Explainable Reinforcement Learning Based on Q-Value Decomposition by Expected State Transitions

Yuta Tsuchiya[1,*], Yasuhide Mori[1] and Masashi Egi[1]

[1]Hitachi, Ltd. 1-280, Higashi-koigakubo, Kokubunzi, Tokyo 185-8601, Japan

Abstract

Because of the high control and planning performance, Reinforcement Learning (RL) is increasingly being adopted in a wide range of applications, including socially responsible tasks. This has led to a rapid advancement in the research of Explainable RL, and one commonly-used approach is past-oriented explanation; highlighting the factors in the current state that drive an agent to take an action. However, the agent is trained in anticipation of future rewards, hence, it is crucial to provide future-oriented explanations that demonstrate what the agent expects the action to achieve. To address the challenges of computational complexity and comprehensiveness of the explanation, we propose a novel method for interpreting the intention behind the agent's actions by introducing state-wise critics, which are additional neural networks that estimate the Q-value for each set of state transitions defined by users. By incorporating the RL model as the target network within the critics' loss function, the resulting explanation is aligned with the Q-value estimated by the agent. To evaluate the validity of the critics, we conducted an experiment of resource pre-allocation to measure the power outage risks. The results indicate that a relatively small number of critics are sufficient even with many combinations of states.

Keywords

Explainable AI, Future-oriented explanation, Interpretability, Q-value, Reinforcement learning, State transition

1. Introduction

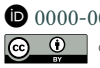
Reinforcement learning (RL) is a powerful framework for learning an optimal policy of an agent so that it can output appropriate actions which lead to high reward under the given environment. Deep reinforcement learning (DRL) has emerged as a result of recent machine learning advances, which expressing agent policies in neural networks to efficiently address problems with large numbers of states and complex probabilistic structures [1]. High performance in control and planning of DRL has been demonstrated in the realm of games and simulations; Alpha Go [2] defeated a world champion in the game of Go, and Agent57 [3] outperforms the average human in playing Atari arcade games. Consequently, there is an expectation that the applications of RL will be expanded to tasks on social infrastructure and medical domains [4].

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), Hyatt Regency, San Francisco Airport, California, USA, March 27-29, 2023.

*Corresponding author.

yuta.tsuchiya.gf@hitachi.com (Y. Tsuchiya); yasuhide.mori.yb@hitachi.com (Y. Mori); masashi.egi.zj@hitachi.com (M. Egi)

0000-0002-4570-957X (Y. Tsuchiya)



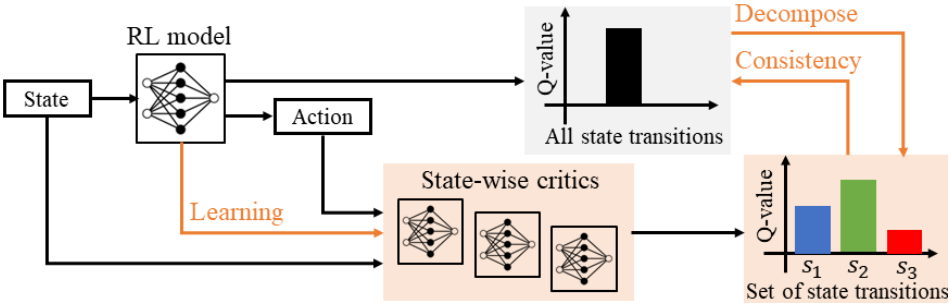


Figure 1: The architecture of our proposed model. Original RL model estimates the expected Q-value for all state transitions. In contrast, the critics decompose the Q-value by the set of state transitions.

In order to apply RL to mission-critical operations, it is required to meet various properties; accountability, fairness, and interpretability. This has led to a rapid advancement in the research of Explainable AI (XAI), which is a technology to explain the reason of decisions made by machine learning models [5]. One commonly-used approach in supervised learning is the visualization of image features through heat maps to understand where the model is focusing on [6]. Similar explanation techniques are being adopted in the field of Explainable Reinforcement Learning (XRL); highlighting the factors in the current state that drive an RL agent to take an action [7]. However, it should be noted that the agent's actions are learned in anticipation of future rewards and desired events, hence, it is crucial to provide "future-oriented explanations" that demonstrate what the agent expects the action to achieve, as opposed to traditional "past-oriented explanations" based on current states.

