figure 5.

As established in Chapter 5, the prominent discourses in AI ethics encompass some overarching focal points. Several of the discourses identified in this thesis contain similar topics and themes with significant overlap, as seen in Figure 5. Conceptualizing this interconnectedness is one of the upshots of a focus on discourses, which revealed the nature and amount of overlap. As seen in Figure 4 however, the discourses are largely isolated with publications within discourse primarily referencing and citing other publications in their respective clusters.

This is one of the interesting findings as the apparent lack of referencing and citing towards publications outside of a given discourse, that nevertheless concern similar topics and themes, represents a challenge that needs to be overcome. This is exemplified in the topics of calling for stakeholder participation and managing the roles and responsibilities of different social actors, which are acclaimed throughout several discourses as critical issues. The importance of stakeholder participation is highlighted in the discourses on *AI governance*, *value sensitive design*, and *ethics auditing*. Nevertheless, as seen in Figure 4, these discourses are only connected in a rudimentary manner with a few scattered references and citations. Thus, progress in research and development of best practices are not necessarily shared throughout the research field as a whole.

In a similar sense, topics pertaining to the management of roles and responsibilities are seen in *AI governance*, *AI ethics praxis*, *responsible AI*, and *moral approaches to AI* discourses. However, establishing a consensus or current state of research on these topics is challenging, as research is progressing on multiple fronts and spread out through multiple discourses. References and citations would imply only a superficial connection between these discourses, however when analyzing the central topics and themes within these discourses, considerably stronger connections were established. Thus, the same issues apply as in the topic of stakeholder participation.

## 7.2 Future Research

Several avenues for future research were noticed during the implementation of this thesis. The topics in this thesis were analyzed using discourse analysis from a habermasian perspective. Further research is needed from the perspective of foucauldian discourse analysis. This would entail researching questions pertaining to different voices and participants that shape the discourse on AI ethics. The roles of different social actors and the imbalance in power relations was partially analyzed in this thesis and the effect of experts in shaping discourse on AI ethics issues was noticed. However, these topics require an in-depth analysis that was not feasible to conduct within the framework of this thesis. Questions of ample representation, seats at the table, the use of power and influence in guiding and shaping AI ethics discourse could serve as interesting future research topics.

Furthermore, Critical Discourse Analysis (CDA) could be used as an approach to complement the findings of this thesis. CDA was excluded as it was outside of the interest and scope of this thesis. The topics discussed could however be analyzed more comprehensively from a critical perspective, with a focus on how social power, its misuse, dominance, and inequality relate to the topics discussed in a social and societal context. In this vein chapter 5.9 analyzed issues in AI ethics through the lens of ethics of care and critical theory, but this was evidently done in a superficial manner, as these approaches only formed parts of the *moral approaches to AI* discourse.

Additional topics that were alluded to during this thesis but were not covered extensively include the relationship between media sources, their portrayal of AI and its impact on the adoption and design of AI systems and technology to a larger extent. Furthermore, the ways in which public discourse shapes new technologies, for instance AI development, could be studied in a societal context.

# BIBLIOGRAPHY

Aastha, Pant, Rashina Hoda, Chakkrit Tantithamthavorn, and Burak Turhan. "Ethics in AI through the Developer's Prism: A Socio-Technical Grounded Theory Literature Review and Guidelines." 2022.

Agbese, Mamia, et al. "Governance of Ethical and Trustworthy Al Systems: Research Gaps in the ECCOLA Method." *IEEE 29th International Requirements Engineering Conference Workshops.* Indiana: IEEE, 2021. 224-229.

Agha, Asif. "Introduction: semiosis across encounters." *Journal of Linguistic Anthropology*, 2005: 1-5.

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion*, 2020: 82-115.

Association for Computing Machinery, US Public Policy Council (USACM). *ACM.* January 12, 2017. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf (accessed May 14, 2023).

Baker, Paul, and Sibonile Ellece. *Key Terms in Discourse Analysis.* New York, N.Y.: Continuum, 2011.

Bakhtin, Mikhail. *Speech Genres and Other Late Essays.* Austin: University of Austin Press, 1986.

Bazerman, Charles. "How Natural Philosophers can Cooperate: The Rhetorical Technology of Coordinated Research in Joseph Priestley's History and Present State of Electricity." In *Textual Dynamics of the Professions*, by Charles Bazerman and James Paradis, 13-44. University of Wisconsin Press, 1991.

Bazerman, Charles. "Intertextual Self-Fashioning: Gould and Lewontin's Representations of the Literature." In *Constructing Experience*, by Charles Bazerman, 194-214. Carbondale and Edwardsville: Southern Illinois University Press, 1994.

Beauchamp, Tom L., and James F. Childress. *Principles of Biomedical Ethics, 8th Edition.* New York: Oxford University Press, 2013.

Bednar, Kathrin, and Sarah Spiekermann. "Eliciting Values for Technology Design with Moral Philosophy: An Empirical Exploration of Effects and Shortcomings." *Science, Technology, & Human Values*, 2022.

Bélisle-Pipon, Jean-Christophe, Erica Monteferrante, Marie-Christine Roy, and Vincent Couture. "Artificial intelligence ethics has a black box problem." *AI and Society*, 2022.

Bezuidenhout, Louise, and Emanuele Ratti. "What does it mean to embed ethics in data science? An integrative approach based on microethics and virtues." *AI and Society*, 2021: 939–953.

Bhatnagar, Sankalp, et al. "Mapping Intelligence: Requirements and Possibilities." *3rd Conference on Philosophy and Theory of Artificial Intelligence.* Springer, Cham, 2018. 117-135.

Bietti, Elettra. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy." *Proceedings of the 2020 Conference on Conference on Fairness, Accountability, and Transparency.* New York: Association for Computing Machinery, 2020. 210–219.

Bijker, Wiebe E., and Trevor Pinch. "The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other." In *The Social Construction of Technological Systems: New Direction in the Sociology of Technology*, by Wiebe E. Bijker, Thomas Parke Hughes and Trevor Pinch, 17-50. Cambridge, Massachusetts: MIT Press, 1987.

Borenstein, Jason, Frances S. Grodzinsky, Ayanna Howard, Keith W. Miller, and Marty J. Wolf. "AI Ethics: A Long History and a Recent Burst of Attention." *Computer*, 2021: 96-102.

Brännström, Mattias, Andreas Theodorou, and Virginia Dignum. "Let it RAIN for Social Good." *Workshop on Artificial Intelligence Safety.* Vienna: IJCAI-ECAI-22, 2022.

Brown, Shea, Jovana Davidovic, and Ali Hasan. "The algorithm audit: Scoring the algorithms that score us." *Big Data & Society*, 2021.

Buchanan, Bruce G. "A (Very) Brief History of Artificial Intelligence." *AI Magazine*, 2005: 53-60.

Choung, Hyesun, Prabu David, and Arun Ross. "Trust and ethics in AI." *AI & Society*, 2022.

Constantinescu, Mihaela, Cristina Voinea, Radu Uszkai, and Constantin Vică. "Understanding responsibility in Responsible AI. Dianoetic virtues and the hard problem of context." *Ethics and Information Technology*, 2021: 803–814.

Cortese, João Figueiredo Nobre Brito, Fabio Gagliardi Cozman, Marcos Paulo Lucca-Silveira, and Adriano Figueiredo Bechara. "Should explainability be a fifth ethical principle in AI ethics?" *AI Ethics*, 2022.

Cox, Andrew. "The Ethics of AI for information professionals: eight scenarios." *Journal of the Australian Library and Information*, 2022: 201-214.

Curtis, Caitlin, Nicole Gillespie, and Steven Lockey. "AI-deploying organizations are key to addressing 'perfect storm' of AI risks." *AI Ethics*, 2022.

de Laat, Paul. "Companies Committed to Responsible AI: From Principles towards Implementation and Regulation?" *Philosophy & Technology*, 2021: 1135–1193.

—. "From Algorithmic Transparency to Algorithmic Accountability? Principles for Rensponsible AI Scrutinized." *18th International Conference on the Ethical and Social Impacts of ICT, Paradigm Shifts in ICT Ethics: Societal Challenges in the Smart Society.* Logroño: Universidad de La Rioja, 2020. 336-340.

