Verifiable Control and Calibrated Trust in Embodied Neuromorphic Agents for Safety-Critical Applications

Sylvester Kaczmarek

Department of Computing
Imperial College London
research@sylvesterkaczmarek.com

Abstract

Agentic systems for physical world applications must satisfy strict bounds on latency, energy, and safety. We present an embodied agent architecture, built on spiking neural networks and integrated with a runtime safety supervisor and a calibrated human-in-the-loop interface. The system provides verifiable control through formal safety envelopes and produces auditable evidence objects at runtime. Hardware-in-the-loop validation on a BrainChip Akida neuromorphic processor demonstrates a 92% mission success rate under combined adversarial and environmental faults, with a median inference latency of 1.2 ms and an energy of 45 μJ per inference. A 90-participant study confirms its utility for human oversight, where integrated explanations increased operator diagnostic accuracy from 61.2% to 88.5% and subjective trust from 2.8 to 4.5 on a five-point Likert scale, with cognitive load assessed via the NASA-TLX framework. The findings establish a practical and verifiable paradigm for embodied agency at the edge, complementing LLM-centric architectures by providing a deployable solution for applications where safety and resource efficiency are paramount.

1 Introduction

Recent advances in large language models have catalyzed the development of agentic AI systems capable of complex reasoning, planning, and tool use. These systems, however, are predominantly designed for and evaluated in digital environments, where computational resources are abundant and the consequences of failure are contained. Their reliance on massive, cloud-hosted models introduces significant and variable latency, a characteristic unsuitable for control loops that require millisecond-scale responsiveness to maintain stability in physical systems. Local language-model agents reduce network delay, yet their compute and memory footprints still exceed typical edge power budgets and make worst-case timing guarantees impractical for millisecond control. Furthermore, the stochastic nature of their generative outputs makes their behavior difficult to verify formally, posing a fundamental challenge for certification in safety-critical applications. This creates a distinct embodied gap for agentic AI, where the requirements of the physical world demand a different architectural foundation.

For agents deployed in high-stakes environments such as autonomous space robotics, trustworthiness is not an abstract goal but a set of non-negotiable engineering requirements. First, the system must provide verifiable safety. Its actions must be guaranteed to remain within a predefined safety envelope, regardless of the internal state of its complex, probabilistic core. This necessitates a mechanism for runtime monitoring and enforcement of formal constraints. Second, the agent must operate in real time. Decisions and interventions must occur on millisecond timescales to interact with and control physical dynamics effectively. This requires a computational architecture with predictable, low-latency performance. Third, the system must enable effective human oversight. This extends beyond simple monitoring to providing human operators with the necessary transparency to understand the agent's reasoning, calibrate their trust in its outputs, and intervene correctly during off-nominal events. The interface must support rapid and accurate human diagnosis without imposing an excessive cognitive load. Together, these imperatives demand an integrated approach to agent design that prioritizes safety, responsiveness, and human-AI collaboration from the outset (Bansal et al. 2021).

This paper presents an integrated assurance framework and an embodied agent architecture designed to meet these safety-critical requirements. At its core is a neuromorphic agent based on a spiking neural network (SNN). Its event-driven, asynchronous nature provides the foundation for the extreme energy efficiency and low-latency processing necessary for edge deployment. This probabilistic agent is governed by a deterministic runtime safety supervisor. The supervisor implements a monitor-shield pattern, continuously checking the agent's state against formal guard conditions and overriding its actions with a predefined safe maneuver if a violation is detected. This provides the verifiable control component of the framework. To enable effective oversight, the system includes an integrated explainability (XAI) module and a human-in-the-loop validation protocol. We demonstrate through a 90-participant study that providing operators with spike-level explanations of the agent's reasoning leads to a quantifiable improvement in diagnostic accuracy and calibrated trust. The complete architecture, illustrated in Figure 1, composes these elements into a single, cohesive system. We present this integrated system not as a replacement for LLM-based agents, but as a necessary and complementary paradigm for embodied agency at the safety-critical edge. We evaluate the framework against three measurable properties: task success under compounded disturbances, end-to-end timing and energy, and the calibration of human trust and performance.

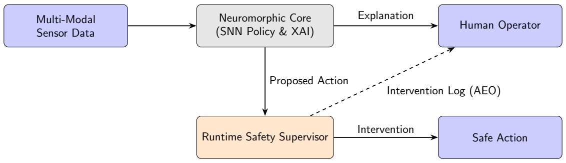

Figure 1: The integrated architecture for a verifiable and trustworthy embodied agent. Sensor data is processed by the neuromorphic core, which generates both a policy action and an explanation. The runtime safety supervisor monitors the agent's state and proposed action against a formal safety envelope, either permitting the action or intervening with a safe maneuver. The explanation is provided to a human operator to support oversight and calibrated trust.

2 The Challenge of Verifiability and Trust at the Edge

The development of trustworthy embodied agents requires a direct confrontation with the physical and operational realities of their intended environments. For systems deployed at the safety-critical edge, these realities impose a set of uncompromising constraints that dictate architectural choices and define the boundaries of feasible autonomy. The challenge is twofold: first, to create an agent that can function effectively within a severely resource-constrained operational envelope, and second, to ensure this agent remains robust and controllable when faced with a complex spectrum of internal failures, environmental hazards, and malicious threats.

