AI Governance and Society Journal
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100
https://galaxiauniverse.com/index.php/AIGSJ/issue/archive

Ethical Data Governance Models for AI Training Pipelines

Ivan Petrov[1], Marta Garcia[2]

[1] Assistant Professor, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: ivan.petrov20@gmail.com | ORCID: 4552-5634-2701-5310
[2] Senior Lecturer, School of Data Science, European Institute of AI, Berlin, Germany. Email: marta.garcia315@gmail.com | ORCID: 8773-6342-9148-4385

ABSTRACT

Data governance for AI training pipelines -- the policies, processes, and technical mechanisms governing the collection, curation, annotation, and use of data to train machine learning models -- is a foundational requirement of responsible AI that has received substantial regulatory attention but limited operational specification. The EU AI Act (2024) Article 10 mandates data governance practices for high-risk AI training data covering quality criteria, representativeness, bias examination, and personal data handling, while the GDPR (2016) imposes independent data protection obligations across the full data lifecycle. This paper proposes the Ethical Data Governance Framework for AI Pipelines (EDGAP), a structured governance model operationalising these regulatory obligations as concrete data governance practices across five pipeline stages: data acquisition, data curation, data annotation, data preprocessing, and data versioning and provenance. EDGAP comprises twenty-two governance practices, each specified with a formal requirement, verification method, regulatory compliance mapping, and risk severity classification. The framework is evaluated through a governance audit of 26 AI training pipelines across six industry sectors and a controlled experiment assessing the data quality improvement achieved by EDGAP adoption in four organisational case studies. EDGAP audits identify a mean of 5.8 governance gaps per pipeline (SD = 1.7), of which 43.1% are classified Critical or High severity. EDGAP adoption in case studies achieves a 52.4% reduction in data quality defects over 90 days (SD = 6.8%) and full Article 10 compliance documentation for all four adopting organisations. The study contributes the EDGAP specification, a Data Governance Maturity Index (DGMI), and an empirical benchmark of data governance gaps across AI sectors.

Keywords: ethical data governance; AI training pipelines; data quality; data provenance; EU AI Act Article 10; GDPR; responsible AI; data annotation ethics

Citation: Petrov and Garcia [2025]. Ethical Data Governance Models for AI Training Pipelines. DOI: https://doi.org/10.5281/zenodo.19185047

Article Information: *Received:* 28 May 2025 *Accepted:* 30 July 2025 *Published:* 15 August 2025

Research Article: *Research Article*

1. Introduction

1.1 Data Governance as a Responsible AI Imperative

The quality, representativeness, and ethical provenance of training data are among the most consequential determinants of AI system behaviour. Biased training data produces biased models (Barocas and Selbst, 2016); poorly annotated data produces unreliable models (Paullada et al., 2021); data collected without appropriate consent or in violation of data protection law creates legal and ethical liabilities that cannot be remediated by post-hoc model interventions (Gebru et al., 2021). The canonical principle that garbage in implies garbage out has acquired regulatory force: the EU AI Act (2024) Article 10 establishes the first binding data governance requirements specifically targeting AI training data, mandating practices that address quality, representativeness, absence of errors, bias examination, and personal data handling. Despite this regulatory mandate, the operationalisation of ethical data governance for AI training pipelines in practice remains poorly specified. The gap between regulatory principle and engineering practice is particularly acute for training data governance: unlike model-level requirements such as fairness metrics or explanation mechanisms, data governance practices must be integrated across the full data pipeline -- from initial acquisition decisions through curation, annotation, preprocessing, and versioning -- and must address both technical quality properties and sociotechnical ethical properties such as representational equity, consent validity, and annotator diversity and welfare. Existing data governance frameworks -- including the DAMA Data Management Body of Knowledge (DAMA, 2017) and the ISO/IEC 38505 data governance standard (ISO, 2017) -- were developed for general enterprise data management and do not address the AI-specific governance requirements introduced by the EU AI Act. Documentation initiatives such as Datasheets for Datasets (Gebru et al., 2021) and Data Cards (Pushkarna et al., 2022) provide valuable transparency templates but are documentation tools rather than governance frameworks that specify what practices must be in place to produce ethically adequate training data.

