machine learning & knowledge extraction

MDPI

Article

Behind the Algorithm: International Insights into Data-Driven AI Model Development

Limor Ziv [1,2] and Maayan Nakash [2,*]

[1] School of Communication, Bar-Ilan University, Ramat Gan 5290002, Israel; limor.ziv@biu.ac.il
[2] Department of Management, Bar-Ilan University, Ramat Gan 5290002, Israel
* Correspondence: maayan.nakash@biu.ac.il

Abstract

Artificial intelligence (AI) is increasingly embedded within organizational infrastructures, yet the foundational role of data in shaping AI outcomes remains underexplored. This study positions data at the center of complexity, uncertainty, and strategic decision-making in AI development, aligning with the emerging paradigm of data-centric AI (DCAI). Based on in-depth interviews with 74 senior AI and data professionals, the research examines how experts conceptualize and operationalize data throughout the AI lifecycle. A thematic analysis reveals five interconnected domains reflecting sociotechnical and organizational challenges—such as data quality, governance, contextualization, and alignment with business objectives. The study proposes a conceptual model depicting data as a dynamic infrastructure underpinning all AI phases, from collection to deployment and monitoring. Findings indicate that data-related issues, more than model sophistication, are the primary bottlenecks undermining system reliability, fairness, and accountability. Practically, this research advocates for increased investment in the development of intelligent systems designed to ensure high-quality data management. Theoretically, it reframes data as a site of labor and negotiation, challenging dominant model-centric narratives. By integrating empirical insights with normative concerns, this study contributes to the design of more trustworthy and ethically grounded AI systems within the DCAI framework.

Keywords: data-centric AI; artificial intelligence; data quality; data governance; AI model development; AI lifecycle

Academic Editor: Isaac Triguero

Received: 1 September 2025
Revised: 3 October 2025
Accepted: 14 October 2025
Published: 17 October 2025

Citation: Ziv, L.; Nakash, M. Behind the Algorithm: International Insights into Data-Driven AI Model Development. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 122. https://doi.org/10.3390/make7040122

1. Introduction

Artificial intelligence (AI) has evolved from a specialized domain within computer science into a transformative force that is reshaping organizational structures, operational workflows, and user experiences across sectors. As predictive analytics, autonomous systems, and generative models gain widespread adoption, AI is becoming a core component of contemporary digital infrastructures [1–5]. At the heart of these systems lies data—not merely as a technical input but as a dynamic, context-sensitive asset that drives model development, influences performance outcomes, and shapes the ethical and social implications of AI deployment [6–8].

Yet despite this centrality, AI research and development continue to prioritize models over datasets. Kumar et al. [9] report that approximately 99% of academic AI research remains model-centric, even as many industry actors shift toward data-centric approaches to address real-world challenges. A similar critique has emerged from MIT

researchers, who argue that while 90% of academic machine learning (ML) efforts focus on algorithmic innovation, only 10% address data preparation and validation—even though practitioners devote most of their time to these tasks [10]. As Sambasivan et al. [5] pointedly observe, "paradoxically, data is the most undervalued and de-glamorised aspect of AI" (p. 1). This imbalance has significant implications: data work is often treated as routine or secondary, attracting limited institutional recognition or investment, which in turn undermines the reliability, fairness, and accountability of AI systems [5,11–15].

The oft-quoted phrase "data is the new oil," first coined by Clive Humby in 2006, captures the idea that raw data, like crude oil, must be processed and refined before it becomes valuable [16]. However, the analogy also emphasizes the need for continuous, real-time data flows to sustain digital systems. This perspective has informed the emergence of "data-centric AI" (DCAI): a paradigm that places dataset quality, contextual relevance, and representativeness at the center of AI system development and evaluation [7,8,17,18]. Within this framework, models are understood as data-driven learning systems whose predictive, classificatory, and decision-making capacities are only as effective as the data on which they are trained [9]. Complementary movements in responsible AI further highlight that data practices are inseparable from questions of governance, bias mitigation, traceability, and transparency [19,20].

Nevertheless, much of the scholarly discourse on data in AI remains conceptual, abstract, or technologically deterministic. While a few empirical studies have addressed the role of data in AI development (e.g., [5,14]), they largely focus on high-stakes applications and were conducted before the widespread emergence of generative artificial intelligence (GenAI) tools in daily professional contexts. Moreover, these studies frequently reduce data work to technical processes or ethical abstractions, without fully engaging with the sociotechnical and organizational dynamics that shape data practices in real-world settings [21].

In particular, there is limited empirical insight into how senior professionals—those who combine technical expertise with strategic responsibility—perceive and navigate the challenges of data in AI development. This includes aligning data practices with business goals, regulatory requirements, and user expectations. Despite its critical importance, the labor involved in curating, annotating, integrating, and contextualizing data remains largely invisible in dominant AI narratives. Yet these activities are foundational to ensuring system reliability, adaptability, and fairness [9,10,22,23].

This study addresses these gaps by foregrounding the experiences and insights of senior AI and data professionals working at the intersection of technical implementation and organizational decision-making. Their experiential knowledge offers a unique window into the complex trade-offs, tensions, and strategies that characterize data-intensive AI development. By focusing on how these professionals conceptualize, prioritize, and operationalize data, the research moves beyond theoretical idealizations to engage with the practical realities of building and deploying AI systems in complex organizational environments. Drawing on these insights, we developed the AI Lifecycle model, a conceptual framework that maps the evolving role of data across all stages of system design and deployment; from data collection and preparation to monitoring, explainability, and long-term system maintenance.

2. Research Question and Objectives

The central research question guiding this study is: *How do strategic professionals in AI and data-related roles experience and manage the role of data in developing AI-driven products and services?* To address this question, the study is structured around three interrelated research objectives:

RO1. To identify the key data-related challenges encountered during AI model training, whether developing new ones or adapting existing ones.

RO2. To examine how these challenges influence AI development and deployment processes.

RO3. To explore the strategies employed by professionals to mitigate data-related risks and constraints.

By foregrounding the perspectives of experts who navigate both technical and organizational facets of AI development, the study contributes a grounded and practice-oriented understanding of data's role in shaping AI outcomes. The findings aim to bridge the empirical gap in the current literature, inform theoretical discussions on the data-model relationship, and offer actionable insights for organizations seeking to design responsible and effective data-driven AI systems.

The remainder of the paper is structured as follows. Section 3 reviews the relevant literature that informs the study's conceptual framework. Section 4 outlines the research methodology, including the study design, sampling strategy, and procedures for data collection and analysis. Section 5 presents the empirical findings in detail, followed by Section 6, which offers a discussion of the results and introduces a data-centric interpretive framework developed in this study. Finally, Section 7 concludes with a summary of key insights, theoretical contributions, practical implications, and limitations that point toward directions for future research.

3. Background and Related Work

3.1. Evolving Capabilities of AI in a Data-Driven World

AI has long been defined as the science and engineering of creating machines capable of replicating human cognitive functions such as perception, reasoning, and learning [24]. This expansive field comprises several subdomains, including expert systems, ML, natural language processing (NLP), computer vision, and robotics [3,4,23]. Over time, AI has experienced a profound paradigm shift: moving from rule-based, symbolic reasoning approaches to probabilistic, data-driven models that learn from large datasets rather than explicit programming. This evolution represents a fundamental epistemological transformation, placing statistical inference and pattern recognition at the core of machine "understanding" [22,25–27]. These methodological and epistemological shifts have laid the groundwork for the most recent wave of AI advancements, which are now transforming the scale, scope, and impact of intelligent systems.

Recent years have witnessed a significant expansion in both the scope and sophistication of AI applications, propelled by three converging trends: the exponential growth of digital data, breakthroughs in deep learning architectures, and advances in high-performance computing infrastructure. Among the most notable recent developments, AI models are increasingly demonstrating multimodal and autonomous capabilities, enabling more context-aware reasoning and broadening the epistemic role of AI across diverse domains. Large Language Models (LLMs) and other GenAI systems now exhibit remarkable proficiency in tasks such as text generation, image synthesis, and complex reasoning. These capabilities are being deployed across an expanding range of sectors—including healthcare diagnostics, supply chain optimization, scientific discovery, and creative industries [28–32]—prompting a re-examination of both the nature and boundaries of machine intelligence.