2.1 Operational Constraints of the Safety-Critical Edge

The cislunar domain serves as an exemplar of a safety-critical edge environment, where the constraints of power, latency, and communication are pushed to their physical limits (Nesnas, Fesq, and Volpe 2021; Izzo et al. 2022). These factors collectively render agent architectures that depend on continuous access to large, remote computational resources non-viable for real-time control.

First, such agents are subject to extreme Size, Weight, and Power (SWaP) limitations. A robotic platform's power budget for a compute subsystem is often less than 10 W. An agent's core inference loop must therefore be exceptionally efficient, with an energy cost on the order of microjoules per decision, to ensure mission longevity, especially during long periods of eclipse where the system must run on limited battery reserves.

Second, communication latency makes remote decision-making for dynamic tasks impossible. The round-trip light-time delay between the Earth and the Moon is approximately 2.5 to 2.7 seconds. We target onboard control loops with median latency near 1–2 ms and a 99th percentile below 5 ms; any architecture must satisfy these bounds without remote calls. This architectural requirement for local, real-time processing fundamentally precludes any system that relies on a remote server for its primary reasoning or action-selection functions.

Third, communication bandwidth is a scarce and power-intensive resource. A high-gain antenna on a deep-space asset might provide a downlink rate of less than 100 kilobits per second. This makes it infeasible to transmit high-volume raw sensor data for analysis. The agent must be capable of performing data fusion and information processing locally, reserving the communication link for high-level summaries and critical alerts.

2.2 The Threat Model Beyond Benign Failures

A trustworthy agent must be resilient to a wide range of active, off-nominal events. The threat space for a cislunar robotic agent is a composite of environmental hazards, internal system degradation, and intelligent adversarial attacks. The environment itself is a source of faults, with radiation inducing single-event upsets (SEUs) that manifest as transient sensor glitches or corrupted data. Internal systems are also subject to failure, including gradual sensor bias drift and actuator degradation.

Finally, the agent must be robust to intentional attacks. Our attacker model assumes white-box access to the perception stack but no access to the supervisor, which is part of the trusted computing base. Perturbation budgets are bounded by an ℓ_∞ timing jitter of ≤ 4 ms on spike times and a point-cloud edit rate of $\leq 2\%$ of points per frame. This includes the adversarial temporal jitter attack, which involves making subtle manipulations to the timing of input events to cause a misclassification (Sharmin et al. 2020). A robust agent must be designed with an explicit threat model that accounts for this full spectrum of failures, including gradient-based attacks (Goodfellow, Shlens, and Szegedy 2014; Madry et al. 2018).

2.3 Defining Verifiable Control and Calibrated Trust

To meet these challenges, an agent must possess two key, measurable properties: verifiable control and the ability to foster calibrated trust with its human supervisors.

We operationalize **verifiable control** as zero unmitigated safety envelope violations across all experimental trials. This is achieved through a runtime safety supervisor that implements a monitor-shield pattern. The monitor evaluates the agent's state and proposed actions against a set of formally specified safety rules. If a proposed action would violate this envelope, the shield intervenes and executes a predefined safe maneuver. We report the monitor and shield decision latency per control cycle to quantify the responsiveness of this safety mechanism.

We operationalize **calibrated trust** as the statistical alignment between an operator's subjective confidence in the agent and their objective performance when using the system. This is measured in our human-subject study by comparing participants' self-reported trust ratings with their diagnostic accuracy. We report these findings with calibration error metrics and the statistical effect sizes of the improvements gained from the system's explanations. The challenge, therefore, is to build an agent that is not only performant and robust, but is also architecturally controllable and demonstrably trustworthy to its human partners.

3 An Integrated Framework for Embodied Agency

The proposed framework for trustworthy embodied agency is an integrated system composed of three primary components. First, a neuromorphic core agent, based on a spiking neural network (SNN), serves as the high-performance policy and perception engine. Second, a deterministic runtime safety supervisor provides a verifiable control layer that enforces a formal safety envelope around the probabilistic core. Third, a human-in-the-loop interface, equipped with a portfolio of explainable AI (XAI) methods, facilitates effective operator oversight and enables the calibration of trust. These components work in concert to create a system that is simultaneously efficient, safe, and interpretable, addressing the core challenges of deploying autonomous agents in high-stakes environments.

3.1 The Neuromorphic Core Agent

The foundation of the agent is a neuromorphic processing architecture that is inherently suited to the constraints of the safety-critical edge. Unlike traditional Artificial Neural Networks (ANNs) that operate on static vectors and require dense matrix multiplications at every time step, SNNs, the third generation of neural network models (Maass 1997), process information using discrete, asynchronous events, or spikes (Gerstner and Kistler 2002). This event-driven paradigm provides a fundamental architectural advantage for embodied applications. Computation and energy are expended only when new information is present, leading to sparse network activity and extreme power efficiency. Furthermore, the explicit representation of time in SNN dynamics allows the network to learn and react to complex temporal patterns in sensor data with millisecond precision, a critical capability for interacting with the physical world.

