Faithful Language-Based Explanations for Decision-Making Agents

Sahar Admoni
Technion – IIT
Haifa, Israel
saharad@campus.technion.ac.il

ABSTRACT

As artificial intelligence systems increasingly operate as autonomous decision-making agents, humans are often required to trust models whose internal reasoning processes are difficult to interpret. This challenge is especially pronounced in sequential decision-making settings, such as reinforcement learning, where behavior emerges over long horizons. While natural language explanations offer a promising interface for human understanding, they introduce a fundamental tension between interpretability and faithfulness, as fluent explanations may fail to reflect the factors that actually drive a model's decisions. My doctoral research studies how to generate language-based explanations for decision-making agents that are both human-usable and faithful to underlying model behavior. I approach this problem from two complementary perspectives: policy-level explanation of reinforcement learning agents through abstractive textual summaries, and evaluation and optimization of explanation faithfulness in large language models via decision–explanation alignment. Together, these directions form a unified agenda toward faithful language-based explanations for sequential decision-making agents.

KEYWORDS

Reinforcement Learning, LLMs, Explainable AI

ACM Reference Format:
Sahar Admoni. 2026. Faithful Language-Based Explanations for Decision-Making Agents. 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/YSGJ1412

1 INTRODUCTION

As machine learning systems are increasingly deployed as autonomous decision-making agents, humans are often required to trust, supervise, or collaborate with models whose internal reasoning processes are difficult to interpret [6, 11]. This challenge is especially acute in sequential decision-making settings such as reinforcement learning (RL) [24], where behavior emerges from long-horizon interactions rather than isolated predictions. In such settings, understanding an agent's strategy, priorities, and failure modes is essential for trust, debugging, and responsible deployment.

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/YSGJ1412

Natural language explanations offer a promising interface for human understanding [16], enabling complex behaviors to be abstracted in a form aligned with human reasoning about goals and actions. However, this promise carries a fundamental risk: language explanations can be fluent and persuasive while failing to reflect the factors that actually drive a model's decisions [8, 25]. As a result, they may provide a false sense of understanding rather than genuine insight.

The overarching goal of my PhD is to develop principled methods for explaining decision-making agents in ways that are both human-usable and faithful to the underlying behavior of the model. While explainable reinforcement learning (XRL) has emerged as an active research area [15, 19], existing approaches span both local and global explanations, yet many global methods remain largely descriptive or demonstration-based, placing substantial interpretive burden on the user. My research investigates language as a primary abstraction layer for explanation, with an explicit focus on how high-level descriptions can be synthesized from agent behavior while remaining faithful to the underlying decision process [9, 26]. Through work on policy-level summarization in reinforcement learning and on decision–explanation alignment in large language models (LLMs), I aim to understand how explanations can scale to complex agents without collapsing into post-hoc rationalizations.

2 POLICY-LEVEL EXPLANATIONS FOR REINFORCEMENT LEARNING AGENTS

My first line of research addresses the problem of explaining reinforcement learning agents at the level of their overall policy. Existing explainable RL (XRL) approaches often rely on saliency maps [7, 20], selected trajectories [4, 12], or symbolic surrogates such as rules or decision trees [18, 22]. While these methods provide local or fragmentary insights, they place a substantial interpretive burden on the user and rarely convey a coherent account of the agent's global strategy. As a result, users are left to infer intent and behavioral regularities from partial evidence [5]. Moreover, these methods often fail to capture the temporal dependencies and long-horizon reasoning characteristic of sequential decision-making.

To address this limitation, I reformulate policy interpretation as a language generation problem. Rather than explaining individual actions or states, my work focuses on generating abstractive textual summaries that describe an agent's behavior across multiple trajectories. These summaries aim to capture recurring strategies, goals, and characteristic failure modes, reflecting how humans naturally seek to understand decision-making agents.

