Incremental Threat Intelligence for Long-Lived Urban Systems

Alan Gonzales, Juan Bryant, Albert Alexander*

* Corresponding author: Albert Alexander

Review Article | Systems, Networks and Secure Computing | ISCSR Research Publishing | Published 23 September 2026

ABSTRACT

How can new threat reports be absorbed without forgetting older campaigns? This review answers by treating continual learning as a property of a sociotechnical workflow rather than a feature that can be read from average accuracy. The focal setting is critical infrastructure updated over decades, where an optimized service can create privacy, security, and distributional harms that are invisible in its technical objective. Evidence from the assigned publications is synthesized with foundational studies of calibration, distribution shift, causal structure, and responsible deployment. Four requirements follow: preserve the lineage of environmental sensing, mobility and service records, infrastructure graphs, threat reports, and planning constraints; measure stability across relevant perturbations; connect confidence to a specific action; and maintain a route for human challenge and correction. The framework distinguishes descriptive performance from decision utility and separates uncertainty about the world from uncertainty created by the model and its evaluator. It also shows why faster inference or richer reasoning is valuable only when it improves a defined decision under a transparent resource budget. The article is a literature review and research agenda, not a report of a newly completed trial.

KEYWORDS

incremental threat intelligence; long-lived urban systems; threat; reports; infrastructure; service; decision

Introduction

How can new threat reports be absorbed without forgetting older campaigns? The problem resists a learned component-only answer; once deployed, a predictor becomes part of a wider decision process. The visible result sits at the end of choices about measurement, encoding, comparison, computation, and intervention. In critical infrastructure updated over decades, these choices interact with institutions and physical processes that the learned component 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. The resulting argument frames continual learning as an evidence problem. The inquiry focuses on the minimum evidence and comparison set required for a credible claim. For the present focus, continual learning therefore becomes a criterion for deciding which evidence deserves further scrutiny.

The unit of analysis is deliberately procedural. Its unit is the coupled system of sensors, ecological processes, data platforms, vendors, agencies, algorithms, and affected communities. 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 further clarifies what can and cannot be transferred among the target studies. Although their application settings differ, they each make some part of an inference chain more documented - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. The key test is whether that documented component remains meaningful when environmental sensing, mobility and service records, infrastructure graphs, threat reports, and planning constraints change, and whether the proposed safeguards lead to resilient planning, privacy controls, contestable recommendations, and staged incident response in place of merely producing another metric. This requirement gives practical content to the article's central question: How can new threat reports be absorbed without forgetting older campaigns?

A Layered Model of Reliability

Four analytically distinct layers organize the problem. 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 resilient planning, privacy controls, contestable recommendations, and staged incident response. 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 continual learning testable because each claim can be assigned to a component and a measurable transition. This requirement gives practical content to the article's central question: How can new threat reports be absorbed without forgetting older campaigns?

A further distinction separates aleatory variation, epistemic uncertainty, and strategic adaptation. 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 deployed operational sequence 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 testing 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 instead of presenting confidence as a decorative number. This requirement gives practical content to the article's central question: How can new threat reports be absorbed without forgetting older campaigns?

A Broader Evidence Base

Several established literatures clarify the design space. Green-infrastructure evidence connects ecosystem services with physical and psychological health rather than treating vegetation as decoration (Tzoulas et al. 2007). Green-infrastructure planning emphasizes connected systems and long-term stewardship over isolated projects (Benedict and McMahon 2006). Differential privacy provides a formal vocabulary for limiting how much an individual's data can change an output (Dwork 2006). Federated optimization reduces raw-data centralization but leaves open questions about leakage, participation, and heterogeneous clients (McMahan et al. 2017). Together these findings imply that continual learning should be evaluated against explicit alternatives and failure costs. In Incremental Threat Intelligence for Long-Lived Urban Systems, this literature supplies a specific test of continual learning within critical infrastructure updated over decades.

The evidence base offers complementary tests rather than one universal metric. Collaborative learning can distribute training while still requiring explicit threat models for shared updates (Shokri and Shmatikov 2015). Federated-learning surveys document unresolved statistical, systems, privacy, and governance problems (Kairouz et al. 2021). Sociotechnical fairness work warns that optimization can erase the institutions and people that give a metric meaning (Selbst et al. 2019). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In Incremental Threat Intelligence for Long-Lived Urban Systems, this literature supplies a specific test of continual learning within critical infrastructure updated over decades.

A credible assurance case can be built from findings that operate at different levels. Urban models accumulate dependencies on changing sensors, vendors, and administrative definitions (Sculley et al. 2015). Privacy risk management extends beyond secrecy to governability, predictability, and individual consequences (NIST 2020). Smart-city scholarship frames urban intelligence as a combination of sensing, modeling, institutions, and citizen action (Batty et al. 2012). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In Incremental Threat Intelligence for Long-Lived Urban Systems, this literature supplies a specific test of continual learning within critical infrastructure updated over decades.

Evidence From the Focal Studies

A useful test case is Trident: Dual-Stream APT Attribution over Heterogeneous Threat Knowledge Graphs. Rather than treating its output as an isolated score, the study frames how dynamic cyber-threat intelligence can attribute campaigns while combining indicators, tactics, and relational context. Its central mechanism is verified event extraction, a locally updated heterogeneous threat graph, an R-GCN structural stream, a Transformer over normalized TTP sequences, confidence-weighted fusion, and replay for incremental learning, and its evidential basis is that a curated corpus of 1,424 reports covering 15 APT groups supports comparison with a graph baseline and tests incremental updating against full retraining. The dual stream addresses both infrastructure reuse and technique overlap while keeping update cost tied to newly arrived intelligence. For critical infrastructure updated over decades, 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. Attribution labels are historically and politically contingent, and a curated report corpus may encode publication, language, and analyst-confirmation biases. (Wang et al. 2026).