As future-oriented explanation method, several studies have discussed to explain the expected reward (represented by Q-values) in the future state. For example, Van et al. have proposed forward simulations, which show the state transitions with the highest probability until the end of the episode, and the obtained rewards are assumed as the intention of the RL model [8]. However, this approach may be insufficient as an explanation since it only provides a single state transition; it may not align with the interests of users, such as future states with low probability yet high rewards, or states that are at high risk due to the actions. Another approach, value function decomposition, is the method to train an additional neural network to output a table of expected rewards for all possible future state and action pairs [9]. While it can exhaustively describe desired outcomes by the action, the additional model learning may not be successful under complex problems with many combinations of state transitions.

In this paper, we propose a novel future-oriented explanation method for interpreting the intention behind an agent's actions by introducing state-wise critics as shown in Fig.1. The critics are additional neural networks that estimate the Q-value for each set of state transitions defined by users. We aim to simultaneously address the challenges of computational complexity and comprehensiveness by decomposing the Q-value of the RL model based on semantically meaningful state types, rather than focusing on a single episode or every state transition, as is done in existing methods. The resulting explanation is aligned with the agent's Q-value by incorporating the RL model as the target network within the critics' loss function. Furthermore, our method can provide contrastive explanations that reveal why the agent thought a certain

action was more effective, as opposed to alternative actions proposed by users.

To evaluate the validity of the derived explanations, we conducted an experiment utilizing a simple planning problem of resource pre-allocation optimization against the disaster risk of power outages. As the RL model, we employed the Deep Deterministic Policy Gradient (DDPG) algorithm, a basic Actor-Critic RL model for continuous action spaces [10]. The results suggest that the critics can provide consistent and valuable explanations by visualizing the trade-offs in Q-values for specific state transitions, and a relatively small number of critics are sufficient to offer users meaningful insights into the intentions behind agent's actions, even for complex problems with numerous state combinations.

In summary our contributions are as follows:

- Provide an explanation for the intention behind actions of the agent by decomposing the Q-value for each meaningful state type for users. This approach does not introduce any changes to the original RL model, while effectively addressing issues of computational complexity and enhancing interpretability.
- Implement the case studies to show the validity of the decomposed Q-values and the usefulness of a contrastive explanation with user alternatives.
- Bridge the gap between human understanding and the intentions of RL models, and serves as a fundamental technology for the development of hybrid intelligent systems.

The rest of this paper is organized as follows. We first discuss the related works to show the relevance of our approach to other XRL techniques. We then describe the basis mechanism of reinforcement learning and the problem of interpretability. After that, the specific procedure for introducing state-wise critics is described. Then, through numerical experiments, we show that our proposed method generates plausible explanations and is useful for interpreting RL's behavioral intentions. Finally, we draw our main conclusions and outline future work.

2. Related Work

XAI techniques for machine learning models have been developed rapidly after the launch of Defense Advanced Research Projects Agency (DARPA) project [5]. While a variety of explanation techniques have been proposed in the field of supervised learning, XRL techniques are gradually developing along with the growing expectations for the expansion of RL applications [11]. XAI methods can be roughly categorized based on two scopes; global and local explanation. Global methods explain the entire, general model behaviour, while local methods focus on a specific decision.

Generating interpretable white-box models is common global XRL approach that is also used in supervised learning [12, 13]. Programmatically Interpretable Reinforcement Learning (PIRL) is a framework inspired by imitation learning to mimic complicated RL's policies by a human-readable programming language [12]. The average performances of the regenerated policies are not as high than the original one. However, in experiments with the The Open Racing Car Simulator (TORCS) dataset, the generated model was less perturbed by noise and blocked sensors. Liu et al. [13] have proposed a framework that approximates the RL policy by Linear Model U-Trees (LMUTs), which have a linear model at each leaf node to improve

its generalization ability. LMUTs are also easier to understand since it generally utilize fewer leaves than ordinal Continuous U-Trees.

