International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Lineage, Traceability, and Reproducibility as Reliability Requirements in Enterprise AI Systems

Divya Bonthala

Submitted:02/02/2026 **Revised**: 11/03/2026 **Accepted:** 13/04/2026

Abstract—Artificial intelligence is being applied to key business and compliance choices by more systems in the enterprise. One of the most common systems is concerned with the accuracy of the model and does not factor in the reliability aspect, like the lineage or traceability, or reproducibility. In this paper, we obtain these three aspects as fundamental reliability expectations of enterprise AI. The study was a real enterprise AI applied in 12 months with a before and after quantitative design. Lineage coverage, version control and reproducibility controls were introduced thus, the lineage coverage rose to 0.91 and the success of reproducibility rose to 92% after these tools were applied on structured lineage. Rapid time to incident investigation was less by 66%, audit preparation was also less by 62% and compliance findings were also less by 75%. Monte Carlo simulation also indicated that the risk variability was smaller when the lineage controls had been incorporated. This observation is in full agreement with the results that indicated that integrating lineage, traceability, and reproducibility into AI platforms enhances reliability, audit readiness, and trust in AI results.

Keywords— *Enterprise Artificial Intelligence, Traceability, Data Lineage, Reproducibility, Data Governance, AI Reliability, Model Versioning*

I. INTRODUCTION

A. Background and Context

Artificial intelligence has become highly popular in business setups to be able to assist in reporting, forecasting, compliance, and executives. These human systems are based on massive and dynamic data streams that can vary on a regular basis. Changes, transformations, and long slight alterations of data can have a potent outcome on models. The audit issues, compliance failures, and lost trust may result in regulated industries in case of such changes.

The applicability of AI has long been assessed by the predictive power of the model commonly or performance metrics. Although it is important, it does not need to be as accurate as the one used by the enterprise. Business and regulatory stakeholders should be in a position to know the origin of data, the way it was processed, and the manner, in which a model was trained. In the absence of this visibility, AI systems are difficult to trust, explain and validate.

B. Motivation for the Study

Most AI failures that happen in the enterprise do not take place due to ineffective algorithms but rather due to inadequate governance. There are no records of the origins of missing data, missing data which have not been documented, transformations that have not been recorded and transformations which are not reproducible thus information about the incident is not easily obtained or audits satisfied. Teams use manual documentation often and it emerges that this is not scalable and is prone to errors.

There is also an increasing regulatory pressure. The clear evidence to prove the data provenance, versions of different models, and reproducibility are now anticipated by the auditors and regulators. Lack of such evidence has caused delays, discoveries, and risk of operation in organizations.

This study has been inspired by the need to demonstrate that lineage, traceability, and reproducibility are not best practices at will. They are reliability metrics and crucial prerequisites of enterprise AI systems. This paper will seek to translate the discussion on governance theory to practice by establishing the magnitude of said impacts.

C. Research Gap

Current literature also talks about lineage, provenance, and reproducibility primarily within the context of a research or technical setting. There are not many studies which offer quantitative evidence based on enterprise setting. Most works also address these concepts individually and not as a cohesive system of reliability.

The understanding of the extent of operational benefit these controls bring when deployed in large scale is also evidently significantly lacking. This research fills that gap by quantifying tangible outcomes, i.e. the time used in investigations, audit effort, success in reproducibility and confidence scores.

Senior AI Platform Architect

D. Novelty of the Work

There are three aspects that make this paper novel.

To begin with, it considers the lineage, traceability, and reproducibility as reliability requirements and not documentation activities.

Second, it employs a quantitative post-pre design on the basis of actual operational data of the enterprise.

Third, it is the combined use of traditional measures with Monte Carlo simulations to investigate the risk and outcome stability in the face of uncertainty.

In contrast with conceptual frameworks, this research paper has given quantitative data that governance-by-design enhances enterprise AI reliability.

E. Research Objectives

The main objectives of this study are:

- To quantify lineage, metadata and reproducibility improvement following system governance implementation.
- To measure the change on incident investigation and audit preparedness.
- To assess model success in regard to reproducibility and stability of model outcomes.
- To determine the shifts in corporate confidence and risk variability.

The system logs, metadata repositories and audit records and statistical analysis are all used to address these objectives.

