Machine Intelligence & Responsible Systems
PERSPECTIVE | COMMISSIONED SYNTHESIS

Documentation as Infrastructure for Accountable AI Systems

Jonathan King - Corresponding author

Abstract

This perspective examines documentation artifacts as operational infrastructure for accountable AI. The organizing question is how model, data, deployment, and incident records can remain connected across a system lifecycle. Ten related scholarly sources are synthesized through a decision-centered framework spanning problem definition, mechanism, measurement, evaluation, implementation, and governance. The review does not invent experiments, pooled estimates, or unreported quantitative results. It instead evaluates the strength and transferability of the available evidence, with particular attention to publishing static transparency documents that are detached from engineering change. The resulting framework links technical or empirical performance to explicit use conditions and identifies tests that should precede wider adoption in organizations deploying learning systems across changing contexts.

Keywords: responsible AI, model cards, datasheets, audits, accountability, documentation

Synthesis lens	Principal question
Mechanism	the handoffs linking datasets, models, evaluations, deployments, and observed incidents
Evaluation	completeness, verifiability, update latency, and whether documented limits predict actual failures
Primary risk	publishing static transparency documents that are detached from engineering change

1. Introduction

Research on documentation artifacts as operational infrastructure for accountable AI sits at the boundary between a promising component and a consequential decision. The core question is how model, data, deployment, and incident records can remain connected across a system lifecycle. Answering it requires more than assembling favorable results. It requires a chain of evidence that makes the problem boundary, mechanism, measurement, comparator, uncertainty, and accountable decision visible to the reader.

This article presents a structured narrative review of ten related publications. Sources were selected for topical relevance, traceable publication metadata, and complementary coverage of technical, empirical, and governance questions. No meta-analytic pooling is attempted because the included studies differ in designs, populations, system boundaries, and outcomes. The purpose is decision-oriented synthesis: to identify what each source can support, where transfer remains uncertain, and which evidence would justify the next stage of use.

The central risk is publishing static transparency documents that are detached from engineering change. A top-line metric or broad policy label can appear decisive while concealing the conditions that produced it. Accordingly, the synthesis separates observation from interpretation and implementation from validation. This separation is especially important when the same system creates benefits for one user or institution while transferring cost, risk, or administrative burden to another.

2. Evidence Base and Problem Boundary

A defensible review begins by defining the unit of analysis. For the present topic, the relevant boundary includes inputs, institutional or technical context, the mechanism producing an output, the person or system acting on that output, and the downstream outcome. Evidence falls outside the boundary when it measures only an intermediate proxy yet is used to claim an end-to-end benefit.

Liu et al. (2022) addresses "The medical algorithmic audit." In this synthesis, that work is used as a source on problem definition within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Königstorfer and Thalmann (2022) addresses "AI Documentation: A path to accountability." In this synthesis, that work is used as a source on conceptual boundary within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Mylrea and Robinson (2023) addresses "Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI." In this synthesis, that work is used as a source on measurement within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Together, these works provide complementary entry points, but their conclusions should not be flattened into a single ranking. Differences in data generation, setting, temporal horizon, and outcome definition may be substantively meaningful. A review should therefore preserve those differences and state where analogy, rather than direct replication, connects one domain to another.

3. Mechanism and System Design

The operative mechanism is the handoffs linking datasets, models, evaluations, deployments, and observed incidents. This formulation discourages component-only reasoning. A model, platform, policy, assay, or infrastructure service acquires practical meaning only when its interfaces and use conditions are specified. The evidence chain should describe what information is available, what transformation occurs, which assumptions make that transformation credible, and who is authorized to act.

Fehr et al. (2024) addresses "A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare." In this synthesis, that work is used as a source on system mechanism within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Naja et al. (2022) addresses "Using Knowledge Graphs to Unlock Practical Collection, Integration, and Audit of AI Accountability Information." In this synthesis, that work is used as a source on representation choice within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Cobbe et al. (2023) addresses "Understanding Accountability in Algorithmic Supply Chains." In this synthesis, that work is used as a source on failure analysis within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

These studies also show why modular claims require interface tests. A component may perform well while the complete pathway fails because inputs drift, users interpret outputs differently, recovery semantics are ambiguous, or institutional incentives change. Design documentation should therefore include not only the intended pathway but also foreseeable deviations, degraded modes, and conditions under which the system should stop or defer.

4. Evaluation and Uncertainty

Evaluation should center on completeness, verifiability, update latency, and whether documented limits predict actual failures. Average performance remains useful, but it should be accompanied by distributions, error categories, relevant subgroups or operating regimes, and uncertainty at the point where a decision changes. Comparators must isolate the value of the proposed addition rather than compare a mature intervention with an intentionally weak baseline.

Sandu et al. (2022) addresses "Time to audit your AI algorithms." In this synthesis, that work is used as a source on comparative evaluation within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Hartmann et al. (2024) addresses "Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society." In this synthesis, that work is used as a source on uncertainty within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

External validation should introduce meaningful shifts in setting, time, population, workload, market structure, or environmental conditions. A nominally independent sample can still be too similar to test transfer. Conversely, a failed transfer is scientifically informative when the changed conditions are measured and the mechanism of failure is examined rather than hidden in an aggregate score.

