Explainable Deep Reinforcement Learning through Introspective Explanations

Nils Wenninghoff[1,*]

[1]Karlsruhe Institute of Technology (KIT), Am Fasanengarten 5, Karlsruhe, 76131, Germany

Abstract

Nowadays modern life is influenced in many ways by artificial intelligence systems. This could be a decision of an artificial intelligence, an autonomous vehicle or functions of consumer devices. In many domains artificial intelligence can act faster, more precise or better than a human could. At the same time, most decision processes of artificial intelligence systems are opaque black box systems. Making these decision processes understandable to humans could increase the reliability, maintainability and trustworthiness of such systems. (World Model based) Deep Reinforcement Learning Agents are often too complex to be explained because decisions are influenced by previous states and actions. To address this complexity, a new multistep explanation will be introduced. The proposed approach uses learned world models and policies to imagine the future states and predicting actions for them. The imagination is used as explanation to understand the plan of the agent and to know and validate the targeted goal. A user study design was created to evaluate the effectiveness of this new form of explanation.

Keywords

Explainable Artificial Intelligence, Deep Reinforcement Learning, World Model, Introspective Explanation

1. Context and Motivation

The rapid advancement of Artificial Intelligence (AI) systems has enhanced their ability to handle complex tasks. However, there's been a notable focus on optimizing task-specific performance metrics, neglecting transparency and explainability, resulting in opaque black box models. Achieving transparent and interpretable AI systems is crucial for ensuring reliability and safety. Within the realm of Deep Reinforcement Learning (DRL), the need for explanations is essential due to its important role in AI. In DRL the agent tries to learn a policy that maximizes the expected cumulative rewards. To do this, some approaches use a learned or given model of the environment the agent is acting in. The task is defined by the reward function which evaluates the actions of the agent with respect to the task. This feedback can be given after each step, sparse at irregular intervals or only at the end of an episode. DRL agents' performance relies on various factors including the selected algorithm, training design, data quality, and reward functions making it challenging to measure their individual contributions. This complexity often leads to optimization through trial-and-error rather than causal decision-making.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

nils.wenninghoff@kit.edu (N. Wenninghoff)

0000-0002-3788-9291 (N. Wenninghoff)

The multidisciplinary research domain Explainable Artificial Intelligence (XAI) addresses these challenges by developing techniques to make the decisions of AI-Systems human understandable [1]. Thereby presenting a dual challenge: technical, in extracting and generating relevant information from opaque AI models; and social, in presenting this information in an optimal manner to various stakeholders.

Traditional XAI techniques fall short in explaining the decision-making process of DRL agents [2]. Unlike deep learning classifiers, which operate in a single step, DRL agents act sequentially to maximize expected returns over time. While classifiers aim for optimal solutions in each state, DRL agents may prioritize less optimal actions to improve overall returns. Sparse rewards pose an additional challenge, as not every step is assigned an individual reward, but instead a number of steps are evaluated with one reward. This fundamental difference necessitates higher-order explanations, such as first-order explanations incorporating agents' goals and beliefs [2]. Recent research has begun addressing these challenges by developing specialized methods for DRL and adapting existing XAI approaches [3, 2]. Deep learning models' black box nature obscures the rationale behind AI decisions making it challenging for users to assess fairness and reliability [4]. This opacity can be dangerous if humans place unwarranted trust in non-transparent systems, potentially compromising reliability [5].

The lack of transparency and interpretability is not only a risk for end-users, but also complicates the development process and the regulation of such AI-systems. Without transparency, it is difficult to ensure legal compliance for developers and operators. As decision-making takes place within the black box, it cannot be guaranteed that the AI will always make legally compliant decisions [1]. Validating this decision-making process is a major challenge in the development process of AI-systems.

