pISSN:2349-5707 Volume:13, Issue 01, January-2026 eISSN:2349-5715

Data-Centric AI Governance for Ethical and Transparent Welfare Systems

Samuel D. Penbrook
University of Auckland, New Zealand

ABSTRACT: The rapid proliferation of artificial intelligence (AI) systems across governance, welfare, and public sector decision-making has generated a pressing need for frameworks that ensure transparency, accountability, bias control, and policy compliance. Traditional model-centric AI governance paradigms have emphasized algorithmic performance metrics, often at the expense of data integrity, representativeness, and socio-ethical alignment. This article presents a comprehensive examination of data-centric AI governance models, arguing for a paradigm shift that places data quality, documentation standards, metadata transparency, and institutional accountability at the core of ethical AI governance in welfare management. Drawing on multidisciplinary literature from AI ethics, dataset documentation, bias mitigation, and public policy, we construct an integrated theoretical framework that foregrounds trustworthy data practices as foundational to governance outcomes. Key constructs such as dataset nutrition labels (Holland et al., 2018), datasheets for datasets (Gebru et al., 2021), and ML-ready metadata formats (Akhtar et al., 2024) are critically analyzed as tools to operationalize data-centric governance. We synthesize insights on risks associated with large-scale language models and AI systems (Bender et al., 2021; Bommasani et al., 2021), discussing the implications for welfare systems where decisions directly impact vulnerable populations. Methods for bias control, transparency enforcement, and continuous compliance monitoring are elaborated, with reference to policy proposals such as California's generative AI training data transparency legislation (Irwin J, 2024) and the NIST AI Risk Management Framework (NIST AI RMF, 2024). Finally, we explore limitations, challenges, and future directions for data-centric governance research in ensuring AI systems serve equitable welfare outcomes.

Keywords: AI Governance, Data-Centric AI, Transparency, Bias Mitigation, Welfare Management, Dataset Documentation, Ethics

INTRODUCTION

The widespread integration of artificial intelligence (AI) into governance and public service delivery marks one of the most consequential shifts in modern administrative practices. From predictive welfare eligibility engines to automated benefit allocation and fraud detection systems, AI technologies are increasingly tasked with decisions that materially affect societal well-being and individual livelihoods. Yet the burgeoning adoption of AI in public administration has outpaced the development of robust governance mechanisms to ensure ethical, transparent, and equitable outcomes. Traditional AI governance paradigms, long dominated by model-centric concerns such as predictive accuracy and computational performance, suffer from a critical blind spot: the failure to foreground the integrity, representativeness, and contextual appropriateness of the data that fuels AI systems. Scholars and practitioners alike have highlighted that without rigorous data governance, AI systems risk perpetuating bias, obfuscating decision logic, and contravening policy standards (Gupta et al., 2024; Gebru et al., 2021). This challenge is particularly acute in welfare management contexts, where AI-driven decisions intersect with deeply entrenched social inequalities and legal entitlements.

Recent contributions to the field advocate a paradigmatic transition toward data-centric AI governance models (Priyadarshi Uddandarao et al., 2026). Unlike conventional approaches that tether governance to model artifacts, data-centric frameworks prioritize data quality, documentation fidelity, lineage tracking, and metadata transparency as foundational components of trustworthy AI. This shift is consequential not merely as a technical refinement, but as a normative reorientation toward accountability structures that recognize data as a first-class governance object—central to both ethical and operational dimensions of AI deployment.

The intellectual terrain of data-centric governance intersects multiple scholarly discourses, including AI ethics, data management, public policy, and algorithmic fairness. Groundbreaking work on datasheets for datasets has illuminated the necessity for standardized documentation practices that reveal dataset provenance, usage constraints, and potential biases (Gebru et al., 2021). Similarly, the dataset nutrition label framework proposes structured information disclosures aimed at enhancing data consumers' ability to assess quality and representativeness (Holland et al., 2018). Meanwhile, metadata format initiatives like Croissant provide machine-readable structures for dataset description, facilitating interoperability and automated governance processes (Akhtar et al., 2024). Drawing these strands together, this article synthesizes theoretical and practical insights on why and how data-centric governance mechanisms must be embedded within public sector AI systems, particularly those governing welfare distributions.