3.2. The Expanding Footprint of AI Adoption in Organizational Contexts

AI technologies are increasingly being integrated into core organizational processes, reflecting their perceived strategic value in enhancing operational efficiency and

innovation [3,4,32–34]. Global surveys from McKinsey [35] and Deloitte [36] reveal that the majority of large enterprises have adopted AI for key functions such as customer service automation, fraud detection, predictive maintenance, and supply chain management. However, the degree and success of AI adoption vary widely across industries, organizational sizes, and levels of digital maturity. Key barriers include limited data infrastructure, resistance to cultural change, and workforce skill gaps [37–39]. Consequently, organizations are compelled to experiment with diverse implementation pathways, ranging from incremental process enhancements to transformative, enterprise-wide initiatives [38,40].

Importantly, AI adoption is not merely a technical implementation challenge but a multifaceted organizational process requiring alignment between technological capabilities, strategic vision, and human factors [4]. Employees often express anxiety over AI's impact on job security, raising concerns about displacement and role redefinition [40,41]. These realities are reflected in emerging AI maturity models, which trace organizational progression from isolated pilot projects to fully integrated AI strategies aligned with long-term business goals [39]. Additionally, effective AI deployment demands that models be trained on use-case-specific datasets, underscoring the critical role of data relevance and context in organizational outcomes [37].

3.3. Critical Risks at the Intersection of AI and Data

While AI holds transformative potential, it also introduces a complex array of ethical, technical, and societal risks [32,42,43]. Prominent concerns include algorithmic bias, lack of interpretability, opaque decision-making processes, unequal access to AI technologies, and threats to individual privacy and civil liberties [2,9,19,20,30,44]. Many of these challenges stem from the data used to train AI systems. Biased, incomplete, or non-representative datasets can result in discriminatory outcomes, particularly in sensitive domains such as hiring, lending, and criminal justice. The "black box" nature of many ML algorithms further complicates accountability, as it is often unclear how specific decisions are made [6,7,13].

Compounding these technical issues are concerns related to data governance and ethical data sourcing [32]. The use of synthetic or repurposed datasets—particularly those lacking transparency or informed consent—can expose organizations to reputational damage and regulatory sanctions [6,18]. In response, international regulatory frameworks such as the EU AI Act and the OECD AI Principles have emphasized the need for responsible data stewardship, human oversight, and algorithmic transparency [19,44,45]. Nevertheless, many organizations continue to struggle with balancing the pressures of rapid AI innovation against the demands of robust ethical governance and regulatory compliance.

3.4. Theoretical Lenses on the Role of Data in AI Systems

The increasing centrality of data in AI development has prompted a shift from model-centric to data-centric paradigms. In traditional model-centric approaches, efforts focus on refining algorithms and architectures to improve performance. By contrast, DCAI emphasizes the importance of data quality, structure, and contextual relevance in training effective models [9]. Without high-quality data, even the most advanced ML models are prone to underperform. DCAI proposes that refining datasets—through annotation, cleaning, augmentation, and validation—can yield greater improvements than adjustments to model design alone [7,8,18,46].

Recent scholarship also reconceptualizes data not as a neutral input, but as a form of socio-technical infrastructure embedded within institutional, political, and cultural frameworks [14,23]. From this perspective, data pipelines are shaped by human choices, including annotation practices, contextual assumptions, and organizational priorities [2,5]. This broader lens highlights the interdependence between data and decision-making systems,

reinforcing the view that AI performance is as much a function of social context as it is of technical design. Consequently, understanding data as infrastructure underscores the need for critical engagement with how data is sourced, structured, and mobilized in the service of AI.

3.5. Synthesis and Positioning of the Present Study

The synthesis presented in Table 1 highlights that, while prior scholarship has substantially advanced knowledge on AI's technical evolution, patterns of organizational adoption, and associated risks, the literature remains fragmented—particularly regarding the interplay between recent technological advances and data practices. The advent of generative models and LLMs has further amplified the centrality of data, introducing new layers of complexity, dependency, and vulnerability that existing research has only partially captured. Contemporary research increasingly recognizes that the capabilities and limitations of AI systems are not determined solely by model architectures but are profoundly shaped by data-related processes, infrastructures, and governance mechanisms [5,7–10,17,18]. Despite this growing recognition, empirical insights into how these challenges are experienced and managed by practitioners remain scarce. By systematically identifying these gaps, the present study situates itself at the intersection of technological innovation and organizational practice, providing an empirically grounded perspective on the evolving role of data across the development and deployment of AI-driven products and services.

Table 1. Summary of Key Domains in Related Work.

Domain (Section)	Key Findings in Prior Work	Unique Contributions	Remaining Gaps
Evolving Capabilities of AI in a Data-Driven World (3.1)	Shift from symbolic to data-driven and generative models; LLMs and GenAI expand AI's scope and complexity [3,4,22–32]	Mapping paradigm shifts; highlighting technical advances and new application domains	Limited empirical evidence on how these advances reshape data work and challenges in practice
AI Adoption in Organizational Contexts (3.2)	Widespread but uneven AI adoption; barriers include data infrastructure, skills, and cultural resistance [3,4,32–41]	Large-scale surveys; maturity models; identification of organizational barriers	Lack of in-depth, cross-sectoral analysis of how data work is managed and aligned with business goals
Critical Risks at the Intersection of AI and Data (3.3)	Risks include bias, lack of transparency, privacy, and governance; regulatory responses emerging [2,6,7,9,13,18–20,30,32,42–45]	Identification of ethical, technical, and societal risks; mapping regulatory frameworks	Few studies examine real-world strategies for mitigating data-related risks in organizational setting
Theoretical Lenses on the Role of Data in AI Systems (3.4)	Shift from model-centric AI to DCAI; data as socio-technical infrastructure [2,5,7–9,14,18,23,46]	Theoretical reframing of data's role; emphasis on annotation, context, and social factors	Scarcity of empirical research on how data-centric approaches are implemented and experienced by practitioners

4. Materials and Methods

4.1. Research Approach and Design

This study adopts a constructivist-interpretivist qualitative research design to explore how professionals in artificial intelligence and data science conceptualize the significance and influence of data throughout the lifecycle of AI-driven technological development. A qualitative approach was deemed particularly suitable given the study's

emphasis on situated knowledge, experiential insight, and meaning-making processes—dimensions that are often tacit, emergent, and deeply embedded within organizational and technological contexts [47–49].

This methodological orientation was selected for its ability to elicit authentic, context-sensitive perspectives and to capture the complexity of professional reasoning that resists quantification. It also enabled an inductive analytical process, allowing themes to emerge organically from participants' narratives rather than being constrained by pre-existing theoretical frameworks [47,50]. To balance structure with flexibility during data collection, the study employed semi-structured in-depth interviews. This method provided a consistent set of guiding questions while allowing for responsiveness to each participant's unique experiences, language, and organizational context. The openness of this format encouraged participants to articulate nuanced reflections and share rich organizational narratives, thereby enhancing the depth and contextual richness of the data [51–53].

All research procedures adhered to rigorous ethical standards in accordance with institutional and scientific community guidelines. The study received approval from the Institutional Review Board (Approval No. 120525582), ensuring compliance with best practices for research involving human participants. Prior to data collection, participants provided informed consent and were assured of full anonymity and confidentiality. They were also informed of their right to withdraw from the study at any time without consequence. No proprietary or commercially sensitive information was collected, and all data were securely stored on password-protected devices accessible only to the research team. The dataset underwent a thorough anonymization process to remove or obscure any identifying details related to participants or their affiliated organizations.

4.2. Sample Characteristics and Composition

The study was conducted between September and October 2024. A total of 74 senior professionals were interviewed, all of whom held leadership roles requiring substantial expertise at the intersection of AI and data. Participants were selected using purposive expert sampling [48,54], based on the criterion that they must hold strategic technological leadership positions with direct responsibility for AI and data-related decision-making. This sampling strategy was chosen to ensure that the insights gathered would reflect the perspectives of individuals with deep, practice-based knowledge and strategic influence in the field.

