Imperial College London

VERIFIABLE CONTROL AND CALIBRATED TRUST IN EMBODIED NEUROMORPHIC AGENTS FOR SAFETY-CRITICAL APPLICATIONS

Sylvester Kaczmarek[1]
[1]Department of Computing, Imperial College London
research@sylvesterkaczmarek.com

THE CHALLENGE

Embodied agents must satisfy strict bounds on latency, energy, and safety. Millisecond control loops and certification constraints make stochastic policies difficult to deploy directly. Operators also need usable explanations under time pressure.

Edge constraints

- **Real time:** millisecond timing bounds.
- **SWaP:** microjoule-scale energy per decision.
- **Safety:** formally specified safety envelope under faults and attacks.
- **Oversight:** explanations that improve diagnosis without increasing workload.

FIGURE 1

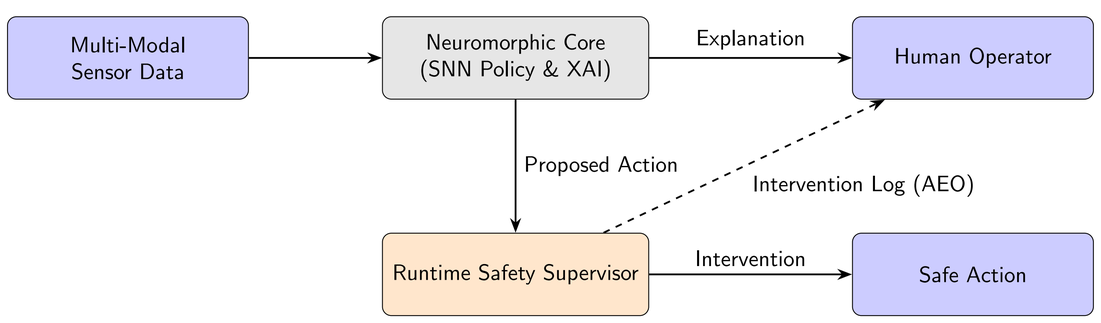

Figure 1. Integrated assurance architecture. A neuromorphic core outputs an action and an explanation. A runtime supervisor enforces a formal safety envelope via monitor and shield logic. Interventions emit an Assurance Evidence Object (AEO) for audit.

APPROACH

- **Neuromorphic core:** spiking policy for low-latency, low-energy inference.
- **Runtime supervisor:** deterministic guards and safe maneuvers enforce the safety envelope at runtime.
- **Human interface:** explanation portfolio (LRP, temporal attention, surrogate rules) to support trust calibration.

Definitions

- **Verifiable control:** zero unmitigated safety envelope violations in trials.
- **Calibrated trust:** higher operator performance with aligned subjective trust, assessed with NASA-TLX workload.

VALIDATION

HIL: policy on BrainChip Akida, supervisor on a co-located MCU, stressed with compounded faults and adversarial perturbations.
Human study: $n = 90$ participants diagnose anomalies with no explanation or one explanation modality.

THREAT MODEL AND SAFETY ENVELOPE

Covered in evaluation

- **Faults:** radiation-like upsets, sensor drift, sensor noise, degraded links.
- **Sensor-space:** LiDAR PGD perturbations with physically plausible edits.
- **Temporal-space:** spike timing jitter bounded to ±4 ms.

Enforcement: guard conditions are checked each control cycle; predicted violations trigger a shielded safe maneuver and an AEO log entry.