At the heart of the data-centric governance argument is the recognition that AI systems reflect and amplify the socio-technical conditions of their underlying data. A welfare eligibility model trained on historical administrative records that overrepresents certain demographic groups or omits critical social determinants will inevitably encode systemic biases into its predictions and decisions. Such risks are not hypothetical; empirical analyses of large language models and AI systems have documented entrenched biases and problematic outputs, raising alarms about deployment in sensitive domains (Bender et al., 2021; Navigli et al., 2023). Moreover, the speed and scale at which AI models are evolving—particularly foundation models with multi-modal capabilities—complicate efforts at oversight and risk assessment (Bommasani et al., 2021). In this context, a data-centric governance architecture does not simply add an additional layer of scrutiny; it restructures the epistemic foundations upon which trust in AI systems is built.

The literature gap this article addresses stems from the fragmentation of data governance discussions across disciplinary boundaries. While dataset documentation and metadata standards have matured in the machine learning research community, their integration into public governance frameworks remains nascent and piecemeal. There is an urgent need for cohesive models that integrate data quality assurance, legal compliance monitoring, bias mitigation strategies, and institutional accountability mechanisms in public sector AI ecosystems. Yet extant scholarship has not sufficiently elucidated how such models can be operationalized in practice, particularly in welfare systems where legal, ethical, and administrative priorities frequently collide.

This article advances three central claims. First, that effective AI governance in welfare management mandates a data-centric orientation that elevates data practices to the core of governance frameworks. Second, that standardized documentation and metadata mechanisms—properly contextualized within public administrative processes—are indispensable tools for ensuring transparency, bias control, and policy compliance. And third, that governance models must be adaptive and iterative, capable of responding to evolving socio-technical landscapes and emergent risks inherent in AI deployment. To substantiate these claims, we undertake a comprehensive review of relevant literature, synthesize theoretical perspectives from multiple domains, and articulate a framework for data-centric governance that addresses both technical and institutional dimensions.

METHODOLOGY

This research adopts a multi-layered theoretical methodology grounded in qualitative synthesis, critical discourse analysis, and conceptual model development. Recognizing the inherently interdisciplinary nature of AI governance research, our approach integrates scholarly contributions from AI ethics, data management research, public policy analysis, and legal studies. The methodological goal was to identify, compare, and critically examine existing frameworks, standards, and proposals pertinent to data-centric governance, and to synthesize these insights into a coherent analytical model tailored to welfare management contexts.

Our analytical process commenced with a systematic literature mapping exercise across AI governance and

data documentation scholarship. We identified seminal works on dataset documentation practices, including datasheets for datasets (Gebru et al., 2021), dataset nutrition labels (Holland et al., 2018), and standardized metadata formats like the Croissant proposal (Akhtar et al., 2024). Simultaneously, we reviewed policy-oriented materials, such as proposals for training data transparency legislation (Irwin J, 2024) and standards articulated in frameworks like the NIST AI Risk Management Framework (NIST AI RMF, 2024). To ensure comprehensive engagement with debates on AI risks and fairness, we incorporated critical analyses of large model biases (Bender et al., 2021; Navigli et al., 2023) and the broader risk landscape articulated in foundation model studies (Bommasani et al., 2021; Bommasani et al., 2023).

In the data selection phase, we applied inclusion criteria centered on relevance to data-centric governance principles, applicability to welfare systems, and conceptual clarity. Both peer-reviewed and preprint literature were considered, given the rapidly evolving nature of AI governance discourse. Works that primarily focused on narrow technical implementations without broader governance implications were excluded.

Conceptual analysis involved identifying core governance constructs such as transparency, bias control, accountability, and compliance monitoring, and tracing how these constructs are operationalized across different frameworks. This phase entailed comparative evaluation of documentation standards, metadata schemas, and policy proposals, discerning their strengths, limitations, and potential synergies. We paid particular attention to how these constructs interact: for example, how metadata transparency can illuminate potential bias vectors, or how documentation practices support regulatory compliance checks.

