Understanding data governance in AI: Mapping governance

Data-centric AI

ODI Research
ADVANCING TRUST IN DATA

June 2024

Contents

About

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

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

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

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

Introduction

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

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

---

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

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

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

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

Through this final report in our series, we aim to spotlight the gaps in literature on governance practices. Building on the insights from our previous reports on the AI data lifecycle and the AI ecosystem, we set an agenda for future research and policymaking. We examine what the literature says about AI data governance practices, highlighting both positive insights and existing gaps. By doing so, we provide an overview of how data governance practices can be optimised to meet the dynamic needs of AI/ML systems in a rapidly evolving digital landscape. Ultimately, we aim to support the development of AI systems that are not only technologically advanced but also ethically sound and socially beneficial, ensuring they can be reliably deployed in diverse real-world applications.

---

[10] Schneider, J., Abraham, R., Meske, C., Vom Brocke, J. (2022), 'Artificial Intelligence Governance For Businesses'.
[11] Roberts, M., et al. (2021), 'Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans'.
[12] Olavsrud, T. (n.d.), 'What is data governance? Best practices for managing data assets'.
[13] Gartner (2024), 'Definition of Data Governance'.
[14] DGI (2020), 'Defining Data Governance'.
[15] Gebru, T., et al. (2021), 'Datasheets for Datasets'.
[16] Priestley, et al. (n 3)
[17] ODI (2023a), 'Data-centric AI'.
[18] ODI (2023b), 'Unlocking the power of good data governance'.
[19] Gerdes, A. (2021), 'A participatory data-centric approach to AI Ethics by Design'.
[20] Jernite, Y., et al. (2022), 'Data Governance in the Age of Large-Scale Data-Driven Language Technology'.
[21] Hughes, S., et al. (2023), 'The BigCode Project Governance Card'.

Synthesis

Governance

In the rapidly advancing fields of AI and ML, data governance is essential for ensuring the ethical, effective and secure management of data. As explained in our first report, effective data governance frameworks are built upon five critical pillars: quality, management, security, access and ethics. Each pillar plays a vital role in maintaining the integrity and utility of data throughout its lifecycle.

1. **Quality:** This pillar emphasises the importance of maintaining high data quality to ensure model effectiveness and reliability. It involves metrics and methods for error identification and correction, ensuring balanced representation, and continuous validation processes during training and deployment to uphold data integrity and mitigate biases.

2. **Management:** Data management involves clear guidelines for organising, storing and retrieving data to ensure consistency and accessibility without compromising its integrity. Effective management includes efficient resource allocation, query optimisation, and standardisation of formats.

3. **Security:** Ensuring data security is paramount in protecting against unauthorised access and breaches. This pillar includes implementing strong security measures such as encryption, input validation, and continuous monitoring.

4. **Access:** Balancing data accessibility with stringent security and privacy requirements is critical. Data governance frameworks specify access controls based on user roles and responsibilities.

5. **Ethics:** Integrating ethical considerations into data governance ensures that data is used responsibly and fairly. This includes addressing biases, ensuring transparency, and maintaining accountability throughout the AI data lifecycle.

These pillars are crucial for structuring our findings as they provide an accurate framework for understanding and evaluating data governance practices across the AI data lifecycle. However, it should be noted that these are not always static categories and that there can be some overlap between them. In fact, it is desirable to ensure that they complement each other, and many of these grey areas will be addressed in our findings.

By focusing on quality, we ensure that data used in AI models is accurate and reliable, directly impacting model performance and fairness. Next, management

practices streamline data handling and accessibility, enhancing efficiency and consistency. Furthermore, security measures protect against data breaches and unauthorised access, maintaining trust and compliance with regulations. Also, a consideration of access controls balance the need for data availability with privacy concerns, ensuring ethical usage. Lastly, the integration of ethical considerations ensures that AI/ML systems are developed and deployed responsibly, addressing biases and promoting transparency.

Together, these pillars form a holistic approach to robust data governance, which is essential for the ethical and effective use of AI technologies. Now we turn to our AI data governance findings relating to the first category: quality.

Quality

Ensuring data quality is crucial for AI and ML, impacting model effectiveness and reliability. This section addresses key themes including metrics and methods for error identification and correction, fairness through balanced representation, compliance with regulations such as GDPR, cost management strategies, and integrating diverse data sources. Additionally, permeating through each of these, we highlight the importance of thorough documentation, tools and techniques for data exploration and preprocessing, and continuous validation processes during training and deployment to maintain data integrity and mitigate biases.

Metrics and frameworks

Quality assessments use metrics to identify flaws and risks in data. Once identified, quality improvement methods include leveraging expert auditing, collective intelligence, and algorithms to correct errors. Assessing data quality and identifying potential issues for resolution using these techniques is key for successful model training. For instance, some common data quality criteria included in these metrics include statistical representativeness, lack of bias, and fair demographic distributions[22].

[22] Zha et al. (n 6).

Methods to improve fairness include balancing underrepresented subgroups in training data. Consequently, constructs used as input/output variables (that is, features and targets/labels) must be regularly assured as valid and representative, especially if the data is continuously ingested and/or updated. Crucial considerations like efficient resource allocation and query optimisation can enable rapid data retrieval and iteration for ongoing data quality assurance, which is especially important in downstream tasks like training and evaluation[23].

Given this, effective governance frameworks are essential for defining how the above metrics are deployed, as they help to maintain data quality throughout the AI data lifecycle. These frameworks encompass personnel, technology and processes to ensure compliance, consistency, and reliability in data quality governance practices[24].

Supporting these frameworks, several libraries and tools have been developed to measure data quality, including Deequ, DQLearn, Pandas Profiling, TensorFlow Validation, TDDA and Great Expectations. While these tools address basic data quality challenges like missing values and outliers, they are limited in their functionalities for EDA and any other data exploration and analysis tasks[25].

Compliance and data sensitivity

One of the fundamental aspects of data quality governance is addressing the sensitivity and regulatory requirements associated with the data being used. Personal data, which relates to individuals, is subject to various regulations, such as the General Data Protection Regulation (GDPR) enacted in the European Union. Practitioners must ensure compliance with these regulations (where applicable according to the jurisdiction) to avoid potential legal and financial consequences, and to maintain the trust of their stakeholders[26].

[23] ibid.
[24] Singh, P. (2023), 'Systematic review of data-centric approaches in artificial intelligence and machine learning'.
[25] Li, P., Chen, Z., Chu, X., Rong, K. (2023), 'DiffPrep: Differentiable Data Preprocessing Pipeline Search for Learning over Tabular Data'.
[26] Schneider et al. (n 10).

Cost management and feasibility

Developing and maintaining high-quality datasets and AI models can require significant time and computational resources, incurring high costs. Practitioners can explore strategies for managing these via governance measures such as active learning, data sampling and subset selection, to alleviate these challenges and optimise the utilisation of resources while maintaining data quality standards. These strategies can then be embedded in planning documentation for the AI/ML development project[27] [28].

Data quality in collection

Integration of diverse data sources and entity resolution[29] [30] processes are crucial for maintaining data integrity[31] [32]. The era of big data introduces new challenges for data governance, necessitating holistic approaches to address issues relating to the velocity, variety and veracity of vast datasets (whether training, validation, test or other data), particularly those integrated from multiple sources[33].

In these datasets, biased or incomplete data can compromise AI models' integrity, highlighting the importance of addressing missing data and mitigating biases during upstream tasks (for example, at the point of collection and creation) as much as possible[34]. These data from diverse sources, such as electronic health records (eHR), offer rich insights but pose challenges due to privacy concerns and the difficulty in obtaining labelled data. Innovative strategies, like training classifiers within hospitals on large eHR datasets and transferring learned information externally, present opportunities to leverage limited sensitive data effectively[35].

---

[27] Wan, Z., Wang, Zhixiang, Chung, C., Wang, Zheng (2023), 'A Survey of Dataset Refinement for Problems in Computer Vision Datasets'.
[28] Hacker, P., Naumann, F., Friedrich, T., Grundmann, S., Lehmann, A., Zech, H., (2022), 'AI Compliance – Challenges of Bridging Data Science and Law'.
[29] Goyal, S., (2021), 'An introduction to Entity Resolution — needs and challenges'.
[30] Entity resolution facilitates the identification and merging of entities across disparate datasets.
[31] Bellomarini, L., et al. (2021), 'Data science with Vadalog: Knowledge Graphs with machine learning and reasoning in practice'.
[32] Paleyes, A., Urma, R.-G., Lawrence, N.D. (2022), 'Challenges in Deploying Machine Learning: A Survey of Case Studies'.
[33] Borrego-Díaz, J., Galán-Páez, J. (2022), 'Explainable Artificial Intelligence in Data Science: From Foundational Issues Towards Socio-technical Considerations'.
[34] Aldoseri, A., Al-Khalifa, K.N., Hamouda, A.M. (2023), 'Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges'.
[35] Bashath, S., et al. (2021), 'A data-centric review of deep transfer learning with applications to text data'.

Detailed documentation is a cornerstone of data quality, providing transparency and facilitating users to make informed decisions[36]. Documentation can encompass data collection methodologies, annotation procedures and dataset provenance to empower stakeholders in navigating the data landscape effectively. Thorough documentation and transparency on data collection, preprocessing and potential biases can enable efficiency and mitigate harm[37].

Data quality in exploration/analysis and preprocessing

Various tools enable exploring data and becoming aware of potential issues during this lifecycle phase, rather than hiding outliers[38]. For this purpose, participatory techniques found in resources like holistic quality toolkits better support quality assessment and data-centric AI beyond basic analysis[39] [40] [41]. These methods can assist with designing expert knowledge-informed rules for handling data, alongside technical solutions such as leveraging test suites[42] [43], which also help identify errors and biases[44]. Specifically, identifying and reducing duplicated data, irrelevant features, or over-detailed representations improves efficiency and quality[45].

