Blending Ontologies and Language Models to Generate Sound and Natural Robot Explanations

Extended Abstract

Alberto Olivares-Alarcos
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Barcelona, Spain
aolivares@iri.upc.edu

Muhammad Ahsan
Institute of Electrical and Control Engineering, National Yang Ming Chiao Tung University
Hsinchu, Taiwan
muhammadahsanfs.ee12@nycu.edu.tw

Satrio Sanjaya
Institute of Electrical and Control Engineering, National Yang Ming Chiao Tung University
Hsinchu, Taiwan
satriosanjaya.ee11@nycu.edu.tw

Hsien-I Lin
Institute of Electrical and Control Engineering, National Yang Ming Chiao Tung University
Hsinchu, Taiwan
sofin@nycu.edu.tw

Guillem Alenyà
Institut de Robòtica i Informàtica Industrial, CSIC-UPC
Barcelona, Spain
aolivares@iri.upc.edu

ABSTRACT

This work introduces an approach to creating semantically grounded and natural robot explanations combining ontology-based reasoning with large language models (LLMs). Ontologies ensure logical consistency and domain grounding, while LLMs provide fluent and interactive explanation generation. Empirical results highlight the potential of ontology–LLM integration to advance explainable robotics, and promote more intuitive human-robot collaboration.

KEYWORDS

applied ontology; explainable collaborative robots; language models

ACM Reference Format:
Alberto Olivares-Alarcos, Muhammad Ahsan, Satrio Sanjaya, Hsien-I Lin, and Guillem Alenyà. 2026. Blending Ontologies and Language Models to Generate Sound and Natural Robot Explanations: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/WAZM4282

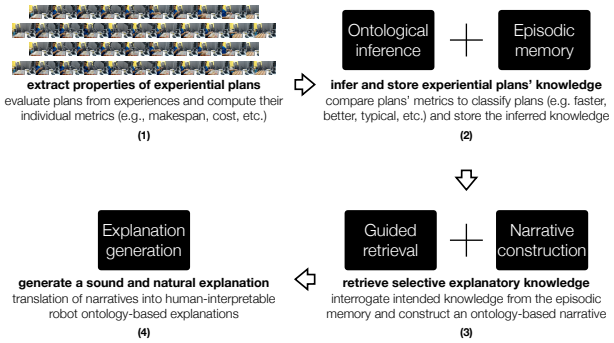

Figure 1: Overview of the methodology for generating sound and natural explanations blending ontologies and LLMs. (1) and (2) operate on n plans, while (3) and (4) on 2 plans (e.g., current and previous executions, current and prototypical, ...).

1 INTRODUCTION

The emergence of large language models (LLMs) offers new opportunities to enable interactive human-robot information exchange [16]. This may contribute to the field of explainable agency (explaining the reasoning behind the behavior of goal-driven agents and robots) [1]. However, foundation models often lack awareness of domain-specific knowledge and internal robot states, and their tendency to generate fluent yet ungrounded text can lead to hallucinations and false information. Retrieval-augmented generation (RAG) is a promising solution to this drawback [2, 3], using domain-specific knowledge sources to improve the reliability of explanations by aligning them with the contextual knowledge of the robot.

Ontologies are great sources of domain-specific knowledge, being a natural fit for the retrieval process in RAG-based explainable systems. Indeed, some authors argue that 'explanation presupposes semantics, and semantics presupposes ontology' [4], which is empirically supported by a solid literature on ontology-based explainable artificial intelligence [7, 12, 15]. Recent works investigated the generation of ontology-based robot explanatory narratives, claiming positive insights and highlighting some limitations (e.g. ontology-based narratives tend to be verbose and complex to read) [9, 10].

Building on this, we investigate the following research question: *How can robots model and reason about their experiences and generate explanations that compare them in a sound and natural manner?* This extended abstract summarizes the design and evaluation of a framework (see Fig. 1) proposed in our more comprehensive work [8].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/WAZM4282

2 METHODOLOGY DESIGN RATIONALE

Background on explainable agency: The literature on explainable agency offers a well-established functional roadmap for designing explainable agents around four core functionalities [5]: (1) evaluating different candidate solutions to a task, (2) systematic indexing and storage of evaluation insights, (3) retrieval of stored knowledge, and (4) transmission of retrieved knowledge. First, during problem solving, the agent must evaluate multiple candidate solutions, analyzing their properties, so that it is possible to discriminate between them. Second, the agent must systematically index and store detailed records of its evaluation and classification in a structured repository, such as an episodic memory. Third, the agent must transform queries into retrieval cues to access relevant knowledge from its memory. Fourth, once the knowledge is retrieved, the agent translates it into human-interpretable language.