2. Related Work

Recent advancements in AI research, especially in DRL, have been notable. However, there's a growing gap between transparent and interpretable solutions and the current state-of-the-art performance across tasks [3]. This often results in models being either high-performing or interpretable, but rarely both. Although there are examples showing the feasibility of achieving both performance and interpretability [6], the widespread lack of interpretability raises significant trust and applicability concerns, particularly in safety-critical domains. It's crucial to ensure the traceability of decision-making processes in robust AI systems for these domains. World models are experiencing increasing research interest and enable the learning of strategies exclusively on imaginary states. The world model learns to represent an abstracted copy of the environment. Particularly in real-world environments or complex simulations, world models can significantly speed up the training process. However, the disadvantage is that only an abstracted version of reality is learned and this may not fully correspond to the real world. Nevertheless, world models have already been used in various control tasks [7]. Humans often explain their actions based on their intentions, with desires explaining overall goals and beliefs detailing the necessary conditions and means to achieve them [8]. In addition research from the psychology domain suggest that humans naturally interpret observed behaviors as goal-directed actions [9]. As explanations involve both an explainer and a recipient, it's crucial to

cater to the needs and capabilities of both. Humans typically rely on goals and solution paths for explanations, and strive to understand others' intentions by knowing their objectives. Adopting a similar approach could enhance the understandability of AI actions for humans. Standard XAI methods cannot be sufficiently applied to DRL agents [10], as they focus on single-step explanations and fail to account for sequential decision processes and temporal dependencies. The demand for explainable models is particularly strong in Explainable Deep Reinforcement Learning (XDRL). XDRL includes different explanation subjects, including the global model, policy or objective model, local responses, and outcomes [3]. Based on [2] explanations can be categorized into different categories or tiers. Zero-Tier explanations rely only on single-step input-output, while first-tier explanations incorporate desired goals and beliefs. Second-tier explanations rationalize an agent's behavior by predicting the intentions of others, while n-th tier explanations illustrate how an agent adjusts its actions based on perceived expectations.

XDRL methods, which utilize expected outcomes as explanations, are currently limited to specific DRL algorithms due to constraints and limitations [10]. However, research has demonstrated the effectiveness of using expected outcomes as explanations [10]. These explanations could address the divers explanatory requirements of different stakeholders. Each stakeholder has an individual need for explanation. Based on these needs, stakeholders can be grouped together. At the same time, however, the individual needs within a group must also be taken into account. For example, an inexperienced developer needs a different explanation than an experienced developer. At the same time, certain needs may exist in several stakeholder groups, resulting in an explanation that addresses the need of different stakeholders [1].

3. Research Focus and Objectives: Defining Questions, Hypotheses, and Goals

The focus of this research project is the development of a new explanation method for deep reinforcement learning and its evaluation as an aid for developers in the development process. The following hypothesis is to be investigated.

For model-based DRL agents, explaining the desired goals and planned paths for each action taken increases the transparency and interpretability of the agent, thereby improving the ability of humans to validate, and understand the agents' actions.

Based on this hypothesis, the following research questions will be addressed.

RQ 1: How can a model-based DRL agent use world models to generate goal-directed trajectories that explain its decision-making process to developers? A trained policy of an DRL agent should be able to solve the task it was trained for. This implies that a series of actions will eventually end up in a goal state, S_G. If the agent is trained well enough, the policy finds for each state S an action A that will eventually end up in a goal state. For example, an autonomous car that is heading towards a crossing. The task is to turn right, and the goal S_G is reached when the car has turned right without breaking a rule or causing an accident. This research question addresses this problem by providing the goal state for each action as it is predicted by the agent. If the agent is able to provide a valid goal state, the agent is working towards the desired goal. This would help to validate that the agent has learned the

intended goal. In many cases, relying solely on the goal state may not suffice to evaluate a policy, particularly in lengthy trajectories or complex environments. Humans prefer goals as explanations for successful or expected actions, while they prefer beliefs (planned actions) for failures or unexpected outcomes [8].

For that, a model-based DRL agent should be developed that learns a model of the environment, which it then uses to predict future states and actions. It also has to decide whether a predicted state has reached the goal. The predicted goal state can then be used as an explanation to a human developer that can evaluate if the agent is learning the intended task. The goal-directed trajectory resulting from the goal state prediction can be used to extend the explanation if needed. By providing the full trajectory the developer can evaluate both the targeted goal and the plan to reach it.

Hafner et al. [7] presented Dreamer, a model-based DRL-algorithm that uses a world model to train the actor. This might be a suitable base that can be extended to create an explainable model-based DRL agent.