Clustering methods[46] [47] detect value similarity, learning-based techniques can identify patterns, and techniques like data reuse, skipping[48] [49] and

---

[36] Fabris, A., Messina, S., Silvello, G., Susto, G.A. (2022), 'Algorithmic fairness datasets: the story so far'.
[37] McMillan-Major, A., Bender, E.M., Friedman, B. (2024), 'Data Statements: From Technical Concept to Community Practice'.
[38] Han, L., Chen, T., Demartini, G., Indulska, M., Sadiq, S., (2023), 'A Data-Driven Analysis of Behaviors in Data Curation Processes'.
[39] 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'.
[40] QMU, (2020), 'QMU Working Paper Series 2020/3'.
[41] A holistic quality toolkit is a set of tools and frameworks that enables a comprehensive assessment and improvement of quality across all stages and aspects of a project or process. It goes beyond basic analysis by facilitating dialogue, reflection and shared understanding among stakeholders to define and uphold quality standards throughout the entire lifecycle.
[42] Lambdatest, (2023), 'Understanding Test Suite & Test Case: Examples and Best Practices'.
[43] A test suite is a collection of test cases that are intended to test and validate the functionality, performance, and behaviour of a software application or system.
[44] Kang, D., et al. (2024), 'Data Management for ML-Based Analytics and Beyond'.
[45] Wang, D., et al. (2019), 'Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI'.
[46] Sarker, I.H., (2021), 'Machine Learning: Algorithms, Real-World Applications and Research Directions'.
[47] Unsupervised ML technique that groups a set of data points into clusters based on similarity, with the goal of maximising intra-cluster similarity and minimising inter-cluster similarity.
[48] Ahmadian, S., Swamy, C., (2016), 'Approximation Algorithms for Clustering Problems with Lower Bounds and Outliers'.
[49] Skipping is a technique used in data processing to selectively ignore or skip over certain data points or elements, often based on predefined criteria or conditions, to reduce computational overhead or improve efficiency.

approximation[50] [51] leverage redundancy. Practitioners can consider embedding terms relating to a reduction of data volume, concerning images, text and other data types, to make governance processes more efficient. Across these considerations, it is key to select features carefully, both by considering domain knowledge and by avoiding overfitting or having too many dimensions[52].

Imputation techniques[53] [54] are appropriate for handling missing values in the data, thus improving quality. To achieve this, statistical methods such as mean, median or mode imputation, as well as more advanced techniques like imputation using regression and k-nearest neighbours[55] [56], are commonly used. Machine learning methods, including libraries like DataWig, offer additional imputation capabilities, which are particularly beneficial for complex datasets. These tools and techniques are all highlighted as important techniques for governing data quality[57].

Standardising units of measurement, date formats and non-numerical data types is essential in data formatting and is also considered during formulation of data governance measures. While outliers are typically removed from datasets, retaining and labelling them may be necessary, depending on the problem description. Data quality, in this context, is defined by the data's fitness for purpose, with frameworks and standards (such as ISO 8000-1:2022) proposed to ensure data quality integrity[58].

Continuous validation of datasets is imperative, to prevent data errors from impacting the overall data quality and downstream performance of AI models. As mentioned, this is especially the case for data that is ingested over time and/or continuously ingested some other way. Incorporating data validation routines into the AI/ML data pipeline helps to catch and address data issues early, mitigating potential problems downstream[59].

---

[50] Chakrabarty, D., Negahbani, M., Sarkar, A., (2022), 'Approximation Algorithms for Continuous Clustering and Facility Location Problems'.
[51] Approximation is the process of finding a solution that is close to the optimal solution, but not necessarily exact, often trading accuracy for computational efficiency or tractability.
[52] Niu, Y., Fan, Y., Ju, X. (2023), 'Critical review on data-driven approaches for learning from accidents: Comparative analysis and future research'.
[53] Jäger, S., Allhorn, A., Bießmann, F., (2021), 'A Benchmark for Data Imputation Methods'.
[54] The process of estimating and filling in missing values in datasets, which is crucial for enabling effective machine learning on incomplete data.
[55] Brownlee, J., (2020), 'kNN Imputation for Missing Values in Machine Learning'.
[56] A technique for estimating and filling in missing values in a dataset by finding the k nearest neighbour data points that have the missing feature value present.
[57] De Silva, D., Alahakoon, D. (2022), 'An artificial intelligence life cycle: From conception to production'.
[58] ibid.
[59] Paleyes et al. (n 32).

Data quality in training and deployment

As Aldoseri et al. (2023) discuss, key data quality dimensions like accuracy, completeness, consistency and relevance need to be ensured[60]. Flaws in training data quality can lead to biased or unreliable model outcomes that lack utility for the problem addressed by the AI/ML system[61] [62].

A key challenge is that real-world training data often contains errors, noise and missing values, even after preprocessing and cleaning[63]. Data governance measures are needed to monitor data quality drift over time if new training data gets added. Automated validation procedures and statistical testing can help detect such deficiencies. If data issues are identified, the training data may need to be re-cleaned or augmented through techniques like imputation of missing values[64].

The criteria used to label or categorise data points can itself be fuzzy or controversial. For example, the definition of 'disadvantaged individuals' may vary based on context. Rigorous governance of data validation and curation processes is required to prevent the introduction of biases or harms[65]. Moreover, the appropriateness of applying supervised learning algorithms to data depends on whether sensible labels can actually be obtained, especially in self-supervised models. Ideally, data governance strategies evaluate if suitable labelling rules or functions exist to allow supervised training; alternatively, unsupervised or semi-supervised methods may be preferable[66].

Data integrity checks play a vital role in maintaining data quality during deployment. These involve verifying updates to input tables[67] [68], ensuring checksums[69] [70] are conducted on historical records, and tracking anomalies to detect deviations from expected patterns[71].

---

[60] Aldoseri et al. (n 34).
[61] ibid.
[62] Whang, S.E., Lee, J.-G. (2020), 'Data collection and quality challenges for deep learning'.
[63] ibid.
[64] Aldoseri et al. (n 34).
[65] Gerdes (n 19).
[66] Bashath et al. (n 35).
[67] Sánchez, J., (2022), 'Introduction to AI Tables'.
[68] Input tables are a tabular data structure that serves as input to a ML model or AI system.
[69] Herson, M., Bolanos, Y., Pandey, N., (2022), 'Building scalable checksums'
[70] Checksums are a type of redundancy check used to detect errors or changes in data during transmission or storage, checksums can be used to verify the integrity of input data tables.
[71] Paleyes et al. (n 32)

Security

Ensuring the security and privacy of data in AI and ML is important to protect against adversarial attacks[72] [73] and to maintain model and system integrity. This section covers key areas such as adversarial defences targeting various stages of the AI data lifecycle, including data preprocessing and training, to enhance model robustness. It addresses the risks posed by data poisoning, where malicious samples compromise datasets, and the importance of using privacy protection measures like differential privacy and encryption to prevent unauthorised access and data breaches.

Data security essentially requires strategies such as secure data storage, input validation, adversarial training, and continuous monitoring to detect and mitigate threats. Privacy-preserving techniques and robust anonymisation safeguard individual identities while maintaining data utility. We also emphasise ongoing adaptation of security measures to address evolving threats and maintain the trust and reliability of AI/ML systems.

Adversarial attacks and defences

Defences against adversarial attacks can target various stages of the AI/ML data lifecycle, including data preprocessing and training (particularly during any regularisation techniques[74] [75] deployed during the latter)[76]. Data preprocessing defences focus on selecting robust features or transforming data representations, while training involves continually augmenting the training data with adversarial examples to improve model robustness.

Data poisoning involves the deliberate injection of malicious samples into training datasets to compromise the performance and reliability of AI/ML models. The proliferation of data poisoning and adversarial attacks poses significant threats to the integrity and security of AI systems, particularly during the data collection phase[77] [78].

---

[72] Dremio, (n.d.), 'Adversarial Attacks in AI'

[73] Adversarial attacks are a technique used to deceive or manipulate ML models by introducing carefully crafted perturbations or modifications to the input data, causing the model to make incorrect predictions or classifications.

[74] Simplilearn, (2024), 'The Best Guide to Regularization in Machine Learning | Simplilearn'

[75] Regularisation techniques are methods used to prevent overfitting of models to the training data, thereby improving their ability to generalise well to new, unseen data.

[76] Xiong, P., et al. (2024), 'It Is All About Data: A Survey on the Effects of Data on Adversarial Robustness'.

[77] Aldoseri et al. (n 34).

[78] Whang et al. (n 62).

Data protection and privacy

Privacy protection measures can be integrated into data collection processes to mitigate risks of unauthorised access, data breaches, or misuse of sensitive information. Through prioritising privacy protection and obtaining informed consent, practitioners can foster trust among data subjects and demonstrate a commitment to ethical AI practices[79].

As mentioned, the collection of personal data for AI applications must adhere to myriad regulations aimed at safeguarding individuals' privacy rights and ensuring fair and transparent data practices[80]. Legislation such as the GDPR in the European Union, the California Consumer Privacy Act (CCPA) in the United States, and analogous laws in other countries, establishes stringent requirements for the lawful collection, processing and storage of personal data[81].

Mitigating re-identification risks requires the adoption of robust anonymisation techniques, which can include data obfuscation[82] [83] and aggregation, to protect individuals' identities while preserving the utility of the data for AI/ML systems[84]. Technical solutions, such as privacy-enhancing technologies (PETs)[85] [86] like differential privacy (DP)[87] [88], can address the challenges of maintaining individual data privacy while learning potentially useful information about populations through privacy-preserving data analysis. Techniques such as anonymisation, perturbation, and cryptographic and distributed protocols are sometimes employed to mitigate privacy risks. Terms defining their proper use can therefore be included in data governance measures[89].

[79] Gerdes (n 19).
[80] Paleyes et al. (n 32).
[81] ibid.
[82] Talend, (n.d.), 'What is Data Obfuscation?'.
[83] A technique used to protect sensitive data by obscuring or modifying it in a way that makes it difficult to interpret or reverse-engineer.
[84] Fabris, A., Messina, S., Silvello, G., Susto, G.A. (2022), 'Algorithmic fairness datasets: the story so far'.
[85] Fischer-Hübner, S., (2009), 'Privacy-Enhancing Technologies'.
[86] Tools, techniques and methodologies designed to protect personal data and maintain the confidentiality and integrity of sensitive information.
[87] Oprisanu, B., (2023) 'Differential Privacy | Privacy-Enhancing Technologies PETs'.
[88] A rigorous mathematical definition of privacy that enables the analysis of datasets while providing strong guarantees against the re-identification of individuals in the data.
[89] Chicco, D., Oneto, L., Tavazzi, E. (2022), 'Eleven quick tips for data cleaning and feature engineering'.

