Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Journal Track

Local and Global Explanations of Agent Behavior: Integrating Strategy Summaries with Saliency Maps (Extended Abstract) *

Tobias Huber[1] , **Katharina Weitz**[1] , **Elisabeth André**[1] and **Ofra Amir**[2]
[1]Universität Augsburg, Universitätsstraße 6a, Augsburg, Germany
[2]Technion - Israel Institute of Technology
{tobias.huber, katharina.weitz, andre}@informatik.uni-augsburg.de, oamir@technion.ac.il

Abstract

With advances in reinforcement learning (RL), agents are now being developed in high-stakes application domains such as healthcare and transportation. Explaining the behavior of these agents is challenging, as they act in large state spaces, and their decision-making can be affected by delayed rewards. In this paper, we explore a combination of explanations that attempt to convey the *global* behavior of the agent and *local* explanations which provide information regarding the agent's decision-making in a particular state. Specifically, we augment strategy summaries that demonstrate the agent's actions in a range of states with saliency maps highlighting the information it attends to. Our user study shows that intelligently choosing what states to include in the summary (global information) results in an improved analysis of the agents. We find mixed results with respect to augmenting summaries with saliency maps (local information).

1 Introduction

The maturing of artificial intelligence (AI) methods has led to the introduction of intelligent systems in areas such as healthcare and transportation [Stone *et al.*, 2016]. Since these systems are used by people in such high-stakes domains, it is crucial for users to be able to understand and anticipate their behavior.

In this paper, we focus on the problem of describing and explaining the behavior of agents operating in sequential decision-making settings, which are trained in a deep reinforcement learning framework. In particular, we explore the usefulness of *global* and *local* post-hoc explanations [Molnar, 2019] of agent behavior. Global explanations describe the overall policy of the agent, that is, the actions it takes in different regions of the state space. An example of such global explanations are strategy summaries [Amir *et al.*, 2019], which show demonstrations of the agent's behavior in a carefully selected set of world states. Local explanations, in contrast, aim to explain specific decisions made by the agent. For instance, saliency maps are used to show users what information the agent is attending to [Huber *et al.*, 2019].

We explore the combination of global and local information describing agent policies. The motivation for integrating the two approaches is their complementary nature: while local explanations can help users understand what information the agent attends to in specific situations, they do not provide any information about its behavior in different contexts. Similarly, while demonstrating what actions the agent takes in a wide range of scenarios can provide users with a sense of the overall strategy of the agent, it does not provide any explanations as to what information the agent considered when choosing how to act in a certain situation.

To examine the benefits of these two complementary approaches and their relative usefulness, we propose a joint local and global explanation approach for RL agents. Specifically, we adapt the HIGHLIGHTS-DIV algorithm for generating strategy summaries [Amir and Amir, 2018] such that it can be applied to deep learning settings, and integrate it with saliency maps that are generated based on Layer-Wise Relevance Propagation (LRP) [Huber *et al.*, 2019]. We combine these two approaches by adding saliency maps, which show what information the agent attends to, to the summary generated by HIGHLIGHTS-DIV.

We evaluate this combination of global and local explanations in a user study that measures the participants' mental model of different agents, whether they can identify the better performing agent, and their satisfaction with the provided explanations. Our results show that participants who were shown HIGHLIGHTS-DIV summaries performed better compared to participants who were shown likelihood-based summaries, and were also more satisfied. We find both potential benefits as well as limitations with respect to adding saliency maps to summaries.

2 Empirical Evaluation

In this section, we describe the user study we conducted to evaluate the combination of global and local explanations of RL agents. For additional details see Huber *et al.* [2021].

Empirical Domain and Agent Training. The environment we used for our experiments is the Atari 2600 game MsPacman included in the Arcade Learning Environment (ALE) [Bellemare *et al.*, 2013], henceforth referred to as Pacman.

*The paper was initially published in the journal of Artificial Intelligence (AIJ): https://doi.org/10.1016/j.artint.2021.103571

Here, Pacman obtains points by eating food pellets while navigating through a maze and escaping ghosts. When Pacman eats a special power pill, the ghosts turn blue for a short time during which Pacman can eat them to get bonus points. The ALE states are based on the raw pixel values of the game and the actions correspond to nine meaningful actions achieved with an Atari 2600 controller. To evaluate participants' ability to differentiate between alternative agents and analyze their strategies, we modified the reward function to obtain agents that behave qualitatively different:

- ***Regular agent***: This agent was trained using the default reward function of the ALE, which measures the increase in score between two states.
- ***Power pill agent***: This agent was trained using a reward function that only assigned positive rewards to eating power pills.
- ***Fear-ghosts agent***: This agent used the default ALE reward function but was given an additional negative reward of -100 when being eaten by ghosts, causing it to more strongly fear ghosts.