In my work on policy-level summarization [2], I introduce a framework that translates an agent's experience buffer into structured natural language representations and synthesizes them into

coherent global summaries using a large language model, building on recent advances in language-based reasoning [10, 27]. To scale to long-horizon environments, the framework employs hierarchical summarization, and to mitigate variability in language generation, it aggregates multiple candidate summaries into a consensus description. Empirical evaluations demonstrate that these textual summaries align closely with expert-written analyses and are strongly preferred by users over demonstration-based explanation baselines [3].

This line of work establishes abstractive textual policy summarization as a viable paradigm for explaining reinforcement learning agents. At the same time, it exposes a fundamental challenge: as explanations become more abstract and fluent, assessing whether they faithfully reflect the agent's actual behavior becomes increasingly difficult. This observation directly motivates my second line of research.

3 FAITHFULNESS AND ALIGNMENT OF NATURAL LANGUAGE EXPLANATIONS

My second line of research examines the faithfulness of natural language explanations produced by large language models. As LLMs are increasingly embedded in user-facing and decision-support systems, their outputs and explanations are often trusted by end users even when independent verification is impractical [13, 23]. To support such trust, LLMs are commonly prompted to generate natural language explanations [14]. However, prior work shows that the features emphasized in these explanations frequently diverge from those that drive the model's decisions [21, 25], raising concerns about the reliability of post-hoc explanations in practice.

In my work on explanation faithfulness [1], I formalize faithfulness as the alignment between the features that drive a model's decision and the features emphasized in its explanation, building on prior definitions [9, 26]. While faithfulness broadly concerns whether explanations reflect the model's decision-making process, prior work suggests that many proposed faithfulness metrics capture forms of self-consistency [17]. To operationalize this notion, I compare feature-attribution distributions computed for a model's output with those computed for its corresponding explanation. This formulation enables explanation quality to be measured directly, rather than inferred indirectly through plausibility or correctness.

To study this alignment at scale, I build on attribution-based approaches that estimate feature influence through counterfactual interventions [25, 26]. However, applying such interventions to LLMs is computationally expensive. Building on this formulation, I introduce a large-scale benchmark that links model decisions, diverse explanations, and attribution vectors across multiple datasets, attribution methods, and model families. Using this benchmark, I show that decision–explanation alignment is largely orthogonal to task accuracy and that ranking-based metrics provide a more reliable signal of alignment than magnitude-based measures. Importantly, I demonstrate that explanation faithfulness can be improved through preference-based optimization without degrading task performance.

This line of work reframes explanation faithfulness as a property that can be systematically measured and optimized, rather than assumed. Together with my work on policy-level summarization, it contributes to a unified research agenda centered on developing language-based explanations that are both interpretable to humans and grounded in the actual behavior of decision-making models.

4 FUTURE DIRECTIONS

Building on my work on language-based explanations and explanation faithfulness, my future research focuses on using large language models to both *improve* and *analyze* decision-making behavior in a transparent manner. In particular, I plan to pursue two closely related directions centered on inference-time control in reinforcement learning and interpretable preference optimization for language models.

Inference-Time Improvement of Reinforcement Learning Agents. My first research direction investigates how large language models can be used to improve the behavior of reinforcement learning agents at inference time, without additional training. In the ReThink project, I study settings in which an existing policy generates candidate actions or trajectories, and an LLM is used to evaluate, rank, or refine these candidates based on task-specific reasoning and constraints. This approach treats the LLM as a deliberative component that operates on top of a fixed policy, enabling performance gains while preserving the original learning process.

A key focus of this work is interpretability. By expressing deliberation and action selection in language, inference-time interventions can produce intermediate rationales that expose why certain decisions are favored. This allows performance improvements to be inspected and analyzed, rather than introduced as opaque modifications to the policy.

Stability and Locality in Preference Optimization for Large Language Models. My second research direction focuses on understanding and controlling the stability of preference optimization methods for large language models. While preference-based optimization has become a central mechanism for aligning models with human judgments, its training dynamics are often poorly understood, and small changes in preference data or optimization settings can lead to disproportionate shifts in model behavior.

