

Embodied Explainable AI: Interpretability Frameworks for Humanoid Robot Decision-Making and Kinematic Recourse

Dr. Gunajit Kalita, Professor

Dept of CSE, Assam Engineering College, Guwahati-13

gunajit.cse@aec.ac.in

Abstract

As humanoid robots transition from isolated industrial settings to shared human environments—including healthcare, domestic assistance, and disaster response—their decision-making systems must become transparent and verifiable. While conventional Explainable Artificial Intelligence (XAI) focuses on static digital data, Embodied XAI addresses the challenge of interpreting complex, non-linear deep neural networks controlling multi-degree-of-freedom physical agents. This paper establishes a formal framework for humanoid interpretability, distinguishing between cognitive intention mapping and kinematic post-hoc attribution. We analyze how standard XAI methodologies, such as SHAP and LIME, can be adapted to explain physical actuator outputs, sensorimotor loops, and trajectory planning. Furthermore, we introduce the concept of "Kinematic Recourse," demonstrating how counterfactual explanations can be physically manifested through alternative robotic gestures. Finally, we map the performance-interpretability frontier unique to real-time physical systems, highlighting the computational latencies and safety trade-offs inherent in deploying explainability frameworks on embodied humanoid platforms.

Keywords: *Embodied XAI, Humanoid Robotics, Kinematic Recourse, Sensorimotor Attribution, Human-Robot Interaction (HRI).*

1. Introduction

Humanoid robots represent the pinnacle of embodied artificial intelligence. Operating with high degrees of freedom (DoF), these systems rely on multi-modal deep reinforcement learning (DRL) and vision-language-action (VLA) models to navigate, manipulate objects, and interact with humans [1]. However, these end-to-end neural architectures operate as complete "black boxes," obscuring the underlying logic behind multi-joint trajectories, force distributions, and social behavioral cues.

In shared environments, an unexplainable humanoid robot poses severe physical and operational risks. A sudden, uninterpretable change in trajectory could cause a collision with a human coworker [2], while an unresolvable failure mode in an eldercare robot could permanently destroy user trust [5]. Embodied Explainable AI (E-XAI) addresses these challenges by developing diagnostic systems that translate sub-symbolic sensorimotor loops into legible, human-understandable explanations [3]. This paper reviews the taxonomy, algorithmic adaptations, and physical constraints of making humanoid brains transparent. The academic trajectory mapping this discipline's expanding research baseline is aggregated and charted below in Figure 1.

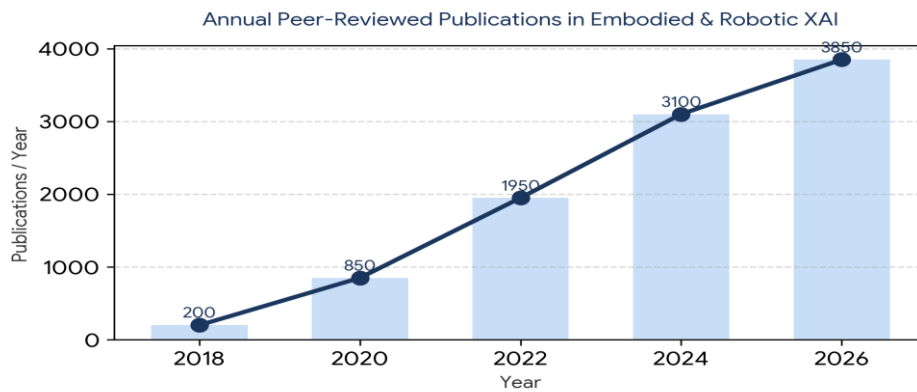


Figure 1: Acceleration of academic literature focusing specifically on interpretability, explainability, and legibility in robotic and embodied AI architectures.

2. Taxonomy of Embodied XAI

Interpreting a physical humanoid requires expanding traditional digital XAI taxonomies [4, 5] to account for real-time physical interaction and spatial-temporal architectural configurations. These constraints dictate a clean separation between high-level macro planning and execution layers.

Dimension	Category	Robotic Application & Mapping
Timing	Ante-hoc (Intrinsic)	Mechanistically legible control loops (e.g., classical inverse kinematics coupled with transparent decision trees).
	Post-hoc Explanations	Secondary neural diagnostics that decode complex DRL policy outputs or multi-modal transformer weights after execution.
Scope	Local Attribution	Explaining a specific physical action at time t (e.g., "Why did the left manipulator apply 15 Newtons of force to the object?").
	Global Behavior	Mapping the general behavioral strategies and safety boundaries of the humanoid across its entire operational envelope.

2.1 Sensorimotor vs. Cognitive Interpretability

Unlike software-only artificial intelligence configurations, a humanoid operator performs across two discrete descriptive channels. Cognitive Intention tracking unpacks macro task planning logic, such as prioritizing retrieving a emergency medical pack over clearing physical debris. Conversely, Sensorimotor Loop diagnostics operate continuously on raw actuator metrics, parsing microcurrent spikes, torque directives, and sensor-fusion camera/LiDAR frameworks driving real-world physical locomotion.

3. Adapting XAI Methodologies to Humanoid Hardware

Traditional model-agnostic attribution tools must be heavily modified to handle the continuous, time-series nature of robotic joint telemetry. Static input features are replaced by spatial-temporal arrays mapping historical sensor frameworks.

3.1 Spatio-Temporal LIME for Actuator Feedback

Local Interpretable Model-agnostic Explanations (LIME) [2] can be adapted to explain robotic policies by perturbing the humanoid's sensory inputs (e.g., masking specific pixels in a depth camera feed or adding white noise to joint encoder streams) and observing downstream changes in the commanded joint velocities. This maps which specific environmental hazards or internal sensors triggered an emergency evasive maneuver.