HIGHLIGHTS is an RL-specific global explanation technique: an algorithm that produces a summary of an agent's behavior by extracting important trajectories from simulations of the agent. Specifically, the states extracted in HIGHLIGHTS are where taking a wrong action can lead to a significant decrease in future rewards according to the agent's Q-values. They conducted a human-subject experiment and participants were more successful at evaluating the performance of agents when presented with HIGHLIGHTS summaries compared to baseline summaries.

Local methods for explaining individual decisions are the subject of this study. We classify this method into past-oriented and future-oriented explanations. Past-oriented explanations generally show the importance of the input state [7]. LMUTs can also be used to analyze the importance of input features and extract rules, e.g., the 'super-pixels' in image inputs are highlighted to illustrate the key regions [13].

Causal Lens focuses on providing future-oriented explanations for an agent's behaviour based on the knowledge of how its actions influence the environment [14]. It can find the differences between the actual action (e.g. 'Why action A?') and the counterfactual (e.g. 'Why not action B?') through a structural causal model (SCM). However, it is challenging to provide an accurate SCM for complex environments. Juozapaitis et al. [15] proposed a decomposition of the scalar reward function into a vector-valued reward function for each sub-reward type. This approach provides a compact explanation of why a particular action is preferred over another from the viewpoint of which sub-reward is considered as important. Our method differs in that the expected reward is decomposed by future state types. Also, while this method modifies the reinforcement learning model itself, our method provides explanations in a completely post-hoc manner.

3. Issues on Interpretability of Reinforcement Learning

3.1. The Basic Formulation of Reinforcement Learning

We first formalize a reinforcement learning environment as a Markov decision process (MDP), $M = \langle \mathcal{S}, \mathcal{A}, p, \gamma, r \rangle$. Here, $\mathcal{S}$ is the set of all environment states, $\mathcal{A}$ is the set of actions, $p(s_{t+1}|s_t, a_t)$ is the state transition function, $\gamma \in [0, 1]$ is the discount factor, and $r(s_t, a_t, s_{t+1}) \in \mathbb{R}$ is the function that calculates the reward R_t for executing action a_t in state s_t arriving at next state s_{t+1}. In the MDP framework, an agent observes the state s_t in the environment and selects an action a_t for s_t. From the result of state transition derived by the action and environment, the agent obtains a reward R_t and moves to the next state s_{t+1}. In order to obtain the maximum cumulative reward, the agent seeks to find the optimal policy $\pi^* : S \to A$ that keeps choosing the action with the highest Q-value; $Q(s_t, a_t)$ is the following evaluation value of the action a_t in the state s_t:

$$Q(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim p(s_{t+1}|s_t, a_t)}[R_t + \gamma \max_{a_{t+1}} Q_{EX}(s_{t+1}, a_{t+1})], \tag{1}$$

where Q_{EX} represents the Q-value at the next state.

Since recording Q-values for all states and actions poses a formidable challenge, deep reinforcement learning that approximates the table of Q-values with a deep neural network (DNN) was proposed and made great progress [1]. There are many kinds of architectures of DNN, for example, in the Deep Q-Network (DQN) model, the output layer of DNN corresponds to the Q-value of each action. Deep Deterministic Policy Gradient (DDPG) [10] stands as a prominent method for dealing with continuous-valued action spaces based on Actor-Critic [16]. DDPG employs two distinct neural networks: the critic aims to accurately predict the Q-value, while the actor learns to generate actions that optimize the estimated Q-value derived by the critic. This study imposes no restrictions on the selection of RL models, and the proposed framework can also be readily applied to any value-based RL algorithm.