Data security in preprocessing and training

Ensuring data security in the preprocessing phase involves safeguarding against various security risks, including data poisoning attacks and Trojan neural networks. Data poisoning attacks involve maliciously tampering with training data to degrade the performance of AI models, while Trojan neural networks embed hidden triggers that can cause unauthorised actions or incorrect predictions when activated[90].

To defend against such threats during preprocessing, practitioners can ensure secure model and data storage through appropriate encryption and access controls. The latter, which is explored further later on, is expressed in data governance measures (such as data strategies) that protect data from unauthorised modifications. In addition to this, implementing input validation safeguards ensure that only 'legitimate' inputs are processed, mitigating the risk of triggering hidden backdoors or Trojan networks[91].

Training data security is imperative, as models learn directly from the data they are exposed to. Data governance processes enabling the aforementioned techniques and technologies like DP, encryption, and federated learning[92] [93] can help to balance trade-offs between data privacy and utility. In particular, federated learning and certain other PETs, such as secure multi-party computation[94] [95], can enable collaborative training (the process of training a ML model across multiple devices or parties) without sharing any raw data[96].

Adversarial training improves model robustness by augmenting training data with adversarial data samples and deploying anomaly detection methods to help uncover and address instances of data tampering or poisoning. For instance, one popular technique called red teaming, which simulates real-world adversarial attacks on the AI/ML system, can further enhance the system's resilience by training it to address these instances[97] [98] [99].

---

[90] Aldoseri et al. (n 34).
[91] ibid.
[92] Martineau, K.2021), 'What is federated learning?'.
[93] Federated learning is a ML technique that trains an algorithm across multiple decentralised edge devices or servers holding local data samples, without exchanging them. This contrasts with traditional centralised ML techniques where all the training data is uploaded to one server.
[94] Frikken, K.B., (2011), 'Secure Multiparty Computation (SMC)'.
[95] Secure multi-party computation is a subfield of cryptography that enables multiple parties to jointly compute a function over their private inputs while keeping those inputs private from each other.
[96] Paleyes et al. (n 32)
[97] Aldoseri et al. (n 34)
[98] Whang et al. (n 62)
[99] Lu, Q., et al. (2023), 'Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering'

Data security in evaluation, deployment and ongoing monitoring

Maintaining data security during the evaluation phase of the AI data lifecycle is essential to safeguard AI/ML systems against potential threats and vulnerabilities. Detecting deviations from expected behaviour in training data can prevent further damage to AI/ML systems via malicious retraining of the underlying model, and can offer valuable insights for enhancing security measures[100]. The dynamic nature of data throughout the AI data lifecycle requires the continuous monitoring and adaptation of security measures. Staying abreast of evolving data patterns and potential security threats enables proactive adjustments to security protocols[101] [102].

AI/ML models, despite anonymisation efforts, can inadvertently disclose sensitive information about training data. Inference attacks[103] [104], such as model inversion[105] [106], exploit model predictions to extract private information or determine the presence of specific data points in the training set[107]. To address this, reusing existing data and pre-trained models significantly can reduce the time, effort and computational resources required for training new models. It can also enable privacy-preserving techniques such as differential privacy and transfer learning to be applied efficiently, to mitigate the risks of inference attacks like the aforementioned model inversion[108].

As with many other pillars, articulating and embedding effective monitoring mechanisms is essential for detecting data anomalies over time. Developing robust monitoring frameworks is crucial for optimising AI/ML systems and addressing any emerging challenges throughout its lifecycle[109].

---

[100] Aldoseri et al. (n 34).
[101] ibid.
[102] De Silva et al. (n 57).
[103] NordVPN, (2022), 'Inference attack definition - Glossary | NordVPN'.
[104] An inference attack is a data mining technique where an attacker analyses and combines seemingly innocuous pieces of information to illegitimately infer sensitive or confidential details about an individual, database, or organisation.
[105] Veale, M., Binns, R., Edwards, L., 2018. 'Algorithms that remember: model inversion attacks and data protection law'.
[106] A type of privacy attack where an adversary attempts to reconstruct or reveal the private training data used to train a machine learning model.
[107] Aldoseri et al. (n 60).
[108] ibid.
[109] Paleyes et al. (n 32).

Management

Effective data management is crucial across the AI data lifecycle, from collection to deployment. In this section, we highlight the importance of outlining the data requirements and ethical principles for each phase, supported by standardised templates. Documentation enhances transparency, accountability and collaboration, covering data cleaning, feature engineering, and metadata for findability and accessibility.

Managing diverse data types involves ensuring quality and preventing malicious entries. Efficient strategies optimise resource use and streamline data ingestion. Furthermore, integrating heterogeneous datasets requires harmonising schemas and handling associated uncertainties.

Data management in planning and documentation

Documentation for data management increases transparency and accountability, builds trust, and enables data discovery and sharing.To stay current, this documentation should ideally be interactive, integrated into tools/workflows, and automatable, particularly as documentation processes require collaboration between stakeholders interacting with data at different points[110] [111].

Separately, documentation can also cover methods and processes for data cleaning, feature engineering, data versions and pipeline software. For these purposes, FAIR principles can be adopted to define key metadata characteristics for findability, accessibility, interoperability and reuse. By contrast, a lack of documentation creates misunderstandings between actors operating in different, but dependent, lifecycle stages[112] [113].

Data integration and storage

Fluctuations in data collection procedures can introduce changes in data over time, adding layers of complexity to data management efforts. Managing data of many heterogeneous types and origins presents a number of challenges for data scientists, including handling schema differences, diverse data types, and ensuring data quality and consistency from varied providers[114].

---

[110] Lu et al. (n 99).
[111] 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'.
[112] Chicco et al. (n 89).
[113] Fabris et al. (n 84).
[114] Paleyes et al. (n 32).

The integrity of pre-existing datasets is also paramount when considering their reuse. Malicious data entries, such as backdoor attacks[115] [116], have the potential to compromise model behaviour, underscoring the necessity of sourcing datasets from trustworthy sources via secure means[117] [118]. Moreover, a considerable portion of compute time in data-processing pipelines is allocated to data ingestion. Therefore, defining and implementing efficient data management strategies is crucial to optimise resource utilisation and streamline the data ingestion process[119].

Data integration poses another challenge, as each relevant data source has its own unique schema and storage convention. Harmonising this disparate information into a cohesive dataset suitable for the AI/ML system requires significant effort[120]. The initial uncertainties inherent to integrated data, and means of knowledge extraction, further amplifies these challenges. However, various probabilistic methods[121] [122] can assess uncertainty levels to support effective integration[123].

Data management in preprocessing and exploration/analysis

As mentioned, curating integrated datasets involves aligning schemas, resolving conflicts, and handling uncertainty to improve quality[124] [125] [126]. Carefully identifying and selecting features from the data early on, in alignment with relevant domain knowledge, mitigates the likelihood of mismatches between data sources and the introduction of biases[127].

---

[115] Lutkevich, B., (2023), 'What is a Backdoor Attack? Tips for Detection and Prevention | Definition from TechTarget'.
[116] Backdoor attacks are a type of malicious entry or vulnerability intentionally inserted into a dataset used to train machine learning models.
[117] Ashmore, R., Calinescu, R., Paterson, C. (2021), 'Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges'.
[118] Murray, D.G., Šimša, J., Klimovic, A., Indyk, I. (2021), 'tf.data: a machine learning data processing framework'.
[119] ibid.
[120] Paleyes et al. (n 32).
[121] Jeevanandam, N., (2022), 'The Importance of Probabilistic Reasoning in AI'.
[122] Probabilistic reasoning is a form of knowledge representation that proves the existence of a mathematical object or structure with certain desired properties. In AI, these are used to examine data using statistical code.
[123] 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'.
[124] Bellomarini et al. (n 123)
[125] Paleyes et al. (n 32)
[126] Grafberger, S., Groth, P., Schelter, S. (2023), 'Automating and Optimizing Data-Centric What-If Analyses on Native Machine Learning Pipelines'.
[127] Niu, Y., Fan, Y., Ju, X. (2024), 'Critical review on data-driven approaches for learning from accidents: Comparative analysis and future research'.

Conducting EDA helps to assess metadata, statistics, distributions, relations and potential issues[128]. Since visualisations and software libraries support analysing complex structured data types, EDA also benefits from a mix of technical and domain expertise and favours simple, interpretable data management and analysis approaches, which improves understanding and transparency across a multidisciplinary team. Overly complex techniques risk decreasing interpretability without much accuracy gain. To help with ensuring the former whilst mitigating the latter, balancing any automation with a 'human-in-the-loop' for checking any results is important[129].

Data management in training

The training phase places heavy data demands in terms of storage, movement and computing, requiring extensive data management. As Bashath et al. (2022) discuss, deep transfer learning involves complex multi-step procedures that combine different data at each training stage[130]. Tracking these data workflows is an important governance need, especially given that large datasets drive requirements for specialised AI hardware like General Processing Units (GPUs) for efficient processing[131]. As data pipelines consume significant resources, it is imperative to align data preprocessing and collection/creation tasks before, and to integrate well with, training[132].

Access

Data access in AI and ML systems is critical and balances the need for data availability with stringent security and privacy requirements. In this section, despite a relative dearth of academic evidence relating to AI data access practices, we outline how effective data access frameworks can ensure that sensitive information is protected while enabling authorised users to perform necessary analyses. Techniques such as role-based access control (RBAC) and attribute-based access control (ABAC) assign permissions based on user roles and attributes, ensuring appropriate data usage.

[128] Chicco et al. (n 89).
[129] ibid.
[130] Bashath et al. (n 35).
[131] Aldoseri et al. (n 60).
[132] Murray et al. (n 118).