In this line of work, I study preference optimization objectives that explicitly constrain how far a model is allowed to deviate from a reference model. Rather than treating preference learning as an unconstrained optimization problem, my goal is to characterize and enforce locality in the optimization process, ensuring that updates remain stable, interpretable, and behaviorally consistent with the reference.

Together, these directions extend my dissertation from faithful explanation toward faithful decision-making, combining inference-time reasoning and interpretable optimization to improve both the performance and transparency of learning-based decision systems.

ACKNOWLEDGMENTS

Funded by the European Union (ERC, Convey, 101078158). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] Sahar Admoni, Ofra Amir, Assaf Hallak, and Yftah Ziser. 2025. Towards Large Language Models with Self-Consistent Natural Language Explanations. *arXiv preprint arXiv:2506.07523* (2025).

[2] Sahar Admoni, Assaf Hallak, Yftah Ziser, Omer Ben-Porat, and Ofra Amir. 2025. From Actions to Words: Towards Abstractive-Textual Policy Summarization in RL. In *The 25th International Conference on Autonomous Agents and Multi-Agent Systems*. https://openreview.net/forum?id=QP9k47Pm2c

[3] Dan Amir and Ofra Amir. 2018. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1168–1176.

[4] Ofra Amir, Finale Doshi-Velez, and David Sarne. 2018. Agent strategy summarization. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1203–1207.

[5] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. 2020. Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 2 (2020), 1–37.

[6] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[7] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. 2018. Visualizing and understanding atari agents. In *International conference on machine learning*. PMLR, 1792–1801.

[8] Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-adjusted simulatability: Can models generate non-trivial explanations of their behavior in natural language? *arXiv preprint arXiv:2010.04119* (2020).

[9] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685* (2020).

[10] Fangjun Li, David C Hogg, and Anthony G Cohn. 2024. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 18500–18507.

[11] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[12] Haozhe Liu, Mingchen Zhuge, Bing Li, Yuhui Wang, Francesco Faccio, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Learning to identify critical states for reinforcement learning from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1955–1965.

[13] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374* (2023).

[14] Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from Large Language Models faithful? *arXiv preprint arXiv:2401.07927* (2024).

[15] Stephanie Milani, Nicholay Topin, Manuela Veloso, and Fei Fang. 2022. A Survey of Explainable Reinforcement Learning. *arXiv preprint arXiv:2202.08434* (2022).

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

[17] Letitia Parcalabescu and Anette Frank. 2024. On Measuring Faithfulness or Self-consistency of Natural Language Explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 6048–6089. https://doi.org/10.18653/v1/2024.acl-long.329

[18] Xiangyu Peng, Mark Riedl, and Prithviraj Ammanabrolu. 2022. Inherently explainable reinforcement learning in natural language. *Advances in Neural Information Processing Systems* 35 (2022), 16178–16190.

[19] Erika Puiutta and Eric Veith. 2020. Explainable reinforcement learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*. Springer, 77–95.

[20] Nikaash Puri, Sukriti Verma, Piyush Gupta, Dhruv Kayastha, Shripad Deshmukh, Balaji Krishnamurthy, and Sameer Singh. 2019. Explain your move: Understanding agent actions using specific and relevant feature attribution. *arXiv preprint arXiv:1912.12191* (2019).

[21] Korbinian Randl, John Pavlopoulos, Aron Henriksson, and Tony Lindgren. 2024. Evaluating the reliability of self-explanations in large language models. In *International Conference on Discovery Science*. Springer, 36–51.

[22] Pedro Sequeira and Melinda Gervasio. 2020. Interestingness elements for explainable reinforcement learning: Understanding agents' capabilities and limitations. *Artificial Intelligence* 288 (2020), 103367.

[23] Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, et al. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561* 3 (2024).

[24] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[25] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2023), 74952–74965.

[26] Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2020. Measuring association between labels and free-text rationales. *arXiv preprint arXiv:2010.12762* (2020).

[27] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* 1, 2 (2023).