Understanding data governance in AI: Exploring the ecosystem

Data-centric AI

ODI Research
ADVANCING TRUST IN DATA

June 2024

Contents

About

This report has been researched and produced by the Open Data Institute (ODI), and published in June 2024.

Its lead author was Thomas Carey-Wilson, with support from Professor Elena Simperl, and Arunav Das and Ioannis Reklos from King's College London (KCL).

It was made possible through the generous support of the Patrick J. McGovern Foundation and Omidyar Network, which contributed funding to enhance our research exploring Data-centric AI and other emerging data topics.

While the funding has facilitated the advancement of crucial research areas, the insights and conclusions of this report do not reflect the official policies of the foundation.

Introduction

The transformative potential of Artificial Intelligence (AI) and Machine Learning (ML) systems continues to impact society, propelled by advancements in Natural Language Processing (NLP) and Computer Vision (CV). These technologies have seen significant progress in recent years, leading to a surge in their applications and capabilities. These technologies, which leverage vast datasets to learn patterns through predictive models[1] [2], herald an era of enhanced automation and decision-making capabilities. However, the performance of AI models is linked not only to engineering aspects, but also to the data quality and the governance frameworks that guide how this data is collected, used and managed[3] [4].

While traditional AI development has emphasised model-centric approaches, focusing on model performance, this often overlooks comprehensive data governance, including consideration for data provenance, usage, tackling biases, and long-term handling, which are vital for creating AI/ML systems that are not only effective but also equitable, safe and compliant. This shift towards a 'data-centric' AI approach addresses these gaps by emphasising the quality of data throughout the AI lifecycle, ensuring better alignment with ethical standards and regulatory requirements[5] [6] [7] [8] [9].

[1] Mitchell (1997), 'Machine Learning textbook'.
[2] Jordan, M.I. (2019), 'Artificial Intelligence—The Revolution Hasn't Happened Yet'.
[3] Priestley, M., O'donnell, F., Simperl, E. (2023), 'A Survey of Data Quality Requirements That Matter in ML Development Pipelines'.
[4] Floridi, L., Cowls, J. (2019), 'A Unified Framework of Five Principles for AI in Society'.
[5] Floridi, L., Chiriatti, M. (2020), "GPT-3: Its Nature, Scope, Limits, and Consequences"
[6] Zha, D., Bhat, Z.P., Lai, K.-H., Yang, F., Hu, X. (2023), "Data-centric Artificial Intelligence: Perspectives and Challenges"
[7] Polyzotis, N., Zaharia, M. (2021), "What can Data-Centric AI Learn from Data and ML Engineering?"
[8] IAP (2024), "Introduction to Data-Centric AI"
[9] van der Schaar Lab (2024), "What is Data-Centric AI?"

However, current data governance frameworks still fall short in covering all phases of AI from design through to deployment, leading to high-performing models in tests that may fail in real-world applications, such as disease diagnostics that perform poorly due to a lack of diversity in training data[10] [11].

Effective data governance, defined as a structured framework of policies, processes and standards, is critical for managing data throughout its lifecycle in any organisation[12] [13] [14] [15]. Data governance addresses crucial issues such as data bias, algorithmic fairness and the ethical use of sensitive information, which are paramount in maintaining the reliability and trustworthiness of AI applications across critical areas like healthcare, finance and public service[16] [17] [18] [19]. Initiatives like BigScience and BigCode have significantly contributed to shaping robust data governance frameworks by establishing best practices that guide the handling and usage of data in AI projects[20] [21].

Our research aligns with these efforts and aims to further the discourse on what data governance looks like across the AI/ML data lifecycle. This report series underscores the importance of a holistic approach to data governance that spans the entire AI lifecycle, ensuring consistent handling of data across all phases. This approach is crucial for the responsible stewardship of data, permeating the entire process of developing AI/ML systems—from data collection and exploration to deployment and beyond.

---

[10] Schneider, J., Abraham, R., Meske, C., Vom Brocke, J. (2023), "Artificial Intelligence Governance For Businesses"

[11] Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., Aviles-Rivero, A.I., Etmann, C., McCague, C., Beer, L., Weir-McCall, J.R., Teng, Z., Gkrania-Klotsas, E., Rudd, J.H.F., Sala, E., Schönlieb, C.-B. (2021), "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans"

[12] Olavsrud, T. (n.d.), "What is data governance? Best practices for managing data assets"

[13] Gartner (2024), "Definition of Data Governance"

[14] DGI (2020), "Defining Data Governance"

[15] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K., (2021), "Datasheets for Datasets"

[16] Priestley, et al. (n 3)

[17] ODI (2023a), "Data-centric AI"

[18] ODI (2023b), "Unlocking the power of good data governance"

[19] Gerdes, A. (2022), "A participatory data-centric approach to AI Ethics by Design"

[20] Jernite, Y., Nguyen, H., Biderman, S., Rogers, A., Masoud, M., Danchev, V., Tan, S., Luccioni, A.S., Subramani, N., Dupont, G., Dodge, J., Lo, K., Talat, Z., Johnson, I., Radev, D., Nikpoor, S., Frohberg, J., Gokaslan, A., Henderson, P., Bommasani, R., Mitchell, M. (2022), "Data Governance in the Age of Large-Scale Data-Driven Language Technology"

[21] Hughes, S., de Vries, H., Robinson, J., Ferrandis, C.M., Allal, L.B., von Werra, L., Ding, J., Paquet, S., Jernite, Y. (2023), "The BigCode Project Governance Card"

Our findings aim to provide actionable insights and recommendations for policymakers, practitioners and researchers, fostering a more structured and unified policy framework that enhances data quality, explainability, transparency, fairness, safety, auditability and compliance across diverse AI applications. By studying the roles of various actors— from data scientists and engineers to domain experts, risk stewards and project managers—this second report in the series analyses how these actors interact within this ecosystem to build a further layer of context toward understanding data governance in AI/ML system development. These interactions are crucial for outlining 'who' is handling data and 'how', where value exchange mechanisms include not only the sharing of datasets but also insights, consent, evaluation, audit of models and best practice across different stages of the AI lifecycle.

In this follow-up report, we extend the discussion from individual lifecycle stages to a broader examination of the data ecosystems that support AI/ML development. By integrating insights from our previous studies and the latest advancements in the field, we aim to offer a comprehensive overview of how data governance practices can be optimised to support the dynamic needs of AI/ML systems in a rapidly evolving digital landscape.

Synthesis

Ecosystem

Actors

Actors in a data ecosystem are defined as 'people, communities and organisations that are stewarding data, creating insights and applications from it, deciding what to do based on it, influencing any of those activities, or are affected by any of those activities'[22]. Developing AI/ML systems involves a diverse set of roles and stakeholders to ensure responsible and effective data governance (and, of course, stewardship) throughout the lifecycle.

As indicated below, some actors contribute to specific stages of the AI data lifecycle, while others are more ambiguous and instead participate in certain activities that cannot be neatly categorised. According to the literature, key roles in this task include:

Data Scientists

These are typically senior roles with several years of experience. They are responsible for problem formulation, conceptualising solutions, identifying representative data sources, and prescribing the problem and solution to more junior data scientists[23]. Data scientists possess skills in AI, ML, data visualisation, and computer science techniques[24].

- **Collection and creation:** Data scientists determine the specific types of data required for projects and oversee the collection and initial data cleaning processes. They ensure that the data collected aligns with the project's needs and is of high quality, involving preprocessing tasks like data cleansing, labelling and transformations. This step is crucial for setting a solid foundation for subsequent modelling efforts[25].

