Decision-Audit Substrate for Safe Embodied AI: The PRISM Logging Standard

Anonymous Author(s)
Submission to IJCAI 2026 Safe Physical AI Workshop

***Abstract*—Safety evaluation of embodied AI systems—autonomous vehicles, surgical robots, industrial manipulation platforms, and lethal autonomous weapons systems—demands two complementary dimensions: safe action selection and post-hoc decision auditability. The latter, essential for certification, accountability, and multi-stakeholder trust, currently lacks a standardized format for production-scale recording of per-decision reasoning in embodied AI. We introduce the PRISM Logging Standard (Lee, 2026d) as a deployable substrate addressing this audit gap. PRISM emits a ∼60-character structured code per decision that captures the decision's reasoning structure across three dimensions: value hierarchy (Schwartz universal values), evidence hierarchy (Walton argumentation schemes), and source hierarchy (Hovland–Kelley source credibility). Empirical validation on 366,120 forced-choice responses from 8 frontier LLMs (Lee, 2026c) reveals that 6 of 8 models elevate the Security value to win-rates 0.951–0.998 in defense-domain decisions, and the gap between Test–Retest Reliability (0.875–0.985) and Paired Consistency Score (0.480–0.659) establishes *framing sensitivity*—not stochastic noise—as a measurable behavioral dimension. PRISM provides (i) audit trails for autonomous-system certification, (ii) complementary signals for runtime safety shields, (iii) shared decision visibility across developers, regulators, and citizens, and (iv) drift-detection substrate for post-update behavioral monitoring. Released under MIT license.**

I. The Audit Gap in Embodied AI Safety

Robotic and autonomous systems increasingly operate in real-world environments. Safety guarantees, however, require both **active safety** (safe action selection) and **audit safety** (post-hoc decision recording). The latter, despite its centrality to certification, incident investigation, and citizen trust, lacks a model-/vendor-comparable standardized format. PRISM Logging Standard addresses this gap.

II. The PRISM Code Format

PRISM emits, per decision, a single line:

```
C:DF/SXs | V:Unc<Ses | E:Pop<Rev | S:Tes<Gov
```

reading: defense domain / societal impact / irreversible / short-term; Universalism-Concern outranked by Security-Societal; Popular Consensus outranked by Systematic Review; Personal Testimony outranked by Government Official.

The format integrates four closed vocabularies: 7-domain × 5-scope × 3-reversibility × 3-time-horizon context (**C:**); Schwartz 10- or 19-value hierarchy (**V:**); Walton 10-type evidence hierarchy (**E:**); Hovland–Kelley 10-class source hierarchy (**S:**). The standard supports three output modes (inline tag, structured JSON, tool call) and is emitted via system prompt addition—no model retraining required.

III. Empirical Behavior Relevant to Embodied AI

The 366,120-response measurement (Lee, 2026c) reveals two findings of direct relevance to physical-AI safety:

Defense-domain Security inflation: 6 of 8 models elevate the Security value to win-rate 0.951–0.998. Crucially, 2 of 4 globally Universalism-first models flip to Security-first under defense framing. For LLM-controlled cyber-physical systems—autonomous weapons, unmanned aerial vehicles, defense-context robotics—this domain-conditioned reorganization of value priorities is invisible to single-context safety evaluation.

Framing sensitivity: Test–Retest Reliability (TRR) is 0.875–0.985 (87.5–98.5%) versus Paired Consistency Score (PCS) 0.480–0.659 (48.0–65.9%), with a 25.3–49.4 percentage-point gap. Stochastic Noise diagnostic is 0% across all models. For embodied AI deployed in safety-critical contexts, this means user-input phrasing differences may systematically shift control-layer decisions—directly relevant to prompt-injection risk and adversarial robustness in autonomous systems.

IV. Integration into Safety Workflows

PRISM logs operate as a complementary layer alongside existing physical-AI safety infrastructure:

- **Runtime safety shield**: PRISM-code patterns matching predefined risk profiles (e.g., medical-domain decision with `V:Care<Sec`) trigger shield activation.
- **Certification audit trail**: Per-decision reasoning records reconstructible during incident investigation and certification review.
- **Drift detection**: V-pair distribution shifts across model versions provide quantitative signals of alignment changes.
- **Anomaly detection**: Unusual frequency of `S:Ano` (anonymous source) pairs signals possible prompt injection or context contamination.
- **Multi-stakeholder visibility**: Developers, regulators, and citizens can review decisions through the same 60-character vocabulary.

PRISM aligns naturally with EU AI Act Articles 12–15 (logging, transparency, human oversight, robustness), though it does not single-handedly satisfy any individual provision.

V. Limitations

PRISM codes are post-hoc self-reports, not causal traces (e.g., chain-of-thought or activation tracing); structured post-hoc reporting remains more auditable than unstructured silence. Production emit reliability under embodied-AI latency

constraints requires further deployment validation; current measurements show $\sim$5–10% token overhead from system-prompt addition. Multilingual vocabulary expansion (currently English and Korean) is future work for global cyber-physical AI deployment.

VI. Conclusion

Embodied AI safety requires both safe action selection and *decision auditability*. PRISM Logging Standard provides the latter through a standardized, compact, model- and vendor-comparable format. Defense-domain Security inflation across 6 of 8 frontier models demonstrates that domain-conditioned behavioral signatures are measurable; PRISM is the production-scale deployable substantiation of that measurement.

Acknowledgments

Author-written Korean drafts were translated and polished into English with LLM assistance. All scientific claims, frameworks, and data analyses originate from the author and prior work (Lee, 2026a/b/c/d).

References

[1] Lee, S. (2026a). *AI Integrity: A New Paradigm for Verifiable AI Governance.* arXiv:2604.11065.

[2] Lee, S. (2026b). *PRISM Risk Signal Framework: Hierarchy-Based Red Lines for AI Behavioral Risk.* arXiv:2604.11070.

[3] Lee, S. (2026c). *Measuring the Authority Stack of AI Systems: Empirical Analysis of 366,120 Forced-Choice Responses Across 8 AI Models.* arXiv:2604.11216.

[4] Lee, S. (2026d). *PRISM Logging Standard: A Common Language for AI Integrity.* ICML AI4GOOD 2026 (under review).

[5] Schwartz, S. H. (1992). Universals in the content and structure of values. *Advances in Experimental Social Psychology*, 25, 1–65.

[6] Walton, D. (2008). *Argumentation Schemes for Presumptive Reasoning.* Cambridge Univ. Press.

[7] Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and Persuasion.* Yale Univ. Press.

[8] European Parliament & Council. (2024). Regulation (EU) 2024/1689 (AI Act).