Course-of-Action Grounding for Knowledge-Guided Safe and Explainable Human-AI Interaction

David J. Lerch[†], Gerrit Holzbach[*], Matthis Leicht[*], Frederik Diederichs[*]
[†] *Karlsruhe Institute of Technology* Karlsruhe, Germany
david.lerch2@kit.edu
[*] *Fraunhofer IOSB* Karlsruhe, Germany
{firstname.lastname}@iosb.fraunhofer.de

***Abstract*—Physical AI systems require safe and explainable human-AI interaction. However, current failure detection methods lack structured knowledge about intended behavior. We propose a framework that combines knowledge-guided failure prediction with course-of-action knowledge graphs (CoA-KG), in which a course of action is a structured sequence of intended actions, phases, and transitions that define normative interaction behavior. The framework detects deviations from expected CoA-KG states, provides structural explanations through KG paths. CoA-KG is used to train explainable models and provide safe fallback behaviors encoded by the CoA-KG, applicable across the industrial, automotive, healthcare and service robotics domains.**

***Index Terms*—Knowledge graphs, Human-robot interaction.**

I. Motivation and Problem Statement

Physical AI systems, including collaborative robots, autonomous vehicles, and service robots, require safe human-AI interaction (HAI) in dynamic, safety-critical environments where failures can cause serious harm to humans [1], [2]. These failures span three levels aligned with robotic cognition stacks [3]: perception (missed safety-critical objects), intention inference (mispredicted human intention), and action selection (unsafe movements). Current runtime monitors are capable of detecting anomalies; however, they lack the capacity to provide a detailed explanation of the underlying causes and the expected outcomes. Out-of-Distribution (OOD) detectors are characterized by an absence of semantic grounding [4], [5]. Knowledge-Guided Failure Prediction (KGFP) enhances perception safety through the detection of semantic misalignment [6], it relies on *implicit* visual knowledge and addresses only perception. To enhance explainability, we use structured knowledge of intended interaction, which is defined as explicit representation of courses of action in a knowledge graph (KG) that defines what should happen. Hierarchical task planning frameworks for human-robot collaboration encode such structured sequences [3]. The knowledge-guided course-of-action grounding is imperative for achieving explainable, safe HAI. The process of inferring expected states through KG traversal involves reasoning over normative interaction models [7].

II. Proposed Method

We propose a framework centered on a *Course-of-Action Knowledge Graph* (CoA-KG) that grounds failure prediction on explicit, structured knowledge of intended interaction. A course of action is defined as a structured sequence of actions, phases, and transitions representing normative interaction behavior [3], [8]. The CoA-KG encodes the normative model of Human-AI interaction: atomic actions (reach, grasp, handover), task phases (approach, collaborate, retreat), predicted human intentions (reach for object, avoid, cooperate), and safety constraints derived from norms like ISO 10218-1:2025 [9] and ISO/TS 15066:2016 [10]. This design builds on knowledge processing infrastructures for cognition-enabled robots [8], behavioral adaptation [11], and ontology-based safety modeling [12], while leveraging knowledge-graph-based intent reasoning for transparent inference of human intentions [13]. CoA-KG thus serves as a runtime specification for correct interactions at any given phase [7]. We extend Knowledge-Guided Failure Prediction (KGFP) [6] beyond perception-only detection to three CoA-KG-grounded levels: perception, intention, and action. At the *perception level*, KGFP's semantic misalignment detection is contextualized using KG-predicted entities expected in the current phase. At the *intention level*, observed human behavior is compared with KG-expected intentions. At the *action level*, planned or executed actions are validated against the KG-expected action sequence. Failure is defined as a *discrepancy between observed state and KG-expected state*. Because failures are defined as KG-deviations, explanations are inherently structural rather than post-hoc [14], [15]. Each prediction traces to a violated KG constraint (e.g., "Expected: handover | Observed: retreat $\rightarrow$ transition constraint violated"). Fallback behaviors are likewise KG-encoded. Each constraint is associated with graduated deterministic responses (reduce speed, pause, increase distance, retreat), following runtime assurance principles [16].