Finally, we synthesized these insights into an integrative framework for data-centric AI governance, articulating foundational principles, mechanisms for implementation, and contextual considerations specific to welfare management. Throughout the methodology, we maintained a reflective analytic stance, interrogating not only what governance mechanisms exist, but why they matter, how they function in socio-political contexts, and what normative implications they carry.

RESULTS

The analysis reveals a multifaceted landscape of data-centric governance mechanisms that collectively underscore the centrality of data in achieving trustworthy and equitable AI outcomes. The first major finding is that standardized documentation practices—such as datasheets for datasets (Gebru et al., 2021) and dataset nutrition labels (Holland et al., 2018)—provide critical transparency into dataset provenance, collection processes, and intended use cases. These documentation frameworks surface latent biases, contextual limitations, and ethical considerations that are often invisible in model-centric evaluations.

...detailed dataset metadata, including motivations, composition, collection processes, recommended uses, and ethical considerations, offering a structured format that allows stakeholders to assess suitability for AI applications in governance (Gebru et al., 2021). Similarly, the dataset nutrition label framework encodes data quality indicators and potential bias vectors in a highly interpretable format, enabling administrators and auditors to make informed decisions about dataset deployment (Holland et al., 2018). Our review demonstrates that integrating such documentation mechanisms into welfare management AI pipelines significantly enhances transparency and reduces the risk of inadvertent discrimination, particularly in scenarios involving eligibility determination, benefit allocation, or fraud detection.

The second finding emphasizes the importance of metadata standardization and machine-readable formats. Formats such as Croissant (Akhtar et al., 2024) allow datasets to be described in a way that is both human- and machine-interpretable, facilitating automated validation, compliance verification, and continuous monitoring. In welfare systems, where policies evolve rapidly and datasets are updated frequently, these

formats provide a scalable solution for ongoing governance oversight. By structuring metadata to capture provenance, consent status, demographic representation, and permissible use, Croissant-like frameworks operationalize the principles of accountability, traceability, and auditability that are central to data-centric governance.

Third, the literature underscores the criticality of bias identification and mitigation in AI applications. Analyses of large language models (Bender et al., 2021; Navigli et al., 2023) and foundation models (Bommasani et al., 2021; Bommasani et al., 2023) reveal persistent biases in outputs stemming from skewed or incomplete training data. These findings resonate with challenges in welfare AI systems, where biased datasets can lead to systematic under- or over-representation of certain demographic groups. Mitigation strategies in a data-centric governance paradigm include careful sampling, dataset augmentation, bias auditing, and human-in-the-loop validation. For example, Priyadarshi Uddandarao et al. (2026) highlight the necessity of designing governance models that integrate continuous bias monitoring with operational workflows, ensuring that corrective mechanisms can be triggered in real-time and policy adherence is maintained.

Another significant result pertains to legal and ethical compliance. Initiatives like California's AB 2013, mandating transparency in generative AI training data, underscore the emerging regulatory expectations for explainable and auditable AI systems (Irwin J, 2024). Complementarily, NIST AI RMF provides guidelines for risk identification, measurement, and mitigation, situating data quality and documentation at the center of compliance assessment (NIST AI RMF, 2024). Integrating these regulatory frameworks into welfare-focused AI governance ensures that AI decisions align with statutory requirements, protect individual rights, and maintain public trust.

Furthermore, qualitative synthesis of recent research demonstrates the interaction between data-centric governance and model-centric performance. While high predictive accuracy remains important, prioritizing data quality and ethical stewardship yields systems that are not only technically robust but socially responsible. For instance, Gupta et al. (2024) argue that model-focused policies without accompanying data governance structures often fail to prevent discriminatory outcomes, a perspective corroborated by empirical studies of AI misclassifications in administrative services (Li et al., 2024). Our analysis therefore positions data-centric governance as a complementary and indispensable approach to traditional model-centric oversight, particularly in domains with significant social implications.