Participants' roles spanned both operational and strategic domains and included executive-level positions such as Chief Technology Officer (CTO), Chief AI Officer (CAIO), Chief Data Officer (CDO), and Chief Information Officer (CIO); individuals responsible for shaping their organizations' AI and data strategies. Additional participants held leadership roles such as Chief Information Security Officer (CISO), Head of Data & AI, and Data Protection Officer (DPO), overseeing technical implementation and regulatory compliance. Other interviewees included team leads and division heads in data science, AI development, cybersecurity, and business innovation—all actively engaged in the design, deployment, and governance of AI systems and data infrastructures.

The sample reflected a broad and diverse geographic distribution, with participants originating from multiple continents. Within Europe, countries represented included the United Kingdom, Ireland, Italy, Germany, France, Spain, Austria, Switzerland, Norway, Latvia, and North Macedonia. North America was represented by participants from the United States, while Africa was represented by South Africa. From Asia, participants were drawn from Israel, and Oceania was represented by Australia. This wide international scope underscores the global relevance of the research topic and contributes to the richness and heterogeneity of perspectives captured in the study. Notably, several participants held multinational roles or worked within globally distributed teams, further

enhancing the diversity of insights and reflecting the transnational nature of contemporary AI and data-driven work environments.

While the vast majority of participants (n = 69; 93.2%) were embedded within organizational structures across diverse sectors; including technology, finance and banking, education, law, healthcare and pharmaceuticals, manufacturing, and energy—a smaller subset (n = 5; 6.8%) operated independently as AI and data consultants. The youngest organization represented was a one-year-old IT consultancy in North Macedonia, while over half the participants (n = 38; 51.4%) worked for companies more than a decade old, including 16 (21.6%) employed by firms with over 20 years of operational history. In terms of organizational scale, six participants (8.1%) worked at small startups with fewer than 20 employees, while ten (13.5%) came from large enterprises with workforces exceeding 10,000 employees.

The selected sample size of 74 participants is considered substantial for an expert-based qualitative study and exceeds the typical range found in similar research involving elite or professional informants. While many qualitative studies reach thematic saturation with 12 to 30 participants—particularly when the population is relatively homogeneous [55–57]—larger samples are increasingly warranted in studies addressing complex, interdisciplinary, and globally distributed phenomena. In management and organizational research, for example, sample sizes of 50 or more have been effectively employed to capture variation across sectors, roles, and institutional contexts [58,59]. In the context of this study, the inclusion of a larger and more diverse pool of senior professionals was essential to ensure the collection of rich, multilayered insights into the evolving and multifaceted landscape of AI and data science—particularly in relation to data governance, regulatory adaptation, and model deployment. This approach aligns with recent methodological recommendations in qualitative research that emphasize maximizing variation and enhancing trustworthiness when investigating global domains [60].

4.3. Data Collection and Analysis

To ensure that participants were meaningfully engaged with the research topic and capable of offering firsthand, experience-based insights, recruitment was conducted through professional knowledge-sharing communities on social media platforms, most notably public LinkedIn profiles of individuals active in the fields of artificial intelligence and data science. These profiles provided transparent, verifiable information about participants' professional roles and expertise. Given the global dispersion of participants and the demanding schedules of senior professionals, the personal interviews were conducted via secure video conferencing platforms, primarily Zoom and Microsoft Teams.

Data collection occurred over four iterative stages, each lasting approximately one week. Following each stage, the interview protocol was refined in light of emerging insights, enabling a deeper and more focused exploration of the research themes over time [47,49]. Sample guiding questions included:

- What are the current challenges and risks involved in managing the data lifecycle—from acquisition to deployment and monitoring of AI models?
- Which of these challenges do you consider most urgent?
- How does data quality influence the accuracy and performance of AI models?
- What are the organizational or project-level consequences of unresolved data quality issues?
- How is your organization addressing evolving regulatory requirements around data privacy and AI compliance?

Participants were encouraged to speak openly and reflectively about their experiences, fostering the emergence of authentic and context-rich narratives. All interviews

were transcribed and securely stored in digital formats. Throughout the data collection and analysis process, researchers maintained reflexive research logs to document emerging themes, interpretive insights, and methodological reflections. The semi-structured interview protocol was carefully designed to balance consistency with flexibility, allowing participants to guide the conversation toward topics most pertinent to their professional contexts. This approach helped minimize the influence of researcher assumptions and facilitated the emergence of participant-driven perspectives. Concurrently, the reflexive logs served as a methodological tool for enhancing interpretive transparency, enabling sustained attention to the interpersonal and contextual dynamics that shaped both data collection and analysis [47,51,52].

The data were analyzed using thematic analysis, supported by the qualitative data analysis software MAXQDA (version 2022.8). The analytic process followed a multi-stage coding strategy: beginning with open coding to identify core ideas, followed by axial coding to explore relationships among themes, and culminating in cross-sectional analysis to compare patterns across participant groups. More specifically, the analysis adhered to six recursive phases: familiarization with the data, generation of initial codes, searching for themes, reviewing themes, defining and naming themes, and final reporting [61,62].

To enhance analytic rigor and minimize interpretive bias, inter-coder reliability procedures were implemented, involving independent coding by multiple researchers and consensus-building discussions [63]. In addition, a comprehensive audit trail, including reflexive field notes, analytic memos, and coding decisions, was maintained throughout the research process to enhance transparency and allow for independent verification of the analytic procedures. Follow-up interviews were conducted when participant responses were ambiguous or required additional context, further supporting the accuracy and credibility of the interpretations. Furthermore, to strengthen the credibility of the findings, authentic and unmediated participant accounts are interwoven throughout the findings chapter, serving as direct evidence to substantiate the analytic claims and provide deeper insight into participants' perspectives.

Figure 1 provides a visual summary of the research process, outlining the key methodological stages from design to analysis.

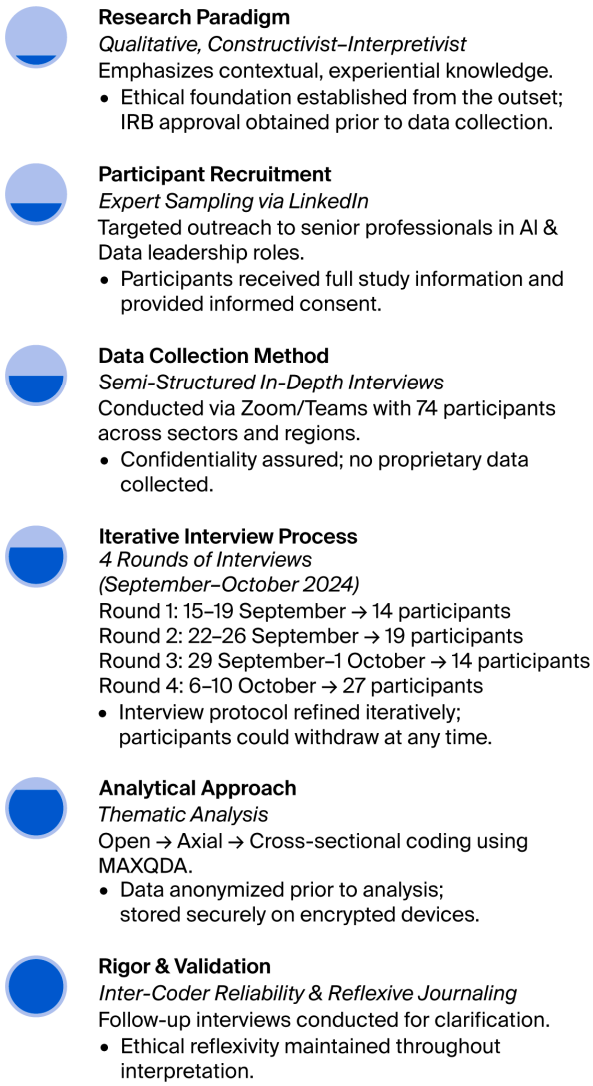

Figure 1. Overview of Research Design and Methodology.

5. Results

This study identified five interrelated thematic domains reflecting the perceptions of senior professionals regarding the challenges and organizational dynamics associated with data-driven AI systems. These themes highlight both technical and organizational dimensions and emphasize the centrality of data as a strategic resource in AI development and deployment. For clarity and synthesis, a concise summary is provided in bullet points at the end of each subsection (Sections 5.1–5.5).