[22] ODI (2023c), 'Defining responsible data stewardship'.
[23] De Silva, D., Alahakoon, D. (2022), 'An artificial intelligence life cycle: From conception to production'.
[24] Coulthart, S., Shahriar Hossain, M., Sumrall, J., Kampe, C., Vogel, K.M. (2023), 'Data-Science literacy for future security and intelligence professionals'.
[25] De Silva and Alahakoon (n 23)

- **Exploration and analysis:** Data scientists use various tools and methodologies for exploratory data analysis (EDA) to gain insights into dataset characteristics, patterns and anomalies. Their expertise in data exploration drives the formulation of hypotheses and guides subsequent analysis tasks. This process is also referred to as data mining[26].

- **Preprocessing**: Data scientists are pivotal in designing and implementing data preprocessing pipelines, which involve deciding on transformation types, the order of transformations, and feature-wise preprocessing strategies. They collaborate with domain experts to interpret results and validate the relevance of preprocessing steps[27] [28]. Post-data preprocessing, data scientists also focus on developing predictive models by selecting suitable modelling techniques, implementing these models, and optimising their performance in alignment with their data science expertise[29] [30].

- **Evaluation:** Once the models are developed, data scientists evaluate their performance against predefined metrics and benchmarks. They conduct tests to validate the models' accuracy, reliability and fairness, ensuring they meet the project's requirements before deployment[31].

- **Deployment:** In the final phase, data scientists collaborate with AI/ML engineers to transition models from a development environment to production. This includes any final optimisation, scaling models for broader use, and ensuring that the deployed models continue to perform well under varied real-world conditions[32].

As we shift from data scientists to data engineers, the focus moves from analysing data to managing its entire lifecycle. While data scientists leverage their expertise in AI and ML to derive insights, data engineers create robust systems that ensure data is clean, accessible and efficiently managed. This partnership is crucial, as it ensures that sophisticated models have a solid, well-maintained data foundation, enabling scalable and effective AI/ML solutions that support informed decision-making across industries.

---

[26] Patel, H., Guttula, S., Gupta, N., Hans, S., Mittal, R., N, L. (2023), 'A Data-centric AI Framework for Automating Exploratory Data Analysis and Data Quality Tasks',
[27] Li, P., Chen, Z., Chu, X., Rong, K. (2023), 'DiffPrep: Differentiable Data Preprocessing Pipeline Search for Learning over Tabular Data'.
[28] Wang, D., Weisz, J.D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., Gray, A. (2019), 'Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI'.
[29] De Silva and Alahakoon (n 23)
[30] Gerdes, A. (2022), 'A participatory data-centric approach to AI Ethics by Design'.
[31] ibid.
[32] ibid.

Data Engineers

Data engineers are highly skilled professionals who play a pivotal role in managing the entire data lifecycle, from planning to deployment. Their expertise encompasses a wide array of responsibilities, all aimed at maintaining high standards of data integrity and usability. Data engineers continuously refine data workflows and contribute significantly to the development of scalable data solutions. By implementing robust data management systems and infrastructure, data engineers ensure that data (and insights from it) remains actionable and secure throughout the AI/ML data lifecycle, helping organisations to make informed decisions[33] [34].

- **Data lifecycle management:** Data engineers are crucial in managing the entire lifecycle of data, which includes stages such as collection and creation, preprocessing, exploration and analysis, and deployment (including post-deployment tasks like ongoing monitoring). They focus on ensuring that each phase is handled with precision, meeting specific data requirements and adhering to ethical principles to avoid issues like bias and unfair model outcomes[35].
- **Infrastructure and pipeline development:** In the context of data engineering and infrastructure, data engineers are responsible for the development and maintenance of the infrastructure that supports data-related workflows. This involves setting up and managing data storage systems, databases, and processing environments that ensure data quality and accessibility for further analysis and model training[36].
- **Specification of data requirements:** Data engineers play a key role in specifying the data requirements for each phase of the data lifecycle (for example, ongoing planning for any further data acquisition beyond initial collection/creation). This can include defining data collection methods, data quality benchmarks, and the protocols for data cleaning and preparation. Such specifications help in standardising data handling practices and ensuring that the data used in AI/ML systems is of high quality and fit for purpose[37].

[33] Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., Jacquet, A. (2023), 'Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering'.
[34] Piorkowski, D., Park, S., Wang, A.Y., Wang, D., Muller, M., Portnoy, F. (2021), 'How AI Developers Overcome Communication Challenges in a Multidisciplinary Team: A Case Study'.
[35] Lu et al. (n 33)
[36] Piorkowski (n 34)
[37] Lu et al. (n 33)

- **Feature engineering:** Data engineers contribute to the creation of custom features and manage feature definitions to ensure consistency and relevance over time. They encounter challenges related to maintaining model reproducibility and adapting to changing data environments[38].
- **Ethical and reliable data handling:** Data engineers are also responsible for implementing data handling practices that conform to ethical guidelines and regulatory requirements. This includes ensuring data privacy, securing data against breaches, and maintaining transparency in data processing and usage[39].

Data engineers orchestrate the complete data lifecycle, ensuring integrity and usability from collection and creation to deployment. They develop and maintain robust infrastructures that support seamless data workflows, enabling effective AI/ML applications. Their rigorous standards for data handling help prevent biases and uphold data security.

On the other hand, domain/subject matter experts integrate industry-specific knowledge into AI/ML projects, guiding everything in this capacity from data collection to model evaluation. Their expertise ensures that AI/ML systems are relevant and ethically aligned, with real-world standards, contributing significantly to how data can shape responsible and effective AI/ML systems.

Domain/Subject Matter Experts (SMEs)

Involving domain experts through participatory activities is crucial for improving data activities and ensuring alignment with domain knowledge[40] [41]. Their deep expertise in specific fields helps the AI team understand real-world problems and provide guidance[42].

- **Planning:** Domain experts are crucial in the initial stages (for example, data discovery during planning activities) where project objectives and requirements are defined. They contribute to formulating the specific needs that the AI solution must meet and ensure that the requirements are aligned with industry standards and practical realities. The literature points out the importance of involving domain experts in participatory activities, such as workshops and brainstorming sessions, to ensure the alignment of any data activities with domain knowledge[43] [44].

---

[38] Orr, L., Sanyal, A., Ling, X., Goel, K., Leszczynski, M. (2021), 'Managing ML pipelines: feature stores and the coming wave of embedding ecosystems'.
[39] Lu et al. (n 33)
[40] Gerdes (n 30)
[41] Piorkowski (n 34)
[42] ibid.
[43] Gerdes (n 30)
[44] Piorkowski (n 34)

- **Collection and creation:** During this phase, SMEs can ensure that the data collected is representative of real-world scenarios and that it addresses the nuances of the domain. Their insights help in selecting and validating the right datasets for model training and testing, ensuring the data reflects true operational conditions[45].
- **Exploration and analysis:** Domain experts can play a critical role in selecting effective representations of data points or attributes, ensuring that the exploration and analysis process aligns with domain-specific requirements[46].
- **Preprocessing:** SMEs collaborate with data scientists to provide insights into the underlying meaning of raw data and help define meaningful features for preprocessing. Their domain knowledge enhances the effectiveness of preprocessing techniques[47].
- **Evaluation:** In the evaluation stage, SMEs are essential for critically assessing the data used in AI models. They contribute significantly to establishing the 'ground truth', which serves as a benchmark for measuring model performance. Their domain-specific insights are invaluable for evaluating the data's quality and relevance, ensuring that the model's outputs are rigorously tested against real-world data and outcomes. This scrutiny helps in identifying any misalignments or biases in the data, which, if unaddressed, could lead to inaccurate model evaluations and, therefore, deployments[48].
- **Governance and ethics:** SMEs can also play a role in the governance of AI projects. Their domain-specific insights are crucial for ensuring ethical compliance and for setting up standards that govern the development and use of data throughout AI/ML systems within specific sectors[49].