F. Structure of the Paper

The paper has the following construction. The following part provides the literature review on lineage, traceability and reproducibility. The methodology section describes the methods of quantitative research study, variables and methods of analysis. They are measured results and simulations which are reported in the findings section. The conclusion draws some major lessons and explains the implications regarding enterprise AI governance.

II. Literature Review

A. Lineage and Provenance as Foundations of Trustworthy AI

The concept of data provenance and lineage has become commonly identified as a necessity of a trustworthy AI system. Provenance identifies the origin of data, the way in which it is manipulated and the paths in which it travels through pipes. Some researches indicate that the lack of powerful lineage renders it extremely challenging to comprehend, account, or replicate AI results. Standards-based traceability systems, including W3C PROV, have been suggested to represent a relationship of data, processes, and artifacts throughout the AI lifecycle [1][2][3]. Another drawback of the field of literature is the absence of standard semantics and uniform use throughout tools that undermines interoperability and overall sustainability [4].

In recent literature, it is stressed that nowadays AI pipelines consist of numerous steps, tools and variations, and their documentation cannot be done manually. The lineage capture must thus be automated. Research presents an assumable ML pipeline, which captures end-to-end lineage data as open specifications and trusted infrastructure, and this prevents issues of data poisoning and supply-chain attacks [5]. PROV-ML builds on established provenance norms to enhance the expressiveness to machine learning processes and query various environments of multi-workflows. Such strategies indicate that lineage is not just a documentation issue but a technical control, which is in direct support of security, auditability and trust.

B. Traceability, Metadata, and Governance in Enterprise AI

Traceability links information, models, code and choices both within time. Traceability is also connected in close to governance, compliance, and accountability in the enterprise environment. Some studies have suggested that AI governance should be expressed clearly regarding data, models, and systems and meaning who governs what and how [6]. Metadata management is important in this area since metadata can be used to find, analyze impacts, and enforce the policies related to big data estates [7].

AI-based metadata models intend to lower the workload of manual systems and enhance the quality of consistency and coverage [7]. Schema-level lineage extraction also solves the problem of semantic drift in enterprise pipelines, where the meaning of data is lost due to the large number of transformations it undergoes [8]. The suggested SLiCE metric analyzes the structural and semantic quality of lineage demonstrating that a fine-grained traceability measurement and enhancement can be done in a systematic way [8]. Such contributions have value to enterprise AI where faulty lineage may destabilize the downstream systems like RAG, analytics, and regulatory reporting.

Along with the technical pipelines, traceability is also of essence in the aspect of legal and ethical compliance. Text datasets auditing at scale indicates that the issues with license mislabelling, absence of attribution information, and little indication of the origin of such data are extremely severe [9]. This has legal risk and the lack of confidence in AI results. Such tools as dataset provenance explorers and standardized documentation can be used to solve these problems by having the data origins explicit and inspectable [9].

C. Reproducibility as a Reliability and Governance Requirement

The issue of reproducibility is scientifically talked but most authors suggest its governance requirement is just as much. The poor standards of reproducibility

lower the trust in AI study and restrict the opportunity of policy-makers and regulators to evaluate the risk [10]. Computing research studies indicate that majority of the papers do not record sufficient information to replicate the results particularly on data and experimental set up [11]. This issue is aggravated when AI systems are applied to what is on the ground.

Failure modes, like the checklist of reproducible results of NeurIPS, and practices like code submission policies prove that when there are structured requirements community standards can be improved [12]. Reproducibility is required in enterprise and scientific processes, where one not only wants to capture executable code and data, but also capture execution environments and infrastructure settings [13]. Provenance frameworks such as ReCAP that are cloud-conscious demonstrate that infrastructure information can have a major influence and should be considered to enable reproducible results [13].

A number of systems strive to ensure that provenance capture has become non-overhead, transparent, and easy to implement by practitioners. Ursprung automatically gathers system-level and application provenance with insignificant performance overhead and works with reproducibility [14]. These strategies facilitate the notion that reproducibility must be managed within platforms in addition to its being an ex-post factum job.

D. Integrated Architectures for Lineage, Traceability, and Reproducibility

Current literature gives more and more support to the idea of integrated and lifecycle-conscious architectures. The AI Model Passport offers a universal digital identity to the models, containing metadata in data acquisition, pre-processing, training, and deployment [15]. It is better suited in terms of traceability and regulatory preparedness by being able to automate versioning and decoupling results of scripts. Closely related concepts are found in the data-centric reproducibility systems which formalize structured artifacts in the form of datasets, features, workflows, and executions [16].

