Causal Provenance for Fatigue-Life Surrogates

Abstract

This evidence synthesis examines simulation lineage in parameterized industrial assemblies. It argues that the appropriate object of evaluation is the digital thread connecting geometry, materials, boundary conditions, solver settings, derived features, and predicted capacity, not a model score in isolation. The review brings together the assigned studies with established work on uncertainty, robustness, provenance, and governance. Across these literatures, a common problem emerges: a fast surrogate can move a design decision far outside the domain where its training simulations were physically credible. The proposed framework separates evidence quality, model behavior, decision policy, and operational monitoring, then asks how each layer changes under distribution shift, adversarial pressure, or incomplete information. It recommends evaluation by slices and repeated trials, explicit reject and escalation policies, preservation of data and reasoning lineage, and prospective monitoring tied to defined actions. The result is a research agenda for systems that are efficient enough to use but also bounded enough to audit. No new experiment is claimed; the article develops a comparative conceptual model and identifies tests that would make future empirical claims more credible.

Introduction

Which data transformations must be traceable before a fatigue surrogate can support design? The problem resists a learned component-only answer; once deployed, a predictor becomes part of a wider decision process. Its behavior reflects a chain of design decisions about evidence, representation, alternatives, resource use, and response. In parameterized industrial assemblies, these choices interact with institutions and physical processes that the estimator only partially represents. A high score can coexist with fragile inputs, an evaluator that rewards the wrong surface feature, or an action rule that turns modest uncertainty into a costly decision. This makes it useful to treat simulation lineage as an evidence problem. The task is therefore to identify the evidence that makes a dependability claim testable. For the present focus, simulation lineage therefore becomes a criterion for deciding which evidence deserves further scrutiny.

The comparison is organized around the operational operational sequence. Its unit is the digital thread connecting geometry, materials, boundary conditions, solver settings, derived features, and predicted capacity. This avoids a recurring category error: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. It also gives the target studies a common basis for comparison. Although their application settings differ, they each make some part of an inference chain more explicit - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. The key test is whether that explicit component remains meaningful when parameterized geometries, finite-element and multiscale outputs, loading histories, material properties, and validation measurements change, and whether the proposed safeguards lead to design screening, uncertainty-aware optimization, targeted simulation, and physical testing rather than merely producing another metric. For the present focus, simulation lineage therefore becomes a criterion for deciding which evidence deserves further scrutiny.

Evidence From the Focal Studies

A useful test case is MuSK: Multi-Scale Knowledge Learning for Provenance-Graph Anomaly Detection. Rather than treating its output as an isolated score, the study frames how stealthy multi-stage attacks can be detected in heterogeneous system-provenance graphs without abundant attack labels. Its central mechanism is self-supervised recovery of masked node attributes, meta-path edges, and positional structure, followed by meta-path-guided subgraph verification, and its evidential basis is that evaluation spans nine provenance settings, including DARPA Transparent Computing hosts, and reports high precision, complete recall on those hosts, and efficiency gains from cached incremental expansion. The design joins local attributes, heterogeneous causal paths, and global position while moving decisions from isolated nodes to suspicious subgraphs. For parameterized industrial assemblies, the transferable lesson is methodological: a system should preserve the path from signal to decision and expose the conditions under which that path may fail. The boundary is equally important. Performance may depend on audit completeness, benign workload diversity, attack coverage, and the stability of hand-selected meta-path semantics. (Ma et al. 2026).

Evidence for simulation lineage becomes concrete in BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models. The paper addresses how reinforcement learning for program repair can overcome sparse execution feedback and coarse sequence-level credit through a three-stage pipeline combining execution-verified supervised traces, sequence- and line-level reward models, and PPO with reward redistribution to critical edit regions. Its evaluation is informative because evaluation spans SWE-bench Verified, Defects4J transfer, HumanEval-Java, and QuixBugs, with reported gains in both repository repair and cross-language generalization. Line-level credit assignment makes test execution informative at the granularity where patches actually change behavior. This does not make the method a universal component; it identifies a design hypothesis that must be re-tested in the article's focal setting. In particular, benchmark success does not by itself establish maintainability, specification fidelity, or robustness to incomplete and flaky tests. The appropriate use of the result is therefore bounded, comparative, and explicit about unresolved deployment conditions (Li et al. 2026).