Domain/SMEs significantly influence AI projects by ensuring that solutions align with industry standards and practical realities from the planning phase. They refine data collection to accurately reflect real-world scenarios and infuse domain knowledge into data preprocessing and evaluation, ensuring AI models are tested against true operational benchmarks.

[45] Gerdes (n 30)
[46] De Silva and Alahakoon (n 23)
[47] Wang et al. (n 28)
[48] Gerdes (n 30)
[49] Schneider, J., Abraham, R., Meske, C., Vom Brocke, J. (2023), 'Artificial Intelligence Governance For Businesses'.

Transitioning from domain expertise to technical execution, AI/ML and software engineers take these conceptual frameworks into the operational realm. They are instrumental in developing and deploying AI/ML systems, employing DevOps principles[50] [51] to streamline and scale these solutions. By continuously monitoring and refining models based on real-world data and user feedback, these engineers bridge the gap between theoretical design and practical application, ensuring the AI/ML systems remain dynamic and effective over time.

AI/ML and software engineers:

These roles are responsible for transforming prototypical AI models into deployed services or solutions that can be accessed by stakeholders and end-users[52]. Although not explicitly related, these roles can reflect the unique data-related challenges presented by AI/ML systems, such as tracking data lineage, which AI/ML engineers manage through specialised techniques[53].

- **Collection and creation:** Technical experts, including ML engineers and software engineers, play a pivotal role in developing and maintaining the automated data collection tools essential for extracting vast amounts of data efficiently[54]. In this way, software engineers focus on upstream datasets closer to the dataset collection and creation processes, contributing to data documentation efforts and addressing issues related to dataset bias and fairness[55].
- **Model development and prototyping:** AI/ML engineers are responsible for turning initial concepts and data-driven insights, developed by data scientists, into prototypical models. This role involves technical expertise in AI algorithms and practical skills in modelling and evaluation, setting the stage for deployment[56].

[50] GitHub (2022), 'DevOps fundamentals: Defining DevOps principles'.
[51] DevOps principles are the core values and practices that bring development, IT operations, and security teams together to collaborate, automate the entire software delivery lifecycle, continuously integrate and deploy code changes, make data-driven decisions through monitoring and measurement, and embrace a culture of learning from failures.
[52] De Silva and Alahakoon (n 23)
[53] Paleyes, A., Urma, R.-G., Lawrence, N.D. (2022), 'Challenges in Deploying Machine Learning: A Survey of Case Studies'.
[54] Coulthart et al. (n 24)
[55] 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'.
[56] De Silva and Alahakoon (n 23)

- **Deployment:** AI/ML engineers play a crucial role in transitioning prototypical AI models into deployed, operational services or solutions. This process includes ensuring that these models are standardised and can be accessed effectively by all stakeholders and end-users. It often involves applying DevOps principles to the specific challenges of AI/ML systems, such as maintaining model accuracy and handling real-time data feeds[57] [58].
- **Monitoring and maintenance:** Post-deployment, AI/ML engineers ensure that systems operate effectively under real-world conditions, a responsibility also referred to as data quality monitoring. They monitor the systems for performance issues, adapt them to changing data environments, and manage the retraining of models to handle concept drift or changes in data distributions. This phase is crucial for maintaining the relevance and accuracy of the AI/ML systems over time, [59].
- **Feedback loops and continuous improvement:** AI/ML engineers are involved in setting up feedback loops that collect performance data and user input to continuously improve the AI models. This involves analysing the operational data to detect and correct biases, improve model performance, and refine the user experience based on actual usage patterns[60].

AI/ML and software engineers play a crucial role in transforming prototypical AI models into operational services, developing and maintaining data collection tools, and addressing data-related challenges such as bias and documentation. They ensure models are effectively deployed, continuously monitored, and improved through feedback loops.

Building on the foundation laid by these engineers, we now turn to project managers, who oversee the coordination and execution of AI projects, ensuring alignment among teams, managing data processes, and maintaining high standards of data quality and model accuracy. Their role is pivotal in facilitating communication, documentation and compliance, ultimately driving the success of AI initiatives.

[57] De Silva and Alahakoon (n 23)
[58] Paleyes et al. (n 53)
[59] ibid.
[60] ibid.

Project managers

Project managers oversee the coordination and execution of AI projects, ensuring alignment between different teams and stakeholders[61]. The role involves managing the project's processes and workflow, including data handling by data scientists and engineers, thereby indirectly influencing how data is utilised and governed[62].

- **Planning and coordination:** Project managers play a critical role in overseeing the coordination and execution of AI projects, ensuring alignment between different teams such as AI developers, data scientists and domain experts. They ensure that all data-related activities—from collection and processing to model training—are aligned with the project's objectives and timelines. By orchestrating the workflows that involve data scientists and engineers, project managers directly influence the efficiency and integrity of data handling, ensuring that the data used is relevant, accurate, and governed according to the project's standards[63].
- **Managing data processes:** Through the AI data lifecycle, project managers are instrumental in managing processes and workflows that include data handling by data scientists and engineers. This role is crucial in ensuring that data collection, processing and exploration tasks are conducted efficiently and align with project goals. Project managers thereby indirectly influence how data is used and governed within the project, ensuring that data-related tasks adhere to planned protocols and timelines[64].
- **Monitoring and quality assurance:** Project managers are also responsible for monitoring the ongoing progress of AI projects, including the quality assurance of both the data used and the AI models developed. They ensure that data quality, model accuracy and performance metrics meet the project's standards and requirements. This oversight helps to identify any potential issues early, allowing for timely interventions to correct course as necessary[65].

[61] De Silva and Alahakoon (n 23)
[62] Piorkowski (n 34)
[63] De Silva and Alahakoon (n 23)
[64] Piorkowski (n 34)
[65] De Silva and Alahakoon (n 23)

- **Facilitating communication and documentation:** Effective communication across various project teams and with stakeholders is essential. Project managers facilitate this by ensuring that all parties are regularly updated on project progress and any changes in project scope or objectives. They also oversee the documentation process, ensuring that all aspects of the AI development process are well-documented and transparent, which is crucial for project audits, compliance, and future referencing[66].

Project managers are integral to the success of AI projects, overseeing the coordination and execution to ensure alignment across various teams and stakeholders. They manage project processes and workflows, influencing how data is handled by data scientists and engineers. Their role spans from planning and coordinating data-related activities to monitoring progress and ensuring data quality and model accuracy. Effective communication and thorough documentation are also key responsibilities, ensuring transparency and compliance throughout the AI development process.

As AI projects progress, a wide range of stakeholders become involved. Business experts, risk stewards, legal counsel and auditors all play crucial roles in managing different aspects of data handling. These stakeholders must continuously adapt their roles to align with the AI/data lifecycle, ensuring effective governance and risk assessment. Collaboration among multidisciplinary teams, including HCI researchers and domain experts, is essential. These experts contribute valuable insights into designing tools and platforms that support data scientists, fostering responsible and effective AI development that meets ethical and regulatory standards.

