

7-29-2025

Understanding the Implementation of Responsible Artificial Intelligence in Organizations: A Neo-Institutional Theory Perspective

David Horneber

Friedrich-Alexander-Universitat Erlangen-Nurnberg, david.horneber@fau.de

Follow this and additional works at: <https://aisel.aisnet.org/cais>

Recommended Citation

Horneber, D. (2025). Understanding the Implementation of Responsible Artificial Intelligence in Organizations: A Neo-Institutional Theory Perspective. *Communications of the Association for Information Systems*, 57, 185-218. <https://doi.org/10.17705/1CAIS.05708>

This material is brought to you by the AIS Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in *Communications of the Association for Information Systems* by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Understanding the Implementation of Responsible Artificial Intelligence in Organizations: A Neo-Institutional Theory Perspective

Cover Page Footnote

This manuscript underwent peer review. It was received 09/25/2024 and was with the authors for six months for two revisions. Gerit Wagner served as Associate Editor.



Understanding the Implementation of Responsible Artificial Intelligence in Organizations: A Neo-Institutional Theory Perspective

David Horneber

Friedrich-Alexander-Universität Erlangen-Nürnberg
david.horneber@fau.de
0000-0001-6866-5146

Abstract:

The rapid development of artificial intelligence (AI) systems has raised concerns about their ethical, legal, and social risks. Despite notable progress in the development of responsible AI frameworks, methods, and tools, research shows that many organizations struggle to effectively implement responsible AI. I review prior research on responsible AI to explain the insufficient implementation of responsible AI in organizations. Drawing on neo-institutional theory, I find that policy-practice decoupling (i.e., organizational responsible AI policies are adopted but not implemented in practice) and means-end decoupling (i.e., organizational responsible AI policies are implemented in practice but do not achieve their intended goals) can explain the ineffective implementation of responsible AI, with AI practitioners playing a key role as institutional entrepreneurs or custodians in driving or inhibiting the implementation of responsible AI. I contribute to the literature on responsible AI by exploring the institutional pressures that drive or inhibit its implementation, synthesizing the challenges to its implementation, and providing an overview of the roles and strategies AI practitioners use to deal with the implementation of responsible AI. I propose several avenues for future research and discuss implications for research and practice.

Keywords: Artificial Intelligence, Responsible AI, AI Ethics, Organizations, Neo-Institutional Theory.

This manuscript underwent peer review. It was received 09/25/2024 and was with the authors for six months for two revisions. Gerit Wagner served as Associate Editor.

1 Introduction

As artificial intelligence (AI) systems become more widespread in society, the companies developing and deploying these technologies face increasing public and regulatory pressure to identify and mitigate potential risks associated with their use (Candelon et al., 2021; Grant & Weise, 2023). In response to these pressures, the concept of responsible AI has emerged as a central approach to ensuring that AI systems are developed and deployed in alignment with ethical, legal, and social norms (Mikalef et al., 2022; Papagiannidis et al., 2025; Vassilakopoulou et al., 2022). Although the topic of responsible AI has been a focus of research and practice in recent years (e.g., Attard-Frost et al., 2022; Dignum, 2019; Mikalef et al., 2022), research shows that companies still face significant challenges in effectively implementing responsible AI (e.g., Morley et al., 2023; Vakkuri et al., 2020; Varanasi & Goyal, 2023). For example, several studies highlight the difficulty of translating abstract ethical principles into practice (e.g., Ali et al., 2023; Ibáñez & Olmeda, 2022; Mittelstadt, 2019). In addition, previous research suggests that methods and tools developed for the development of responsible AI are often not used in practice because they are not known or do not meet the needs of practitioners (e.g., Holstein et al., 2019; Hopkins & Booth, 2021; Morley et al., 2023).

Efforts to advance responsible AI have led to the emergence of a broad and multidisciplinary research landscape. For example, several researchers at the intersection of computer science and ethics have proposed normative principles and conceptual frameworks to guide the ethical development of AI (e.g., Floridi et al., 2018; Mittelstadt et al., 2016). Separately, there is a large body of research in computer science (e.g., Pant, Hoda, Spiegler, et al., 2024) and human-computer interaction (e.g., Rakova et al., 2021) that examines the motivations and challenges of implementing responsible AI in organizations from the perspective of AI practitioners involved in the design, development, and deployment of AI products and services (e.g., Product Manager, AI Engineer). Meanwhile, information systems researchers have focused on understanding responsible AI governance structures in organizations and how such structures can be effectively implemented from a management perspective (e.g., Papagiannidis et al., 2023; Schneider et al., 2022; Seppälä et al., 2021).

To bring coherence to these diverse research efforts, several review papers have attempted to synthesize and align the existing knowledge within and across the different research streams. For example, Papagiannidis et al. (2025) conducted a review of responsible AI principles to propose a unified definition of responsible AI governance and to outline a framework to better understand how responsible AI governance can be implemented in organizations. Similarly, Birkstedt et al. (2023) reviewed the literature on AI governance with the aim of building a consolidating definition of AI governance, capturing the current state of organizational AI governance processes, and identifying avenues for future research. Pant et al. (2024), in turn, provided a review focused on the perspectives of AI practitioners, outlining their needs and challenges in their efforts to implement responsible AI.

Despite the extensive body of research, what remains largely absent is a critical reflection on the implementation of responsible AI in organizations. Several studies suggest that many organizations fail to effectively implement responsible AI in practice (e.g., Morley et al., 2023; Vakkuri et al., 2020; Varanasi & Goyal, 2023). While prior work has begun to explore the challenges faced by AI practitioners, an analysis of how these challenges arise is lacking. In addition, little is known about the external pressures that affect how organizations strategically prioritize their responsible AI initiatives. As a result, it remains unclear whether organizations simply fail to prioritize responsible AI, or whether the existing methods and frameworks developed in the literature are impractical or misaligned with the day-to-day activities of AI practitioners. Accordingly, I argue that to understand why many organizations struggle to effectively implement responsible AI, it is necessary to analyze both: the external pressures that shape how organizations approach responsible AI and the role of AI practitioners in influencing the implementation of responsible AI in practice. To address this gap, I review the existing literature on these topics and pose the following research questions:

RQ1: What are the drivers and inhibitors of implementing responsible AI in organizations?

RQ2: What are the roles and challenges of AI practitioners in implementing responsible AI in organizations?

I used a grounded theory literature review approach to address my research questions (Wolfswinkel et al., 2013). During the analysis, I found that neo-institutional theory provides a valuable lens to explain the

deficit in the implementation of responsible AI in organizations (Ali et al., 2023; Bromley & Powell, 2012; Meyer & Rowan, 1977). While neo-institutional theory did not guide the motivation, design, literature search, and initial open coding, it informed my later stages of analysis and helped me to answer my research questions. My results suggest that the occurrence of policy-practice decoupling (i.e., organizational responsible AI policies are adopted but not implemented into practice) and means-end decoupling (i.e., organizational responsible AI policies are implemented in practice but do not achieve their intended goals) can explain why many organizations fail to implement responsible AI practices effectively. In addition, I found that AI practitioners can take the role of institutional entrepreneurs or custodians to drive or inhibit the implementation of responsible AI in organizations.

I contribute to the literature on responsible AI by (1) unpacking the external drivers and inhibitors that influence the implementation of responsible AI, (2) synthesizing the challenges that AI practitioners face in implementing responsible AI, and explaining how they arise, and (3) providing an overview of how AI practitioners in different roles influence the implementation of responsible AI. Without effective responsible AI practices, companies will continue to fail to address the risks of the systems they develop and deploy. The findings of my study provide a foundation for researchers, practitioners, and policymakers to develop and implement measures that enable companies to effectively mitigate the risks of their AI systems.

2 Theoretical Background

2.1 Responsible AI

Responsible AI refers to the practice of developing, deploying, governing, and using AI systems in accordance with ethical, social, and legal norms (Mikalef et al., 2022; Minkkinen et al., 2023; Vassilakopoulou et al., 2022). After several AI incidents have highlighted the risks that AI systems can pose to individuals and society, the topic of responsible AI has received increased attention from researchers, organizations, and policymakers in recent years (Papagiannidis et al., 2025). Although there are clear definitions of responsible AI (Papagiannidis et al., 2025; Vassilakopoulou et al., 2022), research and practice often use different terms to describe efforts aimed at minimizing the risks associated with AI systems. For example, the terms trustworthy AI (e.g., Li et al., 2023), ethical AI (e.g., Mirbabaie et al., 2022), or AI governance (e.g., Birkstedt et al., 2023) are commonly used by researchers, policymakers, and practitioners. However, these terms are closely related in meaning and are often used interchangeably (Papagiannidis et al., 2025). As the term Responsible AI has become established and clearly defined in the IS community (e.g., Mikalef et al., 2022; Vassilakopoulou et al., 2022, 2022), I use it in my paper to ensure conceptual clarity.

Previous research has pursued different goals and approaches to the development and deployment of responsible AI, resulting in several research streams that span across disciplines. Early work emerged at the intersection of computer science and ethics, with the goal of identifying the ethical implications of AI and developing conceptual frameworks for mitigating the risks involved (e.g., Ashok et al., 2022; Floridi et al., 2018; Mittelstadt et al., 2016). At the same time, a more technically-oriented body of literature has developed within the computer science community. This stream focuses on creating methods and tools – such as model cards (Mitchell et al., 2019) or SHAP explanations (Shapley Additive Explanations) (Lundberg & Lee, 2017) – to support the development of responsible AI systems in practice. Building on these foundational efforts, a stream of research in information systems has emerged that seeks to understand and design organizational governance structures for responsible AI (e.g., Papagiannidis et al., 2023; Schneider et al., 2022; Seppälä et al., 2021). This literature examines how such structures can be implemented within companies to ensure alignment with ethical and societal expectations. In addition, a fourth stream has developed within computer science (e.g., Pant, Hoda, Spiegler, et al., 2024) and human-computer interaction (e.g., Rakova et al., 2021), exploring the motivations and challenges of AI practitioners in putting responsible AI principles into practice in organizations.

While several review papers have synthesized these research streams to advance the conceptual (Birkstedt et al., 2023; Papagiannidis et al., 2025), technical (Li et al., 2023), and individual (Pant, Hoda, Tantithamthavorn, et al., 2024) foundations of responsible AI, comparatively little is known about how organizations actually engage with these ideals in practice. Several studies suggest that the implementation of responsible AI in organizations is often lacking and that AI practitioners face numerous challenges in operationalizing responsible AI frameworks and using responsible AI methods (e.g., Morley et al., 2023; Vakkuri et al., 2020; Varanasi & Goyal, 2023). Therefore, the goal of my research is to review

this literature and examine how these challenges arise and why many companies fail to effectively implement responsible AI.

To contextualize the implementation of responsible AI in organizations, I follow the work of Ali et al. (2023) and draw on neo-institutional theory. In the next section, I explain the basics of neo-institutional theory and describe how its concepts of institutional decoupling, institutional entrepreneurs, and institutional custodians inform my analysis of the implementation of responsible AI practices in organizations.

2.2 Neo-Institutional Theory and the Concept of Institutional Decoupling

Neo-institutional theory focuses on understanding how institutions shape organizational structures (Meyer & Rowan, 1977; Orlikowski & Barley, 2001). Institutions can be understood as “organized, established procedures” (Jepperson & Meyer, 2021, p. 37) that are represented through regulatory, normative, or cultural-cognitive elements in society (Scott, 2014). Regulatory elements are the formal laws and policies that are typically monitored and informally or formally sanctioned if violated. Normative aspects are the socially accepted values and norms of a domain that empower or constrain behaviors. Finally, cultural-cognitive elements are the shared conceptions and beliefs of a society that form the basis of their social reality and influence their behaviors and interactions.

Neo-institutional theory suggests that organizations are pressured to consider and incorporate these elements into their structures to gain legitimacy and secure their survival (Meyer & Rowan, 1977). However, complying with and embedding institutional elements is often not aligned with organizations' efficiency goals. Furthermore, institutions can be competing and mutually inconsistent (Boxenbaum & Jonsson, 2017; Meyer & Rowan, 1977). To resolve these issues, Meyer and Rowan (1977) suggested that companies decouple their organizational practices (i.e., activities, routines, and behaviors that occur within an organization) from their organizational structures (i.e., formal, visible elements of an organization that are designed to comply with existing institutions). This allows them to gain legitimacy by formally responding to institutional pressures and changing their structures without changing their work activities (Bromley & Powell, 2012; Meyer & Rowan, 1977).

While decoupling can be a useful strategy for organizations, it also carries a potential threat by possibly undermining their legitimacy if detected (Boxenbaum & Jonsson, 2017). Organizational decoupling can appear in different forms and at multiple levels (Bromley & Powell, 2012; Schnyder, 2018). Traditionally, decoupling has been studied as a gap between policies and organizational practices, with policies existing either at the macro- (i.e., supranational or national) or at the micro- (i.e., organizational) level. Building on this, Bromley and Powell (2012) proposed that decoupling can also occur as a gap between means and ends. Means-end decoupling describes the phenomena of when policies are implemented in practice (means) but they are disconnected from the intended goals (ends) (Bromley & Powell, 2012). Similar to policy-practice decoupling, these policies and ends can exist either at the macro- or organizational level (Schnyder, 2018).

Previous studies have demonstrated that policy-practice decoupling is likely to occur when the policies conflict with the organization's existing identities and proficiency goals when organizations and their leaders are not convinced of the policies, when organizations lack the capacities to implement them, or when a lack of sanctions or high internal power allows organizational members to resist them (Boxenbaum & Jonsson, 2017; Bromley & Powell, 2012). Accordingly, newly established policies with weak external enforcement are more likely to be only symbolically adopted (Bromley & Powell, 2012).

Means-end decoupling, on the other hand, is often driven by competing institutional pressures that have concrete organizational consequences due to varying transparency and accountability demands (Bromley & Powell, 2012). To serve these demands, practices are often implemented symbolically, without concrete utility. Additionally, it often occurs in highly complex settings where it is not clear which practices lead to desired outcomes and how outcomes can be defined and measured (Bromley & Powell, 2012; Wijen, 2014). Hence, although internal constituents often champion and pursue certain practices, these practices often do not lead to the expected outcomes (Bromley & Powell, 2012; Jabbouri et al., 2019).