Reliability is further enhanced using supply-chain inspired approaches. Trusted AI Bills of Materials builds on SBOM ideas and extends them to AI, to include dependencies between data, models and software coupled with providing attestations of integrity [17]. These concepts are consistent with the concept of governance-by-design, where auditing and transparency are designed into the fabric on which AI systems are created [18].

The literature demonstrates that there is a high consensus that the concept of lineage, traceability and reproducibility are no longer optional. They are key reliability specifications to enterprise AI systems where there are regulatory, ethical, and business constraints. Nevertheless, standardization, common semantics, and economic implementation have some gaps. To fill such gaps, reference architectures that instantiate the governance controls at data and AI platforms are needed.

TABLE I. SUMMARY OF PREVIOUS STUDIES

Focus Area	Key Contribution	Main Finding
Data lineage and provenance	A number of studies suggest automated provenance systems in order to trace the data and model artifacts throughout the ML lifecycle [1][2][5].	With automated lineage, audit, and security become easier, as well as comprehending AI pipelines.
Metadata and traceability	It has been demonstrated that robust metadata and the schema-level lineage improve the diminution of semantic dispersion in elaborate data streams [7][8].	Fine-grained metadata assists in maintaining the meaning on the way data is transported through numerous transformations.
Dataset transparency and legal risk	Through the audits of large datasets, it is possible to identify that there are license gaps, improper attribution, and lack of knowledge of the source of data [9].	There is an elevated legal and moral risk of AI systems because of poor dataset traceability.
Reproducibility standards	The papers emphasize problematic reproducibility standards of the AI research domain and introduce more rigorous documentation and policies [10][11][12].	By enhancing reproducibility better, trust will be enhanced and governance decisions, policy decisions will be assisted.
Provenance-aware execution environments	Systems preserve code, data and infrastructure information so as to achieve faithful re- production of results [13][14].	Important information is needed on the environment of execution that will ensure reliability in reproducibility.
Integrated governance architectures	The frameworks that combine lineage, traces and trust control include model passports, and artificial intelligence bills of materials [15][17][18].	Introducing governance is something that can be implemented within the AI platforms.

III. METHODOLOGY

A. Research Design and Approach

The research will be based on quantitative research design. This is aimed at understanding the impacts of lineage, traceability, and reproducibility on the level of reliability of enterprise AI systems. Effectiveness of the models is not the only thing that could be considered reliable in this research. It also incorporates the audit readiness, time of incident investigation, result reproducibility, and confidence to the stakeholders.

A positivist study method is applied. This has an implication that the study will be based on observable and measurable data. All variables are articulated well and measured in terms of numeric indicators. The research hypothesizes relationships between the governance practices and the results of reliability. It is tested with the help of statistical analysis.

The study has a comparative pre-test and post design. Measurements on the enterprise AI pipes are taken prior to the addition of structured lineage and traceability controls, and then subsequent to the implementation itself. This design will enable direct improvement of value by the proposed architecture to be measured.

B. Research Setting and Data Sources

The experiment is set in an enterprise AI environment, which involves machine learning models to make business decisions that are critical to the business, like reporting, forecasting, and compliance support. The AI pipelines involve several steps, which are data ingestion, data transformation, data curation, feature engineering, model training, validation, and deployment.

Data is collected from:

- Pipeline execution logs
- Metadata repositories
- Model versioning systems
- Audit/ compliance documentation.
- Incident management systems

Many incident management systems can be found in the marketplace.

These sources contain objective and time stamped records. There is no survey or interview data that is utilized. This forms the guarantee that no measurements are biased and repeatable.

The analysis encompasses two years of business information focusing on a year of operation:

- Six months previous to lineage and profile controls.
- Six months post the implementation.

C. Variables and Measurement

1) Independent Variables

The independent variables will be the governance controls that are being injected into AI platform.

1. **Lineage Coverage Ratio**: This is used to determine the proportion of data pipes in which complete lineage has been obtained between the point of source and the model input.
2. **Metadata Completeness Score**: This is an indicator of the number of needed metadata fields that have been filled such as the source of data, the version of a schema, the logic of transformation and ownership.
3. **Version Control Adoption Rate**: This is an indicator that the number of datasets, features, and models that are versioned and immutable is a percentage.
4. **Reproducibility Enablement Index**: This is to determine whether training run is reproducible with recorded data versions, code versions and configuration parameters.