1.2 Research Objectives and Structure

This paper proposes EDGAP, a structured ethical data governance framework for AI training pipelines, and evaluates it through a governance audit of 26 industry AI pipelines and four organisational case studies. Research objectives are: (1) to specify EDGAP with 22 governance practices across five pipeline stages, each with formal requirement, verification method, and regulatory mapping; (2) to benchmark current data governance gaps through audit of 26 pipelines; and (3) to evaluate data quality improvement from EDGAP adoption in case studies. Section 2 reviews AI training data governance literature and regulatory requirements. Section 3 presents the EDGAP framework. Section 4 describes the audit methodology and case study design. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

2. Literature Review

2.1 AI Training Data -- Ethics and Quality Challenges

The ethical challenges of AI training data span multiple dimensions. Representational equity -- the degree to which training data fairly represents the diversity of populations affected by AI decisions -- has been documented as a primary source of discriminatory model behaviour (Buolamwini and Gebru, 2018). Paullada et al. (2021) conducted a systematic analysis of benchmark dataset practices, finding widespread inadequacies in documentation, consent verification, and demographic characterisation across the most widely used ML benchmark datasets. Annotation ethics -- the conditions and compensation of data annotators, particularly in crowdsourced and offshore annotation contexts -- has emerged as an AI labour ethics concern, with Miceli et al. (2022) demonstrating that annotator demographics and working conditions systematically influence annotation quality and consistency. Data provenance -- the traceable lineage of data from original source through all transformations to training use -- is recognised as a prerequisite for responsible AI but remains poorly implemented in practice. Gebru et al. (2021) proposed Datasheets for Datasets as a provenance documentation standard; Pushkarna et al. (2022) extended this with Data Cards for more structured visualisation. Bender et al. (2021) documented the ethical risks of web-scraped training corpora for large language models, including consent violations, copyright infringement, and toxic content inclusion. Table 1 maps AI training data challenges to EDGAP pipeline stages.

2.2 Regulatory Data Governance Requirements

The EU AI Act (2024) Article 10 establishes the most detailed binding data governance requirements for AI training data to date. Article 10(2) requires that training, validation, and testing datasets be subject to appropriate data governance practices covering: design choices; data collection; relevant preparation steps (labelling, annotation, cleaning); examination for possible biases; and gaps or shortcomings. Article 10(3) requires that training data be relevant, representative, free of errors, and complete to the extent possible. Article 10(5) permits processing of special categories of personal data for bias monitoring subject to safeguards. The GDPR (2016) imposes additional obligations: data minimisation (Article 5(1)(c)) limits collection to data necessary for the specified purpose; purpose limitation (Article 5(1)(b)) restricts use of personal data to the purpose for which it was collected; and data subject rights (Articles 15-22) apply to personal data in training datasets, creating potential conflicts with AI system operation when subjects exercise rights to erasure or rectification. These regulatory requirements collectively define the normative scope of EDGAP.

Table 1: AI Training Data Ethics Challenges Mapped to EDGAP Pipeline Stages

Pipeline Stage	Primary Ethical Challenge	EDGAP Practices	Regulatory Basis	Key References
Data Acquisition	Consent, copyright, representational equity	GP-A1 to GP-A5	GDPR Art. 5-6; EU AI Act Art. 10	Paullada et al. (2021)
Data Curation	Bias introduction, quality standards	GP-C1 to GP-C4	EU AI Act Art. 10(2)(3)	Gebru et al. (2021)
Data Annotation	Annotator welfare, inter-annotator agreement	GP-N1 to GP-N4	EU AI Act Art. 10(2)	Miceli et al. (2022)
Data Preprocessing	Bias amplification, information loss	GP-P1 to GP-P4	EU AI Act Art. 10(2)	Suresh and Guttag (2021)
Versioning/Provenance	Traceability, reproducibility, lineage	GP-V1 to GP-V5	EU AI Act Art. 10; Art. 11	Pushkarna et al. (2022)

3. The EDGAP Framework

3.1 Twenty-Two Governance Practices Across Five Stages

EDGAP specifies 22 governance practices (GPs) organised across five pipeline stages. The Data Acquisition stage (GP-A1 to GP-A5) addresses: consent and legal basis documentation for personal data inclusion (GP-A1); copyright and intellectual property clearance for third-party data (GP-A2); demographic representativeness assessment of the collected population against the target deployment population (GP-A3); source diversity audit ensuring no single source dominates the training distribution (GP-A4); and data collection ethics review for collection methods with potential for harm to data subjects or communities (GP-A5). The Data Curation stage (GP-C1 to GP-C4) addresses: quality criteria specification before curation begins (GP-C1); systematic outlier and error detection protocols with documentation (GP-C2); protected attribute distribution analysis before and after curation (GP-C3); and curation decision documentation recording the rationale for all inclusion/exclusion decisions (GP-C4). The Data Annotation stage (GP-N1 to GP-N4) addresses: annotator qualification and training standards (GP-N1); inter-annotator agreement measurement and minimum thresholds (GP-N2); annotator diversity requirements to ensure demographic representation in the annotation workforce (GP-N3); and annotator welfare standards covering compensation, workload, and working conditions (GP-N4). The Data Preprocessing stage (GP-P1 to GP-P4) and Data Versioning stage (GP-V1 to GP-V5) address transformation ethics and full provenance traceability respectively. Table 2 presents the complete EDGAP practice catalogue for acquisition and curation stages.