Unlike efficiency- or performance-driven initiatives, previous research indicates that efforts to implement responsible AI are primarily legitimacy-oriented, driven by emerging institutional pressures (Agbese et al., 2023; Ibáñez & Olmeda, 2022). Given this emphasis on legitimacy as a key driver of organizational responsible AI initiatives, I argue that neo-institutional theory provides a valuable lens for examining the implementation of responsible AI in organizations (Ali et al., 2023). This theoretical perspective is particularly relevant because legitimacy-seeking behaviors often lead to institutional decoupling.

Moreover, I argue that the tendency toward institutional decoupling is further exacerbated by the current responsible AI landscape, which is characterized by high levels of regulatory uncertainty and complexity. Therefore, I propose that the concept of institutional decoupling is useful for understanding not only why organizations commit to responsible AI, but also how and why gaps may arise between these commitments and their practical implementation.

While the concept of institutional decoupling allows me to analyze why gaps exist between organizations' commitments to responsible AI and their implemented practices, it provides limited insight into the challenges that arise from these gaps or the mechanisms through which they might be actively navigated by AI practitioners. Addressing these issues requires focusing on the individual agency of AI practitioners to understand what challenges they face and how they drive or inhibit the implementation of responsible AI in organizations. In the next section, I discuss the role of individual agency in driving or inhibiting change and introduce the concepts of institutional entrepreneurs and custodians as a basis for analyzing the role of AI practitioners in this paper.

2.3 Institutional Entrepreneurs and Custodians

While the early work on neo-institutional theory has focused on understanding how institutional pressures affect organizational structures, over the years researchers have begun to examine how institutions emerge and change (Greenwood & Suddaby, 2006). To explain how individuals, groups, and organizations can influence these processes, researchers have established the concept of institutional entrepreneurship (Dimaggio, 1988; Maguire et al., 2004). Institutional entrepreneurs are “*actors who have an interest in particular institutional arrangements and who leverage resources to create new institutions or to transform existing ones*” (Maguire et al., 2004, p. 657). To envision and promote change and innovation, actors must look beyond their institutional field (Hardy & Maguire, 2017). However, given their embeddedness within an institutional field, the question arises how can they imagine and promote new practices despite the pressures of their institutional field?

Already a vast amount of literature has explored this paradox of embedded agency. Previous research has shed light on the characteristics and strategies of actors and the conditions of institutional fields that enable institutional entrepreneurs to overcome constraints and establish new practices. For instance, prior work suggests that actors who are cognitively or materially immune to existing institutional logics (Lepoutre & Valente, 2012), who possess high levels of reflexivity (Mutch, 2007), or who occupy positions that give them legitimacy with various stakeholders (Maguire et al., 2004) are more likely to exhibit behaviors that challenge existing practices. Moreover, disruptive events in a field (Hoffman, 1999), emerging fields (Maguire et al., 2004), or fields with tensions and contradictions (Seo & Creed, 2002) are associated with a greater likelihood of institutional entrepreneurship. Strategies to convince others of the desired change may include the use of discursive practices to set a rationale for new institutional arrangements, mobilizing and harnessing resources to incentivize the diffusion of new practices, or building inter-actor relationships to form cooperation and create a foundation for collective action (Hardy & Maguire, 2017).

Besides individual or collective actors creating and changing institutions, they can also use their agency to maintain them (Lawrence & Suddaby, 2006). Institutional maintenance “*involves supporting, repairing or recreating the social mechanisms that ensure compliance [with the existing institutions]*” (Lawrence & Suddaby, 2006, p. 230). While institutions are traditionally viewed as self-sustaining, without institutional maintenance, they may suffer from declining relevance, efficiency, and legitimacy (Siebert et al., 2017). Especially in times of institutional change, deinstitutionalization, or high institutional complexity, institutional maintenance can play an important role in preserving institutions (Colombero & Boxenbaum, 2019). Because institutional change typically involves the redistribution of power and resources (Hirsch & Bermis, 2009), the actors who seek to maintain institutions are typically those who benefit from them (Currie et al., 2012). While different terms have been used to describe individuals who work to maintain institutionalized practices (Wright et al., 2021), the notion of institutional custodians is the most widely used in the literature on neo-institutional theory (Dacin & Dacin, 2008; Montgomery & Dacin, 2020). Lawrence and Suddaby (2006) have identified six types of strategies that actors can use to ensure the maintenance of institutions. The first three – enabling, policing, and deterring – are concerned with the creation (enabling) and observance (policing) of rules and barriers (detering) that support existing institutions. The other three – valorizing and demonizing, mythologizing, and embedding and routinizing – deal with the reproduction of existing institutions to anchor their legitimacy and strengthen compliance.

Because institutional maintenance contradicts institutional change, previous work has shown that it can contribute to organizational decoupling (Hirsch & Bermiss, 2009).

Given that the implementation of responsible AI in organizations involves the introduction of new practices and challenges existing routines, it constitutes a process of institutional change (Ali et al., 2023). Several studies suggest that this institutional change is influenced by AI practitioners, who either drive or inhibit the implementation of responsible AI (Ali et al., 2023; Lancaster et al., 2023; Rakova et al., 2021). To better understand how AI practitioners use their agency to influence this process of institutional change, I use the concepts of institutional entrepreneurs and custodians.

3 Methodology

My study aims to explain why organizations often fail to implement responsible AI effectively. Accordingly, my study can be placed at the intersection of an organizing and broad theorizing review, according to Leidner (2018). To achieve the objective of my paper and to synthesize and describe the existing literature on the topic, I used the grounded theory literature review approach from Wolfswinkel et al. (2013). Using grounded theory as an analysis method enabled me to get an in-depth understanding of the studies in my review and allowed me to analyze how the themes identified hinder the effective implementation of responsible AI in organizations. Wolfswinkel et al. (2013) proposed a five-step approach to rigorously review the literature according to the grounded theory method: define, search, select, analyze, and present. In the following subsections, I describe how I conducted my literature review according to the first four steps. I then proceed to section four, where I present my findings.

3.1 Define

First, I defined the scope of my literature search. I used the following criteria to decide whether an article should be included in my review. (1) Articles should focus on the drivers, inhibitors, challenges, roles of AI practitioners, and practices for implementing responsible AI in organizations. This criterion was used to filter out all articles that address responsible AI from a perspective other than how responsible AI practices are understood and implemented in organizations. (2) Articles had to be empirical research papers. This criterion was set to ensure that the findings presented are grounded in real-world observations and provide a reliable basis for understanding how responsible AI is used in practice. (3) Articles had to be published in peer-reviewed journals or conference proceedings. This criterion was defined to exclude all gray literature on the topic and to guarantee that the papers included have been rigorously reviewed by experts in the field. (4) Articles had to be published after 2015. This criterion was used because 2016 was the year when the issue of responsible AI became more relevant in society. This shift followed the public discussion of the first high-profile cases of harmful AI systems. Notable examples include the biased recidivism prediction (Angwin et al., 2016) or Microsoft's racist "Tay" Chatbot (Schwartz, 2019). As a result of this increased attention, many public and private organizations have since developed and implemented responsible AI initiatives (De Laat, 2021; Jobin et al., 2019). (5) Articles had to be written in English and the full text had to be available online. This last criterion was established to ensure the accessibility of the articles. Furthermore, I excluded all articles that focused exclusively on the concrete use of a tool or method because they typically centered on improving the tools or methods rather than providing a rich description of companies' responsible AI activities. For example, I excluded the article by Heger et al. (2022) because it focused exclusively on the needs and challenges of AI practitioners with respect to documentation practices. As such, it did not include findings on how AI practitioners drive or inhibit the implementation of responsible AI and the challenges they face in implementing responsible AI.

Second, I determined the field of my search. Because responsible AI is an interdisciplinary topic at the intersection of information systems, computer science, ethics, and organizational studies, I did not restrict my search to specific journals or databases. Instead, I used the following search systems to conduct my literature search: Scopus, ACM Digital Library (The ACM Guide to Computer Literature), Web of Science, EBSCOhost (EconLit, Business Source Complete, Psychology, and Behavioral Science Collection), ProQuest (ABI/INFORM Collection) and AIS eLibrary. I did not use the IEEE Digital Library for my search because it has limited search capabilities for systematic searching (Gusenbauer & Haddaway, 2020). Moreover, its articles are indexed in three other search systems I used for my literature search (i.e., Scopus, Web of Science, Pro Quest). To expand my search and identify additional articles, I performed a forward and backward search of all included articles (Webster & Watson, 2002). The forward search was conducted by screening the title and abstract of all papers that cited the articles included in my review via Google Scholar.

Third, I specified the search terms. I iteratively developed the search terms by testing out different combinations of search terms and continuously adding more synonyms. To ensure the thoroughness of my search, I have used different spellings for each term. My final search string consisted of the terms displayed in Table 1.

Table 1. Search String

<p>("Responsible AI" OR "Responsible Artificial Intelligence" OR "AI responsib*" OR "Artificial intelligence responsib*" OR "Ethical AI" OR "Ethical Artificial intelligence" OR "AI ethic*" OR "Artificial Intelligence ethic*" OR "Trustworthy AI" OR "Trustworthy Artificial Intelligence" OR "AI trustworth*" OR "Artificial Intelligence trustworth*" OR "AI Safety" OR "Artificial Intelligence Safety" OR "AI Governance" OR "Artificial Intelligence Governance" OR "AI Compliance" OR "Artificial Intelligence Compliance" OR "Human-Centered AI" OR "Human-Centered Artificial intelligence" OR "AI alignment" OR "Artificial Intelligence alignment" OR "Accountable AI" OR "Accountable Artificial Intelligence" OR "AI accountab*" OR "Artificial Intelligence accountab*" OR "Fair AI" OR "Fair Artificial Intelligence" OR "AI fair*" OR "Artificial Intelligence fair*" OR "Explainable AI" OR "Explainable Artificial Intelligence" OR "AI explainab*" OR "Artificial Intelligence explainab*" OR "Transparent AI" OR "Transparent Artificial Intelligence" OR "AI transparen*" OR "Artificial Intelligence transparen*" OR "Responsible ML" OR "Responsible Machine Learning" OR "ML responsib*" OR "Machine learning responsib*" OR "Ethical ML" OR "Ethical Machine Learning" OR "ML ethic*" OR "Machine learning ethic*" OR "Trustworthy ML" OR "Trustworthy Machine Learning" OR "ML trustworth*" OR "Machine learning trustworth*" OR "ML Safety" OR "Machine Learning Safety" OR "ML Governance" OR "Machine Learning Governance" OR "ML Compliance" OR "Machine Learning Compliance" OR "Human-Centered ML" OR "Human-Centered Machine Learning" OR "ML alignment" OR "Machine Learning alignment" OR "Accountable ML" OR "Accountable Machine Learning" OR "ML accountab*" OR "Machine learning accountab*" OR "Fair ML" OR "Fair Machine Learning" OR "ML fair*" OR "Machine learning fair*" OR "Explainable ML" OR "Explainable Machine Learning" OR "ML explainab*" OR "Machine learning explainab*" OR "Transparent ML" OR "Transparent Machine Learning" OR "ML transparen*" OR "Machine learning transparen*") AND ("organization*" OR "organisation*" OR "compan*" OR "corporat*" OR "enterprise*" OR "business" OR "firm")</p>
--

As shown in Table 1, I combined the terms "responsible" (and its synonyms) and "AI" (and its synonyms) within a single set of quotation marks. I found that separating these terms produced search results that were too broad because they were often used in unrelated contexts. For example, combining the terms "responsible" and "AI" with an AND- or OR-operator yielded many results that only touched on the ethical implications of algorithms, without putting responsible AI at the center of the discussion. Finally, to set the focus on responsible AI in organizations, I combined synonyms for the term "responsible AI" and synonyms for the term "organization" with an AND-operator.

3.2 Search

After defining the scope, field, and search terms, I conducted the search using the terms and search systems described above. In all systems, I limited the search to the abstract, study title, and keywords to improve the precision of my results. Whenever possible, I filtered the inclusion criteria directly via the search query (e.g., the period after 2015). I conducted the search initially in October 2023 and updated it in September 2024. Accordingly, my sample does not contain any articles published after August 2024. Figure 1 gives an overview of the number of articles identified for each search system.

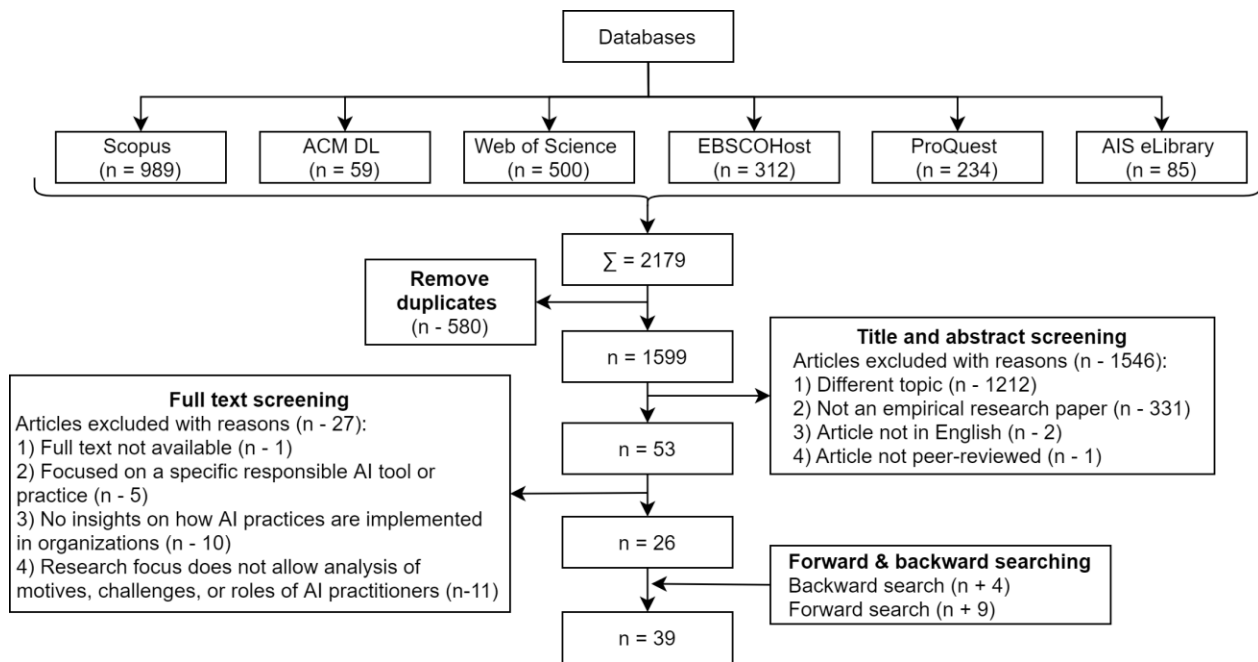


Figure 1. Literature Search Process

3.3 Select

In this step, I first filtered out all duplicates based on the digital object identifier (DOI) (if available) or the study title. Next, the author and a student assistant screened the title and abstract of each study independently. Based on the predefined selection criteria they determined which study should be included or excluded. A total of 29 studies were included by consensus, 1537 studies were excluded by consensus, and no consensus was reached for 33 studies. Accordingly, a Cohen's Kappa intercoder reliability value of 0.627 was obtained, indicating substantial agreement (Cohen, 1960). I used Cohen's Kappa because the data was nominal, the studies were screened by two coders, and I had no missing codes (Nili et al., 2020). I calculated intercoder reliability using ReCal2 (Freelon, 2010).