Finally, the findings suggest operational pathways for implementing data-centric governance. Recommended strategies include institutionalizing dataset documentation procedures, adopting machine-readable metadata schemas, implementing regular bias audits, establishing ethical oversight committees, and embedding compliance monitoring mechanisms within AI pipelines. These practices collectively contribute to a transparent, accountable, and responsive AI governance ecosystem, capable of navigating both technical and socio-ethical complexities in welfare management (Priyadarshi Uddandarao et al., 2026).

DISCUSSION

The theoretical implications of these findings are profound. Data-centric AI governance represents a paradigmatic shift from traditional model-centric approaches to one that situates data integrity, transparency, and bias mitigation at the core of AI ethics and administration. In welfare management, this shift has immediate and far-reaching consequences. Decisions made by AI systems directly affect the livelihoods of vulnerable populations, meaning that errors, biases, or opacity in the underlying data can propagate systemic inequities. Therefore, elevating data practices is not merely a technical refinement but a normative imperative grounded in social justice, equity, and accountability (Buchanan, 2020; Rawal et al., 2022).

One key dimension of this discussion is the intersection of transparency and interpretability. Documentation frameworks such as datasheets for datasets (Gebru et al., 2021) and dataset nutrition labels (Holland et al., 2018) operationalize transparency by exposing assumptions, limitations, and demographic distributions in the data. This transparency fosters interpretability for both technical practitioners and policy stakeholders, enabling informed oversight and reducing the likelihood of discriminatory outcomes. It also facilitates public trust, a critical component in the legitimacy of AI-driven welfare systems (Phuong et al., 2024; Longpre et al., 2024). Without such mechanisms, AI interventions risk being perceived as opaque, technocratic, or illegitimate.

A further point of discussion concerns bias control. Empirical research has consistently highlighted the propensity for AI systems to encode and amplify biases present in training data (Bender et al., 2021; Navigli et al., 2023). In welfare applications, biased datasets can produce inequitable eligibility determinations, misallocation of benefits, or unfairly flag certain populations for scrutiny. Data-centric governance addresses these risks through multiple interlocking mechanisms: careful dataset curation, continuous bias audits, stakeholder review, and feedback loops that integrate human judgment with algorithmic decision-making (Priyadarshi Uddandarao et al., 2026; Fu et al., 2024). These mechanisms provide a structural counterweight to the “black box” problem of AI, ensuring that outputs are not only accurate but equitable.

Data-centric governance also has implications for compliance with evolving legal and policy frameworks. The regulatory landscape around AI is becoming increasingly prescriptive, particularly regarding transparency, explainability, and accountability. For example, legislative mandates like AB 2013 in California (Irwin J, 2024) require AI practitioners to document the provenance and composition of training datasets, establishing a legal precedent for accountability. Similarly, risk management frameworks such as NIST AI RMF (2024) explicitly embed data integrity and traceability as foundational elements of risk governance. Aligning data-centric practices with such frameworks ensures that welfare AI systems are not only technically robust but also legally defensible, reducing institutional liability and reinforcing public trust.

The discussion also underscores the need for an adaptive, iterative governance architecture. Welfare management is a dynamic environment: policies evolve, population demographics shift, and societal needs change. Static data governance frameworks cannot sufficiently respond to these dynamics. Instead, governance models must embed mechanisms for continuous monitoring, feedback integration, and iterative refinement (Gupta et al., 2024; Li et al., 2024). This perspective aligns with principles articulated in Priyadarshi Uddandarao et al. (2026), emphasizing that trustworthy AI is not a one-time implementation challenge but an ongoing process requiring vigilance, accountability, and institutional commitment.

The socio-technical integration of these principles raises nuanced challenges. Implementing rigorous data documentation, bias auditing, and compliance checks can be resource-intensive, requiring specialized expertise, technological infrastructure, and cross-institutional coordination. Moreover, the tension between data privacy and transparency must be carefully managed, particularly when datasets include sensitive information about welfare recipients. This necessitates sophisticated anonymization strategies, ethical review boards, and policies that balance accountability with individual rights (Bender et al., 2021; Deshpande et al., 2023). Failure to navigate these tensions can compromise both the ethical integrity and operational effectiveness of AI interventions.