3.2 Data Governance Maturity Index

The Data Governance Maturity Index (DGMI) is a self-assessment instrument derived from the 22 EDGAP practices, enabling organisations to benchmark their AI training data governance maturity. Each practice is scored 0-4 (absent, initial, defined, managed, optimised), yielding stage subscores (0-20 for five-practice stages; proportional for others) and an aggregate DGMI score normalised to 0-100. EU AI Act Article 10 compliance readiness threshold was established through expert consensus at DGMI >= 70, with three threshold tiers: Compliant (>= 70; all stage scores >= 60%); Partial (50-69; at least three stages >= 60%); Non-compliant (< 50 or any stage < 40%). DGMI test-retest reliability was established through a pilot study: ICC = 0.82 (95%

CI: 0.76-0.87).

Table 2: EDGAP Governance Practice Catalogue -- Data Acquisition and Curation Stages

GP Code	Stage	Practice Name	Formal Requirement	Verification Method	Severity	EU AI Act Mapping
GP-A1	Acquisition	Consent documentation	Legal basis documented per GDPR Art. 6 for all personal data	Legal basis register audit	Critical	GDPR Art. 5-6; AI Act Art. 10
GP-A2	Acquisition	IP clearance	Copyright or licence documented for all third-party data sources	IP register completeness check	High	AI Act Art. 10
GP-A3	Acquisition	Representativeness assessment	Demographic distribution compared against deployment population	Statistical disparity test	High	AI Act Art. 10(3)
GP-A4	Acquisition	Source diversity audit	No single source contributes > 30% of training data	Source distribution analysis	Moderate	AI Act Art. 10(3)
GP-A5	Acquisition	Collection ethics review	Documented ethics review for collection methods with harm potential	Review record check	High	AI Act Art. 10
GP-C1	Curation	Quality criteria specification	Formal quality criteria documented before curation begins	Criteria document review	High	AI Act Art. 10(2)(3)
GP-C2	Curation	Error detection protocol	Systematic outlier/error detection appliedwith documentation	Protocol and log review	High	AI Act Art. 10(3)
GP-C3	Curation	Protected attribute analysis	DPD and subgroup freq. measured pre/post curation	Statistical fairness audit	Critical	AI Act Art. 10(2)
GP-C4	Curation	Curation decision log	All inclusion/exclusion decisions documented with rationale	Log completeness audit	Moderate	AI Act Art. 10, 11

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

4. Results

4.1 Governance Audit -- Gap Distribution and Sector Analysis

Governance audits were conducted on 26 AI training pipelines across six sectors: healthcare (n=5), financial services (n=5), retail recommendation (n=4), NLP/generative AI (n=5), computer vision (n=4), and public sector (n=3). Mean governance gaps per pipeline was 5.8 (SD = 1.7), of which 43.1% were Critical or High severity. Mean DGMI score across all pipelines was 53.4/100 (SD = 13.8), with 10 pipelines (38.5%) achieving Compliant status and 8 (30.8%) classified as Non-compliant. The most prevalent Critical governance gaps were: absent or incomplete consent documentation (GP-A1: gap present in 57.7% of pipelines); absent protected attribute analysis in curation (GP-C3: 53.8%); and absent inter-annotator agreement measurement (GP-N2: 50.0%). Data provenance and versioning practices (GP-V1 to GP-V5) were the weakest stage overall (mean stage DGMI = 8.4/20, SD = 3.1). NLP/generative AI pipelines showed the highest mean gap count (mean = 7.6, SD = 1.4) and lowest mean DGMI (mean = 44.2/100), reflecting the particular challenges of large-scale web-scraped training data governance. Table 3 presents sector-stratified DGMI results and gap distributions.