The two coders discussed the studies for which no agreement could be reached, and a joint decision was made. Of the 33 studies for which consensus could not be reached, 24 were included. 9 were excluded. As a result, 53 studies were retained out of the 1599 studies that were screened. Articles that were excluded based on their topic during title and abstract screening typically focused on the development of responsible AI artifacts, the ethical implications of specific AI artifacts, or the development or assessment of normative frameworks for the responsible governance of AI systems.

Following the title and abstract screening, the first author read the full text of all 53 included articles and excluded another 27 articles. Articles excluded in this step were either (1) unavailable, (2) focused on the development and implementation of a specific practice (3) centered on identifying responsible AI practices in organizations at a high-level, without providing insights on how they are implemented or (4) dealt with understanding responsible AI practices in organizations from a very specific perspective (e.g., moral engagement or disengagement of AI practitioners) that did not allow me to analyze the motives, challenges, or roles of AI practitioners.

Finally, I conducted a forward and backward search as recommended by Webster and Watson (2002). This yielded another 13 studies included in my analysis. Overall, I found 39 articles suitable for my review. The study selection process is illustrated in Figure 1. A description of the study profiles and an overview of all studies included in my review are provided in Appendix A.

3.4 Analyze

Following Wolfswinkel et al. (2013) I conducted the data analysis based on the principles of grounded theory (Gioia et al., 2013). In the first step, the author imported all 39 articles into the qualitative data analysis software MAXQDA and started reading the articles in random order to identify relevant excerpts

(VERBI Software, 2023). The excerpts were assigned with first-order codes that were iteratively developed and adapted during the open coding process. Although the full articles were read, the focus of coding was on the results and discussion sections. Open coding was guided by the objective of my review, which was to understand the drivers, inhibitors, roles, and challenges of the implementation of responsible AI in organizations. Because each article could address several aspects of interest in answering my research question, I assigned multiple codes to each article.

As initial concepts emerged and a clearer understanding of the data was developed, I began to identify similarities and differences among the initial codes and organized them into second-order themes. To define the first-order concepts and second-order themes, I followed the suggestion from Glaser and Strauss (1967) and constantly compared the excerpts with each other. As I identified the second-order themes, I continually returned to the first-order concepts and revised them as needed. During this process and inspired by the paper of Ali et al. (2023), I found that neo-institutional theory provides a useful lens for examining the implementation of responsible AI practices in organizations. Hence, I selectively recoded the second-order themes. For the second-order themes where neo-institutional theory was not applicable, I defined my own concepts to describe the themes. Finally, I looked for relations among the second-order themes and combined them into aggregate dimensions.

Overall, my analysis moved from an inductive approach to identifying the first-order codes to an increasingly abductive approach. I used deductive reasoning to make sense of my first-order codes, while simultaneously using the data to uncover new aspects and relationships within the concepts drawn from neo-institutional theory (Timmermans & Tavory, 2012; Wolfswinkel et al., 2013). An example of the data structure is presented in Table 2. The complete data structure is illustrated in Appendix B.

Table 2. Example of the Data Structure

Excerpt	First-order codes	Second-order themes	Aggregate dimensions
<i>"Ethical AI' was seen as a marketing advantage, with one participant suggesting that it is a 'very, very good influencing tool [where] users might choose [my company] over the competition."</i> (Widder & Nafus, 2023, p. 7)	Demand for responsible AI	Responsible AI implementation drivers	External pressures
<i>"According to the findings, these [ethical] requirements are considered relatively new and have had little impact on the market, meaning they currently have no financial value due to a lack of customer demand."</i> (Agbese et al., 2023, p. 5)	Demand for functional AI	Responsible AI implementation inhibitors	
<i>"The imbalance of knowledge and power created strife for some practitioners, who felt constrained in their own moral and ethical autonomy."</i> (Orr & Davis, 2020, p. 727)	Lack of authority	Challenges of policy-practice decoupling	Challenges to the implementation of responsible AI
<i>"The importance of clear roles and responsibilities regarding AI development and data governance was widely highlighted. However, these were still relatively unclear for many [...]"</i> (Seppälä et al., 2021, p. 8)	Unclear roles and responsibilities		
<i>"Another challenge practitioners faced was encountering new RAI values during late stages of implementation."</i> (Varanasi & Goyal, 2023, p. 9)	Reactive responsible AI structures	Challenges underlying means-end decoupling	
<i>"Translating ethics from principles to practice can be a challenging task. While ethical principles provide a foundation for ethical decision-making, applying these principles in practice can be complex."</i> (Pant, Hoda, Spiegler, et al., 2024, p. 80:15)	Operationalization of responsible AI policies		
<i>"One of the biggest challenges practitioners reported was that as responsible AI ethical tensions are identified, overly rigid organizational incentives may demotivate addressing them, [...]"</i> (Rakova et al., 2021, p. 7:16)	Disincentivization of responsible AI	Institutional custodians	Roles of AI practitioners
<i>"Others were able to influence others by reframing any problem as an issue of product quality."</i> (Ali et al., 2023, p. 222)	Reframing ethical issues	Institutional entrepreneurs	

4 Results

Figure 2 provides an overview of my findings. As illustrated in Figure 2 I identified several competing institutional pressures that can drive or inhibit the implementation of responsible AI. My analysis suggests that the tension between these drivers and inhibitors contributes to companies adopting (i.e., policy-practice decoupling) or implementing (i.e., means-end decoupling) their responsible AI policies only symbolically. Moreover, as shown in Figure 2, I found that AI practitioners can use their agency to either inhibit (institutional custodians) or drive (institutional entrepreneurs) the implementation of responsible AI in organizations. Finally, I identified several challenges to the implementation of responsible AI that result from policy-practice, or that can lead to means-end decoupling.

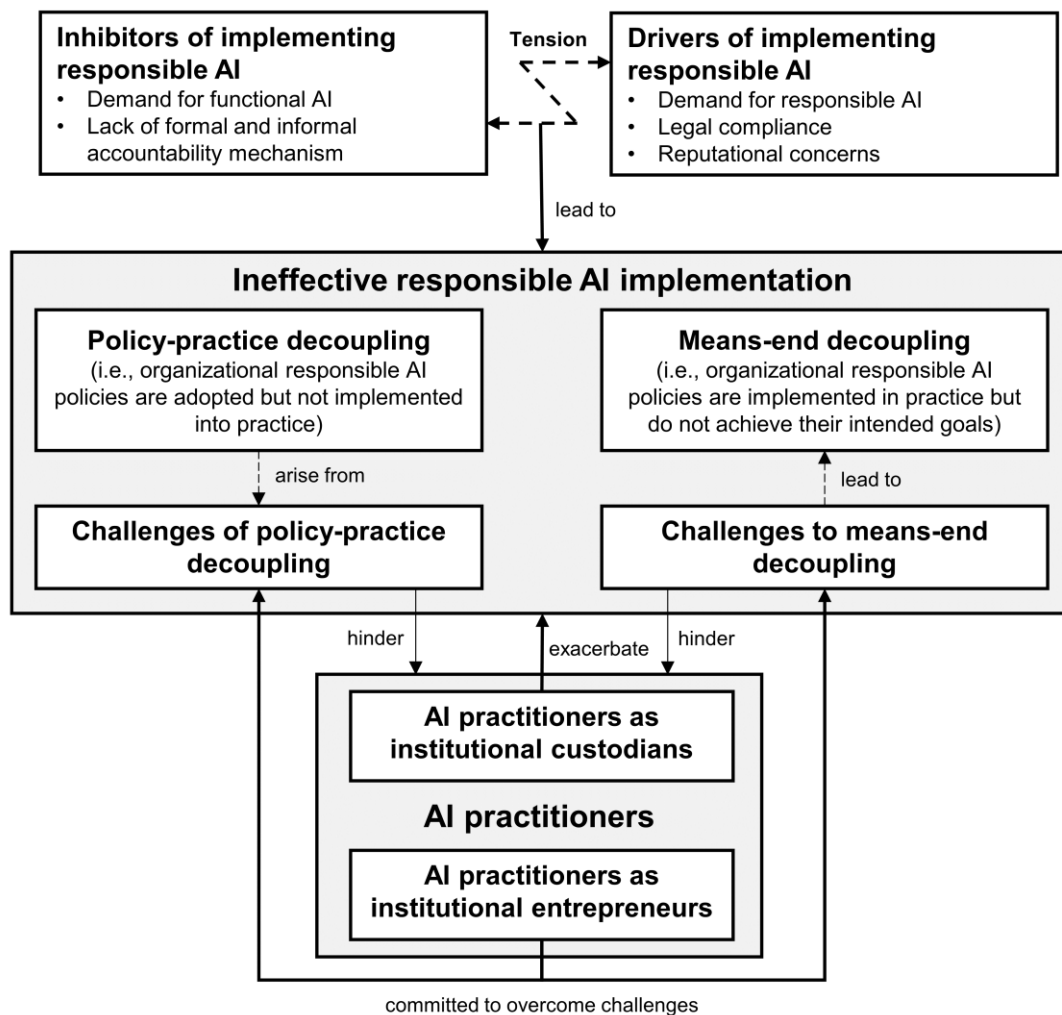


Figure 2. Framework for Explaining the Ineffective Implementation of Responsible AI

In the next section, I first discuss the drivers and inhibitors of responsible AI implementation. I then discuss the roles that AI practitioners play in implementing responsible AI. Finally, I outline the challenges that impede the implementation of responsible AI.

4.1 Drivers and Inhibitors for the Implementation of Responsible AI

In analyzing the drivers and inhibitors of implementing responsible AI, I found that there are tensions between all of the identified drivers and inhibitors. First, concerning the market demand for responsible AI, I found that there is potential for companies to use responsible AI activities as a marketing device to gain a competitive advantage (e.g., Morley et al., 2023; Widder & Nafus, 2023). However, most studies indicate that at the moment, the fulfillment of responsible AI policies is not demanded in the market and companies themselves are not concerned about adverse outcomes of the AI systems they are developing (e.g.,

Agbese et al., 2023; Baker-Brunnbauer, 2021). Second, I found tension in the regulatory requirements. On the one hand, organizations need to be compliant with existing regulations (e.g., GDPR) and they are concerned about future regulations (e.g., EU AI Act) (e.g., Hinton, 2023; Rakova et al., 2021). On the other hand, several studies state that there is a lack of regulatory accountability mechanisms (e.g., Ibáñez & Olmeda, 2022; Khan et al., 2023). Lastly, prior work indicates that organizations are inclined to adopt responsible AI practices due to reputational concerns and branding as an ethical employer (e.g., Griffin et al., 2024b; Sloane & Zakrzewski, 2022). However, due to weak informal accountability mechanisms and low demand for responsible AI, there is often no need for organizations to closely align their responsible AI policies with their development practices.

Therefore, I propose that policy-practice decoupling occurs in institutional fields, where formal (e.g., regulation) and informal (e.g., negative media attention) accountability mechanisms are lacking, and responsible AI is not demanded by the market. Conversely, means-end decoupling tends to occur in fields where regulation is more pressing (e.g., with respect to privacy) and the competition to meet financial, functional, and ethical requirements is stronger. Overall, my analysis suggests that the tension between external drivers and inhibitors is a key determinant of the ineffective implementation of organizationally responsible AI policies in practice.

4.2 The Role of AI Practitioners in Implementing Responsible AI

4.2.1 The Role of AI Practitioners as Institutional Custodians

Although the implementation of responsible AI is largely driven by external institutional pressures, my analysis indicates that certain AI practitioners play a crucial role in shaping how these practices are implemented in organizations. Prior work shows that AI practitioners can act as institutional custodians. In this role, they seek to preserve established AI development practices, which typically focus on efficiency and technical performance. AI practitioners who act as institutional custodians typically hold powerful positions in the AI development process (e.g., product managers) and they can use this position to allocate resources (e.g., Pant, Hoda, Spiegler, et al., 2024; Xivuri & Twinomurizi, 2023) and set incentives (e.g., Griffin et al., 2024a; Rakova et al., 2021) to prevent AI practitioners in subordinate positions from performing responsible AI work. As a justification for not considering organizationally responsible AI policies in AI development and deployment, these practitioners typically argue that there is a goal conflict between ethical and business goals (Hopkins & Booth, 2021; Sanderson et al., 2023). For example, collecting additional data to mitigate bias can be at odds with project-related goals such as timeliness and cost. Moreover, they argue that considering responsible AI policies can degrade the performance of the AI system they are developing (e.g., Avnoon et al., 2023; Widder & Nafus, 2023). In addition, I found that these AI practitioners often view the AI systems they develop as inherently value-neutral (Holstein et al., 2019; Lancaster et al., 2023), and they sometimes view certain problems, such as biased AI, as inevitable (Lancaster et al., 2023). Thus, they do not want to invest resources to implement responsible AI policies.

