Human-in-the-Loop Assurance with Neuromorphic Explanations for Safety-Critical Systems

Sylvester Kaczmarek
Department of Computing, Imperial College London
London, United Kingdom
research@sylvesterkaczmarek.com

Abstract

Private oversight of safety-critical AI is limited by proprietary models and opaque decision traces. We present a compact assurance artefact that links neuromorphic explanations to verifiable runtime checks, enabling oversight without exposing model internals. Each decision emits a signed *Assurance Evidence Object* (AEO) binding input provenance, safety limits, and an explanation with a verifier-checked fidelity score. In a controlled study with $n = 90$, explanations raised operator diagnostic accuracy from $61.2\,\%$ to $> 82\,\%$ and trust from 2.8 to > 4.1 on a five-point scale. Hardware-in-the-loop measurements show $1.2\,\mathrm{ms}$ median latency and $45\,\mu\mathrm{J}$ per inference, with stable behavior under timing jitter and bit-flip surrogates. We provide a verifier and a one-page Trust Card that packages these signals for procurement and audit evidence review.

1 Governance need for verifiable assurance

Auditors, insurers, and buyers need evidence that a deployed system respects safety limits and that a human can supervise it effectively, yet model internals are often inaccessible [2,6]. We attach verifiable evidence to each critical decision and summarize it in a governance-ready *Trust Card*, inspired by model reporting cards [5].

2 Assurance Evidence Object pipeline

AEO fields. For each decision d_t, the system emits an AEO including: (i) input hash (provenance), (ii) decision and confidence, (iii) safety-limit check results (pass/fail per rule), (iv) explanation digest (hash of the explanation payload), (v) fidelity score (defined below), (vi) monitor-state digest (hash of runtime monitor alerts and sampling status), (vii) trusted timestamp, (viii) cryptographic signature, and (ix) chain identifier. These fields support end-to-end auditability from input provenance to safety conformance and operator-facing rationale.

Neuromorphic explanations. We use *neuromorphic explanations* to mean explanations native to spiking models, specifically spike-level LRP attribution heatmaps [9] (which input features contributed most) and temporal attention windows over the decision horizon (when the decisive evidence occurred).

Verification. A lightweight, weight-agnostic verifier checks signature validity and chain continuity, input–decision binding, limit conformance, and *monitor coverage* (that required runtime monitors were active and produced a state digest for the relevant window) [10]. *Fidelity score* is the normalized agreement between the explanation's top-k attributed features and the model's behavior under perturbations of those features [7]; scores below threshold τ fail verification, guarding against uninformative explanations [8]. We quantify fidelity by masking the top-k attributed features and

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: First EurIPS Workshop on Private AI Governance.

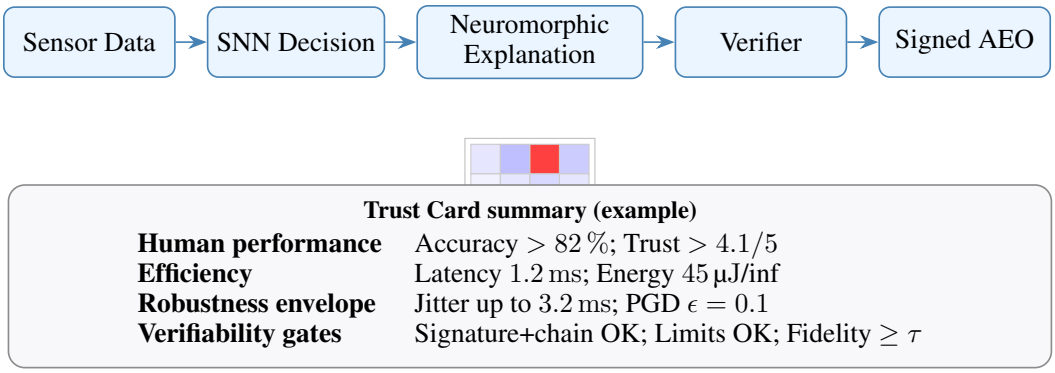

Figure 1: Per-decision assurance pipeline (top) producing a signed AEO. The explanation stage includes an example attribution map. The Trust Card (bottom) aggregates verifiable fields and outcomes into a one-page artifact for evidence review.

measuring the relative drop in model confidence. Verification cost is linear in the number of checks and requires no access to proprietary weights. Figure 1 summarizes the evidence pipeline and the resulting Trust Card.