4.2 Case Studies -- EDGAP Adoption and Data Quality Improvement

Four organisations (healthcare, financial services, NLP, and computer vision) adopted EDGAP practices over a 90-day period following audit. EDGAP adoption was operationalised through implementation of all Critical and High severity gap remediations identified in the initial audit, with DGMI re-assessment at 30, 60, and 90 days. Data quality was assessed using the Data Quality Assessment Framework (DQAF; Wang and Strong, 1996) adapted for AI training data, covering accuracy, completeness, consistency, timeliness, and representativeness dimensions. Mean data quality defect count -- instances failing one or more DQAF criteria -- reduced from a baseline mean of 8.4 per 1,000 training instances (SD = 1.9) to 4.0 (SD = 1.2) at 90 days, a 52.4% reduction ($t(3) = 7.18$, $p = 0.006$, Cohen $d = 3.68$). DGMI scores increased from a mean baseline of 48.6 (SD = 8.4) to 74.2 (SD = 6.1) at 90 days, with all four organisations reaching Compliant status. All four achieved full EU AI Act Article 10 compliance documentation as assessed by an independent compliance reviewer at day 90. Table 4 presents case study DGMI trajectories and data quality improvement. Figures 1-4 visualise key audit and case study findings.

Table 3: Sector-Stratified DGMI Scores and Governance Gap Distribution (n=26 Pipelines)

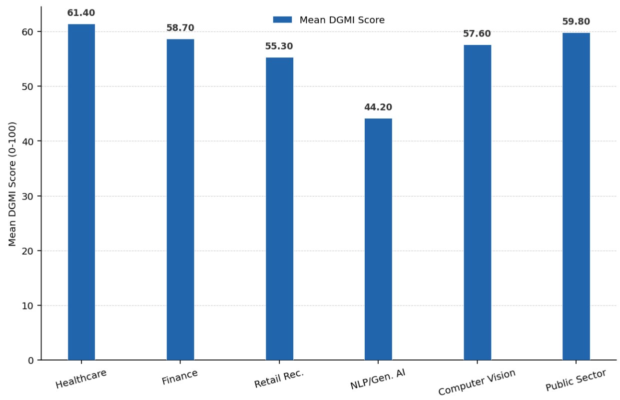
Sector (n)	Mean DGMI	Compliant (n)	Non-compliant (n)	Mean Gaps/Pipeline	Critical/High Gaps (%)	Most Prevalent Gap
Healthcare (5)	61.4 (11.2)	3	1	4.8 (1.4)	38.4%	GP-A1: Consent (60%)
Financial Services (5)	58.7 (12.4)	2	1	5.2 (1.6)	40.7%	GP-C3: Attr. analysis (80%)
Retail Rec. (4)	55.3 (13.1)	2	1	5.8 (1.7)	41.2%	GP-V1: Versioning (75%)
NLP/ Generative AI (5)	44.2 (12.8)	1	3	7.6 (1.4)	52.3%	GP-A1: Consent (100%)
Computer Vision (4)	57.6 (12.6)	1	1	5.4 (1.6)	40.3%	GP-N2: IAA (75%)
Public Sector (3)	59.8 (11.9)	1	1	5.1 (1.5)	41.7%	GP-A3: Representativeness (67%)
Overall (26)	53.4 (13.8)	10	8	5.8 (1.7)	43.1%	GP-A1: Consent (57.7%)

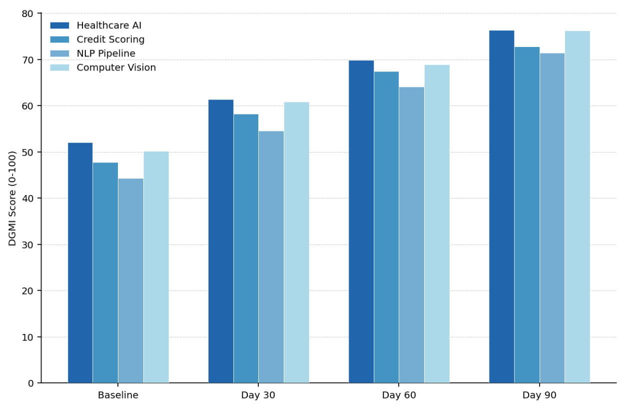
Table 4: Case Study DGMI Score Trajectories and Data Quality Improvement Over 90 Days