The strongest contribution of Machine Learning-Based Fatigue Life Evaluation of the Pump Spindle Assembly With Parametrized Geometry is not a generic claim that learning improves performance. It is the more specific proposition that how fatigue life can be estimated for an assembled pump spindle whose geometry, assembly choices, and loads vary together. The proposed route is a hybrid ensemble over linear, tree, boosting, and gradient-boosting models, with rule-based sample generation and repeated Monte Carlo validation. Support comes from the fact that the evaluation platform considers both high- and low-cycle fatigue, multiple design parameters, external loads, and standard regression performance measures. The work moves fatigue learning from material coupons toward parameterized industrial assemblies. Read through the lens of simulation lineage, the study also illustrates why a reported average must remain connected to the workflow that produced it. Simulation-derived samples and random validation can overstate transfer to manufacturing variation, damage histories, and operating regimes outside the design envelope. (Wang et al. 2023).

A Layered Model of Reliability

The conceptual framework begins by separating four layers. The evidence layer records observations and their provenance. The representation layer turns those observations into features, graphs, tokens, or simulated states. The inference layer produces predictions, explanations, translations, anomaly scores, or candidate actions. The decision layer maps those outputs to design screening, uncertainty-aware optimization, targeted simulation, and physical testing. Reliability can fail at any boundary. Better inference cannot repair evidence that omitted the relevant condition, and a calibrated probability cannot rescue an action rule whose costs were never specified. Conversely, a conservative decision policy may reduce harm even when the underlying model remains imperfect. This layered view makes simulation lineage testable because each claim can be assigned to a component and a measurable transition. This point narrows the research task for parameterized industrial assemblies: compare alternatives under the same information and resource constraints.

Another important separation concerns random variation, limited knowledge, and feedback from strategic actors. Aleatory variation concerns irreducible differences among cases. Epistemic uncertainty concerns missing knowledge, limited samples, or unsupported regions of the input space. Strategic adaptation occurs when users, adversaries, markets, or institutions respond to the operational tool itself. These sources demand different controls. Repeated measurement can characterize variation; new evidence and conservative extrapolation can reduce epistemic uncertainty; red teaming and feedback-aware validation address adaptation. Combining them into one confidence score hides which intervention is appropriate. A defensible system therefore reports uncertainty in a form tied to an available response in place of presenting confidence as a decorative number. The implication is especially consequential in parameterized industrial assemblies, where a small modeling choice can redirect attention and resources.

From Evidence Capture to Escalation

A uniform inference budget is rarely the best operational policy. Easy, well-supported cases can take a fast path. Shifted, ambiguous, or high-cost cases should activate additional precision, retrieval, alternative drafts, simulation, or expert review. This conditional design is preferable to applying maximum computation everywhere because resource cost is part of implementation quality. It is also preferable to an unconditional fast path because efficiency can amplify a systematic error at scale. The trigger must be evaluated as carefully as the expensive model: coverage, false reassurance, subgroup effects, latency, and the consequences of abstention all matter. The output is not simply a prediction but a package containing evidence links, uncertainty, the action taken, and the conditions that caused escalation. This point narrows the research task for parameterized industrial assemblies: compare alternatives under the same information and resource constraints.

A traceable deployment in parameterized industrial assemblies begins with typed evidence. Each record should identify its source, time, collection conditions, transformations, and relation to other records. Graphs are useful when relations carry meaning, but graph construction is itself a hypothesis. An edge may denote temporal adjacency, physical causation, code dependency, shared infrastructure, or analyst assertion; treating these as interchangeable creates false certainty. The architecture should therefore maintain edge types and confidence, retain alternative links when evidence is ambiguous, and version both the data schema and the rules that construct the graph. A model may aggregate this structure, but the original evidence must remain recoverable for audit. The implication is especially consequential in parameterized industrial assemblies, where a small modeling choice can redirect attention and resources.

A Broader Evidence Base

Several established literatures clarify the design space. Critical-plane analysis demonstrates why multiaxial phase relations cannot be collapsed into a single load magnitude (Fatemi and Socie 1988).