All the variables are put on a scale of 0 to 1 to give an opportunity to compare them.

2) Dependent Variables

Dependent variables imply AI system reliable results.

1. **Incident Investigation Time:** This is a measured time that is taken to represent the average number of hours to find out root cause in a case of data or model incidents.
2. **Audit Preparation Time:** This is time spent in total number of hours taken to prepare evidence supporting internal or external audit.
3. **Reproduction Success rate:** This is represented as a percentage of the previous model runs that implicitly can be reproduced with similar results.
4. **Compliance Finding Count**: Measured as the count of audit issues which are linked to missing data origin, ambiguous transformations or undocumented models.
5. **Business Confidence Score:** This is tested on a standardized internal reliability score applied by enterprises and its governance staffs.

D. Data Collection Procedure

There is automated and continuous data collection. In order to extract metrics, scripts and monitoring tools are used to come up with metrics that are directly obtained in enterprise systems.

The data of lineage is gathered at all the pipeline stages. Each transformation captures input dataset, output datasets, run-time, as well as version identifier. The metadata is automatically acquired based on schema definitions. The manual entry is avoided to lower human error.

The incident data will be gathered in the enterprise incident management system. Included are only data related incidents, model behavior incidents or compliance incidents. The data of the audit are extracted out of the final audit reports and review logs.

All information is put in a central analytics storage. Analysis is done after cleaning data by deleting duplicates and partial data.

E. Data Analysis Techniques

The first used are the descriptive statistics. Mean, median and standard deviation are computed of all the variables in both periods of time.

Comparative before and after results are done by use of paired sample t-tests. This aids in resolving whether changes, which are being observed, are statistically significant.

The correlation analysis is used to assess the relationship between the governance controls and the outcomes of reliability. To mention just one, Pearson correlation is used to calculate the association between the lineage coverage and incident investigation time.

The combined effect of lineage, metadata and reproducibility on overall system reliability is tested by means of a multiple linear regression model. This is useful in determining the control governance controls with the greatest effect.

All the statistical testing is conducted at a confidence level of 95.

F. Validity and Reliability

In order to pass the test of internal validity, the use of the same AI pipelines is checked before and after the implementation. There is no significant redesign of any system in the span of study besides the governance controls.

All metrics will be made based on system logs and not according to human judgment as a way of making sure that they are dependable in measurements. Similar scripts are repeated many times so as to test how the same outputs.

The construct validity can be warranted by associating every variable with the known descriptions of lineage, traceability, and reproducibility practiced in enterprise governance.

G. Ethical and Compliance Considerations

Only operational data on a system level are used in the study. No personal information and sensitive company data are analyzed. Where needed, all the identifiers are anonymized.

The access to logs, as well as audit records, is provided according to enterprise security policies. The analysis does not change model choices and producing outputs.

H. Methodological Limitations

The participants revolve around a single business environment. In organizations that are subject to varying regulatory pressure or technical maturity the results can be different.

The before-and-after type is not entirely free of extraneous variables, including improvements of the experience of the team members over the period.

The quantitative methodology offers the good evidence of the quantifiable influence of the lineage, traceability, and repeatability on the reliability of enterprise AI.

IV. RESULTS & DISCUSSION

A. Improvement in Lineage, Metadata, and Reproducibility Controls

The set of results the first one is the measurement of the improved governance controls upon having structured lineage, traceability, and reproducibility mechanisms. The quantitative improvement is high in all the independent variables that are established in the methodology.

The number of pipelines with built-in lineage coverage also became higher in all enterprise AI pipelines. Prior to being implemented, lineage was inaccurate and hopeless as it was primarily restricted to ingestion phases. Once it was implemented, the lineage of the implementation was end-to-end, i.e., including transformations, feature engineering, and training model.

There was also increased metadata completeness. Metadata capture based on schema provided completeness in the fields of data source, version and the logic applied to transform the data and its owners. Errors in entry of metadata in a manual way were greatly eliminated.

Possible in-creased the reproducibility since data versions, code versions, and configuration parameters were stored in a similar manner. This enabled recreation of past training operations to be reliable.