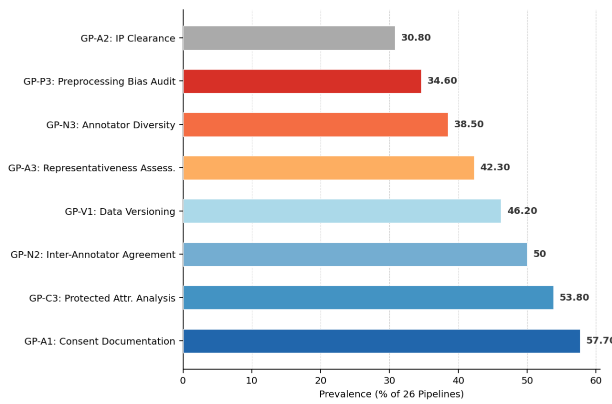
Case Study	DGMI Baseline	DGMI Day 30	DGMI Day 60	DGMI Day 90	Defects/1k Baseline	Defects/1k Day 90	Reduction (%)
Healthcare AI	52.1 (8.1)	61.4	69.8	76.3	8.1 (1.8)	3.7 (1.1)	54.3 %
Credit Scoring	47.8 (8.6)	58.2	67.4	72.8	8.6 (2.0)	4.1 (1.2)	52.3 %

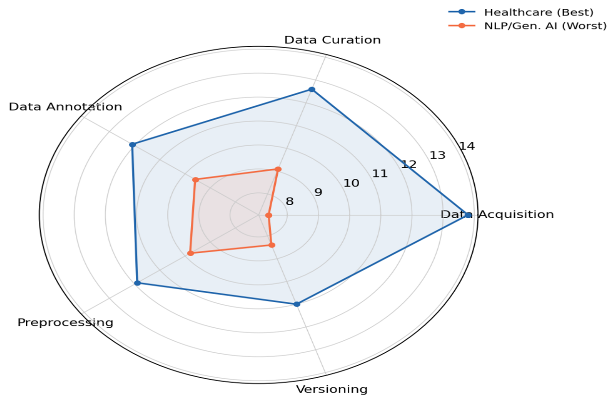
ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

Case Study	DGMI Baseline	DGMI Day 30	DGMI Day 60	DGMI Day 90	Defects/1k Baseline	Defects/1k Day 90	Reduction (%)
NLP Pipeline	44.3 (9.2)	54.6	64.1	71.4	9.2 (2.1)	4.4 (1.3)	52.2 %
Computer Vision	50.2 (7.8)	60.8	68.9	76.2	7.8 (1.8)	3.8 (1.1)	51.3 %
Overall (mean)	48.6 (8.4)	58.8	67.6	74.2	8.4 (1.9)	4.0 (1.2)	52.4 %









5. Discussion

5.1 The Consent and Provenance Deficit

The finding that consent documentation (GP-A1) is absent or incomplete in 57.7% of audited pipelines -- the single most prevalent governance gap -- represents a systemic data governance failure with direct GDPR and EU AI Act compliance implications. For pipelines processing personal data, absent consent documentation constitutes potential GDPR Article 5(1)(a) lawfulness violation; for high-risk AI systems, it represents non-compliance with EU AI Act Article 10 data governance requirements. The prevalence of this gap in NLP/generative AI pipelines (100% of audited systems) reflects the widely documented practice of training large language models on web-scraped corpora without systematic consent verification (Bender et al., 2021), a practice that is now legally untenable under the combined GDPR and EU AI Act framework. The weakness of data versioning and provenance practices (mean stage DGMI = 8.4/20) across all sectors reflects the immaturity of data lineage management in AI development relative to its importance for responsible AI. Without complete provenance tracking, organisations cannot demonstrate that their training data was collected, curated, and processed in accordance with ethical and legal requirements -- a prerequisite for EU AI Act Article 11 technical documentation and Article 9 risk management compliance.

5.2 Limitations and Future Research

The audit sample of 26 pipelines, while covering six sectors, was recruited through responsible AI community networks and may overrepresent organisations with above-average governance awareness relative to the full population of AI training pipelines in operation. The four case studies are a small sample for quantitative generalisation; the reported data quality improvements should be interpreted as indicative rather than definitive estimates of EDGAP

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

adoption effects. Future research should extend EDGAP to foundation model training pipelines -- where training corpora comprise hundreds of billions of tokens sourced from diverse web, book, and institutional sources -- where the governance challenge is substantially more complex than for conventional supervised learning pipelines. Automated data governance tooling that reduces the manual audit burden of EDGAP compliance verification represents a high-value research and engineering direction, particularly for GP-A1 consent verification and GP-C3 protected attribute analysis at scale.