Computational homogenization clarifies which mesoscale details may be summarized and which must remain explicit (Geers, Kouznetsova, and Brekelmans 2010). Model-calibration theory separates parameter uncertainty, observation error, and structural discrepancy (Kennedy and O'Hagan 2001). Global sensitivity analysis measures interactions that one-factor-at-a-time studies can miss (Saltelli et al. 2010). Together these findings imply that simulation lineage should be evaluated against explicit alternatives and failure costs. In Causal Provenance for Fatigue-Life Surrogates, this literature supplies a specific test of simulation lineage within parameterized industrial assemblies.

The evidence base offers complementary tests rather than one universal metric. Random forests offer nonlinear prediction and useful baselines with comparatively modest tuning demands (Breiman 2001). Gradient boosting provides a strong tabular-data benchmark against which more complex ensembles should be justified (Friedman 2001). Additive feature attribution can audit model behavior, although attribution is not a causal explanation (Lundberg and Lee 2017). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In Causal Provenance for Fatigue-Life Surrogates, this literature supplies a specific test of simulation lineage within parameterized industrial assemblies.

A credible assurance case can be built from findings that operate at different levels. Topology optimization supplies a disciplined link between design variables, constraints, and structural objectives (Bendsoe and Sigmund 2003). FAIR data practice matters when simulation lineage and material metadata determine whether a surrogate can be reproduced (Wilkinson et al. 2016). Classical fatigue mechanics links cyclic loading, microstructural damage, crack initiation, and crack growth across scales (Suresh 1998). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In Causal Provenance for Fatigue-Life Surrogates, this literature supplies a specific test of simulation lineage within parameterized industrial assemblies.

An Evaluation Protocol

The action and its costs should determine the reported metrics. Discrimination measures whether cases are ranked correctly; calibration asks whether confidence corresponds to observed frequency; selective risk asks what error remains after deferral; robustness measures performance across defined perturbations; and utility assigns costs to actions. These are not interchangeable. For simulation lineage, the minimum report should include overall performance, slice-level results, uncertainty intervals, coverage-risk curves, stability across runs, and failure examples linked back to parameterized geometries, finite-element and multiscale outputs, loading histories, material properties, and validation measurements. If the system updates incrementally, the evaluation must also test forgetting, stale evidence, and rollback. If an automatic judge supplies labels, a sample needs independent human review and content-preserving perturbations that reveal whether the judge is grading substance. This requirement gives practical content to the article's central question: Which data transformations must be traceable before a fatigue surrogate can support design?

A claim-to-test matrix should precede metric selection. Each intended claim - such as improved robustness, safer abstention, faster updating, or better structural localization - needs a target population, comparator, outcome, and boundary condition. Random splits are insufficient when related samples share a patient, repository, instrument run, season, organization, or simulated design family. Grouped and temporal splits better approximate the novelty encountered after operational use. Stress tests should vary one factor at a time when attribution is important, then combine factors to measure interaction. Repeated runs are necessary whenever decoding, optimization, retrieval, or numerical kernels can vary. Reporting only the best seed or a pooled mean makes operational instability disappear. Seen through simulation lineage, the issue is not merely technical; it determines which claims remain open to challenge.

Next Steps for Evidence Building

Long-term progress depends on cumulative records of both successful and failed approaches. Benchmarks should preserve negative results, failed branches, and changed definitions so later work does not learn only from successful demonstrations. Shared reporting should include data lineage, versioned validation code, confidence intervals, and enough error material to reproduce major conclusions. Prospective studies should predefine monitoring and stopping rules, then measure how people adapt to the application. Finally, researchers should compare simple baselines with complex architectures under equivalent information. Complexity is justified when it adds a measurable capability - for example, structural localization, calibrated deferral, or incremental updating - not merely when it creates a richer narrative around the same errors. For the present focus, simulation lineage therefore becomes a criterion for deciding which evidence deserves further scrutiny.

Future validation sets should be organized around plausible mechanisms of failure. Cases should represent unsupported regions, ambiguous evidence, conflicting modalities, rare but costly conditions, and realistic adversarial transformations. Each case needs provenance and a reason for inclusion. The second priority is intervention testing: when uncertainty is detected, compare additional computation, retrieval, alternative reasoning paths, model ensembles, and human escalation under the same resource budget. This reveals whether a proposed safeguard actually changes the decision or only changes the explanation. The third priority is transfer. A method should be re-evaluated across sites, devices, languages, organizations, or physical regimes before broad claims are made. This point narrows the research task for parameterized industrial assemblies: compare alternatives under the same information and resource constraints.