Dignum, Virginia. "Ethics in artificial intelligence: introduction to the special issue." *Ethics and Information Technology*, 2018b: 1-3.

Dignum, Virginia, et al. "Ethics by Design: Necessity or Curse?" *In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.* New York: Association for Computing Machinery, 2018a. 60–66.

Djeffal, Christian, Markus B. Siewert, and Stefan Wurster. "Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies." *Journal of European Public Policy*, 2022: 1799-1821.

Dobbe, Roel, Thomas Krendl Gilbert, and Yonatan Mintz. "Hard choices in artificial intelligence." *Artificial Intelligence*, 2021.

Dobrev, Dimiter. "A Definition of Artificial Intelligence." 2003.

Doya, Kenji, Arisa Ema, Hiroaki Kitano, Masamichi Sakagami, and Stuart Russell. "Social impact and governance of AI and neurotechnologies." *Neural Networks*, 2022: 542-554.

Eisenhart, Christopher, and Barbara Johnstone. *Rhetoric in Detail : Discourse Analyses of Rhetorical Talk and Text. Discourse Approaches to Politics, Society and Culture.* Amsterdam: John Benjamins Publishing Co., 2008.

European Commision. "Ethics Guidelines for Trustworthy AI." *High-Level Expert Group on Artificial Intelligence.* April 8, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed January 10, 2023).

European Commission. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts.* Proposal, Brussels: European Commission, 2021.

European Commission, Directorate-General for Research and Innovation, European Group on Ethics in Science and New Technologies. *Statement on artificial intelligence, robotics and 'autonomous' systems.* Brussels: Publication Office of the European Union, 2018.

Fahmideh, Mahdi, et al. "Ethics of AI: A Systematic Literature Review of Principles and Challenges." *In Proceedings of ACM Conference (EASE).* New York: ACM, 2022.

Fairclough, Norman. *Analysing Discourse: Textual Analysis for Social Research.* London; New York: Routledge, 2003.

—. *Discourse and Social Change.* Cambridge: Polity, 1992.

Falco, Gregory, et al. "Governing AI safety through independent audits." *Nature Machine Intelligence*, 2021: 566–571.

Fast, Ethan, and Eric Horvitz. "Long-term trends in the public perception of artificial intelligence." *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence.* San Francisco California: AAAI Press, 2017. 963–969.

Fenwick, Ali, and Gabor Molnar. "The importance of humanizing AI: using a behavioral lens to bridge the gaps between humans and machines." *Discover Artificial Intelligence*, 2022.

Flathmann, Christopher, Beau G. Schelble, Rui Zhang, and Nathan J. McNeese. "Modeling and Guiding the Creation of Ethical Human-AI Teams." *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* New York: Association for Computing Machinery, 2021. 469–479.

Floridi, Luciano, and Josh Cowls. "A Unified Framework of Five Principles for AI in Society." *Harvard Data Science Review*, 2019.

Foucault, Michel. *The Archaeology of Knowledge.* London: Tavistock, 1972.

Future of Life Institute. *Open Letter, AI Principles.* August 11, 2017. https://futureoflife.org/open-letter/ai-principles/ (accessed May 14, 2023).

Gabriel, Iason. "Artificial Intelligence, Values, and Alignment." *Minds & Machines*, 2020: 411–437.

Gan, Isabel, and Sara Moussawi. "A Value Sensitive Design Perspective on AI Biases." *Hawaii International Conference on System Sciences.* Hawaii: University of Hawaii at Manoa, 2022. 1-10.

Georgieva, Ilina, Claudio Lazo, Tjerk Timan, and Anne Fleur van Veenstra. "From AI ethics principles to data science practice: a reflection and a gap analysis based on recent frameworks and practical experience." *AI Ethics*, 2022: 697-711.

Gergen, Kenneth. "The Social Constructivist Movement in Modern Psychology." *American Psychologist*, 1985.

Giuffrida, Iria. "Liability for AI decision-making: some legal and ethical considerations." *Fordham Law Review*, 2019: 439-456.

Goertzel, Ben. "Human-level artificial general intelligence and the possibility of a technological singularity: A reaction to Ray Kurzweil's The Singularity Is Near, and McDermott's critique of Kurzweil." *Artificial Intelligence*, 2007b: 1161-1173.

Goertzel, Ben, and Pei Wang. "Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms - Proceedings of the AGI Workshop 2006." *AGI Workshop.* Amsterdam: IOS Press, 2007a. 36.

Graetz, Georg, Pascual Restrepo, and Oskar Nordström Skans. "Technology and the labor market." *Labour Economics*, 2022.

Green, Ben. "Data science as political action: Grounding data science in a politics of justice." *Journal of Social Computing*, 2021a: 249-265.

Green, Ben. "The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice." *Journal of Social Computing*, 2021b: 209-225.

Greene, Daniel, Anna Hoffmann, and Luke Stark. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." *Hawaii International Conference on System Sciences.* Manoa: University of Hawaii, 2019. 2122-2131.

Habermas, Jürgen. *The theory of Communicative Action vol. 1: Reason and the Rationalization of Society.* Boston: Beacon Press, 1984.

Hagendorff, Thilo. "A Virtue-Based Framework to Support Putting AI Ethics into Practice." *Philosophy & Technology*, 2022a.

Hagendorff, Thilo. "Blind spots in AI ethics." *AI Ethics*, 2022b: 851–867.

Hagendorff, Thilo. "The Ethics of AI Ethics: An Evaluation of Guidelines." *Minds & Machines*, 2020: 99-120.

Halme, Erika, et al. "How to Write Ethical User Stories? Impacts of the ECCOLA Method." *Agile Processes in Software Engineering and Extreme Programming.* Springer, Cham, 2021. 36–52.

Harrison, Andrew, Dayana Spagnuelo, and Ilaria Tiddi. "An Ontology for Ethical AI Principles." *Semantic Web Journal*, 2021.

Hauptman, Allyson I., Beau G. Schelble, and Nathan J. McNeese. "Adaptive Autonomy as a Means for Implementing Shared Ethics in Human-AI Teams." 2021.

Häußermann, Johann Jakob, and Christoph Lütge. "Community-in-the-loop: towards pluralistic value creation in AI, or—why AI needs business ethics." *AI Ethics*, 2022: 341-362.

Häußler, Helena. "The Underlying Values of Data Ethics Frameworks: A Critical Analysis of Discourses and Power Structures." *Libri*, 2021: 307-319.

Herzog, Christian. "Three Risks That Caution Against a Premature Implementation of Artificial Moral Agents for Practical and Economical Use." *Science and Engineering Ethics*, 2021.

High-Level Expert Group on Artificial Intelligence. "Ethics Guidelines for Trustworthy AI." *European Commision.* April 8, 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed January 10, 2023).

Himmelreich, Johannes, and Lim Désirée. "AI and Structural Injustice: Foundations for Equity, Values, and Responsibility." In *The Oxford Handbook of AI Governance*, by Justin B. Bullock, et al. Oxford: Oxford University Press, 2022.

Hodges, Brian David, Ayelet Kuper, and Scott Reeves. "Discourse analysis." *British Medical Journal*, 2008.

Holtman, Koen. "Demanding and Designing Aligned Cognitive Architectures." *arXiv:2112.10190*, 2021.

Holzscheiter, Anna. "Discourse as Capability: Non-State Actors' Capital in Global Governance." *Millennium*, 2005: 723–746.

IBM. "Everyday Ethics for Artificial Intelligence." October 4, 2018. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf (accessed May 14, 2023).

Janks, Hilary. "Deconstruction and reconstruction: Diversity as a productive resource." *Discourse: Studies in the Cultural Politics of Education*, 2005: 31-43.

Jantunen, Marianna, et al. "Building a Maturity Model for Developing Ethically Aligned AI Systems." *Papers of the 44th Information Systems Research Seminar in Scandinavia.* IRIS Association, 2021.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The global landscape of AI ethics guidelines." *Nature Machine Intelligence*, 2019: 389-399.

Johnson, Brittany, and Justin Smith. "Towards Ethical Data-Driven Software: Filling the Gaps in Ethics Research & Practice." *2nd International Workshop on Ethics in Software Engineering Research and Practice.* Madrid: IEEE/ACM, 2021. 18-25.