The core agent's architecture is a hybrid SNN designed for multi-modal data fusion. The policy has approximately 1.2 million synapses distributed across 5 layers. Inputs are integrated over 20 ms windows with a 1 ms stride, and the system operates on a control cycle of 2 ms. Input layers, composed of computationally efficient Leaky Integrate-and-Fire (LIF) neurons, are responsible for encoding raw sensor data from various modalities into spike trains. Deeper, integrative layers of the network employ Izhikevich neurons, which offer a richer set of dynamical behaviors, such as bursting and adaptation. This heterogeneity allows the network to capture a wider range of temporal features, from the sharp transients of an impact event to the slow drifts of sensor degradation. The network topology includes convolutional SNN layers for extracting spatial features from camera imagery and recurrent connections for processing sequential data from telemetry and inertial sensors.

The agent's resilience is built into its core design through the Hierarchical Temporal Defense (HTD) framework (Kaczmarek 2025a). This multi-layered security architecture provides intrinsic robustness against a wide spectrum of perturbations by combining input-level Bayesian filtering, homeostatic neuronal thresholds, and volatility-gated synaptic plasticity.

Furthermore, the agent is designed for long-term autonomy through the Information-Theoretic Adaptive Morphology (ITAM) framework. This mechanism, guided by principles from the information bottleneck method (Tishby, Pereira, and Bialek 2000), governs the network's structural plasticity, allowing it to adapt to novel, unforeseen events by dynamically adding or pruning neurons and connections. ITAM plasticity is enabled only during specific adaptation phases, with add and prune thresholds set to the 95th and 5th percentiles of the estimated mutual information change over a 2-second horizon. The entire agent is trained on our Cislunar Anomaly and Risk Dataset (CARD) (Kaczmarek 2025b), which comprises over 150 mission scenarios (avg. 10 s duration) with 100 Hz IMU and 10 Hz LiDAR streams, labeled across 5 categories (e.g., collision, sensor degradation). Training proceeds for 200 epochs using the Adam optimizer with a learning rate of 1e-3, L1 regularization for sparsity, and a comprehensive data augmentation protocol (including noise injection $\sigma = 0.05$, rotation $\pm 15^\circ$, and temporal warping) to improve generalization. The training objective minimizes a composite loss $\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{adv}$, where $\mathcal{L}_{CE}$ is cross-entropy, $\mathcal{L}_{L1}$ targets 5% sparsity ($\lambda_1 = 10^{-4}$), and $\mathcal{L}_{adv}$ is the loss on PGD-generated adversarial examples ($\lambda_2 = 0.5$).

3.2 The Runtime Safety Supervisor

The neuromorphic core agent is a high-performance, probabilistic system. To provide the deterministic safety guarantees required for critical applications, it is enclosed within a runtime safety supervisor. This supervisor provides the verifiable control component of the framework, implementing a monitor-shield pattern that decouples the complex, learned policy of the agent from the enforcement of absolute safety constraints (Bloem et al. 2015; Ames et al. 2017).

The first component of the supervisor is the **monitor**. The monitor is a lightweight, deterministic module that runs in parallel with the core agent. It continuously observes a set of critical state variables from the agent and its environment. These variables include the agent's physical state (e.g., position, velocity, orientation), the state of its internal systems (e.g., motor torques, battery temperature), and its proposed next action as determined by the SNN policy. The monitor's function is to evaluate this information against a set of formal guard conditions at every control cycle.

These guard conditions collectively define the agent's **safety envelope**. The safety envelope is a set of simple, human-authored, and formally verifiable rules that specify the bounds of safe operation. These rules are derived from the system's engineering specifications and mission constraints. For the cislunar rover application, examples of such rules include hard limits on kinematic variables (e.g., '$velocity < 0.5$ m/s'), physical constraints (e.g., '`motor_current` < 5 A'), and environmental interactions (e.g., '`predicted_time_to_collision` $>$ 2 seconds'). The logic of the safety envelope is intentionally simple, consisting of straightforward threshold checks and logical predicates that can be formally verified to be correct and can be executed with minimal computational overhead.

The second component of the supervisor is the **shield**. The shield is a simple, finite-state machine that is activated only when the monitor detects a violation of the safety envelope. If the agent's proposed action would lead to a state that breaches a guard condition, the shield intervenes. It preempts and discards the agent's proposed action and instead issues a command for a predefined, provably safe maneuver. The set of safe maneuvers is small and conservative. For a mobile robot, the default safe action is typically '`halt_motion`', which brings the vehicle to a controlled stop. Other safe actions might include '`retract_arm`' or '`enter_low_power_state`'. The shield's logic is also deterministic and formally verifiable, ensuring that its intervention will always transition the system to a known-safe state. To prevent rapid switching (oscillations) between policy and shield control, the supervisor implements a latching mechanism that maintains the safe state for a minimum duration of 2 seconds or until stability conditions are met. The monitor and shield logic execute in 0.35 ms median and 0.48 ms at the 99th percentile per cycle on the host microcontroller; these bounds were verified with cycle-accurate timers. The guard rules are unit-tested against over 10,000 synthetic states per rule, and the supervisor's finite-state machine (5 states, 12 transitions) is validated via exhaustive state enumeration to always transition to a safe sink state.