Based on these findings, I propose that AI practitioners acting as institutional custodians can inhibit the implementation of responsible AI practices in two ways. First, they reinforce existing norms, actively deepening non-compliance and exacerbating the decoupling between organizational policies and practices. Second, they may resist emerging norms, leading to means-end decoupling. In both cases, these dynamics impede the effective implementation of responsible AI.

4.2.2 The Role of AI Practitioners as Institutional Entrepreneurs

Besides AI practitioners seeking to maintain existing AI development practices, my findings indicate that AI practitioners with expertise in responsible AI can take on the role of an institutional entrepreneur and drive the implementation of responsible AI in organizations (Ali et al., 2023; Rakova et al., 2021; Varanasi & Goyal, 2023). Previous research suggests that these practitioners are motivated by a personal or professional ethos to take responsibility for the AI products or services that they are developing (e.g., Avnoon et al., 2023; Wang et al., 2023). While in organizations with more mature responsible AI structures, institutional entrepreneurs may hold a formal responsible AI position. In organizations with less mature responsible AI structures, they often take on this role voluntarily (Rakova et al., 2021; Varanasi & Goyal, 2023).

In their mission to advance the implementation of responsible AI, AI practitioners are actively working to address the challenges of responsible AI work (see Tables 3 and 4). While not all challenges can be fully

overcome or tackled by AI practitioners, existing research has identified several strategies to mitigate at least some of the structural and knowledge barriers. One strategy can be to build and rely on relationships with managers and decision-makers to convince them of the need to consider ethical issues. As the work of Ali et al. (2023) and Rakova et al. (2021) suggests, having a network of high-trust relationships can help obtain resources to implement responsible AI policies in organizations. Furthermore, educating AI practitioners in leadership and operational positions about ethical risks and carefully communicating responsible AI measures can help raise awareness of the risks and demystify the perception that the risks are unmanageable (Wang et al., 2023). Other strategies can be to reframe responsible AI policies as quality issues (Ali et al., 2023), or to piggyback on existing development practices to obtain support from leadership for the implementation of responsible AI practices (Deng et al., 2023). Lastly, institutional entrepreneurs can use bridging activities, such as interdisciplinary responsible AI workshops, to foster a shared understanding of responsible AI issues among diverse AI practitioners and help them translate and contextualize abstract ethical principles into practice (Deng et al., 2023).

While using such strategies can bring forward the implementation of responsible AI practices in organizations, several studies suggest that the effort that institutional entrepreneurs take often goes unrecognized and uncompensated (e.g., Deng et al., 2023; Rakova et al., 2021). Moreover, because institutional entrepreneurs often have to navigate tensions between business goals and organizationally responsible AI policies, they often face unrealistic expectations. As a result, they often suffer from stress related to their work on responsible AI (e.g., Ali et al., 2023; Rakova et al., 2021).

4.3 The Challenges of Implementing Responsible AI

4.3.1 The Challenges of Policy Practice Decoupling

When organizationally responsible AI policies are decoupled from organizational practices, it creates an environment of ambiguity in which AI practitioners who act on the existing institutions typically struggle to navigate the gap between what is formally required and what is informally expected. This leads to several challenges that can hinder the implementation of responsible AI in organizations. Table 3 provides an overview of these challenges.

Table 3. Overview of Responsible AI Implementation Challenges Resulting from the Decoupling of Policy and Practice

Challenges	References
Unclear roles and responsibilities	Morley et al. (2023); Papagiannidis et al. (2023); Rakova et al. (2021); Seppälä et al. (2021); Widder & Nafus (2023)
Responsibility avoidance	Ali et al. (2023); Baker-Brunnbauer (2021); Deng et al. (2023); Griffin et al. (2024b); Hopkins & Booth (2021); Lancaster et al. (2023); Orr & Davis (2020); Popova et al. (2024); Sanderson et al. (2023); Seguel & Vaast (2021); Seppälä et al. (2021); Vakkuri, Kemell, Tolvanen, et al. (2022); Widder & Nafus (2023)
Lack of leadership support	Ali et al. (2023); Griffin et al. (2024b); Kelley (2022); Lancaster et al. (2023); Mayer et al. (2021); Merhi (2023); Morley et al. (2023); Rakova et al. (2021); Ryan et al. (2022)
Lack of authority	Ali et al. (2023); Avnoon et al. (2023); Griffin et al. (2024a); Lancaster et al. (2023); Orr & Davis (2020); Popova et al. (2024); Rakova et al. (2021); Ryan et al. (2022); Wang et al. (2023)

My analysis indicates that the adoption of an organizationally responsible AI policy generally increases awareness of the importance of responsible AI among AI practitioners (e.g., Khan et al., 2023; Stahl et al., 2022). However, I also find that there are typically no formalized positions dedicated to implementing these policies. Hence, there is no clarity about whose job or responsibility it is to take care of responsible AI goals and AI practitioners are not incentivized to carry out responsible AI work (e.g., Morley et al., 2023; Rakova et al., 2021). As a result, AI practitioners do not see themselves as responsible for implementing organizationally responsible AI policies, even if they are aware of the ethical risks that may exist. Consequently, many AI practitioners are shifting the responsibility for potential negative outcomes up (e.g., users) or down (e.g., data providers) the AI supply chain (e.g., Popova et al., 2024; Seppälä et al., 2021).

Furthermore, even when AI practitioners are self-motivated to do responsible AI work, they often do not have the necessary authority to make decisions (e.g., Lancaster et al., 2023; Orr & Davis, 2020). Instead, they have to negotiate with managers or product owners to integrate ethics-orientated aspects in the development or deployment process (Ali et al., 2023; Lancaster et al., 2023). This is an issue because as explained in subsection 4.2, in an environment of policy-practice decoupling the support from leadership is

typically lacking (e.g., Ali et al., 2023; Kelley, 2022). Consequently, the implementation of organizationally responsible AI policies is de-prioritized, and AI practitioners have to deal with a lack of resources such as time, personnel, or data if they want to address the ethical issues they encounter (e.g., Hartikainen et al., 2022; Merhi, 2023; Pant, Hoda, Spiegler, et al., 2024).

4.3.2 The Challenges Underlying Means-End Decoupling

Despite implementing responsible AI practices, many organizations still fail to meet the intended goals of their responsible AI policies. I identified several challenges in the literature that can explain this gap between means and ends. The challenges are presented in Table 4.

Table 4. Overview of Responsible AI Implementation Challenges Underlying the Decoupling of Means and Ends

Challenges	References
AI risk blind spots	Agbese et al. (2023); Avnoon et al. (2023); Baker-Brunnbauer (2021); Deng et al. (2023); Griffin et al. (2024a, 2024b); Hinton (2023); Holstein et al. (2019); Hopkins & Booth (2021); Morley et al. (2023); Orr & Davis (2020); Popova et al. (2024); Ryan et al. (2022); Seguel & Vaast (2021); Seppälä et al. (2021); Treacy (2023); Vakkuri et al. (2020); Widder & Nafus (2023); Xivuri & Twinomurizi (2023)
Operationalization of responsible AI policies	Ali et al. (2023); Deng et al. (2023); Figueras et al. (2022); Ghatar et al. (2023); Griffin et al. (2024a); Ibáñez & Olmeda (2022); Kelley (2022); Khan et al. (2023); Merhi (2023); Morley et al. (2023); Pant, Hoda, Spiegler, et al. (2024); Ryan et al. (2022); Sanderson et al. (2022, 2023); Varanasi & Goyal (2023)
Goal conflicts between responsible AI goals	Figueras et al. (2022); Griffin et al. (2024a); Hinton (2023); Hopkins & Booth (2021); Khan et al. (2023); Merhi (2023); Pant, Hoda, Spiegler, et al. (2024); Rakova et al. (2021); Ryan et al. (2022); Sanderson et al. (2022, 2023); Seguel & Vaast (2021); Stahl et al. (2022)
Contextualization of responsible AI goals	Ghatar et al. (2023); Holstein et al. (2019); Ibáñez & Olmeda (2022); Kelley (2022); Khan et al. (2023); Pant, Hoda, Spiegler, et al. (2024); Sanderson et al. (2023); Treacy (2023); Varanasi & Goyal (2023)
Unpredictability of AI risk	Ghatar et al. (2023); Hartikainen et al. (2022); Holstein et al. (2019); Hopkins & Booth (2021); Merhi (2023); Pant, Hoda, Spiegler, et al. (2024); Sanderson et al. (2022); Treacy (2023); Wang et al. (2023)
Measuring and monitoring the impact of responsible AI	Ali et al. (2023); Deng et al. (2023); Figueras et al. (2022); Holstein et al. (2019); Kelley (2022); Mayer et al. (2021); Pant, Hoda, Spiegler, et al. (2024); Popova et al. (2024); Rakova et al. (2021); Sanderson et al. (2022); Varanasi & Goyal (2023)
Fragmented responsible AI structures	Ali et al., 2023; Baker-Brunnbauer, 2021; Ghatar et al., 2023; Holstein et al., 2019; Ibáñez & Olmeda, 2022; Kelley, 2022; Papagiannidis et al., 2023; Varanasi & Goyal, 2023; Xivuri & Twinomurizi, 2023
Reactive responsible AI structures	Ali et al. (2023); Hartikainen et al. (2022); Hinton (2023); Treacy (2023); Varanasi & Goyal (2023); Wang et al. (2023)
Reorganizations	Ali et al. (2023)
Lack of tools and methods	Figueras et al. (2022); Holstein et al. (2019); Hopkins & Booth (2021); Khan et al. (2023); Morley et al. (2023); Sanderson et al. (2022)

First, I found that the implementation of responsible AI practices often fails due to a lack of knowledge (e.g., Hinton, 2023; Ibáñez & Olmeda, 2022). Addressing ethical issues in AI requires interdisciplinary knowledge of AI and ethics. Prior work suggests that without the support of institutional entrepreneurs, AI practitioners frequently suffer from blind spots regarding the potential risks of AI, leading them to overlook or inadequately address these issues (e.g., Morley et al., 2023; Vakkuri, Kemell, Tolvanen, et al., 2022). Even when AI practitioners are aware of ethical goals and the need to handle them, they often struggle to operationalize and implement them in practice (e.g., Ghatar et al., 2023; Pant, Hoda, Spiegler, et al., 2024). This is because the goals of responsible AI can be very abstract, and they can have multiple meanings. For example, while achieving transparency can focus on disclosing information about the input data, design goals, and methods of development and operation of the AI systems (Loi et al., 2021), it can also focus on making the results of AI outcomes understandable (Lipton, 2018). Consequently, AI practitioners may dispute over the operationalization of certain goals, pursue the wrong goals, or take ineffective or superficial measures to achieve them (e.g., Figueras et al., 2022; Varanasi & Goyal, 2023).

The operationalization of responsible AI policies is further complicated by the fact that ethical issues are often application- and context-dependent. This makes it difficult for AI practitioners who typically develop only a part of the final system, to contextualize responsible AI policies and to anticipate the implications of AI systems in specific use cases (Holstein et al., 2019; Khan et al., 2023). Moreover, the understanding of responsible AI principles and goals can vary from culture to culture. Therefore, it can be a problem to

consider all cultural and geopolitical sensitivities when implementing responsible AI policies (e.g., Kelley, 2022; Treacy, 2023).

Because AI systems are data-driven, computationally complex, adaptive, and autonomous, their outcomes and their consequences are often unpredictable (Vakkuri, Kemell, Kultanen, et al., 2022). Hence, despite the aim of addressing organizational responsible AI policies, all outcomes might not be anticipated and mitigated during the development of the system (e.g., Pant, Hoda, Spiegler, et al., 2024; Vakkuri, Kemell, Kultanen, et al., 2022). Moreover, prior work suggests that concerning certain responsible AI issues, AI practitioners encounter challenges in identifying effective methods or tools to address them (e.g., Holstein et al., 2019; Sanderson et al., 2022). This can be due to a lack of awareness of certain tools or methods (Figueras et al., 2022). But it can also be because the tools or methods to perform, simplify, or automate responsible AI practices do not yet exist (Holstein et al., 2019; Sanderson et al., 2023). Consequently, even when AI practitioners are aware of certain problems and want to address them, they often have to devote considerable effort to finding ways to solve them, or they may not be able to address them technically at all.

While many of these knowledge-related challenges can be addressed by institutional entrepreneurs who build support structures to provide expertise in responsible AI, previous research indicates that even organizations with such structures often fail to achieve their intended goals. One reason for this gap is that many of these responsible AI structures in organizations are reactive (e.g., Ali et al., 2023; Varanasi & Goyal, 2023). For example, dedicated responsible AI experts or teams are frequently brought in late in the development process, leading to significant efforts to course correct, rather than proactively pursuing organizational responsible AI policies from the outset (e.g., Rakova et al., 2021; Wang et al., 2023).

In addition, even if companies have taken steps to implement responsible AI practices there are often no standardized or formalized processes around the implemented practices (e.g., Holstein et al., 2019; Ibáñez & Olmeda, 2022). As a result, responsible AI practices in organizations are often siloed and only implemented ad hoc by individual teams, rather than as centralized structures that can be leveraged by all teams and that provide guidance to AI practitioners on what responsible AI policies to focus on and how to achieve them (Ali et al., 2023; Holstein et al., 2019). In this context, studies report mixed findings concerning the success of centralized responsible AI teams to support the implementation of responsible AI practices. On the one hand, such teams can help individual project- or product-related AI teams by offering knowledge, resources, and building bridges between different parties working on responsible AI issues (Mayer et al., 2021; Wang et al., 2023). On the other hand, AI practitioners are often unaware that such teams exist, and when consulted, they sometimes have difficulties enforcing and taking responsibility from other teams (Kelley, 2022; Varanasi & Goyal, 2023).