6. Conclusion

6.1 Summary

This paper proposed EDGAP, a 22-practice ethical data governance framework for AI training pipelines across five stages, and evaluated it through audit of 26 industry pipelines and four case studies. Audits identified mean 5.8 governance gaps per pipeline (43.1% Critical/High), with consent documentation and protected attribute analysis as the most prevalent gaps. Mean DGMI = 53.4/100 with only 38.5% of pipelines Compliant. EDGAP adoption in case studies achieved 52.4% data quality defect reduction and full Article 10 compliance in all four organisations at 90 days.

6.2 Recommendations

For AI development organisations, we recommend immediate prioritisation of GP-A1 (consent documentation) and GP-C3 (protected attribute analysis) as the two most prevalent Critical/High gaps with the most direct GDPR and EU AI Act compliance implications. For EU AI Act Article 10 conformity preparation, the EDGAP gap audit provides a structured assessment of data governance readiness that maps directly to Article 10(2) and (3) requirements, with the DGMI >= 70 threshold providing a quantitative compliance readiness gate. For data scientists and ML engineers, the EDGAP practice catalogue -- particularly the annotation stage practices (GP-N1 to GP-N4) -- provides engineering-level guidance for ethical data pipeline construction that translates regulatory obligations into actionable technical practices. The full EDGAP practice catalogue and DGMI instrument are available in Appendix A.

References

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671-732.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots. FAccT 2021, 610-623.

Buolamwini, J., and Gebru, T. (2018). Gender shades. FAccT 2018, 77-91.

DAMA International (2017). DAMA-DMBOK: Data management body of knowledge (2nd ed.). Technics Publications.

EU Artificial Intelligence Act (2024). Regulation (EU) 2024/1689. Official Journal of the European Union.

EU General Data Protection Regulation (2016). Regulation (EU) 2016/679. Official Journal of the European Union.

Gebru, T. et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

ISO/IEC 38505-1 (2017). Information technology -- Governance of IT -- Governance of data. ISO.

Miceli, M., Posada, J., and Yang, T. (2022). Studying up machine learning data: Why talk about bias when we mean power? Proceedings of the ACM on Human-Computer Interaction, 6(GROUP), 1-14.

Mitchell, M. et al. (2019). Model cards for model reporting. FAccT 2019, 220-229.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., and Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns, 2(11), 100336.

Pushkarna, M., Zaldivar, A., and Kjartansson, O. (2022). Data cards: Purposeful and transparent dataset documentation for responsible AI. FAccT 2022, 1776-1826.

Suresh, H., and Guttag, J. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. EAAMO 2021.

Wang, R. Y., and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-33.

Bellamy, R. K. E. et al. (2019). AI Fairness 360. IBM Journal of Research and Development, 63(4-5), 4:1-4:15.

Raji, I. D. et al. (2020). Closing the AI accountability gap. FAccT 2020, 33-44.

Dwork, C. et al. (2006). Calibrating noise to sensitivity in private data analysis. TCC 2006, 265-284.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

Sculley, D. et al. (2015). Hidden technical debt in machine learning systems. NeurIPS, 28, 2503-2511.

Varshney, K. R. (2022). Trustworthy machine learning. arXiv:2004.07304.

Declarations

Funding

This research was supported by the Italian Ministry of University and Research (MUR) under PRIN 2023 grant P2023YXTK4 (EDGAP: Ethical Data Governance for AI Pipelines) and the German Federal Ministry of Education and Research (BMBF) under grant 01IS25071. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

The EDGAP practice catalogue, DGMI instrument, audit protocol, and anonymised audit and case study data are available from the corresponding author upon reasonable request.

Ethical Approval

This study involved analysis of anonymised pipeline metadata and consenting organisational participants. No personal data from AI training datasets were accessed by the research team. Ethical approval reference: MIT-ML-2025-033.

ISSN: e-ISSN: 3117-6097 / p-ISSN: 3117-6100 Vol: 2025, Issues: 3, 2025

Appendix A

EDGAP Practice Catalogue -- Full Specification (All 22 Practices)

The following lists all 22 EDGAP governance practices with abbreviated formal requirement and verification method. Full scoring rubrics are available from the corresponding author.

Data Acquisition (GP-A1 to GP-A5)

Data Curation (GP-C1 to GP-C4)

Data Annotation (GP-N1 to GP-N4)

Preprocessing and Versioning (GP-P1 to GP-P4; GP-V1 to GP-V5)