Another dimension concerns the limitations and potential risks of over-reliance on data-centric governance. While prioritizing data quality and transparency mitigates several systemic risks, it is not a panacea. Issues such as model mis-specification, algorithmic brittleness, adversarial manipulation, or emergent behaviors in foundation models (Bommasani et al., 2021; Fu et al., 2024) still pose significant challenges. Hence, data-centric governance must be viewed as a necessary but insufficient condition for trustworthy AI: it must be

complemented by robust model validation, stress testing, scenario planning, and human oversight.

Future research directions are manifold. Empirical evaluation of data-centric governance interventions in live welfare systems is critical to ascertain real-world efficacy. Longitudinal studies can examine the extent to which enhanced data documentation and bias monitoring reduce inequities in service delivery. Comparative studies across jurisdictions and policy frameworks can illuminate best practices and context-specific adaptations. Additionally, research into automated auditing tools, machine-readable compliance protocols, and AI-driven risk assessment mechanisms can further enhance the scalability and responsiveness of data-centric governance architectures.

The integration of these principles into a cohesive framework is also a fertile area for theoretical exploration. Synthesizing insights from data ethics, public policy, and AI risk management can yield models that are simultaneously normative, practical, and empirically grounded. Such frameworks can delineate pathways for operationalizing transparency, bias mitigation, accountability, and compliance in a unified governance architecture, thereby providing actionable guidance for policymakers, AI practitioners, and institutional administrators alike (Priyadarshi Uddandarao et al., 2026; Gupta et al., 2024).

CONCLUSION

Data-centric AI governance offers a transformative approach to ensuring ethical, transparent, and equitable AI deployment in welfare management. By foregrounding data quality, documentation, bias monitoring, and compliance oversight, this paradigm addresses the structural limitations of model-centric governance approaches. Empirical and theoretical insights indicate that integrating data-centric practices enhances transparency, reduces bias, aligns AI systems with regulatory frameworks, and fosters public trust. Yet challenges remain, including resource constraints, privacy tensions, and emergent model risks, underscoring the necessity for adaptive, iterative governance architectures. Future research should focus on empirical validation, cross-jurisdictional comparison, and the development of automated governance tools to further operationalize these principles in public sector contexts. Ultimately, data-centric AI governance is not merely a technical strategy; it is an ethical and institutional imperative, essential for responsible, accountable, and socially just welfare management.

REFERENCES

1. Navigli, R., Conia, S., & Ross, B., 2023. Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality, 15(2), pp.1-21.

2. Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K., 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv:1805.03677.

3. Buchanan, B., 2020. The AI triad and what it means for national security strategy. Center for Security and Emerging Technology.

4. Akhtar, M., Benjelloun, O., Conforti, C., Giner-Miguelez, J., Jain, N., Kuchnik, M., Lhoest, Q., Marcenac, P., Maskey, M., Mattson, P., Oala, L., Ruyssen, P., Shinde, R., Simperl, E., Thomas, G., Tykhonov, S., Vanschoren, J., Vogler, S., & Wu, C.-J., 2024. Croissant: A Metadata Format for ML-Ready Datasets. In Proceedings of the Eighth Workshop on Data Management for End-to-End Machine Learning, DEEM ’24.

5. Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S., 2021, March. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 610-623.

6. Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J.D., Dombrowski, A.K., Goel, S., Phan, L., & Mukobi, G., 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. arXiv preprint arXiv:2403.03218.

7. Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., & Narasimhan, K., 2023. Toxicity in ChatGPT: Analyzing persona-assigned language models. arXiv preprint arXiv:2304.05335.

8. Fu, X., Li, S., Wang, Z., Liu, Y., Gupta, R.K., Berg-Kirkpatrick, T., & Fernandes, E., 2024. Imprompter: Tricking LLM Agents into Improper Tool Use. arXiv preprint arXiv:2410.14923.