A recurring concern among participants was the cautious and often hesitant approach organizations take toward adopting AI technologies. As one expert in NLP, working as an independent data analytics consultant, explained: *"Due to the multitude of associated challenges, organizations aren't quick to adopt AI technologies"* (P21). This cautious sentiment was echoed by a CDO at a major U.S. bank, who shared:

> *"Right now, we're not rushing to adopt this innovative technology. In fact, we're even quite hesitant to implement older AI-based systems that are already in use by other banks for underwriting and loans. The only AI-based application we've agreed to implement is a personal virtual assistant—a kind of chatbot—that doesn't require broad, unrestricted access to our customer data"* (P9).

A similar perspective was voiced by the VP of Emerging Tech at a large French insurance firm: *"We still don't have an AI solution that gives us full confidence in setting premiums. Many processes, even today, are done manually"* (P56). In contrast, concerns were raised about the opposite tendency—organizations that adopt AI too hastily. One participant warned:

> *"Companies that rush to adopt AI applications don't understand the implications of the risks. We're going to start seeing more and more lawsuits against companies for non-compliance with regulations or for discrimination. Only then will executives start to wake up"* (P50).

The complexity of developing and deploying AI-based systems was frequently described using vivid language. Participants referred to these efforts as involving *"pain points"* (P70), a *"burning issue"* (P61), and *"super-significant challenges"* (P35). These difficulties contribute to *"reduced trust in AI technologies among users"* (P13), a concern explored further in the following subsections.

5.1. Data Preparation Challenges

Participants consistently emphasized the technical and operational burdens associated with collecting, cleaning, and preparing data for AI development. One expert highlighted the challenge of integrating heterogeneous data sources: *"Integrating data from multiple channels—with high variability in formats, schemas, and standards—can lead to inconsistencies, making it difficult to ensure 'clean' and usable data for analysis and modeling"* (P65). Another added: *"Handling data that comes from diverse sources and exists in different formats can be challenging, especially when combining structured and unstructured data"* (P66).

The infrastructural demands of managing such data were also underscored. A CAIO at a large U.S. tech firm explained:

> *"As data increases in volume and variety, maintaining an efficient and cost-effective infrastructure that can handle both large-scale storage and processing becomes a major challenge—particularly when real-time access is required"* (P31).

Beyond infrastructure, financial implications were also raised. According to the Head of Data at a major pharmaceutical company:

> *"Many times, after completing the data cleaning process, we discover that the data is not relevant at all, and we have to start over—sometimes even purchase entirely different datasets. This costs our organization a great deal of money! Millions of dollars are sometimes wasted due to inaccurate data"* (P43).

This point was reinforced by the CTO of a mid-sized tech company: *"Often, organizations lack the data needed to train their models and find themselves in a bind—so they settle for off-the-shelf models that cover about 90% of what they were aiming for in terms of key business performance metrics"* (P24).

Several participants stressed the difficulty of discerning relevance within large datasets. As one consultant put it: *"The biggest challenge is understanding what's actually relevant [for model training] and, conversely, which data can be discarded"* (P8). Labeling data for supervised learning was also flagged as a resource-intensive process. One participant stated: *"Accurately labeling large-scale raw data, while also anonymizing it without losing its meaning and utility, presents major challenges"* (P70). Others described labeling as *"tedious and labor-intensive work that takes a lot of time"* (P47). A CDO in an information services company in the US noted: *"Often, people are hired temporarily just for this manual effort, and then let go afterward"* (P60).

As a mitigation strategy, many highlighted the need for a skilled, data-literate workforce: *"Organizations should focus on recruiting and developing a workforce with exceptionally strong data literacy"* (P67). Another expert described internal quality assurance practices:

"We frequently validate our data collection, cleaning, and processing workflows using internal tools" (P49).

In summary, participants highlighted the complexity and demands of preparing data in developing AI-enabled solutions:

- Integration of heterogeneous and variable data sources adds complexity and inconsistency.
- Large-scale data processing and infrastructure requirements create operational burdens.
- Data cleaning, relevance assessment, and labeling are time- and labor-intensive.
- Skilled, data-literate teams and internal validation practices are crucial for managing these challenges.

5.2. Data Quality Risks and Mitigation Strategies

Informants consistently emphasized that high-quality data is indispensable across the AI lifecycle, representing a shared concern among both C-level executives and compliance officers. One participant characterized it as a *"pressing and significant concern"* (P39), while a Director of Data Science at a major technology firm noted: *"It's a core priority that receives substantial resources and dedicated personnel"* (P44). Despite this recognition, ensuring and sustaining data quality was described as a costly endeavor. As two experts put it: *"Maintaining high-quality data for AI systems involves high operational costs"* (P65), and *"It's extremely expensive"* (P72).

Poor data quality was described in concrete terms—*"missing, incomplete, or partial data"* (P29), *"inaccurate, irrelevant, duplicate, or inconsistent records"* (P26), and *"incorrectly labeled or annotated datasets"* (P45). A founder of a French business consultancy summarized: *"It's a garbage in, garbage out situation"* (P42). The consequences of poor-quality data were seen as far-reaching. A data scientist at a global food and beverage company noted: *"It introduces a lot of noise into the models and severely limits their accuracy"* (P45). Others pointed to broader organizational impacts, such as *"delayed time-to-market"* (P29), *"increased operational costs"* (P41), *"eroded trust in AI systems"* (P40), and *"regulatory non-compliance"* (P38). A CTO emphasized the pivotal role of data as the foundation upon which effective AI model performance depends: *"You can't separate data from proper model development and optimization"* (P24).

To mitigate these risks, participants described a range of validation and monitoring practices. One expert outlined a multi-stage approach:

> *"We start by ensuring that the data is well-defined, diverse, and representative. Then we validate the expected inputs for each model and scan them thoroughly before training begins. … We've built Power BI dashboards that alert us in real time to poor data quality. … We implement automated data cleaning—or at the very least, automated alerts that flag errors"* (P73).

This emphasis on proactive quality control was echoed by others:

> *"The ability to track data from acquisition through every stage of its lifecycle—preprocessing, modeling, and deployment—is critical. … I use automated tools to detect missing values, inconsistencies, and anomalies, followed by enrichment processes to ensure completeness and accuracy"* (P65).

Finally, the importance of robust governance frameworks was highlighted. A Head of Data at a leading German tech firm stressed, *"Every organization needs to implement strict data governance mechanisms. These ensure proper preparation, responsible use, and effective management of data. Governance also establishes ownership and clearly defines which roles are accountable for data quality"* (P41). Nevertheless, it was evident that participants expressed considerable dissatisfaction with the current solutions available. Many underscored that existing

tools and frameworks fall short in addressing the complex and persistent challenges associated with data quality. These approaches were often viewed as insufficiently effective or comprehensive, lacking the adaptability and depth required to meet the rigorous demands of data-intensive environments.

In summary, data quality emerged as a central concern affecting model performance and organizational outcomes:

- Inaccurate, incomplete, or inconsistent data jeopardizes AI reliability.
- Poor-quality data increases operational costs and delays in deployment.
- Governance frameworks and validation tools help monitor and maintain quality.
- Existing solutions are still insufficient to fully address complex, persistent challenges.

5.3. Privacy, Security, and Data Leakage Concerns

Concerns around data privacy, leakage, and cybersecurity threats were prominent across interviews. A CISO in the cybersecurity sector pointed to a growing organizational worry: *"The main concern companies have today is how to ensure that employees don't share personal or sensitive information with various chatbots"* (P2). A DPO at a UK public firm warned about vendor misuse: *"Some vendors claim to be AI providers, but in reality, they're collecting everyone's data to train their models and sell them to big tech companies"* (P10). Similarly, a data and product development expert at a legal services firm in Israel emphasized the need for internal control:

> *"A company purchasing LLMs must be absolutely certain that its data is securely handled by the supplier. But managers often struggle to control what employees input into AI tools, so it's crucial for every organization to have a clear policy on this matter"* (P15).

This underscores the need to *"ensure that every department complies with national privacy regulations, particularly GDPR* [General Data Protection Regulation]*"* (P2).

Cybersecurity threats were described as both emerging and serious. *"One of the most significant risks in deploying these models is related to security vulnerabilities. AI models can introduce new threats, such as adversarial attacks, making strong security measures absolutely essential"* (P66). Another AI and Data expert elaborated: *"Deployed AI models may be exposed to adversarial attacks, where malicious users attempt to manipulate predictions or access sensitive information"* (P65). A CIO at a South African IT firm added a broader concern: *"There's so much data that it's hard to ensure none of it leaks or gets lost"* (P3).