Transparency and privacy

Transparency plays a pivotal role in facilitating decision-making processes by providing access to crucial information, such as training data, models and applications. However, achieving transparency must be balanced with privacy considerations, especially when it involves access to personal data[133]. In contexts such as auditing procedures, transparency via access may compete with privacy, requiring careful deliberation over and adherence to ethical and legal frameworks[134].

Data sharing and its distribution

Comprehensive documentation helps datasets to be discovered and opens them for us by a diverse array of stakeholders, including companies, public administrations, universities and civic organisations. Through pertinent information about datasets (such as their origin, characteristics and permissible uses), practitioners can foster a culture of transparency and collaboration in the AI ecosystem.

The distinction between internal and external distribution of datasets within organisations underscores the importance of early consideration of distribution plans, particularly from a privacy standpoint. Both internal and external distribution are subject to review and approval according to internal company policies, highlighting the need for robust privacy safeguards. Proactive measures to map and appropriately address privacy concerns in data distribution can help practitioners navigate regulatory requirements and build trust with stakeholders[135].

Access controls and safeguards

Implementing access controls and safeguards protects sensitive data like personal information[136]. Balancing open data access for research with responsible controls for authorised users and uses is important. Before releasing datasets, enhancing metadata – for example, renaming variables to improve clarity and informativeness – can make data more usable and easily understood. For example, renaming variables to accurately reflect their meanings can prevent misinterpretation, and facilitate secondary studies[137].

[133] Gerdes (n 19).
[134] Hacker et al. (n 28).
[135] Heger et al. (n 111).
[136] Gerdes (n 19).
[137] Chicco et al. (n 89).

To ensure effective data governance, it is essential to implement various data access techniques tailored to different user needs and roles. Role-based access controls (RBAC) is a widely used method in which access permissions are assigned based on roles within an organisation. Each role has specific permissions, ensuring that users only access data necessary for their job. For instance, data stewards and data engineers might have access to raw data for preprocessing and anonymisation, while ML engineers and researchers access labelled or annotated data for model training and development[138].

On the other hand, attribute-based access controls (ABAC) allows access based on a user's attributes, such as their role, project membership, or data sensitivity levels. This method offers more granular control than RBAC. For instance, access to certain datasets can be restricted based on the project's requirements or the sensitivity of its data, thereby enhancing security and compliance[139].

Ethics

Establishing a robust ethical framework is essential for AI/ML system development, guiding every stage from problem setting to deployment. Ethical considerations must inform data collection and modelling to ensure fairness, accuracy and transparency, addressing hidden biases and societal impacts. As in other governance pillars, planning includes detailed documentation, stakeholder collaboration, and adherence to privacy regulations like GDPR.

Effective governance ensures that protected groups are represented, and risks are assessed continuously. Fairness and bias mitigation are crucial, and require techniques like data reweighting and adversarial learning. Comprehensive documentation provides context and helps avoid misuse, while continuous monitoring and multidisciplinary oversight ensure ethical integrity throughout the AI lifecycle.

[138] Lu et al. (n 99).
[139] ibid.

Data ethics across the lifecycle

Fundamental ethical principles to consider include fairness, transparency, accountability, safety, reliability, privacy, human values, diversity, and stakeholder needs. Effective planning analyses principles that might conflict, and defines appropriate tradeoffs[140] [141]. For this reason, incorporating AI ethics into model training involves planning to address biases in all stages[142].

Planning can establish broad ethical frameworks that narrow to granular ethical considerations for problem setting, algorithms, data representations, and stakeholders. Guidelines help make ethical choices during development and use. Codes of ethics from organisations provide ethical rules for AI system development[143] [144]. It is crucial to recognise that ethical considerations extend beyond algorithmic fairness to encompass the entire AI data lifecycle, from data collection to model deployment[145].

The data collection and data preprocessing phases are crucial for ensuring fairness and data accuracy. As data typically reflects societal biases, it must be demonstrated that it is accurate, representative and neutral. Leslie (2019) emphasises the necessity of establishing a continuous chain of human responsibility across the entire AI project delivery workflow. If practitioners do not adhere to this recommendation, complications may arise due to hidden biases in data that are difficult to foresee and detect[146] [147].

In selecting data and training datasets, it is essential to account for data protection measures, ensure transparency, and establish the means to mitigate potential harmful biases[148]. Issues of representation, inclusion and diversity are central to a fair AI/ML community. Due to historical biases stemming from structural inequalities, some populations and their perspectives are underrepresented in certain domains.

---

[140] ibid.
[141] De Silva et al. (n 57).
[142] Gowda, S.C.M., Joshi, S., Zhang, H., Ghassemi, M. (2021), 'Pulling Up by the Causal Bootstraps: Causal Data Augmentation for Pre-training Debiasing'.
[143] Gerdes (n 65).
[144] Lu et al. (n 99).
[145] Chicco et al. (n 89).
[146] Leslie, D., (2019), 'Understanding artificial intelligence ethics and safety'.
[147] Gerdes (n 19).
[148] ibid.

Risk assessment and governance

Risk assessment frameworks tailored for AI ethics can be used to systematically evaluate risks related to ethics, unintended harms, biases, discrimination, and other potential negative societal impacts. Stakeholder interviews can help to identify these risks, supported by standards like ISO 23894, which provides risk assessment processes[149] [150]. The planning stage may analyse risks across the pipeline – from problem specification to data collection, model development, evaluation and deployment. Sources of bias span data, algorithms, evaluation metrics, generalisation, and the integration of knowledge sources[151] [152].

Ethical considerations should always inform data collection and modelling activities. The Alan Turing Institute emphasises that a continuous chain of human responsibility must be established across the entire AI project delivery workflow. If not, complications may arise due to hidden biases in data that are difficult to foresee and detect[153].

Governance frameworks should also ensure that protected groups are well represented in the data. The review of datasets should consider privacy and transparency, ensuring that all data providers are informed that their data is used and for what purpose, and data users that informed consent has been given for the use of data for specific purposes. This approach helps in identifying and mitigating biases, ensuring that the constructs used as input and output variables are valid and reliable[154].

Governance and documentation

From the outset, AI/ML systems should be verified against ethical requirements, not just technical validity. System explainability enables tracing data use, system behaviours, and decisions. Model transparency and explanations can be customised for different users[155] [156]. Planning data and model documentation using standardised templates improves transparency on quality, limitations, trade-offs and risks to developers and users. In this way, documentation facilitates the identification of ethical issues[157] [158].

---

[149] Lu et al. (n 99).
[150] Simbeck, K. (2023), 'They shall be fair, transparent, and robust: auditing learning analytics systems'.
[151] Aldoseri et al. (n 60).
[152] Khan, M.J., Breslin, J.G., Curry, E. (2023), 'Towards Fairness in Multimodal Scene Graph Generation: Mitigating Biases in Datasets, Knowledge Sources and Models'.
[153] Gerdes (n 19).
[154] Fabris et al. (n 84).
[155] Lu et al. (n 99).
[156] Simbeck (n 150).
[157] Heger et al. (n 111).
[158] Miceli, M., Posada, J., Yang, T., (2022), 'Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power?'.

Datasets are often taken as factual information that supports objective computation and pattern extraction. However, research in human–computer interaction and critical data studies argues that this belief is superficial and potentially harmful. Proper documentation should discuss and explain features, providing context about who collected and annotated the data – and how, and for which purpose. This information helps dataset users select appropriate datasets for their tasks and avoid unintentional misuse[159].

Fairness and bias mitigation

Fairness in AI and ML systems is crucial across various tasks, including classification, ranking, matching, anomaly detection and NLP. This involves ensuring equitable treatment across different sub-populations, such as racial groups or genders, by balancing features to produce fair outcomes[160]. Addressing biases in word embeddings, language models, machine translation and speech recognition is essential to prevent perpetuating stereotypes and inequities throughout the AI data lifecycle.

Issues of representation, inclusion and diversity are central to the fair ML community. Due to historical biases stemming from structural inequalities, some populations and their perspectives are underrepresented in certain domains. Data is a human-influenced entity, determined by discretionary decisions on measurement, sampling and categorisation, shaping how, and by whom, data will be collected and annotated. Some ecological studies suggest that data science professionals are aware of how curation and annotation choices influence their data and its relation to the underlying phenomena[161].

Training data often reflects societal biases, necessitating governance to ensure ethical AI development. Internet-scraped data used in training large language models can embed stereotypes and prejudices, requiring techniques like data reweighting to mitigate bias[162] [163]. Subtle biases, such as confounding correlations, need careful auditing and analysis to uncover. Adversarial learning can jointly enhance model fairness and robustness in these scenarios[164]. Governance strategies should promote transparency, implement safeguards, educate users on model limitations, and involve multidisciplinary oversight to address technical and social considerations throughout the AI lifecycle[165].

---

[159] Heger et al. (n 111).
[160] Fabris et al. (n 84).
[161] Heger et al. (n 111).
[162] Corchado, J.M., F, S.L., V, J.M.N., S, R.G., Chamoso, P. (2023), 'Generative Artificial Intelligence: Fundamentals'.
[163] Aldoseri et al. (n 60).
[164] Whang et al. (n 62).
[165] Corchado et al. (n 162).

Summary

Each pillar addresses unique aspects of data governance, ensuring that data is not only effectively managed and secured but also ethically utilised and accessible to authorised stakeholders. Table 1, below, encapsulates the key findings from our governance analysis, providing an overview of each pillar.

It details the core descriptions, key themes, involved stakeholders, challenges, tools and techniques for each pillar. This structured representation aims to facilitate a deeper understanding of what the academic literature is saying about data governance practices across the AI data lifecycle.

Next, we bring together insights from across the literature covered in the three reports, to a discussion on the gaps identified, for the purpose of informing further research in this area.