For training the Pacman agents, we used the OpenAI baselines [Dhariwal *et al.*, 2017] implementation of the DQN-algorithm [Mnih *et al.*, 2015]. Each agent was trained for 5 Million steps. At the end of this training period, the best-performing policy is restored. As basis for our survey, we then recorded a stream of $10,000$ steps for each agent. We computed the average in-game score of each trained agent over the entire stream. This allows us to objectively determine which agent achieved the most points and therefore gives us a ground-truth for the agent comparison task.

Experimental Conditions. To evaluate the benefits of integrating global and local explanations, and their relative importance, we used four experimental conditions:

- **Likelihood-based Summaries (*L*)**: The summaries in this condition consisted of states that the agent was likely to encounter during gameplay. To generate these summaries, we uniformly sample state-action pairs from the streams of the Pacman agents playing the game. Because of the random component of these summaries, it is possible that a single summary is, by chance, particularly good or particularly bad. Therefore, we generated 10 different likelihood-based summaries and randomly assigned them to participants in this condition.
- **HIGHLIGHTS-DIV summaries (*H*)**: Participants were shown summaries generated by the HIGHLIGHTS-DIV [Amir and Amir, 2018] algorithm. The algorithm generates a summary that includes the most "important" states the agent encounters in simulations, while also attempting to increase diversity in the summary by adding states only if they are sufficiently different from states already included in the summary. A state is considered to be important if there is a huge difference between q-values for different actions.
- **Likelihood-based Summaries+Saliency (*L+S*)**: These summaries included the same states as those shown in the *L* summaries. However, each image was overlayed with a local saliency map generated by a variant of LRP [Huber *et al.*, 2019], which analyzes the particular decision of the agents.
- **HIGHLIGHTS-DIV summaries+Saliency (*H+S*)**: Analogously, these summaries included the same states as those shown in the *H* summaries, where each image additionally was overlayed with a saliency map.

Each summary included 5 base states chosen either based on likelihood or by HIGHLIGHTS-DIV. For each state, we included a surrounding context window of 10 states that occurred right before and after the chosen state.

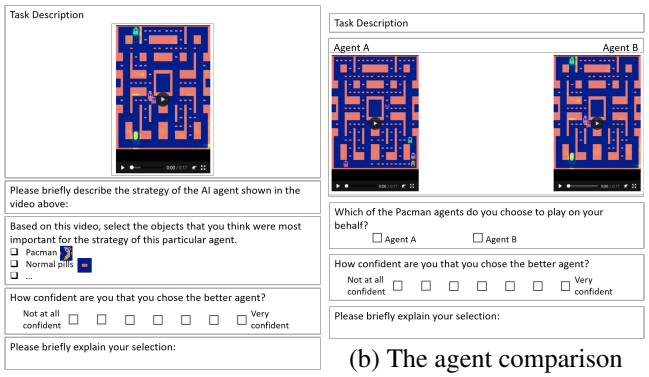
Participants and Procedure. We recruited participants through Amazon Mechanical Turk ($N = 133$, the majority of participants were between the ages of 25 and 44, 47 females). Participation was limited to people from the US, UK, or Canada with task approval rate greater than 97%. Each participant was assigned to one of the four conditions (*L*:33, *H*:33 *L+S*:34, *H+S*:33).

Participants were first asked to answer demographic questions (age, gender) and questions regarding their experience with Pacman and their views on AI. Then, they were shown a tutorial explaining the rules of Pacman and were asked to play the game to familiarize themselves with it. To verify that participants understood the rules, they were required to successfully complete a quiz. In conditions *L+S* and *H+S*, we also included an explanation and a quiz about saliency maps. If the participants answered wrongly in the quiz, they had to redo it until it was correct. Then, they proceeded to the main experimental tasks. Participants received a $4 base payment and an additional bonus of 10 cents for each correct answer. The study protocol was approved by the Institutional Review Board at the Technion (#2018-059).

Main Tasks. We aimed to investigate three aspects related to explanations: (1) participants' mental models of the agents, (2) participants' ability to assess agents' performance (appropriate trust), and (3) participants' satisfaction with the explanations presented.