RQ 2: How can introspective explanations of goal-directed trajectories help experts and developers to evaluate the performance of a World Model-based Deep Reinforcement Learning (WMBDRL) agent? In environments with complex tasks, the training process of DRL agents can be difficult. The reward can be used to evaluate, but one major difficulty is the credit assignment problem. Debugging and evaluating such agents can present considerable challenges, as it is often unclear whether the agent has genuinely learned the intended task or if it found the correct solution through trial and error. Additionally, assessing the agent's assumptions about the world can be difficult. Particularly for agents that act instantaneously without planning ahead, distinguishing between following a long-term plan and merely reacting until reaching a state where it knows how to proceed or complete the task poses a significant challenge. This challenge increases if there is only a sparse or delayed reward, in stochastic environments or with task that need many steps to solve it. Introspective explanations could offer valuable insights into differentiating whether an agent is capable of acting goal oriented or is simply engaging in random actions, even for agents that act instantaneously without prior planning.

4. Methodology and Rationale: Investigating the Research Hypothesis

The research approach combines experimental and empirical methods, divided into a technical and a social dimension. In the technical dimension, an explanation algorithm will be developed using an experimental research method to determine whether providing goal and plan-based explanations is possible. The social dimension evaluates the effectiveness of this type of explanation for developers, analyzing the performance of agents in the development process.

Recent reviews provide a comprehensive overview of XAI, XDRL, and DRL [3, 11]. We are conducting a systematic literature review on world models in DRL, focusing on papers published after 2018. The review focuses on identifying different types of world models. The review will identify the different strength and weaknesses of the different approaches.

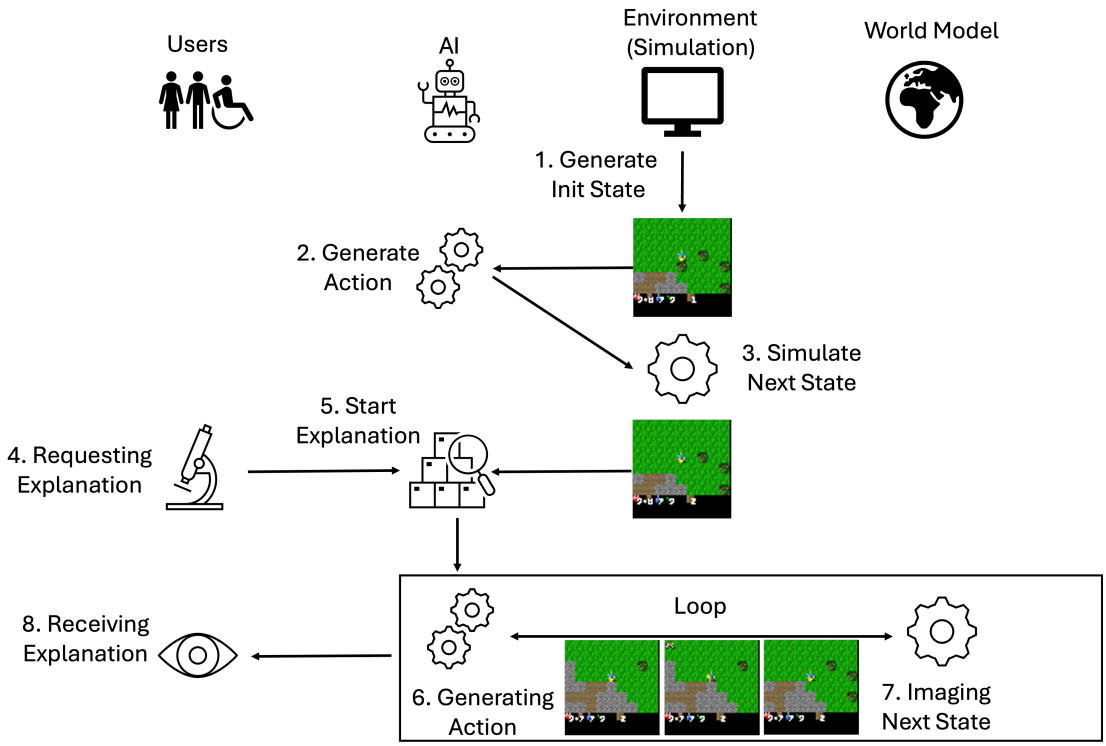

Figure 1: Illustration of the Introspective Explanation Process. The figure depicts a sequential process wherein the simulation generates the initial state. Subsequently, the AI selects an action, and the environment simulates the resulting next state. Upon user request for an explanation of the last action, the AI engages with the world model to construct an imagined future trajectory, which is then presented to the user.