FIGURE 2

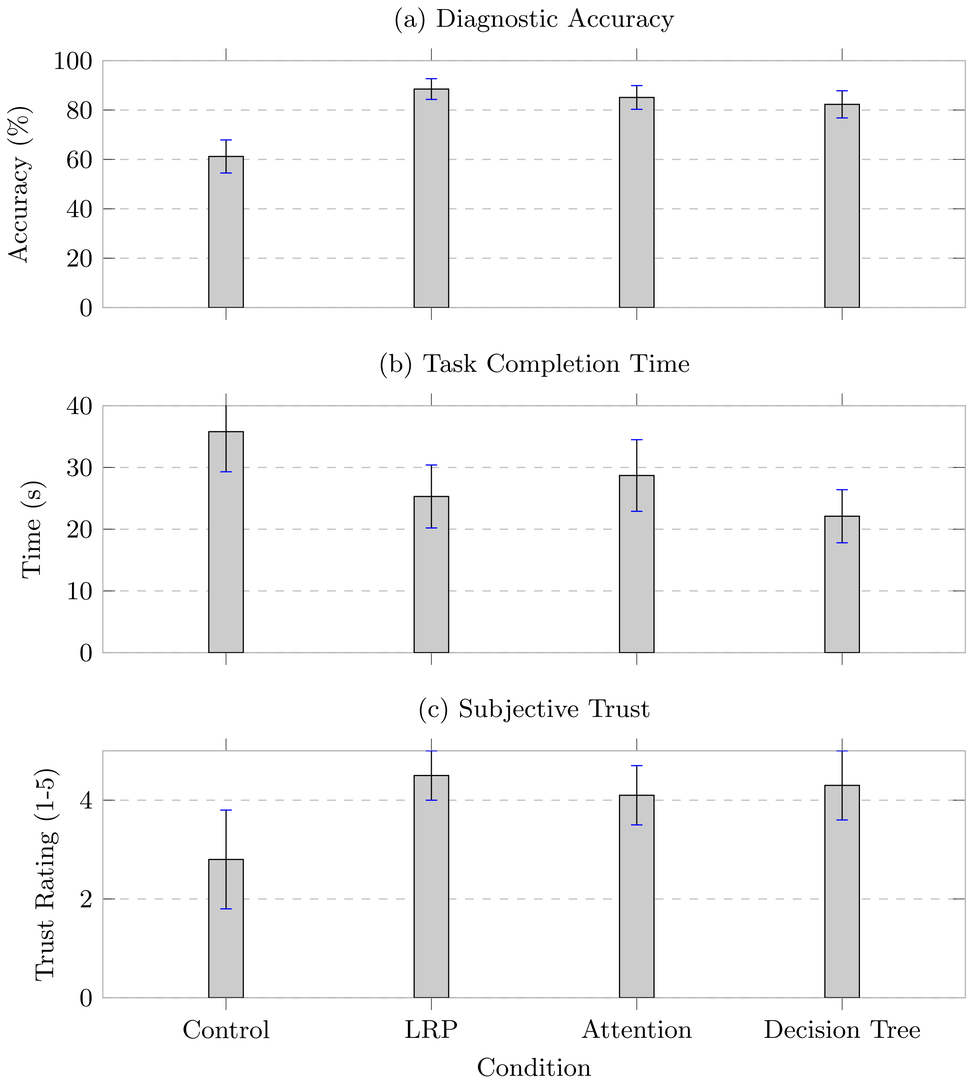

Figure 2. Human-subject outcomes ($n = 90$). Explanations improve accuracy, reduce time, and increase trust (95% CIs).

RESULTS

System performance under combined stress conditions

Table 1. Key metrics

Metric	Proposed	Baseline SNN	LSTM
Mission success rate	**92%**	43%	35%
Attack success rate (LiDAR PGD)	**22.4%**	78.1%	85.6%
End-to-end latency (P99)	4.8 ms	>30 ms	>40 ms
Energy per inference	45 µJ	67 µJ	8.5 mJ
Unmitigated safety envelope violations	**0**	—	—

Human oversight outcomes

- **Accuracy:** 61.2% (control) to 88.5% (best explanation condition).
- **Trust:** 2.8 to 4.5 on a 5-point Likert scale.
- **Efficiency and workload:** time 35.8 s to 22.1 s; NASA-TLX 68 to 41 (best workload condition).

Takeaway: safety is enforced by auditable runtime logic, and explanations measurably improve operator performance.

SELECTED PUBLICATION

S. Kaczmarek, Verifiable Control and Calibrated Trust in Embodied Neuromorphic Agents for Safety-Critical Applications, *Proc. AAAI 2026 Workshops (TrustAgent)*, 2026.