



Integrity by Design: Building Trustworthy Artificial Intelligence for Infrastructure and Clinical Decision Support

Sai Nethra Betgeri, Ph.D.,

Department of Computer Science and Engineering, University of Louisville, Louisville, Kentucky, United States.

Article Details

Article Type: Commentary Article

Received date: 26th August, 2026

Accepted date: 05th September, 2026

Published date: 07th September, 2026

***Corresponding Author:** Sai Nethra Betgeri, Ph.D., Department of Computer Science and Engineering, University of Louisville, Louisville, Kentucky, United States.

Citation: Betgeri, S. N., (2026). Integrity by Design: Building Trustworthy Artificial Intelligence for Infrastructure and Clinical Decision Support. *J Inform Techn Int*, 4(1): 116. doi: <https://doi.org/10.33790/jiti1100116>

Copyright: ©2026, This is an open-access article distributed under the terms of the [Creative Commons Attribution License 4.0](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Artificial intelligence is increasingly used to guide decisions in infrastructure management, health care, and sustainable materials research. In these settings, predictive accuracy alone is not sufficient. Trust also depends on the integrity of data, methods, explanations, and deployment practices. This commentary draws on multidisciplinary machine-learning research to argue for an integrity-by-design approach in which provenance, domain knowledge, uncertainty, reproducibility, fairness, and human oversight are treated as core system requirements. The discussion highlights practical principles for developing AI systems that remain scientifically defensible and socially responsible beyond the laboratory.

Keywords: Trustworthy Artificial Intelligence; Data Integrity; Machine Learning; Explainability; Reproducibility; Clinical Decision Support; Infrastructure Management

Artificial intelligence beyond predictive accuracy

Machine-learning systems now influence decisions about deteriorating infrastructure, clinical risk, environmental sustainability, and resource allocation. These applications can produce substantial public benefit, but they also reveal a central challenge: a model may be statistically impressive while remaining operationally untrustworthy. High accuracy does not repair incomplete records, biased sampling, data leakage, unstable feature definitions, or deployment conditions that differ from the training environment. For high-consequence applications, integrity must therefore describe the complete lifecycle of the system—not merely the final prediction.

My research across infrastructure analytics, clinical decision-support systems, and sustainable materials has reinforced this principle. Although these domains differ in their data and objectives, they share a common requirement: model outputs must be traceable to credible evidence and interpretable within the relevant domain. A useful AI system should not only answer “What is likely to happen?” but also “Why is this result plausible, how certain is it, and when should a human decline to rely on it?” [1].

Data integrity is the first layer of trustworthy AI

Trustworthy modeling begins before algorithm selection. Infrastructure records may combine inspection reports, pipe characteristics, maintenance histories, environmental variables, and sensor measurements collected at different times and resolutions. Clinical data may integrate electronic health records, laboratory results, imaging, biomarkers, and genomic information. Each source carries missingness, measurement error, coding changes, and potential demographic or institutional bias. If these limitations are concealed during preprocessing, the model can transform uncertainty into false confidence.

An integrity-centered workflow should document data provenance, inclusion and exclusion rules, preprocessing decisions, variable definitions, missing-data strategies, and dataset versions. Training, validation, and test partitions must prevent temporal, patient-level, or asset-level leakage. Researchers should report performance across meaningful subgroups and operating conditions rather than relying on one aggregate metric. In both infrastructure and health applications, uncertainty estimates and calibration are often as important as discrimination because decision makers must understand the reliability of an individual prediction [2].

Domain knowledge strengthens accountability

Purely data-driven models may learn associations that are convenient but physically, clinically, or scientifically implausible. Integrating domain knowledge can reduce this risk. In infrastructure research, physics-informed features and deterioration mechanisms can constrain or contextualize predictions. In clinical decision support, established pathways, temporal relationships, and expert review can help distinguish meaningful signals from correlations created by documentation practices. In materials research, chemical composition, curing conditions, and known performance relationships should guide feature design and the interpretation of model outputs.

Explainability should similarly be used as a verification instrument rather than as decoration. Feature-importance measures, local

explanations, sensitivity analyses, and counterfactual checks can reveal whether a model relies on appropriate evidence. However, an explanation method does not automatically make a system fair or causal. Explanations must be evaluated for stability and communicated in language appropriate for engineers, clinicians, researchers, and affected users.

Reproducibility and governance must continue after publication

Scientific integrity requires enough information for others to assess and reproduce the work. Authors should report dataset construction, model architecture, hyperparameter selection, validation design, evaluation metrics, software versions, random seeds, and limitations. Where privacy or ownership prevents release of raw data, researchers can still provide data dictionaries, synthetic examples, analysis code, model cards, and detailed protocols. Reproducibility is not simply a reporting exercise; it exposes hidden assumptions and strengthens confidence in the conclusions.

Deployment adds another layer of responsibility. Models should be monitored for data drift, performance degradation, security threats, and unequal impact. Version control, audit logs, access controls, incident-response procedures, and defined responsibilities are essential parts of AI governance. Human oversight must be substantive: users need authority to question or override a recommendation, and the system should identify cases that fall outside its validated scope. These practices align with risk-management approaches that emphasize validity, reliability, transparency, accountability, privacy, and fairness [1–3].

A practical agenda for integrity-by-design

The next generation of applied AI research should evaluate systems on more than predictive performance. A credible contribution should demonstrate: (1) traceable and fit-for-purpose data; (2) validation that reflects the intended setting; (3) interpretable and uncertainty-aware outputs; (4) explicit analysis of bias, privacy, security, and misuse; (5) reproducible methods; and (6) continuing human governance. These criteria should be incorporated at project conception rather than added immediately before publication or deployment [3–4].

The field also benefits from closer collaboration among computer scientists, domain experts, social scientists, institutional leaders, and end users. Such collaboration helps define the real decision problem, identify harms that technical metrics may overlook, and determine whether an AI intervention improves practice. Journals devoted to information technology and integrity have an important role in advancing this broader standard by encouraging transparent reporting, negative and replication findings, interdisciplinary evaluation, and responsible discussion of limitations.

Conclusion

Artificial intelligence earns trust not through complexity, but through evidence, transparency, and accountable use. Across infrastructure, health care, and sustainable engineering, integrity-by-design offers a practical path from promising models to reliable decision-support tools. When data provenance, domain knowledge, uncertainty, reproducibility, and human oversight are treated as first-class requirements, AI can support innovation without weakening scientific or public trust.

References

1. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1; 2023.
2. World Health Organization. Ethics and Governance of Artificial Intelligence for Health. Geneva: WHO; 2021.
3. UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO.
4. Collins, G. S., Moons, K. G. M., & Dhiman, P., et al. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 385:e078378.