Johnstone, Barbara. *Discourse Analysis, 2nd edn.* Malden, MA: Wiley-Blackwell, 2008.

Kazim, Emre, and Adriano Soares Koshiyama. "A high-level overview of AI ethics." *Patterns*, 2021.

Kerasidou, Charalampia (Xaroula), Angeliki Kerasidou, Monika Buscher, and Stephen Wilkinson. "Before and beyond trust: reliance in medical AI." *Journal of Medical Ethics*, 2022: 852-856.

Kitchenham, Barbara A., Tore Dybå, and Magne Jørgensen. "Evidence-based software engineering." *26th International Conference on Software Engineering.* 2004. 273-281.

Kitchenham, Barbara, and Stuart Charters. "Guidelines for performing Systematic Literature Reviews in Software Engineering." 2007.

Kok, Joost N., Egbert J. Boers, Walter A. Kosters, Peter van der Putten, and Mannes Poel. "Artificial intelligence: definition, trends, techniques, and cases." *Artificial intelligence*, 2009: 270-299.

Kopec, Matthew, et al. "The Effectiveness of Embedded Values Analysis Modules in Computer Science Education: An Empirical Study." *ArXiv abs/2208.05453*, 2022.

Korteling, J. E., van de Boer-Visschedijk G. C., Blankendaal R. A. M., Boonekamp R. M., and Eikelboom A. R. "Human- versus Artificial Intelligence." *Frontiers in Artificial Intelligence*, 2021.

Kristeva, Julia. *Desire in Language: A Semiotic Approach to Literature and Art.* New York: Columbia University Press, 1980.

Larsson, Stefan. "On the Governance of Artificial Intelligence through Ethics Guidelines." *Asian Journal of Law and Society*, 2020: 437-451.

Latzko-Toth, Guillaume & Bonneau, Claudine & Millette, Melanie. "Small Data, Thick Data: Thickening Strategies for Trace-Based Social Media Research." In *The SAGE Handbook of Social Media Research Methods*, by Sloan Luke and Anabel Quan-Haase, 199-214. Thousand Oaks: SAGE Publications Ltd, 2016.

Lera-Leri, Roger, Filippo Bistaffa, Marc Serramia, Maite Lopez-Sanchez, and Juan Rodriguez-Aguilar. "Towards Pluralistic Value Alignment: Aggregating Value Systems Through ℓp-Regression." *Proceedings of the 21st International Conference* *on Autonomous Agents and Multiagent Systems.* Richland: International Foundation for Autonomous Agents and Multiagent Systems, 2022. 780-788.

Liao, Vera Q., and Shyam S. Sundar. "Designing for Responsible Trust in AI Systems: A Communication Perspective." *ACM Conference on Fairness, Accountability, and Transparency.* New York: Association for Computing Machinery, 2022. 1257-1268.

Lieto, Antonio, Mehul Bhatt, Alessandro Oltramari, and David Vernon. "The role of cognitive architectures in general artificial intelligence." *Cognitive Systems Research*, 2018: 1-3.

Lindgren, Simon, and Jonny Holmström. "A Social Science Perspective on Artificial Intelligence: Building Blocks for a Research Agenda." *Journal of Digital Social Research*, 2020: 1-15.

Lukkien, Dirk R. M., et al. "Toward Responsible Artificial Intelligence in Long-Term Care: A Scoping Review on Practical Approaches." *The Gerontologist*, 2021.

MacKenzie, Donald, and Judy Wajcman. "Introductory essay: the social shaping of technology." In *The social shaping of technology*, by Donald MacKenzie and Judy Wajcman, 3-27. Buckingham: Open University Press, 1999.

Martin, Kirsten. "Creating Accuracy and The Ethics of Predictive Analytics." 2021.

Martinez-Martin, Nicole, Henry T Greely, and Mildred K Cho. "Ethical Development of Digital Phenotyping Tools for Mental Health Applications: Delphi Study." *JMIR mHealth and uHealth*, 2021.

Mayes, Patricia. "The discursive construction of identity and power in the critical classroom: Implications for applied critical theories." *Discourse & Society*, 2010: 189-210.

McCarthy, John. "From here to human-level AI." *Artificial Intelligence*, 2007: 1174-1182.

McLennan, Stuart, Amelia Fiske, Daniel Tigard, Ruth Müller, Sami Haddadin, and Alena Buyx. "Embedded ethics: a proposal for integrating ethics into the development of medical AI." *BMC Med Ethics*, 2022.

Mittelstadt, Brent. "Principles alone cannot guarantee ethical AI." *Nature Machine Intelligence*, 2019: 501-507.

Mökander, Jakob, and Luciano Floridi. "Operationalising AI governance through ethics-based auditing: an industry case study." *AI Ethics*, 2022b: 1-18.

Mökander, Jakob, and Luciano Floridi. "Operationalising AI governance through ethics-based auditing: an industry case study." *AI Ethics*, 2022: 1-18.

Morley, Jessica, Anat, Garcia, Francesca Elhalal, Libby Kinsey, Jakob Mökander, and Luciano Floridi. "Ethics as a Service: A Pragmatic Operationalisation of AI Ethics." *Minds & Machines*, 2021a: 239-256.

Morley, Jessica, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. "Operationalising AI ethics: barriers, enablers and next steps." *AI & Society*, 2021b.

Morley, Jessica, Luciano Floridi, Libby Kinsey, and Anat 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*, 2020: 2141-2168.

Munn, Luke. "The uselessness of AI ethics." *AI Ethics*, 2022.

Nabavi, Ehsan, and Chris Browne. "Five Ps: Leverage Zones towards Responsible AI." *arXiv:2205.01070*, 2022.

Nyrup, Rune. "From General Principles to Procedural Values: Responsible Digital Health Meets Public Health Ethics." *Frontiers in Digital Health*, 2021.

Orlikowski, Wanda J., and Susan V. Scott. "Exploring Material-Discursive Practices." *Journal of Management Studies*, 2015: 697-705.

Othman, Kareem. "Public acceptance and perception of autonomous vehicles: a comprehensive review." *AI Ethics*, 2021: 355-387.

Ouchchy, Leila, Allen Coin, and Veljko Dubljević. "AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media." *AI & Society*, 2020: 927-936.

Paraman, Pradeep, and Sanmugam Anamalah. "Ethical artificial intelligence framework for a good AI society: principles, opportunities and perils." *AI and Society*, 2022: 1-17.

Parker, Ian. *Social Constructionism, Discourse and Realism.* London: SAGE Publications Ltd., 1998.

Petersen, Eike, et al. "Responsible and Regulatory Conform Machine Learning for Medicine: A Survey of Challenges and Solutions." *IEEE Access*, 2021.

Rahwan, Iyad, et al. "Machine behaviour." *Nature*, 2019: 477–486.

Raji, Inioluwa Deborah, Peggy Xu, Colleen Honigsberg, and Daniel Ho. "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance." *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.* New York: Association for Computing Machinery, 2022. 557–571.

Ratti, Emanuele, and Mark Graves. "Cultivating Moral Attention: a Virtue-Oriented Approach to Responsible Data Science in Healthcare." *Philosophy & Technology*, 2021: 1819–1846.

Remes, Liisa. "Diskurssianalyysin perusteet." In *Laadullisen Tutkimuksen Käsikirja*, by Jari Metsämuuronen, 315. Helsinki: International Methelp, 2006.

Reséguier, Anaïs, and Rowena Rodrigues. "AI ethics should not remain toothless! A call to bring back the teeth of ethics." *Big Data & Society*, 2020.

Resseguier, Anais, and Rowena Rodrigues. "Ethics as attention to context: recommendations for the ethics of artificial intelligence." *Open Research Europe*, 2021.

Roberts, Laura Weiss, Katherine A. Green Hammond, Cynthia M.A. Geppert, and Teddy D. Warner. "The Positive Role of Professionalism and Ethics Training in Medical Education: A Comparison of Medical Student and Resident Perspectives." *Academic Psychiatry*, 2004: 170-182.