A particularly stark warning came from a senior software engineer at a U.S. cybersecurity firm, who expressed concern about open-source LLMs:

> *"It's unclear what these models were trained on or what cybersecurity risks they might contain—like backdoors, exploits, or critical vulnerabilities. … Most only conduct basic security checks. … We haven't seen a major AI-driven breach yet, but I'm certain it's only a matter of time"* (P20).

In summary, concerns about misuse, employee practices, and system vulnerabilities converged into a shared sense of organizational risk, reflected in the following issues:

- Employees may inadvertently expose sensitive information through AI tools.
- Vendors risk misusing organizational data for training and resale.
- AI models create new cybersecurity threats, including adversarial attacks.
- Open-source models carry hidden vulnerabilities that are difficult to assess.

5.4. Ethical and Technical Challenges of Bias and Opacity

Unintended algorithmic bias was consistently identified as a critical challenge in AI system development. One participant emphasized the ethical implications of such bias: *"There is always a risk of unintended outcomes* [from algorithmic models]*, such as biased*

decision-making, which requires ongoing ethical evaluation" (P66). Another expert added a longer-term perspective, highlighting the issue of model drift:

> *"Even well-trained models can produce biased results, which may have unintended social or ethical consequences that are often difficult to detect in real time. There's also the issue of model drift—over time, AI models may become less effective due to changes in underlying data or external factors, leading to inaccurate predictions"* (P65).

Real-world examples illustrate how contextual factors can distort model outputs. A senior data science lead at a U.S.-based company specializing in smart water sensors shared:

> *"We expect real-time alerts when a leak occurs. Our main pain point is when the algorithmic model simply gets it wrong—sometimes it flags a problem where there isn't one, and other times it misses actual issues. … For instance, during the Super Bowl in the U.S., people's water usage patterns change dramatically. The AI-based sensors misinterpret this as a leak, introducing bias into the model"* (P55).

Bias can also emerge from the early stages of data preparation. A data scientist specializing in image processing at a small environmental services company in Norway explained it as follows: *"Proper data preparation is critical for producing reliable model outcomes. Even slight pixel-level inconsistencies in training data can cause major disruptions, leading to bias and rendering the results irrelevant"* (P28). Importantly, several participants stressed that bias is not only a technical issue but also a human one. As one interviewee put it: *"We recognize that humans are biased, and therefore so are the developers of these models"* (P67). Another recurring concern was the lack of transparency and explainability in AI systems. One expert described this challenge succinctly: *"There's a deep lack of understanding about what AI actually does and how it works. In practice, it's a black box"* (P67).

To address these ethical and technical issues, some organizations are adopting enhanced validation protocols. A CAIO in the healthcare sector described their approach:

> *"Our organization conducts rigorous quality assurance processes based on multiple logic layers throughout the training phase, because we don't fully trust the model outputs. We flag errors as they arise and perform unique model validation for each dataset—we don't just feed data into the model blindly. … Any company that deals with large volumes of data and wants to integrate AI applications must have someone on staff with strong formal training in data science"* (P17).

This view was reinforced by the Head of Cybersecurity at a U.S. civil engineering institution, who warned:

> *"In many companies, the people working with AI systems are data scientists who lack a strong foundation in advanced statistical methods. This exposes them to significant risks without even realizing it. The key is to hire expert statisticians who can handle the data before it enters the models. Only then can we better understand the algorithms, correct for bias, and remain alert to emerging issues"* (P6).

In summary, bias and opacity were seen as critical barriers to trust and reliability in AI:

- Algorithmic bias can arise from data inconsistencies and contextual misinterpretations.
- Human bias and limited statistical expertise compromise model reliability.
- Model drift and opacity reduce trust and long-term validity.
- Organizations respond with layered validation protocols and expert oversight.

5.5. Organizational Responses to AI Regulation

The topic of AI regulation generated considerable discussion, particularly among participants from highly regulated sectors such as finance, healthcare, legal services, and

the public sector. A CEO of a small startup developing a proprietary LLM shared: *"Many of our customers frequently ask about the compliance of our AI-based product with regulatory and legal standards"* (P18). Despite growing awareness, many acknowledged that regulatory adaptation is still in its early stages. As the Head of Data and AI at a financial firm remarked: *"Most companies are not there yet"* (P14). In the absence of comprehensive mandates, *"organizations often rely on internal policies"* (P59) or *"assign dedicated teams to track upcoming regulatory developments"* (P10).

Most current efforts appear to focus narrowly on privacy compliance. One expert observed: *"Organizations are mostly just making sure their model development complies with privacy regulations, and not much more"* (P5). Several participants mentioned early engagement with the EU AI Act. One described this initial response as follows: *"Some companies are just beginning to explore the broader implications of the EU AI Act. For now, most are focusing on transparency and documentation of training processes"* (P2). A data protection consultant in Italy confirmed the limited practical readiness: *"Although we're seeing a shift toward AI Act compliance, we currently lack the practical tools to address it. For now, it's mostly about documentation"* (P8).

A deeper concern, however, centered on the lack of enforcement mechanisms: *"There's no clear way to enforce these regulatory frameworks, so in practice, anyone can do whatever they want. That's the main problem with regulation—it lacks enforceable mechanisms"* (P74). Another interviewee noted how responsibility is often shifted to external AI vendors: *"At this point, the burden of responsibility is largely pushed back onto the major suppliers, like OpenAI. I rely on their certifications, and AI regulation always ends up at the bottom of our priority list"* (P16).

Nonetheless, some organizations are beginning to adopt more structured and proactive approaches. One participant described a comprehensive strategy:

> *"We maintain continuous monitoring. A dedicated team tracks regulatory changes and updates our processes accordingly. Our data governance framework was designed to be flexible from the outset, allowing us to quickly implement necessary changes in response to new regulations. Beyond that, we ensure our teams receive ongoing training on the latest regulatory requirements and emphasize the importance of compliance throughout the entire project lifecycle. I believe that built-in tools for automated compliance will significantly enhance productivity and reduce the risk of non-compliance"* (P66).

Another participant concluded with a broader call to action: *"A thorough analysis of regulatory requirements is essential, along with active involvement from stakeholders and international experts in systems thinking"* (P71).

In summary, regulatory adaptation remains limited, but proactive measures are emerging:

- Most organizations are in early stages of regulatory readiness.
- Internal policies and monitoring teams are used to track evolving requirements.
- Current efforts primarily focus on privacy compliance (e.g., GDPR).
- Enforcement mechanisms are unclear, often shifting responsibility to vendors.

6. Data-Centric Framework

This study investigated how strategic professionals in AI and data science conceptualize the role of data in shaping AI-enabled solutions. Drawing on in-depth interviews with 74 senior experts, the findings offer a grounded, practice-oriented perspective that disrupts dominant model-centric paradigms in AI research. While much academic focus remains on algorithmic innovation and model development, this study repositions data as the principal site of complexity, uncertainty, and strategic decision-making in real-world AI development.

To articulate this reorientation, a conceptual model of the AI lifecycle was constructed based on practitioners' narratives and grounded in their lived experiences (see Figure 2). The model delineates a data-centric process that spans from initial collection and preparation to deployment, monitoring, and explainability; each phase involves distinct professional roles and tightly coupled interdependencies. Rather than presenting a linear pipeline, the model emphasizes the recursive and evolving nature of data work, portraying data not as a static input but as an active infrastructure that is continuously shaped by, and shaping, technical and organizational decisions. Whether through feature engineering, error correction, or interpretability practices, data emerges as both the foundation and connective tissue of AI systems. This centrality is further detailed in Table 2, which outlines the key components of the model and illustrates how data-related tasks, challenges, and decisions permeate every stage of the lifecycle.