3.2. The Interpretability of Agent's Action

Because of the high control and planning performance, RL is increasingly being adopted in a wide range of applications, including socially responsible tasks as well as gaming and simulation. However, the decision-making process of RL models is complex and difficult for humans to interpret, thus, the field of XRL has been received much attention and various types of explanation have been discussed [11]. In this study, we focus on local future-oriented explanations that provide insights into the anticipated states and rewards associated with individual actions. The agent is picking actions to maximize the cumulative expected reward, and therefore, explanation of the events that follow a specific action choice is vital to reveal the agent intention. Notably, psychological studies suggest that about 70% of everyday explanations are intention-based [17].

The key to comprehending the agent's behavioral intentions is the Q-value: it directly relates to the cumulative expected reward that the agent seeks to optimize. Nonetheless, the Q-value estimated by the ordinal RL model is the expected value of all next state transitions as (1), and therefore, it is challenging to interpret specific future states and rewards. Existing studies have discussed the utilization of the most probable future state transition [8] or a table for all possible states and actions [9]. However, the former approach of forward simulations may not provide sufficient information to explain the agent's intent comprehensively and not meet the diverse interests of the user. In addition, it may be difficult to simulate the state transition in a real environment. Although the latter approach, value function decomposition, allows for comprehensive explanations by the table, it may not be readily applicable to complex real-world problems involving a lot of combinations of states.

4. Future-Oriented Explanation by State-Wise Critics

4.1. Introduction of State-Wise Critics

This paper introduces a novel explanation method to reveal the expected states driving the actions by incorporating state-wise critics for each user-defined set of state transitions as shown in Fig.1. The critics are additional neural networks that estimate the Q-value under the specific state transitions by taking the state s_t and the agent's corresponding action a_t as input.

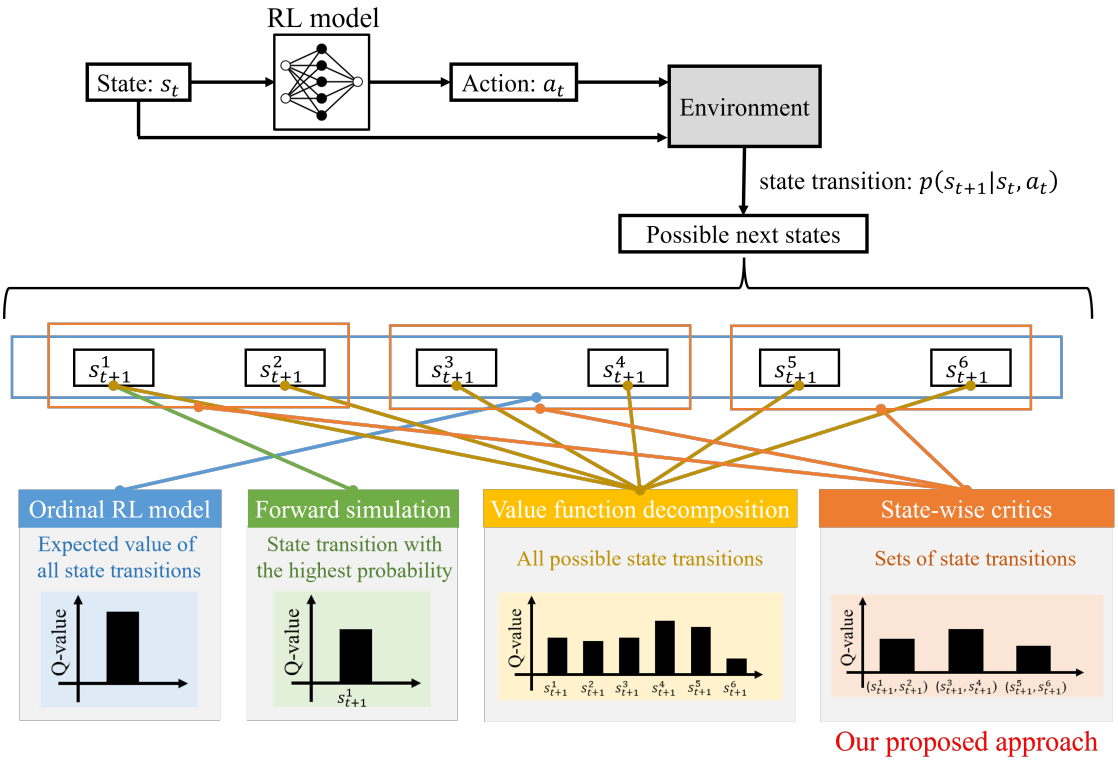


Figure 2: The comparison of each explanation methods regarding the focus of state transitions and obtained Q-values.