Existing WMBDRL algorithms offer components for introspective explanations, which will be integrated into a new framework. After the development of the explanation algorithm, a technical analysis evaluates their explanatory performances based on the CO-12 properties introduced in [12]. Followed by a user study with DRL-experts and developers assessing the effectiveness of introspective explanations for DRL-agent performance analysis. The hypothesis is that introspective explanations for WMBDRL can support experts in evaluating the agents performance by providing the agents plan and by making transparent which assumption the agent has regarding the effects of his actions on the environment.

5. Preliminary Findings and Current Contributions

Explainability approaches often overlook insights from social studies and psychology, prioritizing technical feasibility [13]. Addressing explainability challenges requires prioritizing the human perspective. Inspired by how humans naturally explain actions, psychological research suggests that actions are goal-directed [9]. By understanding how actions contribute to goals, humans comprehend them better. This perspective can be applied to explain AI actions, making

internal goals, plans, and assumptions understandable.

Introspective explanations use the AI's imagination to predict changes in its environment as the interaction progresses. Only internal information and components are used for this. Without accessing external data, the AI demonstrates how it would respond in various scenarios and reveals its underlying assumptions. For an AI that is embedded in a larger system, such explanations could be integrated in larger system-explanations [14].

To create a DRL agent that is able to create an introspective explanation, it has to be able to imagine the future and to predict how its actions will impact the environment. World models could be a potential solution to this challenge. World models are frequently employed to speed up the training processes [7] of DRL. Many of these functionalities align with the requirements for introspective explanations. However, world models have not been used for this purpose. Simulations could represent a potential alternative. These are already frequently used for training AI models. However, simulations are not always available and, depending on their complexity, they require more resources and time.

To evaluate the new explanation type, we developed a prototype that is able to create introspective explanations. For that, we trained a DreamerV3 agent [7] and use the policy and world model during inference to generate visual explanations. To generate an explanation, the explainer receives one or more past steps, the world model predicts the subsequent state. Based on the state prediction, the policy predicts an action. Following each step, the reward is approximated, and a prediction is made regarding whether the new state is a goal state. Figure 1 shows this iterative process, that can continue indefinitely or until the goal state is reached. Particularly in the early stages of training, the explainer may struggle to envision a viable path towards the goal.

Especially in stochastic or non-deterministic environments, the prediction will likely differ from the true future trajectory. Since imagination is independent of the real environment, such explanations could show whether the agent has learned causal relationships that would lead to the solution of the problem or the fulfillment of the task, thus reducing the stochastic dependence on the environment. In experiments this hypothesis will be further investigated.

The technical possibility of generating visual introspective explanations was ensured by an initial prototype. However, further technical and social evaluations are needed. Potential limitations or challenges include the low resolution of the imagined images (64x64 and lacking sharpness). Our initial experiments suggest that the lack of a persistent imaginary world is a major weakness of current world models. For example, if the agent follows a circular path in its imagination, the imagined world does not persist. This could make it more difficult to solve tasks in which localization plays a decisive role and hinder the reaching of goals. A relevant use case would be solving a maze or route planning in an autonomous vehicle. This problem results from the world models used and is not a result of the explanation. The development of persistent world models would solve this problem.

6. Future Research Directions and Expected Contributions

World models and the domain of WMBDRL has experienced a lot of research interest in the recent past. Currently, there is no review paper that is collecting the different approaches and

applications. To address this, we are currently working on a review on different applications and approaches of WMBDRL.