Rousi, Rebekah. "With Clear Intention—An Ethical Responsibility Model for Robot Governance." *Frontiers in Computer Science*, 2022.

Russell, Stuart J. "Rationality and intelligence." *Artificial Intelligence*, 1997: 57-77.

Russell, Stuart J., et al. *Artificial Intelligence: A Modern Approach.* Harlow: Pearon, 2022.

Ryan, Mark, and Bernd Carsten Stahl. "Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications." *Journal of Information, Communication and Ethics in Society*, 2021: 61-86.

Saariluoma, Pertti, Hannu Karvonen, and Rebekah Rousi. "Techno-Trust and Rational Trust in Technology – A Conceptual Investigation." *Human Work Interaction Design. Designing Engaging Automation.* Springer, Cham, 2018. 283–293.

Salles, Arleen, Kathinka Evers, and Michele Farisco. "Anthropomorphism in AI." *AJOB Neuroscience*, 2020: 88-95.

Sanderson, Conrad, Qinghua Lu, David Douglas, Xiwei Xu, Liming Zhu, and Jon. Whittle. "Towards Implementing Responsible AI." 2022.

Schiff, Daniel: Borenstein, Jason, Justin Biddle, and Kelly Laas. "AI Ethics in the Public, Private, and NGO Sectors: A Review of a Global Document Collection." *IEEE Transactions on Technology and Society*, 2021: 31-42.

Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural Networks*, 2015: 85-117.

Schmitz, Anna, Maram Akila, Dirk Hecker, Maximilian Poretschkin, and Stefan Wrobel. "The why and how of trustworthy AI: An approach for systematic quality assurance when working with ML components." *Automatisierungstechnik*, 2022: 793-804.

Schopmans, Hendrik R. "From Coded Bias to Existential Threat: Expert Frames and the Epistemic Politics of AI Governance." *AAAI/ACM Conference on AI, Ethics, and Society.* New York: Association for Computing Machinery, 2022. 627–640.

Schwartz, Reva, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* Special Publication 1270, National Institute of Standards and Technology, 2022.

Seger, Elizabeth. "In Defence of Principlism in AI Ethics and Governance." *Philosophy and Technology*, 2022.

Seppälä, Akseli, Teemu Birkstedt, and Matti Mäntymäki. "From Ethical AI Principles to Governed AI." *International Conference on Information Systems.* Austin, Texas: ICIS, 2021.

Sigfrids, Anton, Mika Nieminen, Jaana Leikas, and Pietari Pikkuaho. "How Should Public Administrations Foster the Ethical Development and Use of Artificial Intelligence? A Review of Proposals for Developing Governance of AI." *Frontiers in Human Dynamics*, 2022.

Simon, Herbert A. "Artificial intelligence: an empirical science." *Artificial Intelligence*, 1995: 95-127 .

—. *Utility and Probability.* London: Palgrave Macmillan, 1990.

Siqueira de Cerqueira, José, Anayran Azevedo, Heloise Tives, and E.D. Canedo. "Guide for Artificial Intelligence Ethical Requirements Elicitation - RE4AI Ethical Guide." *Hawaii International Conference on System Sciences.* Honolulu: University of Hawaii at Manoa, Association for Information Systems IEEE Computer Society, 2022.

Sison, Alejo José G., and Dulce M. Redín. "A neo-aristotelian perspective on the need for artificial moral agents (AMAs) ." *AI & Society*, 2021: 1-19.

Smith, Brad, and Harry Shum. "Foreword." In *The Future Computed: Artificial Intelligence and its Role in Society*, by Microsoft, 9. Independently published, 2018.

Stahl, Bernd Carsten. "Whose Discourse? A Comparison of the Foucauldian and Habermasian Concepts of Discourse in Critical IS Research." *AMCIS 2004*

*Proceedings.* New York: Americas Conference on Information Systems, 2004. 4329-4336.

Stahl, Bernd Carsten, Josephina Antoniou, Mark Ryan, Kevin Macnish, and Tilimbe Jiya. "Organisational responses to the ethical issues of artificial intelligence." *AI & Society*, 2022: 23–37.

Stenseke, Jakob. "Artificial virtuous agents: from theory to machine implementation." *AI & Society*, 2021.

Stenseke, Jakob. "Interdisciplinary Confusion and Resolution in the Context of Moral Machines." *Science and Engineering Ethics*, 2022.

Strümke, Inga, Marija Slavkovik, and Vince Istvan Madai. "The social dilemma in artificial intelligence development and why we have to solve it." *AI Ethics*, 2022: 655-665.

Tannen, Deborah, Heidi E. Hamilton, and Deborah Schiffrin. *Handbook of Discourse Analysis.* Hoboken: Wiley-Blackwell, 2015.

The 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, First Edition." 2019. https://ethicsinaction.ieee.org/wp-content/uploads/ead1e.pdf.

Trocin, Cristina, Patrick Mikalef, Zacharoula Papamitsiou, and Kieran Conboy. "Responsible AI for Digital Health: a Synthesis and a Research Agenda." *Information Systems Frontiers*, 2021.

Turing, Alan. "Computing Machinery and Intelligence." *Mind*, 1950: 433-460.

Ugwudike, Pamela. "Predictive Algorithms in Justice Systems and the Limits of Tech-Reformism." *International Journal for Crime, Justice and Social Democracy*, 2022: 85-99.

Ulnicane, Inga, Damian Okaibedi Eke, William Knight, Ogoh George, and Bernd Carsten Stahl. "Good governance as a response to discontents? Déjà vu, or lessons for AI from other emerging technologies." *Interdisciplinary Science Reviews*, 2021b: 71-93.

Ulnicane, Inga, William Knight, Tonii Leach, Bernd Carsten Stahl, and Winter-Gladys Wanjiku. "Framing governance for a contested emerging technology: insights from AI policy." *Policy and Society*, 2021a: 158-177.

Umbrello, Steven, and Ibo van de Poel. "Mapping value sensitive design onto AI for social good principles." *AI Ethics*, 2021: 283-296.

Vakkuri, Ville, and Pekka Abrahamsson. "The Key Concepts of Ethics of Artificial Intelligence." *IEEE International Conference on Engineering, Technology and Innovation.* Stuttgart: IEEE, 2018. 1-6.

Vakkuri, Ville, et al. "Time for AI (Ethics) maturity model is now." *Proceedings of the 2021 Workshop on Artificial Intelligence Safety.* RWTH Aachen: CEUR Workshop Proceedings, 2021b.

Vakkuri, Ville, Kai-Kristian Kemell, Joni Kultanen, and Pekka Abrahamsson. "The Current State of Industrial Practice in Artificial Intelligence Ethics." *IEEE Software*, 2020a.

Vakkuri, Ville, Kai-Kristian Kemell, Marianna Jantunen, and Pekka Abrahamsson. "This Is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments." *Agile Processes in Software Engineering and Extreme Programming.* Springer, 2020b. 195-210.

Vakkuri, Ville, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. "ECCOLA — A method for implementing ethically aligned AI systems." *Journal of Systems and Software*, 2021a.

Vakkuri, Ville, Kai-Kristian Kemell, Pekka Abrahamsson, Minna M. Rantanen, and Jani Koskinen. "AI Ethics in Industry: A Research Framework." 2019.

van Maanen, Gijs. "AI Ethics, Ethics Washing, and the Need to Politicize Data Ethics." *Digital Society*, 2022.

Varona, Daniel, and Juan Luis Suárez. "Discrimination, Bias, Fairness, and Trustworthy AI." *Applied Sciences*, 2022.

Vica, Constantin, Cristina Voinea, and Radu Uszkai. "The emperor is naked: Moral diplomacies and the ethics of AI." *Információs Társadalom*, 2021: 83-96.

Viljoen, Salomé. "The Promise and Limits of Lawfulness: Inequality, Law, and the Techlash." *SSRN: https://ssrn.com/abstract=3725645*, 2020.

Villegas-Galaviz, Carolina, and Kirsten Martin. "Moral Approaches to AI: Missing Power and Marginalized Stakeholders." *SSRN: https://ssrn.com/abstract=4099750*, 2022a.

Villegas-Galaviz, Carolina, and Kirsten Martin. "Moral Approaches to AI: Missing Power and Marginalized Stakeholders." *SSRN: https://ssrn.com/abstract=4099750*, 2022.