A key function of this architecture is the generation of an auditable trail for accountability and post-hoc analysis. Every time the shield intervenes, it generates a signed, immutable log entry called an *Assurance Evidence Object* (AEO). This data structure contains a timestamp, the specific guard condition that was violated, a snapshot of the system state at the time of the violation, the unsafe action proposed by the core agent, and the safe action executed by the shield. This creates an unambiguous record of every safety-critical intervention, which is essential for debugging, incident analysis, and providing evidence for certification and regulatory compliance, analogous to the goals of formal assurance cases (Graydon 2018).

3.3 The Human-in-the-Loop Interface and XAI Portfolio

While the runtime supervisor provides a guarantee of low-level safety, effective management of complex, off-nominal situations still requires the expertise of a human operator. The third component of the framework is a human-in-the-loop interface designed to provide operators with the necessary transparency to understand the agent's behavior, make informed decisions, and build a calibrated sense of trust in the autonomous system. This is the calibrated trust component of the framework.

The interface presented to participants in our human-subject study is designed to mimic a mission control dashboard. It features a multi-panel display showing synchronized, time-series plots of key telemetry streams from the agent's sensors. A dedicated alert panel highlights when the neuromorphic core has detected an anomaly, displaying the system's confidence in its detection. The core of the interface, however, is the XAI module, which provides a portfolio of explanations to give operators insight into the agent's decision-making process. Recognizing that different tasks and user expertise levels may benefit from different levels of abstraction, the system offers three distinct types of explanations, as validated in our prior work.

The first explanation method is **Layer-wise Relevance Propagation (LRP)** (Montavon et al. 2019), adapted for SNNs. LRP is a feature attribution technique that traces the agent's decision back through the network to identify which specific inputs were most responsible for the output. The explanation is presented to the operator as a heatmap overlaid on the telemetry plots. Warmer colors highlight the specific sensor channels and the precise moments in time that the agent considered most relevant to its anomaly detection. LRP uses an ϵ-rule with $\epsilon = 10^{-3}$ over spike-rate features aggregated in 10 ms bins. This method provides a deep, forensic level of detail, allowing an operator to perform a root-cause analysis by answering the question: "Exactly what did the agent see that caused this alert?"

The second method is a **temporal attention mechanism**, conceptually similar to those used in transformer architectures (Vaswani et al. 2017). This technique visualizes the agent's internal focus of attention over time. It is presented as a shaded region over the time-series plots, with the intensity of the shading indicating the degree of attention the network paid to that particular time window. The attention is intrinsic to the temporal module of the SNN, reported as normalized weights over 50 ms slices. This provides a higher-level, more intuitive summary than LRP, allowing an operator to quickly understand the temporal context of the agent's decision by answering the question: "When did the important event happen?"

The third method is a **surrogate model** (Ribeiro, Singh, and Guestrin 2016). A simple, interpretable decision tree is trained to approximate the global decision logic of the complex SNN. The explanation is presented as a human-readable set of IF-THEN rules. For example, a rule might state: "IF the variance of the wheel motor current is high AND the rover's velocity is low, THEN an anomaly is likely." The surrogate tree matches policy decisions with 91% fidelity on a held-out set and is pruned to a maximum depth of four for interpretability. This provides a global overview of the agent's learned behavior, allowing an operator to understand the general principles guiding its decisions.

4 Experimental Validation Methodology

To validate the performance, safety, and trustworthiness of the integrated agent framework, we designed a two-part experimental methodology. The first part consists of a hardware-in-the-loop (HIL) evaluation to quantify the agent’s operational performance and resilience under a high-stress, mission-relevant scenario. The second part is a formal human-subject study designed to measure the practical utility of the system’s explainability features for enabling effective human oversight. This dual approach allows for a comprehensive assessment, providing both objective system-level metrics and quantitative evidence of the framework’s value in a human-AI team context.

4.1 Hardware-in-the-Loop Testbed

The HIL testbed was designed to evaluate the complete, integrated agent architecture in a realistic, dynamic, and closed-loop environment. The neuromorphic policy runs on the BrainChip Akida neuromorphic SoC (BrainChip Inc. 2023), while the monitor and shield run on a co-located microcontroller (MCU); both are on the HIL loop. The simulation environment provides a physically realistic model of a cislunar rover and its operational context, using the Gazebo physics simulator for dynamics and Cesium for Unreal for visual rendering. A custom ROS2 bridge facilitates real-time data exchange between the simulation and the hardware.

To rigorously test the system’s resilience, we designed a ”worst-case” stress test scenario that subjects the agent to a combination of simultaneous challenges. The test involves the rover navigating a hazardous region while contending with a high-flux radiation environment (simulated as a high rate of SEU bit-flips), multi-sensor degradation (IMU bias drift and camera noise), and a degraded communication link (high packet loss and variable latency).