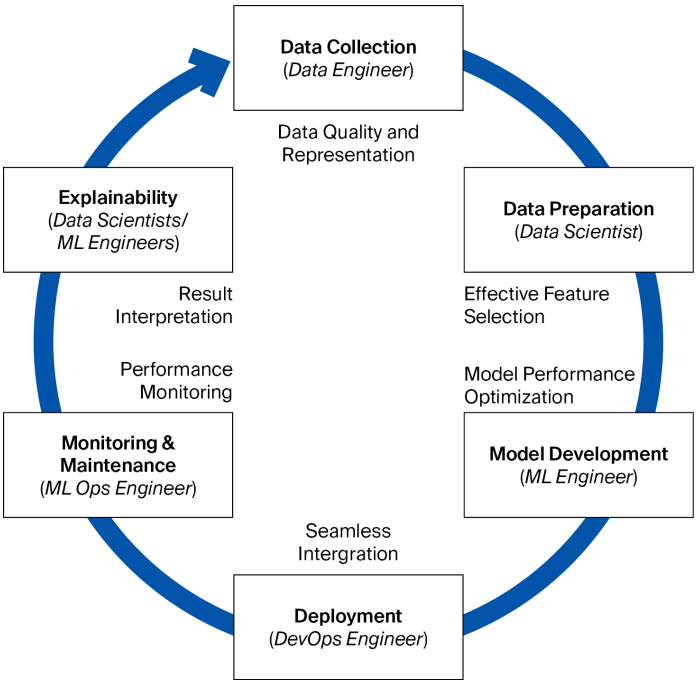

Figure 2. Conceptual Model of the AI Lifecycle: A Data-Centric Perspective.

Table 2. Data-Centric Components of the AI Lifecycle Model.

Stage	Role(s) Involved	Data-Centric Focus and Description
Data Collection	Data Engineer	Initiating the lifecycle, this stage involves sourcing, aggregating, and validating raw data. The quality, representativeness, and accessibility of data at this point fundamentally shape all downstream AI processes.
Data Preparation	Data Scientist	Data is cleaned, transformed, and structured to ensure usability. Feature selection—identifying the most relevant variables—is a critical data-driven task that directly impacts model performance.
Model Development	ML Engineer	While focused on algorithmic design, this stage remains data-dependent, as model training, tuning, and validation rely entirely on the quality and structure of the input data.
Deployment	DevOps Engineer	Although technical in nature, deployment requires careful handling of data pipelines to ensure that real-time or batch data flows into the model as intended.
Monitoring & Maintenance	ML Ops Engineer	Ongoing evaluation of model performance is driven by continuous data input. Monitoring for data drift, anomalies, or shifts in distribution is essential to maintain reliability.
Explainability & Interpretation	Data Scientists/ML Engineers	Interpreting model outputs requires understanding how data influenced decisions. Explainability tools often rely on data-centric techniques to trace and justify predictions.

6.1. Data Challenges in the Eyes of Strategic Experts (RO1)

The first research objective examined the primary challenges experts face when working with data during AI model development. Participants uniformly portrayed data work as laborious and iterative—entailing a complex web of manual cleaning, source reconciliation, formatting, labeling, and infrastructure management. Unlike abstract definitions of data quality, their accounts emphasized the embodied and organizational effort required to render data usable at scale. These accounts provide empirical weight to recent critiques of model-centrism (e.g., [8,9,11]), and reposition data readiness—not model architecture—as the true bottleneck of AI pipelines.

Moreover, the findings indicate that many organizations remain preoccupied with basic data hygiene, often delaying attention to higher-level ethical considerations such as bias mitigation or algorithmic fairness. This suggests that data quality cannot be meaningfully pursued without first achieving a threshold of operational maturity. Rather than viewing veracity, consistency, and completeness as low-level concerns, the study reveals them as strategic dependencies upon which more advanced AI ambitions rest [32]. In this way, the invisibility of data labor [5,12,17,22] emerges not only as a theoretical issue but as a concrete obstacle to responsible AI development.

6.2. The Strategic Impact of Data on AI Development (RO2)

The second research objective explored how data-related challenges influence AI system development and organizational outcomes. Across interviews, a clear consensus emerged: data quality, contextual relevance, and organizational fit consistently outweigh the influence of model sophistication. Participants linked poor data to cascading effects—ranging from lower model accuracy and prolonged development cycles to regulatory exposure and reputational damage. These consequences illustrate that data challenges are not confined to technical performance but affect broader strategic imperatives, including time-to-market, stakeholder alignment, and compliance trajectories.

Crucially, professionals did not view data as a fixed asset to be optimized once models are in place. Instead, they described data as an evolving entity that must be aligned with changing organizational goals and domain conditions. This fluidity aligns with the principles of DCAI, which emphasize continuous data refinement as a primary lever of model performance [7,9,23]. Beyond technical barriers, participants emphasized organizational frictions, such as cross-team miscommunication, inconsistent annotation protocols, and underinvestment in domain-specific knowledge. These dynamics reinforce arguments by D'Ignazio and Klein [64] that data infrastructures are inherently social and value-laden, requiring more than engineering rigor; they demand cultural fluency and interdisciplinary negotiation.

6.3. Strategies and Limitations in Addressing Data Risks (RO3)

The third research objective focused on how professionals attempt to manage data-related risks. Participants described a variety of mitigation strategies, such as automated QA pipelines, annotation guidelines, and validation dashboards. Yet these were often seen as stopgap solutions rather than systemic fixes. Particularly, reliance on outsourced data services or pre-trained models was met with caution, as these approaches often lack the domain specificity required for nuanced applications. The recurring view was that data cannot be abstracted from its context—local expertise and in-house stewardship are essential to ensuring relevance, traceability, and trust.

These findings contribute to a growing literature on the operationalization of AI, which underscores the importance of integrating models into complex, real-world environments [32,65]. As Sambasivan et al. [5] argue, most existing tools were not designed to meet the specific challenges posed by modern AI data pipelines. This study supports that

view and further highlights the strategic importance of robust data governance—not as a compliance checkbox, but as a dynamic and ongoing framework for managing responsibility, adaptability, and cross-functional alignment. A unifying insight from the interviews was the emergence of a shared professional ethos: a recognition that trustworthy AI depends not just on sophisticated models but on rigorous, socially informed data practices. This ethos represents an implicit call to reorient both academic and industrial priorities toward the data layer as the true locus of AI capability and risk.

7. Conclusions

7.1. Theoretical Contributions

This study makes a significant theoretical contribution by redirecting scholarly attention in AI research from algorithmic optimization to the foundational role of data. While much of the academic and industry discourse has emphasized model performance and computational power, our findings reinforce an emerging shift toward a data-centric paradigm. Data is revealed not as a neutral input, but as a dynamic, labor-intensive, and context-dependent infrastructure—one that critically shapes the effectiveness, fairness, and accountability of AI systems.

By foregrounding data as a site of human judgment, negotiation, and labor, the study enriches sociotechnical understandings of AI as a system deeply embedded in institutional, organizational, and cultural contexts. This reconceptualization aligns with and is operationalized through the AI Lifecycle model developed in this research, which maps the iterative and interdependent phases of AI development—from problem formulation and data acquisition to model deployment and post-deployment monitoring. Crucially, our findings demonstrate that data-related decisions recur across the entire lifecycle and must therefore be treated as a central component of system design and evaluation. Ethical imperatives such as transparency, bias mitigation, and accountability cannot be appended as afterthoughts; they must be embedded within data governance, stewardship, and epistemic framing from the outset. This perspective advances theoretical discourse by bridging normative principles with empirical data work, and by emphasizing the socio-organizational processes through which AI systems become trustworthy and functional.

7.2. Practical Implications

On a practical level, the study suggests that organizations may benefit from a strategic realignment in AI development—shifting from a narrow focus on model accuracy to a broader attention to data quality, governance, and context-sensitive workflows. Many of the most pressing technical, ethical, and legal challenges associated with AI arise from upstream data issues, rather than algorithmic flaws. Anchoring responsible AI practices within the AI Lifecycle model, particularly in its early and middle phases, can help anticipate and mitigate downstream risks such as bias, opacity, and performance failures.

The findings indicate that successful AI implementation depends on domain-specific, context-aware data work that integrates technical expertise, organizational knowledge, and regulatory awareness. Robust AI governance appears to require cross-functional collaboration among data scientists, domain specialists, legal advisors, compliance officers, and decision-makers. Policy instruments such as the EU AI Act are important, yet effective implementation also demands enforceable operational standards, sector-specific guidelines, and capacity-building initiatives that institutionalize responsible data practices. Investments in data pipelines, documentation protocols, metadata standards, and collaborative workflows are therefore likely to enhance sustainability and credibility, though practical constraints and organizational realities may limit the extent to which these measures are fully adopted.

