Understanding data governance in AI: A lifecycle perspective

June 2024

Data-centric AI

ODI Research
ADVANCING TRUST IN DATA

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, who 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 field of Artificial Intelligence (AI) offers the potential for transformative advancements across various sectors, fuelled by techniques like Machine Learning (ML), Natural Language Processing (NLP) and Computer Vision (CV). These technologies harness vast datasets to 'learn' patterns and create predictive models based on them, heralding an era of enhanced automation and decision-making[1] [2]. However, the performance of AI models is not just an engineering feature; the outputs from the model are fundamentally dependent on data and therefore on the rules underlying its governance[3] [4].

Data governance is defined as the structured framework of policies, processes and standards (for example, from high-level policies like data strategies to more granular documentation like datasheets for datasets) that guide the handling of data across an organisation, or between two or more[5] [6] [7] [8]. Getting these policies, processes, and standards right enables the ethical, effective development and deployment of AI/ML systems, and is becoming increasingly recognised as important, underscoring the need for robust data governance to navigate associated challenges.

[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] Olavsrud, T. (n.d.), "What is data governance? Best practices for managing data assets"
[6] Gartner (2024), "Definition of Data Governance"
[7] DGI (2020), "Defining Data Governance"
[8] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumé III, H., Crawford, K., (2021), "Datasheets for Datasets"

Concerns over data bias, algorithmic fairness, and the ethical use of sensitive information have brought greater attention to this area. Inadequate data governance frameworks not only risk embedding biases and unethical practices within the development of AI/ML systems, but can also jeopardise the trust and reliability of AI applications in critical areas such as healthcare, finance and public services[9] [10] [11] [12].

Initiatives like BigScience and BigCode have made significant strides in addressing these challenges in AI data governance, laying a foundation of best practices and project frameworks[13] [14]. Our research recognises the importance of these pioneering efforts in shaping the AI data governance landscape. The diverse range of AI/ML systems, from recommendation engines to diagnostic tools, requires nuanced governance models tailored to specific needs and contexts. By building on the foundational work of these initiatives, our study aims to further our understanding of the current state of AI data governance during system development, framing prospective future work in this space.

Despite these advancements, significant challenges persist, primarily around data governance, where current practices lack standardisation and a unified policy framework, leading to inconsistencies and challenges in data quality, explainability, transparency, fairness, safety, auditability and compliance. This situation stems chiefly from two key deficiencies: a narrow focus that extends beyond the AI model itself and the absence of a holistic AI lifecycle-based governance framework.

[9] Priestley, et al. (n 3)
[10] ODI (2023a), "Data-centric AI"
[11] ODI (2023b), "Unlocking the power of good data governance"
[12] Gerdes, A. (2022), "A participatory data-centric approach to AI Ethics by Design"
[13] 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"
[14] 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"

Traditionally, AI development has been model-centric, focusing mainly on the performance of AI models without adequately considering broader data governance issues that arise throughout the AI lifecycle. This model-centric approach often overlooks aspects like data provenance, usage, and long-term management, which are critical for ensuring that AI systems are not only effective but also fair, safe and compliant with regulatory standards[15] [16] [17] [18] [19].

Moreover, the existing data governance frameworks do not sufficiently cover the entire AI lifecycle from design through to deployment and beyond. Consequently, even if an AI model performs well technically, it might still fail in real-world applications due to poor data governance practices[20]. For example, an AI model used for diagnosing diseases might show high accuracy in controlled tests but fail in real-life scenarios if the training data did not include a diverse patient population, leading to biased or inaccurate outcomes[21].

Accordingly, through a systematic scoping review methodology, our research provides a holistic exploration of AI data governance practices, to map the gaps in our understanding of this area. In leveraging a data-centric AI perspective, we seek to answer the research question: *How is data governance implemented across the AI data lifecycle by those involved in developing AI/ML systems?*

To do this, we present a series of three reports. In the first two, we outline areas of context that are useful in structuring our account of the AI data governance landscape (the AI data lifecycle and ecosystem), while the third covers our findings on AI data governance practices. In this first report, we map out the AI data lifecycle, delving into the various stages from data collection and preprocessing to model training and deployment. By adopting a data-centric AI perspective, we emphasise the importance of high-quality, well-governed data as the cornerstone for creating ethical and effective AI/ML systems.

---

[15] Floridi, L., Chiriatti, M. (2020), "GPT-3: Its Nature, Scope, Limits, and Consequences"
[16] Zha, D., Bhat, Z.P., Lai, K.-H., Yang, F., Hu, X. (2023), "Data-centric Artificial Intelligence: Perspectives and Challenges"
[17] Polyzotis, N., Zaharia, M. (2021), "What can Data-Centric AI Learn from Data and ML Engineering?"
[18] IAP (2024), "Introduction to Data-Centric AI"
[19] van der Schaar Lab (2024), "What is Data-Centric AI?"
[20] Schneider, J., Abraham, R., Meske, C., Vom Brocke, J. (2023), "Artificial Intelligence Governance For Businesses"
[21] 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"

Our findings reveal significant gaps in the literature concerning data governance across the AI lifecycle, particularly in the areas of data access, documentation, and some ethical considerations. These gaps underscore the need for a more holistic approach to data governance that spans the entire AI lifecycle, to ensure that the right policies, processes and standards are aligned and integrated at every stage. Our research contributes to the ongoing discourse on responsible AI development, offering insights and recommendations for policymakers, practitioners and researchers. By highlighting the complexities of AI data governance, we aim to further the development of AI/ML systems that are not only technologically advanced but also ethically sound and socially beneficial.

Background

The field of Artificial Intelligence (AI) is revolutionising various aspects of our lives, ushering in an era of automation, intelligent decision-making, and advanced data analysis[22]. Techniques like Machine Learning (ML), Natural Language Processing (NLP) and Computer Vision (CV) are at the heart of this transformation. These techniques leverage vast amounts of data to identify patterns and relationships, enabling the underlying model to 'learn' from data and make increasingly accurate predictions given the introduction of new data[23] [24].

While AI has undeniable potential, there are mounting concerns regarding data bias, algorithmic fairness, and the potential misuse of sensitive information[25] [26] [27] [28] [29]. The ethical implications of AI necessitate robust governance frameworks to ensure that AI/ML systems are developed and deployed in a responsible and trustworthy manner[30] [31].

While in this research we refer to these technologies as AI/ML systems for brevity (and because a large majority of the literature discusses ML specifically), it is important to note that there is a significant degree of nuance, with differences in the use cases and models used. Reference to an 'AI/ML system' can apply to TikTok's ML-enabled personalised content recommendation engine as much as to NLP-powered chatbots, language translation, and text generation applications [32] [33]. There are notable differences between the 'model' used for producing outputs and the broader 'system' this model is deployed within.

---

[22] Misuraca, G., Van, N.C. (2020), "AI Watch - Artificial Intelligence in public services"
[23] Mitchell et al. (n 1)
[24] Jordan (n 2)
[25] Aldoseri, A., Al-Khalifa, K.N., Hamouda, A.M. (2023), "Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges"
[26] Priestley, et al. (n 3)
[27] Sher, G., Benchlouch, A. (2023), "The privacy paradox with AI"
[28] NIST (2022), "There's More to AI Bias Than Biased Data"
[29] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2021), "A Survey on Bias and Fairness in Machine Learning"
[30] IBM (2021), "IBM Documentation"
[31] Sharma, G.D., Yadav, A., Chopra, R. (2020), "Artificial intelligence and effective governance: A review, critique and research agenda"
[32] Fang, S. (2023), "TikTok – Transform Entertainment with AI"
[33] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., (2021), "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜"

An AI/ML model generally consists of input data, an algorithm[34] [35] that identifies patterns, and a classification output. This can be seen in a model that analyses chest radiology scans (data) and categorises them (via an algorithm) as either normal or abnormal (classification output)[36]. Different types of models have discriminative (i.e., telling the difference between things, as in the above radiology use case) or generative (i.e., creating new things) purposes[37] [38].

One popular example in the latter case is the GPT (Generative Pre-trained Transformer) model, which is a type of generative language model that drives the kind of NLP-powered chatbots mentioned earlier. Unlike purely discriminative models, it is designed to generate new, human-like text in response to a given prompt. The model generates this text by predicting the most likely next word, or sequence of words using the natural language patterns 'learned' from training data[39].

The performance measurements of a model offer useful but limited insight into the overall effectiveness or practical application of the full AI/ML system. The full AI/ML system encompasses the entire solution that utilises the model along with other necessary components for its real-world deployment and use. When models are introduced into live 'human' contexts, they rarely work in isolation. Numerous aspects like contextual model parameters[40] [41], labellers[42] [43], front end[44] [45], and far more, are crucial in determining an AI/ML system's overall performance and effectiveness beyond accuracy in processing input data to produce a certain output[46] [47].

---

[34] Dremio, (2024), "Neural Network Architecture | Dremio"
[35] Especially in ML, these algorithms follow a neural network architecture: "Neural networks are essentially learning algorithms, using layers of nodes (or 'neurons') to find and learn patterns in input data"
[36] Goodfellow, I., Bengio, Y., Courville, aA. (2016), "Deep Learning"
[37] ATI (2024a), "Generative models vs Discriminative models for Deep Learning"
[38] Jebara, T. (2004), "Generative Versus Discriminative Learning"
[39] Floridi et al. (n 15)
[40] Thieme, A., Morrison, C. (2022), "AI Models vs. AI Systems: Understanding Units of Performance Assessment"
[41] The parameters and settings that govern how the AI model is deployed and integrated into the broader AI/ML application or workflow, beyond just the model's standalone performance metrics
[42] AWS, (2024), "What is Data Labeling? - Data Labeling Explained - AWS"
[43] The component responsible for annotating and labelling the raw data for training the AI model
[44] Thieme et al. (n 40)
[45] The user-facing application that integrates the AI model's predictions and capabilities
[46] Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., Liu, X. (2023), "AI vs. Human -- Differentiation Analysis of Scientific Content Generation"
[47] Thieme (n 40)

This concept can be seen in many issues related to data governance. Even if an AI model has high accuracy on a specific task, poor data governance practices can undermine the overall effectiveness of the AI/ML system. For instance, if the training data used to develop the model contains biases or lacks diversity, the model may perform well on certain inputs but fail to generalise to real-world scenarios, leading to unfair or unintended outcomes[48]. Inadequate guidance relating to data provenance and monitoring practices can also prevent organisations from understanding the limitations of their AI/ML systems and making informed decisions about their deployment. In these cases, the model's technical performance does not necessarily translate to practical effectiveness, highlighting the importance of considering the entire AI/ML system and its surrounding processes[49].