Villegas-Galaviz, Carolina, and Kirsten Martin. "Moral Distance, AI, and the Ethics of Care." *SSRN: https://ssrn.com/abstract=4003468*, 2022b.

Waelen, Rosalie. "Why AI Ethics Is a Critical Theory." *Philosophy & Technology*, 2022.

Whittlestone, Jess, Rune Nyrup, Anna Alexandrova, and Stephen Cave. "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions." *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* New York, NY, USA: Association for Computing Machinery, 2019. 195–200.

Wohlin, Claes. "Guidelines for snowballing in systematic literature studies and a replication in software engineering." *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering.* New York: Association for Computing Machinery, 2014. 1-10.

Wood, Linda A., and Rolf O. Kroger. *Doing Discourse Analysis : Methods for Studying Action in Talk and Text.* Thousand Oaks: SAGE Publications, Incorporated, 2000.

World Health Organization. *Ethics and governance of artificial intelligence for health: WHO guidance.* Guidance, World Health Organization, 2021.

Yampolskiy, Roman V. "Predicting future AI failures from historic examples." *Foresight*, 2019: 138-152.

Yazdannik, Ahmadreza, Alireza Yousefy, and Sepideh Mohammadi. "Discourse analysis: A useful methodology for health-care system researches." *Journal of Education and Health Promotion*, 2017.

Zicari, Roberto V., et al. "How to Assess Trustworthy AI in Practice." *ArXiv abs/2206.09887*, 2022.

Zuber, Niina, Jan Gogoll, Severin Kacianka, Alexander Pretschner, and Julian Nida-Rümelin. "Empowered and embedded: ethics and agile processes." *Humanities and Social Sciences Communications*, 2022.

# Appendices

## A BACKWARD SNOWBALL ITERATION 1

| Year | Author | Title | Exclusion Criteria |
|---|---|---|---|
| 2018 | Whittaker, Meredith, Crawford, Kate, Dobbe, Roel, Fried, Genevieve, Kaziunas, Elizabeth, Mathur, Varoon, West, Sarah Myers, Richardson, Rashida, Schultz, Jason & Schwartz, Oscar | AI Now Report 2018 | Publication Year |
| 2018 | Nemitz, Paul | Constitutional democracy and technology in the age of artificial intelligence | Publication Year |
| 2018 | Calo, Ryan | Artificial Intelligence Policy: A Primer and Roadmap | Publication Year |
| 2018 | Floridi, Luciano, Cowls, Josh, Beltrametti, Monica, Chatila, Raja, Chazerand, Patrice, Dignum, Virginia, Lütge, Christoph, Madelin, Robert, Pagallo, Ugo, Rossi, Francesca, Schafer, Burkhard, Valcke, Peggy & Vayena, Effy | AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations.<br>Minds and Machines. | Publication Year |

|  |  |  |  |
|---|---|---|---|
| 2018 | Filipović, Alexander, Koska, Christopher & Paganini, Claudia | Developing a Professional Ethics for Algorithmists: Learning from the Examples of Established Ethics | Publication Year |
| 2009 | Beauchamp, Tom L. & Childress, James F. | Principles of biomedical ethics | Publication Year |
| 2010 | Bosk, Charles L. | Bioethics, Raw and Cooked: Extraordinary Conflict and Everyday Practice | Publication Year |
| 2006 | Beauchamp, Tom L. & DeGrazia, David | Principles and Principlism | Publication Year |
| 1939 | Marshall, Thomas H. | The recent history of professionalism in relation to social structure and social policy | Publication Year |
| 1989 | Frankel, Mark S. | Professional codes: Why, how, and with what impact? | Publication Year |
| 2009 | Campbell Black, Henry, A. Garner, Bryan A. | Black's law dictionary | Publication Year |
| 2007 | MacIntyre, Alasdair | After Virtue: A Study in Moral Theory | Publication Year |
| 1986 | Gillon, Raanan | Do doctors owe a special duty of beneficence to their patients? | Publication Year |

| | | | |
|---|---|---|---|
| 1993 | Pellegrino, Edmund D. & Thomasma, David C. | The virtues in medical practice | Publication Year |
| 2010 | Van den Bergh, Joachim & Deschoolmeester, Dirk | Ethical Decision Making in ICT: Discussing the Impact of an Ethical Code of Conduct | Publication Year |
| 2009 | Manders-Huits, Noëmi & Zimmer, Michael | Values and Pragmatic Action: The Challenges of Introducing Ethical Intelligence in Technical Design Communities | Publication Year |
| 1991 | McDowell, Banks | Ethical Conduct and the Professional's Dilemma: Choosing Between Service and Success | Publication Year |
| 2002 | Iacovino, Livia | Ethical Principles and Information Professionals: Theory, Practice and Education | Publication Year |
| 2017 | Boddington, Paula | Towards a code of ethics for artificial intelligence research | Publication Year |
| 2017 | Zarsky, Tal | Incompatible: The GDPR in the Age of Big Data | Publication Year |
| 2015 | Balkin, Jack M. | Information Fiduciaries and the First Amendment | Publication Year |

|  |  |  |  |
|---|---|---|---|
| 2010 | Kish-Gephart, Jennifer J., Harrison, David A. & Treviño, Linda Klebe | Bad apples, bad cases, and bad barrels: Meta-analytic evidence about sources of unethical decisions at work | Publication Year |
| 2018 | Wakabayashi, Daisuke & Shane, Scott | Google Will Not Renew Pentagon Contract That Upset Employees | Publication Year |
| 2018 | Conger, Kate & Wakabayashi, Daisuke | Google Employees Protest Secret Work on Censored Search Engine for China | Publication Year |
| 1981 | Parker, Donn B. | Ethical Conflicts in Computer Science and Technology | Publication Year |
| 2006 | Carrese, Joseph A. & Sugarman, Jeremy | The Inescapable Relevance of Bioethics for the Practicing Clinician | Publication Year |
| 2016 | Brotherton, Stephen, Kao, Audiey & Crigger, Bette-Jane | Professing the Values of Medicine: The Modernized AMA Code of Medical Ethics | Publication Year |
| 2010 | Greenfield, Bruce & Jensen, Gail M. | Beyond a code of ethics: phenomenological ethics for everyday practice | Publication Year |
| 2018 | Panensky, S. A. & Jones, R. | Does IT Go Without Saying? | Publication Year |

| | | | |
|---|---|---|---|
| 1998 | Perlman, D. T | Who Pays the Price of Computer Software Failure Notes and Comments | Publication Year |
| 2016 | Mittelstadt, Brendt, Allo, Patrick, Taddeo, Mariarosaria, Wachter, Sandra & Floridi, Luciano | The ethics of algorithms:<br>Mapping the debate | Publication Year |
| 1991 | Jones, Thomas M. | Ethical Decision Making by Individuals in Organizations: An Issue Contingent Model | Publication Year |
| 2016 | Metcalf, Jacob & Crawford, Kate | Where are human subjects in Big Data research? The emerging ethics divide | Publication Year |
| 2016 | Burrell, Jenna | How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms | Publication Year |
| 2016 | Floridi, Luciano | Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions | Publication Year |
| 2018 | Edmond, Awad, Sohan, Dsouza, Richard, Kim, Jonathan, Schulz, Joseph, Henrich, | The Moral Machine experiment | Publication Year |

| | | | |
|---|---|---|---|
| | Azim, Shariff, Jean-François, Bonnefon & Iyad Rahwan | | |
| 2006 | Ess, Charles | Ethical pluralism and global information ethics | Publication Year |
| 1997 | van den Hoven, Jeroen | Computer Ethics and Moral Methodology | Publication Year |
| 1955 | Gallie, Walter B. | Essentially Contested Concepts | Publication Year |
| 1990 | Richardson, Henry S. | Specifying Norms as a Way to Resolve Concrete Ethical Problems | Publication Year |
| 2003 | Turner, Leigh | Bioethics in a Multicultural World: Medicine and Morality in Pluralistic Settings | Publication Year |
| 2015 | Rhodes, Rosamond | Good and not so good medical ethics | Publication Year |
| 1990 | Clouser, K. Danner & Gert, Bernard | A Critique of Principlism | Publication Year |
| 1992 | Degrazia, David | Moving Forward in Bioethical Theory: Theories, Caes, and Specified Principlism | Publication Year |
| 1993 | Orentlicher, David | The Influence of a Professional Organization on Physician Behavior | Publication Year |