Institutional Conditions for Reliability

Oversight requires its own capacity, interface, and evaluation plan. Reviewers need enough context to challenge the output, but not so much generated rationale that weak evidence is obscured by volume. Escalation queues require prioritization and workload limits. A reject option that sends nearly everything to experts is safe only on paper; a reject option that rarely fires may be equally ineffective. For parameterized industrial assemblies, oversight should therefore be evaluated with realistic staffing, time pressure, and disagreement. The system should record the final action and its reason without turning that record into surveillance unrelated to the stated purpose. Contestability, privacy, and security are properties of this institutional process as much as of the predictive component. This requirement gives practical content to the article's central question: Which data transformations must be traceable before a fatigue surrogate can support design?

A governance layer translates validation results into permissions and constraints. A operational use specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the deployed pipeline. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream outcomes. A rising override rate may indicate model drift, a changed pipeline, or new user behavior; treating it only as operator noncompliance wastes evidence. Incident review should reconstruct the entire chain, including the model and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. Seen through simulation lineage, the issue is not merely technical; it determines which claims remain open to challenge.

Conclusion

Simulation lineage is credible only when it is attached to an clearly stated evidence chain and decision policy. Across the reviewed studies, several mechanisms can strengthen that chain: structured exploration, typed graphs, conditional revision, adaptive computation, physical constraints, and repeated validation. A reported benchmark gain still leaves the boundary of safe transfer to be established. For parameterized industrial assemblies, the practical standard should be a traceable process that measures shift and instability, supports design screening, uncertainty-aware optimization, targeted simulation, and physical testing, and preserves enough evidence for an independent reviewer to reconstruct failure. Future work should compare safeguards under realistic resource and staffing limits, publish negative and subgroup results, and treat monitors and automated judges as components requiring their own validation. The final standard is not uninterrupted automation but evidence-linked action with clearly stated deferral and independent verifiability. Seen through simulation lineage, the issue is not merely technical; it determines which claims remain open to challenge.

References

1. Ma, Mingjun, et al. "MuSK: Multi-Scale Knowledge Learning for Provenance-Graph Anomaly Detection." *Computer Networks* 289 (2026): 112728.

2. Li, Yuanhao, et al. "BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models." *arXiv preprint arXiv:2605.09134* (2026).

3. Wang, Lizhe, et al. "Machine Learning-Based Fatigue Life Evaluation of the Pump Spindle Assembly With Parametrized Geometry." *ASME 2023 International Mechanical Engineering Congress and Exposition* 87684 (2023): V011T12A022.

4. Fatemi, Ali, and Darrell F. Socie. "A Critical Plane Approach to Multiaxial Fatigue Damage Including Out-of-Phase Loading." *Fatigue & Fracture of Engineering Materials & Structures*, vol. 11, no. 3, 1988, pp. 149-165.

5. Geers, M. G. D., V. G. Kouznetsova, and W. A. M. Brekelmans. "Multi-Scale Computational Homogenization: Trends and Challenges." *Journal of Computational and Applied Mathematics*, vol. 234, no. 7, 2010, pp. 2175-2182.

6. Kennedy, Marc C., and Anthony O'Hagan. "Bayesian Calibration of Computer Models." *Journal of the Royal Statistical Society: Series B*, vol. 63, no. 3, 2001, pp. 425-464.

7. Saltelli, Andrea, et al. "Variance Based Sensitivity Analysis of Model Output: Design and Estimator for the Total Sensitivity Index." *Computer Physics Communications*, vol. 181, no. 2, 2010, pp. 259-270.

8. Breiman, Leo. "Random Forests." *Machine Learning*, vol. 45, 2001, pp. 5-32.

9. Friedman, Jerome H. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics*, vol. 29, no. 5, 2001, pp. 1189-1232.

10. Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, vol. 30, 2017.

11. Bendsoe, Martin P., and Ole Sigmund. *Topology Optimization: Theory, Methods, and Applications*. Springer, 2003.

12. Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018.

13. Suresh, S. *Fatigue of Materials*. 2nd ed., Cambridge University Press, 1998.