Proper Scoring for High-Stakes Classification Pipelines

Deborah King, Stephanie Wright, Rebecca Scott*

* Corresponding author: Rebecca Scott

Review Article | Journal of Algorithmic Discovery and Applied AI | IASCI Research Publishing | Published 23 September 2026

ABSTRACT

This evidence synthesis examines evaluation incentives in probabilistic forecasts and threat labels. It argues that the appropriate object of evaluation is the chain from public signals and historical records to probability, label, action, and feedback into future data, 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: overconfident predictions can shift markets or defenses and thereby invalidate the data-generating process assumed by the model. 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.

KEYWORDS

proper scoring; high-stakes classification pipelines; evaluation; future; data; shift; monitoring

Introduction

Do chosen metrics reward honest uncertainty or brittle point decisions? A satisfactory answer must look beyond the predictor, because operational use joins several technical and institutional stages. Observation, representation, hypothesis retention, computational effort, and action are separate choices, even when an interface makes them appear as one answer. In probabilistic forecasts and threat labels, 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. The resulting argument frames evaluation incentives as an evidence problem. The task is therefore to identify the evidence that makes a assurance claim testable. The implication is especially consequential in probabilistic forecasts and threat labels, where a small modeling choice can redirect attention and resources.

The argument proceeds from a operational sequence perspective. Its unit is the chain from public signals and historical records to probability, label, action, and feedback into future data. The benefit is that it avoids attributing every problem to the predictor: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. A process-level unit also supports a more precise comparison of the assigned studies. Although their application settings differ, they each make some part of an inference chain more clearly stated - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. The comparison examines whether that clearly stated component remains meaningful when sequential outcomes, behavioral signals, threat reports, relational graphs, and model confidence change, and whether the proposed safeguards lead to calibrated forecasts, deferred attribution, scenario analysis, and controlled updating instead of merely producing another metric. This requirement gives practical content to the article's central question: Do chosen metrics reward honest uncertainty or brittle point decisions?

What the Core Studies Establish

Trident: Dual-Stream APT Attribution over Heterogeneous Threat Knowledge Graphs asks how dynamic cyber-threat intelligence can attribute campaigns while combining indicators, tactics, and relational context. It uses 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. The relevant evidence 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. In the present review, this work matters because evaluation incentives cannot be evaluated as an abstract virtue; it must change how evidence is collected and how an action is withheld. Attribution labels are historically and politically contingent, and a curated report corpus may encode publication, language, and analyst-confirmation biases. (Wang et al. 2026).

A useful test case is BoostAPR: Boosting Automated Program Repair via Execution-Grounded Reinforcement Learning with Dual Reward Models. Rather than treating its output as an isolated score, the study frames how reinforcement learning for program repair can overcome sparse execution feedback and coarse sequence-level credit. Its central mechanism is a three-stage pipeline combining execution-verified supervised traces, sequence- and line-level reward models, and PPO with reward redistribution to critical edit regions, and its evidential basis is that 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. For probabilistic forecasts and threat labels, 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. Benchmark success does not by itself establish maintainability, specification fidelity, or robustness to incomplete and flaky tests. (Li et al. 2026).

Evidence for evaluation incentives becomes concrete in A New Playing Method of the Guessing Football Lottery. The paper addresses how buyer opinions and alternative ticket structures might affect football-lottery prediction and betting behavior through a proposed 'sky ladder' playing strategy evaluated through feasibility analysis and simulations using back-propagation and LSTM neural networks. Its evaluation is informative because the reported simulations are used to argue that the new format offers flexibility while controlling the risks faced by buyers and lottery operators. The paper links predictive modeling with product design and behavioral response instead of treating match probabilities as the only object of study. 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, simulation performance cannot establish bettor welfare or market effects without prospective calibration, behavioral data, and comparison against simple probabilistic baselines. The appropriate use of the result is therefore bounded, comparative, and explicit about unresolved deployment conditions (Liu et al. 2020).

Relevant Foundations

Several established literatures clarify the design space. Poisson score models provide a transparent baseline for separating team strengths from match randomness (Maher 1982). Dependence corrections and time weighting show how modest domain structure can improve football forecasts (Dixon and Coles 1997). Dynamic team-strength models make temporal change explicit instead of absorbing it into unexplained error (Rue and Salvesen 2000). Elo ratings offer a parsimonious sequential benchmark for match-result prediction (Hvattum and Arntzen 2010). Together these findings imply that evaluation incentives should be evaluated against explicit alternatives and failure costs. In Proper Scoring for High-Stakes Classification Pipelines, this literature supplies a specific test of evaluation incentives within probabilistic forecasts and threat labels.

The evidence base offers complementary tests rather than one universal metric. The Brier score evaluates probabilistic accuracy rather than rewarding only the most likely categorical outcome (Brier 1950). Proper scoring rules encourage honest probabilities and expose overconfident forecasts (Gneiting and Raftery 2007). Calibration links repeated probability statements to observed frequencies over an appropriate reference class (Dawid 1982). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In Proper Scoring for High-Stakes Classification Pipelines, this literature supplies a specific test of evaluation incentives within probabilistic forecasts and threat labels.