Table 1. Summary of AI Data Governance Considerations

Pillar	Description	Key Themes	Actors Involved	Challenges	Tools and Techniques
Quality	• Maintaining high data quality for AI effectiveness and reliability. • Metrics and methods for error identification and correction. • Ensuring balanced representation in datasets. • Continuous validation processes during training and deployment.	Error identification; balanced representation; continuous validation	Data Scientists; Data Engineers; Domain Experts	Balancing representation; detecting biases	Deequ; DQLearn; Pandas Profiling; TensorFlow Validation
Management	• Comprehensive guidelines for organising, storing and retrieving data to ensure consistency and accessibility. • Efficient resource allocation and standardisation of formats. • Clear guidelines for each phase, from planning to deployment.	Organising; storing; retrieving; resource allocation; standardisation	Data Engineers; Project Managers; AI/ML Engineers	Standardising data formats; efficient resource allocation	Efficient data storage systems; query optimisation tools
Security	• Implementing strong security measures such as encryption, input validation and continuous monitoring. • Privacy-preserving techniques and robust anonymisation. • Continuous monitoring to detect and mitigate threats.	Encryption; input validation; continuous monitoring	Security Experts; Data Engineers; Compliance Officers	Protecting against breaches; ensuring data privacy	Encryption tools; anonymisation techniques
Access	• Balancing data accessibility with stringent security and privacy requirements. • Implementing role-based and attribute-based access controls. • Ensuring that data access is compliant with privacy regulations.	Role-based access control (RBAC); attribute-based access control (ABAC)	Data Stewards; IT Administrators; Legal Teams	Managing data accessibility without compromising security	RBAC; ABAC
Ethics	• Integrating ethical considerations into data governance to ensure responsible and findable, accessible, interoperably and reusable (FAIR) data usage. • Addressing biases and ensuring transparency. • Maintaining accountability throughout the AI data lifecycle.	Addressing biases; transparency; accountability	Ethics Committees; Data Scientists; Legal Teams	Ensuring fairness; maintaining transparency	Ethical AI frameworks; fairness auditing tools

Discussion and recommendations

The scoping review revealed significant gaps in the existing literature regarding data governance practices across the AI data lifecycle. This section delves into these gaps, highlighting areas requiring further investigation for the development of robust governance frameworks.

Incompleteness in mapping governance practices

The review identified inconsistencies in the literature regarding data governance practices at each stage of the AI data lifecycle. While some studies provide insights into documentation and techniques relevant to specific governance aspects (for example, data quality, security), a survey of these practices across all lifecycle stages remains absent. However, some insights stood out as important for further study from the literature that was analysed. These are listed below.

Standardise documentation

To enhance consistency and facilitate knowledge sharing across the AI development community, standardised documentation templates and guidelines should be explored, developed and potentially adopted for, or across, each stage of the AI data lifecycle. This approach ensures that data governance practices are uniformly applied and easily understood by all stakeholders, promoting transparency and accountability. The benefits of this aim are clear:

- **Improved transparency and accountability:** Standardised documentation practices increase transparency by clearly outlining data provenance, characteristics, and permissible uses, which is key for upstream and downstream tasks. This transparency aids in understanding the context of data collection and processing, which is crucial for informed decision-making during data exploration and analysis[166] [167].

[166] Simbeck (n 150)
[167] Heger et al. (n 111)

- **Enhanced consistency and reproducibility:** By using standardised templates, such as datasheets and model cards, documentation becomes more consistent and easier to reproduce. These templates ensure that essential metadata is captured systematically, facilitating the replication of experiments and the validation of results across different stages of the AI data lifecycle[168] [169].

- **Facilitation of knowledge sharing:** Standardised documentation templates provide a common framework that facilitates knowledge-sharing among various stakeholders in the AI development community. This common framework includes dataset documentation, model service documentation, and information sheets, which help in disseminating critical information efficiently[170].

Implementation strategies

The literature from this scoping review suggests several implementation strategies that researchers and policymakers should consider exploring in greater detail to enhance data governance practices in AI development. These strategies can help improve the consistency, transparency and accountability of documentation throughout the AI data lifecycle:

- **Interactive and automated documentation:** The literature emphasises the importance of integrating documentation into existing tools and workflows to maintain its accuracy and relevance. Researchers and policymakers should investigate the promotion of interactive and automated documentation tools such as wikis, version control repositories, and automated AI systems. These tools streamline information retrieval and updates, support continuous improvement, and foster collaboration among data science teams[171].

- **Role-specific documentation:** Highlighted as a best practice, tailoring documentation to the specific roles within the AI development ecosystem is crucial. Developing role-specific guidelines and training programmes ensures that all stakeholders understand their documentation responsibilities and can effectively contribute to data governance. Researchers should explore how this approach can clarify roles and duties, enhancing the overall efficiency and accuracy of the governance process[172].

[168] Chicco et al. (n 89)
[169] Fabris et al. (n 84)
[170] Miceli et al. (n 158).
[171] Heger et al. (n 111).
[172] Lu et al. (n 99).

- **A place for ad-hoc documentation:** The literature suggests that ad-hoc documentation, created as needed rather than through a standardised process, is common in many AI development projects and holds value for practitioners. Researchers should examine how widespread this practice is and its impact on the consistency and effectiveness of data governance. Understanding the role and benefits of ad-hoc documentation can help determine how it can complement standardised practices[173] [174].
- **Adoption of FAIR principles:** Implementing the FAIR (Findability, Accessibility, Interoperability and Reusability) principles in documentation practices is recommended. These principles ensure that datasets and models are easily discoverable and usable by others, enhancing the quality and usability of AI/ML systems. Policymakers should advocate for the adoption of FAIR principles to promote better data management and sharing practices, which are crucial for transparency and accountability in AI development[175].

Although the existing literature does not provide a comprehensive review of implementation practices for standardised documentation, these highlighted strategies offer a valuable starting point for researchers and policymakers. By focusing on interactive and automated documentation, role-specific guidelines, investigation of ad-hoc documentation and the adoption of FAIR principles, stakeholders can work towards more effective and ethical data governance frameworks in AI development. These practices, drawn from current research, provide a pathway to improving transparency, accountability and collaboration in the AI community[176] [177].

[173] Simbeck (n 150).
[174] Heger et al. (n 111).
[175] Micheli, M., Hupont, I., Delipetrev, B., Soler-Garrido, J., (2023), 'The landscape of data and AI documentation approaches in the European policy context'.
[176] Simbeck (n 150).
[177] Heger et al. (n 111).

Downstream lifecycle stages

While academic research has yet to thoroughly explore AI data governance during downstream phases of the AI data lifecycle (that is the lifecycle from training and onward), solutions are emerging to address this need. One such solution is DataPerf, an open source platform and set of benchmarks specifically designed to manage data quality, security and access control during the model evaluation and training stages[178]. This further reinforces the need for more academic research reviewing the overall commonly-used methods of designing, developing and implementing data governance during these later tasks[179].

Here, we outline several areas for policymakers and researchers to consider:

Exploring data management in downstream stages

Further research is needed to investigate how data management practices evolve during the evaluation and training stages, to ensure the integrity and security of data used for model development. Current literature suggests several areas of focus:

- **Data quality governance:** Effective data quality governance practices during the downstream stages are critical for the development of reliable AI models. Ensuring data integrity involves continuous monitoring and validation of the data used for training and evaluation. This includes implementing robust data validation frameworks that can detect and correct anomalies in real time[180].
- **Security protocols:** Security is paramount during the evaluation and training stages to protect sensitive data from breaches and misuse. Research should explore advanced encryption techniques and secure data-access protocols to safeguard data at these critical phases. Establishing secure environments for data processing, such as isolated training environments, can significantly reduce the risk of data leaks[181].

[178] Mazumder et al. (2023), 'DataPerf: Benchmarks for Data-Centric AI Development. Presented at the Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track'.
[179] Gerdes (n 65).
[180] Aldoseri et al. (n 60).
[181] Schneider et al. (n 10).

- **Access-control mechanisms:** As explored in greater detail below, governing who has access to data during the evaluation and training stages is crucial for maintaining data integrity and security. Implementing role-based access control (RBAC) ensures that only authorised personnel can access sensitive data. This approach helps in mitigating the risk of unauthorised access and potential data corruption[182].

Examining ecosystem interactions

Research should also delve into the data exchanges and interactions occurring within the AI development ecosystem during evaluation and training stages. Key areas of exploration include:

- **Data exchange protocols:** Understanding how data is shared between different stakeholders in the AI ecosystem is vital for identifying potential governance challenges. Studies should examine the protocols and standards used for data exchange, and how these impact data integrity and security. This includes exploring the role of data stewards in overseeing data exchanges and ensuring compliance with governance frameworks[183].
- **Collaboration between roles:** The AI development ecosystem involves multiple roles, including data scientists, data engineers and data stewards, each of whom plays a part in managing data. Research should investigate how these roles interact and collaborate during the downstream stages, focusing on communication channels and collaboration tools that facilitate effective data governance[184].
- **Governance challenges and strategies:** Identifying potential data governance challenges during the evaluation and training stages is essential for developing mitigation strategies. This includes exploring issues such as data bias, ethical considerations and compliance with regulatory standards. Developing governance frameworks that address these challenges can help in ensuring the ethical and responsible use of data[185].

[182] Schneider et al. (n 10).
[183] De Silva et al. (n 57).
[184] 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'.
[185] Lu et al. (n 99).

While the literature on data management during downstream lifecycle stages of AI development is sparse, industry practices and emerging solutions offer a foundation for further exploration. By focusing on data quality management, security protocols and access control mechanisms, and by examining ecosystem interactions, future research can significantly enhance our understanding of data management in these crucial stages. Establishing robust data governance practices during model evaluation and training will ensure the development of reliable, secure and ethically sound AI systems.

Limited guidance on data access governance

Data access governance refers to the policies, processes and standards that determine how data can be accessed, by whom, and under what conditions. In the context of AI, data access governance ensures that data is used responsibly throughout its lifecycle at each point where it changes hands, from data source to practitioner – and indeed, between practitioners. Governance of this dynamic is crucial for maintaining security, privacy and compliance with regulations, while enabling the effective use of data in AI development[186].

A significant gap identified in the literature is the lack of comprehensive literature specifically addressing data access governance. Ensuring safe and responsible data access throughout the AI data lifecycle is essential. While some regulations, like the EU's GDPR, provide guidelines on data subject rights and access control, there is a need for a more thorough exploration of data access governance in AI development contexts[187]. Further research is needed to holistically explore, and potentially establish, best practice for aspects such as:

[186] Paleyes et al. (n 32).
[187] Gerdes (n 19).