Wider stakeholders

Various stakeholders, including business experts, risk stewards, legal counsel, software developers, auditors, regulators, end-users, and impacted individuals or communities, may be involved in different aspects of handling data during projects to develop AI/ML systems. These stakeholders must adapt and evolve their roles and frameworks to align with the AI/data lifecycle. For instance, quality assurance and audit functions are critical internal and independent stakeholders that need to continuously update their approaches to ensure effective risk assessment and governance[67] [68].

[66] Piorkowski (n 34)
[67] Lu et al. (n 33)
[68] Piorkowski (n 34)

The planning stage often involves a multidisciplinary team with expertise in computer science, data science, human-computer interaction, cognitive science, and domain-specific knowledge[69]. Collaboration and communication between heterogeneous roles is essential to enable responsible and effective AI development, aligned with ethical principles, legal regulations and stakeholder needs.

HCI researchers contribute to developing tools and features that support data scientists in their exploration and analysis endeavours. Their insights inform the design of interactive visual analytics tools, coding environments, and collaborative platforms tailored to the needs of data science workflows[70].

Various stakeholders, including business experts, risk stewards, legal counsel, software developers, auditors, regulators, end-users and impacted communities, are integral to AI/ML projects, and each of these handles different aspects of data management. During the planning stage, multidisciplinary teams with expertise in computer science, data science, human-computer interaction, cognitive science and domain-specific knowledge collaborate to ensure responsible and effective AI development. Therefore, it is clear that the sectoral context also influences the roles and responsibilities of actors in AI/ML development.

Sectoral Context

As highlighted in the previous report, the differences between industry entities and academic institutions in developing AI/ML systems significantly affect actos' roles and responsibilities. In industry, there is a greater emphasis on roles that align with business outcomes, such as project managers, data engineers and AI/ML engineers, who are tasked with integrating AI solutions into broader business processes and ensuring these solutions meet stringent operational and regulatory standards. This set-up needs a structured data governance model that supports the complex hierarchy and functional demands of business environments, ensuring data integrity, compliance and operational efficiency[71] [72].

[69] Aldoseri, A., Al-Khalifa, K.N., Hamouda, A.M. (2023), "Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges"
[70] Wang et al. (n 28)
[71] Priestley, et al. (n 3)
[72] Schneider et al. (n 49)

Conversely, in academic settings, the actors primarily include researchers and PhD students, whose roles are more fluid and defined by the needs of experimental and theoretical research rather than strict business objectives. The focus here is on innovation and the advancement of knowledge, which allows for a more flexible approach to data governance that encourages exploration and experimentation[73]. This difference fundamentally affects how data is managed, with academic projects often lacking the formal governance structures found in industry, which are crucial for scalable, reliable AI deployment in commercial applications[74]. These distinctions underline the varying challenges and priorities faced by actors in academic versus industry contexts, shaping the development and implementation strategies of AI/ML systems.

Value exchanges

Alongside the key actors involved, a feedback loop joins these entities together to amplify the value exchange between them. That is: 'the links between the assets, how data is accessed and used and shared, helps to highlight points where data value can be realised'. In this section, we situate these exchanges materially, providing a brief overview of the kind of tools that facilitate these exchanges and what data/insights are shared in the process.

As discussed, the AI data lifecycle ecosystem involves multiple stakeholders, including data scientists, domain experts, business experts, strategy consultants and software developers. Effective value exchanges or tools and media for sharing insights and data across these roles are crucial. Several key mechanisms emerge from the literature:

Data collection and preparation tools

In AI/ML systems, the richness and diversity of data sources are foundational for developing robust models. Recognised datasets, such as those from large-scale web data extractions and specialised databases, are critical, as they often come from varied yet reliable sources, ensuring a broad representation of phenomena[75]. For managing these vast datasets, platforms like AWS SageMaker, Microsoft ML and TensorFlow TFX are indispensable, as they provide comprehensive solutions for data storage, model hosting and deployment, significantly reducing the operational challenges associated with maintaining ML models[76].

[73] Gerdes (n 30)
[74] Priestley et al. (n 3)
[75] Corchado, J.M., F, S.L., V, J.M.N., S, R.G., Chamoso, P. (2023), 'Generative Artificial Intelligence: Fundamentals'.
[76] Paleyes et al. (n 53)

Enhancing the efficiency of data handling are tools such as Goods and Juneau, which automate the discovery and cataloguing of datasets. These tools streamline the integration of data from heterogeneous sources, including relational and graph databases, and are pivotal in managing the complexity of large data volumes[77].

Alternative data collection methods also enrich the dataset landscape. Platforms like Amazon Mechanical Turk facilitate the creation of diverse datasets through crowdsourcing, while simulators such as Hermoupolis and Crash to Not Crash provide synthetic data generation capabilities, enhancing both the diversity and the volume of data available for AI/ML training[78].

In supporting these technological efforts, effective documentation is crucial, as it enhances communication among dataset creators, consumers and stakeholders, thereby improving the transparency and understanding of data sources[79]. Specialised services such as Amazon SageMaker Ground Truth and Google Cloud Labeling further streamline the data-labelling process, ensuring the availability of high-quality datasets for model training and evaluation. Tools like Snorkel and Snuba exemplify innovation in automating the generation of labelling functions, which accelerates the preparation of training data[80].

Moreover, frameworks like Facets and TensorFlow Data Validation (TFDV) provide schema-based validation and anomaly detection, which are essential for maintaining data integrity and consistency across the AI development lifecycle[81] [82]. Libraries such as Deequ and mlinspect offer declarative inspection of ML pipelines, enabling data scientists to validate preprocessing steps and verify the integrity of data transformations effectively[83].

Collaboration among data scientists, subject matter experts and engineers is vital in streamlining data preparation tasks. By leveraging diverse expertise, collaborative efforts can enhance the efficiency and effectiveness of data collection and preprocessing activities[84].

---

[77] Bellomarini, L., Fayzrakhmanov, R.R., Gottlob, G., Kravchenko, A., Laurenza, E., Nenov, Y., Reissfelder, S., Sallinger, E., Sherkhonov, E., Vahdati, S., Wu, L. (2022), 'Data science with Vadalog: Knowledge Graphs with machine learning and reasoning in practice'.
[78] Whang, S.E., Lee, J.-G. (2020), 'Data collection and quality challenges for deep learning'.
[79] Heger et al. (n 55)
[80] Whang and Lee (n 78)
[81] ibid.
[82] Li et al. (n 27)
[83] Whang and Lee (n 78)
[84] Wang et al. (n 28)

Data exploration and analysis tools

Data scientists have extensively used notebook environments like Jupyter Notebook, JupyterLab and Google Colaboratory to integrate code, visualisations and descriptive text, enhancing literate programming and data science workflow documentation. Additionally, modern low-code and no-code platforms such as Microsoft Power BI, Tableau and Knime further democratise data science by enabling users with minimal coding expertise to perform complex data analyses and visualisations.

These environments support a wide range of data activities from data exploration to machine learning model deployment, facilitating the robust data management practices that are essential for scalable AI solutions[85] [86]. These tools exemplify the trend towards more accessible and integrated data science workflows, addressing the needs for transparency, compliance and stakeholder engagement in AI development[87] [88].