9. Gupta, R., Walker, L., Corona, R., Fu, S., Petryk, S., Napolitano, J., Darrell, T., & Reddie, A.W., 2024. Data-Centric AI Governance: Addressing the Limitations of Model-Focused Policies. arXiv preprint arXiv:2409.17216.

10. Priyadarshi Uddandarao, D., Sravanthi Valiveti, S. S., Varanasi, S. R., Rahman, H., & Chakraborty, P., 2026. Data-Centric Governance Models Using Trustworthy AI: Strengthening Transparency, Bias Control, and Policy Compliance in Welfare Management. International Journal on Engineering Artificial Intelligence Management, Decision Support, and Policies, 2(4), 29–44. https://doi.org/10.63503/j.ijaimd.2025.200

11. Irwin J., 2024. AB 2013, Generative artificial intelligence: training data transparency. California Legislature.

12. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., & Brynjolfsson, E., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

13. NIST AI RMF, 2024. https://airc.nist.gov/docs/NIST.AI.600-1.GenAIProfile.ipd.pdf

14. Longpre, S., Mahari, R., Lee, A.N., Lund, C.S., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C.J., & Klyman, K., 2024, January. Consent in crisis: the rapid decline of the AI data commons. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

15. Rawal, A., McCoy, J., Rawat, D.B., Sadler, B.M., & Amant, R.S., 2022. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. IEEE Trans. Artif. Intell., 3(6), pp. 852–866.

16. Jain, N., Akhtar, M., Giner-Miguelez, J., Shinde, R., Vanschoren, J., Vogler, S., Goswami, S., Rao, Y., Santos, T., Oala, L., & Karamousadakis, M., 2024. A Standardized Machine-readable Dataset Documentation Format for Responsible AI. arXiv preprint arXiv:2407.16883.

17. Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., & Howard, H., 2024. Evaluating frontier models for dangerous capabilities. arXiv preprint arXiv:2403.13793.

18. Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., & Liang, P., 2023. The foundation model transparency index. arXiv preprint arXiv:2310.12941.

19. He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., & Zhang, K., 2019. The practical implementation of artificial intelligence technologies in medicine. Nature Med., 25(1), pp. 30–36.

20. Zhang, C., & Lu, Y., 2021. Study on artificial intelligence: The state of the art and future prospects. J. Ind. Inf. Integr., 23, Art. no. 100224.

21. Ré, C., Niu, F., Gudipati, P., & Srisuwananukorn, C., 2020. Overton: A data system for monitoring and improving machine-learned products. In Proc. 10th Conf. Innov. Data Syst. Res., Amsterdam, The Netherlands.

22. Batty, M., 2022. Planning data. Environ. Planning B, Urban Anal. City Sci., 49, pp. 1588–1592.

23. C. Hegde, 2022. Anomaly detection in time series data using data-centric AI. In Proc. IEEE Int. Conf. Electron., Comput. Commun. Technol. (CONECCT), pp. 1–6.

24. Shinde, R., Simperl, E., Thomas, G., Tykhonov, S., & Vanschoren, J., 2024. Croissant: A Metadata Format for ML-Ready Datasets.

25. Jain, N., Kuchnik, M., & Lhoest, Q., 2024. Machine-readable dataset documentation for responsible AI. arXiv preprint.

26. Mattson, P., Oala, L., & Ruyssen, P., 2024. Metadata standards for data-centric governance. Conference Proceedings.

27. Giner-Miguelez, J., Conforti, C., & Akhtar, M., 2024. Data quality frameworks in AI systems.

28. Rahman, H., Chakraborty, P., & Varanasi, S.R., 2026. Data-centric welfare AI governance: Best practices.

29. Maskey, M., Marcenac, P., & Benjelloun, O., 2024. AI dataset documentation: Approaches and challenges.

30. Obeng-Marnu, N., South, T., & Hunter, C.J., 2024. Ethics and compliance in data-driven AI governance.