Access-control mechanisms

The available literature indicates that defining who can access data at different stages of the lifecycle and setting appropriate authorisation levels is an important consideration for practitioners. This includes understanding the roles and responsibilities of various stakeholders involved in the AI ecosystem, such as data scientists, data engineers and domain experts[188]. Broadly, this access may be divided into two types:

- **Raw data access:** Strict access controls should be in place for raw data, especially when dealing with sensitive information like medical records or financial data. Role-based access control (RBAC) can be implemented, where only authorised personnel with specific roles, such as data stewards or data engineers, have access to the raw data for preprocessing and anonymisation[189].
- **Labelled/annotated data access:** Once the data is preprocessed and labelled/annotated, access can be granted to a broader set of stakeholders, such as machine learning engineers and researchers, for model training and development purposes. Attribute-based access control (ABAC) can be employed, where access is granted based on attributes like project membership, data sensitivity levels, or specific use cases[190].

Data access with anonymisation and pseudonymisation techniques

Implementing strategies to protect sensitive data while allowing access to it for AI development is crucial. Techniques like anonymisation and pseudonymisation are critical to balancing data utility and privacy, particularly for systems operating in very security-sensitive domains (for example, healthcare, finance)[191]. Different data types and levels of sensitivity may have different access requirements, such as:

- **Anonymisation for training data access:** When dealing with sensitive data like medical records or personal information, anonymisation techniques like data masking, generalisation, or differential privacy can be applied to the training data used for model development. This ensures that individual identities are protected while preserving the utility of the data for AI training[192]. Access to

[188] Schneider et al. (n 10).
[189] Bashath et al. (n 35).
[190] Bellomarini et al. (n 123).
[191] Paleyes et al. (n 32).
[192] Bashath et al. (n 35).

anonymised training data can be granted to a broader set of stakeholders, such as machine learning engineers and researchers, as the risk of re-identification is minimised[193].

- **Pseudonymisation for fine-tuning data access:** In scenarios where personalised or context-specific data is required for fine-tuning AI models, pseudonymisation techniques like tokenisation or encryption can be employed. This allows the data to be used for model optimisation while maintaining a level of privacy and reversibility if needed. Access to pseudonymised fine-tuning datasets can be more restricted, as there is a higher risk of re-identification if the pseudonymisation keys are compromised[194].

The literature suggests that a choice between anonymisation and pseudonymisation for data access should be based on the level of data sensitivity, regulatory requirements, and the intended use case. For example, highly sensitive data like medical records may require strict anonymisation before access is granted for training purposes, while less sensitive data, such as customer behaviour data, could be pseudonymised for evaluation purposes like fine-tuning[195].

Audit trails and accountability

Maintaining detailed records of data access is essential for ensuring traceability and accountability. This involves creating transparent documentation and data lineage to track how data is used, transformed and shared across different stages of AI development[196]. As made clear in the literature, the following practices enhance audit trails and accountability:

- **Transparent documentation:** Documentation captures the provenance, characteristics and permissible uses of datasets. This transparency helps in understanding the context of data collection and processing, aiding informed decision-making during exploration and analysis[197].

[193] Fabris et al. (n 84).
[194] Paleyes et al. (n 32).
[195] Fabris et al. (n 84).
[196] Heger et al. (n 111).
[197] Gerdes (n 19).

- **Data lineage and provenance tracking:** Recording information about data provenance, processing history, and ownership, provides transparency and supports auditing data flows and quality issues. Standardised schemas for documenting datasets ensure interoperability and effective metadata management[198].

- **Continuous documentation:** Practitioners benefit from documentation being interactive, integrated into tools and workflows, and, where possible, automatable to remain current. This approach increases transparency and accountability, building trust and enabling data discovery and sharing[199].

Ensuring data access is governed in a way that protects privacy and security, while enabling innovation and compliance with legal standards, is crucial. While the literature reveals useful insights relating to data access mechanisms described above, the information is fragmented, and this is the pillar with the least consistent reference to the AI data lifecycle stages. Addressing these areas through further primary research, leveraging structured frameworks can contribute to robust data governance frameworks that support ethical and effective AI development[200].

For instance, the ODI's (2021) framework for facilitating safe access to sensitive data is an approach for mapping the means and considerations driving governance of AI data access. This framework is centred around the 'data-use journey', which delineates the stages through which data flows within an entity: collection, ingress, management, egress and usage. By understanding and managing these stages, the framework ensures that data (particularly sensitive data, given the risks involved) is accessed responsibly and securely at every point in its lifecycle[201].

At the core of this framework are the four levers of control: 'who', 'what', 'how' and 'what for'. These levers define the mechanisms that can be used to regulate access to sensitive data:

- **Who gets access:** stringent authentication and authorisation processes ensure only qualified individuals can access the data.

- **What data can be accessed:** this is managed through data minimisation and modification techniques like anonymisation and pseudonymisation.

[198] Fabris et al. (n 84).
[199] Gerdes (n 19).
[200] Schneider et al. (n 10).
[201] ODI, (2021), 'How do data institutions facilitate safe access to sensitive data?'.

- **How data is accessed:** technical controls such as secure virtual environments and controlled data streams ensure that sensitive data remains protected during use.
- **What for:** the permissible purposes for data use, aligning data access with ethical guidelines and legal requirements.

This kind of structure presents a useful tool for researchers, policymakers and even practitioners to effectively map factors supporting data access throughout AI/ML system development pipelines.

The practitioner landscape

The reviewed literature predominantly outlines high-level data governance tasks and activities without delving into the specific roles and responsibilities of practitioners within the AI development ecosystem. Understanding these roles is crucial for assigning ownership and accountability for governance implementation.

Identified roles in AI development

The reviewed literature predominantly outlines high-level data governance tasks and activities without delving into the specific roles and responsibilities of practitioners within the AI development ecosystem. Understanding these roles is crucial for assigning ownership and accountability for governance implementation[202]. These roles include:

- **Data Scientists:** Responsible for problem formulation, data collection, initial data cleaning, exploratory data analysis and developing predictive models. They play a pivotal role in ensuring data quality and relevance throughout the AI lifecycle[203].
- **Data Engineers:** Focus on managing the data lifecycle, including data collection, preprocessing, and infrastructure development. They ensure data integrity and usability, and their work is crucial for maintaining scalable data solutions[204].

[202] Coulthart, S., Shahriar Hossain, M., Sumrall, J., Kampe, C., Vogel, K.M. (2023), 'Data-Science literacy for future security and intelligence professionals'.
[203] Ashmore et al. (n 117).
[204] Paleyes et al. (n 32).

- **Domain/Subject Matter Experts:** Provide industry-specific knowledge that guides data collection, preprocessing, and model evaluation. Their expertise ensures that AI systems are relevant and aligned with real-world standards and ethical considerations[205].
- **AI/ML Engineers:** Transform prototypical AI models into deployed services, develop and maintain data collection tools, and ensure models are effectively monitored and updated based on real-world feedback[206].
- **Project Managers:** Oversee the coordination and execution of AI projects, ensuring alignment between different teams. They manage workflows and ensure that data-related activities adhere to project standards and timelines[207].

While we have been able to identify these roles from across different studies, there is a lack of literature that maps these roles and their responsibilities in detail and across lifecycle stages. This gap poses a challenge in understanding the holistic implementation of data governance within AI development[208].

Gaps in current literature

While some studies provide insights into specific governance aspects and roles, no single study offers a comprehensive review of the actors involved in the data ecosystem of AI/ML development. Future research should aim to:

- **Map roles and responsibilities:** Establish a clear picture of the various actors involved in the AI development ecosystem (including data scientists, data engineers and data stewards) and their responsibilities for data governance practices at each stage of the lifecycle[209].
- **Develop role-specific guidance:** Tailor data governance best practice and training programmes to the specific needs and responsibilities of different practitioners within the AI development ecosystem. This approach would ensure that all stakeholders understand their roles in governance implementation, promoting accountability and effective data management[210].

[205] Borrego-Díaz et al. (n 117).
[206] Paleyes et al. (n 32).
[207] De Silva et al. (n 57).
[208] Coulthart et al. (n 202).
[209] Paleyes et al. (n 32).
[210] Bellomarini et al. (n 123).

Addressing these gaps will provide a more detailed understanding of the practitioner landscape, facilitating better assignment of roles and responsibilities in AI data governance[211]. But simply describing roles and responsibilities does not provide a necessarily holistic view, as it may forego how these interact with one another. Well-established methodologies such as data ecosystem mapping (DEM) can support efforts to explore the actors taking part in a data ecosystem, and join up the exchanges of 'soft' – knowledge, insights – and 'hard' – the data itself[212].

Tools like these can help organisations visualise and analyse the complex web of data flows, dependencies and stakeholder interactions within their AI data ecosystems. Mapping out these relationships can identify potential risks, bottlenecks or governance gaps, and implement appropriate controls and policies to ensure secure, ethical, and responsible data handling practices across the entire AI data lifecycle.

Beyond ML: A focus on broader AI systems

The body of literature reviewed for this scoping review predominantly concentrates on the ML component of the AI/ML spectrum, which includes technologies like NLP. This focus on ML, while significant, often neglects other, equally crucial, forms of AI. Specifically, rule-based systems and knowledge-based systems, which rely on predefined rules and logic to process data and make decisions, are underrepresented in current data governance research. These systems do not learn from data in the same way as ML models but are still integral to many AI/ML systems[213].

In the domain of medicine, rule-based systems power prominent expert systems that support medical professionals. MYCIN, developed in the 1970s at Stanford University and a well-known example of such a system, employed a vast knowledge base of infectious diseases and appropriate antibiotic therapies, formulated as rules. Given a patient's symptoms and medical history, the system could recommend potential diagnoses and treatment plans, aiding doctors in complex decision-making[214]. Similarly, rule-based expert systems are still crucial in other safety-critical areas. For instance, in aviation, planes' Terrain Awareness and Warning System (TAWS) uses a set of rules to identify potential dangers based on factors like altitude, terrain data and flight path[215].

---

[211] Paleyes et al. (n 32).
[212] ODI (2022), 'Mapping data ecosystems: methodology'.
[213] Borrego-Díaz et al. (n 117).
[214] Clancey, W.J., (1994), 'Notes on Epistemology of a rule-based expert system'.
[215] Airbus, (2022), 'Safety Innovation #3: Terrain Awareness Warning System (TAWS)'.