5. Translation and Governance

Translation to organizations deploying learning systems across changing contexts depends on more than technical readiness. Decision rights, documentation, maintenance, monitoring, appeal, and recovery are part of the intervention. The governance requirement is document ownership, version control, review gates, and public-facing accountability. Each control should be connected to a named failure mode and should identify an owner, evidence source, review interval, and escalation path.

Lund et al. (2025) addresses "Standards, frameworks, and legislation for artificial intelligence (AI) transparency." In this synthesis, that work is used as a source on implementation within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Micheli et al. (2023) addresses "The landscape of data and AI documentation approaches in the European policy context." In this synthesis, that work is used as a source on governance within documentation artifacts as operational infrastructure for accountable AI. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for organizations deploying learning systems across changing contexts. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

A credible implementation plan also defines sunset and revalidation conditions. New data, software updates, institutional restructuring, population change, or shifts in incentives can invalidate previously reasonable assumptions. Monitoring should therefore test the validity of the evidence chain, not merely whether a service remains online or a headline metric remains stable.

6. Research Agenda

Future studies should preregister or otherwise predefine the intended decision, principal outcomes, exclusions, and analysis plan when feasible. They should report missingness, sensitivity analyses, negative or null findings, and the cost of errors to differently situated stakeholders. Reproducible artifacts should include sufficient metadata to reconstruct data provenance, model or analytical versions, parameter choices, and the sequence from evidence to conclusion.

Cross-disciplinary work needs a shared object of explanation rather than a collection of adjacent vocabularies. For documentation artifacts as operational infrastructure for accountable AI, that shared object is the complete decision pathway. Technical, clinical, social, environmental, or economic expertise should each change the design or interpretation of the study. When evidence remains incomplete, the useful output is a bounded hypothesis, a transparent uncertainty statement, or a request for additional measurement.

7. Conclusion

The literature supports a disciplined approach to documentation artifacts as operational infrastructure for accountable AI: define the decision, preserve system boundaries, test consequential failure modes, connect uncertainty to action, and assign accountable control. The strongest conclusion is not that one method is universally superior. It is that credible use in organizations deploying learning systems across changing contexts requires an auditable link from source evidence to design choice, evaluation result, and governed decision.

References

Cobbe, J., Veale, M., & Singh, J. (2023). Understanding Accountability in Algorithmic Supply Chains. Scholarly publication. https://doi.org/10.31235/osf.io/p4sey

Fehr, J., Citro, B., Malpani, R., Lippert, C., & Madai, V. I. (2024). A trustworthy AI reality-check: the lack of transparency of artificial intelligence products in healthcare. Frontiers in Digital Health, 6, 1267290. https://doi.org/10.3389/fdgth.2024.1267290

Hartmann, D., Pereira, J. R. L. D., Streitbörger, C., & Berendt, B. (2024). Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. AI and Ethics, 5(4), 3617-3638. https://doi.org/10.1007/s43681-024-00595-3

Königstorfer, F., & Thalmann, S. (2022). AI Documentation: A path to accountability. Journal of Responsible Technology, 11, 100043. https://doi.org/10.1016/j.jrt.2022.100043

Liu, X., Glocker, B., McCradden, M. M., Ghassemi, M., Denniston, A. K., & Oakden-Rayner, L. (2022). The medical algorithmic audit. The Lancet Digital Health, 4(5), e384-e397. https://doi.org/10.1016/s2589-7500(22)00003-6

Lund, B., Orhan, Z., Mannuru, N. R., Bevara, R. V. K., Porter, B., Vinaih, M. K., & Bhaskara, P. (2025). Standards, frameworks, and legislation for artificial intelligence (AI) transparency. AI and Ethics, 5(4), 3639-3655. https://doi.org/10.1007/s43681-025-00661-4

Micheli, M., Hupont, I., Delipetrev, B., & Soler-Garrido, J. (2023). The landscape of data and AI documentation approaches in the European policy context. Ethics and Information Technology, 25(4). https://doi.org/10.1007/s10676-023-09725-7

Mylrea, M., & Robinson, N. (2023). Artificial Intelligence (AI) Trust Framework and Maturity Model: Applying an Entropy Lens to Improve Security, Privacy, and Ethical AI. Entropy, 25(10), 1429. https://doi.org/10.3390/e25101429

Naja, I., Markovic, M., Edwards, P., Pang, W., Cottrill, C., & Williams, R. (2022). Using Knowledge Graphs to Unlock Practical Collection, Integration, and Audit of AI Accountability Information. IEEE Access, 10, 74383-74411. https://doi.org/10.1109/access.2022.3188967

Sandu, I., Wiersma, M., & Manichand, D. (2022). Time to audit your AI algorithms. Maandblad Voor Accountancy en Bedrijfseconomie, 96(7/8), 253-265. https://doi.org/10.5117/mab.96.90108