Another major challenge is to define metrics to assess the achievement of responsible AI policies and monitor their fulfillment over time (e.g., Holstein et al., 2019; Rakova et al., 2021). Even if leadership is supportive of the implementation of responsible AI, they typically require AI practitioners to provide quantitative evidence of the impact and success of their activities, otherwise, organizational responsible AI policies may be de-prioritized (e.g., Ali et al., 2023; Rakova et al., 2021). Furthermore, because AI systems can learn and change over time, it is important to continually monitor adherence to responsible AI goals to address problems that may arise after deployment (Sanderson et al., 2023). Without such continuous monitoring, AI practitioners and leadership may see responsible AI practices as one-shot activities that do not require continuous engagement. Finally, constant reorganization of development and responsible AI teams can hinder the success of institutional entrepreneurs' implementation efforts (Ali et al., 2023).

5 Discussion

5.1 Theoretical Implications

The goal of my review was to examine why organizations are failing to effectively implement responsible AI. By systematically analyzing the existing literature on the topic and drawing on neo-institutional theory, I make three major contributions.

First, I set the focus on organizational decoupling as a phenomenon that can explain why the implementation of responsible AI practices in organizations is often lacking (Ali et al., 2023). While the existing literature on responsible AI has identified various challenges that prevent the implementation of responsible AI in organizations (e.g., Pant, Hoda, Spiegler, et al., 2024; Rakova et al., 2021; Varanasi &

Goyal, 2023), it remains unclear how these challenges arise. Hence, I contribute to the literature by uncovering the institutional pressures that drive or inhibit the implementation of responsible AI in organizations, and show how these can lead to different types of organizational decoupling. My results suggest that in less mature institutional fields (i.e., in fields with little or no market and regulatory pressures for responsible AI), organizational responsible AI policies are often decoupled from organizational responsible AI practices, and in more mature institutional fields (i.e., in fields with existing and competing market and regulatory pressures) organizational responsible AI practices are often decoupled from their intended goals. Because the responsible AI drivers and inhibitors can vary depending on the specific responsible AI policies, different forms of decoupling may simultaneously occur in an organization. Given that these different forms of organizational decoupling present unique challenges, my results call into question the notion of a uniform approach to responsible AI and highlight the complexity of implementing responsible AI in practice.

Second, I synthesize and outline the challenges that AI practitioners face when implementing responsible AI. While several studies have developed frameworks (e.g., Li et al., 2023), methods (e.g., Mitchell et al., 2019), and tools (e.g., Bellamy et al., 2018) to help AI practitioners implement responsible AI practices and address the risks of AI systems, my findings show that these are not well known in the industry and do not meet the needs of AI practitioners (e.g., Deng et al., 2023; Morley et al., 2023). Thus, I contribute to the literature by demonstrating that there is a gap between research and practice concerning the provision and use of responsible AI methods and tools. Moreover, my findings indicate that many of the challenges AI practitioners face when implementing responsible AI can be traced back to an interdisciplinary knowledge gap concerning ethics and AI. Accordingly, I suggest that to overcome the challenges posed by responsible AI issues, not only technical solutions are needed, but also activities and practices to promote knowledge dissemination and collaboration across different positions. As such, I support the call from previous research to “*pursue ethics as a process, not technological solutionism*” (Mittelstadt, 2019, p. 505). This is particularly important because my findings suggest that most existing responsible AI practices in organizations are reactive and there are no standardized or formalized processes in place. Therefore, research on responsible AI needs to shift from developing high-level frameworks to designing processes that can be used to proactively monitor and embed ethical values in AI systems.

Finally, I shed light on AI practitioners’ roles in the implementation of responsible AI in organizations. Although several studies have mapped out how AI practitioners understand and deal with their responsibilities in the development and deployment of AI (e.g., Orr & Davis, 2020; Seguel & Vaast, 2021), the role that AI practitioners play in the implementation of responsible AI is not yet well understood. Synthesizing the findings from previous studies, I propose that AI practitioners can take on two roles – as institutional custodians or as institutional entrepreneurs – in which they can influence the implementation of responsible AI. By explaining the motivation of AI practitioners in these roles and describing the strategies they use to inhibit or promote the implementation of responsible AI, I extend the literature on the topic. Looking at the differences in the motivations of institutional custodians and entrepreneurs from an institutional perspective, my review reveals that they emerge from the existing and dominant market-oriented (e.g., goal conflicts between business and responsible AI policies) and technocratic (e.g., value-neutrality of AI) logics on the one hand, and an emerging human-centered (e.g., personal ethos) logic on the other. In addition, I contribute to the literature by demonstrating the constraints of a bottom-up approach to responsible AI. While my findings suggest that institutional entrepreneurs can play an important role in advocating for the implementation of responsible AI and, in particular, in overcoming the challenges of policy-practice decoupling, I also find that AI practitioners’ bottom-up approach to implementing responsible AI practices often fails to achieve its intended goals due to the challenges that lead to means-end decoupling. Accordingly, I propose that without organizational support structures, a bottom-up approach will not be successful in resolving the challenges underlying means-end decoupling. Furthermore, my review provides insights into the consequences for AI practitioners who are motivated to drive the implementation of responsible AI and thereby, are striving to change existing institutions. As previous research reveals, AI practitioners can take several strategies to bring about change in organizations (e.g., Ali et al., 2023; Deng et al., 2023; Varanasi & Goyal, 2023). However, this work is often associated with high levels of occupational stress and is typically not rewarded by management. Hence, I point out the adverse consequences that can emerge for AI practitioners who are engaging in responsible AI work.

5.2 Avenues for Future Research

In conducting my literature review, I identified several avenues for future research, which I have organized into three overarching themes that reflect the three main aspects of my findings. Table 5 summarizes the key research questions within each theme.

Table 5. Overview of Future Research Opportunities

Research goal	Sample research question
Understanding the pressures that shape Responsible AI implementation	<ul style="list-style-type: none"> • How do conflicts between coercive, normative, and cultural-cognitive pressures impact the implementation of responsible AI in organizations? • What role do industry norms and certifications play in shaping organizations' approaches to responsible AI? • Do organizations gain legitimacy when adopting and implementing responsible AI policies? • How do organization-level responsible AI efforts influence user trust in specific AI systems developed by the organization?
Overcoming the challenges of responsible AI	<ul style="list-style-type: none"> • How can responsible AI structures be integrated into core organizational processes to shift from reactive compliance to proactive ethical innovation? • How to develop metrics to effectively measure the alignment between organizationally responsible AI policies and their implementation? • What structures best disseminate responsible AI knowledge across technical and non-technical teams to bridge ethical-technical divides? • How can responsible AI methods and tools be tailored to the different needs of AI practitioners?
Understanding the institutional work of AI practitioners	<ul style="list-style-type: none"> • What are the institutional logics that influence the responsible AI work of AI practitioners? • How do AI practitioners reproduce or challenge dominant institutional logic to drive or inhibit the implementation of responsible AI in organizations? • How do organizational characteristics enable or constrain AI practitioners' ability to act as institutional entrepreneurs or custodians in their responsible AI work? • How do AI practitioner communities legitimize or challenge responsible AI norms, and to what extent do these external logics influence practitioners' institutional work in their organizations?

5.2.1 Understanding the Pressures that Shape the Responsible AI Implementation

At the moment the institutional field of responsible AI is highly fragmented (e.g., uncertainty about AI regulations, upcoming industry standards concerning AI auditing and certification practices, increasing public interest in AI through growing media attention, and the availability of open-source AI systems). My review suggests that policy-practice and means-end decoupling are driven by the institutional pressures emerging from these fields. However, because the studies in my review were predominantly exploratory and did not include in-depth case studies, further research is needed to better understand how certain coercive, normative, or cultural-cognitive pressures may influence the development of AI in different organizations and the resulting responsible AI structures. For example, research on means-end decoupling in corporate sustainability practices indicates that strong compliance pressures can exacerbate the mismatch between means and ends, as companies lack the freedom to effectively address their intended goals (Wijen, 2014). If current AI regulations have a similar effect, it may be better to define regulations that give companies more freedom on how to address responsible AI policies, while maintaining accountability mechanisms for failing to meet responsible AI policies. Therefore, I call for more research on how different drivers and inhibitors can influence the implementation of responsible AI.

Besides this, there is also a lack of understanding of the organizational outcomes of companies' current responsible AI practices. Neo-institutional theory suggests that organizations can gain legitimacy by matching their AI practices with societal values (Meyer & Rowan, 1977). However, because many organizations fail to reach their intended responsible AI policies, accusations of AI washing may arise if the public perceives that companies' responsible AI policies and communications are misleading (Seele & Schultz, 2022). While in a few studies of my review, AI practitioners accused their companies of AI washing (Kelley, 2022; Lancaster et al., 2023), there is a lack of knowledge on how firms' responsible AI activities are perceived by the public. Hence, future research can analyze how the public perceives responsible AI activities and how they affect organizations' responsible AI legitimacy. Better understanding the factors that influence organizations' responsible AI legitimacy can help guide

companies on what responsible AI actions to pursue and it can inform policy-makers on whether there is a need for stronger coercive pressures.

5.2.2 Understanding Organizational Approaches to Overcoming the Challenges of Responsible AI

My findings indicate that the two types of organizational decoupling create organizational structures that can inhibit the implementation of responsible AI in organizations. While I suggest that the challenges of policy-practice decoupling can be overcome either through increased external pressure or the bottom-up work of institutional entrepreneurs, solutions to the challenges of means-end decoupling are not well understood. My findings show that responsible AI structures are often siloed and reactive, and thus, are not effective in fulfilling organizationally responsible AI policies. For example, despite many studies reporting a lack of centralized support structures, I found mixed findings concerning the effectiveness of centralized responsible AI teams. Accordingly, further research is needed to explore how effective responsible AI structures need to be designed. In doing so, future research should focus on specific responsible AI policies and draw on findings from other fields that have examined how social goals are achieved in organizations (e.g., corporate social responsibility). Findings in this area can inform organizations about how to effectively implement responsible AI practices.

As described in the theoretical implications section, I identified a gap between research and practice concerning the development and provision of responsible AI frameworks, methods, and tools by researchers and their usefulness and dissemination in practice (e.g., Morley et al., 2023). For example, as Holstein et al. (2019) outlined, quantitative fairness metrics are often insufficient in practice because they do not match end users' understanding of fairness in AI systems. Instead, AI practitioners prefer to have tools for prototyping and rapid testing, or for simulating the behavior of AI systems, to assess fairness issues at the system level rather than at the model level (Holstein et al., 2019). Consequently, I call for more research that examines the interdisciplinary needs of AI practitioners in different positions (e.g., design, technical, and project management roles). In addition, future research is needed to support AI practitioners break down responsible AI policies and to better disseminate knowledge about responsible AI in practice. To do this, scholars can focus on specific domains, use cases, and types of AI systems to provide context-specific responsible AI guidelines and goals. Furthermore, research can examine the use of participatory design practices in organizations and how these may be better integrated to foster collaboration across technical and user-facing roles and address ethical and technical knowledge gaps. A better understanding of these aspects can help AI practitioners to navigate the challenges that cause means-end decoupling.

5.2.3 Understanding the Institutional Work of AI Practitioners

My review indicates that AI practitioners can take on different roles in the implementation of responsible AI. Although the literature has provided first insights into the institutional work of AI practitioners, more research is needed to understand the impact of their work as institutional custodians or institutional entrepreneurs and the strategies they are using to sustain or change the institutions around the development and deployment of AI systems. Because previous research has largely interviewed AI practitioners from different organizations, it has not provided a deep dive into the work of AI practitioners in specific contexts. Hence, future research can take a case study approach to get a detailed understanding of the impact and strategies of AI practitioners who are inhibiting or championing the implementation of responsible AI. This approach can also help to better understand the institutional logics that underpin the implementation of responsible AI. While my literature review has provided initial insights into these, future research needs to further explore which logics are in place and how AI practitioners interpret and reproduce them in their daily work. Findings in this area can shed light on the motivations behind the institutional work of AI practitioners and how their activities can maintain or change AI development or deployment practices.

Because most of the existing research on the perspectives and work practices of AI practitioners has not considered the environment in which they work (e.g., organizational culture, hierarchical structure), there is a lack of understanding of how these factors facilitate and affect the work of AI practitioners. Thus, future work should analyze how organizational characteristics may influence the actions and success of institutional custodians and entrepreneurs. For example, my findings suggest that power dynamics or personal relationships within an organization play an important role in whether or not the institutional work of AI practitioners is successful. In exploring these aspects, future research could further investigate the

coexistence of institutional custodians and entrepreneurs to gain insights into how they negotiate or collaborate in the implementation of responsible AI. Finally, future research can focus on examining the impact of professional AI practitioner communities on the implementation of responsible AI. Many AI practitioners participate in these communities to share knowledge, discuss technical developments, and adopt technical practices. By actively participating in these communities, AI practitioners engaged in institutional work may also play a role in reinforcing or reshaping the norms, values, and institutional logic that define these communities. As these logics evolve, they may in turn influence how other practitioners approach their work, potentially affecting the broader institutional environment. Thus, studying how AI practitioner communities may shape the perspectives and actions of AI practitioners, and how AI practitioners may, in turn, influence the logic and norms of the communities, can help to better understand how responsible AI is diffused and institutionalized across organizations.

5.3 Practical Implications

My study has several practical implications for organizations and policymakers. My findings suggest that to successfully implement responsible AI practices, firms must involve institutional custodians in the change process, address conflicting institutional logics on responsible AI, and ensure that the incentives for AI practitioners are aligned with their responsible AI policies. In addition, they should support bottom-up responsible AI structures initiated by institutional entrepreneurs and provide AI practitioners with the necessary resources and legitimacy within their teams to effectively implement their responsible AI policies. Furthermore, companies' responsible AI practices must cover the complete lifecycle of an AI system and responsible AI policies should be addressed proactively. To achieve this, companies need to foster better collaboration between different stakeholders to create a common understanding of the risks and goals of responsible AI. My results reveal that centralized responsible AI teams are often isolated and that the risks of AI systems are highly context-dependent. Hence, organizations should ensure that each development team has at least one responsible AI expert who is familiar with the system and its potential risks. These decentralized responsible AI experts can be supported by centralized responsible AI resources (e.g., experts focusing on company-wide data governance) and practices (e.g., impact assessments). With this combination of centralized and decentralized responsible AI structures, development teams can make their own decisions about how best to address the risks of the AI systems they are building, while still having access to enterprise-wide resources and best practices to help them achieve responsible AI goals. Concerning policy implications, my findings imply that current regulations are too weak to encourage companies to effectively implement responsible AI. Moreover, the uncertainty due to missing regulations can further negatively affect the implementation. Accordingly, stronger regulations with explicit accountability mechanisms are needed. However, these regulations should take into account the complexity of different AI risks and should not impose ineffective goals and additional bureaucratization on companies, to not expand the mismatch between responsible AI means and ends.