| | | | |
|---|---|---|---|
| | | Symposium on the Legal and Ethical Implications of Innovative Medical Technology | |
| 2005 | Maxine A. Papadakis, Arianne Teherani, Mary A. Banach, Timothy R. Knettler, Susan L. Rattner, David T. Stern, J. Jon Veloski & Carol S. Hodgson | Disciplinary Action by Medical Boards and Prior Behavior in Medical School | Publication Year |
| 1982 | Toulmin, Stephen | How medicine saved the life of ethics | Publication Year |
| 2013 | van de Poel, Ibo | Translating Values into Design Requirements | Publication Year |
| 2015 | Gillon, Raanan | Defending the four principles approach as a good basis for good medical practice and therefore for good medical ethics | Publication Year |
| 2017 | Friedman, Batya, Hendry, David G. & Borning, Alan | A Survey of Value Sensitive Design Methods | Publication Year |
| 1992 | Friedman, Batya & Kahn, Peter H. | Human agency and responsible computing: Implications for computer system design | Publication Year |
| 2013 | Shilton, Katie | Values Levers: Building Ethics into Design | Publication Year |

| | | | |
|---|---|---|---|
| 1983 | MacKinnon, Kevin S. | Computer Malpractice: Are Computer Manufacturers, Service Burreaus, and Programmers Really the Professionals They Claim to Be | Publication Year |
| 2004 | Bynum, Terrell Ward | Ethical decision making and case analysis in computer ethics | Publication Year |
| 1985 | Ladd, John | The quest for a code of professional ethics: an intellectual and moral confusion | Publication Year |
| 2018 | McNamara, Andrew, Smith, Justin & Murphy-Hill, Emerson | Does ACM's code of ethics change ethical<br>decision making in software development? | Publication Year |
| 1996 | Brief, Arthur P., Dukerich, Janet M., Brown, Paul R. & Brett, Joan F. | What's wrong with the treadway commission report? Experimental analyses of the effects of personal values and codes of conduct on fraudulent financial reporting | Publication Year |
| 2007 | Helin, Sven & Sandström, Johan | An Inquiry into the Study of Corporate Codes of Ethics | Publication Year |

|  |  |  |  |
|---|---|---|---|
| 1996 | McCabe, Donald L., Trevino, Linda K. & Butterfield, Kenneth D. | The Influence of Collegiate and Corporate Codes of Conduct on Ethics-Related Behavior in the Workplace | Publication Year |
| 2007 | Jin, K. Gregory, Drozdenko, Ronald & Bassett, Rick | Information Technology Professionals' Perceived Organizational Values and Managerial Ethics: An Empirical Study | Publication Year |
| 2015 | Shilton, Katie | "That's Not An Architecture Problem!": Techniques and Challenges for Practicing Anticipatory Technology Ethics | Publication Year |
| 2016 | Goertzel, Karen M. | Legal liability for bad software | Publication Year |
| 2014 | Laplante, Phillip A. | Licensing professional software engineers: seize the opportunity | Publication Year |
| 2000 | Pour, Gilda, Griss, Martin L. & Lutz, Michael | The push to make software engineering respectable | Publication Year |
| 2008 | Seidman, Stephen B. | The Emergence of Software Engineering Professionalism | Publication Year |
| 1983 | Abbott, Andrew | Professional Ethics | Publication Year |

| | | | |
|---|---|---|---|
| 1978 | O’Connor, J. E. | Computer Professionals: The Need for State Licensing | Publication Year |
| 1998 | Gotterbarn, Don | The Ethical Computer Practitioner—Licensing the Moral Community: A Proactive Approach | Publication Year |
| 2018 | Suchman, Lucy | Corporate Accountability | Publication Year |
| 2016 | Angwin, Julia, Larson, Jeff & Kirchner, Lauren | Machine Bias | Publication Year |
| 2018 | European Group on Ethics in Science and New Technologies | Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems | Publication Year |
| 2018 | Holstein, Kenneth, Vaughan, Jennifer W., Daumé III, Hal, Dudík, Miro & Wallach, Hanna | Improving fairness in machine learning systems: What do industry practitioners need? | Publication Year |
| 2019 | OECD | Forty-two countries adopt new OECD Principles on Artificial Intelligence | Type of Publication |
| 2019 | European Commission | High Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI | Type of Publication |
| 2019 | Wachter, Sandra and Mittelstadt, Brent | A Right to Reasonable Inferences: Re-Thinking | Type of Publication |

| | | | |
|---|---|---|---|
| | | Data Protection Law in the Age of Big Data and AI | |
| 2019 | Wong, J. C. | Demoted and sidelined: Google walkout organizers say company retaliated | Type of Publication |
| 2019 | Benkler, Yochai | Don't let industry write the rules for AI | Type of Publication |
| 2019 | IEEE | The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design | Type of Publication |

# B FORWARD SNOWBALL ITERATION 1

| **Year** | **Author** | **Title** | **Exclusion Criteria** |
|---|---|---|---|
| 2019 | Glass, Philip | EINE SKIZZE ZUR RECHTLICHEN VERBINDLICHKEIT «ETHISCHER» KI-PRINZIPIEN | Language |
| 2019 | Whittlestone, Jess & Nyrup, Rune & Alexandrova, Anna & Dihal, Kanta & Cave, Stephen | Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research | Type of Publication |

| | | | |
|---|---|---|---|
| 2019 | AI Now Institute | AI Now 2019 Report | Type of Publication |
| 2019 | Redmond, Niamh & Michael Cooke NUIM Mac | Smart and Healthy Ageing through People Engaging in supporting Systems | Type of Publication |
| 2020 | Greene, Travis | The WEDF Student Initiative Submission Towards a Normative Evaluation of Personalization | Type of Publication |
| 2020 | Rockwell, Geoffrey | On IRIE Vol. 29 | Type of Publication |
| 2020 | Brey, Philip & Bottis, Maria | D5. 6: Recommendations for the enhancement of the existing legal frameworks for genomics, human enhancement, and AI and robotics | Type of Publication |
| 2020 | Christen, Markus, Heitz, Christoph, Kleiber, Tom & Loi, Michele | Code of ethics for data-based value creation | Type of Publication |
| 2020 | Stenberg, Louise & Nilsson, Svante | Factors influencing readiness of adopting AI: A qualitative study of how the TOE framework applies to AI adoption in governmental authorities | Type of Publication |
| 2020 | Donner, O. J. J. | The Influence of Ethical Guidelines for AI on Developers of AI | Type of Publication |

| | | | |
|---|---|---|---|
| 2020 | Wachter, Sandra, Brent Mittelstadt & Chris Russell | WHY FAIRNESS CANNOT BE AUTOMATED: BRIDGING THE GAP BETWEEN EU NON-DISCRIMINATION LAW AND AI | Type of Publication |
| 2020 | Gascón Marcén, Ana | Los principios para el desarrollo de la Inteligencia Artificial en Japón y las avenidas para la cooperación con la Unión Europea | Language |
| 2020 | Moratinos, Guillermo Lazcoz | Análisis de la propuesta de reglamento sobre los principios éticos para el desarrollo, el despliegue y el uso de la inteligencia artificial, la robótica y las tecnologías conexas | Language |
| 2020 | van Maanen, Gijs | Ethics washing: een introductie | Language |
| 2020 | Mühlhoff, Rainer | Automatisierte Ungleichheit | Language |
| 2020 | 甘偵蓉 & 許漢 | AI 倫理的兩面性初探-人類研發 AI 的倫理與 AI 倫理 | Language |
| 2020 | Bondu, Mathilde | Ethique et objets connectés | Language |