Important differences affecting the shape and rules governing the development of AI/ML systems can also relate to who is developing and deploying them. Priestley et al (2023) point out that the way ML data is managed[50] [51] tends to differ between teams of academic and industry practitioners. In academic teams developing ML systems, data is normally managed in-house at the discretion of a relatively small team of researchers. Whereas in (especially larger) industry settings, systems relating to data collection, storage and processing exist across different functional areas. The latter reinforces a greater need for formal data management guidelines (for example, standards and policies) to ensure consistency across complex organisational structures[52]. In a related point, Longepre et al (2023) identified that data licensing[53] [54] information for academic/non-commercial ML datasets are generally more accurate than those of commercial ML datasets, identifying that academic practitioners drive overall ML dataset curation efforts[55].

---

[48] Mehrabi et al. (n 29)
[49] Floridi (n 4)
[50] ODI (n 11)
[51] Data management is a distinct but necessary part of broader data governance. This is described in further detail later.
[52] Priestley, et al. (n 3)
[53] Butler, J., (2024), "Content Licensing Model Types for Generative AI Training Data"
[54] Another important part of data governance; licences are legal frameworks *governing* how datasets can be accessed, used and redistributed by AI/ML companies and researchers
[55] Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., Hooker, S. (2023), "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI."

What data?

These differences, and more, are key to explaining the importance of common data governance practices for developing AI/ML systems (this point is discussed in greater detail in the next subsection). For instance, inconsistent data governance practices across business functions can result in data quality issues, leading to unreliable, and potentially unsafe, systems[56] [57]. Accordingly, the need for responsible collection, use and sharing of data underlies various AI/ML use cases and their corresponding models/systems. But before exploring data governance, it is important to clarify what we mean by data in this context.

Data can be produced as a byproduct of a deployed AI/ML system's operation. For example, log data generated by an AutoML system[58] [59] during its automated model-building process can provide valuable insights into the data, feature engineering, and model selection decisions made by the system. While not the primary input data used for training a model, this log data can be an important secondary data source for understanding the AI/ML system on an ongoing basis. Other sources of data produced by deployed AI/ML systems can include those related to ongoing monitoring of model performance, such as indicators from key performance metrics, data[60] [61] and model drift[62] [63] detection, and implementing alerting systems to identify issues[64].

[56] Priestley, et al. (n 3)
[57] Floridi (n 4)
[58] Databricks, (2024), "Databricks AutoML - Automated Machine Learning"
[59] An ML system that automates the repetitive and time-consuming aspects of the model development process, allowing data scientists and other users to focus more on the high-level problem-solving and model customisation
[60] Machiraju, S., (2021), "Why data drift detection is important and how do you automate it in 5 simple steps"
[61] Data drift occurs when the statistical properties or distribution of the input data used by a deployed model changes compared to the data the model was originally trained on
[62] Paka, A., (2020), "How to Detect Model Drift in MLOps Monitoring"
[63] Model drift, also known as concept drift, refers to changes in the underlying relationships and patterns that the model has learned, leading to a divergence between the model's predictions and the true target variable. This can happen even if the input data distribution remains stable
[64] Pykes, K. (2023), "A Guide to Monitoring Machine Learning Models in Production"

Each of these examples, and more, can be counted as data, which provides useful insights during the operation of AI/ML systems. When *developing* these systems, however, the absolutely essential data is that which is used to create the model for its required purpose; for example, the following kinds of datasets[65] [66] [67].

- **Training dataset:** The primary data used to train the ML model. The model learns patterns and relationships from this data, adjusting its internal parameters to minimise the error between its predictions and the known target outputs.
- **Validation dataset:** Separate data used to evaluate the model's performance during training. These validation data are used to tune hyperparameters and fine-tune the model parameters using the results from the validation step, preventing overfitting and ensuring the model generalises well to live scenarios.
- **Test dataset:** An independent dataset reserved for final evaluation of the trained model's performance. These test data are used to provide an unbiased estimate of the model's real-world accuracy and capabilities.
- **Fine-tuning dataset:** When deploying a pre-trained model to a new domain or application, fine-tuning data is used to further adapt and specialise the model's parameters for the new context. This usage is more common in generative AI paradigms and LLMs, which often leverage external knowledge bases to enhance their performance in specific domains[68] (Petroșanu et al, 2023).
- **Benchmark dataset:** Standard, publicly available data that are used to compare the performance of different ML models on common tasks. These benchmark data allow for objective evaluation and ranking of model capabilities.

These datasets play crucial roles throughout the AI/ML development lifecycle (i.e., the steps that the data undergoes through the development and operation of the AI/ML system), from initial model training to evaluation and improvement of AI/ML systems, even after deployment.

[65] MLCommons (2024), "Benchmark Work"
[66] Pruneski, J.A., Williams, R.J., Nwachukwu, B.U., Ramkumar, P.N., Kiapour, A.M., Martin, R.K., Karlsson, J., Pareek, A. (2022), "The development and deployment of machine learning models"
[67] Shah, T. (2017), "About Train, Validation and Test Sets in Machine Learning"
[68] Petroșanu, D.-M., Pîrjan, A., Tăbușcă, A. (2023), "Tracing the Influence of Large Language Models across the Most Impactful Scientific Works"

When we refer to ‘data’ throughout this report, we are referring to these datasets directly involved in developing and optimising the underlying models in AI/ML systems, not those relating to ancillary tasks like system logs or KPIs. We rationalise this focus because the former is fundamental for developing the core functionality of AI/ML systems (i.e., training models), and directly relevant to reinforcing or mitigating harms that may arise from active misuse or poor development[69]. Certain operational data, like the aforementioned logs, may be referenced where relevant to AI data governance during system development, but this is not the core focus of the study for this reason.

For instance, if training data does not adequately cover, or control for, the full scope of the problem the model is intended to solve, the model may perform poorly on real-world scenarios that it was not exposed to during training. This is known as a dataset with ‘class imbalances’, or an ‘imbalanced dataset’[70]. Addressing these imbalances is an integral part of any data governance. We see an example of this problem in Roberts et al (2021), who found that their deep learning model was inadequate for diagnosing Covid-19 because it was trained on a dataset containing scans of both patients who were lying down and those who were standing up. Those patients lying down were often more sick, causing the algorithm to wrongly associate risk of Covid with the person's position in the scan[71].

Furthermore, if the data used to train an AI/ML model is biased (eg, overrepresenting certain demographics or underrepresenting others), the model may exhibit biased behaviour and make unfair or discriminatory decisions[72]. For example, in 2014, Amazon began work on a nascent AI-enabled recruitment tool but, being trained on data consisting of predominantly male resumes, the system penalised phrases like ‘women’s’ and downgraded candidates from all-female colleges. Despite attempts to neutralise the bias in the tool by editing the programme running it, Amazon proved unable to eliminate all chances of bias and ended the project in 2018 due to concerns over perpetuating discrimination[73].

Even though neither of these systems trained on faulty data was deployed at scale, one can imagine the potential negative impact if they had. Thus, an essential driver of the effectiveness and responsible deployment of all AI/ML systems hinges on the handling of the data on which they are trained. This underscores the critical role of **data governance** and **data stewardship** in the AI development process[74].

---

[69] OECD (2021), "Artificial Intelligence, Machine Learning and Big Data in Finance"
[70] Zinkevich, M. (2019), "Rules of Machine Learning"
[71] Roberts et al. (n 21)
[72] Mehrabi et al. (n 29)
[73] Dastin, J. (2018), "Insight"
[74] ODI (n 10)

Data governance vs data stewardship

Data governance refers to a structured framework of policies, processes and standards that ensure the effective, ethical, and secure management of an organisation's data assets[75] [76] [77]. In this report, we will structure our analysis of governance considerations along the four 'pillars' of data governance, plus one other particularly important aspect, understood as the following[78]:

1. **Quality:** Effective data governance prioritises ensuring the accuracy, consistency and reliability of data throughout its lifecycle.
2. **Management:** Data governance sets clear guidelines for organising, storing and retrieving data, ensuring it is easily handled by authorised users without compromising its integrity.
3. **Security:** Given the frequent occurrence of data breaches, data governance is crucial for implementing strong security measures to prevent unauthorised access and breaches.
4. **Access:** Data governance balances security with accessibility, specifying who can access what data based on their roles and responsibilities.
5. **Ethics:** Data governance integrates ethical considerations, ensuring that data is used responsibly and fairly.

On the other hand, **data stewardship** refers to how data is collected, used and shared in more practical terms (i.e., not just with reference to policies, processes and standards). While data stewards are responsible for implementing and enforcing data governance initiatives, the required specific skills and training are not necessarily part of the data governance domain. There are many definitions of data stewardship, but core to each of them is the dynamic of looking after data on behalf of others[79].

Responsible data stewardship is therefore defined by the Open Data Institute (ODI) as 'an iterative, systemic process of ensuring that data is collected, used and shared for public benefit, mitigating the ways that data can produce harm, and addressing how it can redress structural inequalities.'[80].

[75] Olavsrud (n 5)
[76] Gartner (n 6)
[77] DGI (n 7)
[78] ODI (n 11)
[79] ODI (2023c), "Defining responsible data stewardship"
[80] ibid.

For instance, the organisation UK Biobank is a large-scale biomedical database and research resource that collects detailed health and genetic information from half a million UK participants[81]. This initiative illustrates the complex balance between respecting participant privacy and consent, and maximising the public utility of data for research purposes. It is important to acknowledge that while UK Biobank strives to serve as a model for data stewardship, there have been challenges and learning opportunities, including discussions around data sharing practices. These experiences provide valuable insights into the ongoing dialogue about ethical data use in research.

As described, AI/ML systems are complex and use vast quantities of data for training, validation and test purposes. While data stewardship is no doubt an important framing to understand how data is managed on a day-to-day basis, it is data governance that provides the overarching structure, policies and standards essential for efficient, ethical and secure handling of data at an organisational level. Without this governance-related structure, responsible data stewardship can become directionless, especially given the aforementioned complexity of AI/ML systems.

If prior to 2014, Amazon had developed and adhered to an organisational data governance framework outlining that the data it used to train its ML-enabled HR management system must be sourced from a better balance of male and female resumes (featuring operational guidance on, for example, who should manage this data acquisition), the identified bias may have been mitigated sufficiently for the system to be workable.

Accordingly, the objective of this work is to understand the current considerations behind, and means of defining, data governance across AI/ML systems. This work is particularly salient in the context of the emerging science of data-centric AI, which underscores the primacy of high-quality, well-governed data as the foundation for developing more accurate and equitable AI solutions.