5.4 Limitations

My research is subject to several limitations. First, my review is restricted to studies focusing on a specific topic (i.e., the implementation of responsible AI in the industry), method (i.e., empirical research), time frame (i.e., after 2015), and quality criteria (i.e., peer-reviewed). While I have chosen these criteria to gain in-depth insights into how organizations adopt and implement responsible AI, I may have overlooked relevant work from non-empirical papers and gray literature. Moreover, I may have excluded papers in the study selection process that do not explicitly mention dealing with responsible AI in organizations. Second, despite focusing on peer-reviewed articles I did not assess the studies in my review for explicit quality criteria. Hence, the quality and rigor of the studies in my review may vary, which can influence the strength of my findings. Third, due to the subjective nature of grounded theory data analysis, my interpretation of the data may be questioned. Nevertheless, to increase the credibility of my interpretations, I present the full data structure in Appendix B. Fourth, I used neo-institutional theory as a lens to study the implementation of responsible AI in organizations. Thus, my results depend on that lens, and they would have been different if I had used another theoretical perspective. However, I find this lens a valuable tool for better understanding why responsible AI practices in organizations are often not effectively implemented. Fifth, conducting my review at the intersection of an organizing and broad theorizing review (Leidner, 2018) the focus of my analysis was on breadth rather than depth. Accordingly, my findings do not provide the detail that a narrower focus and in-depth analysis using the grounded theory approach might have provided. Nevertheless, I believe that my results provide a good synthesis of the literature and offer opportunities for future research to explore the topic further.

6 Conclusion

The results of this paper show that while the public discussion of AI risks has led firms to recognize the importance of developing AI responsibly, pursuing the goals of responsible AI is still too often seen as conflicting with performance or project-oriented goals. Therefore, stronger market and regulatory pressures are needed to set incentives for organizations to align their development practices with their responsible AI policies. To overcome the challenges that lead to means-end decoupling, companies need to establish responsible AI structures that enable AI practitioners to integrate ethical considerations into their daily work. I suggest that this can be done by establishing a mix of centralized and decentralized responsible AI structures. This approach gives AI practitioners the freedom to decide how to manage the risks of the AI systems they develop or deploy, while also supporting them with enterprise-wide resources and best practices. Overall, it can be concluded that significant progress has been made in research and practice to enable companies to develop and deploy AI systems responsibly. However, more work is needed by academia, policymakers, and industry to further research, develop, and implement measures that effectively manage the risks associated with AI systems.

Acknowledgments

The author would like to thank Timo Mayer for his assistance with the literature search and screening process and Xhovana Prenga for her careful proofreading of the manuscript. Furthermore, the author would like to thank Sven Laumer and Sebastian Schötteler for their helpful comments and suggestions on earlier versions of the paper.

Declaration of AI

During the preparation and revision of this work the author used ChatGPT and DeepL Write in order to proofread the manuscript. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

- Agbese, M., Mohanani, R., Khan, A., & Abrahamsson, P. (2023). Implementing AI ethics: Making sense of the ethical requirements. *Proceedings of the 27th International Conference on Evaluation and Assessment in Software Engineering* (pp. 62–71).
- Ali, S. J., Christin, A., Smart, A., & Katila, R. (2023). Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. *2023 ACM Conference on Fairness, Accountability, and Transparency*, 217–226.
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica. Retrieved from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management*, 62, 102433.
- Attard-Frost, B., De los Ríos, A., & Walters, D. R. (2022). The ethics of AI business practices: A review of 47 AI ethics guidelines. *AI and Ethics*, 3(2), 1-18.
- Avnoon, N., Kotliar, D. M., & Rivnai-Bahir, S. (2023). Contextualizing the ethics of algorithms: A socio-professional approach. *New Media & Society*, 26(10), 5962-5982.
- Baker-Brunnbauer, J. (2021). Management perspective of ethics in artificial intelligence. *AI and Ethics*, 1(2), 173–181.
- Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development*, 63(4/5), 4:1-4:15
- Birkstedt, T., Minkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, 33(7), 133–167.
- Boxenbaum, E., & Jonsson, S. (2017). Isomorphism, diffusion and decoupling: Concept evolution and theoretical challenges. In R. Greenwood, C. Oliver, T. Lawrence, & R. Meyer (Eds.), *The SAGE handbook of organizational institutionalism* (pp. 77–97). SAGE Publications Ltd.
- Bromley, P., & Powell, W. W. (2012). From smoke and mirrors to walking the talk: Decoupling in the contemporary world. *Academy of Management Annals*, 6(1), 483–530.
- Candelon, F., Charme di Carlo, R., De Bondt, M., & Evgeniou, T. (2021, September). *AI regulation is coming*. Harvard Business Review. Retrieved from <https://hbr.org/2021/09/ai-regulation-is-coming>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
- Colombero, S., & Boxenbaum, E. (2019). Authentication as institutional maintenance work. *Journal of Management Studies*, 56(2), 408–440.
- Currie, G., Lockett, A., Finn, R., Martin, G., & Waring, J. (2012). Institutional work to maintain professional power: Recreating the model of medical professionalism. *Organization Studies*, 33(7), 937–962.
- Dacin, T. M., & Dacin, P. A. (2008). Traditions as institutionalized practice: Implications for deinstitutionalization. In *The SAGE handbook of organizational institutionalism* (pp. 326–351). SAGE.
- De Laat, P. B. (2021). Companies committed to responsible AI: From principles towards implementation and regulation? *Philosophy & Technology*, 34(4), 1135–1193.
- Deng, W. H., Yildirim, N., Chang, M., Eslami, M., Holstein, K., & Madaio, M. (2023). Investigating practices and opportunities for cross-functional collaboration around AI fairness in industry practice. *2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 705–716).
- Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer International Publishing AG.

- Dimaggio, P. (1988). Interest and agency in institutional theory. In *Research on institutional patterns: Environment and culture* (pp. 3–21). Ballinger Publishing Co.
- Figueras, C., Verhagen, H., & Cerratto Pargman, T. (2022). Exploring tensions in responsible AI in practice: An interview study on AI practices in and for Swedish public organizations. *Scandinavian Journal of Information Systems*, 34(2).
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.
- Freelon, D. (2010). *ReCal2: Reliability for 2 coders*. Retrieved from <http://dfreelon.org/utills/recalfront/recal2/>
- Ghatar, P. A., Pappas, I., & Vassilakopoulou, P. (2023). Practices for responsible AI: Findings from interviews with experts. *Proceedings of the 29th Americas Conference on Information Systems*.
- Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods*, 16(1), 15–31.
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Routledge.
- Grant, N., & Weise, K. (2023, April 7). In *A.I. race, Microsoft and Google choose speed over caution*. The New York Times. Retrieved from <https://www.nytimes.com/2023/04/07/technology/ai-chatbots-google-microsoft.html>
- Greenwood, R., & Suddaby, R. (2006). Institutional entrepreneurship in mature fields: The Big Five accounting firms. *Academy of Management Journal*, 49(1), 27–48.
- Griffin, T. A., Green, B. P., & Welie, J. V. M. (2024a). The ethical agency of AI developers. *AI and Ethics*, 4(2), 179–188.
- Griffin, T. A., Green, B. P., & Welie, J. V. M. (2024b). The ethical wisdom of AI developers. *AI Ethics*, 5, 1087–1097
- Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods*, 11(2), 181–217.
- Hardy, C., & Maguire, S. (2017). Institutional entrepreneurship and change in fields. In R. Greenwood, C. Oliver, T. Lawrence, & R. Meyer (Eds.), *The SAGE handbook of organizational institutionalism* (pp. 261–280). SAGE Publications Ltd.
- Hartikainen, M., Väänänen, K., Lehtiö, A., Ala-Luopa, S., & Olsson, T. (2022). Human-centered AI design in reality: A study of developer companies' practices. *Nordic Human-Computer Interaction Conference* (pp. 1–11).
- Heger, A. K., Marquis, L. B., Vorvoreanu, M., Wallach, H., & Wortman Vaughan, J. (2022). Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 1–29.
- Hinton, C. (2023). The state of ethical AI in practice: A multiple case study of Estonian public service organizations. *International Journal of Technoethics*, 14(1), 1–15.
- Hirsch, P. M., & Bermiss, Y. S. (2009). Institutional “dirty” work: Preserving institutions through strategic decoupling. In T. B. Lawrence, R. Suddaby, & B. Leca (Eds.), *Institutional work* (pp. 262–283). Cambridge University Press.
- Hoffman, A. J. (1999). Institutional evolution and change: Environmentalism and the U.S. chemical industry. *Academy of Management Journal*, 42(4), 351–371.
- Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).

- Hopkins, A., & Booth, S. (2021). Machine learning practices outside big tech: How resource constraints challenge responsible development. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 134–145).
- Ibáñez, J. C., & Olmeda, M. V. (2022). Operationalising AI ethics: How are companies bridging the gap between practice and principles? An exploratory study. *AI & Society*, 37(4), 1663–1687.
- Jabbouri, R., Truong, Y., Schneckenberg, D., & Palmer, M. (2019). Organizational decoupling: A systematic literature review and directions for future research. *EURAM Proceedings*. EURAM, Lisbon, Portugal.
- Jepperson, R. L., & Meyer, J. W. (2021). *Institutional theory: The cultural construction of organizations, states, and identities*. Cambridge University Press.
- Jobin, A., Lenca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399.
- Kelley, S. (2022). Employee perceptions of the effective adoption of AI principles. *Journal of Business Ethics*, 178(4), 871–893.
- Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M., & Abrahamsson, P. (2023). AI ethics: An empirical study on the views of practitioners and lawmakers. *IEEE Transactions on Computational Social Systems* (pp. 1–14).
- Lancaster, C. M., Schulenberg, K., Flathmann, C., McNeese, N. J., & Freeman, G. (2023). “It’s everybody’s role to speak up... but not everyone will”: Understanding AI professionals’ perceptions of accountability for AI bias mitigation. *ACM Journal on Responsible Computing*, 1(1), 5.
- Lawrence, T. B., & Suddaby, R. (2006). Institutions and institutional work. In S. Clegg, C. Hardy, T. Lawrence, & W. Nord (Eds.), *The SAGE handbook of organization studies* (pp. 215–254). SAGE Publications Ltd.
- Leidner, D. (2018). Review and theory symbiosis: An introspective retrospective. *Journal of the Association for Information Systems*, 19(6), 552–567.
- Lepoutre, J. M. W. N., & Valente, M. (2012). Fools breaking out: The role of symbolic and material immunity in explaining institutional nonconformity. *Academy of Management Journal*, 55(2), 285–313.
- Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9), 1–46.
- Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36–43.
- Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: Explaining and justifying inscrutable algorithms. *Ethics and Information Technology*, 23(3), 253–263.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768–4777).
- Maguire, S., Hardy, C., & Lawrence, T. B. (2004). Institutional entrepreneurship in emerging fields: HIV/AIDS treatment advocacy in Canada. *Academy of Management Journal*, 47(5), 657–679.
- Mayer, A.-S., Haimerl, A., Strich, F., & Fiedler, M. (2021). How corporations encourage the implementation of AI ethics. *Proceedings of the 29th European Conference on Information Systems*. European Conference of Information Systems (ECIS), Marrakech, Morocco.
- Merhi, M. I. (2023). An assessment of the barriers impacting responsible artificial intelligence. *Information Systems Frontiers*, 25(3), 1147–1160.
- Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83(2), 340–363.
- Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and ‘the dark side’ of AI. *European Journal of Information Systems*, 31(3), 257–268.

- Minkkinen, M., Zimmer, M. P., & Mäntymäki, M. (2023). Co-shaping an ecosystem for responsible AI: Five types of expectation work in response to a technological frame. *Information Systems Frontiers*, 25(1), 103–121.
- Mirbabaie, M., Brendel, A. B., & Hofeditz, L. (2022). Ethics and AI in information systems research. *Communications of the Association for Information Systems*, 50(1), 726–753.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229).
- Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507.
- Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2), 1–21.
- Montgomery, A. W., & Dacin, M. T. (2020). Water wars in Detroit: Custodianship and the work of institutional renewal. *Academy of Management Journal*, 63(5), 1455–1484.
- Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: Barriers, enablers and next steps. *AI & Society*, 38(1), 411–423.
- Mutch, A. (2007). Reflexivity and the institutional entrepreneur: A historical exploration. *Organization Studies*, 28(7), 1123–1140.
- Nili, A., Tate, M., Barros, A., & Johnstone, D. (2020). An approach for selecting and using a method of inter-coder reliability in information management research. *International Journal of Information Management*, 54, 102154.
- Orlikowski, W. J., & Barley, S. R. (2001). Technology and institutions: What can research on information technology and research on organizations learn from each other? *MIS Quarterly*, 25(2), 145–165.
- Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by artificial intelligence practitioners. *Information, Communication & Society*, 23(5), 719–735.
- Pant, A., Hoda, R., Spiegler, S. V., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in the age of AI: An analysis of AI practitioners' awareness and challenges. *ACM Transactions on Software Engineering and Methodology*, 33(3), 1–35.
- Pant, A., Hoda, R., Tantithamthavorn, C., & Turhan, B. (2024). Ethics in AI through the practitioner's view: A grounded theory literature review. *Empirical Software Engineering*, 29(3), 67.
- Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers*, 25(1), 123–141.
- Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), 101885.
- Popova, K., Figueras, C., Höök, K., & Lampinen, A. (2024). Who should act? Distancing and vulnerability in technology practitioners' accounts of ethical responsibility. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1), 1–27.
- Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–23.
- Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., Newnham, G., Hajkovicz, S., Robinson, C., & Hansen, D. (2023). AI ethics principles in practice: Perspectives of designers and developers. *IEEE Transactions on Technology and Society*, 4(2), 171–187.
- Sanderson, C., Lu, Q., Douglas, D., Xu, X., Zhu, L., & Whittle, J. (2022). Towards implementing responsible AI. *2022 IEEE International Conference on Big Data (Big Data)*, 5076–5081.
- Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2022). Artificial intelligence governance for businesses. *Information Systems Management*, 1–21.