These tools also offer coding environments that support interactive data exploration, visualisation and storytelling, empowering data scientists to conduct comprehensive EDA tasks[89]. Collaborative platforms and features integrated into coding environments facilitate asynchronous sharing and collaboration among data science teams, enhancing knowledge sharing and collaboration during the exploration and analysis phase (as mentioned, also known as data mining)[90].

Furthermore, data formats, such as traditional graph data formats like RDF and Property Graphs, offer versatile frameworks for analysing structured data, enhancing exploratory and analysis capabilities. Beyond these, electronic health records (EHRs) provide rich, textual data detailing patient histories, which are crucial for medical data analysis, though presenting challenges in labelling and privacy[91]. Similarly, ML systems leverage diverse datasets, from social media platforms to real-world data, which must be handled with care to address privacy concerns and bias[92] [93].

---

[85] Gerdes (n 30)
[86] Schneider et al. (n 49)
[87] Lu et al. (n 33)
[88] Coulthart et al. (n 24)
[89] Wang et al. (n 28)
[90] ibid.
[91] Bashath, S., Perera, N., Tripathi, S., Manjang, K., Dehmer, M., Streib, F.E. (2022), 'A data-centric review of deep transfer learning with applications to text data'.
[92] Wan, Z., Wang, Zhixiang, Chung, C., Wang, Zheng (2023), 'A Survey of Dataset Refinement for Problems in Computer Vision Datasets'.
[93] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021), 'A Survey on Bias and Fairness in Machine Learning'.

These varied data formats collectively enhance the scope and depth of data analysis, catering to specific domain needs while navigating ethical and operational challenges.

Tools like Deequ, Pandas Profiling and TFDV enable data scientists to assess the quality of datasets by identifying issues such as missing values, outliers and data inconsistencies. These tools support basic data analysis and profiling functionalities, albeit with limitations in addressing advanced data-quality challenges[94].

Data visualisation and presentation tools

Presentation tools like PowerPoint and Mural serve as vital mediums for visually representing information and explaining complex concepts, models and predictions to non-technical stakeholders. Data scientists create specialised visualisations, such as confusion matrices, feature importance plots, and model comparisons, to convey insights effectively[95].

Data-driven visualisations using statistical metrics like Earth Mover's distance and interactive visual analytics tools, such as ZQL query language, facilitate the exploration and interpretation of complex datasets, enabling data scientists to derive meaningful insights[96].

Documentation and standardisation tools

To standardise and streamline documentation across the AI lifecycle, researchers propose various approaches. These include dataset documentation, model service documentation, information sheets, questionnaires, composable widgets, and checklists. Some employ data-driven components that can automatically analyse data samples, metadata and model outputs to generate insights about datasets and model behaviour[97]. Collaborative platforms and automated AI (AutoAI) systems can enhance collaboration among data science teams. AutoAI can automate tasks like data cleaning and model building during individual work phases, while advising teams based on collective efforts during collaborative phases[98].

---

[94] Patel et al. (n 26)
[95] Piorkowski (n 34)
[96] Patel et al. (n 26)
[97] Micheli, M., Hupont, I., Delipetrev, B. and Soler-Garrido, J. (2023), 'The landscape of data and AI documentation approaches in the European policy context | Ethics and Information Technology'.
[98] Wang et al. (n 28)

Efficient data documentation practices, supported by dedicated locations such as wikis and version control repositories, streamline information retrieval and knowledge sharing, improving transparency and reproducibility in data analysis workflows[99]. Advanced EDA techniques, including clustering methods and visualisation tools, enable analysts to uncover hidden patterns, relationships and trends within datasets, supporting hypothesis generation and model development[100].

[99] Heger et al. (n 55)
[100] Chicco, D., Oneto, L., Tavazzi, E. (2022), 'Eleven quick tips for data cleaning and feature engineering'.

Concluding remarks

Our examination of the AI/ML ecosystem highlights the critical roles played by various actors, from data scientists and engineers to project managers and wider stakeholders, in ensuring effective data governance throughout the AI lifecycle. This report supports the necessity of coordinated efforts among these actors to address data quality, ethical considerations and compliance requirements. The involvement of domain experts and stakeholders, such as business leaders, risk stewards and legal advisors, further enriches the governance framework by integrating diverse perspectives and expertise.

By extending our analysis to the broader AI data ecosystem, we have identified key value exchanges and interactions that enhance the robustness of AI/ML systems. Tools and practices facilitating data collection, exploration, analysis and documentation play a pivotal role in maintaining data integrity and transparency. Our findings advocate for a holistic approach to data governance, emphasising continuous improvement through feedback loops and collaborative efforts across multidisciplinary teams.

As we move forward, the insights from this report will inform our ongoing research into optimising data governance practices. Future work, in our forthcoming third report of the series, will focus on developing actionable recommendations to bridge gaps identified in current frameworks, to foster a more unified and effective governance model. Our ultimate goal is to support the development of AI/ML systems that are not only technologically advanced but also ethically sound and socially beneficial, ensuring they can be reliably deployed in diverse real-world applications.

We hope that these insights will contribute to the broader discourse on responsible AI development and encourage further collaboration among researchers, practitioners and policymakers. By collectively addressing the challenges and opportunities presented by AI data governance, we can pave the way for an AI future that is both innovative and aligned with societal values.

Annex A. Methodology

A scoping review is a type of knowledge synthesis that maps the existing literature on a particular topic[101]. Unlike systematic reviews, which focus on answering specific research questions with a narrow scope, scoping reviews aim to provide a broad overview of the available evidence, identify gaps in knowledge, and explore diverse perspectives within the literature[102] [103]. This methodological approach is particularly useful for emerging or complex topics where existing research may be diverse and heterogeneous in nature, which, as we show, AI/ML data governance proves to be[104] [105].

Research question

The primary research question guiding this scoping review is: How is data governance implemented across the AI data lifecycle by those involved in developing AI/ML systems?

Sub questions

1. What are the key steps of the AI data lifecycle?
2. Who is involved in each step of the AI data lifecycle and what is the relationship between each actor?
3. What are the governance considerations to be made at each stage of the AI data lifecycle?

[101] Mak, S., Thomas, A. (2022), 'Steps for Conducting a Scoping Review'.
[102] Levac, D., Colquhoun, H., O'Brien, K.K. (2010), 'Scoping studies: advancing the methodology'.
[103] Arksey, H., O'Malley, L. (2002), 'Scoping studies: towards a methodological framework'.
[104] Floridi et al. (n 4)
[105] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., Barnes, P. (2020), 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing'.

Conceptual framework

Our conceptual framework for this research question structures our search syntax by progressing through four stages:

1. AI terms (broad)
2. Data terms (specific; governance, ecosystem, lifecycle)
3. Evidence type
4. Exclusions

This framework ensures a structured approach to identifying relevant literature while excluding less relevant studies.

Using this conceptual framework, we generated a sequence of relevant keywords under each stage for syntax to deploy in academic databases. Firstly, starting broad, we cover terms relating to the general AI domain. Secondly, we then included terms detailing data governance, ecosystem and lifecycle-related categories. Next, we covered some desired types of evidence, from studies outlining processes, pipelines, practices and even any work on knowledge graphs[106] [107], thus worthy of inclusion. Finally, through a trial-and-error approach, we added some exclusionary keywords to filter out irrelevant domains that appeared consistently in preliminary searches. The final search syntax is outlined in the next subsection.

Next, to structure our synthesis of the literature, the conceptual framework for the AI data lifecycle, encompassing stages from planning to deployment, is justified by insights drawn from the literature. We observe common descriptions of certain stages in the AI data lifecycle. Bringing these common stages together results in the following:

[106] ATI (2024b), 'Knowledge graphs'.