Figure 2 presents a comparison between the proposed and existing approaches regarding the focus of the subsequent states and obtained Q-values for an action a_t. Consider the six possible states at the next time: $(s^1_{t+1}, ..., s^6_{t+1})$ after the agent performs action a_t. The original RL model outputs the expected value for all state transitions as (1). Forward simulations present the most probable state (s^1_{t+1} in our example) as the agent's intended future state. However, this approach may overlook states with low probabilities yet high rewards. Value function decomposition enables the visualization of Q-values for each of all possible next states (strictly, including further time steps). However, this approach may not be suitable when states are represented by continuous values. Our approach addresses both computational complexity and interpretability issues by providing Q-values for each user-defined set of state transitions. As shown in Fig.2, the six possible states are grouped into three sets, i.e., one critic assigned to two states. This enables users to obtain explanations by the granularity they are interested in.

4.2. Training Procedure of Critics

The proposed critic is designed to estimate the Q-value calculated by the explanation-target RL model under a specific set of state transitions. In order to achieve the accurate decomposition and the consistency of the derived explanation, each critic is trained on historical data for

its corresponding state transitions only, and adding the Q-value of the RL model to the error function of each critic. The learning process can execute concurrently with or after the training of the RL model. Note that the computational time required for each critic training is directly proportional to the number of critics employed. Nevertheless, since each critic can be trained independently, parallel processing can be employed to expedite the overall learning process.

The suggested learning steps for critics are as follows:

1. Sample some historical data from training data memory of RL model (generally, the replay buffer). Each data consist of the state-action pairs and the corresponding next state and reward, i.e., $\langle s_t, a_t, s_{t+1}, R_t \rangle$. The appropriate sample size depends on the complexity of the problem. Note that the critics for state transitions with low probability are also trained enough.
2. Extract the corresponding state transition data for each critic from sampled data set. Various methods can be employed to select the data, such as using conditional expressions or pre-labeling the historical data.
3. Train critics to minimize the following temporal difference (TD) error:

$$E(s_t, a_t) = (R_t + \gamma \max_{a_{t+1}} Q_{tar}(s_{t+1}, a_{t+1}) - Q(s_t, a_t))^2, \tag{2}$$

   where Q_{tar} is the Q-value estimated by a target network, which is commonly introduced in training RL models for stabilization of the learning process [1].

In a commonly used approach, each critic typically employs its own target network, which is a neural network with the same structure as the critic. However, this approach causes the learning objective to deviate from the Q-value of the original RL model, as shown in Fig. 3. The critic estimates the Q-value Q_{pred} for a specific type of a state transition s^1_{t+1}, and aims to minimize the error between its estimation and the original RL model's prediction of Q_{RL} for s^1_{t+1}. However, since the critics are trained on a data set that contains only state transitions of s^1, the target network with copied parameters of the critic assumes that only the state transition of s^1 occurs at $t+2$ when the Q-value for the next action, Q_{tar_critic}, is estimated. In contrast, the target network based on the original RL model estimates the expected Q-value, Q_{tar_RL} for all state transitions at $t+2$, resulting in a discrepancy between these target networks. Therefore, to ensure consistency with the original model, we incorporate the RL model as the target network into the error function (2) of each critic.

Upon sufficient repetition of the aforementioned learning process, the error function of each critic converges, thereby reducing the deviation from the Q-value of the original RL model. When each critic corresponds to the independent state transitions and all states included by RL model are represented by the entire critics as shown in Fig. 2, the average of Q-values estimated by critics will converge to the Q-value of ordinal RL model $Q(s_t, a_t)$ as follows:

$$Q(s_t, a_t) \approx \mathbb{E}[Q^{(s^1_{t+1}, s^2_{t+1})}(s_t, a_t) + Q^{(s^3_{t+1}, s^4_{t+1})}(s_t, a_t) + Q^{(s^5_{t+1}, s^6_{t+1})}(s_t, a_t)]. \tag{3}$$

It should be noted that critics do not necessarily have to handle independent state transitions, and some states may be included in multiple sets. Even in such case, the predicted values of each critic can still correctly converge to the Q-value of the RL model under the specific set of state transitions.