[81] UK Biobank (2024), "UK Biobank"

Data-centric AI

In recent years, a paradigm shift has emerged within the AI landscape, with the concept of **data-centric AI** gaining traction[82] [83] [84] [85]. This approach emphasises the importance of data as the primary driver of AI development. Rather than solely focusing on optimising algorithms, data-centric AI advocates for engineering, curating, augmenting, monitoring and governing high-quality datasets to maximise model performance and generate more reliable outcomes from AI/ML systems. Data-centric AI is often situated in contrast to model-centric AI where AI/ML models are changed to fit the data[86] [87] [88] [89].

The data-centric AI approach recognises that curating data for AI/ML systems is not a one-off requirement but an ongoing process that continues throughout the system's lifecycle. Rather than considering the data as a fixed input, it involves constantly monitoring, evaluating and improving the datasets used to train and fine-tune the models. As briefly discussed earlier, this could include identifying and mitigating biases, addressing gaps or class imbalances in the data, and incorporating new, relevant data sources as they become available[90]. By emphasising the importance of data quality and curation, data-centric AI aims to produce models and wider systems that are more robust, reliable and generalisable.

This scoping review explores the existing literature on data governance practices within the context of the AI data lifecycle, thereby adopting a data-centric AI perspective. To understand these governance practices, we first document the data lifecycle and ecosystem aspects that are clear from the literature (i.e., the actors involved and the value exchanges, defined as the flows of data, information and knowledge between them[91]).

Based on this, governance practices can be expressed in a much more concrete way when related back to the structure of the lifecycle and the elements of the ecosystem, contextualising data governance to the 'how' and 'who' associated with it.

---

[82] King, J., Meinhardt, C. (2024), "White Paper Rethinking Privacy in the AI Era: Policy Provocations for a Data-Centric World"
[83] Pradhan, R. (2024), "The Paradigm Shift from Model-Centric to Data-Centric AI"
[84] Business Insider (2023), "Cleanlab Announces New Data-Centric AI Software to Reinvent Data Quality and Data Science"
[85] Iyengar, G. (2023), "Redefining Ecosystems Of Exchange: Marketplaces Powered By Data-Centric AI"
[86] Zha et al. (n 16)
[87] Polyzotis et al. (n 17)
[88] IAP (n 18)
[89] van der Schaar Lab (n 19)
[90] IAP (n 18)
[91] ODI (2022), "Mapping data ecosystems: methodology"

These aspects are valuable as structuring tools, especially when tackling such complex technologies like AI/ML systems. Ultimately, by identifying gaps and inconsistencies in current research, this review strives to illuminate areas requiring further investigation.

Existing literature on AI governance often adopts a broad perspective, ultimately focusing on various subjects beyond data governance. For instance, through interviews with practitioners, Piorkowski et al (2021) find that data scientists do not use documentation for data management consistently throughout the AI/ML system development lifecycle[92]. However, this is couched in a study of the artefacts that AI/ML developers use for communication in multidisciplinary teams, not specifically for data governance purposes. Zhang et al (2020) conduct a similar study, finding an inconsistent use of tools and related documentation practices throughout data science workflows. But again, the focus of the research is a study of how practitioners generally 'collaborate' with one another, not specifically on data governance[93].

Other research like Bellomarini et al (2021) provides an important account of the key challenge for practitioners in combining heterogeneous sources of data during the upstream data collection and creation stage, yet the overt purpose of the article is to detail a solution for these challenges called the 'Vadalog system'[94]. Similarly, Siddiqi et al (2023), early on present several challenges around the usability and scalability of data cleaning operations when pre-processing data for AI/ML systems, but these are summarised to set the stage for 'SAGA', their framework to solve these challenges[95].

These recent studies are representative of two observed challenges with the literature in this space. Either, specific data governance considerations are addressed partially in larger observational studies focused primarily on more general themes like 'communication' and 'collaboration' within teams (and, frequently not labelling these practices as relevant to data governance). Or, data governance challenges are given as key context, but in support of an experimental study solely focused on prescribing a certain approach to solve these challenges.

---

[92] 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"

[93] Zha et al. (n 16)

[94] 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"

[95] Siddiqi, S., Kern, R., Boehm, M. (2023), "SAGA: A Scalable Framework for Optimizing Data Cleaning Pipelines for Machine Learning Applications"

While such studies provide valuable insights, a more holistic description of the landscape of data governance activities, considerations and enabling ecosystem throughout the AI data lifecycle is essential to confirm where there are gaps, and therefore further work to do in building more robust data governance practices.

This scoping review split across three reports, of which this is the first, seeks to address this gap by exploring:

- **Data governance considerations** at each stage of the AI data lifecycle, encompassing planning, collection, exploration, pre-processing, training, evaluation and deployment.
- **Cross-cutting governance categories** such as data quality, security, access, management and ethical considerations.
- The existing knowledge base on the **implementation and effectiveness of data governance practices** in ensuring responsible AI development.

Structure

This work is split into three reports: two focused on enabling aspects of data governance (lifecycle and ecosystem) and one focused on governance. This is an important differentiator to previous work, as we emphasise a more holistic view of data governance practices through the lens of the AI data lifecycle and its supporting ecosystem of actors and value exchanges.

As mentioned, this report will focus on defining and explorating what the literature tells us about the AI data lifecycle. Following our introduction, this methodology section details the search strategies employed to identify relevant literature. The subsequent sections delve into the findings of the review, synthesising key themes related to data governance practices for each stage of the AI data lifecycle and across cross-cutting themes.

The discussion section critically analyses the identified gaps and inconsistencies in the existing literature. Finally, the conclusion summarises the key takeaways and underscores the importance of ongoing research in this critical domain. The reference list provides an overview of the reviewed studies.

Our contribution

This scoping review synthesises insights from 55 relevant articles identified through searches on Scopus, ACM Digital Library, and a supplementary hand search. From this, several gaps become apparent in available literature relating to data governance considerations across the AI lifecycle:

- There is little academic research on governance practices relating to AI data access between practitioners within teams (i.e., for purposes of AI/ML system development) and between practitioners and external sources (for example, ad hoc data sharing agreements, or guidance about sourcing training data from platforms like Hugging Face Hub).
- The research offers piecemeal insights into governance techniques and documentation relevant to different stages of the AI data lifecycle. However, it lacks a unified guide that connects these practices across all stages towards a singular governance objective. For instance, it explores forms of documentation that succinctly communicate the quality of data at the collection and creation[96] [97] and the exploration and analysis lifecycle stages[98]. But there is no information relating to what happens to these documents in downstream tasks and whether the same, or similar, governance tools are retained for assessing data quality throughout the lifecycle.
- Much of the literature discusses high-level tasks without specifying the types of practitioners performing them. This may be down to the fluidity or contingency of roles in an AI/ML project team, but confirmation of this speculation is itself unclear.
- Little has been published about data management considerations during downstream lifecycle stages like evaluation and deployment.

[96] Fabris, A., Messina, S., Silvello, G., Susto, G.A. (2022), "Algorithmic fairness datasets: the story so far"
[97] McMillan-Major, A., Bender, E.M., Friedman, B. (2023), "Data Statements: From Technical Concept to Community Practice"
[98] 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"

- Minimal information was found on how ethical data governance considerations are applied in practice during data collection/creation despite the importance of this for responsible data stewardship. While this does not definitively mean an absence of real-world practices, it confirms the finding of scholars such as Piorkowski et al (2021) and Zhang et al (2020) that documentation per se is applied sporadically, and produced largely as-needed for certain tasks[99] [100].

Through a detailed analysis of 52 studies, chosen from an initial collection of 1000 within the prominent academic databases of Scopus and ACM Digital Library, this review explores the literature surrounding data governance practices throughout the AI data lifecycle, aiming to:

- **Highlight the gaps in the current knowledge base:** Identifying areas where further research is needed to address existing challenges in data governance for AI development.
- **Inform the development of fit-for-purpose data governance frameworks:** Providing a foundation for establishing frameworks that facilitate responsible and ethical AI development.
- **Policy-focused discourse:** Urging researchers, practitioners and policymakers to delve into critical discussions about areas in the AI data lifecycle that are currently underrepresented with respect to data governance, with the goal of establishing best practices.

Overall, this highlights literature gaps on contextualising governance implementation throughout the lifecycle. The above points are outlined more in the third edition in this report series, which focuses on governance. We now turn to our methodology, methods and overall approach for this scoping review.

[99] Piorkowski et al. (n 92)
[100] Zhang, A.X., Muller, M., Wang, D. (2020), "How do Data Science Workers Collaborate? Roles, Workflows, and Tools"

Synthesis

Lifecycle

While we have outlined the AI data lifecycle in terms of wider AI/ML systems, as mentioned earlier, there are distinct efforts between different stakeholder groups, particularly in terms of academic vs industry efforts.

Firstly, academic AI/ML model development often focuses on advancing the state-of-the-art in algorithms and techniques, with a strong emphasis on publishing novel research in peer-reviewed journals and conferences[101]. The primary goal is to make theoretical breakthroughs that can push the boundaries of what is possible with AI/ML[102]. In contrast, industry efforts are typically more focused on building practical, deployable AI/ML systems that can deliver tangible business value, with a stronger emphasis on data, infrastructure and end-user requirements[103].

In academia, the assessment criteria for success are heavily weighted towards peer recognition, as measured by publications, citations and awards[104]. Industry projects, on the other hand, are primarily judged on their ability to generate business profits and achieve strategic objectives[105].

The people involved in academic AI/ML research are typically domain experts with deep technical backgrounds, while industry teams often have a more diverse mix of roles, including data scientists, software engineers, product managers and domain experts[106]. This can lead to differences in communication styles and priorities between the two settings[107].

Secondly, much of the literature surveyed discusses parts of the AI data lifecycle as though models are trained from scratch and datasets are collected or created from some external source prior to model training. However, as we show, a very common occurrence is the use of *pre-trained* models (as in the case of the aforementioned GPT model) that are then refined using fine-tuning datasets.

---

[101] Ashmore, R., Calinescu, R., Paterson, C. (2021), "Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges"
[102] ibid.
[103] Shrestha, S. and Das, S. (2022), "Exploring gender biases in ML and AI academic research through systematic literature review"
[104] Gerdes (n 12)
[105] Luna-Reyes, L.F., Harrison, T.M. (2023), "An Enterprise View for Artificial Intelligence Capability and Governance: A System Dynamics Approach"
[106] Schneider et al. (n 20)
[107] ibid.

This is especially advantageous for resource-constrained industry stakeholders, as pre-trained models drastically reduce the overheads associated with developing models from scratch.

Both of these points of nuance are found in the literature, and are addressed below where relevant. We now turn to a synthesis of findings from the portion of the scoping review on the AI data lifecycle.

Planning

The planning stage of the AI data lifecycle involves a set of activities and considerations to ensure responsible and effective data stewardship for AI/ML systems throughout the lifecycle[108] [109] [110] [111] [112].