[107] Which organise disparate data sources and can facilitate explainability of data science workflows and data ecosystems through visual means

Table 1. AI data lifecycle framework

AI data lifecycle stage	Related stage and source
Planning	'Project initiation and design'[108] 'Design, Planning, and Prototyping'[109] 'Problem Formulation'[110] 'Identify and Formulate the Problem' and 'Review Data and AI Ethics'[111] 'Planning phase'[112]
Collection and creation	'Data collection'[113] 'Data Collection'[114] 'Data Discovery'[115] 'External Data Acquisition'[116] 'Collection phase'[117] 'Data Ingestion'[118]
Exploration and analysis	'Data analysis and preprocessing'[119] 'Error Detection and Repair'[120] 'Data Preparation' and 'Data Exploration'[121] 'Preparation phase' and 'Analysis phase'[122] 'Data Preparation & Cleansing'[123]
Preprocessing	'Data analysis and preprocessing'[124] 'Data Cleaning and Data Annotation'[125] 'Data Preparation – Data Cleaning and Data Labeling'[126] 'Data Preprocessing', 'Data Augmentation'[127] 'Preparation phase'[128] 'Data Preparation & Cleansing', 'Data Transformation'[129]

[108] ICO (2024), 'Annex A: Fairness in the AI lifecycle'.
[109] Chai, C., Wang, J., Luo, Y., Niu, Z., Li, G. (2023), 'Data Management for Machine Learning: A Survey'.
[110] Khan, M.J., Breslin, J.G., Curry, E. (2023), 'Towards Fairness in Multimodal Scene Graph Generation: Mitigating Biases in Datasets, Knowledge Sources and Models'.
[111] De Silva et al. (n 193)
[112] Shah, S.I.H., Peristeras, V., Magnisalis, I. (2021), 'DaLiF: a data lifecycle framework for data-driven governments'.
[113] ICO (n 108)
[114] Khan et al. (n 110)
[115] Chai et al. (n 109)
[116] De Silva et al. (n 111)
[117] Shah et al. (n 112)
[118] Cognilytica, (2020), 'AI Data Engineering Lifecycle Checklist'.
[119] ICO (n 108)
[120] Chai et al. (n 109)
[121] De Silva et al. (n 111)
[122] Shah et al. (n 112)
[123] Cognilytica (n 312)
[124] ICO (n 302)
[125] Khan et al. (n 304)
[126] Chai et al. (n 303)
[127] De Silva et al. (n 193)
[128] Shah et al. (n 306)
[129] Cognilytica (n 312)

Training	‘Model development‘[130] ‘Model Training and Evaluation‘[131] ‘Model Training & Inference – In-ML‘[132] ‘Build Initial AI Model‘, ‘Data Augmentation‘, and ‘Build Multiple AI Models‘[133]
Evaluation	‘Model evaluation‘[134] ‘Model Evaluation‘[135] ‘Model Management – Model Storage, Versioning, Query, and Diagnosis‘[136] ‘Evaluate Primary Metrics‘ and ‘Evaluate Secondary Metrics‘[137] ‘Use, re(use), and feedback‘[138]
Deployment	‘Model deployment and monitoring‘[139] ‘Model Deployment and Inference‘[140] ‘Model Deployment and Serving‘[141] ‘AI Model Deployment and Risk Assessment‘, ‘Post-deployment Review’[142] ‘Use, re(use), and feedback‘[143] ‘Data and Production Orchestration‘[144]

[130] ICO (n 302)
[131] Khan et al. (n 304)
[132] Chai et al. (n 303)
[133] De Silva et al. (n 193)
[134] ICO (n 302)
[135] Khan et al. (n 304)
[136] Chai et al. (n 303)
[137] De Silva et al. (n 193)
[138] Shah et al. (n 306)
[139] ICO (n 302)
[140] Khan et al. (n 304)
[141] Chai et al. (n 303)
[142] De Silva et al. (n 193)
[143] Shah et al. (n 306)
[144] Cognilytica (n 312)

Synthesising the descriptions from each of these stages, we arrive at the following common definitions of how each stage is supported:

1. **Planning:** The beginning of the lifecycle involves defining the problem, considering the data needed and how it will be collected. The paper illustrates how datasets are deeply integrated into MLR work practices, not just for training and testing models but also for organising the scientific field around shared problems.

2. **Collection and creation:** The increasing concentration on fewer datasets within task communities and the significant adoption of datasets from other tasks, as revealed in the paper, suggests a need for more diversified dataset creation. This supports expanding the lifecycle stage to focus not only on data collection but also on the creation of new datasets to address task-specific needs and reduce reliance on datasets not originally designed for the task at hand.

3. **Exploration and analysis:** Given the paper's findings on dataset usage patterns and the critical examination of benchmark datasets, there's a clear indication that more attention should be given to exploring and analysing datasets to understand their biases, limitations, and fitness for the intended application. This aligns with integrating an exploration and analysis phase in the lifecycle to assess datasets' suitability and potential impacts on model performance and fairness.

4. **Preprocessing:** The necessity of dataset preprocessing is supported by discussions on the need for datasets to closely align with real-world tasks for accurate scientific progress measurement and model deployment. Effective preprocessing ensures that datasets are cleaned, annotated and formatted in ways that maximise their utility and relevance to specific AI/ML tasks.

5. **Training:** The focus on benchmark datasets for training models, as discussed, highlights the training stage's significance. However, the paper also calls for a broader perspective on training practices, considering the diverse and often complex nature of real-world applications beyond benchmark performance.

6. **Evaluation:** The findings on the concentration of research efforts on established benchmarks and the ethical concerns associated with biased datasets underscore the importance of rigorous evaluation methods that go beyond traditional benchmarking to include ethical and societal considerations.

7. **Deployment:** Finally, the deployment stage is implicitly validated by the paper through discussions on the ecological validity of AI/ML research and the impact of deployed models on society. It suggests a need for models to be tested in diverse, real-world settings to ensure they perform ethically and effectively outside controlled environments.

Using these two conceptual frameworks (one for the search and one for the synthesis), we now turn to the syntax and search results.

Syntax and search strategy

The syntax used for the search strategy is as follows:

("artificial" AND "intelligence" OR "ai" OR "machine" AND "learning" OR "ml" OR "data-centric" AND "ai" OR "data-centric" AND "artificial" AND "intelligence")

AND ("data" AND "governance" OR "data" AND "management" OR "data" AND "ecosystem*" OR "data" AND "ethic*" OR "data" AND "quality" OR "data" AND "security" OR "data" AND "access" OR "data" AND "collection" OR "data" AND "acquisition" OR "data" AND "curation" OR "data" AND "wrangling" OR "data" AND "mining" OR "data" AND "preprocessing")

AND ("process*" OR "pipeline*" OR "practic*" OR "best" AND "practice*" OR "guid*" OR "knowledge" AND "graph")

AND NOT ("metaverse" OR "iot" OR "internet" AND "of" AND "things" OR "quantum" AND "computing" OR "nanotechnology" OR "edge" AND "comput*")

We conducted an initial search on Scopus and ACM Digital Library, focusing on articles and conference papers (hereafter labelled 'studies') published between 2019 and 2024. There are two reasons for this limitation:

1. DCAI as a subject area, and movement galvanised by Andrew Ng, has taken shape since 2021[145] [146] [147] [148] [149] [150] [151] [152]. From this point, the notion of shifting away from purely model-centric approaches to AI/ML and towards enhancing the quality of data has grown, becoming more visible to researchers and practitioners.