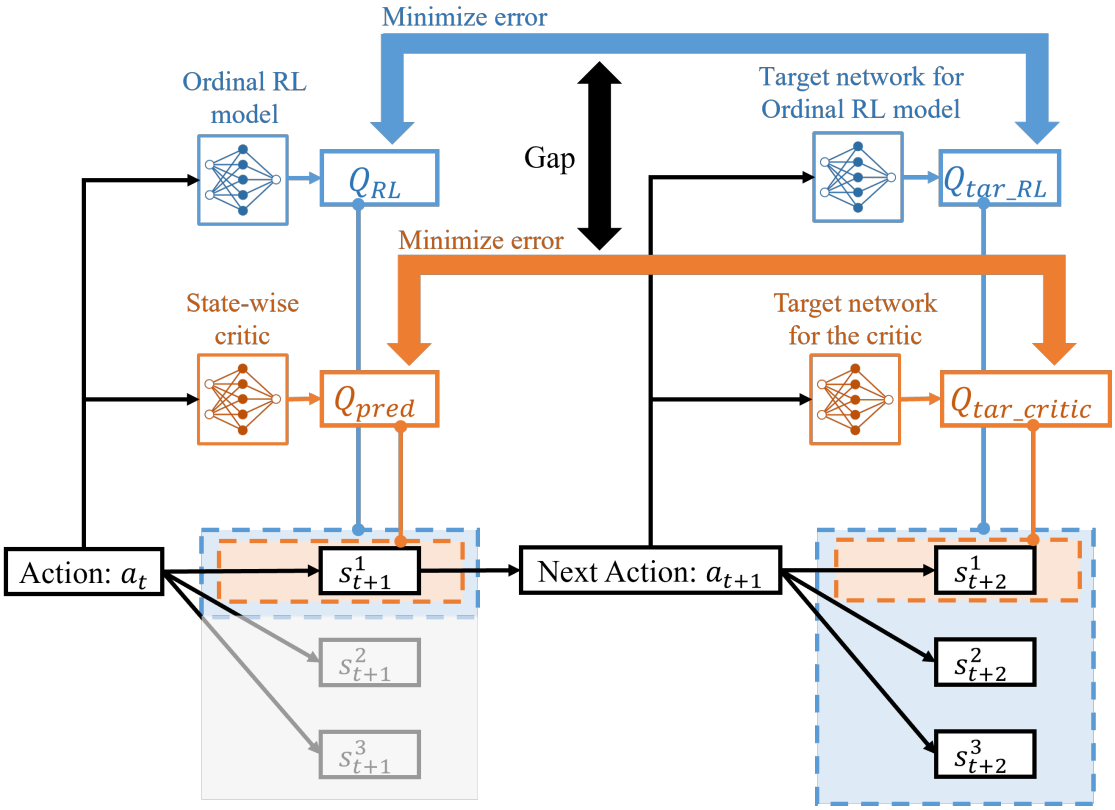


Figure 3: An example flow of the gap of error minimization between ordinal RL model and state-wise critic for s^1_{t+1}. Note that current state s_t and reward R_t are omitted for simplicity.

4.3. Explanation Procedure by Q-Value Decomposition

State-wise critics enable users to gain insights into the effectiveness and risks in specific states associated with the agent's actions. These critics are versatile and can address a variety of user concerns about the agent's behavior. For example, users can understand the most expected state transition by the critic with the largest Q-value. To contrast, the state with the lowest Q value indicates the high-risk future state. Furthermore, presenting additional information, such as probability, allows users to understand whether a state has "a relatively low reward and a high probability" or the opposite.