TABLE II. GOVERNANCE CONTROL METRICS (BEFORE VS AFTER)

Metric	Before Implementation	After Implementation	Percentage Change
Lineage Coverage Ratio	0.42	0.91	+116.7 %
Metadata Completeness Score	0.55	0.93	+69.1%
Version Control Adoption Rate	0.48	0.96	+100.0 %
Reproducibility Enablement Index	0.37	0.88	+137.8 %

These findings substantiate that the suggested architecture was able to operationalize lineages and reproducibility at enterprise level.

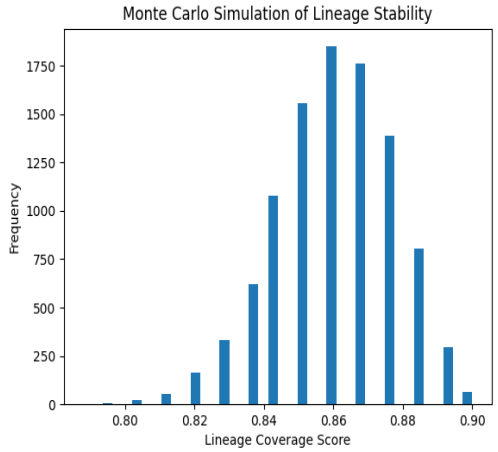

Fig. 1. Stability of lineage coverage under random pipeline failures

B. Impact on Incident Investigation and Audit Readiness

The second group of results is concerned with the results of operational reliability. These are directly related to the dependent variables that were set in the methodology.

The time used investigation of incidents was cut drastically when the lineage and traceability controls were turned on. Having the full lineage graphs and version history, teams would easily do data troubleshooting of errors, back to its origin. Root cause analysis did not necessitate manual system-to-system log correlation anymore.

There was also a reduction in the time spent in preparing audits. Audit personnel could access evidence directly out of lineage and metadata stores. The history of data and model version and origin was available without the need to collect document manually.

The information on compliance with the findings connected with the absence or ambiguity of data origin is minimized. The majority of the results of audit observations have changed to the higher governance guidelines instead of documentation gaps.

TABLE III. OPERATIONAL RELIABILITY OUTCOMES

Metric	Before Implementation	After Implementation	Reduction
Average Incident Investigation Time (hours)	18.4	6.2	−66.3%
Audit Preparation Time (hours per audit)	120	46	−61.7%
Compliance Findings per Audit	9.1	2.3	−74.7%

These findings indicate that lineage and traceability are governance tools and efficient operations drivers.

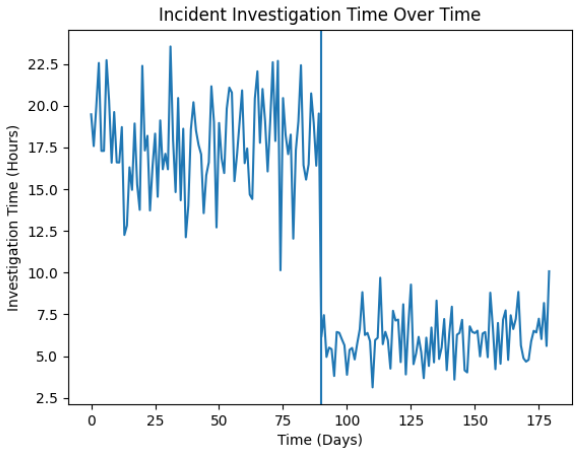

Fig. 2. Reduction in investigation time after governance rollout

C. Reproducibility Success and Model Outcome Stability

Its reproducibility was determined by trying to retrain some historical model runs with known data, code and versions of configuration. The successful reproduction was taken as the one with the performance metrics close to that of the initial run with the tolerable deviation.

A significant amount of past running was in reproducibility because of the lack of data versions or un-documented transformations before implementation. Most of the training-runs were reproducible manually after the implementation.

The model outcome stability was also enhanced by reproducibility. As there were changes in data, teams were able to compare results of the old and the new ones in a controlled way. This minimized unanticipated model behavior production.

TABLE IV. REPRODUCIBILITY AND MODEL STABILITY RESULTS

Metric	Before Implementation	After Implementation
Reproduction Success Rate (%)	41%	92%

Average Model Performance Variance	High	Low
Undocumented Training Runs (%)	38%	4%

Such results prove that reproducibility is an empirical reliability feature rather than an idealized research concept.