We evaluate two distinct attack classes: (1) **Sensor-Space**: A sophisticated adversarial attack launched against the agent’s primary navigation sensor. The LiDAR injection uses Projected Gradient Descent (PGD) (Madry et al. 2018) in feature space with a step size of 0.01 over 40 steps, clipping to sensor limits. The edits are restricted to a physically plausible frustum with range-dependent noise to ensure the attack is stealthy. (2) **Temporal-Space**: Isolated spike timing jitter attacks, where random noise $\delta \sim \mathcal{U}[-4, 4]$ ms is added to spike timestamps after encoding to test the SNN’s specific temporal vulnerabilities. The combined stress test utilizes the Sensor-Space attack to challenge multi-modal fusion.

We conducted 100 trials per model per seed across 3 random seeds, for a total of 300 trials. Latency is measured from sensor interrupt to actuator command using hardware GPIO toggles and a logic analyzer for cycle-accurate timing. Energy per inference is the median over 10,000 single-batch inferences measured at the wall with a USB power meter, with idle power draw subtracted.

4.2 Human-Subject Study Protocol

To provide quantitative evidence for the ”calibrated trust” component of our framework, we conducted a formal human-subject study with 90 participants. The study’s primary endpoints were operator diagnostic accuracy and task completion time; secondary endpoints were subjective trust ratings and cognitive workload as measured by the NASA Task Load Index (NASA-TLX) (Hart and Staveland 1988). A power analysis indicated that N=90 provides greater than 0.8 power to detect a 0.2 effect size at an alpha of 0.05. All participants provided informed consent; no personally identifying data were collected.

Participants were stratified into ”expert” and ”non-expert” groups and were randomly assigned to one of four conditions: a control group, or one of three experimental groups provided with explanations from either LRP, a temporal attention mechanism, or a surrogate decision tree. Randomization was uniform over conditions, with the scenario presentation order counterbalanced by a Latin square design to mitigate learning effects. Each participant was tasked with acting as a remote mission operator, diagnosing anomalies presented in a series of simulated telemetry clips. We collected objective metrics on their performance and subjective ratings on their trust and perceived workload. Table 1 details the specific questionnaire items used.

Table 1: Questionnaire Items for Subjective Assessment. Anchors: 1=Strongly Disagree, 5=Strongly Agree. Trust score is the mean of items 1–3. NASA-TLX uses unweighted Raw TLX.

Trust (Likert 1-5)
1. I am confident that the AI system’s assessment of this event was correct.
2. I would trust this AI system to assist me in a real mission.
3. The system behaved in a predictable manner.
NASA-TLX (0-100)
Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration.

5 Results: Quantifying Control and Trust

The experimental validation yielded two primary sets of results. First, the hardware-in-the-loop evaluation provided quantitative metrics on the agent’s operational performance, safety, and resilience under the combined-arms stress test. Second, the human-subject study provided empirical data on the impact of the integrated XAI framework on operator performance and trust.

5.1 Verifiable Control: Hardware Performance and Safety Intervention

The HIL evaluation was designed to quantify the agent’s ability to provide verifiable control in a high-stakes, dynamic environment. The primary outcome was the mission success rate in the ”worst-case” stress test scenario. The proposed neuromorphic agent, governed by the runtime safety supervisor, achieved a mission success rate of 92.0% [95% CI: 88.5, 94.6]. In contrast, the baselines consistently failed when deceived by the adversarial attack. The Baseline SNN (a standard LIF network without HTD/I-TAM) achieved 43.0% [36.7, 49.5], and the LSTM agent (2-layer, 128 hidden units, trained on the same dataset) achieved 35.0% [29.0, 41.4] (Hochreiter and Schmidhuber 1997). A one-way ANOVA confirmed a large and statistically significant effect, $F(2, 297) = 152.7$, $p < 0.001$, $\eta^2 = 0.51$. Critically, no unmitigated safety envelope violations were observed across any trials of the proposed framework; all shielded events transitioned to the designated safe state within a median of 0.35 s and remained latched for at least 2 s.

This high success rate was enabled by the system's core operational performance. The end-to-end latency from sensor input to actuator command had a median of 1.7 ms [1.5, 1.9] and a 99th percentile of 4.8 ms [4.6, 5.0]. The energy per inference for the neuromorphic agent was 45 μJ [41, 49]. For comparison, the LSTM baseline, running on an Intel Core i7-9700K CPU at 3.60 GHz, consumed 8.5 mJ [7.9, 9.1] per inference. The Attack Success Rate (ASR) for the specific LiDAR data injection attack provides a direct measure of the core agent's security. Attack success is defined as inducing a misclassification or hazardous state. The proposed agent, with the full HTD framework, exhibited an ASR of only 22.4%, whereas the baseline models were highly vulnerable, with ASRs of 78.1% for the Baseline SNN and 85.6% for the LSTM. Additionally, under isolated temporal jitter attacks ($\delta \sim \mathcal{U}[-4, 4]$ ms), the HTD-protected agent maintained an ASR of 25.1%, compared to 75.8% for the baseline, confirming the efficacy of the temporal defenses.