Insights from the social sciences: Miller [6] argues that the field of explainable artificial intelligence (XAI) can build on existing research in philosophy, psychology, and cognitive science investigating how people define, generate, select, evaluate, and present explanations. Miller reviewed key works, highlighting major findings and discussing how they can be integrated into XAI research and practice. The most important findings state that explanations are *contrastive*, *selected*, and *social*. They are contrastive because they respond to counterfactual questions such as why one plan is better than another. Explanations are selective, typically presenting only part of the reasons, drawn from broader knowledge according to specific criteria. Finally, explanations are social, transferring knowledge in conversational form as part of interactions between agents. These qualities are inherent in the nature of the explanations themselves, thus, they are not just desirable but essential for the design and development of explainable robots.

Guidance principles for the methodology design: Inspired by these ideas, the proposed methodology produces *contrastive* and *social* (interactive) explanations that are *selected* according to different levels of *specificity* or detail. Furthermore, our proposal comprises four functionalities or steps: (1) robot experiences analysis and computation of their individual plan's properties, (2) ontology-based episodic storage and inference for plan comparison and classification, (3) selective and systematic retrieval of *sound* ontology-based narratives using a state-of-the-art approach [11], and (4) LLM-based generation of *natural* explanations. Note that there is a relationship between the three properties and the four functionalities (steps). Contrastive explanations require the analysis of the qualities of multiple alternative options (step 1). The (biased) selection of explanation content depends on the knowledge storage and retrieval (steps 2 and 3), and can be learned from users' preferences. Finally, social explanations demand transmission in a human-interpretable form (step 4).

3 EVALUATION RESULTS AND DISCUSSION

Experiential robot episodes were recorded (steps 1 and 2) using a lab mock-up of a real industrial task, a collaborative human-robot visual inspection of the surface of metallic SSD cases [14]. The state-of-the-art ontology-based narratives (step 3) have been identified as verbose and complex to understand. Hence, a statistical analysis was conducted to evaluate whether the LLM-based explanations (step 4) were better in terms of *brevity* and *clarity*, while preserving the *semantic coherence*. The performance of the proposed approach is consistently superior to that of the baseline ontology-based narratives. Considering **brevity**, our explanations are a 33%, 86% and 93% shorter on average than the baseline narratives for the specificity (detail) level 1, 2, and 3, respectively. In the case of the **clarity**, the proposed method produces explanations that are a 19% and 76% easier to read than the baseline narratives at levels 2 and 3. Importantly, this considerable gain in brevity and clarity does not come at the cost of **semantic coherence**. Indeed, the semantic cosine similarity was consistently above 0.7, indicating a strong semantic correlation between the generated explanations and the baseline narratives.

Practical view on the brevity improvement: The explanations discussed in this work would often be transmitted from robots to humans, who would need to understand and retain the explanatory content. To retain information, people are comfortable with a speaking pace of 150-160 words per minute (wpm), while the pace for silent reading is 250-400 wpm [13]. Let's consider the lowest bound of those intervals, and the average explanation lengths obtained in the evaluation. For the baseline narratives, humans would spend almost **5 minutes** listening to an average narrative of specificity 2, and **12 minutes** for one of specificity 3. Our method reduces those times to **42** and **55 seconds**, respectively. In the case of silent human reading, the baseline narratives of specificity 2 and 3 would require **3** and **7 minutes**, while our explanations only **25** and **32 seconds**. This is a huge advancement towards the development of technically rigorous and socially acceptable explainable robots.

Discussion on an extension to interactive scenarios: Blending the use of ontologies with the use of LLMs, our methodology has proven to effectively produce sound and natural explanations, improving the brevity and clarity of ontology-based narratives. Furthermore, the use of ontology-informed LLMs also brings a powerful interactive capability. For some structured scenarios, patterns of explanations or pre-formatted messages may be useful. However, in non-structured cases such as human-robot interactive scenarios, robot explanations must adapt to user goals, clarification requests, and evolving dialogue context. For these cases, our methodology can be easily extended by slightly modifying the fourth and last step. Once the initial explanation is generated, further user prompts can be used to refine the explanation (e.g., asking for a shorter or clearer version). Although the current evaluation dimensions (i.e. brevity, clarity, and semantic coherence) can be useful to validate this adaptive capacity to some extent, a complete evaluation of user-specific requirements may ultimately require dedicated user studies.