- Schnyder, G. (2018). Investigating new types of “decoupling”: MSP in law and corporate practice. *Academy of Management Proceedings*, 2018(1), 16963.
- Schwartz, O. (2019, November 25). In 2016, Microsoft's racist chatbot revealed the dangers of online conversation. *IEEE Spectrum*. <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>
- Scott, W. R. (2014). *Institutions and organizations: Ideas, interests, and identities* (4th ed.). SAGE.
- Seele, P., & Schultz, M. D. (2022). From greenwashing to machinewashing: A model and future directions derived from reasoning by analogy. *Journal of Business Ethics*, 178(4), 1063–1089.
- Seguel, P., & Vaast, E. (2021). Repertories of evaluation in AI ethics: Plurality in professional responsibility and accountability. *Proceedings of the 42nd International Conference on Information Systems (ICIS)*. Austin, Texas, United States.
- Seo, M.-G., & Creed, W. E. D. (2002). Institutional contradictions, praxis, and institutional change: A dialectical perspective. *The Academy of Management Review*, 27(2), 222–247.
- Seppälä, A., Birksted, T., & Mäntymäki, M. (2021). From ethical AI principles to governed AI. *ICIS 2021 Proceedings*. Austin, Texas, United States.
- Siebert, S., Wilson, F., & Hamilton, J. R. A. (2017). “Devils may sit here:” The role of enchantment in institutional maintenance. *Academy of Management Journal*, 60(4), 1607–1632.
- Sloane, M., & Zakrzewski, J. (2022). German AI start-ups and “AI ethics”: Using a social practice lens for assessing and implementing socio-technical innovation. *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 935–947).
- Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K., & Jiya, T. (2022). Organisational responses to the ethical issues of artificial intelligence. *AI & Society*, 37(1), 23–37.
- Timmermans, S., & Tavory, I. (2012). Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory*, 30(3), 167–186.
- Treacy, S. (2023). Mechanisms and constraints underpinning ethically aligned artificial intelligence systems: An exploration of key performance areas. *Human-Centric Intelligent Systems*, 3(3), 189–196.
- Vakkuri, V., Kemell, K.-K., Kultanen, J., & Abrahamsson, P. (2020). The current state of industrial practice in artificial intelligence ethics. *IEEE Software*, 37(4), 50–57.
- Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., & Abrahamsson, P. (2022). Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *Electronic Journal of Business Ethics and Organization Studies*, 1.
- Vakkuri, V., Kemell, K.-K., Tolvanen, J., Jantunen, M., Halme, E., & Abrahamsson, P. (2022). How do software companies deal with artificial intelligence ethics? A gap analysis. *The International Conference on Evaluation and Assessment in Software Engineering 2022*, 100–109.
- Varanasi, R. A., & Goyal, N. (2023). “It is currently hodgepodge”: Examining AI/ML practitioners’ challenges during co-production of responsible AI values. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–17).
- Vassilakopoulou, P., Parmiggiani, E., Shollo, A., & Grisot, M. (2022). Responsible AI: Concepts, critical perspectives and an information systems research agenda. *Scandinavian Journal of Information Systems*, 34(2), 89–104.
- VERBI Software. (2023). *MAXQDA 2024* [Computer software]. maxqda.com
- Wang, Q., Madaio, M., Kane, S., Kapania, S., Terry, M., & Wilcox, L. (2023). Designing responsible AI: Adaptations of UX practice to meet responsible AI challenges. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–16).
- Webster, J., & Watson, R. T. (2002). Guest editorial: Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2), xiii–xxiii.

- Widder, D. G., & Nafus, D. (2023). Dislocated accountabilities in the “AI supply chain”: Modularity and developers’ notions of responsibility. *Big Data & Society*, 10(1), 1–13.
- Wijen, F. (2014). Means versus ends in opaque institutional fields: Trading off compliance and achievement in sustainability standard adoption. *Academy of Management Review*, 39(3), 302–323.
- Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, 22(1), 45–55.
- Wright, A. L., Meyer, A. D., Reay, T., & Staggs, J. (2021). Maintaining places of social inclusion: Ebola and the emergency department. *Administrative Science Quarterly*, 66(1), 42–85.
- Xivuri, K., & Twinomurizi, H. (2023). How AI developers can assure algorithmic fairness. *Discover Artificial Intelligence*, 3(1), 27.

Appendix A: Overview of Studies

Analyzing the study profiles of the articles in my review, I find that research on the topic of responsible AI in organizations is still at an early stage with all studies being published between 2019 and 2024. Most studies were published in journals and conferences that focus on the intersection of information technology and societal issues (16) (e.g., ACM FAccT, AI & Society, AI and Ethics), highlighting the interdisciplinary nature of the topic. The remaining studies were published in the fields of computer science and software engineering (9), information systems (7), human-computer interaction (6), and business ethics (1). In terms of methodology, 34 of the articles used a qualitative approach. Most of these studies (31) conducted semi-structured interviews with AI practitioners. Among these 31 studies, three conducted semi-structured interviews, and workshops or focus groups with AI practitioners. Out of the three remaining studies, one utilized a survey to collect qualitative data via open-ended questions, one applied the analytic hierarchy process as a research method, and one conducted only workshops. In addition, five studies used a survey as a research method, of which two used it in combination with a qualitative approach, while three used the survey as a stand-alone method. The articles sampled a broad range of AI practitioners across different hierarchy levels working in companies of various sizes (e.g., small, medium-sized, large) and across different domains. Table A1 presents an overview of all studies included in the review.

Table A1. Overview of Studies

Author(s)	Research method	Sample	Codes derived
Agbese et al. (2023)	Qualitative, semi-structured interviews	Ten AI practitioners from various organizations in leadership positions (e.g., Product Management Director)	Demand for responsible AI, Legal compliance, Reputational concerns, Demand for functional AI, AI risk blind spots
Ali et al. (2023)	Qualitative, semi-structured interviews	25 AI practitioners from various organizations in different positions (e.g., Engineers, Program Managers)	Lack of formal or informal accountability, Responsibility avoidance, Lack of leadership support, Lack of authority, Lack of resources, Measuring and monitoring the impact of responsible AI, Operationalization of responsible AI policies, Fragmented responsible AI structures, Reactive responsible AI structures, Reorganizations, Disincentivization of responsible AI, Occupational stress, Goal conflicts between business and responsible AI goals, Leveraging relationships to influence leadership, Reframing ethical issues, Sensitizing and educating AI practitioners
Avnoon et al. (2023)	Qualitative, semi-structured interviews	25 AI practitioners (50 data scientists and five managers of data scientists) from various organizations, and five university professors teaching data scientists	Lack of authority, AI risk blind spots, Personal ethos, Professional ethos, Goal conflicts between business and responsible AI goals
Baker-Brunnbauer (2021)	Qualitative, semi-structured interviews	Nine AI practitioners from various organizations in leadership positions	Reputational concerns, Legal compliance, Lack of formal or informal accountability, Demand for functional AI, AI risk blind spots, Responsibility avoidance, Goal conflicts between business and responsible AI goals, Fragmented responsible AI structures
Deng et al. (2023)	Qualitative, semi-structured interviews, and workshops	Interviews: 17 AI practitioners from 12 organizations in different positions Workshops: Five workshops attended by 12 AI practitioners in different positions (six of whom had already participated in the interviews and six of	Responsibility avoidance, Lack of resources, Measuring and monitoring the impact of responsible AI, AI risk blind spots, Operationalization of responsible AI policies, Disincentivization of responsible AI, Personal ethos, Sensitizing and educating AI practitioners, Bridging

		whom only participated in the workshops)	activities, Piggybacking, Occupational stress, Additional labor
Figueras et al. (2022)	Qualitative, semi-structured interviews	13 AI practitioners from six organizations (all in the public sector) in different positions (e.g., AI Researcher, Product Owner)	Lack of resources, Goal conflicts between responsible AI goals, Operationalization of responsible AI policies, Lack of tools and methods, Goal conflicts between business and responsible AI goals, Measuring and monitoring the impact of responsible AI
Ghatar et al. (2023)	Qualitative, semi-structured interviews	12 AI practitioners from various organizations in different positions	Demand for responsible AI, Lack of resources, Contextualization of responsible AI goals, Operationalization of responsible AI policies, Unpredictability of AI risk, Fragmented responsible AI structures
Griffin et al. (2024a)	Qualitative, semi-structured interviews	40 AI developers from various organizations	Lack of formal or informal accountability, Lack of authority, AI risk blind spots, Goal conflicts between responsible AI goals, Operationalization of responsible AI policies, Disincentivization of responsible AI, Technology value-neutrality, Personal ethos, Professional ethos
Griffin et al. (2024b)	Qualitative, semi-structured interviews	40 AI developers from various organizations	Reputational concerns, Lack of leadership support, AI risk blind spots, Responsibility avoidance, Goal conflict between responsible AI goals, Technology value-neutrality, Goal conflicts between business and responsible AI goals, Disincentivization of responsible AI, Personal ethos
Hartikainen et al. (2022)	Qualitative, semi-structured interviews	12 AI practitioners from various organizations in different positions	Demand for functional AI, Lack of resources, Unpredictability of AI risk, Reactive responsible AI structures
Hinton (2023)	Qualitative, semi-structured interviews	Eight AI practitioners from various organizations (all public) in different positions (e.g., Developer, Project Lead)	Lack of formal or informal accountability, Legal compliance, Lack of resources, AI risk blind spots, Reactive RAI structures, Goal conflicts between responsible AI goals
Holstein et al. (2019)	Mixed-method, semi-structured interviews, and a survey	Interviews: 35 AI practitioners from ten organizations in different positions (e.g., ML Engineer, UX Researcher) Survey: 267 AI practitioners from different companies in different positions	Demand for functional AI, Lack of resources, AI risk blind spots, Unpredictability of AI risk, Contextualization of responsible AI goals, Lack of tools and methods, Measuring and monitoring the impact of responsible AI, Fragmented responsible AI structures, Disincentivization of responsible AI, Technology value-neutrality, Additional labor
Hopkins and Booth (2021)	Qualitative, semi-structured interviews	17 AI practitioners from various organizations in different positions (e.g., Chief Technology Officer, Data Engineer)	Lack of formal or informal accountability, Lack of resources, Responsibility avoidance, Unpredictability of AI risk, AI risk blind spots, Goal conflicts between business and responsible AI goals, Lack of tools and methods, Goal conflicts between responsible AI goals
Ibáñez and Olmeda (2022)	Qualitative, semi-structured interviews, and focus groups	Interviews: 22 AI practitioners from various organizations in leadership positions (e.g., Head of AI, AI Consulting Director)	Demand for functional AI, Lack of formal or informal accountability, Legal compliance, Reputational concerns, Operationalization of

		Focus groups: Two focus groups attended by ten AI practitioners who already participated in the interviews	responsible AI policies, Contextualization of responsible AI goals, Fragmented responsible AI structures, Personal ethos
Kelley (2022)	Qualitative, semi-structured interviews	49 AI practitioners from 24 organizations (all in the financial industry) in different positions (e.g., Junior Data Scientist, Chief Analytics Officer)	Lack of formal or informal accountability, Lack of leadership support, Lack of resources, Operationalization of responsible AI policies, Contextualization of responsible AI goals, Measuring and monitoring the impact of responsible AI, Fragmented responsible AI structures, Disincentivization of responsible AI
Khan et al. (2023)	Quantitative, survey	70 AI Practitioners from various organizations in different positions (e.g., AI Engineer, Project Manager) and 29 legislators	Lack of formal or informal accountability, Goal conflicts between responsible AI goals, Contextualization of responsible AI goals, Operationalization of responsible AI policies, Lack of tools and methods,
Lancaster et al. (2023)	Qualitative, semi-structured interviews	20 AI practitioners from various organizations in different positions (e.g., Front-End Developer, Data Scientist)	Demand for functional AI, Lack of resources, Lack of leadership support, Responsibility avoidance, Lack of authority, Personal ethos, Professional ethos, Disincentivization of responsible AI, Technology value-neutrality, Inevitability of AI risks, Goal conflicts between business and responsible AI goals
Mayer et al. (2021)	Qualitative, semi-structured interviews	Nine AI practitioners from six organizations in leadership positions (e.g., AI Strategy Expert, Product Manager)	Lack of leadership support, Measuring and monitoring the impact of responsible AI
Merhi (2023)	Mixed-method: Analytic Hierarchy Process	Seven AI practitioners (four project managers, two system analysts, one data analyst) from various organizations	Demand for functional AI, Lack of formal or informal accountability mechanism, Lack of leadership support, Lack of resources, Operationalization of responsible AI policies, Goal conflicts between responsible AI goals, Unpredictability of AI risk
Morley et al. (2023)	Mixed-method, semi-structured interviews, and a survey	Interviews: Six AI practitioners from various organizations in different positions Survey: 54 AI practitioners from various organizations in different positions	Lack of formal or informal accountability mechanism, Demand for responsible AI, Legal compliance, Unclear roles and responsibilities, Lack of leadership support, Lack of resources, AI risk blind spots, Operationalization of responsible AI policies, Personal ethos, Occupational stress, Lack of tools and methods
Orr and Davis (2020)	Qualitative, semi-structured interviews	21 AI practitioners from various organizations in different positions	Legal compliance, Reputational concerns, Responsibility avoidance, Lack of authority, Lack of resources, AI risk blind spots, Inevitability of AI risks, Goal conflicts between business and responsible AI goals, Disincentivization of responsible AI, Professional ethos
Pant et al. (2024)	Quantitative, survey	100 AI practitioners from various organizations in different positions (e.g., Data Scientist, AI Engineer)	Lack of resources, AI risk blind spot, Unpredictability of AI risk, Goal conflicts between responsible AI goals, Operationalization of responsible AI policies, Contextualization of responsible AI