This method has another useful application in the form of contrastive explanation, which involves comparing the Q-value for the agent's actions with that derived by the user-suggested actions. Typically, a critic, that represents a set of state transitions with a substantial difference in Q-value between the action of the agent and users, can be extracted to comprehend why the agent deemed the action more appropriate than the user's.

The following are additional guidelines for defining critics. In cases where states are represented by continuous values, upper and lower bounds can be set to define the range of values that each critic is responsible for. It is important to note that critics do not need to be defined

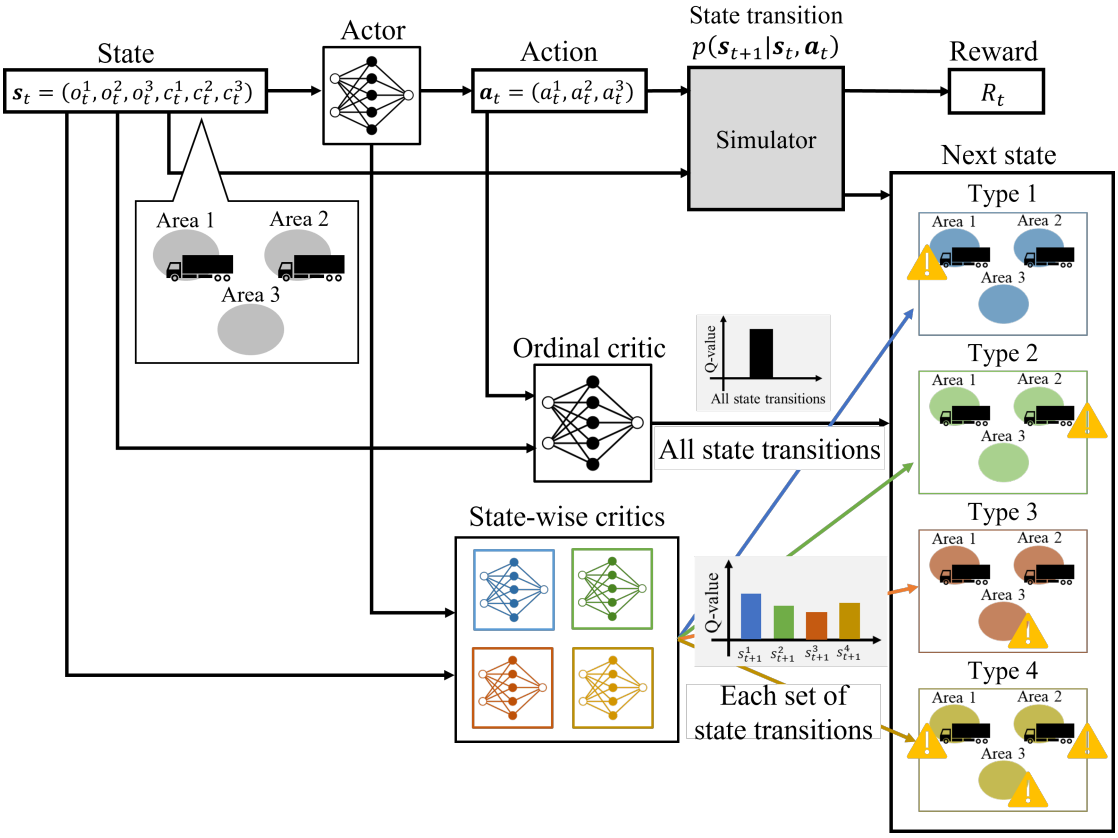


Figure 4: The schematic diagram of this experiment.

for each time step separately. State transitions that share similar characteristics across multiple time steps can be assigned to the same critic.

5. Numerical Experiment

5.1. Experimental Setup

In order to verify the validity of the generated explanations under the problem with a lot of state transitions, we conducted experiments on resource pre-allocation problems in power systems to mitigate disaster risks. Figure 4 shows a schematic diagram of this experiment. We employed DDPG, a RL model that can handle actions with continuous values [10]. "Actor" observes the state with the amount of power outage o_t and existing resources c_t in three areas $\boldsymbol{s}_t = (o_t^1, o_t^2, o_t^3, c_t^1, c_t^2, c_t^3)$, and specify resource allocation such as power-supply cars in each area by the action $\boldsymbol{a}_t = (a_t^1, a_t^2, a_t^3)$. Once the allocation is completed, the simulator executes one of the state transitions (accident cases) shown in Table 1. There are differences in the accident area, probability, and the magnitude of damage among the accident cases. Rewards will be given if resources are pre-positioned to cover the amount of damage in the affected area.