2. It is well documented that from 2020, AI/ML systems' pace of development has accelerated substantially[153] [154]. Innovation policy literature suggests that rapid technological change can disrupt existing innovation processes and create a 'race into the unknown', increasing uncertainties and dependencies[155] [156]. Thus, to control, as much as possible, for any unseen qualitative changes in how AI/ML systems are developed arising from this increased pace of change (and other confounding factors around this time like the proliferation of working from home in 2020 via Covid-19), we place the earliest limitation on the literature a year before this point. This is so that we can capture any literature originating in the immediate roots of this acceleration (i.e., not discounting relevant evidence from mid-late 2019, etc).

On Scopus (given its multidisciplinary variety), the search was limited to the disciplines of Computer Science, Social Science, Decision Science, Business, Management and Accounting, Arts and Humanities, and 'Multidisciplinary'. The search results were ranked by relevance, with the top 500 articles screened in each database.

[145] IAP (n 18)
[146] Jakubik, J., Vössing, M., Kühl, N., Walk, J., Satzger, G. (2024), 'Data-Centric Artificial Intelligence'.
[147] Datta, S. (2023), 'Potential Impact of Data-Centric AI on Society'.
[148] Patel et al. (n 162)
[149] Zha et al. (n 16)
[150] Brown, S. (2022), 'Why it's time for 'data-centric artificial intelligence''.
[151] Strickland, E., (2022), 'Andrew Ng: Unbiggen AI - IEEE Spectrum'.
[152] Ng, A., (2021), 'Data-Centric AI Competition'.
[153] Chow, A., Perrigo, B. (2023), 'The AI Arms Race Is On. Start Worrying'.
[154] McKendrick, J. (2021), 'AI Adoption Skyrocketed Over the Last 18 Months'.
[155] Beer, P., Mulder, R.H. (2020), 'The Effects of Technological Developments on Work and Their Implications for Continuous Vocational Education and Training: A Systematic Review'.
[156] EEA (2015), 'Global megatrends update: 4 Accelerating technological change'.

Screening

In line with best practice as outlined by Collins et al. (2015), this scoping review undertook a two-phased approach to refine the search results based on predefined inclusion and exclusion criteria, a method chosen to explore data governance across AI/ML systems. The first phase includes compiling literature from the academic databases according to their title, abstract and keywords. The second phase involved scanning the whole document from this shortlist for a deeper review of their relevance[157].

For this second phase, we thematically coded literature according to three coding frameworks relating to the lifecycle, ecosystem and governance-related observations. Each coding framework is structured according to the general stages of the AI data lifecycle to ground descriptions of the activities, ecosystem components and governance in the right context. These are shown below:

Table 2. 'Lifecycle' coding framework

Category	Stage
Planning	Li.Pl
Collection/Creation	Li.Co
Exploration/Analysis	Li.Ex
Preprocessing	Li.Pr
Training	Li.Tr
Evaluation	Li.Ev
Deployment	Li.De

[157] Collins, A., Coughlin, D., Miller, J., Kirk, S. (2015), 'The production of quick scoping reviews and rapid evidence assessments'.

Table 3. ‘Ecosystem’ coding framework

Category	Actors	Value Exchange
Planning	Ec.Ac.Pl	Ec.VE.Pl
Collection/Creation	Ec.Ac.Co	Ec.VE.Co
Exploration/Analysis	Ec.Ac.Ex	Ec.VE.Ex
Preprocessing	Ec.Ac.Pr	Ec.VE.Pr
Training	Ec.Ac.Tr	Ec.VE.Tr
Evaluation	Ec.Ac.Ev	Ec.VE.Ev
Deployment	Ec.Ac.De	Ec.VE.De

Table 4. ‘Governance’ coding framework

Category	Quality	Security	Management	Access	Ethics
Planning	Qu.Pl	Se.Pl	Ma.Pl	Ac.Pl	Et.Pl
Collection/Creation	Qu.Co	Se.Co	Ma.Co	Ac.Co	Et.Co
Exploration/Analysis	Qu.Ex	Se.Ex	Ma.Ex	Ac.Ex	Et.Ex
Preprocessing	Qu.Pr	Se.Pr	Ma.Pr	Ac.Pr	Et.Pr
Training	Qu.Tr	Se.Tr	Ma.Tr	Ac.Tr	Et.Tr
Evaluation	Qu.Ev	Se.Ev	Ma.Ev	Ac.Ev	Et.Ev
Deployment	Qu.De	Se.De	Ma.De	Ac.De	Et.De

Phase 1 - Title, abstract and keywords

Inclusion criteria

- **Discussion of data practices and processes for AI/ML systems:** Articles must address data practices and processes relating to AI/ML systems. This includes providing technical descriptions of the data lifecycle, governance considerations, and ecosystem actors and their relationships.
- **Publication timeline:** Articles must have been published between 2019 and 2024, which was enabled automatically by customising the search filters.
- **Peer-reviewed:** Articles must be from peer-reviewed journals or conference proceedings, ensuring a level of academic rigour and quality. This was enabled automatically by customising the search filters.

Exclusion criteria

- **Inversion of conceptual framework:** Special attention is paid to excluding articles that only discuss the use of AI/ML technologies in data science pipelines, inverting the conceptual framework outlined in the methodology (i.e., where the focus is instead on how data is used for the development of AI/ML systems).
- **Not a review:** Articles that are review pieces, such as book reviews or literature reviews, are excluded from consideration.

Phase 2 - Full text

Inclusion criteria

- **Acceptance of observational and experimental studies:** Observational and experimental studies are considered for inclusion in this phase.
- **Presence of useful context:** Experimental studies must contain at least an Introduction, Background, or Literature Review section that provides useful context about data science pipelines/lifecycles for AI. This ensures that selected studies offer sufficient background information to inform the scoping review.

Exclusion criteria

- **Overly technical descriptions:** Articles that consist solely of technical descriptions of specific models or systems, without providing narrative context about data governance for AI/ML systems, are excluded. This criterion ensures that selected articles offer socio-technical understanding of data practices and processes in AI contexts.
- **Use case focus:** Articles that are focused on very specific technical use cases, without providing useful narrative about data governance per se, are excluded. This criterion ensures that selected articles contribute to a broader understanding of data governance across diverse AI applications.

By adhering to these criteria in the screening process, our goal is to spotlight where there are gaps in the available evidence relating to data governance practices throughout the AI data lifecycle. The approach not only ensures relevance and academic rigour, but also aligns closely with our research's core aim: to uncover and analyse the governance practices that underpin the development and deployment of AI/ML systems.

Data sources

The primary data sources for this scoping review were Scopus and ACM Digital Library. The former was selected due to its multidisciplinarity and the latter for its crucial focus on computer science literature. Additionally, a hand search[158] (i.e., manually scanning reference lists and relevant academic journals for further literature) was conducted to supplement the search results from the databases.

[158] "Handsearching is a critical part of the review to find materials not found through traditional searches. It is a manual process to examine and identify further relevant studies and includes:
- Perusing the pages of key journals, conferences and other sources
- Checking reference lists of identified articles and documents" (James Cook University 2024)

Results

Results were ranked by relevance and the top 500 results from each academic database were screened, leading to the examination of 1,000 studies overall.