			goals, Measuring and monitoring the impact of responsible AI, Personal ethos, Professional ethos, Goal conflicts between business and responsible AI goals
Papagiannidis et al. (2023)	Qualitative, semi-structured interviews (multiple case study)	15 AI practitioners from three organizations in different positions (e.g., Chief AI Officer, Machine Learning Engineer).	Unclear roles and responsibilities, Lack of resources, Fragmented responsible AI structures
Popova et al. (2024)	Qualitative, semi-structured interviews	23 AI practitioners from various organizations (private and public) in different positions (e.g., Associate Professor, Product Owner).	Lack of resources, Responsibility avoidance, Lack of authority, AI risk blind spots, Measuring and monitoring the impact of responsible AI, Technology value-neutrality, Personal ethos, Professional ethos, Occupational stress
Rakova et al. (2021)	Qualitative, semi-structured interviews, and one workshop	Interviews: 26 AI practitioners from 19 organizations in different positions (e.g., AI Strategy, Legal, Engineering).	Lack of formal or informal accountability, Legal compliance, Reputational concerns, Unclear roles and responsibilities, Lack of leadership support, Lack of authority, Lack of resources, Measuring and monitoring the impact of responsible AI, Goal conflicts between responsible AI goals, Disincentivization of responsible AI, Personal ethos, Professional ethos, Leveraging relationships to influence leadership, Occupational stress, Additional labor
Ryan et al. (2022)	Mixed-method, systematic literature analysis, and three workshops	Three workshops with 19 AI practitioners from various organizations in different positions (e.g., AI Developer, Education)	Demand for responsible AI, Legal compliance, Reputational concerns, Lack of leadership support, Lack of authority, Goal conflicts between responsible AI goals, AI risk blind spots, Operationalization of responsible AI policies, Goal conflicts between business and responsible AI goals, Technology value-neutrality, Personal ethos, Professional ethos, Occupational stress
Sanderson et al. (2023)	Qualitative, semi-structured interviews	21 AI practitioners from one organization (public) in different positions (e.g., Principal Science Engineer, Research Scientist)	Goal conflicts between responsible AI goals, Operationalization of responsible AI policies, Contextualization of responsible AI goals, Responsibility avoidance, Demand for functional AI, Personal ethos, Lack of resources
Sanderson et al. (2022)	Qualitative, semi-structured interviews	21 AI practitioners from one organization (public) in different positions (e.g., Principal Science Engineer, Research Scientist)	Lack of resources, Lack of tools and methods, Legal compliance, Operationalization of responsible AI policies, Unpredictability of AI risk, Goal conflicts between responsible AI goals, Measuring and monitoring the impact of responsible AI
Segal and Vaast (2021)	Qualitative, semi-structured interviews	62 AI practitioners from various organizations in different positions (e.g., Data Scientist, Software Engineer)	Responsibility avoidance, AI risk blind spots, Goal conflicts between responsible AI goals, Personal ethos, Professional ethos
Seppälä et al. (2021)	Qualitative, semi-structured interviews	13 AI practitioners from 12 organizations in leadership positions (e.g., Analytics Lead, Head of Data Scientist)	Unclear roles and responsibilities, Responsibility avoidance, AI risk blind spots
Sloane and Zakrzewski (2022)	Qualitative, semi-structured interviews	64 AI practitioners from various AI start-up organizations in different positions (e.g., (Co-)	Demand for responsible AI, Legal compliance, Reputational concerns, Personal ethos

		Founder, Accelerators)	
Stahl et al. (2022)	Qualitative, semi-structured interviews (multiple case study)	42 AI practitioners from ten organizations in different positions	Goal conflicts between responsible AI goals, Goal conflicts between business and responsible AI goals, Professional ethos
Treacy (2023)	Qualitative, semi-structured interviews	Ten AI practitioners from four organizations in different positions (e.g., Data Scientist, AI Engineer)	AI risk blind spots, Contextualization of responsible AI goals, Unpredictability of AI risk, Reactive responsible AI structures, Goal conflicts between business and responsible AI goals
Vakkuri et al. (2020)	Quantitative, survey	249 respondents from 211 organizations in different positions	Demand for functional AI, lack of formal or informal accountability mechanism, Responsibility avoidance, Personal ethos
Vakkuri et al. (2022)	Qualitative, survey, and open-ended questions	249 respondents from 39 organizations in different positions (e.g., AI Specialist, Software Engineer)	Legal compliance, AI risk blind spots, Goal conflicts between business and responsible AI goals
Varanasi and Goyal (2023)	Qualitative, semi-structured interviews	23 AI practitioners from ten organizations in different positions (e.g., Engineer, UX Designer)	Demand for responsible AI, Reputational concerns, Legal compliance, Demand for functional AI, Measuring and monitoring the impact of responsible AI, Contextualization of responsible AI goals, Operationalization of responsible AI policies, Goal conflicts between responsible AI goals, Reactive responsible AI structures, Fragmented responsible AI structures, Personal ethos, Professional ethos, Occupational stress, Additional labor
Wang et al. (2023)	Qualitative, semi-structured interviews	23 AI practitioners (15 UX practitioners and eight responsible AI experts) from one organization	Lack of resources, Lack of authority, Unpredictability of AI risk, Reactive responsible AI structures, Goal conflicts between business and responsible AI goals, Personal ethos, Sensitizing and educating AI practitioners, Occupational stress, Additional labor
Widder and Nafus (2023)	Qualitative, semi-structured interviews	27 AI practitioners from various organizations in different positions (e.g., ML Engineer, Project Manager)	Demand for responsible AI, Reputational concerns, Unclear roles and responsibilities, Responsibility avoidance, AI risk blind spots, Goal conflicts between business and responsible AI goals, Disincentivization of responsible AI, Technology value-neutrality, Personal ethos
Xivuri and Twinomurinzi (2023)	Qualitative, semi-structured interviews	10 AI developers from various organizations	Lack of formal or informal accountability mechanisms, Legal compliance, Lack of resources, AI risk blind spots, Fragmented responsible AI structures, Goal conflicts between business and responsible AI goals

Appendix B: Data Structure

Table B1. Data Structure

Excerpt	First-order codes	Second-order themes	Aggregate Dimensions
<i>"Ethical AI' was seen as a marketing advantage, with one participant suggesting that it is a 'very, very good influencing tool [where] users might choose [our company] over the competition."</i> (Widder & Nafus, 2023, p. 7)	Demand for responsible AI	Responsible AI implementation drivers	External pressures
<i>"The value of ethical requirements is assessed in terms of regulatory measures like GDPR."</i> (Agbese et al., 2023, p. 66)	Legal compliance		
<i>"The most prevalent incentives for action were catastrophic media attention and decreasing media tolerance for the status quo."</i> (Rakova et al., 2021, p. 7:10)	Reputational concerns		
<i>"However, principles of ... [human-centered AI] are not reflected in the ways of working, but values are focused on the technical excellence."</i> (Hartikainen et al., 2022, p. 8)	Demand for functional AI	Responsible AI implementation inhibitors	
<i>"The third most common challenging factor is lacking monitoring bodies, and it was highlighted by (68.7%) of the survey participants."</i> (Khan et al., 2023, p. 6)	Lack of formal or informal accountability mechanism		
<i>"The importance of clear roles and responsibilities regarding AI development and data governance was widely highlighted. However, these were still relatively unclear for many [...]"</i> (Seppälä et al., 2021, p. 8)	Unclear roles and responsibilities	Challenges of policy-practice decoupling	Challenges to the implementation of responsible AI
<i>"Here, ethics was not claimed to be unimportant, but participants acknowledged it as someone else's responsibility, be it a particular group of developers, ethicists or social scientists, users of the resulting system, or another group that the technology practitioners did not belong to."</i> (Popova et al., 2024, p. 157:13)	Responsibility avoidance		
<i>"Many participants mentioned the main problem they faced, which came with several downstream effects: a lack of support from leadership."</i> (Ali et al., 2023, p. 220)	Lack of leadership support		
<i>"The imbalance of knowledge and power created strife for some practitioners, who felt constrained in their own moral and ethical autonomy."</i> (Orr & Davis, 2020, p. 727)	Lack of authority		
<i>"Most of the participants noted that lack of time is a challenge they face in considering and following ethics during the development of AI-based systems"</i> (Pant et al., 2024, p. 13)	Lack of resources		
<i>"The results discussion demonstrates that AI practitioners have an abstract and relatively narrow understanding of ethical principles and how these can be translated into practice."</i> (Morley et al., 2023, p. 417)	AI risk blind spots	Challenges to means-end decoupling	
<i>"Implementation of RAI values was also not a straight forward process as implementing certain RAI values created conflict with other RAI values."</i> (Varanasi & Goyal, 2023, p. 9)	Goal conflicts between responsible AI goals		
<i>"At the same time, the motivation of individual AI practitioners who might be willing to effect change from the bottom up is undermined by a lack of conceptual clarity about the meaning of ethical principles"</i> (Morley et al., 2023, p. 418)	Operationalization of responsible AI policies		
<i>"The cross-cultural context where the AI system is used needs to be considered: 'The problem is that</i>	Contextualization of responsible AI		

<i>sometimes when we design AI, we don't take that care to think about which system it's going to be put into. We just do it as a prototype to then to be deployed everywhere'.</i> (Sanderson et al., 2023, p. 178)	policies		
<i>"Experts highlighted it was difficult make a system fully secure due to the constantly evolving threats that these systems are subjected to [...]"</i> (Treacy, 2023, p. 194)	Unpredictability of AI risk		
<i>"Moreover, a survey respondent explicitly considered the lack of tools for ethical transparency and AI biases as significant challenges to AI ethics."</i> (Khan et al., 2023, p. 6)	Lack of tools and methods		
<i>"Respondents reported that many challenges in their prevalent work practices arise from the inability to adequately use existing metrics to account for the goals of responsible AI work"</i> (Rakova et al., 2021, p. 7:12)	Measuring and monitoring the impact of responsible AI		
<i>"In general, there is an absence of formalisation of procedures and policies. Companies do not have a guide of their own with indications for the ethical design, development, Disincentivization of responsible AI and control of these systems."</i> (Ibáñez & Olmeda, 2022, p. 1767)	Fragmented responsible AI structures		
<i>"Another challenge practitioners faced was encountering new RAI values during late stages of implementation."</i> (Varanasi & Goyal, 2023, p. 9)	Reactive responsible AI structures		
<i>"Multiple participants mentioned the frequency of reorganizations, or "reorgs," which rearranged the structure of their teams and/or the teams they worked with."</i> (Ali et al., 2023, p. 223)	Reorganizations		
<i>"One of the biggest challenges practitioners reported was that as responsible AI ethical tensions are identified, overly rigid organizational incentives may demotivate addressing them, [...]"</i> (Rakova et al., 2021, p. 7:16)	Disincentivization of responsible AI		
<i>"Management is focusing on cost savings, growth, process optimization and sees the social impact of AI systems as a competitive element and not as a social problem."</i> (Baker-Brunnbauer, 2021, p. 179)	Goal conflicts between business and responsible AI goals		
<i>"The repeated refrain that the algorithm is 'just a tool' and it 'depends on how you use it' implies that many developers believe the technologies they use, develop, and deploy are morally neutral—neither good nor bad."</i> (Griffin et al., 2024a, p. 186)	Technology value-neutrality	Institutional custodians	
<i>"[...] other participants viewed bias as an inevitable outcome of human-designed systems. Because these participants consider bias unavoidable, they often do not feel obligated to or capable of acting against these biased outcomes."</i> (Lancaster et al., 2023, p. 5:12)	Inevitability of AI risks		Roles of AI practitioners
<i>"One of the enablers for such bottom-up innovation was individuals' sense of responsibility towards producing AI/ML models that did not contribute to any harm in society"</i> (Varanasi & Goyal, 2023, p. 5)	Personal ethos		
<i>"If their company valued taking social responsibility for their outputs, it made them more inclined to feel accountable for AI bias."</i> (Lancaster et al., 2023, p. 14)	Professional ethos	Institutional entrepreneurs	
<i>"If resources were not allocated for ethics work, ethics entrepreneurs would sometimes attempt to convince the product team to allocate funds out of their project budgets for these kinds of activities."</i> (Ali et al., 2023, p. 222)	Leveraging relationships to influence leadership		

<i>"Others were able to influence others by reframing any problem as an issue of product quality." (Ali et al., 2023, p. 222)</i>	Reframing ethical issues		
<i>"In those cases, ethics entrepreneurs had to spend time 'pre-emptively assuaging their fears' and make sure their suggestions did not 'make things harder for that team'." (Ali et al., 2023, p. 222)</i>	Sensitizing and educating AI practitioners		
<i>"Participants shared that they bridged these evaluation gaps by initiating and organizing meetings for cross-functional teams to align model-focused and user-focused fairness evaluations." (Deng et al., 2023, p. 708)</i>	Bridging activities		
<i>"... we find that participants employed "piggybacking" as a tactic to push fairness work forward in organizations that might not otherwise provide resources or incentives for fairness work." (Deng et al., 2023, p. 710)</i>	Piggybacking		
<i>"In these cases, the interviewees were economically and emotionally vulnerable: while recognizing ethical concerns, they also recognized that they would face negative consequences for both acting and not acting [...]" (Popova et al., 2024, p. 157:16)</i>	Occupational stress		
<i>"Again, the prevalent work practices reveal individuals working on responsible AI taking on the extra labor of trying to translate their work into ill-fitting terms and metrics that are not designed to measure or motivate success on responsible AI outcomes." (Rakova et al., 2021, p. 7:13)</i>	Additional labor		

About the Authors

David Horneber is a PhD student in Information Systems at the Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany. His research interests include the responsible development and use of artificial intelligence systems, persuasive technologies in e-commerce, and authenticity-driven social networking sites. His work has been published in the *Business & Information Systems Engineering Journal* and various conference proceedings including ICIS and ECIS.

Copyright © 2025 by the Association for Information Systems. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than the Association for Information Systems must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or fee. Request permission to publish from: AIS Administrative Office, P.O. Box 2712 Atlanta, GA, 30301-2712 Attn: Reprints or via e-mail from publications@aisnet.org.