3 Evaluation

Human study. We conducted a between-subjects study with four conditions ($n = 90$): (i) AI output only, (ii) AI + explanation, (iii) AI + verifier status only, (iv) AI + explanation + verifier status. Operators diagnosed simulated robotic anomalies (accuracy is correct fault-class identification). Explanations improved diagnostic accuracy from $61.2\,\%$ to $> 82\,\%$ (ANOVA with Tukey HSD; $p < 0.01$, large effect) and increased self-reported trust from 2.8 to > 4.1 on a five-point scale. Calibration error decreased, and NASA-TLX indicated lower cognitive workload [3].

Hardware and robustness. Deployed on a neuromorphic edge processor, the system achieved $1.2\,\mathrm{ms}$ median decision latency and $45\,\mu\mathrm{J}$ per inference under load. Safety limits held under timing-jitter perturbations up to $3.2\,\mathrm{ms}$, synthetic bit-flip surrogates at memory interfaces [1], and iterative PGD input attacks ($\epsilon = 0.1$, 20 steps) [4].

4 The Trust Card

The one-page report aggregates fields a reviewer can act on: the threat model and safety limits, a summary of the evaluation and attack suite, validated human-in-the-loop outcomes, latency and energy under load, the explanation fidelity threshold τ, pass/fail gates, and verifier results for sampled AEOs.

5 Governance use and conclusion

Procurement can require a Trust Card and sampled signed AEOs at delivery milestones, supporting structured evidence review against explicit pass/fail gates. The same artifacts can support audit workflows and risk assessments, alongside independent testing, process controls, and incident reporting. This approach turns human-in-the-loop evidence and runtime limits into verifiable, governance-ready signals without exposing proprietary model weights, enabling practical private oversight [2].

Acknowledgments

The author thanks the workshop reviewers for constructive feedback, and the study participants.

References

[1] Kaczmarek, S. (2025). A Bio-Inspired Hierarchical Temporal Defense for Securing Spiking Neural Networks. In *Proceedings of the NeurIPS 2025 Workshops: Machine Learning and the Physical Sciences (ML4PS)*.

[2] Ball, D. W. (2025). A Framework for the Private Governance of Frontier Artificial Intelligence. *arXiv preprint* arXiv:2504.11501.

[3] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*, pp. 139–183. North-Holland.

[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*.

[5] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pp. 220–229. ACM.

[6] Weil, G., Pistillo, M., Van Arsdale, S., Ikegami, J., Onuma, K., Okawa, M., & Osborne, M. A. (2024). Insuring emerging risks from AI. *LawAI Working Paper Series*, No. 4-2024.

[7] Samek, W., Binder, A., Montavon, G., Bach, S., & Müller, K.-R. (2015). Evaluating the visualization of what a deep neural network has learned. *arXiv preprint* arXiv:1509.06321.

[8] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:1810.03292.

[9] Sun, M., Zhang, D., Ge, Z., Wang, J., Li, J., Fang, Z., & Xu, R. (2024). EventRPG: Event data augmentation with relevance propagation guidance. *arXiv preprint* arXiv:2403.09274.

[10] Cofer, D., Amundson, I., Sattigeri, R., Passi, A., Boggs, C., Gilham, L., Byun, T., & Rayadurgam, S. (2020). Run-time assurance for learning-enabled systems. In *NASA Formal Methods (NFM 2020)* (LNCS 12229), pp. 361–368. Springer. `https://doi.org/10.1007/978-3-030-55754-6_21`.