1. Problem definition and stakeholder engagement

Most AI/ML systems begin with a business or academic problem to be solved. At the very early stages (and throughout the lifecycle), stakeholder collaboration and participatory activities with domain experts are emphasised to support any data-related activities by ensuring alignment with relevant domain knowledge[113]. Involving subject matter experts can help to bridge the gap between data scientists and business stakeholders, facilitating the translation of business problems into well-defined data science problems[114].

Furthermore, broader stakeholder engagement, involving employees, customers, governments and society, is also emphasised in the literature to address ethical concerns and align with stakeholder needs and expectations related to data collection and usage[115] [116]. Tensions may arise between stakeholder rights (for example, to data protection) and an organisation's intention to collect valuable data for AI models. However, effective foresight which bears these risks in mind, supports appropriate mitigation measures.

---

[108] Gerdes (n 12)
[109] Schneider et al. (n 20)
[110] Luna-Reyes et al. (n 105)
[111] Ashmore et al. (n 91)
[112] Whang, S.E., Lee, J.-G. (2020), "Data collection and quality challenges for deep learning"
[113] Gerdes (n 12)
[114] Piorkowski et al. (n 82)
[115] Schneider et al. (n 20)
[116] 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"

2. Governance and compliance

Developing an enterprise-wide view of data governance, including decision-making structures, processes and documentation practices, is important for effective data management[117], in AI contexts[118] [119] and beyond. This involves a collaborative and iterative process of designing, and making decisions on, aspects of the data lifecycle, associated risks, organisational structures for accountability, and resource allocation[120].

In this way, establishing ongoing governance mechanisms – such as an ethics committee, if within the means of the team – is considered an important potential step to mitigate challenges and raise AI potential in organisations[121] [122]. These committees can develop standard processes for decision-making, approve and monitor AI projects, and address ethical concerns related to data collection and usage[123].

Legal and regulatory compliance is also an important consideration, with efforts like explainable AI (XAI) driven by regulations like the General Data Protection Regulation (GDPR), granting individuals rights to explanations involving the way that their personal data is collected, managed and used[124] [125]. Most organisations (depending on the jurisdiction) need to ensure compliance with the necessary data protection and transparency requirements when collecting and using personal data for AI/ML systems. In addressing concerns regarding fairness and bias in AI/ML systems, it is important to highlight that the implementation of regulations such as the GDPR focuses predominantly on explainability and data protection rather than directly mandating checks for dataset balance or bias.

While it is accurate that legal and compliance teams may not specifically verify dataset imbalances, the responsibility to address such concerns can be informed by broader regulatory and ethical guidelines that indirectly influence data handling practices.

---

[117] Heger et al. (n 98)
[118] NIST (2023), "AI Risk Management Framework"
[119] Referring to the holistic organisational, business and societal environment in which an AI/ML system is being developed and deployed, beyond just the technical AI capabilities themselves
[120] Luna-Reyes et al. (n 93)
[121] Schneider et al. (n 20)
[122] Lu et al. (n 104)
[123] Lu et al. (n 104)
[124] Schneider et al. (n 20)
[125] Bibal, A., Lognoul, M., de Streel, A., Frénay, B. (2020), "Impact of Legal Requirements on Explainability in Machine Learning"

For instance, Schneider et al (2023) and Gerdes (2022) discuss the importance of documentation and transparency throughout the AI lifecycle, which includes data collection and preparation stages where bias could be addressed[126] [127].

Moreover, organisations are increasingly aware of, and are adopting, data governance frameworks that include fairness audits and the use of tools like model cards and datasheets for datasets, which provide transparency on dataset characteristics and model performance[128]. These practices help to mitigate biases by ensuring that the datasets used are well-understood and documented concerning their limitations and potential biases.

3. Data and system type

Planning for data collection methods, sources, representation formats, and data cleaning/preprocessing strategies is also crucial[129] [130] [131].
For instance, practitioners may consider whether primary and secondary data sources, a range of data collection techniques (for example, surveys, simulations, existing databases), and different data representations (for example,numerical, text, images, audio, and video) are required, particularly with respect to the aforementioned challenge[132]. The logistical feasibility of sourcing these different types of data can be quite different. A key aspect of this feasibility can be the costs associated with collection methods such as experiments, simulations, questionnaires, and observations for primary data, and official statistics, existing databases, and incident reports for secondary data[133].

[126] Schneider et al. (n 20)
[127] Gerdes (n 12)
[128] ibid.
[129] Zhou, Z., Wei, L., Yuan, J., Cui, J., Zhang, Z., Zhuo, W., Lin, D. (2023), "Construction safety management in the data-rich era: A hybrid review based upon three perspectives of nature of dataset, machine learning approach, and research topic"
[130] Singh, P. (2023), "Systematic review of data-centric approaches in artificial intelligence and machine learning"
[131] Bellomarini et al. (n 94)
[132] Zhou et al. (n 129)
[133] Whang et al. (n 112)

Planning for model management systems to track the data pipeline, transformations, and provenance is important for maintaining transparency and accountability[134] [135]. These systems are analogous to versioning in software development and support the sharing of models and data lineage information. They are especially beneficial as the AI/ML system becomes increasingly complex, reducing the feasibility of them being managed manually or with simpler tools[136].

Tools like datasheets and dataset nutrition labels are noted to thoroughly document dataset characteristics, intended use cases, and potential limitations or risks[137] [138] [139]. These documentation frameworks can aid in improving data curation practices and facilitating more informed data selection and utilisation by stakeholders[140] [141]. The need for adaptable and context-specific documentation approaches is highlighted across the literature, modified to technical aspects of the data such as its type (such as text, image, etc), whether it is static or streaming, or how broadly it will be distributed[142].

Acknowledging the distinction between AI models developed in research settings and their applications within AI/ML systems highlights a crucial aspect of AI/ML system development. Pre-trained models, originating from both public and private research teams, can bridge the gap between theoretical research and practical applications. The use of pre-trained models, such as those discussed by Besharati and Izadi (2022) and exemplified by Llama 2's comprehensive training on publicly available data sources, illustrate the efficiency these models bring to the AI data lifecycle[143] [144].

By incorporating these models, which are parameterised and enriched based on domain-specific semantics and other data-driven aspects[145], developers can leverage vast amounts of pre-processed and semantically rich data, accelerating the development process and enhancing model performance across various domains.

---

[134] Schneider et al. (n 20)
[135] Kang, D., Guibas, J., Bailis, P., Hashimoto, T., Sun, Y., Zaharia, M. (2024), "Data Management for ML-Based Analytics and Beyond"
[136] IBM (n 30)
[137] Gerdes (n 12)
[138] Heger et al. (n 98)
[139] Fabris et al. (n 96)
[140] Heger et al. (n 98)
[141] Fabris et al. (n 96)
[142] Heger et al. (n 98)
[143] Besharati, M.R., Izadi, M. (2022), "DD-KARB: data-driven compliance to quality by rule based benchmarking"
[144] Corchado, J.M., F, S.L., V, J.M.N., S, R.G., Chamoso, P. (2023), "Generative Artificial Intelligence: Fundamentals"
[145] Besharati et al. (n 143)

This not only streamlines the transition from model development to system application but also highlights the importance of ethical considerations in data source selection[146] [147]and the potential for fine-tuning to adapt models to specific organisational needs.

Another notable practice involves not just the collection of new data, but the repurposing of existing datasets to align with specific AI tasks. This approach, driven by the availability of datasets rather than the active gathering of new data, reflects a pragmatic shift towards utilising what is already accessible. For instance, Whang and Lee (2020) highlight the extensive effort and computing resources devoted to acquiring datasets containing billions of training examples, aiming to ensure model performance across diverse conditions. This is indicative of a broader trend where the challenge is less about collecting new data and more about finding and repurposing datasets that are relevant to the task at hand[148].

Similarly, Bashath et al. (2022) describe how, in the absence of labelled data, especially in contexts like healthcare where privacy and data sharing are concerns, repurposing and transfer learning from large eHR datasets within a hospital can prove beneficial for training classifiers on related tasks with limited data availability[149]. This pragmatic approach to dataset utilisation not only accelerates the development process but also mitigates the constraints associated with data privacy and the intensive labour of data collection, cleaning, and labelling[150] [151].

Lastly, the planning phase often includes the selection of benchmark datasets, which are vital for assessing model performance and sometimes generalisability in downstream tasks. However, the applicability of these datasets can vary significantly across different domains. Benchmark datasets offer standardised metrics that facilitate the comparison and enhancement of model performance broadly[152]. They are crucial for pre-training and as a baseline to track progress on defined problems.

---

[146] Fabris et al. (n 96)
[147] Schneider et al. (n 20)
[148] Whang et al. (n 112)
[149] 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"
[150] Whang et al. (n 112)
[151] Bashath et al. (n 149)
[152] Fabris et al. (n 96)

However, it's essential to consider the domain-specific relevance of these benchmark datasets. For example, while general benchmark datasets like those in the UCI ML repository might be suitable for tasks such as language understanding, they may not always be appropriate for domain-specific applications like patient image scanning, which require datasets tailored to the unique characteristics and requirements of the medical field[153] [154]. Thus, the selection of benchmark datasets should be strategically aligned with the intended application area to ensure relevance and effective performance measurement.

4. Cost and risk management

In the planning stage, it is crucial to consider the costs and feasibility of data-related activities. A study by Paleyes et al (2022) identifies high costs associated with data acquisition, preprocessing, model development, integration, and change management as a primary reason for failed AI deployments. This highlights the need for thorough feasibility analyses and cost assessments during the planning phase[155]. Separately, Whang and Lee (2020) report that data preparation tasks, including collection, cleaning, and preparation for ML, can consume 80-90% of the total project time and effort[156]. While both high costs and extensive time investment in data preparation are significant, they impact AI projects differently. The former often pertains to financial outlays across multiple stages including change management, whereas the latter primarily concerns the time and labour invested in early stages of data handling. Addressing both dimensions – cost efficiency and effective time management in data curation – appears to be important for the success of AI/ML system development projects.

Additionally, the decreasing costs of data storage have enabled the growth of data integration, analysis, visualisation tools, and computation-intensive processes available to organisations[157]. However, the planning stage still carefully evaluates the feasibility and costs associated with managing and processing large volumes of data for AI/ML systems. Techniques like active learning, where the model iteratively selects the next sample for an oracle or domain expert to clean, and the use of pre-trained models are sometimes deployed as potential cost-saving approaches during data preparation[158]. Exploring such methods to optimise data curation costs can be considered in the planning phase.