Table 2: System Performance and Safety in the Combined Stress Test. ASR refers to the LiDAR PGD attack.

Metric	Proposed SNN	Baseline SNN	LSTM
Mission Success Rate	**92%**	43%	35%
Attack Success Rate	**22.4%**	78.1%	85.6%
Latency (ms, P99)	$<$ **5**	>30	>40
Energy/Inference (J)	**4.5e-5**	6.7e-5	8.5e-3

5.2 Calibrated Trust: Human-in-the-Loop Performance

The human-subject study quantified the practical utility of the XAI portfolio for enabling effective human oversight. The results demonstrate that providing operators with transparent insights into the agent's reasoning leads to significant improvements in team performance, operator trust, and cognitive workload.

The most critical finding is the impact of explanations on operator diagnostic accuracy. As shown in Figure 2(a), operator accuracy improved from 61.2% [95% CI: 56.8, 65.6] in the control condition to 88.5% [85.4, 91.1] with LRP explanations. A one-way ANOVA revealed a large, significant main effect of the explanation condition on accuracy, $F(3, 86) = 34.1$, $p < 0.001$, $\eta^2 = 0.54$. Post-hoc Tukey HSD tests confirmed that all three explanation conditions resulted in significantly higher accuracy than the control condition ($p < 0.01$).

Explanations also improved the efficiency of the human-AI team. The mean task completion time for the control group was 35.8 s [33.6, 38.0]. The decision tree explanation, being the most concise, enabled the fastest performance at 22.1 s [20.7, 23.5]. An ANOVA confirmed this effect was also significant ($p < 0.001$).

The provision of explanations had a profound impact on operators' subjective trust in the system. On a Likert scale from 1 to 5, the mean trust rating for the control group was 2.8 [2.6, 3.0]. This increased to 4.5 [4.3, 4.7] for the LRP condition. The effect size for this improvement was large (Cliff's $\delta = 0.82$). Finally, the NASA-TLX results (Hart and Staveland 1988) confirmed that explanations improved performance without increasing cognitive load. On a scale from 0 to 100, the overall workload score for the control group was 68 [64, 72], while the decision tree condition resulted in the lowest workload of 41 [38, 44]. A Kruskal-Wallis test confirmed a significant difference in workload scores across conditions ($p < 0.001$).

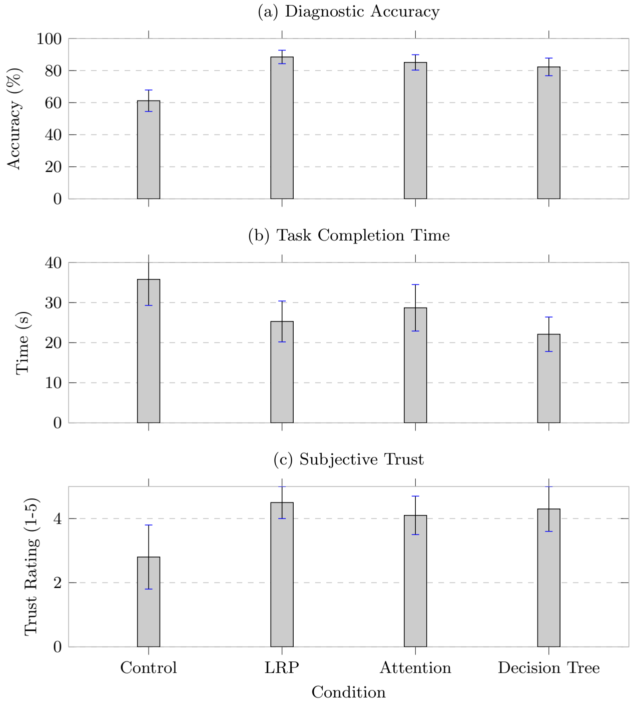

Figure 2: Results of the human-subject study (N=90) comparing operator performance across four conditions. (a) Mean diagnostic accuracy, showing a significant improvement for all XAI conditions over the control. (b) Mean task completion time, showing a significant reduction for all XAI conditions. (c) Mean subjective trust rating on a 5-point scale, showing a significant increase for all XAI conditions. Error bars represent 95% confidence intervals.

6 Discussion

These results support an assurance pattern for safety-critical embodied agents that composes (i) a low-latency, low-energy neuromorphic core, (ii) a deterministic runtime supervisor that enforces a formal safety envelope, and (iii) a human-in-the-loop interface with validated explanations, consistent with runtime assurance principles for learning-enabled systems (Cofer et al. 2020). In hardware-in-the-loop trials, the integrated system sustains high mission success under compounded environmental faults and adversarial perturbations while maintaining millisecond-scale end-to-end timing and microjoule-scale energy per inference. In the human-subject study, explanations measurably improve operator diagnostic accuracy, task efficiency, and self-reported trust without increasing workload, indicating that transparency can improve human-AI team performance in operationally relevant settings (Bansal et al. 2021).