| | | | |
|---|---|---|---|
| 2020 | Klopotowska, Joanna E. | AI WILL SEE YOU NOW | Language |
| 2020 | 목광수 | 인공지능 개발자 윤리-덕성 기반의 모델을 제안하며 | Language |
| 2020 | TATO, NICOLÁS SANTIAGO | TRABAJO FINAL DE MAESTRÍA | Language |
| 2020 | Mühlhoff, Rainer | Automatisierte Ungleichheit: Ethik der Künstlichen Intelligenz in der biopolitische Wende des Digitalen Kapitalismus | Language |
| 2020 | Hagendorff, Thilo | The Ethics of AI Ethics: An Evaluation of Guidelines | Duplicate |
| 2020 | Zhang, Lawrence | Initiatives in AI Governance | Type of Publication |
| 2021 | W. H. O. Guidance | Ethics and Governance of Artificial Intelligence for Health | Duplicate |
| 2021 | Green, Ben | The contestation of tech ethics: A sociotechnical approach to ethics and technology in Action | Duplicate |
| 2021 | Bietti, Elettra | From Ethics Washing to Ethics Bashing: A Moral | Duplicate |

|  |  |  |  |
|---|---|---|---|
|  |  | Philosophy View on Tech Ethics |  |
| 2021 | Gray, Joanne & Witt, Alice | A feminist data ethics of care framework for machine learning: The what, why, who and how. | Duplicate |
| 2021 | Kelley, Stephanie | Employee Perceptions of Effective AI Principle Adoption | Duplicate |
| 2021 | Anderson, Marc & Fort, Karën | D1. 2 Legal and ethical requirements for human-machine interaction | Type of Publication |
| 2021 | W. H. O. Guidance | Ethics and Governance of Artificial Intelligence for Health | Type of Publication |
| 2021 | Pryzant, Reid | Natural Language Processing for Computing the Influence of Language on Perception and Behavior | Type of Publication |
| 2021 | Kililo, Damaris | Impacts of Artificial Intelligence on Job Security | Type of Publication |
| 2021 | Laine, Joakim | Ethics-based AI auditing core drivers and dimensions: A systematic literature review | Type of Publication |
| 2021 | Elliott-Renhard | Deux machina: a cross-disciplinary approach to | Type of Publication |

| | | | |
|---|---|---|---|
| | | artificial intelligence for regulatory understanding | |
| 2021 | Hinton, Charlene Palomer | The State of Ethical AI in Practice: A Multiple Case Study of Estonian Public Service Organizations | Type of Publication |
| 2021 | Smuha, Nathalie A | The Human Condition in An Algorithmized World: A Critique through the Lens of 20th-Century Jewish Thinkers and the Concepts of Rationality, Alterity and History | Type of Publication |
| 2021 | Al-Sultan, Narjes | How can Ethical Artificial Intelligence be understood from the perspective of the key principles of transparency, accountability, responsibility, fairness, privacy, and data governance? | Type of Publication |
| 2021 | Mäntymäki, Matti & Birkstedt, Teemu | ART of AI governance: Pro-ethical conditions driving ethical governance | Type of Publication |
| 2021 | ACS | The Ethics and Risks of AI Decision-Making | Type of Publication |

| | | | |
|---|---|---|---|
| 2021 | Dimitri, Alexandre | Artificial Intelligence in the Practice of Monitoring Worker Exhaustion | Language |
| 2021 | Avnoon, Netta, Kotliar, Dan & Rivnai-Bahir, Shira | What Do We Talk about When We Talk about Algorithmic Ethics? The Case of Israeli Data Scientists | Language |
| 2021 | Bruneault, Frédérick, & Andréane Sabourin Laflamme | Éthique de l'intelligence artificielle et ubiquité sociale des technologies de l'information et de la communication: comment penser les enjeux éthiques de l'IA dans nos sociétés de l'information? | Language |
| 2021 | Häußler, Helena | Data Ethics Frameworks | Language |
| 2021 | John-Mathews, Jean-Marie | L'Éthique de l'Intelligence Artificielle en Pratique. Enjeux et Limites | Language |
| 2021 | Llano-Alonso, Fernando H | L'etica dell'intelligenza artificiale nel quadro giuridico dell'Unione europea | Language |
| 2021 | de Cerqueira, Siqueira, Antonio, José, Tives, Heloise Acco & Canedo, Edna Dias | Ethical Guidelines and Principles in the Context of Artificial Intelligence | Language |

|  |  |  |  |
|---|---|---|---|
| 2022 | Selwyn, Neil | AI, education and ethics–starting a conversation | Type of Publication |
| 2022 | Häußermann, Johann Jakob | The Politics of Innovation | Type of Publication |
| 2022 | Nguyen Hoang Nam, Phuong | Data ethics in the context of data literacy–an analysis of educational approaches for higher education | Type of Publication |
| 2022 | Domin, Heather Elaine | Principles for Facial Recognition Technology: A Content Analysis of Ethical Guidance | Type of Publication |
| 2022 | Aaltonen, Venla | Ethics and safety of community-based geospatial data processes in the resilient urban South | Type of Publication |
| 2022 | Sjøberg, Marius Christopher | Responsible AI and Its Effect on Organizational Performance | Type of Publication |
| 2022 | Kallioinen, Emilia | The Making of Trustworthy and Competitive Artificial Intelligence: A Critical Analysis of the Problem Representations of AI in the European Commission's AI Policy | Type of Publication |
| 2022 | Kelley, Stephanie | UNDERSTANDING AND PREVENTING | Type of Publication |

| | | | |
|---|---|---|---|
| | | ARTIFICIAL INTELLIGENCE ETHICS ISSUES IN FINANCIAL SERVICES ORGANIZATIONS: THREE STUDIES | |
| 2022 | Stewart, Elizabeth K. | Alexa, Should I Trust You? A Theory of Trustworthiness for Artificial Intelligence | Type of Publication |
| 2022 | Susskind, Jamie | The Digital Republic: On Freedom and Democracy in the 21st Century | Access |
| 2022 | Steinacker, Léa | Code Capital | Language |
| 2022 | von Roesgen, Felix | Das Potential diskursiver Instrumente zur Implementierung von Corporate Digital Responsibility: Eine kritische Analyse | Language |
| 2022 | Beaudouin, Valérie & Julia Velkovska | Appel à propositions «Éthique de l'IA»: enquêtes de terrain | Language |
| 2022 | Gil, Aranguren & del Rocío, Mónica | La sostenibilidad frente a las prácticas de Greenwashing: análisis en el sector textil | Language |

| | | | |
|---|---|---|---|
| 2022 | 汪琛, 孙启贵 & 徐飞 | 国际人工智能伦理治理研究态势分析与展望 | Language |
| 2022 | 周慎, 朱旭峰 & 梁正 | 全球可持续发展视域下的人工智能国际治理 | Language |
| 2022 | Crépel, Maxime & Cardon, Dominique | Robots vs algorithmes | Language |
| | Хисамова, Зарина Илдузовна | О некоторых аспектах международного сотрудничества в области разработки этических стандартов и регуляторики искусственного интеллекта и больших данных | Language |
| 2022 | Flores, Yarasca & Fátima Sarita, Susana | Nivel de conocimiento y aplicación de los principios éticos del Establecimiento de Salud los Licenciados, Ayacucho, 2021 | Language |
| 2022 | Benbouzid, Bilel & Cardon, Dominique | Contrôler les IA | Language |
| 2022 | Gatt, Monika | Sein und Zahl–der Dialog | Language |
| 2022 | Haas, Normen & Sessler, Marie-Luise | Integration moralischer Anforderungen in den | Language |

| | | | |
|---|---|---|---|
| | | agilen Entwicklungsprozess KI-basierter Anwendungen am Beispiel von Scrum | |
| 2022 | Иджитканлар, Тан & Кугурульо, Федерико | Устойчивость искусственного интеллекта: взгляд урбаниста сквозь призму концепции умного и устойчивого города | Language |
| 2022 | Lana, Maurizio | Artificial Intelligence Systems and problems of the concept of author. Reflections on a recent book | Language |
| 2022 | Santosh, K. C. & Casey Wall | AI and Ethical Issues | Access |
| 2022 | Schiff, Daniel S., Laas, Kelly, Biddle, Justin B. & Borenstein, Jason | Global AI Ethics Documents: What They Reveal About Motivations, Practices, and Policies | Access |
| 2022 | Vakkuri, Ville, Kemell, Kai-Kristian, Tolvanen, Joel, Jantunen, Marianna, Halme, Erika & Abrahamsson, Pekka | How Do Software Companies Deal with Artificial Intelligence Ethics? A Gap Analysis | Access |
| 2022 | Kernbach, Julius M., Hakvoort, Karlijn, Ort, Jonas, | The Artificial Intelligence Doctor: | Access |