---

[153] Bashath et al. (n 149)
[154] Fabris et al. (n 96)
[155] Paleyes, A., Urma, R.-G., Lawrence, N.D. (2022), "Challenges in Deploying Machine Learning: A Survey of Case Studies"
[156] Whang et al. (n 112)
[157] Coulthart, S., Shahriar Hossain, M., Sumrall, J., Kampe, C., Vogel, K.M. (2023), "Data-Science literacy for future security and intelligence professionals"
[158] Whang et al. (n 112)

In light of cost and feasibility considerations, synthetic data generation can prove a cost-effective alternative to traditional data collection methods. Generative models, capable of producing synthetic data that closely mimics real-world phenomena, offer a scalable solution for data augmentation and imputation across various learning tasks. This can not only reduce the financial hurdles associated with gathering and annotating vast datasets but also facilitates the exploration of novel AI applications without the constraints of data scarcity[159].

Moving from planning to action, the next step is to collect and create data, essential for developing effective AI models.

Collection and creation

The collection and creation phase of the AI data lifecycle encompasses a series of activities aimed at collecting and creating both raw data and data labelled by crowdworkers for downstream analysis and model development.

1. Collection

Machine learning, particularly deep learning, relies heavily on large volumes of training data drawn from the same underlying distribution to effectively learn latent patterns[160]. However, the task of data collection is not just about compiling data; it involves navigating complex environments to discover existing data sources relevant to the purpose of the AI/ML system[161]. The synthesis also includes insights into specific methodologies and approaches within the collection phase. Practical considerations, such as dataset sizes and computational limitations for data storage, further underscore the importance of efficient data collection strategies[162].

2. Creation

This stage extends beyond the mere gathering of raw data (or, data collection); it also involves the integration of heterogeneous data from various sources, including relational databases, graph databases, and web data extraction[163].

---

[159] Watson, D.S. (2023), "On the Philosophy of Unsupervised Learning"
[160] Bashath et al. (n 149)
[161] Paleyes et al. (n 155)
[162] 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"
[163] Bellomarini et al. (n 94)

This also encompasses integration techniques, which refers to sophisticated processes that tackle the challenge of merging heterogeneous data sources, which is essential for preparing a coherent dataset that accurately represents the domain of interest[164].

From discovering and integrating datasets via schema to leveraging generative adversarial networks (GANs) for data augmentation, various strategies can be employed to enrich and diversify the available data, in essence 'creating' it as much as collecting it[165] [166]. It is important to recognise that data augmentation can also occur during the 'pre-processing' stage as data is prepared for training and evaluation. In this sense, data augmentation can be deployed in upstream or downstream tasks to iterate on existing, or create new, data.

Furthermore, advancements in technology have led to the emergence of synthetic data generation techniques. As mentioned earlier, synthetic data serves as a supplement or alternative to conventional data collection methods, reducing expenses and circumventing privacy and sensitivity constraints associated with real-world data. This approach can be instrumental in data augmentation and imputation, bolstering the robustness and diversity of datasets for supervised, reinforcement, or self-supervised learning tasks[167]. Conversely, crowdsourcing plays a crucial role in establishing ground truth for predictive AI, fine-tuning, and evaluating generative AI models, by leveraging human intelligence to label data and improve data quality[168].

In this way, techniques such as data augmentation can also address shortfalls in dataset completeness and facilitate the training of robust ML models[169]. The proliferation of digital ecosystems and the advent of citizen data science projects underscores the dynamic nature of possibilities for data creation, where users themselves contribute to the generation and curation of data[170].

Crowdsourcing has emerged as an important part in the generation and curation of data for AI/ML systems, harnessing the collective intelligence and effort of a vast number of participants to label, validate and enhance datasets. This is particularly useful in creating ground truth data[171], which is pivotal for the training, fine-tuning, and evaluation of predictive AI models.

---

[164] Whang et al. (n 112)
[165] ibid.
[166] Singh (n 130)
[167] Watson (n 159)
[168] Whang et al. (n 112)
[169] Ashmore et al. (n 101)
[170] Borrego-Díaz, J., Galán-Páez, J. (2022), "Explainable Artificial Intelligence in Data Science: From Foundational Issues Towards Socio-technical Considerations"
[171] Real-world information used as a benchmark to train and evaluate ML models

By engaging a diverse crowd to annotate data, crowdsourcing mitigates challenges associated with data labelling, ensuring the development of AI models with higher accuracy and reliability. Moreover, the utilisation of crowdsourced data in fine-tuning generative AI can reflect the importance of human input in refining AI outputs, aligning the models more closely with real-world applications and potentially ethical considerations[172].

In parallel, the collective fine-tuning of LLMs represents a novel collaboration in AI development. This process, where individuals contribute by donating prompts, reviewing outputs, and providing corrective feedback, is another community-driven approach to enhancing AI models. Such collective efforts not only improve the models' performance but also ensure they are more attuned to the nuanced demands of specific tasks and ethical standards. This practice can highlight the symbiotic relationship between human intelligence and AI capabilities, where collaborative input from a broad user base propels the refinement and applicability of AI technologies across various domains[173].

3. Integration

In contemporary AI research, the process of integrating data stands as a critical pillar for robust model development and deployment. Bellomarini et al (2022) underscore the significance of amalgamating data from disparate sources, including relational and graph databases, alongside web data extraction methodologies. This amalgamation is imperative to ensure that the data landscape mirrors real-world scenarios, thereby fostering model generalisability and efficacy[174].

Paleyes et al (2022) shed light on the challenges inherent in navigating the vast data ecosystems prevalent in organisational settings. The task of discovering and organising data sources is underscored as a formidable task, necessitating efficient integration strategies to streamline the data collection process[175]. Fundamentally, the integration of data addresses the foundational assumptions of ML elucidated by Bashath et al (2022)[176]. Ensuring that training and testing data originate from analogous distributions, coupled with the abundance of training data to capture latent patterns, forms the bedrock of effective model development.

[172] Whang et al. (n 112)
[173] Corchado et al. (n 144)
[174] Bellomarini et al. (n 94)
[175] Paleyes et al. (n 155)
[176] Bashath et al. (n 149)

Maintaining a clear record of data provenance and lineage is crucial for navigating the complex data ecosystems. Asudeh et al (2020) emphasise the significance of data investigation tools that incorporate data profiling, context and provenance. These tools are instrumental in ensuring that AI models are built upon a foundation of transparent and traceable data sources, enabling developers to effectively manage and streamline the data collection process[177].

Moreover, Borrego-Díaz and Galán-Páez (2022) outline the strategic significance of continuous data selection and curation, describing its role in shaping problem-solving approaches and solution design within the AI landscape[178]. This insight applies to the previous subsection on ‘creation’ as much as the current part on ‘integration’.

With the completion of the collection and creation phase, we shift our focus to exploring and analysing the data, where a deeper understanding is developed to refine data quality and, as per our data-centric AI perspective, further inform further model development.

Exploration and analysis

The exploration phase of the AI data lifecycle involves gaining an in-depth understanding of the dataset through various analysis techniques. This phase is critical for ensuring high quality data that leads to accurate models. The components of this part are various depending on the nature of the data and the problems addressed by the AI/ML system.

Broadly, the goals of exploration and analysis include developing an understanding of the structure and relationships of the data to guide any further collection, cleaning, preprocessing and model development[179] [180] [181]. For this purpose, visualisations and summaries often help to communicate insights from this phase[182] [183]. Documentation of data properties, assumptions, context and provenance during exploration can enable reuse[184].

---

[177] Asudeh, A., Jagadish, H.V. (2020), "Fairly evaluating and scoring items in a data set"
[178] Borrego-Díaz et al. (n 170)
[179] Siddiqi et al. (n 95)
[180] Ashmore et al. (n 101)
[181] Gowda, S.C.M., Joshi, S., Zhang, H., Ghassemi, M. (2021), "Pulling Up by the Causal Bootstraps: Causal Data Augmentation for Pre-training Debiasing"
[182] 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"
[183] Ashmore et al. (n 91)
[184] Heger et al. (n 98)

Key activities in the exploration phase to understand the data include:

- **Data profiling and summary statistics:** Calculating statistics such as maximum, minimum, mean, median, mode, variance and distributions, provides an overview of the dataset[185] [186]. Visualisations like histograms and boxplots are also used. This allows detection of initial data quality issues like outliers[187].
- **Feature analysis:** Understanding the meaning, type and number of unique values, cardinality (as examples) for each feature. Plotting and visualising different features helps identify anomalous or erroneous values[188].
- **Sampling:** Taking a subset of data by printing top/bottom samples or random sampling helps understand variety in the data. Advanced sampling techniques like stratified and cluster sampling can better capture distribution and rare variations[189].
- **Missing value analysis:** Identifying features with missing values, and imputing them if needed[190]. Patterns in missing values can also be analysed[191].
- **Multivariate analysis:** Using techniques like correlation analysis, clustering and dimensionality reduction to find interactions between features[192] [193].
- **Data validation:** Detecting inconsistent, duplicate or anomalous data points and handling outliers. May require domain knowledge of expected value ranges[194].
- **Bias detection:** Checking for population bias via coverage of demographic groups and behavioural bias (for example, proxy variables)[195].

---

[185] Bellomarini et al. (n 94)
[186] Chicco, D., Oneto, L., Tavazzi, E. (2022), "Eleven quick tips for data cleaning and feature engineering"
[187] ibid.
[188] Patel et al. (n 162)
[189] ibid.
[190] Chicco et al. (n 186)
[191] Whang et al. (n 112)
[192] Chicco et al. (n 186)
[193] De Silva, D., Alahakoon, D. (2022), "An artificial intelligence life cycle: From conception to production"
[194] Chicco et al. (n 186)
[195] Asudeh et al. (n 177)

- **Label analysis:** Understanding quality of labels, presence of noise, informativeness and potential biases[196].
- **Domain-specific analyses:** Customised analysis based on data types - for example, tabular, text, image, time-series, graph[197].

Human-centred perspectives often enable user learning and sensemaking during exploratory analysis[198] [199]. So, close collaboration between data scientists and domain experts during exploration and curation enables developing trustworthy data and data products true to the purpose of the system[200]. Manual feature engineering and selection is common, but increasingly automated tools assist in exploration[201]. Regardless, ongoing challenges to any exploration and analysis of the data include the scale of the data that requires sampling[202], streaming data where properties change[203], and evolving user knowledge affecting analysis[204].

Following the exploration and analysis, we progress to the pre-processing stage, where the insights gained are utilised to clean, transform and refine the data, preparing it for effective model training and evaluation.

Pre-processing

Pre-processing is a critical phase in the AI data lifecycle that prepares the raw data for use in model training and evaluation. It involves various techniques to clean, transform, and augment the data to improve quality and address issues that could negatively impact model performance[205] [206] [207].