Task 1: Eliciting Mental Models through Retrospection. By mental model, we understand the cognitive representation that the participant has about a complex model [Halasz and Moran, 1983; Norman, 2014], in our case, the agent. We used a task reflection method inspired by prior works [Anderson *et al.*, 2019; Sequeira *et al.*, 2019; Hoffman *et al.*, 2018]. This task asked the participants to analyze the behavior of the three different AI agents. Specifically, participants were shown the video summary (according to the condition they were assigned to) and were asked to briefly describe the strategy of the AI agent (textual) and to select up to 3 objects that they think were most important to the strategy of the agent. They were also asked how confident they were in their responses, and to justify their reasoning (textual). Fig. 1 (a) shows a sketch of a **retrospection task**. The ordering of the agents was randomized.

Task 2: Measuring Appropriate Trust through Agent Comparison. We understand appropriate trust to be a well-calibrated trust that matches the true capabilities of a technical system [Lee and See, 2004]. We measured appropriate trust using an **agent comparison task**. Here, participants were shown summaries of two of the three agents



(a) The retrospection task.

(b) The agent comparison task.

Figure 1: Sketches of the two tasks.

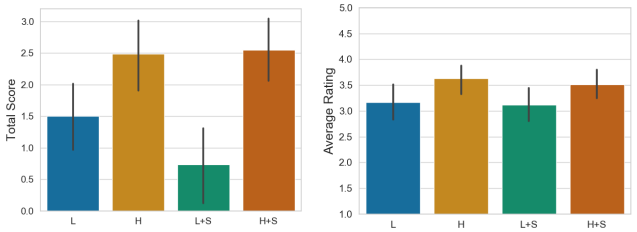
at a time and were asked to indicate which agent performs better (similar to tasks used in [Amir and Amir, 2018; Selvaraju *et al.*, 2020]). They thus made three comparisons. We did not ask the participant directly about their trust in the two agents shown. Instead, the participants had to choose one of the two agents that they would like to play on their behalf (see Fig. 1 (b)). This implicit question reveals which agent participants consider more reliable and qualified for the task. As in the retrospection task, they were asked to indicate their level of confidence and provide a textual justification for their decision. The ordering of the three agent comparisons was randomized.

Explanation satisfaction questions We measured participants' subjective satisfaction using questions adapted from the explanation satisfaction questionnaire [Hoffman *et al.*, 2018]. We did this separately after completing the retrospection task and after completing the agent comparison task, as we hypothesized there may be differences in the usefulness of the explanation methods for these different tasks.

Analysis. To make sure that the participants involved in our analysis did in fact watch the videos of the agents, we recorded whether the participants clicked play on each video in addition to how often each video was paused. We remove participants from each of the two main tasks if they did not watch enough videos. For evaluating the object selection in the retrospection task we use a scoring system, where two of the authors involved in agent training assigned a score to each item for each agent before the study. Selecting more than three items results in a score of zero. To evaluate participants' textual responses we use summative content analysis [Hsieh and Shannon, 2005]. An independent coder (not one of the authors) classified the textual responses into several categories which were not mutually exclusive. To evaluate the correctness of participants' descriptions of the agents' strategies, we implemented a simple scoring system. For each agent and each category, we decided whether it is correct, irrelevant, or wrong, based on predefined 'ground-truth' answers that two of the authors, who were involved in the training of the agents, wrote for each agent before the study started. The exact scoring functions can be found in the open-sourced code (https://github.com/HuTobias/HIGHLIGHTS-LRP).

Task	Variable	Effect of strategy summaries: $H > L$	$H{+}S > L{+}S$	Effect of saliency maps: $L{+}S > L$	$H{+}S > H$
retrospection task	score	0.008^{*}	$3.3e-05^{*}$	0.965	0.514
	satisfaction	0.021^{*}	0.035^{*}	0.677	0.710
	text score				$0.088^{\dagger}$
agent comparison task	score	0.014^{*}	0.180	$0.062^{\dagger}$	0.307
	satisfaction	0.147	0.235	0.627	0.833

Table 1: Summary of all significance tests (calculated with Mann-Whitney tests). The * denotes statistically significant differences and † denotes a p-value < 0.1.



(a) Total score of the object selection for all three agents.

(b) Average of all explanation satisfaction questions.

Figure 2: Average results of the participants in each condition in the retrospection task.