Evidence for continual learning 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 Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift is not a generic claim that learning improves performance. It is the more specific proposition that how an inference system should revise its quantization policy when deployment data no longer resembles calibration data. The proposed route is an adaptive quantization strategy that treats numerical precision as a deployment control rather than a fixed compression choice. Support comes from the fact that the study is framed around robustness under distribution shift and therefore makes the stability of low-precision inference, not compression alone, the central outcome. It connects efficient inference with the broader problem of shift-aware reliability. Read through the lens of continual learning, the study also illustrates why a reported average must remain connected to the workflow that produced it. Its conclusions should be tested across architectures, bit widths, hardware kernels, and shifts that were not represented during calibration. (Sang 2026).

Green Infrastructure Integration in Smart City Planning and Development asks how smart-city data and optimization can support green infrastructure without reducing ecological planning to a static map. It uses dynamic coupling of ecological parameters and urban data, multi-objective adaptive weighting, and reinforcement learning for evolving planning choices. The relevant evidence is that the accessible paper summary reports experiments in large eastern Chinese cities and evaluates ecological benefit, resource efficiency, and cost control. The framework places ecological and technical feedback in the same planning loop. In the present review, this work matters because continual learning cannot be evaluated as an abstract virtue; it must change how evidence is collected and how an action is withheld. Transfer across climates, governance systems, and distributional priorities requires evidence beyond a limited urban sample and model-selected objectives. (Liu and Ying 2026).

An Evaluation Protocol

Metric choice must be derived from the decision. 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 continual learning, the minimum report should include overall performance, slice-level results, uncertainty intervals, coverage-risk curves, stability across runs, and failure examples linked back to environmental sensing, mobility and service records, infrastructure graphs, threat reports, and planning constraints. If the system updates incrementally, the assessment 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. In critical infrastructure updated over decades, this distinction should be tested before it changes an operational decision.

Evaluation begins by writing each claim in testable form. 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 implementation. 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. The article's emphasis on continual learning makes that boundary observable and, in principle, auditable.

From Evidence Capture to Escalation

The next design choice concerns where additional computation is worth its cost. 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 deployment 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. The article's emphasis on continual learning makes that boundary observable and, in principle, auditable.

The evidence architecture for critical infrastructure updated over decades 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 retain 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. Seen through continual learning, the issue is not merely technical; it determines which claims remain open to challenge.

Institutional Conditions for Reliability

Human intervention works only when the review task is feasible and consequential. 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 critical infrastructure updated over decades, 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. In critical infrastructure updated over decades, this distinction should be tested before it changes an operational decision.

A governance layer translates validation results into permissions and constraints. A implementation specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the system. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream

outcomes. A rising override rate may indicate model drift, a changed workflow, or new user behavior; treating it only as operator noncompliance wastes evidence. Incident review should reconstruct the entire chain, including the estimator and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. The article's emphasis on continual learning makes that boundary observable and, in principle, auditable.

Next Steps for Evidence Building

A complementary research program should improve how evidence accumulates over time. 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 evaluation 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 system. 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, continual learning therefore becomes a criterion for deciding which evidence deserves further scrutiny.

The next generation of benchmarks should sample mechanisms and boundary conditions explicitly. 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 requirement gives practical content to the article's central question: How can new threat reports be absorbed without forgetting older campaigns?

Conclusion

Continual learning 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 assessment. The evidence also warns against treating the development benchmark as a license for unrestricted use. For critical infrastructure updated over decades, the practical standard should be a traceable pipeline that measures shift and instability, supports resilient planning, privacy controls, contestable recommendations, and staged incident response, 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. Reliability is demonstrated through warranted answers, consequential uncertainty, and a record that another observer can inspect. The article's emphasis on continual learning makes that boundary observable and, in principle, auditable.

References

1. Wang, Zijun, et al. "Trident: Dual-Stream APT Attribution over Heterogeneous Threat Knowledge Graphs." Computers & Security 171 (2026): 105094.

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. Sang, Yinghao. "Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift." Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy (2026): 362-368.

4. Liu, Kun, and Yong Ying. "Green Infrastructure Integration in Smart City Planning and Development." Proceedings of the 2025 5th International Conference on Smart Transportation and City Engineering (STCE 2025) (2026).

5. Tzoulas, Konstantinos, et al. "Promoting Ecosystem and Human Health in Urban Areas Using Green Infrastructure: A Literature Review." Landscape and Urban Planning, vol. 81, no. 3, 2007, pp. 167-178.

6. Benedict, Mark A., and Edward T. McMahon. Green Infrastructure: Linking Landscapes and Communities. Island Press, 2006.

7. Dwork, Cynthia. "Differential Privacy." Automata, Languages and Programming, Springer, 2006, pp. 1-12.

8. McMahan, Brendan, et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data." Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017, pp. 1273-1282.

9. Shokri, Reza, and Vitaly Shmatikov. "Privacy-Preserving Deep Learning." Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015, pp. 1310-1321.

10. Kairouz, Peter, et al. "Advances and Open Problems in Federated Learning." Foundations and Trends in Machine Learning, vol. 14, nos. 1-2, 2021, pp. 1-210.

11. Selbst, Andrew D., et al. "Fairness and Abstraction in Sociotechnical Systems." Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 59-68.

12. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." Advances in Neural Information Processing Systems, vol. 28, 2015.

13. National Institute of Standards and Technology. NIST Privacy Framework: A Tool for Improving Privacy through Enterprise Risk Management, Version 1.0. U.S. Department of Commerce, 2020.

14. Batty, Michael, et al. "Smart Cities of the Future." European Physical Journal Special Topics, vol. 214, 2012, pp. 481-518.