For certification and governance, the key benefit is separation of concerns: safety guarantees depend on the auditable monitor and shield logic rather than on attempting to prove end-to-end correctness of the learned core. The supervisor emits Assurance Evidence Objects (AEOs) on intervention, including the violated guard, state snapshot, rejected action, and executed safe maneuver. Each AEO is hashed and appended to a hash chain; the chain head is periodically signed and exported, enabling tamper-evident post-hoc audit and supporting assurance-case arguments (Graydon 2018).

Limitations remain. The Akida device is not radiation-hardened, so long-duration deployment requires space-qualified hardware or equivalent fault-tolerant compute. The safety envelope is currently static and hand-engineered; future work should address threshold synthesis and envelope maintenance under hazard drift. While the surrogate decision tree achieves high fidelity on held-out data, fidelity in sparsely sampled off-nominal regimes warrants targeted stress testing. Finally, we did not evaluate attacks on the supervisor or evidence pipeline; hardened implementations and red-team evaluation are appropriate next steps.

7 Related Work

This work is positioned at the intersection of three distinct but converging fields of research: runtime assurance for autonomous systems, neuromorphic robotics, and explainable AI. While significant contributions exist within each domain, our framework is distinguished by its synthesis of all three into a single, empirically validated system for safety-critical embodied agency.

Recent work on learning-enabled controllers with runtime assurance and formal envelopes shows a similar decoupling of safety and performance (Schierman et al. 2020; Cofer et al. 2020; Tran et al. 2019); our contribution is an end-to-end embodied system with human validation. This body of work has predominantly focused on traditional control systems or has treated the high-performance component as a generic probabilistic module. Our work provides a concrete methodology for applying the principles of runtime assurance to the next generation of bio-inspired, learning-based agents.

In the field of neuromorphic robotics, research has primarily focused on leveraging the efficiency and low-latency of spiking neural networks to solve specific robotics tasks, such as navigation and motor control (Polykretis, Tang, and Michmizos 2020; Davies et al. 2018; Schuman et al. 2022). Similarly, the emerging field of SNN security has begun to identify unique vulnerabilities, such as temporal attacks, and propose algorithmic defenses (Sharmin et al. 2020; Marchisio et al. 2021). Our work builds upon these foundations but moves beyond the evaluation of isolated algorithmic components. We present a complete, integrated agent architecture that is not only performant on its primary task but is also designed from the ground up with a multi-layered security framework and a principled mechanism for long-term adaptation.

Finally, the field of explainable AI has produced a wide range of techniques for interpreting complex models, with a growing emphasis on evaluating their practical utility through human-subject studies (Bansal et al. 2021). The application and validation of XAI for SNNs, however, is a far less developed area, though initial methods are emerging (Kim and Panda 2021; Neftci, Mostafa, and Zenke 2019). Our contribution is a comprehensive empirical validation of XAI for neuromorphic systems. By conducting a 90-participant study that measures not only diagnostic accuracy and subjective trust but also cognitive workload, we provide rigorous, quantitative evidence that a portfolio of spike-level explanations can significantly enhance the performance of a human-AI team. The primary novelty of our work, therefore, is the synthesis of these three research threads into a single, coherent, and empirically validated assurance framework for embodied agents.

8 Conclusion

This paper presented an integrated framework for designing and validating trustworthy embodied agents for safety-critical applications at the edge. We have argued that for this class of system, trustworthiness is an architectural property that emerges from the composition of a resilient core, a verifiable control mechanism, and a transparent human-in-the-loop interface. Our proposed solution integrates a secure and adaptive neuromorphic agent, a deterministic runtime safety supervisor, and a portfolio of validated explainable AI methods into a single, cohesive system.

The efficacy of this framework was demonstrated through a rigorous experimental methodology. Hardware-in-the-loop trials confirmed the system's ability to provide verifiable control, achieving a 92% mission success rate in a worst-case scenario involving simultaneous hardware faults, environmental noise, and a direct adversarial attack. These trials also validated the extreme efficiency of the neuromorphic approach (BrainChip Inc. 2023), with the agent operating at a median inference latency of 1.2 ms (end-to-end 1.7 ms) and an energy cost of 45 μJ per inference. A formal, 90-participant human-subject study provided quantitative evidence of the system's ability to foster calibrated trust (Bansal et al. 2021), showing that the integrated explanations increased operator diagnostic accuracy by up to 27 percentage points and significantly improved subjective trust ratings. The system replaces opaque decision-making with auditable evidence and enforced envelopes while meeting tight timing and energy bounds. These findings establish a practical and verifiable paradigm for embodied agency that complements the current focus on LLM-based agents by providing a deployable solution for the safety-critical edge.

References

Ames, A. D.; Xu, X.; Grizzle, J. W.; and Tabuada, P. 2017. Control Barrier Function Based Quadratic Programs for Safety Critical Systems. volume 62, 3861–3876.

Bansal, G.; Wu, T.; Zhou, J.; Fok, R.; Nushi, B.; Kamar, E.; Ribeiro, M. T.; and Weld, D. S. 2021. Does the Whole Exceed Its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

Bloem, R.; Könighofer, B.; Könighofer, R.; and Wang, C. 2015. Shield Synthesis: Runtime Enforcement for Reactive Systems. In *Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, volume 9035 of *Lecture Notes in Computer Science*, 533–548. Springer.