7.3. Limitations and Future Research

As a qualitative, expert-driven study, the findings presented here are interpretive and contextually situated. They primarily reflect the experiences of professionals in strategic and technical leadership roles, which may limit the representation of operational perspectives, including those of data annotators, compliance teams, or end-users directly interacting with AI systems. The study also captures a specific moment within an evolving regulatory and technological landscape, potentially constraining the generalizability of its conclusions across different times and sectors. Furthermore, although the research draws on a substantial set of expert interviews, validation of secondary data sources remains somewhat limited, and despite efforts to address potential biases during data collection, some degree of interpretive bias inherent to qualitative inquiry cannot be entirely excluded.

Future research should extend this work by employing comparative case studies across different industries, organizational structures, and national settings. Longitudinal designs could offer deeper insight into how data governance practices evolve in response to shifting regulatory frameworks such as the EU AI Act. In particular, empirical investigation into the enactment and adaptation of AI Lifecycle models in real-world environments—using ethnographic or participatory approaches—could reveal the frictions, improvisations, and negotiations that shape responsible AI development in practice. Interdisciplinary collaborations that bridge AI research, organizational theory, and legal scholarship will be essential for unpacking the complex infrastructures that underpin trustworthy and effective AI systems.

Author Contributions: Conceptualization, M.N.; methodology, L.Z. and M.N.; software, M.N.; validation, L.Z. and M.N.; formal analysis, M.N.; investigation, L.Z.; resources, L.Z.; data curation, M.N.; writing—original draft preparation, M.N.; writing—review and editing, M.N.; visualization, M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request, subject to confidentiality constraints related to the anonymity of participants and their organizations.

Acknowledgments: The authors gratefully acknowledge Tomer Yatzkan from the "Humane AI" startup for his substantial and sustained contributions. His expertise and close collaboration were instrumental in shaping the research outcomes.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
DCAI	Data-centric AI
ML	Machine Learning
GenAI	Generative artificial intelligence
LLMs	Large language models
CAIO	Chief AI Officer

References

1. Gill, K.S. The end AI innocence: Genie is out of the bottle. *AI Soc.* **2025,** *40*, 257–261. https://doi.org/10.1007/s00146-025-02267-0.
2. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. https://doi.org/10.1145/3457607.

3. Nakash, M.; Bolisani, E. Knowledge management meets artificial intelligence: A systematic review and future research agenda. In *European Conference on Knowledge Management*; Academic Conferences International Limited: Reading, UK, 2024; pp. 544–552. https://doi.org/10.34190/eckm.25.1.2443.
4. Nakash, M.; Bolisani, E. The transformative impact of AI on knowledge management processes. *Bus. Process Manag. J.* **2025**, *31*, 124–147. https://doi.org/10.1108/BPMJ-11-2024-1137.
5. Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; Aroyo, L.M. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan, 8–13 May 2021; pp. 1–15. https://doi.org/10.1145/3411764.3445518.
6. Goyal, M.; Mahmoud, Q.H. A systematic review of synthetic data generation techniques using generative AI. *Electronics* **2024**, *13*, 3509. https://doi.org/10.3390/electronics13173509.
7. Patel, K. Ethical Reflections on Data-Centric AI: Balancing Benefits and Risks. Available at SSRN 4993089 2024. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4993089 (accessed on 4 February 2025).
8. Zha, D.; Bhat, Z.P.; Lai, K.H.; Yang, F.; Hu, X. Data-centric AI: Perspectives and challenges. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), Minneapolis, MN, USA, 27–29 April 2023; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 2023; pp. 945–948. https://doi.org/10.1137/1.9781611977653.ch106.
9. Kumar, S.; Datta, S.; Singh, V.; Singh, S.K.; Sharma, R. Opportunities and challenges in data-centric AI. *IEEE Access* **2024**, *12*, 33173–33189. https://doi.org/10.1109/ACCESS.2024.3369417.
10. Stonebraker, M.; Rezig, E.K. Machine learning and big data: What is important? *IEEE Data Eng. Bull.* **2019**, *42*, 3–7..
11. Mazumder, M.; Banbury, C.; Yao, X.; Karlaš, B.; Gaviria Rojas, W.; Diamos, S.; Janapa Reddi, V. Dataperf: Benchmarks for data-centric ai development. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 5320–5347. https://doi.org/10.48550/arXiv.2207.10062.
12. Nagle, T.; Redman, T.C.; Sammon, D. Only 3% of companies' data meets basic quality standards. *Harv. Bus. Rev.* **2017**, *95*, 2–5.
13. Narayanan, A.; Kapoor, S. Why an overreliance on AI-driven modelling is bad for science. *Nature* **2025**, *640*, 312–314. https://doi.org/10.1038/d41586-025-01067-2.
14. Slota, S.C.; Fleischmann, K.R.; Greenberg, S.; Verma, N.; Cummings, B.; Li, L.; Shenefiel, C. Good systems, bad data?: Interpretations of AI hype and failures. *Proc. Assoc. Inf. Sci. Technol.* **2020**, *57*, e275. https://doi.org/10.1002/pra2.275.
15. Wagstaff, K. Machine learning that matters. *arXiv* **2012**, arXiv:1206.4656. https://doi.org/10.48550/arXiv.1206.4656.
16. Batty, M. Planning data. *Environ. Plan. B Urban Anal. City Sci.* **2022**, *49*, 1588–1592. https://doi.org/10.1177/23998083221105496.
17. Jarrahi, M.H.; Memariani, A.; Guha, S. The principles of data-centric AI (DCAI). *arXiv* **2022**, arXiv:2211.14611. https://doi.org/10.48550/arXiv.2211.14611.
18. Whang, S.E.; Roh, Y.; Song, H.; Lee, J.G. Data collection and quality challenges in deep learning: A data-centric ai perspective. *VLDB J.* **2023**, *32*, 791–813. https://doi.org/10.1007/s00778-022-00775-9.
19. Camilleri, M.A. Artificial intelligence governance: Ethical considerations and implications for social responsibility. *Expert Syst.* **2024**, *41*, e13406. https://doi.org/10.1111/exsy.13406.
20. Radanliev, P. AI ethics: Integrating transparency, fairness, and privacy in AI development. *Appl. Artif. Intell.* **2025**, *39*, 2463722. https://doi.org/10.1080/08839514.2025.2463722.
21. Sartori, L.; Theodorou, A. A sociotechnical perspective for the future of AI: Narratives, inequalities, and human control. *Ethics Inf. Technol.* **2022**, *24*, 4. https://doi.org/10.1007/s10676-022-09624-3.
22. Paullada, A.; Raji, I.D.; Bender, E.M.; Denton, E.; Hanna, A. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* **2021**, *2*, 100336. https://doi.org/10.1016/j.patter.2021.100336.
23. Zha, D.; Bhat, Z.P.; Lai, K.H.; Yang, F.; Jiang, Z.; Zhong, S.; Hu, X. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–42. https://doi.org/10.1145/3711118.
24. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4th ed; Pearson Education: London, UK, 2020; Available online: http://lib.ysu.am/disciplines_bk/efdd4d1d4c2087fe1cbe03d9ced67f34.pdf (accessed on 15 December 2024).
25. Domingos, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine will Remake Our World*; Basic Books: New York, NY, USA, 2015; Available online: https://www.redalyc.org/pdf/6380/638067264018.pdf (accessed on 15 December 2024).
26. Domingos, P. Machine learning for data management: Problems and solutions. In Proceedings of the 2018 International Conference on Management of Data, Houston, TX, USA, 10–15 June 2018; p. 629. Available online: https://doi.org/10.1145/3183713.3199515 (accessed on 27 January 2025).
27. Holzinger, A. Introduction to machine learning & knowledge extraction (make). *Mach. Learn. Knowl. Extr.* **2019**, *1*, 1–20. https://doi.org/10.3390/make1010001.

28. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Liang, P. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258.
29. Fui-Hoon Nah, F.; Zheng, R.; Cai, J.; Siau, K.; Chen, L. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *J. Inf. Technol. Case Appl. Res.* **2023**, *25*, 277–304. https://doi.org/10.1080/15228053.2023.2233814.
30. Hagos, D.H.; Battle, R.; Rawat, D.B. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Trans. Artif. Intell.* **2024**, *5*, 5873–5893. https://doi.org/10.1109/TAI.2024.3444742.
31. Schwartz, D.; Te'eni, D. AI for knowledge creation, curation, and consumption in context. *J. Assoc. Inf. Syst.* **2024**, *25*, 37–47. https://doi.org/10.17705/1jais.00862.
32. Sinha, S.; Lee, Y.M. Challenges with developing and deploying AI models and applications in industrial systems. *Discov. Artif. Intell.* **2024**, *4*, 55. https://doi.org/10.1007/s44163-024-00151-2.
33. Kirchner, K.; Bolisani, E.; Kassaneh, T.C.; Scarso, E.; Taraghi, N. Generative AI Meets Knowledge Management: Insights From Software Development Practices. *Knowl. Process Manag.* **2025**, 1–13. https://doi.org/10.1002/kpm.70004.
34. Peretz, O.; Nakash, M. From Junior to Senior: Skill Requirements for AI Professionals Across Career Stages. In Proceedings of the International Conference on Research in Business, Management and Finance, Rome, Italy, 5–7 December 2025; Volume 2, pp. 9–10. https://doi.org/10.33422/icrbmf.v2i1.1178.
35. McKinsey. The State of AI in 2023: Generative AI's Breakout Year. 2023. Available online: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year (accessed on 22 May 2025).
36. Deloitte. 2024 Year-End Generative AI Report. 2024. Available online: https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-generative-ai-in-enterprise.html (accessed on 22 May 2025).
37. Kurup, S.; Gupta, V. Factors influencing the AI adoption in organizations. *Metamorphosis* **2022**, *21*, 129–139. https://doi.org/10.1177/09726225221124035.
38. McElheran, K.; Li, J.F.; Brynjolfsson, E.; Kroff, Z.; Dinlersoz, E.; Foster, L.; Zolas, N. AI adoption in America: Who, what, and where. *J. Econ. Manag. Strategy* **2024**, *33*, 375–415. https://doi.org/10.1111/jems.12576.
39. Sadiq, R.B.; Safie, N.; Abd Rahman, A.H.; Goudarzi, S. Artificial intelligence maturity model: A systematic literature review. *PeerJ Comput. Sci.* **2021**, *7*, e661. https://doi.org/10.7717/peerj-cs.661.
40. Romeo, E.; Lacko, J. Adoption and integration of AI in organizations: A systematic review of challenges and drivers towards future directions of research. *Kybernetes* **2025**, 1–22. https://doi.org/10.1108/K-07-2024-2002.
41. Wu, T.J.; Liang, Y.; Wang, Y. The buffering role of workplace mindfulness: How job insecurity of human-artificial intelligence collaboration impacts employees' work–life-related outcomes. *J. Bus. Psychol.* **2024**, *39*, 1395–1411. https://doi.org/10.1007/s10869-024-09963-6.
42. Araujo, T.; Helberger, N.; Kruikemeier, S.; De Vreese, C.H. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI Soc.* **2020**, *35*, 611–623. https://doi.org/10.1007/s00146-019-00931-w.
43. Pflanzer, M.; Dubljević, V.; Bauer, W.A.; Orcutt, D.; List, G.; Singh, M.P. Embedding AI in society: Ethics, policy, governance, and impacts. *AI Soc.* **2023**, *38*, 1267–1271. https://doi.org/10.1007/s00146-023-01704-2.
44. Cabrera, B.M.; Luiz, L.E.; Teixeira, J.P. The Artificial Intelligence Act: Insights regarding its application and implications. *Procedia Comput. Sci.* **2025**, *256*, 230–237. https://doi.org/10.1016/j.procs.2025.02.116.
45. Finocchiaro, G. The regulation of artificial intelligence. *Ai Soc.* **2024**, *39*, 1961–1968. https://doi.org/10.1007/s00146-023-01650-z.
46. Redman, T.C. If your data is bad, your machine learning tools are useless. *Harv. Bus. Rev.* **2018,** *2*. Available online: https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless (accessed on 18 November 2024).
47. Dodgson, J.E. About research: Qualitative methodologies. *J. Hum. Lact.* **2017**, *33*, 355–358. https://doi.org/10.1177/0890334417698693.
48. Douglas, H. Sampling techniques for qualitative research. In *Principles of Social Research Methodology*; Islam, M.R., Khan, N.A., Baikady, R. Eds.; Springer: Singapore, 2022; pp. 415–426. https://doi.org/10.1007/978-981-19-5441-2_29.
49. Mohajan, H.K. Qualitative research methodology in social sciences and related subjects. *J. Econ. Dev. Environ. People* **2018**, *7*, 23–48. https://doi.org/10.26458/jedep.v7i1.571.
50. Gummesson, E. *Qualitative Methods in Management Research*; Sage: London, UK, 2000. Available online: https://www.researchgate.net/publication/215915855_Qualitative_Research_Methods_in_Management_Research (accessed on 19 November 2024).
51. Creswell, J.W.; Poth, C.N. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*; Sage Publications: London, UK, 2016; ISBN 978-1-5063-3020-4.
52. Pathak, V.; Jena, B.; Kalra, S. Qualitative research. *Perspect. Clin. Res.* **2013**, *4*, 192. https://doi.org/10.4103/2229-3485.115389.

53. Scanlan, C.L. *Preparing for the Unanticipated: Challenges in Conducting Semi-Structured, In-Depth Interviews*; Sage Publications Limited: London, UK, 2020; pp. 67–80. https://doi.org/10.4135/9781529719208.
54. Ahmad, M.; Wilkins, S. Purposive sampling in qualitative research: A framework for the entire journey. *Qual. Quant.* **2024**, *59*, 1–19. https://doi.org/10.1007/s11135-024-02022-5.
55. Guest, G.; Bunce, A.; Johnson, L. How many interviews are enough? An experiment with data saturation and variability. *Field Methods* **2006**, *18*, 59–82. https://doi.org/10.1177/1525822X05279903.
56. Hagaman, A.K.; Wutich, A. How many interviews are enough to identify metathemes in multisited and cross-cultural research? Another perspective on Guest, Bunce, and Johnson's (2006) landmark study. *Field Methods* **2017**, *29*, 23–41. https://doi.org/10.1177/1525822X16640447.
57. Marshall, B.; Cardon, P.; Poddar, A.; Fontenot, R. Does sample size matter in qualitative research?: A review of qualitative interviews in IS research. *J. Comput. Inf. Syst.* **2013**, *54*, 11–22. https://doi.org/10.1080/08874417.2013.11645667.
58. Boddy, C.R. Sample size for qualitative research. *Qual. Mark. Res. Int. J.* **2016**, *19*, 426–432. https://doi.org/10.1108/QMR-06-2016-0053.
59. Bouncken, R.B.; Czakon, W.; Schmitt, F. Purposeful sampling and saturation in qualitative research methodologies: Recommendations and review. *Rev. Manag. Sci.* **2025**, 1–37. https://doi.org/10.1007/s11846-025-00881-2.
60. Dworkin, S.L. Sample size policy for qualitative studies using in-depth interviews. *Arch. Sex. Behav.* **2012**, *41*, 1319–1320. https://doi.org/10.1007/s10508-012-0016-6.
61. Braun, V.; Clarke, V. Thematic analysis. In *Encyclopedia of Quality of Life and Well-Being Research*; Springer International Publishing: Cham, Switzerland, 2024; pp. 7187–7193. https://doi.org/10.1007/978-3-031-17299-1_304024.
62. Guest, G.; MacQueen, K.M.; Namey, E.E. *Applied Thematic Analysis*; Sage publications: London, UK, 2011. Available online: https://antle.iat.sfu.ca/wp-content/uploads/Guest_2012_AppliedThematicAnlaysis_Ch1.pdf (accessed on 25 July 2024).
63. O'Connor, C.; Joffe, H. Intercoder reliability in qualitative research: Debates and practical guidelines. *Int. J. Qual. Methods* **2020,** *19*, 1–13. https://doi.org/10.1177/1609406919899220.
64. D'Ignazio, C.; Klein, L.F. *Data Feminism*; MIT Press: Cambridge, MA, USA, 2023; ISBN 9780262547185.
65. Brandao, P.R. The Impact of Artificial Intelligence on Modern Society. *AI* **2025**, *6*, 190. https://doi.org/10.3390/ai6080190.