In contrast to the exploration and analysis stage, in particular exploratory data analysis (EDA), which is primarily aimed at *understanding* the characteristics, patterns, and anomalies within the dataset.

---

[196] Xiong, P., Tegegn, M., Sarin, J.S., Pal, S., Rubin, J. (2023), "It Is All About Data: A Survey on the Effects of Data on Adversarial Robustness"
[197] Bellomarini et al. (n 94)
[198] Han, L., Chen, T., Demartini, G., Indulska, M., Sadiq, S., (2023), "A Data-Driven Analysis of Behaviors in Data Curation Processes"
[199] Kang et al. (n 135)
[200] Gerdes (n 12)
[201] Wang et al. (n 182)
[202] Patel et al. (n 162)
[203] Heger et al. (n 98)
[204] Kang et al. (n 135)
[205] Patel et al. (n 162)
[206] Ashmore et al. (n 101)
[207] De Silva et al. (n 193)

Instead, the pre-processing phase focuses on *preparing* the data for model training and evaluation. Specifically, while EDA provides insights and informs decisions on how to handle the data, pre-processing involves the active 'cleaning' and 'transformation' of data. In practice, there is evidence that practitioners may jump between these stages iteratively as they understand more about the data as it changes[208].

1. Data Cleaning

A key pre-processing activity is data cleaning to fix errors, remove noise and inconsistencies, and handle missing values[209] [210] [211]. Common approaches include forms of statistical methods, constraint checking, entity resolution and ML techniques. Cleaning helps assure the data accurately represents the target domain and phenomena[212]. Several key methods for this include:

- **Smoothing noisy data:** Techniques like filtering using statistical metrics, and aggregation using summary statistics, can reduce the noise present in the raw datasets that may negatively impact model training[213]. Additionally, outlier detection methods can identify anomalous or poisoned data points and discard them to avoid distorting the models[214] [215].
- **Filling missing values:** Imputation using basic statistical approaches such as mean, median, and mode can fill in missing values in the data, avoiding biases that may arise from incomplete data and improving model robustness. More advanced predictive model-based techniques can also fill missing values by estimating replacements based on correlations and patterns in the existing data from the exploration and analysis phase[216] [217] [218].

[208] Piorkowski et al. (n 92)
[209] Schneider et al. (n 20)
[210] Patel et al. (n 162)
[211] Narayan, A., Chami, I., Orr, L., Ré, C. (2022), "Can Foundation Models Wrangle Your Data?"
[212] Ashmore et al. (n 91)
[213] Chicco et al. (n 186)
[214] Whang et al. (n 112)
[215] Xiong et al. (n 196)
[216] Chicco et al. (n 186)
[217] Li, P., Chen, Z., Chu, X., Rong, K. (2023), "DiffPrep: Differentiable Data Preprocessing Pipeline Search for Learning over Tabular Data"
[218] Whang et al. (n 112)

- **Debiasing:** Removing or reweighting biased attributes that encode unfair discrimination can help reduce biases present in the training data that may propagate through the models. Oversampling underrepresented demographic groups in the data can also mitigate issues arising from imbalanced class distributions[219] [220].

- **Resolving inconsistencies:** Entity resolution techniques can identify data that references the same real-world entities but are incorrectly represented as distinct. Constraints encoding data integrity rules can also detect conflicting, inaccurate values that need resolution. Fixing these inconsistencies improves data quality[221] [222].

2. Data Transformation

Data transformation formats the data for easier consumption by algorithms. This can involve normalisation, discretisation, aggregation, dimensionality reduction, and feature extraction or engineering[223] [224]. Appropriate transformations can simplify models, reduce overfitting and improve generalisability[225].

- **Embedding:** Latent feature representations obtained through dimensionality reduction techniques or neural network-based models themselves which can capture useful structures and properties of complex raw data such as text, graphs, and images[226] [227]. Depending on the techniques deployed, this can sometimes be implemented as part of the training phase but this does use significantly more computational overhead.

[219] Asudeh et al. (n 177)
[220] Whang et al. (n 112)
[221] Asudeh et al. (n 177)
[222] Whang et al. (n 112)
[223] Patel et al. (n 162)
[224] De Silva et al. (n 179)
[225] Ashmore et al. (n 101)
[226] Murray, D.G., Šimša, J., Klimovic, A., Indyk, I. (2021), "tf.data: a machine learning data processing framework"
[227] Orr, L., Sanyal, A., Ling, X., Goel, K., Leszczynski, M. (2021), "Managing ML pipelines: feature stores and the coming wave of embedding ecosystems"

- **Labelling:** Manual human annotation, crowdsourcing, heuristic rules[228] [229], and weak supervision[230] [231] can systematically generate labels for the data to enable supervised[232] [233] training of models that require large labelled datasets[234]. High-quality labels are essential for model training or fine tuning[235].
- **Attribute selection:** Feature selection techniques eliminate redundant, irrelevant or uninformative attributes to improve model accuracy, stability and interpretability by reducing noise and focusing on the signals that are useful to the purpose of the AI/ML system[236] [237].
- **Normalisation:** Normalising features to similar scales, such as via min-max scaling, can improve model performance and stability[238] [239].
- **Generalisation:** Aggregating raw input features into higher-level representations can reduce high dimensionality in the data that may negatively impact model performance, while potentially constructing some useful summary statistics[240] [241].
- **Aggregation:** Grouping raw input examples into aggregates based on similarity or semantics provides useful summary statistics that can serve as informative features for modelling[242] [243].
- **Splitting:** Partitioning the data into complementary training, validation and test sets enables robust evaluation of model performance on holdout datasets, preventing overfitting [244] [245]. Proper splitting reduces overfitting and improves generalisation.

---

[228] Simplilearn, (2023), "Introduction To The Heuristic Function In AI | Simplilearn"
[229] "Heuristics is a method of problem-solving where the goal is to come up with a workable solution in a feasible amount of time."
[230] Stanford (2019), "Weak Supervision: A New Programming Paradigm for Machine Learning"
[231] "Weak supervision is about leveraging higher-level and/or noisier input from subject matter experts (SMEs)"
[232] Google (2024b), "What is Supervised Learning?"
[233] "Supervised learning is a category of machine learning that uses labelled datasets to train algorithms to predict outcomes and recognize patterns"
[234] Whang et al. (n 112)
[235] Yoon, H., Kim, H., (2023), "Label-Noise Robust Deep Generative Model for Semi-Supervised Learning."
[236] Chicco et al. (n 186)
[237] Zhang, R., Nie, F., Li, X., Wei, X. (2019), "Feature selection with multi-view data: A survey"
[238] Li et al. (n 217)
[239] Murray et al. (n 226)
[240] Johnson, J.M., Khoshgoftaar, T.M. (2023), "Data-Centric AI for Healthcare Fraud Detection"
[241] Kang et al. (n 135)
[242] Johnson et al. (n 240)
[243] Kang et al. (n 135)
[244] Whang et al. (n 112)
[245] Overfitting is defined as a condition where a machine learning model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data, underfitting being the opposite where too little of the data detail and noise is learned leading to over-generalisation.

Similarly to the collection and creation stage, data augmentation can also be implemented during pre-processing to artificially expand the dataset by generating new samples through transformations like rotation, flipping, cropping, and noise injection[246] [247]. As before, this addresses limited data availability and class imbalance, and improves model robustness and generalisation[248] [249].

Following sufficient data pre-processing, the training phase can commence, leveraging the refined dataset to 'teach' models on making accurate predictions and performing designated tasks. This progression from preparation to learning underscores the critical transition where raw data transforms into actionable insights.

Training

The training phase is a critical stage in the development of ML models where algorithms learn from data to perform tasks and make predictions. Yet it is not relevant for more static knowledge-based AI systems[250] [251]. This phase involves several data-related key activities.

Broadly, a key predicate to training involves selecting the type of model architecture and training algorithms to use, based on factors like problem type and relevant data characteristics[252] [253]. The prepared dataset is then used to train the chosen model and algorithms to optimise parameters and minimise errors. Training involves iterative exposure of data to the model to gradually enhance its ability to generalise and make accurate predictions[254].

The training process is complex, often involving transfer learning[255] [256] from pre-trained models and semi-supervised or active learning when labelled data is scarce[257] [258].

---

[246] Paleyes et al. (155)
[247] Ashmore et al. (n 101)
[248] Wan, Z., Wang, Zhixiang, Chung, C., Wang, Zheng (2023), "A Survey of Dataset Refinement for Problems in Computer Vision Datasets"
[249] Ashmore et al. (n 101)
[250] Russell, S., Norvig, P. (n.d.), "Artificial Intelligence - A Modern Approach (3rd Edition)"
[251] While machine learning models require a training phase to learn from data, knowledge-based systems rely more on explicitly encoded rules and domain knowledge, rather than learning from examples
[252] Ashmore et al. (n 101)
[253] Kang et al. (n 135)
[254] Coulthart et al. (n 157)
[255] Marques (2022), "Transfer Learning - an overview"
[256] “Transfer learning is a technique that utilises a trained model's knowledge to learn another set of data.”
[257] Schneider et al. (n 20)
[258] Wan et al. (n 248)

Batching methods (ie, processing the data in larger 'batches') are often used to improve efficiency, with the previously mentioned validation and test datasets eventually used to help assess model performance during training to avoid overfitting. This is described in the next section[259] [260].

1. Model types

The way that data is used during the training process varies significantly across different ML paradigms. In supervised learning, the algorithms are trained on labelled datasets where the target variable to be predicted is known beforehand[261]. Extensive data preprocessing is carried out to extract informative features that help with predicting the target variable[262]. The model is then trained through iterative exposure to the dataset to minimise errors in predicting the labels, with performance evaluated using metrics like accuracy, precision and recall to measure how well the target variable is predicted[263].

In contrast, unsupervised learning algorithms operate on unlabelled data, where the aim is to uncover hidden patterns, clusters and representations within the data distribution, rather than predicting a specific target[264]. As Watson (2023) discusses, common unsupervised tasks include clustering data points into groups, reducing data dimensionality, and building generative models that can create new data samples mimicking the training distribution[265]. Without ground truth labels for the discovered patterns, unsupervised learning models are harder to evaluate, often relying on measures of cluster or feature coherence, compression performance, or the downstream utility of extracted representations.

Reinforcement learning utilises a wholly different data paradigm, where agents learn behaviours through active interactions within an environment aimed at maximising a numerical 'reward signal'[266]. Training data is dynamically generated from the AI/ML agent's exploratory 'actions' and continuously stored rather than provided as a static predefined dataset[267].