Phase 1 – Title, abstract and keywords

The initial search on Scopus yielded 1,142 results and on ACM Digital Library resulted in 168,496 results. These were reduced to 608 and 10,525, respectively, after applying the search filter to limit results to studies from between 2019 – 2024 and academic articles and conference papers.

Next, also using the search filter on both databases, the results from both were ranked by relevance and the top 500 results from each were screened (ie, 1,000 in total). From this shortlist, on Scopus 32, articles were selected based on title, abstract and keywords according to the phase 1 criteria. On ACM Digital Library, 54 articles were selected based on the screening of their title, abstract and keywords. As shown below in Figure 1., this resulted in the retaining of 86 articles overall for this phase.

Phase 2 – Full text

After scanning the full text of these 86 articles, 15 were retained from Scopus and 25 from ACM Digital Library after application of the phase 2 screening criteria listed above. After scanning the reference lists of some of these articles, and looking at relevant journals such as the Journal of Big Data and ACM Computing Surveys, the supplementary hand search uncovered 15 relevant articles.

After removing three duplicates, in total 52 articles were identified as relevant for further analysis in this scoping review. As mentioned, each of these phases are shown below in Figure 2.

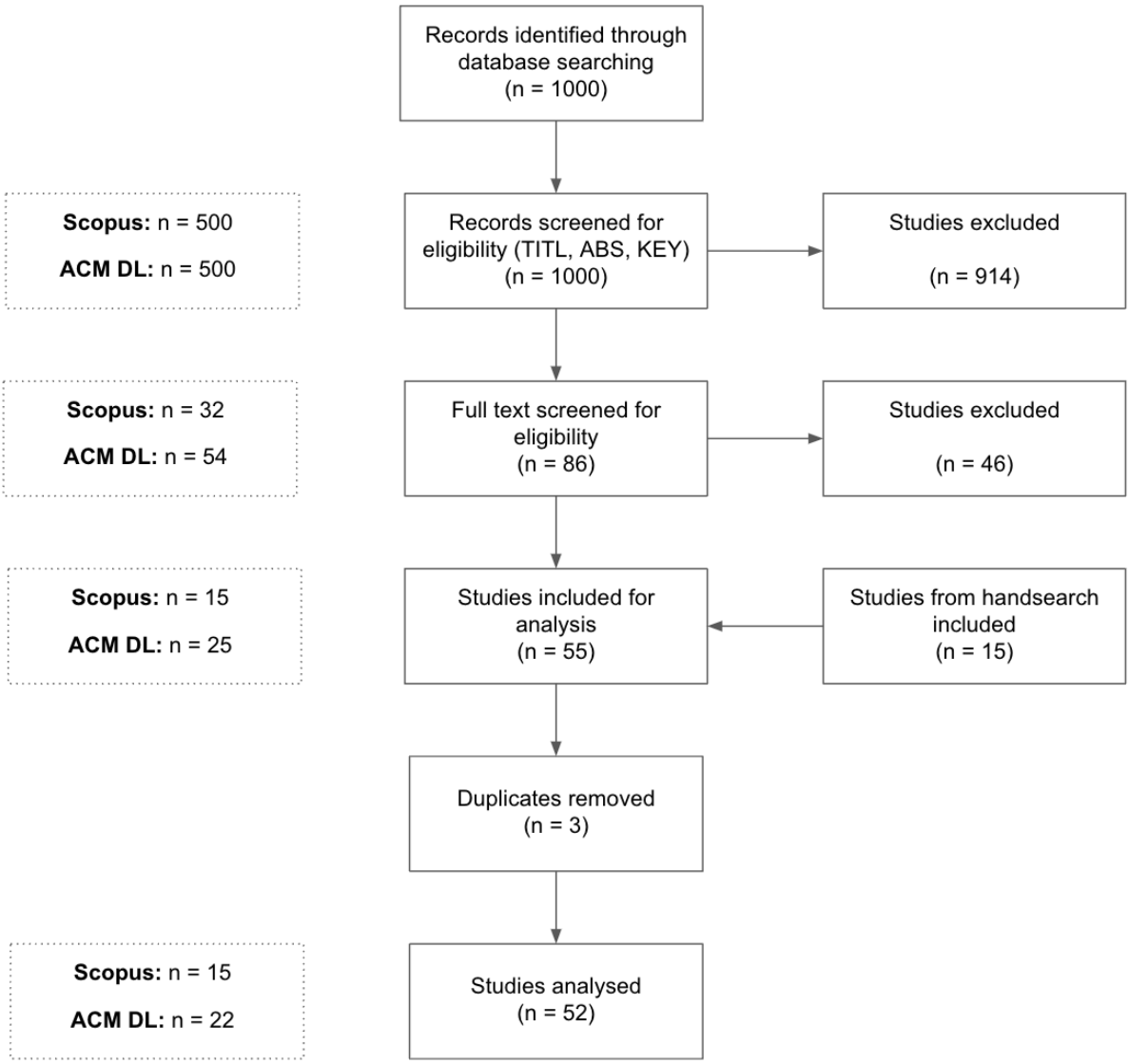

Figure 2. Diagram of scoping review results

The final selection of papers comprised both journal articles and conference papers, reflecting a diverse array of scholarly contributions.

From journals, we collected 32 articles with notable contributions such as Aldoseri et al (2023) in 'Applied Sciences' discussing the rethinking of data strategy for AI[159], Alzubaidi et al (2021) in the 'Journal of Big Data'

[159] Aldoseri et al. (n 25)

providing a review of deep learning concepts[160], and Chicco et al (2022) in 'PLOS Computational Biormalogy' offering tips for data cleaning and feature engineering[161].

Complementing the journal articles, we uncovered 20 conference papers providing insights into the dynamic and evolving nature of AI research. For instance:

- Biswas et al (2022) at the 44th International Conference on Software Engineering (ICSE '22) explores deep transfer learning, underscoring the importance of adaptable AI models[162].
- Gowda et al (2021) presented at the 30th ACM International Conference on Information & Knowledge Management (CIKM '21), which addresses the critical issue of bias in AI through data augmentation techniques[163].
- Contributions to the Proceedings of the ACM on Management of Data by Grafberger et al (2023) and Li et al (2023) delve into optimising AI data pipelines, indicating the technical strides being made in data management for AI[164] [165].
- Heger et al (2022) at the ACM Conference on Human Factors in Computing Systems (CHI) highlight the importance of data documentation, pointing to the human-centred aspects of AI development[166].

Through this diverse collection of scholarly work, our review captures the gaps in AI data governance, providing a solid foundation for future research and practice in the field.

[160] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L. (2021), 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions'.
[161] Chicco et al. (n 186)
[162] Biswas et al. (n 262)
[163] Gowda et al. (n 181)
[164] Grafberger, S., Groth, P., Schelter, S. (2023), 'Automating and Optimizing Data-Centric What-If Analyses on Native Machine Learning Pipelines'.
[165] Li et al. (n 217)
[166] Heger et al. (n 98)

Annex B. Scope and limitations

This systematic scoping review acknowledges that the field of AI is constantly evolving, and the definition of “AI/ML systems” itself encompasses a diverse range of technologies. While the review strives to capture the nuances between various ML models (for example, supervised vs. unsupervised learning, predictive vs generative AI), it does prioritise providing a high-level overview of data governance practices across the AI data lifecycle.

The simplified and linear representation of the AI data lifecycle may not fully capture the complexities and iterative nature of real-world implementations. However, this framework still serves a valuable purpose by providing clarity for understanding data governance processes within the context of AI development.