3 Results

We analyzed our main hypotheses using the non-parametric Mann-Whitney test [McKnight and Najab, 2010], as our dependent variables are not normally distributed. During the retrospection task, users obtained significantly higher scores and were more satisfied in the HIGHLIGHTS conditions *H* and *H+S* compared to the likelihood-based conditions *L* and *L+S* (Tab. 1, Fig. 2). Furthermore, there was a trend indicating that a combination of saliency maps and HIGHLIGHTS (condition *H+S*) was the most helpful for describing the agents' strategies textually (Tab. 1, Fig. 3). In the agent comparison task, users in condition *H* performed significantly better then users in condition *L* and there is a positive trend for users in *L+S* compared to *L* (Tab. 1, Fig. 4). For additional results including effect sizes refer to Huber *et al.* [2021].

4 Discussion

Strategy Summarization. The results of our study reinforce prior findings [Amir and Amir, 2018] showing that summaries generated by HIGHLIGHTS-DIV significantly improve participants' performance in the agent comparison task compared to likelihood-based summaries, and generalizes this result to RL agents based on neural networks. Furthermore, they show that HIGHLIGHTS-DIV summaries were more useful for analyzing agent strategies and were preferred by participants. Overall, in our study, the choice of states that are shown to participants was more important than the inclusion of local explanations in the form of saliency maps.

Limitations and Potential of Saliency Maps. In contrast to previous studies about saliency maps for image classification tasks, which found weak positive effects for saliency maps [Alqaraawi *et al.*, 2020; Selvaraju *et al.*, 2020], there are no significant differences between the saliency and non-saliency conditions in our study. When examining partici-

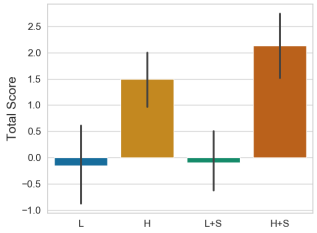
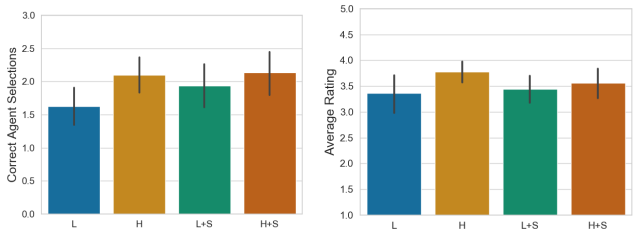


Figure 3: Participants' score for the textual descriptions of agents' strategies in the retrospection task (summed over all three agents).



(a) Number of correct agent selections.

(b) Average of all explanation satisfaction questions.

Figure 4: Average results of the participants in each condition in the agent comparison task.

pants' answer justifications, we observe that most participants do not mention utilizing the saliency maps, which may provide a partial explanation for their lack of contribution to participants' performance. Participants' comments also reflect their dissatisfaction with saliency maps, e.g., "I do not believe that the green highlighting was useful or relevant".

Based on participants' comments, we note some possible usability barriers of saliency maps. First, when saliency maps are shown as part of a video, it may be difficult for users to keep track of the agent's attention, compared to displays of static saliency maps, as done in previous studies [Selvaraju *et al.*, 2020; Anderson *et al.*, 2019; Alqaraawi *et al.*, 2020]. We tried to mitigate this by using a selective saliency map generation algorithm (LRP-argmax) and interpolating between selected saliency maps to reduce the amount of information, as well as allowing participants to pause the video at any time. However, this does not seem to be enough.

Second, participants were not accustomed to interpreting saliency maps, which can be non-intuitive to non-experts. One of the participants commented: "...I don't know if I would prefer an AI that 'looked' around more at the board, or focused more in a small area to accomplish a task". It is possible that prior studies which used saliency maps for interpreting image classification [Alqaraawi *et al.*, 2020; Selvaraju *et al.*, 2020] did not encounter this problem due to the more intuitive nature of the task. Interpreting a visual highlighting for image classification only requires identifying objects that contributed to the classification, while in RL there is an added layer of complexity as interpretation also requires making inferences regarding how the highlighted regions affect the agent's long-term policy.