III. Expected Contributions and Impact

With CoA-KG for failure prediction we aim at grounded, explainable and safe AI for human-AI interaction. We target safety and reliability to detect when intentions/actions deviate from expected courses, with built-in explainability supporting *trust* [17] and *behavioral guarantees* [18], [19] in safety-critical domains. We extend the knowledge guidance from foundation model embeddings to CoA-KGs to make failure prediction both traceable and explainable. Current limitations include the need for domain-specific CoA-KG engineering and missing evaluation on real-world HAI scenarios.

REFERENCES

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.

[2] S. Haddadin, A. Albu-Schäffer, and G. Hirzinger, "Requirements for safe robots: Measurements, analysis and new insights," *Int. J. Robot. Res.*, vol. 28, no. 11–12, pp. 1507–1527, Nov. 2009.

[3] L. Johannsmeier and S. Haddadin, "A hierarchical human-robot interaction-planning framework for task allocation in collaborative industrial assembly," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017, pp. 372–378.

[4] J. Yang, K. Wang, P. Zou, Z. Zhou, K. Ding, W. Peng, H. Wang, G. Chen, B. Li, Y. Sun, X. Du, K. Zhou, W. Zhang, D. Hendrycks, Y. Li, and Z. Liu, "OpenOOD: Benchmarking generalized out-of-distribution detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[5] J. Guérin, K. Delmas, R. Ferreira, and J. Guiochet, "Out-of-distribution detection is not all you need," *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 12, pp. 14 829–14 837, 2023.

[6] J. P. Zimmermann, G. Holzbach, and D. Lerch, "Knowledge-guided failure prediction: Detecting when object detectors miss safety-critical objects," *arXiv preprint arXiv:2603.25499*, Mar. 2026.

[7] M. Leucker and C. Schallhart, "A brief account of runtime verification," *J. Log. Algebr. Program.*, vol. 78, no. 5, pp. 293–303, 2009.

[8] M. Tenorth and M. Beetz, "KnowRob: A knowledge processing infrastructure for cognition-enabled robots," *Int. J. Robot. Res.*, vol. 32, no. 5, pp. 566–590, May 2013.

[9] ISO, "ISO 10218-1:2025 Robotics—Safety requirements—Part 1: Industrial robots," International Organization for Standardization, Geneva, Switzerland, Feb. 2025.

[10] ISO, "ISO/TS 15066:2016 Robots and robotic devices—Collaborative robots," International Organization for Standardization, Geneva, Switzerland, 2016.

[11] S. Nikolaidis, D. Hsu, and S. Srinivasa, "Human-robot mutual adaptation in collaborative tasks: Models and experiments," *Int. J. Robot. Res.*, vol. 36, no. 5–7, pp. 618–634, May 2017.

[12] E. Prestes, J. L. Borges, and M. I. Pereira, "Towards a core ontology for robotics and automation," *Robot. Auton. Syst.*, vol. 61, no. 11, pp. 1353–1364, Nov. 2013.

[13] Y. Chen, L. Zhang, and W. Wang, "Enhancing intention prediction and interpretability in service robots with LLM and KG," *Scientific Reports*, vol. 14, art. no. 27 916, Nov. 2024.

[14] A. B. Arrieta, N. Diaz-Rodrizalez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[15] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *Proc. Int. Conf. Auton. Agents Multiagent Syst. (AAMAS)*, 2019, pp. 1078–1088.

[16] A. Desai, S. Guo, A. Seshia, D. Sadigh, and S. S. Sastry, "SOTER: A runtime assurance framework for programming safe robotics systems," in *49th Annual IEEE/IFIP Int. Conf. Dependable Syst. Networks (DSN)*, 2019.

[17] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, Spring 2004.

[18] S. A. Seshia, D. Sadigh, and S. S. Sastry, "Toward verified artificial intelligence," *Commun. ACM*, vol. 65, no. 7, pp. 46–55, Jun. 2022.

[19] M. Luckcuck, M. Farrell, L. A. Dennis, C. Dixon, and M. Fisher, "Formal specification and verification of autonomous robotic systems: A survey," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–41, Oct. 2019.