Importance of rule-based systems

Rule-based AI systems operate on a set of predefined rules to derive conclusions from input data (hence, here we drop the 'AI/ML' terminology). These systems are essential in applications where consistency, transparency, and adherence to regulatory frameworks are paramount. For example, in financial services, rule-based systems are used for compliance monitoring and fraud detection. Similarly, expert systems in healthcare assist in diagnostic processes by applying medical knowledge, encoded in rules, to patient data[216].

Data governance challenges

The limited focus on rule-based systems in data governance frameworks presents a significant oversight. These systems, although not adaptive like ML models, require rigorous data governance to ensure they operate correctly and fairly. The key challenges include:

- **Data integrity and consistency:** Ensuring the data used in rule-based systems is accurate and up-to-date is critical. Unlike ML models that can adapt to new data, rule-based systems rely on the accuracy of their initial rule set and the data they process[217].
- **Transparency and explainability:** Rule-based systems are inherently more transparent than ML models since their decision-making process is based on explicit rules. However, maintaining this transparency requires documentation and governance of the rules and data used[218].
- **Security and compliance:** Ensuring that rule-based systems comply with regulatory requirements and security standards is crucial, particularly in sensitive applications like healthcare and finance. This includes safeguarding the data from unauthorised access, and ensuring the rules are implemented consistently[219].

[216] De Silva et al. (n 57).
[217] Borrego-Díaz et al. (n 117).
[218] Simbeck (n 150).
[219] Paleyes et al. (n 32).

Expanding data governance frameworks

To address these challenges, existing data governance frameworks need to be expanded to explicitly include rule-based AI systems. This involves several key steps:

- Incorporating rule-based systems into governance policies: Data governance policies should be updated to cover the specific requirements of rule-based systems, including the integrity, consistency and transparency of the rules and data used[220].
- Developing role-specific guidelines: Tailoring data governance best practice and training programmes to the specific needs and responsibilities of different practitioners, including those working with rule-based systems, will help ensure effective governance across all types of AI systems[221].
- Creating comprehensive documentation standards: Establishing standards for documenting the rules and logic used in these systems, as well as the data they process, is essential for maintaining transparency and accountability[222].

Emerging Areas: Federated learning and Privacy-Enhancing Technologies (PETs)

In addition to rule-based systems, there are other emerging areas within AI that require attention in data governance frameworks:

- **Privacy-enhancing technologies (PETs):** These technologies, such as differential privacy, homomorphic encryption, and secure multi-party computation, are designed to enhance privacy in ML applications. PETs enable data to be used for training ML models while ensuring individual privacy. Integrating PETs into data governance frameworks will help to maintain privacy standards and compliance with regulations like GDPR and CCPA[223].
- **Federated learning:** This approach is a specific PET that involves training ML models across multiple decentralised devices or servers holding local data samples without exchanging them. It enhances privacy and security by keeping the data itself localised. Governance

[220] De Silva et al. (n 57).
[221] Bellomarini et al. (n 123).
[222] Fabris et al. (n 84).
[223] Fabris et al. (n 84).

frameworks should include guidelines for managing data in federated learning environments, ensuring data privacy, model integrity, and compliance with local regulations[224].

Future research directions

To ensure the responsible and ethical development of AI technologies, future research should broaden its scope to include all types of AI systems. Key areas for future research include:

- **Extending data governance frameworks:** Research should focus on expanding existing data governance frameworks to explicitly address the unique data considerations and challenges associated with rule-based AI systems. This includes exploring how these systems can be integrated into broader AI governance models[225].
- **Investigating data governance for specific applications:** Detailed studies are needed to understand the nuances of data governance in specific applications that rely on rule-based systems, such as expert systems and decision-support systems. This research should aim to develop tailored governance strategies that ensure the integrity, security and fairness of these systems[226].
- **Evaluating ecosystem interactions:** Exploring the interactions within the AI development ecosystem, particularly how data exchanges and governance challenges manifest during the deployment and operation of rule-based systems, will provide valuable insights for developing robust governance frameworks[227].

By broadening the scope of data governance research to encompass all types of AI systems, including rule-based and knowledge-based systems, as well as emerging areas like federated learning and PETs, the field can ensure the development of AI technologies that are not only advanced but also responsible, ethical, and aligned with societal values. This approach to data governance is crucial in fostering trust and reliability in AI applications across various sectors[228].

[224] Whang et al. (n 62).
[225] Paleyes et al. (n 32).
[226] Borrego-Díaz et al. (n 117).
[227] De Silva et al. (n 57).
[228] Coulthart et al. (n 202).

Closing remarks

The culmination of this series of reports on data-centric AI governance underscores the critical importance of robust data governance frameworks that span the entire AI lifecycle. The transformative potential of AI and ML systems is undeniable; they promise enhanced automation and decision-making capabilities across various sectors. However, this potential can only be fully realised when data governance is prioritised, ensuring that the data fuelling these systems is of high quality, ethically sourced, and managed with stringent standards throughout its lifecycle.

Our exploration began with an in-depth look at the AI data lifecycle, highlighting the need for a data-centric approach. This perspective shifts the focus from merely optimising model performance to ensuring that the data used is curated and governed. This approach addresses fundamental issues such as data provenance, bias and long-term management, which are pivotal for developing AI/ML systems that are not only effective but also equitable, safe and compliant with regulatory standards.

The second phase of our research delved into the AI ecosystem, examining the interactions between various stakeholders, including data scientists, engineers, domain experts and policymakers. This ecosystem perspective emphasised the importance of collaborative governance, where data sharing, transparency, and ethical considerations are embedded into every interaction. Effective governance frameworks must facilitate these interactions, ensuring that data is handled responsibly and that the value exchange among stakeholders is optimised.

In this final report, we integrated insights from the AI lifecycle and ecosystem analyses to address the gaps on AI data governance identified in the current literature and practices, advocating for a unified policy approach that encompasses all stages of the AI lifecycle. By doing so, we aim to foster an environment in which AI/ML systems can be developed and deployed responsibly, with an emphasis on continuous monitoring and improvement to maintain ethical integrity and social benefit.

Our findings underscore the need for ongoing research and the development of best practices that can adapt to the rapidly evolving AI landscape. Policymakers, practitioners and researchers must collaborate to refine and implement governance models that ensure the trustworthiness and reliability of AI applications. This collaboration is essential for navigating the complex ethical, legal and technical challenges that arise in AI development and deployment.

In conclusion, the series highlights the paramount importance of a holistic approach to data governance in AI. By ensuring that data governance frameworks are robust and comprehensive, we can support the creation of AI/ML systems that are not only technologically advanced, but also aligned with societal values and ethical standards. This approach will enable the responsible stewardship of data throughout the AI lifecycle, fostering trust and reliability in AI applications across diverse real-world contexts.

Annex A. Methodology

A scoping review is a type of knowledge synthesis that maps the existing literature on a particular topic[229]. 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[230] [231]. This methodological approach is particularly useful for emerging or complex topics where existing research may be diverse and heterogeneous in nature. As we have shown, this is true of AI/ML data governance[232] [233].

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 governance considerations need to be made at each stage of the AI data lifecycle?

[229] Mak, S., Thomas, A. (2022), 'Steps for Conducting a Scoping Review'.
[230] Levac, D., Colquhoun, H., O'Brien, K.K. (2010), 'Scoping studies: advancing the methodology'.
[231] Arksey, H., O'Malley, L. (2002), 'Scoping studies: towards a methodological framework'.
[232] Floridi et al. (n 4).
[233] 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 covered terms relating to the general AI domain. Secondly, we 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[234] [235]. Finally, through a trial-and-error approach, we added 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:

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

[235] Knowledge graphs 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'[236] 'Design, planning and prototyping'[237] 'Problem formulation'[238] 'Identify and formulate the problem' and 'Review data and AI ethics'[239] 'Planning phase'[240]
Collection and creation	'Data collection'[241] 'Data collection'[242] 'Data discovery'[243] 'External data acquisition'[244] 'Collection phase'[245] 'Data ingestion'[246]
Exploration and analysis	'Data analysis and preprocessing'[247] 'Error detection and repair'[248] 'Data preparation' and 'data exploration'[249] 'Preparation phase' and 'Analysis phase'[250] 'Data preparation and cleansing'[251]

[236] ICO (2024), 'Annex A: Fairness in the AI lifecycle'.
[237] Chai, C., Wang, J., Luo, Y., Niu, Z., Li, G. (2023), 'Data Management for Machine Learning: A Survey'.
[238] Khan, M.J., Breslin, J.G., Curry, E. (2023), 'Towards Fairness in Multimodal Scene Graph Generation: Mitigating Biases in Datasets, Knowledge Sources and Models'.
[239] De Silva et al. (n 57).
[240] Shah, S.I.H., Peristeras, V., Magnisalis, I. (2021), 'DaLiF: a data lifecycle framework for data-driven governments'.
[241] ICO (n 234).
[242] Khan et al. (n 236).
[243] Chai et al. (n 235).
[244] De Silva et al. (n 57).
[245] Shah et al. (n 238).
[246] Cognilytica, (2020), 'AI Data Engineering Lifecycle Checklist'.
[247] ICO (n 234).
[248] Chai et al. (n 235).
[249] De Silva et al. (n 57).
[250] Shah et al. (n 238).
[251] Cognilytica (n 244).