Even though saliency maps do not significantly increase participants' scores in the simple object selection part of the retrospection task, they do result in improved scores in the textual strategy description. The difference between the HIGHLIGHTS-DIV conditions *H+S* and *H* is similar to the one observed by Anderson *et al.* [2019], who also used a strategy description task. The poor result of our likelihood-based condition *L+S* can be explained by the fact that Anderson et al. implicitly chose meaningful states, which we only did with our global explanation method in the HIGHLIGHTS-DIV conditions. A possible reason for the difference between the object selection and the textual strategy description subtasks is the higher complexity of strategy description. It requires participants to not only identify the correct objects but also to describe how they are used. Under this assumption, the increased performance of participants in condition *H+S* suggests that saliency maps are useful for putting the objects in the correct context. For example, participants' textual descriptions showed that, while the non-saliency groups know that Pacman is important (most likely based on the fact that it is important for them as players), they do not identify it as a central source of information for the agent.

In the agent comparison task, we observed that saliency maps alone improve participants' ability to place appropriate trust in different agents when comparing conditions *L* and *L+S*. There, performance was comparable to the performance of participants in the HIGHLIGHTS-DIV conditions, *H* and *H+S*. The lacking improvement of condition *H+S* compared to *H* might be explained by the accessibility issues of saliency maps mentioned earlier.

Combination of Local and Global Explanations. It is important to note that the positive effects of saliency maps in the retrospection task are only visible in the HIGHLIGHTS-DIV condition *H+S*, reinforcing our claim that the choice of states is crucial for explaining RL agents. Therefore, even if the limitations of saliency maps mentioned above are addressed, the potential benefits might only be visible and likely reinforced by a combination with strategy summarization techniques. We note that studies that evaluate local explanations typically implicitly make a global decision about which states to present local explanations for [Anderson *et al.*, 2019; Madumal *et al.*, 2020]. Our results suggest that this implicit choice may have a substantial impact on participants' understanding of agent behavior.

In the retrospection task, we observed that local explanations in the form of saliency maps were useful for identifying what objects the agent attends to (e.g. Pacman), while strategy summaries were more useful for identifying the agent's goals (e.g. special power pills). The local saliency maps contribute to users' understanding of the agents *attention*, as they reflect the information the agent attends to, while strategy summaries contribute to users' understanding of the agent's *intentions*, as they reflect how the agent acts. Taken together, our results suggest that there is potential for a combined explanation framework in the future if the accessibility issues of saliency maps are addressed.

References

[Alqaraawi *et al.*, 2020] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for

convolutional neural networks: A user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 275–285, New York, NY, USA, 2020. Association for Computing Machinery.

[Amir and Amir, 2018] Dan Amir and Ofra Amir. Highlights: Summarizing agent behavior to people. In *Proc. of the 17th International conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2018.

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

[Anderson *et al.*, 2019] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Alan Fern, and Margaret Burnett. Explaining reinforcement learning to mere mortals: An empirical study. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 1328–1334. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Bellemare *et al.*, 2013] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, jun 2013.

[Dhariwal *et al.*, 2017] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. https://github.com/openai/baselines, 2017. Accessed: 2020-05-01.

[Halasz and Moran, 1983] Frank G Halasz and Thomas P Moran. Mental models and problem solving in using a calculator. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 212–216, 1983.

[Hoffman *et al.*, 2018] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[Hsieh and Shannon, 2005] Hsiu-Fang Hsieh and Sarah E Shannon. Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288, 2005.

[Huber *et al.*, 2019] Tobias Huber, Dominik Schiller, and Elisabeth André. Enhancing explainability of deep reinforcement learning through selective layer-wise relevance propagation. In *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 188–202. Springer, 2019.

[Huber *et al.*, 2021] Tobias Huber, Katharina Weitz, Elisabeth André, and Ofra Amir. Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *Artificial Intelligence*, 301:103571, 2021.

[Lee and See, 2004] John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80, 2004.

[Madumal *et al.*, 2020] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. Explainable reinforcement learning through a causal lens. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, pages 2493–2500. AAAI Press, 2020.

[McKnight and Najab, 2010] Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1, 2010.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[Molnar, 2019] Christoph Molnar. Interpretable machine learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book, 2019. Accessed: 2020-05-01.

[Norman, 2014] Donald A Norman. Some observations on mental models. In *Mental models*, pages 15–22. Psychology Press, 2014.

[Selvaraju *et al.*, 2020] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.*, 128(2):336–359, 2020.

[Sequeira *et al.*, 2019] Pedro Sequeira, Eric Yeh, and Melinda T Gervasio. Interestingness elements for explainable reinforcement learning through introspection. In *IUI Workshops*, page 7, 2019.

[Stone *et al.*, 2016] Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, Press William, Saxenian AnnaLee, Shah Julie, Tambe Milind, and Teller Astro. Artificial intelligence and life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel*, 2016.