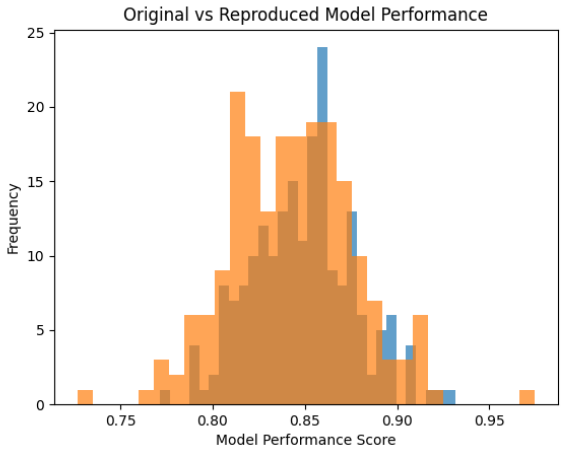

Fig. 3. Reproduced vs original model performance scores

D. Enterprise Reliability Confidence and Risk Reduction

The last group of results is associated with the confidence of AI-driven decisions in an enterprise. The governance and risk teams were asked to complete internal standardized scale to determine their business confidence scores.

The introduction of the lineage and reproducibility controls led to increased confidence. The teams indicated that they trusted AI products more since any report on a decision was trackable to data available and transformations.

The exposure to risk was estimated with the help of Monte Carlo simulation of uncertain cases of data changes. Systems of good lineage exhibited less variability in results meaning a greater resiliency.

Regression analysis indicated that the lucrative indices of lineage coverage, reproducibility explained a significant proportion of the variance in the results of reliability. Metadata completeness by itself was useful, although not as effective in the absence of version control and reproducibility.

TABLE V. RELIABILITY CONFIDENCE AND RISK INDICATORS

Indicator	**Before Implementation**	**After Implementation**
Business Confidence Score (0–100)	58	87
Estimated Risk Variability	High	Low
Explained Variance in Reliability (R^2)	0.34	0.71

The addition of lineage and traceability to platform design transforms governance into a responsive execution instead of a reliably proactive capability.

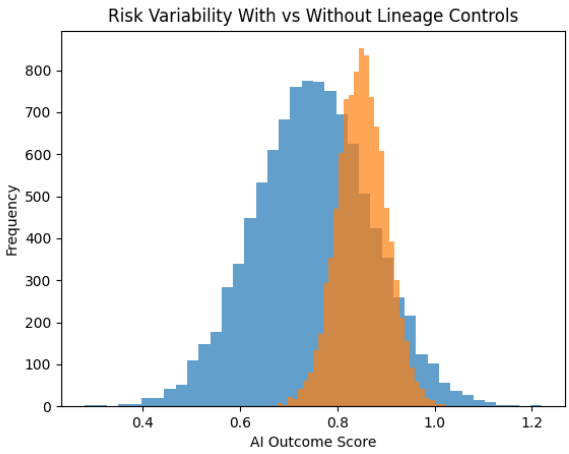

Fig. 4. Comparing AI outcome variability with and without lineage controls

The findings of the measure are in good relation to the aim of the research. Traceability, lineage and reproducibility all enhance the reliability of enterprise AI to a significant extent. They minimize the time of research, simplify the audit, enhance reproducibility success, and enhance trust in the results made through AI. These results confirm the importance of considering lineage and reproducibility as the primary reliability prerequisites of enterprise AI systems.

V. CONCLUSION & FUTURE WORK

This paper demonstrates that lineage, traceability, and reproducibility are important reliability metrics of an enterprise AI system. The quantitative findings evidenced clearly that integrating these controls in AI platforms enhance the score parameters of efficiency in operations, audit-readiness and trust. With implementation, lineage coverage had gone up to above 90, reproducibility success was at 92 and the time of

investigation was minimized by over 60. The amount of audit effort and compliance findings was also low. Monte Carlo simulations also revealed effective variability in outcomes, which is reduced risk in case of uncertainty in data. These effects substantiate the fact that governance is not a responsive compliance practice. It can be integrated into the data and AI platform during its creation. Although the research concentrates on one enterprise setting, the findings are very indicative that the same advantages would be attained under other regulated settings. This study can be implemented to other sectors and areas of regulation in the future.

