Enhancing User Understanding of Reinforcement Learning Agents Through Visual Explanations

Doctoral Consortium

Yotam Amitai
Technion - I.I.T
Haifa, Israel
yotama@campus.technion.ac.il

ABSTRACT

With the rapid advancement of Artificial Intelligence, the frequency of interaction between people and autonomous agents is on the rise. Effective human-agent collaboration requires that people understand the agent's behavior. Failing to do so may cause reduced productiveness, misuse, frustration, and even danger. Current explainable AI methods prioritize interpreting the local decisions of an agent, putting less emphasis on the challenge of conveying global behavior. Furthermore, there is a growing demand for explanation methods for agents in sequential decision-making frameworks such as reinforcement learning. Agent strategy summarization methods are used to describe the strategy of an agent to its user through demonstration. The summary's purpose is to maximize the user's understanding of the agent's aptitude by showcasing its behavior in a set of world states, chosen by some importance criteria. Extracting the crucial states from the execution traces of the agent in such a way as to best portray the agent's behavior is a challenging task. My thesis tackles this objective by adding to the equation the context in which the user interacts with the agent. This research proposes novel methods for generating summary-based explanations for reinforcement learning agents

KEYWORDS

Explainable Reinforcement Learning, Interactive, Contrastive

ACM Reference Format:
Yotam Amitai. 2023. Enhancing User Understanding of Reinforcement Learning Agents Through Visual Explanations : Doctoral Consortium. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023,* IFAAMAS, 3 pages.

RESEARCH STATEMENT

The prevalence of AI agents in our everyday lives is on the rise, from transportation solutions to algorithmic trading or medical care recommendations. These agents can significantly improve and advance society in numerous domains, such as healthcare, education, and transportation. However, these agents are not independent of their environment. Most importantly, their ability to function together with and alongside people plays a crucial role in both their success and their adoption. People interacting with agents must be able to predict and comprehend their behavior. For instance, a driver of an autonomous car must be prepared for a scenario in which the car malfunctions and the driver must take over, while a physician must understand an agent-proposed treatment to ensure it suits the patient.

User misunderstanding of agent behavior can incur significant negative outcomes to their interaction, such as mistrust, reduced effectiveness, and even dangerous situations. Alas, as people's mental models of complex system behaviors are typically incomplete, parsimonious, and unstable, understanding the behavior of agents can be a difficult task.

Not only is it beneficial to explain AI behavior to people, it will also likely be *necessary*. The EU has already considered the "right to explanation" as part of the General Data Protection Regulation (GDPR), stating with respect to automated decision-making that "[the data subject should have] the right ... to obtain an explanation of the decision reached"[6]. Explanation standards will need to be devised for more complex AI agents. This thesis can help form the basis for creating such standards and will develop methods that provide capabilities for adhering to them.

With the recognition of the importance of human understanding of agents' behavior, there is a growing interest in developing "explainable AI" methods [5, 8]. Existing methods primarily focus on providing explanations for specific decisions made by a machine learning model (e.g., whether a tumor is benign), by showing the features that contributed most to the decision.

Fewer works have addressed the problem of explaining agents' actions in sequential decision-making settings [9]. These focus mainly on "local" explanations, e.g., showing what information a game-playing agent attends to in a specific game state [7]. Less abundant in the field are methods which are concerned with demonstrating the "global" behavior of a model. Jacobs et al. [12] exposed the need for global explanations, as expressed by clinicians stating they prefer understanding the model as a whole, at the beginning as opposed to assessing each decision individually.

Thus, current state-of-the-art explainable AI methods do not adequately address the challenge of conveying the *global* behavior of agents operating in large state spaces over an extended time duration. Moreover, most explainable AI approaches focus on the technology and lack careful consideration of users' needs. They are thus at risk of being useful only to the designers of the algorithms and not to their intended users [14].

One method for conveying agent behavior to the end user is through strategy summarization methods [2]. This visual explanation method allows the user a glimpse at the agent performing its task in a selected set of world-states based on some criteria, such as the importance of a decision [1, 10] or an ability to reconstruct the agents' policy [11]. Using these methods, a visual summary

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

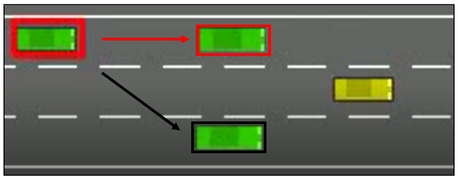

Figure 1: Visualizing agent disagreements: Two agents start at the same state (top left), where their policies diverge: one agent (in red box) chose to stay in the top lane, while the other agent (in black box) switched to the bottom lane.

can be generated for each agent, allowing users to gain insights into its capabilities and strategy without the need to witness them firsthand or observe the agent for an extended time duration.

In this thesis, we plan to develop new visual explanation methods inspired by the principles for "good" explanations described in the social sciences [13, 15]. These principles suggest that good explanations have several of the following elements: *i)* contrastive, *ii)* containing a selected subset of causes, *iii)* dependent on social context, *iv)* describing the abnormal, *v)* truthful, *vi)* consistent with prior beliefs, or *vii)* being general and probable. We aim to develop methods chiefly based on agent strategy summarization techniques. One benefit of following this approach is that the visual output provides a rich context for the agent's behavior in its environment, as opposed to textual or rule-based explanations, and allows the user to comprehend the entire interaction scene and derive from it further information. I intend to extend this approach in significant ways to provide additional capabilities, by drawing on insights and methodologies from AI, cognitive science, and human-computer interaction literature.