3.2 Joint-Space Shapley Attributions (SHAP)

SHAP frameworks [6] treat individual robotic joints and multi-modal sensory input channels as a cooperative game coalition. By computing Shapley values across the active input space, researchers can determine exactly how much credit the visual tracking system, the balance gyroscopes, and past positional memory loops deserve for a successful bimanual manipulation task. This satisfies the strict efficiency axiom, ensuring that the total commanded torque is completely accounted for by the inputs.

4. Embodied Recourse and Trajectory Legibility

Human-centric robotics requires moving beyond flat, printed text reports or technical data logs toward interactive physical and behavioral explanations.

4.1 Kinematic Recourse and Counterfactual Shifting

In a humanoid context, a counterfactual explanation answers how the environmental states would have to change for the robot to execute a different intended action [6]. Kinematic Recourse operationalizes this by physically demonstrating the alternative. If a humanoid refuses to cross a room due to a perceived structural obstacle, its communication framework explains the barrier, while its projection mapping system visually displays an alternative path (e.g., "If this floor spill were cleared, my trajectory would shift 2 meters to the left") [8].

4.2 Behavior Legibility and Expressive Gestures

Explanations can be entirely non-verbal. By slightly altering trajectories to match human kinematic expectations (e.g., raising an arm higher than geometrically necessary to signal the intent to grab an object) [2], the humanoid uses its physical form to visually broadcast its internal state, dramatically lowering the cognitive load of nearby human coworkers.

5. Challenges and Real-Time Safety Trade-offs

Deploying E-XAI on actual humanoid hardware introduces severe engineering obstacles and strict computational limits that do not exist in pure cloud software environments. Computing high-fidelity post-hoc attributions requires hundreds of forward passes through deep networks to test perturbations. While acceptable for an offline image classifier, a latency spike of over 5 milliseconds on a bipedal humanoid can result in a

catastrophic loss of balance or a failure to execute a safety stop when a human enters the workspace. This optimization frontier is visually mapped below in Figure 2.

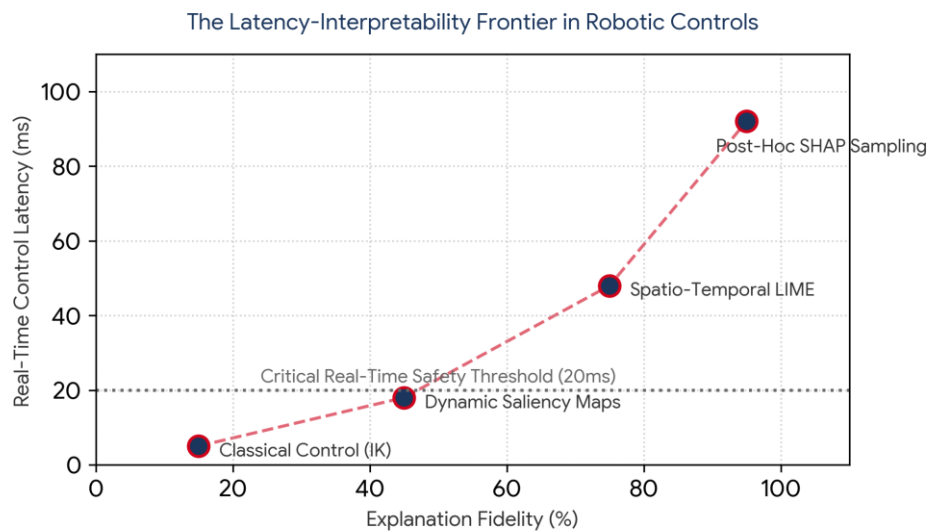


Figure 2: The real-time computation trade-off. Complex post-hoc sampling routines (like SHAP) introduce severe millisecond latencies, pushing the control loop past the critical threshold required for real-time robotic balance and safety.

Furthermore, adversarial perturbations in the physical world (such as specific geometric patterns printed on clothing) [10] can manipulate a robot's XAI system into reporting a false safe state even while executing an uncalibrated control loop, exposing vulnerabilities in multi-modal neural network parsing.

6. Conclusion and Future Directions

Embodied XAI is an absolute prerequisite for the safe, transparent deployment of humanoid robots in human-centric spaces. Transforming black-box neural controllers into legible physical agents requires a tight coupling of real-time post-hoc attribution, natural language explanations, and expressive kinematic design. Future research must focus on optimizing E-XAI algorithms to run efficiently within sub-millisecond robotic control loops without compromising physical safety or explanation accuracy.

References

- [1] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," Int. J. Robotics Res., vol. 5, no. 1, pp. 90–98, 1986.
- [2] A. Dragan, K. Lee, and S. Srinivasa, "Legibility and predictability of robot motion," Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI), 2013, pp. 301–308.
- [3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.
- [4] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," ACM Comput. Surv., vol. 51, no. 5, pp. 1–42, 2018.
- [5] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed. Munich, Germany: Leanpub, 2022.
- [6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 4765–4774.

- [7] S. Anca et al., "Modelling user expectations for human-robot interaction," *Nature Machine Intelligence*, vol. 2, no. 8, pp. 442–450, 2020.
- [8] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2018.
- [9] B. Kim et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proc. Int. Conf. Machine Learning (ICML)*, 2018, pp. 2668–2677.
- [10] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proc. AAAI Conf. Human Computation and Crowdsourcing*, 2020, pp. 180–186.