---

[259] Kang et al. (n 135)
[260] Ashmore et al. (n 91)
[261] Niu, Y., Fan, Y., Ju, X. (2024), "Critical review on data-driven approaches for learning from accidents: Comparative analysis and future research"
[262] Biswas, S., Wardat, M., Rajan, H. (2022), "The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large"
[263] Zhou et al. (n 129)
[264] Niu et al. (n 261)
[265] Watson (n 159)
[266] Corchado et al. (n 144)
[267] Wan et al. (n 248)

The model is optimised to develop policies that maximise cumulative reward through the rounds of training. As we show next, evaluation thus focuses on the agent's resultant abilities to achieve specified goals or rewards.

As the training phase concludes, we move to the evaluation stage, where the effectiveness of models is assessed using new datasets to ensure they perform well in 'real-world' scenarios and maintain the integrity of predictions across diverse conditions.

Evaluation

The evaluation phase is a critical stage in the AI data lifecycle where models are validated on new data to assess their ability to generalise beyond the training data[268] [269] [270] [271] [272] [273] [274] [275] [276]. Key activities in this phase include:

- **Data validation and test:** Before evaluating models, the validation and test data is checked to ensure it is representative and correctly formatted[277]. This includes exploratory analysis to check for anomalies and confirm that data matches expectations.
- **Model Evaluation:** The model is tested using a designated 'holdout' validation dataset, and performance is quantified using task-specific metrics identified in relevant literature. For instance, classification tasks may utilise metrics like accuracy and F1 score, while object detection tasks could be evaluated using mean average precision[278]. It is advised to complement these quantitative evaluations with qualitative assessments conducted on real-world examples, allowing for insights from domain experts[279]. Employing multiple metrics helps to navigate trade-offs, such as precision versus recall. Additionally, visualisations and interactive tools are useful for facilitating human-in-the-loop analysis, enhancing the evaluation process.

[268] Ashmore et al. (n 101)
[269] Besharati et al. (n 143)
[270] Biswas et al. (n 262)
[271] Kang et al. (n 135)
[272] Mehrabi et al. (n 29)
[273] Schneider et al. (n 20)
[274] Singh (n 130)
[275] Wan et al. (n 248)
[276] Zha et al. (n 16)
[277] Siddiqi et al. (n 95)
[278] Biswas et al. (n 262)
[279] Ashmore et al. (n 101)

- **Generalisation testing:** Testing on new data evaluates how well the models generalise beyond the training data. The goal is assessing real-world performance and controlling for overfitting. Techniques like contextually relevant perturbations can improve generalisation testing[280] [281]. For example, geographic splits create distinct training and test regions. In this way, generalisation metrics quantify real-world performance and overfitting.

- **Human-in-loop evaluation:** Interactive evaluation enables the updating of models based on user feedback. This addresses concept drift and allows models to adapt to users' evolving understanding[282]. Humans may provide labels, confidence scores or qualitative assessments. Models are retrained or updated based on this feedback.

- **Benchmarking:** Standard benchmarks (via benchmark datasets) allow comparing models and tracking progress. Benchmarks can validate if solutions meet minimum requirements[283]. For example, General Language Understanding Evaluation (GLUE) is a popular NLP benchmark for assessing natural language understanding[284]. Benchmarks can validate if models meet minimum requirements and are suitable for an application[285]. Public leaderboards like PapersWithCode track benchmark results to show progress.

- **Monitoring and logging:** Recording evaluation metrics, model parameters and training dynamics provides traceability. This supports reproducing and diagnosing failures . Monitoring systems record evaluation metrics, model configurations, and training dynamics throughout experiments[286] [287]. This supports reproducing successes and investigating failures by providing full provenance and audit trails. Logging evaluation data enables the analysis of experiment trends over time.

Following thorough evaluation, the deployment phase transitions the tested models into real-world applications, where they are implemented, monitored and continuously improved to meet evolving data and operational demands.

[280] ibid.
[281] Schneider et al. (n 20)
[282] Kang et al. (n 135)
[283] Besharati et al. (n 143)
[284] Wang et al. (n 182)
[285] Besharati et al. (n 143)
[286] Wan et al. (n 248)
[287] Singh (n 130)

Deployment

In the deployment phase of the AI data lifecycle, various critical activities ensure the effective implementation, monitoring and maintenance of AI/ML systems. A synthesis of insights from recent literature highlights key considerations and challenges:

- **Adapting to new data:** Post-deployment, strategies for model adaptation to changing data trends are crucial. Techniques such as scheduled regular retraining and continual learning facilitate the integration of new data while addressing concept drift. Failure to detect and mitigate concept drift can significantly impact model performance[288].
- **Data monitoring and quality assurance:** Establishing rigorous monitoring systems is crucial to continually assess the quality of data being processed by AI/ML systems. This includes mechanisms to detect shifts in data distribution, known as concept drift, which can significantly affect model performance. Proactive monitoring helps in identifying when datasets need refreshing or when models require retraining to maintain their accuracy and reliability[289].
- **Data-driven model updates:** Regularly updating AI models based on new data insights is essential for adapting to changing trends and requirements. This may involve incorporating new data sources, retraining models with updated datasets, or fine-tuning models to align with current data landscapes. Such updates are critical to ensure the AI/ML system remains effective over time, necessitating a careful balance to avoid disruptions[290].
- **Ensuring data explainability and transparency:** The deployment phase must also prioritise the explainability of AI decisions, especially those based on data insights. Transparent models, offering clear explanations on how data inputs influence outputs, foster trust and understanding among users, making it easier for them to interpret AI-driven decisions[291].

[288] Paleyes et al. (155)
[289] ibid.
[290] ibid.
[291] Simbeck, K. (2023), "They shall be fair, transparent, and robust: auditing learning analytics systems"

- **Data governance and ethics review:** A thorough review of the AI/ML system's data governance practices ensures compliance with ethical standards and industry regulations. This review process evaluates how data is collected, stored and used, ensuring that the AI/ML system's deployment is responsible and respects data privacy and integrity[292].
- **Feedback loops for continuous data quality improvement:** Implementing feedback mechanisms allows for the continuous enhancement of data quality and the AI model's performance. This involves collecting user feedback, analysing system outputs, and integrating these insights into future data collection and model training processes, thereby closing the loop in the AI data lifecycle[293] [294].

Concluding remarks

Our analysis has described the key steps of the AI data lifecycle. Our research has systematically highlighted the various stages of data handling – from collection and preprocessing to deployment – and identified key areas where practices could be enhanced to support more robust AI systems. Notably, we've observed that while some stages like data collection are well-documented, other downstream stages, such as long-term system evaluation, may require further attention.

We have summarised our findings on the key steps of the AI data lifecycle through a visual resource hosted on visual workspace Miro. A preview of this resource with a link to the full visualisation is depicted below in **Figure 1.**

[292] De Silva et al. (n 193)
[293] Biswas et al. (n 262)
[294] Wang et al. (n 182)

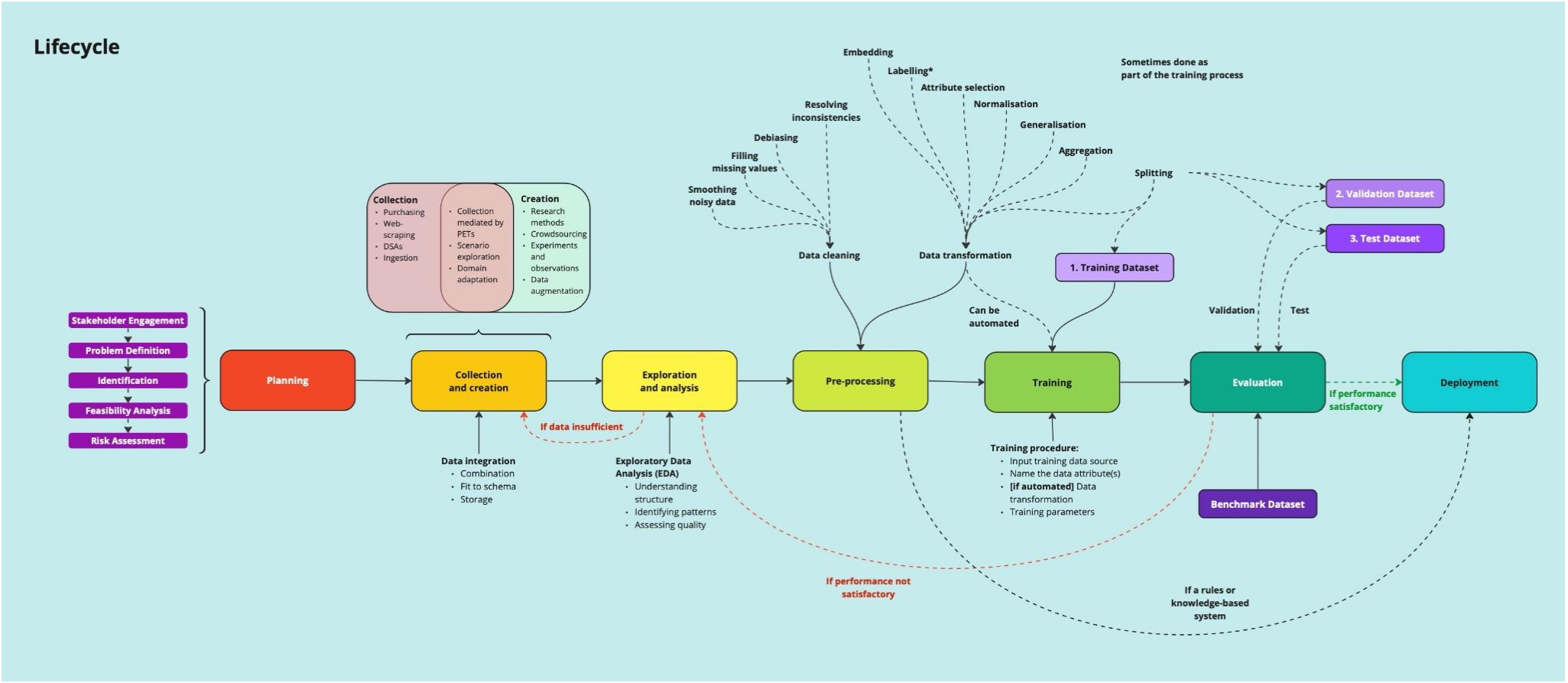

Figure 1. AI data lifecycle preview (explore the visual resource here)

Our findings reinforce a need for research that spans the entire AI data lifecycle, ensuring that AI technologies are not only advanced in their capabilities but also robust and reliable in their long-term application. This calls for ongoing efforts in education and knowledge sharing among AI developers, policymakers and the broader public to foster a well-informed community around AI operations and their impacts.

The next phase of our research will extend these insights into the broader AI data ecosystem, exploring the actors and value exchanges between them when developing AI/ML systems. This forthcoming report will aim to deepen our understanding of how data is shared and utilised within the AI landscape, further setting the stage for our final report examining the current understanding of AI data governance practices across the landscape.