PRELIMINARY RESULTS

We now describe current work, efforts, and results achieved through this line of work.

Agent Comparisons. Providing comparison methods to help portray critical differences between agents, by comparing them in a dependent manner, such as to enable users to choose which agent better suits their needs.

In my first PhD project [3], we *i)* introduced and formalize the problem of comparing agent policies; *ii)* developed DISAGREEMENTS, an algorithm for generating visual contrastive summaries of agents' behavioral conflicts, optimized for the agent comparison task, and *iii)* evaluated our approach through human-subject experiments, demonstrating that summaries generated by DISAGREEMENTS lead to improved user performance compared to HIGHLIGHTS summaries. A comparison example is shown in Figure 1.

Contrastive Explanations. Generating visual contrastive explanations for a single agent's decisions by visualizing alternative paths it could have taken instead, i.e., had it not chosen the specific action that it had.

In my thesis, I focus on helping participants develop correct mental models of the agent's preferences and understanding the trade-offs between alternative actions. To this end, we developed a new local explanation method, "contrastive highlights", which draws inspiration from global policy summaries. This method visualizes both the trajectory chosen by the policy, along with a simulated one highlighting a path had the agent chosen a different, contrastive, action for a given state. For example, the contrastive action may be the second-best action as predicted by the agent. This approach aims to provide more information regarding the decision made by the agent by showing side-by-side the *outcomes* of the chosen action and an alternative one.

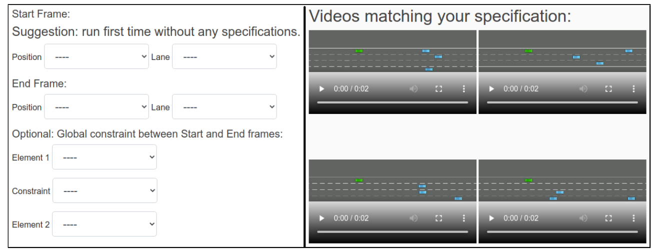

Figure 2: Left: Query Interface; Right: Explanation Interface.

Interactive Explanations. Allowing user preferences to shape the explanations generated. This will be achieved through self-selected summary states or choosing from multiple off-the-shelf expert explanations generated in advance. These explanations can be based on both domain-specific and domain-agnostic methods such as clustering.

We developed an interactive XRL tool that aims to assist users to comprehend an agent in a global manner [4] (Figure 2). Using iterative pilot studies, we were able to design the tool according to laypeople's needs and cognitive capabilities. Our tool generates clips of the agent interacting with its environment. The user controls which clips will be presented by feeding queries that specify properties of clips of interest. The interaction with the tool resembles a dialogue: the user enters a query, receives a handful of clips that answer it, the user can then refine her query, and the process continues.

SUMMARY & FUTURE PLANS

To summarize, this thesis will focus on developing user-focused visual explanation methods for conveying the behavior of agents in sequential decision-making settings to a human counterpart, be it a layperson end-user wishing to further grasp its capabilities or the developers themselves for debugging intentions. In addition to further improving and expanding my existing contributions, I plan on pursuing additional research directions such as:

Visualising Domain Attributes. Conveying the aptitude of an agent in the domain is not enough, there is a need for enhancing the user's understanding of the domain itself and its dynamics.

Visual Summaries for Non-Visual Domains. Developing generic methods for conveying changes in tabular data states to meaningful visualizations. Highlighting trends and goal emphasis can be used to better and more intuitively portray progression in such feature spaces, thus broadening the scope of visual summaries.

Visual explanations for ad-hoc human-robot teamwork. Leveraging the strength and intuitiveness of visual explanations for online human-robot teamwork tasks in collaborative settings.

All proposed approaches and methods have been or will be tested through user studies.

REFERENCES

[1] Dan Amir and Ofra Amir. 2018. HIGHLIGHTS: Summarizing Agent Behavior to People. In *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

[2] Ofra Amir, Finale Doshi-Velez, and David Sarne. 2019. Summarizing agent strategies. *Autonomous Agents and Multi-Agent Systems* 33, 5 (2019), 628–644.

[3] Yotam Amitai and Ofra Amir. 2022. "I Don't Think So": Summarizing Policy Disagreements for Agent Comparison. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*.

[4] Yotam Amitai, Guy Avni, and Ofra Amir. 2022. Interactive Explanations of Agent Behavior. In *ICAPS 2022 Workshop on Explainable AI Planning*.

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

[6] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[7] Sam Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. 2017. Visualizing and Understanding Atari Agents. *arXiv preprint arXiv:1711.00138* (2017).

[8] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

[9] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. 2021. Explainability in deep reinforcement learning. *Knowledge-Based Systems* 214 (2021), 106685.

[10] Sandy H Huang, Kush Bhatia, Pieter Abbeel, and Anca D Dragan. 2018. Establishing Appropriate Trust via Critical States. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3929–3936.

[11] Sandy H Huang, David Held, Pieter Abbeel, and Anca D Dragan. 2017. Enabling robots to communicate their objectives. *Autonomous Robots* (2017), 1–18.

[12] Maia Jacobs, Jeffrey He, Melanie F Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. *arXiv preprint arXiv:2102.00593* (2021).

[13] Tim Miller. 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2018).

[14] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, Vol. 36.

[15] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.