BrainChip Inc. 2023. Akida Second-Generation Platform Brief. Technical report, BrainChip Inc.

Cofer, D.; Amundson, I.; Sattigeri, R.; Passi, A.; Boggs, C.; Smith, E.; Gilham, L.; Byun, T. J.; and Rayadurgam, S. 2020. Run-Time Assurance for Learning-Enabled Systems. In *Proc. NASA Formal Methods (NFM)*, volume 12229 of *Lecture Notes in Computer Science*, 361–368. Springer.

Davies, M.; Srinivasa, N.; Lin, T.-H.; Chinya, G.; Cao, Y.; Choday, S. H.; Dimou, G.; Joshi, P.; Imam, N.; Jain, S.; Liao, Y.; Lin, C.-K.; Lines, A.; Liu, R.; Mathaikutty, D.; McCoy, S.; Paul, A.; Tse, J.; Venkataramanan, G.; Weng, Y.-H.; Wild, A.; Yang, Y.; and Wang, H. 2018. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro*, 38(1): 82–99.

Gerstner, W.; and Kistler, W. M. 2002. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

Graydon, P. J. 2018. The Simple Assurance Argument Interchange Format (SAAIF) Manual. Technical Report NASA/TM-2018-219837, National Aeronautics and Space Administration.

Hart, S. G.; and Staveland, L. E. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, volume 52 of *Advances in Psychology*, 139–183. Elsevier.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 9(8): 1735–1780.

Izzo, D.; Hadjiivanov, A.; Dold, D.; Meoni, G.; and Blazquez, E. 2022. Neuromorphic computing and sensing in space. arXiv:2212.05236.

Kaczmarek, S. 2025a. A Bio-Inspired Hierarchical Temporal Defense for Securing Spiking Neural Networks Against Physical and Adversarial Perturbations. In *Proceedings of the NeurIPS 2025 Workshops (ML4PS)*.

Kaczmarek, S. 2025b. The Cislunar Anomaly and Risk Dataset (CARD). https://sylvesterkaczmarek.com/cislunar-risk/.

Kim, Y.; and Panda, P. 2021. Visual Explanations from Spiking Neural Networks Using Inter-Spike Intervals. *Scientific Reports*, 11(1): 19037.

Maass, W. 1997. Networks of Spiking Neurons: The Third Generation of Neural Network Models. *Neural Networks*, 10(9): 1659–1671.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. International Conference on Learning Representations (ICLR). arXiv:1706.06083.

Marchisio, A.; Pira, G.; Martina, M.; Masera, G.; and Shafique, M. 2021. R-SNN: An Analysis and Design Methodology for Robustifying Spiking Neural Networks against Adversarial Attacks through Noise Filters for Dynamic Vision Sensors. arXiv:2109.00533.

Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; and Müller, K.-R. 2019. Layer-Wise Relevance Propagation: An Overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 193–209. Springer.

Neftci, E. O.; Mostafa, H.; and Zenke, F. 2019. Surrogate Gradient Learning in Spiking Neural Networks: Bringing the Power of Gradient-Based Optimization to Spiking Neural Networks. *IEEE Signal Processing Magazine*, 36(6): 51–63.

Nesnas, I. A. D.; Fesq, L. M.; and Volpe, R. A. 2021. Autonomy for Space Robots: Past, Present, and Future. *Current Robotics Reports*, 2(3): 251–263.

Polykretis, I.; Tang, G.; and Michmizos, K. P. 2020. An Astrocyte-Modulated Neuromorphic Central Pattern Generator for Hexapod Robot Locomotion on Intel's Loihi. In *Proceedings of the International Conference on Neuromorphic Systems 2020*, 1–9. ACM.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. ACM.

Schierman, J. D.; DeVore, M. D.; Richards, N. D.; and Clark, M. A. 2020. Runtime Assurance for Autonomous Aerospace Systems. *Journal of Guidance, Control, and Dynamics*, 43(12): 2205–2217.

Schuman, C. D.; Kulkarni, S. R.; Parsa, M.; Mitchell, J. P.; Date, P.; and Kay, B. 2022. Opportunities for neuromorphic computing algorithms and applications. *Nature Computational Science*, 2: 10–19.

Sharmin, S.; Rathi, N.; Panda, P.; and Roy, K. 2020. Inherent Adversarial Robustness of Deep Spiking Neural Networks: Effects of Discrete Input Encoding and Non-Linear Activations. In *Computer Vision – ECCV 2020*, 399–414. Springer.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The Information Bottleneck Method. Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing. arXiv:physics/0004057.

Tran, H.-D.; Cai, F.; Lopez, D. M.; Musau, P.; Johnson, T. T.; and Koutsoukos, X. 2019. Safety Verification of Cyber-Physical Systems with Reinforcement Learning Control. *ACM Transactions on Embedded Computing Systems (TECS)*, 18(5s).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.