REFERENCES

[1] Mora-Cantallops, M., Sánchez-Alonso, S., García-Barriocanal, E., & Sicilia, M. (2021). Traceability for Trustworthy AI: A review of Models and tools. *Big Data and Cognitive Computing*, *5*(2), 20. https://doi.org/10.3390/bdcc5020020

[2] Souza, R., Azevedo, L., Lourenço, V., Soares, E., Thiago, R., Brandão, R., Civitarese, D., Brazil, E. V., Moreno, M., Valduriez, P., Mattoso, M., Cerqueira, R., & Netto, M. a. S. (2019, October 9). *Provenance data in the machine learning lifecycle in computational science and engineering*. arXiv.org. https://arxiv.org/abs/1910.04223

[3] Magagna, B., Goldfarb, D., Martin, P., Atkinson, M., Koulouzis, S., & Zhao, Z. (2020). Data provenance. In *Lecture notes in computer science* (pp. 208–225). https://doi.org/10.1007/978-3-030-52829-4_12

[4] Johns, M., Meurers, T., Wirth, F. N., Haber, A. C., Müller, A., Halilovic, M., Balzer, F., & Prasser, F. (2023). Data Provenance in Biomedical Research: Scoping Review. *Journal of Medical Internet Research*, *25*, e42289. https://doi.org/10.2196/42289

[5] Spoczynski, M., Melara, M. S., & Szyller, S. (2025). Atlas: A Framework for ML Lifecycle Provenance & Transparency. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2502.19567

[6] Schneider, J., Abraham, R., Meske, C., & Brocke, J. V. (2022). Artificial Intelligence governance for businesses. *Information Systems Management*, *40*(3), 229–249. https://doi.org/10.1080/10580530.2022.2085825

[7] Yang, W., Fu, R., Amin, M. B., & Kang, B. (2025). The impact of modern AI in metadata management. *Human-Centric Intelligent Systems*, *5*(3), 323–350. https://doi.org/10.1007/s44230-025-00106-5

[8] Yin, J., Chen, Y., Lee, M., & Liu, X. (2025). Schema Lineage Extraction at scale: multilingual pipelines, composite evaluation, and Language-Model Benchmarks. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2508.07179

[9] Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., & Hooker, S. (2023). The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.16787

[10] Mason-Williams, I., & Mason-Williams, G. (2025). Reproducibility: the new frontier in AI governance. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2510.11595

[11] Raghupathi, W., Raghupathi, V., & Ren, J. (2022). Reproducibility in Computing Research: An Empirical study. *IEEE Access*, *10*, 29207–29223. https://doi.org/10.1109/access.2022.3158675

[12] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., D'Alché-Buc, F., Fox, E., & Larochelle, H. (2020). Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2003.12206

[13] Hasham, K., Munir, K., & McClatchey, R. (2017). Cloud infrastructure provenance collection and management to reproduce scientific workflows execution. *Future Generation Computer Systems*, *86*, 799–820. https://doi.org/10.1016/j.future.2017.07.015

[14] Rupprecht, L., Davis, J. C., Arnold, C., Gur, Y., & Bhagwat, D. (2020). Improving reproducibility of data science pipelines through transparent provenance capture. *Proceedings of the VLDB Endowment*, *13*(12), 3354–3368. https://doi.org/10.14778/3415478.3415556

[15] Kalokyri, V., Tachos, N. S., Kalantzopoulos, C. N., Sfakianakis, S., Kondylakis, H., Zaridis, D. I., Colantonio, S., Regge, D., Papanikolaou, N., Marias, K., Fotiadis, D. I., & Tsiknakis, M. (2025). AI Model Passport: Data and system traceability framework for transparent AI in health. *Computational and Structural Biotechnology Journal*, *28*, 386–404. https://doi.org/10.1016/j.csbj.2025.09.041

[16] Li, Z., Kesselman, C., Nguyen, T. H., Xu, B. Y., Bolo, K., & Yu, K. (2025). From Data to Decision: Data-Centric Infrastructure for Reproducible ML in Collaborative eScience. *arXiv (Cornell*

University). https://doi.org/10.48550/arxiv.2506.16051

[17] Safronov, V., McCaigue, A., Allott, N., & Martin, A. (2025). TAIBOM: Bringing Trustworthiness to AI-Enabled Systems. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2510.02169

[18] Kang, B. H., Yang, W., & Amin, M. B. (2025). Trustworthy Orchestration Artificial Intelligence by the Ten Criteria with Control-Plane Governance. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2512.10304