Preprocessing	‘Data analysis and preprocessing’[252] ‘Data cleaning and data annotation’[253] ‘Data preparation – data cleaning and data labelling’[254] ‘Data preprocessing’, ‘Data Augmentation’[255] ‘Preparation phase’[256] ‘Data preparation and cleansing’, ‘Data transformation’[257]
Training	‘Model development‘[258] ‘Model training and evaluation‘[259] ‘Model training & inference – in-ML‘[260] ‘Build initial AI model‘, ‘Data augmentation‘, and ‘Build multiple AI models‘[261]
Evaluation	‘Model evaluation‘[262] ‘Model evaluation‘[263] ‘Model management – Model storage, versioning, query, and diagnosis‘[264] ‘Evaluate primary metrics‘ and ‘Evaluate secondary metrics‘[265] ‘Use, re(use) and feedback‘[266]
Deployment	‘Model deployment and monitoring‘[267] ‘Model deployment and inference‘[268] ‘Model deployment and serving‘[269] ‘AI Model deployment and risk assessment‘, ‘Post-deployment review’[270] ‘Use, re(use) and feedback‘[271] ‘Data and production orchestration‘[272]

[252] ICO (n 234).
[253] Khan et al. (n 236).
[254] Chai et al. (n 235).
[255] De Silva et al. (n 57).
[256] Shah et al. (n 238).
[257] Cognilytica (n 244).
[258] ICO (n 234).
[259] Khan et al. (n 236).
[260] Chai et al. (n 235).
[261] De Silva et al. (n 57).
[262] ICO (n 234).
[263] Khan et al. (n 236).
[264] Chai et al. (n 235).
[265] De Silva et al. (n 57).
[266] Shah et al. (n 238).
[267] ICO (n 234).
[268] Khan et al. (n 236).
[269] Chai et al. (n 235).
[270] De Silva et al. (n 57).
[271] Shah et al. (n 238).
[272] Cognilytica (n 244).

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 is 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 impact on model performance and fairness.

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

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

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

7. **Deployment:** Finally, the deployment stage is implicitly validated by the paper through discussions on the ecological validity of AI/ML

research and the impact of deployed models on society. It suggests a need for models to be tested in diverse, real-world settings to ensure they perform ethically and effectively outside controlled environments.

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

Syntax and search strategy

The syntax used for the search strategy is as follows:

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

AND ('data' AND 'governance' OR 'data' AND 'management' OR 'data'

AND 'ecosystem*' OR 'data' AND 'ethic*' OR 'data' AND 'quality' OR 'data' AND 'security' OR 'data' AND 'access' OR 'data' AND 'collection' OR 'data' AND 'acquisition' OR 'data' AND 'curation' OR 'data' AND 'wrangling' OR 'data' AND 'mining' OR 'data' AND 'preprocessing')

AND ('process*' OR 'pipeline*' OR 'practic*' OR 'best' AND 'practice*'

OR 'guid*' OR 'knowledge' AND 'graph')

AND NOT ('metaverse' OR 'iot' OR 'internet' AND 'of' AND 'things'

OR 'quantum' AND 'computing' OR 'nanotechnology' OR 'edge' AND 'comput*')

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

1. DCAI as a subject area, and as a movement galvanised by Andrew Ng, has taken shape since 2021[273] [274] [275] [276] [277] [278] [279] [280]. From this point, the notion of shifting away from purely model-centric approaches to AI/ML, and towards enhancing the quality of data, has grown, becoming more visible to researchers and practitioners.

2. It is well documented that from 2020, AI/ML systems' pace of development has accelerated substantially[281] [282]. Innovation policy literature suggests that rapid technological change can disrupt existing innovation processes and create a 'race into the unknown', increasing uncertainties and

[273] IAP (n 8).
[274] Jakubik, J., Vössing, M., Kühl, N., Walk, J., Satzger, G. (2024), 'Data-Centric Artificial Intelligence'.
[275] Datta, S. (2023), 'Potential Impact of Data-Centric AI on Society'.
[276] Patel et al. (n 39).
[277] Zha et al. (n 6).
[278] Brown, S. (2022), 'Why it's time for 'data-centric artificial intelligence'.
[279] Strickland, E., (2022), 'Andrew Ng: Unbiggen AI - IEEE Spectrum'.
[280] Ng, A., (2021), 'Data-Centric AI Competition'.
[281] Chow, A., Perrigo, B. (2023), 'The AI Arms Race Is On. Start Worrying'.
[282] McKendrick, J. (2021), 'AI Adoption Skyrocketed Over the Last 18 Months'.

dependencies[283] [284]. 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, including 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 (not discounting relevant evidence from mid-late 2019, for example).

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.

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 academic databases by title, abstract and keywords. The second involved scanning the whole document from this shortlist for a deeper review of its relevance[285].

For this second phase, we thematically coded literature according to three coding frameworks relating to 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:

[283] 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'.
[284] EEA (2015), 'Global megatrends update: 4 Accelerating technological change'.
[285] Collins, A., Coughlin, D., Miller, J., Kirk, S. (2015), 'The production of quick scoping reviews and rapid evidence assessments'.

Table 2. ‘Lifecycle’ coding framework

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

Table 3. ‘Ecosystem’ coding framework

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

Table 4. ‘Governance’ coding framework

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

Phase 1 - Title, abstract and keywords

Inclusion criteria

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

Exclusion criteria

- **Inversion of conceptual framework:** Special attention is paid to excluding articles that only discuss the use of AI/ML technologies in data science pipelines, inverting the conceptual framework outlined in the methodology (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 ensures that selected articles offer socio-technical understanding of data practices and processes in AI contexts.
- **Use case focus:** Articles focused on very specific technical use cases, without providing useful narrative about data governance per se, are excluded. This 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. A hand search[286] (manually scanning reference lists and relevant academic journals for further literature) was conducted to supplement the search results from the databases.

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

Results

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

Phase 1 – Title, abstract and keywords

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

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

Phase 2 – Full text

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

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

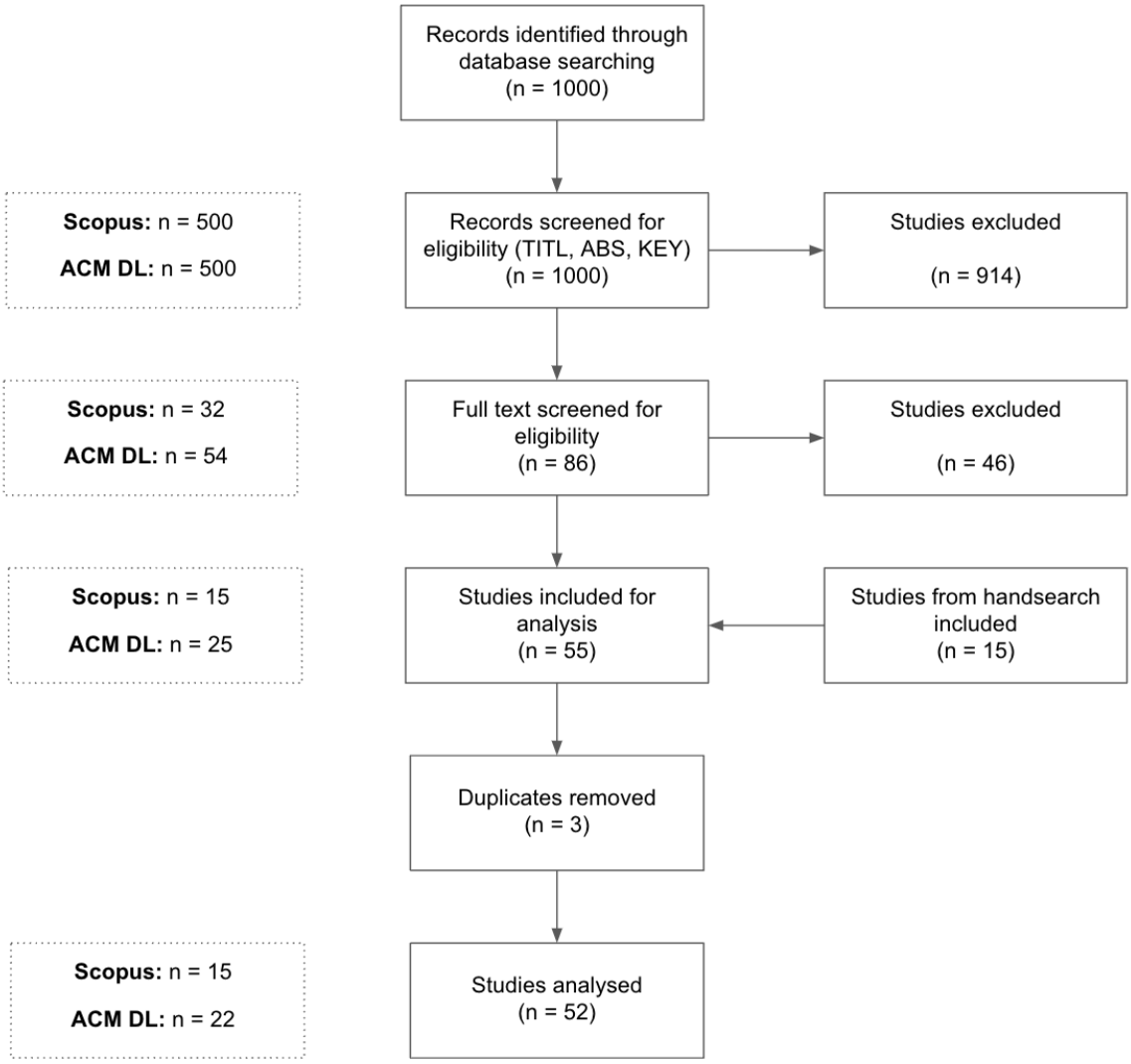

Figure 2. Diagram of scoping review results

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. Notable contributions included Aldoseri et al (2023) in 'Applied Sciences' discussing the rethinking of data strategy for AI[287], Alzubaidi et al (2021) in the 'Journal of Big Data' providing a review of deep learning concepts[288], and Chicco et al (2022) in 'PLOS Computational Biology' offering tips for data cleaning and feature engineering[289].

[287] Aldoseri et al. (n 34).
[288] 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'.
[289] Chicco et al. (n 89).

Complementing the journal articles, we uncovered 20 conference papers that provided 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), explored deep transfer learning, underscoring the importance of adaptable AI models[290].
- 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[291].
- 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[292] [293].
- Heger et al (2022), at the ACM Conference on Human Factors in Computing Systems (CHI), highlighted the importance of data documentation, pointing to the human-centred aspects of AI development[294].

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.

[290] 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'.
[291] Gowda et al. (n 142).
[292] Grafberger et al. (n 126).
[293] Li et al. (n 25).
[294] Heger et al. (n 111).

Annex B. Scope and limitations

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

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