A credible assurance case can be built from findings that operate at different levels. Kelly's criterion connects probabilistic advantage to bankroll growth while making estimation error consequential (Kelly 1956). Football-specific evaluation research shows that metric choice can change apparent model quality (Constantinou and Fenton 2013). Dataset shift can degrade both predictions and uncertainty estimates when seasons, teams, or markets change (Ovadia et al. 2019). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In Proper Scoring for High-Stakes Classification Pipelines, this literature supplies a specific test of evaluation incentives within probabilistic forecasts and threat labels.

Conceptual Foundations

Reliability analysis must not collapse aleatory variability, epistemic uncertainty, and strategic response into one category. 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 system 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 evaluation 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. In probabilistic forecasts and threat labels, this distinction should be tested before it changes an operational decision.

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 calibrated forecasts, deferred attribution, scenario analysis, and controlled updating. 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 testing incentives testable because each claim can be assigned to a component and a measurable transition. The implication is especially consequential in probabilistic forecasts and threat labels, where a small modeling choice can redirect attention and resources.

Testing Claims Rather Than Scores

A defensible protocol starts from clearly stated claims instead of available metrics. 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. The implication is especially consequential in probabilistic forecasts and threat labels, where a small modeling choice can redirect attention and resources.

No single metric can stand in for the downstream action. 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 validation incentives, the minimum report should include overall performance, slice-level results, uncertainty intervals, coverage-risk curves, stability across runs, and failure examples linked back to sequential outcomes, behavioral signals, threat reports, relational graphs, and model confidence. If the application updates incrementally, the validation 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 point narrows the research task for probabilistic forecasts and threat labels: compare alternatives under the same information and resource constraints.

Deployment Controls

Measurements become useful only when they alter operational use limits. A operational use specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the application. 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 learned component and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. The implication is especially consequential in probabilistic forecasts and threat labels, where a small modeling choice can redirect attention and resources.

Placing a person in the loop does not by itself create effective control. 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 probabilistic forecasts and threat labels, 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 model. This point narrows the research task for probabilistic forecasts and threat labels: compare alternatives under the same information and resource constraints.

A Traceable System Architecture

Operational design in probabilistic forecasts and threat labels 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. For the present focus, evaluation incentives therefore becomes a criterion for deciding which evidence deserves further scrutiny.

Computational effort should vary with evidential need. 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 field use 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. Seen through evaluation incentives, the issue is not merely technical; it determines which claims remain open to challenge.

Experiments the Field Still Needs

Research should begin with cases selected for what they can reveal about failure, not merely for availability. 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. The implication is especially consequential in probabilistic forecasts and threat labels, where a small modeling choice can redirect attention and resources.

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. In probabilistic forecasts and threat labels, this distinction should be tested before it changes an operational decision.

Conclusion

Evaluation incentives is credible only when it is attached to an documented evidence chain and decision policy. The focal literature contributes several ways to improve the chain: structured exploration, typed graphs, conditional revision, adaptive computation, physical constraints, and repeated evaluation. Their limitations make clear that benchmark scope and field use scope are different. For probabilistic forecasts and threat labels, the practical standard should be a traceable workflow that measures shift and instability, supports calibrated forecasts, deferred attribution, scenario analysis, and controlled updating, 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 documented deferral and independent verifiability. The implication is especially consequential in probabilistic forecasts and threat labels, where a small modeling choice can redirect attention and resources.

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. Liu, Shunqi, et al. "A New Playing Method of the Guessing Football Lottery." IOP Conference Series: Materials Science and Engineering 790.1 (2020): 012100.

4. Maher, M. J. "Modelling Association Football Scores." Statistica Neerlandica, vol. 36, no. 3, 1982, pp. 109-118.

5. Dixon, Mark J., and Stuart G. Coles. "Modelling Association Football Scores and Inefficiencies in the Football Betting Market." Journal of the Royal Statistical Society: Series C, vol. 46, no. 2, 1997, pp. 265-280.

6. Rue, Havard, and Oyvind Salvesen. "Prediction and Retrospective Analysis of Soccer Matches in a League." Journal of the Royal Statistical Society: Series D, vol. 49, no. 3, 2000, pp. 399-418.

7. Hvattum, Lars Magnus, and Halvard Arntzen. "Using ELO Ratings for Match Result Prediction in Association Football." International Journal of Forecasting, vol. 26, no. 3, 2010, pp. 460-470.

8. Brier, Glenn W. "Verification of Forecasts Expressed in Terms of Probability." Monthly Weather Review, vol. 78, no. 1, 1950, pp. 1-3.

9. Gneiting, Tilmann, and Adrian E. Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation." Journal of the American Statistical Association, vol. 102, no. 477, 2007, pp. 359-378.

10. Dawid, A. P. "The Well-Calibrated Bayesian." Journal of the American Statistical Association, vol. 77, no. 379, 1982, pp. 605-610.

11. Kelly, J. L., Jr. "A New Interpretation of Information Rate." Bell System Technical Journal, vol. 35, no. 4, 1956, pp. 917-926.

12. Constantinou, Anthony C., and Norman E. Fenton. "Solving the Problem of Inadequate Scoring Rules for Assessing Probabilistic Football Forecast Models." Journal of Quantitative Analysis in Sports, vol. 9, no. 3, 2013, pp. 209-222.

13. Ovadia, Yaniv, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift." Advances in Neural Information Processing Systems, vol. 32, 2019.