| | | | |
|---|---|---|---|
| | Clusmann, Hans, Neuloh, Georg, & Delev, Daniel | Considerations for the Clinical Implementation of Ethical AI | |
| 2022 | Siapka, Anastasia | Towards a Feminist Metaethics of AI | Access |
| 2022 | Kleizen, Bjorn, van Dooren, Wim & Verhoest, Koen | Chapter 6: Trustworthiness in an era of data analytics: what are governments dealing with and how is civil society responding? | Access |
| 2021 | Ruttkamp-Bloem, Emma | Re-imagining Current AI Ethics Policy Debates: A View from the Ethics of Technology | Access |
| 2021 | Langlois, Lyse & Régis, Catherine | Analyzing the contribution of ethical charters to building the future of artificial intelligence governance | Access |
| 2021 | Crawford, Kate | The atlas of AI: Power, politics, and the planetary costs of artificial intelligence | Access |
| 2021 | Ko, Young-Hwa & Leem, Choon-Seong | The influence of AI technology acceptance and ethical awareness towards intention to use | Language |

| | | | |
|---|---|---|---|
| 2022 | Burnett, Donna, ElHaber, Nicole, Alahakoon, Damminda, Karnouskos, Stamatis & De Silva, Daswin | Advancing an Artificial Intelligence Ethics Framework for Operator 4.0 in Sustainable Factory Automation | Access |
| 2022 | Pardoux, Éric | Ethical Design for AI in Medicine | Access |
| 2021 | Spiekermann, Sarah | From value-lists to value-based engineering with IEEE 7000™ | Access |
| 2021 | Brown, Shea, Davidovic, Jovana & Hasan, Ali | The algorithm audit: Scoring the algorithms that score us | Duplicate |
| 2020 | Wright, Steven A. | AI in the Law: Towards Assessing Ethical Risks | Access |
| 2021 | Sidorenko, E. L., Khisamova, Z. I. & Monastyrsky, U. E. | The main ethical risks of using artificial intelligence in business | Access |
| 2021 | Vakkuri, Ville, Kemell, Kai-Kristian & Abrahamsson, Pekka | Technical briefing: Hands-on session on the development of trustworthy AI software | Access |
| 2022 | Sigfrids, A., Nieminen, M., Leikas, J. & Pikkuaho, P. | How Should Public Administrations Foster the Ethical Development and Use of Artificial Intelligence | Duplicate |

| | | | |
|---|---|---|---|
| 2022 | López Belloso, María | Women's Rights Under AI Regulation: Fighting AI Gender Bias Through a Feminist and Intersectional Approach | Access |
| 2020 | Aitken, Mhairi, Ng, Magdalene, Toreini, Ehsan, van Moorsel, Aad, Coopamootoo, Kovila PL & Elliott, Karen | Keeping it human: a focus group study of public attitudes towards AI in banking | Duplicate |
| 2020 | Fiske, Amelia, Tigard, Daniel, Müller, Ruth, Haddadin, Sami, Buyx, Alena & McLennan, Stuart | Embedded ethics could help implement the pipeline model framework for machine learning healthcare applications | Access |
| 2020 | Burr, Christopher & Floridi, Luciano | The ethics of digital well-being: A multidisciplinary perspective | Duplicate |
| 2021 | Leonelli, Sabina & Beaulieu, Anne | Data and society: A critical introduction | Access |
| 2021 | Hagendorff, Thilo | Ethics of Machine Learning. A Critical Appraisal of the State of the Art | Access |
| 2022 | Nnamani, Ngozi | Harnessing Technology for Hospitality and Tourism | Access |
| 2021 | Cassenti, Daniel N. & Kaplan, Lance M. | Robust uncertainty representation in human-AI collaboration | Access |

| | | | |
|---|---|---|---|
| 2020 | BRUNEAULT, Frédérick & SABOURIN LAFLAMME, Andréane | AI Ethics: How Can Information Ethics Provide a Framework to Avoid Usual Conceptual Pitfalls? An Overview | Duplicate |
| 2021 | Nooraie, Reza Yousefi, Lyons, Patrick G., Baumann, Ana A. & Saboury, Babak | Equitable Implementation of Artificial Intelligence in Medical Imaging: What Can be Learned from Implementation Science? | Access |
| 2022 | Koukouvinou, Panagiota & Holmström, Jonny | AI MANAGEMENT BEYOND THE NARRATIVES OF DYSTOPIAN NIGHTMARES AND UTOPIAN DREAMS: A SYSTEMATIC REVIEW AND SYNTHESIS OF THE LITERATURE | Access |
| 2022 | Arunagiri, Aravindhan & Udayaadithya, Avadhanam | Governing Artificial Intelligence in Post-Pandemic Society | Access |
| 2022 | Ullmann, Stefanie | Gender Bias in Machine Translation Systems | Access |
| 2021 | Ruttkamp-Bloem, Emma | The Quest for Actionable AI Ethics | Access |

|  |  |  |  |
|---|---|---|---|
| 2021 | Petit, Nicolas & De Cooman, Jerome | Models of Law and Regulation for AI | Access |
| 2021 | Schmager, Stefan | TRUST-WORTHY AI | Type of Publication |
| 2021 | Corrêa, Nicholas & De Oliveira, Nythamar | Good AI for the Present of Humanity Democratizing AI Governance | Duplicate |
| 2022 | Martin, Kirsten | Ethical Theories and Data Analytics | Access |
| 2022 | Guzmán-Velásquez, C. & Lalinde-Pulido, J. G. | Digital Ethics: Toward a Socially Preferable Development of AI Systems | Access |
| 2022 | Knox, Jeremy, Hoel, Tore & Yuan, Li | From Principles to Processes: Lessons for Higher Education From the Development of AI Ethics | Access |
| 2021 | Strümke, Inga, Slavkovik, Marija & Madai, Vince Istvan | The social dilemma in artificial intelligence development and why we have to solve it | Duplicate |
| 2021 | Cihon, Peter, Klein-altenkamp, Moritz J., Schuett, Jonas & Baum, Seth D. | AI certification: Advancing ethical practice by reducing information asymmetries | Duplicate |

| | | | |
|---|---|---|---|
| 2021 | Jantunen, Marianna, Halme, Erika, Vakkuri, Ville, Kemell, Kai-Kristian, Rousi, Rebekah, Mikkonen, Tommi, Nguyen Duc, Anh & Abrahamsson, Pekka | Building a Maturity Model for Developing Ethically Aligned AI Systems | Duplicate |
| 2020 | McLennan, Stuart, Fiske, Amelia, Celi, Leo Anthony, Müller, Ruth, Harder, Jan, Ritt, Konstantin, Haddadin, Sami & Buyx, Alena | An embedded ethics approach for AI development | Access |
| 2021 | Maas, Matthijs M. | Artificial Intelligence Governance under Change Foundations, Facets, Frameworks | Type of Publication |
| 2021 | Kelley, Stephanie | Effective Adoption and Implementation of AI Principles | Access |

|  |  |  |  |
|---|---|---|---|
| 2022 | Steinacker, Léa | Code Capital: A Sociotechnical Framework to Understand the Implications of Artificially Intelligent Systems from Design to Deployment | Access |
| 2022 | Sounderajah, Viknesh, Melissa D. McCradden, Xiaoxuan Liu, Sherri Rose, Hutan Ashrafian, Gary S. Collins, James Anderson, Patrick M. Bossuyt, David Moher, and Ara Darzi | Ethics methods are required as part of reporting guidelines for artificial intelligence in healthcare | Access |
| 2022 | Floridi, Luciano, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen | capAI-A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act | Type of Publication |
| 2022 | Sanderson, Conrad, Qinghua Lu, David Douglas, Xiwei Xu, Liming Zhu, and Jon Whittle | Towards Operationalising Responsible AI: An Empirical Study | Duplicate |