For this last report, we will be providing a detailed analysis and actionable recommendations, aiming to identify areas where better data governance practices, and research on them, can underpin the effectiveness and ethical alignment of AI systems globally. As we progress in this series, our ultimate objective is to help shape an AI future that is both responsibly managed and technologically innovative.

Annex A. Methodology

A scoping review is a type of knowledge synthesis that maps the existing literature on a particular topic[295]. 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[296] [297]. 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[298] [299].

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?

[295] Mak, S., Thomas, A. (2022), "Steps for Conducting a Scoping Review"
[296] Levac, D., Colquhoun, H., O'Brien, K.K. (2010), "Scoping studies: advancing the methodology"
[297] Arksey, H., O'Malley, L. (2002), "Scoping studies: towards a methodological framework"
[298] Floridi et al. (n 15)
[299] 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 of the 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[300] [301], 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:

[300] ATI (2024b), "Knowledge graphs"
[301] 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'[302] 'Design, planning, and prototyping'[303] 'Problem Formulation'[304] 'Identify and formulate the problem' and 'Review data and AI ethics'[305] 'Planning phase'[306]
Collection and creation	'Data collection'[307] 'Data collection'[308] 'Data discovery'[309] 'External data acquisition'[310] 'Collection phase'[311] 'Data Ingestion'[312]
Exploration and analysis	'Data analysis and pre-processing'[313] 'Error detection and repair'[314] 'Data preparation' and 'Data exploration'[315] 'Preparation phase' and 'Analysis phase'[316] 'Data preparation & cleansing'[317]
Pre-processing	'Data analysis and pre-processing'[318] 'Data cleaning and Data annotation'[319] 'Data preparation - Data cleaning and Data labeling'[320] 'Data pre-processing", "Data augmentation'[321] 'Preparation phase'[322]

[302] ICO (2024), "Annex A: Fairness in the AI lifecycle"
[303] Chai, C., Wang, J., Luo, Y., Niu, Z., Li, G. (2023), "Data Management for Machine Learning: A Survey"
[304] Khan, M.J., Breslin, J.G., Curry, E. (2023), "Towards Fairness in Multimodal Scene Graph Generation: Mitigating Biases in Datasets, Knowledge Sources and Models"
[305] De Silva et al. (n 193)
[306] Shah, S.I.H., Peristeras, V., Magnisalis, I. (2021), "DaLiF: a data lifecycle framework for data-driven governments"
[307] ICO (n 302)
[308] Khan et al. (n 304)
[309] Chai et al. (n 303)
[310] De Silva et al. (n 193)
[311] Shah et al. (n 306)
[312] Cognilytica, (2020), "AI Data Engineering Lifecycle Checklist"
[313] ICO (n 302)
[314] Chai et al. (n 303)
[315] De Silva et al. (n 193)
[316] Shah et al. (n 286)
[317] Cognilytica (n 312)
[318] ICO (n 302)
[319] Khan et al. (n 304)
[320] Chai et al. (n 303)
[321] De Silva et al. (n 193)
[322] Shah et al. (n 306)

	‘Data Preparation & Cleansing’, ‘Data Transformation’[323]
Training	‘Model development‘[324] ‘Model training and evaluation‘[325] ‘Model training & inference – In-ML‘[326] ‘Build initial AI model‘, ‘Data augmentation‘, and ‘Build multiple AI models‘[327]
Evaluation	‘Model evaluation‘[328] ‘Model evaluation‘[329] ‘Model management – model storage, versioning, query, and diagnosis‘[330] ‘Evaluate primary metrics‘ and ‘Evaluate secondary metrics‘[331] ‘Use, re(use), and feedback‘[332]
Deployment	‘Model deployment and monitoring‘[333] ‘Model deployment and inference‘[334] ‘Model deployment and serving‘[335] ‘AI model deployment and risk assessment‘, ‘Post-deployment review’[336] ‘Use, re(use), and feedback‘[337] ‘Data and production orchestration‘[338]

[323] Cognilytica (n 312)
[324] ICO (n 302)
[325] Khan et al. (n 304)
[326] Chai et al. (n 303)
[327] De Silva et al. (n 193)
[328] ICO (n 302)
[329] Khan et al. (n 304)
[330] Chai et al. (n 303)
[331] De Silva et al. (n 193)
[332] Shah et al. (n 306)
[333] ICO (n 302)
[334] Khan et al. (n 304)
[335] Chai et al. (n 303)
[336] De Silva et al. (n 193)
[337] Shah et al. (n 306)
[338] 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. **Pre-processing:** 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 pre-processing 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[339] [340] [341] [342] [343] [344] [345] [346]. 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.

[339] IAP (n 18)
[340] Jakubik, J., Vössing, M., Kühl, N., Walk, J., Satzger, G. (2024), "Data-Centric Artificial Intelligence"
[341] Datta, S. (2023), "Potential Impact of Data-Centric AI on Society"
[342] Patel et al. (n 162)
[343] Zha et al. (n 16)
[344] Brown, S. (2022), "Why it’s time for ‘data-centric artificial intelligence’"
[345] Strickland, E., (2022), “Andrew Ng: Unbiggen AI - IEEE Spectrum”
[346] Ng, A., (2021), “Data-Centric AI Competition”

2. It is well documented that from 2020, AI/ML systems' pace of development has accelerated substantially[347] [348]. Innovation policy literature suggests that rapid technological change can disrupt existing innovation processes and create a 'race into the unknown', increasing uncertainties and dependencies[349] [350]. 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.

[347] Chow, A., Perrigo, B. (2023), "The AI Arms Race Is On. Start Worrying"
[348] McKendrick, J. (2021), "AI Adoption Skyrocketed Over the Last 18 Months"
[349] 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"
[350] 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[351].

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
Pre-Processing	Li.Pr
Training	Li.Tr
Evaluation	Li.Ev
Deployment	Li.De

[351] 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
Pre-Processing	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
Pre-Processing	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[352] (i.e., manually scanning reference lists and relevant academic journals for further literature) was conducted to supplement the search results from the databases.

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.

[352] "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)

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**.

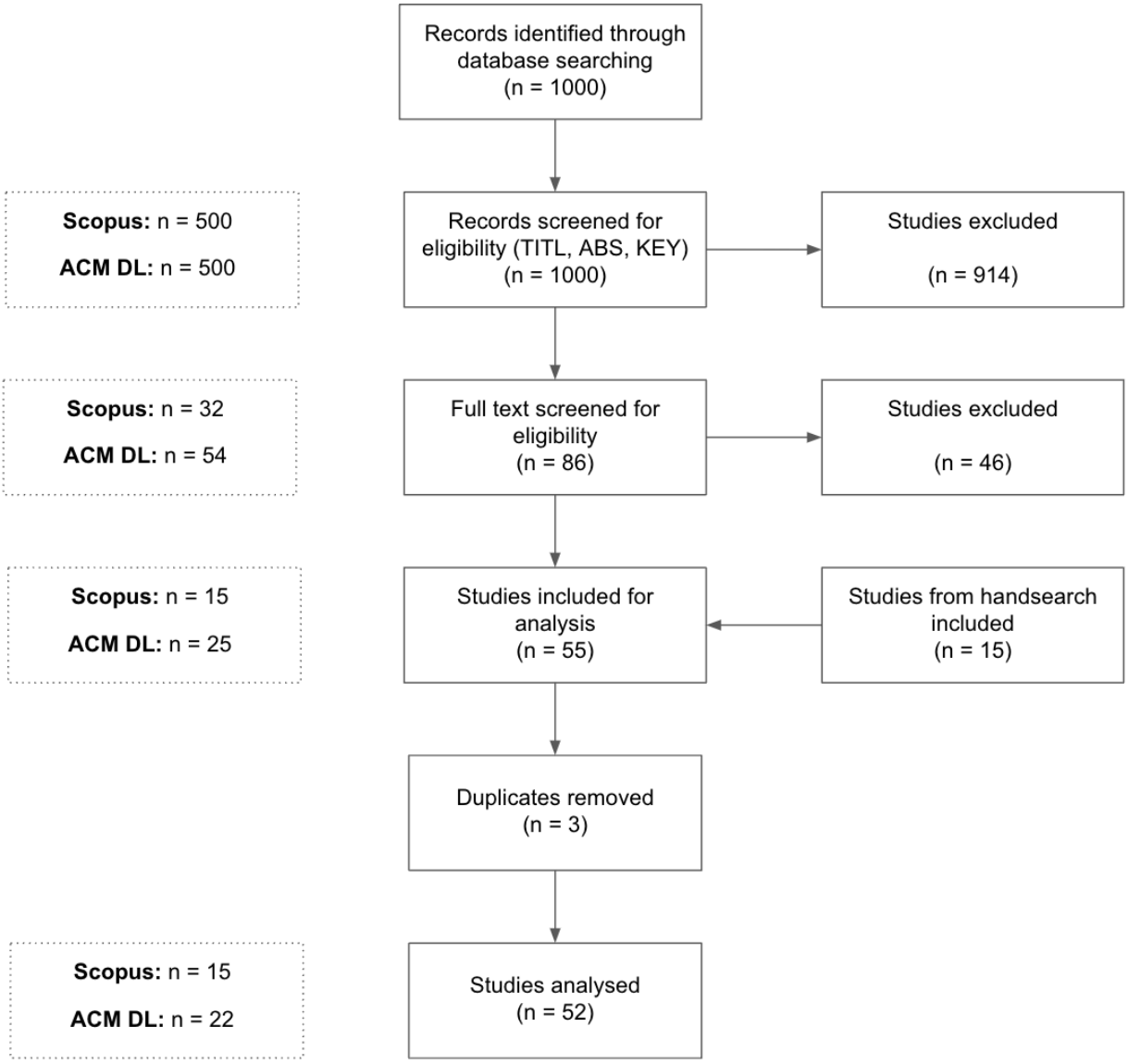

Figure 2. Diagram of scoping review results

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[353], Alzubaidi et al (2021) in the 'Journal of Big Data' providing a review of deep learning concepts[354], and Chicco et al (2022) in 'PLOS Computational Biormalogy' offering tips for data cleaning and feature engineering[355].

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[356].
- 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[357].
- 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[358] [359].
- 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[360].

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.

---

[353] Aldoseri et al. (n 25)
[354] 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"
[355] Chicco et al. (n 186)
[356] Biswas et al. (n 262)
[357] Gowda et al. (n 181)
[358] Grafberger, S., Groth, P., Schelter, S. (2023), "Automating and Optimizing Data-Centric What-If Analyses on Native Machine Learning Pipelines"
[359] Li et al. (n 217)
[360] 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 (eg, 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.