We've developed the first prototype of the introspective explanation model by extending a DreamerV3 agent. The next step is a user study to evaluate the effectiveness of introspective visual explanations for experts in deep reinforcement learning. Participants of this study are required to have knowledge and experience in training deep reinforcement learning.

In the study, participants will evaluate the AI's ability to solve a task in the future while viewing a sequence of the AI's performance. Data will be collected through Likert scale ratings, open-ended responses, pre- and post-questionnaires. Finally, a statistical analysis will be conducted to compare the participants' analysis performance with and without our explainability technique. The user study will be submitted to the ethics committee and undergo a data protection review prior to being conducted.

We expect that this project will generate a new form of explanation that is able to make the decision process of DRL agents understandable even in stochastic or non-deterministic environments. In addition, findings on the effectiveness of this type of explanation for developers are expected from the user study.

We plan to conduct a second study employing a similar methodology, aiming to assess whether the explanation technique enhances end-users' understanding of AI actions. In this study, participants will be presented with sequences of the AI playing a game, both with and without the additional explanation. Subsequently, participants will be asked to predict the AI's next actions in problem-solving scenarios and evaluate their confidence in assessing the AI's behavior.

We are planning a second study using a similar methodology to determine whether the explanation technique improves end users' understanding of AI actions. Building on this, it could be investigated whether these explanations lead to a false sense of confidence, but this investigation is not planned as part of this project.

Our first prototype is based on a DreamerV3. However, other algorithms are possible in addition to Dreamer. It can be investigated whether other models can be used to generate this type of explanation. Furthermore, it can be investigated how other existing XAI methods can be applied to the imagination and whether this can improve the explanation.

Acknowledgments

This research was supported by the Innovation Campus for Future Mobility (www.icm-bw.de), by the Helmholtz Association within the Core Informatics project. This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research and by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

References

[1] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, K. Baum, What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective

on xai and a conceptual model guiding interdisciplinary xai research, Artificial Intelligence 296 (2021) 103473. doi:10.1016/j.artint.2021.103473.

[2] R. Dazeley, P. Vamplew, F. Cruz, Explainable reinforcement learning for broad-xai: a conceptual framework and survey, Neural Computing and Applications 35 (2023) 16893–16916. doi:10.1007/s00521-023-08423-1.

[3] G. A. Vouros, Explainable deep reinforcement learning: state of the art and challenges, ACM Computing Surveys 55 (2022) 1–39. doi:10.1145/3527448.

[4] A. Bertrand, R. Belloum, J. R. Eagan, W. Maxwell, How cognitive biases affect xai-assisted decision-making: A systematic review, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, AIES '22, ACM, 2022, pp. 78–91. doi:10.1145/3514094.3534164.

[5] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[6] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al., Inner monologue: Embodied reasoning through planning with language models, arXiv preprint arXiv:2207.05608 (2022).

[7] D. Hafner, J. Pasukonis, J. Ba, T. Lillicrap, Mastering diverse domains through world models, arXiv preprint arXiv:2301.04104 (2023). doi:https://doi.org/10.48550/arXiv.2301.04104.

[8] B. F. Malle, How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction, The MIT Press, 2004. doi:10.7551/mitpress/3586.001.0001.

[9] G. Csibra, G. Gergely, 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans, Acta Psychologica 124 (2007) 60–78. doi:10.1016/j.actpsy.2006.09.007.

[10] H. Yau, C. Russell, S. Hadfield, What did you think would happen? explaining agent behaviour through intended outcomes, Advances in Neural Information Processing Systems 33 (2020) 18375–18386. doi:10.5555/3495724.3497267.

[11] T. Speith, A review of taxonomies of explainable artificial intelligence (xai) methods, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, ACM, 2022, pp. 2239–2250. doi:10.1145/3531146.3534639.

[12] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023) 1–42. doi:10.1145/3583558.

[13] T. Miller, P. Howe, L. Sonenberg, Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences, arXiv preprint arXiv:1712.00547 (2017).

[14] M. Schwammberger, R. Mirandola, N. Wenninghoff, Explainability Engineering Challenges: Connecting Explainability Levels to Runtime Explainability, Springer Nature Switzerland, 2024, pp. 205–218. doi:10.1007/978-3-031-63803-9_11.