4 CONCLUSION

This work establishes that integrating ontological abstraction and episodic storage of robot experiences with selective narrative construction and the priming capabilities of LLMs enables the generation of concise, clear, and semantically faithful explanations. The proposed hybrid approach facilitates the design of robots capable of providing explanations that are both technically reliable and socially intuitive. Future research will focus on extending the approach and its evaluation to interactive scenarios, and conducting experiments with users to assess the impact of such explanations on subjective perceptions of trust and collaboration.

ACKNOWLEDGMENTS

This work was partially supported by the European Union under the project ARISE (HORIZON-CL4-2023-DIGITAL-EMERGING-01-101135959), and by the Spanish National Research Council (CSIC) under grant number CSIC-BILAT23120.

REFERENCES

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable Agents and Robots: Results from a Systematic Literature Review. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1078–1088.

[2] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 6491–6501.

[3] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs.CL]

[4] Giancarlo Guizzardi and Nicola Guarino. 2024. Explanation, semantics, and ontology. *Data & Knowledge Engineering* 153 (2024), 102325.

[5] Pat Langley. 2024. From Explainable to Justified Agency. In *Explainable Agency in Artificial Intelligence*, Silvia Tulli and David W. Aha (Eds.). CRC Press, 1–20.

[6] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[7] Muhammad Raza Naqvi, Linda Elmhadhbi, Arkopaul Sarkar, Bernard Archimede, and Mohamed Hedi Karray. 2024. Survey on ontology-based explainable AI in manufacturing. *Journal of Intelligent Manufacturing* 35, 8 (01 Dec 2024), 3605–3627.

[8] Alberto Olivares-Alarcos, Muhammad Ahsan, Satrio Sanjaya, Hsien-I Lin, and Guillem Alenyà. 2026. Ontological grounding for sound and natural robot explanations via large language models. arXiv:2602.13800 [cs.RO] https://arxiv.org/abs/2602.13800

[9] Alberto Olivares-Alarcos, Antonio Andriella, Sergi Foix, and Guillem Alenyà. 2023. Robot explanatory narratives of collaborative and adaptive experiences. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11964–11971.

[10] Alberto Olivares-Alarcos, Sergi Foix, Júlia Borràs, Gerard Canal, and Guillem Alenyà. 2024. Ontological Modeling and Reasoning for Comparison and Contrastive Narration of Robot Plans. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2405–2407.

[11] Alberto Olivares-Alarcos, Sergi Foix, Júlia Borràs, Gerard Canal, and Guillem Alenyà. 2025. Ontological foundations for contrastive explanatory narration of robot plans. arXiv:2509.22493 [cs.RO] https://arxiv.org/abs/2509.22493

[12] Andrew Ponomarev and Anton Agafonov. 2024. Ontology-Based Explanations of Neural Networks: A User Perspective. In *Artificial Intelligence in HCI*, Helmut Degen and Stavroula Ntoa (Eds.). Springer Nature Switzerland, Cham, 264–276.

[13] Keith Rayner, Elizabeth R. Schotter, Michael E. J. Masson, Mary C. Potter, and Rebecca Treiman. 2016. So Much to Read, So Little Time: How Do We Read, and Can Speed Reading Help? *Psychological Science in the Public Interest* 17, 1 (2016), 4–34.

[14] Satrio Sanjaya, Muhammad Ahsan, Alberto Olivares-Alarcos, Hsien-I Lin, and Guillem Alenyà. 2025. An Investigation of Defect Inspection Performance through Human-Robot Collaboration. In *2025 International Automatic Control Conference (CACS)*. 1–6.

[15] Alexander Smirnov and Andrew Ponomarev. 2023. Ontology-Based Explanations of Neural Networks for Collaborative Human-AI Decision Support Systems. In *Proceedings of the Seventh International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'23)*, Sergey Kovalev, Igor Kotenko, and Andrey Sukhanov (Eds.). Springer Nature Switzerland, Cham, 353–362.

[16] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics* 3, 4 (2023), 100131.