Table 1
All possible state transitions (accident cases) in the experiment.

	Area	Probability	Damage (MW)
Type 1	Area 1	0.2	100 to 200
Type 2	Area 2	0.3	50 to 120
Type 3	Area 3	0.1	50 to 120
Type 4	All areas	0.4	100 to 150

Table 2
Correspondence table of actions and the number of resources. From the second line, at least one variable is less than 0.

	Area 1 c_t^1	Area 2 c_t^2	Area 3 c_t^3
$\forall a_t > 0$	15	15	15
$a_t^1 > a_t^2 \geq a_t^3$	20	15	10
$a_t^1 > a_t^3 \geq a_t^2$	20	10	15
$a_t^2 > a_t^1 \geq a_t^3$	15	20	10
$a_t^2 > a_t^3 \geq a_t^1$	10	20	15
$a_t^3 > a_t^1 \geq a_t^2$	15	10	20
$a_t^3 > a_t^2 \geq a_t^1$	10	15	20
$a_t^1 = a_t^2 > a_t^3$	20	20	5
$a_t^1 = a_t^3 > a_t^2$	20	5	20
$a_t^2 = a_t^3 > a_t^1$	5	20	20

Actor's action is three-dimensional and represented by a tanh function that outputs values from -1 to 1. Each element of $\boldsymbol{a}_t$ represents the priority of whether to allocate resources to the corresponding area. Based on the relative values of each action shown in Table 2, the number of resources allocated to each area is determined, and the total number of resources with existing resources (allocated at the previous time) is obtained as $(c_{t+1}^1, c_{t+1}^2, c_{t+1}^3)$.

For state transitions, the accident type is first determined based on probability, and then the amount of damage is randomly selected within the range shown in Table 1. As a result, there are numerous possible combinations of damage $|\mathcal{O}|$ as follows:

$$|\mathcal{O}| = \underbrace{100}_{Case1} + \underbrace{70}_{Case2} + \underbrace{70}_{Case3} + + \underbrace{50 \times 50 \times 50}_{Case4} = 125240. \tag{4}$$

Note that previous time damage is also added cumulatively to $(o_{t+1}^1, o_{t+2}^2, o_{t+3}^3)$.

The reward R_t is calculated based on the amount of damage $(o_{t+1}^1, o_{t+1}^2, o_{t+1}^3)$ caused by the state transition and resource allocation $(c_{t+1}^1, c_{t+1}^2, c_{t+1}^3)$ through the action as follows:

For accident Type 1 to 3 (the reward is only calculated for the accident area):

$$R_t = \begin{cases} 1.0 & if\ 10 \times c_{t+1} \geq o_{t+1} \\ (10 \times c_{t+1} - o_{t+1})/100 & others. \end{cases} \tag{5}$$

For the case of Type 4 accidents (the reward is calculated in all areas):

$$R_t = \begin{cases} 0.35 & if\ 10 \times c_{t+1} \geq o_{t+1} \\ (10 \times c_{t+1} - o_{t+1})/200 & others. \end{cases} \tag{6}$$

DDPG's critic (ordinal critic) estimates the expected value of the Q-value for all state transitions: the amount of damage shown in Table 1 and the resource allocations. However, it is challenging to interpret which accident types Actor's resource allocation is most effective for. To address this issue, we introduce four critics for each accident type as shown in Fig. 4, which are trained to follow the ordinal critic. The reinforcement learning model was trained 1,500 times with episodes consisting of three time steps, and the critics were also trained at the same time.

Then, we define the validity of the explanation as follows:

1. The relationship between each action and its corresponding Q-value for each state is highly reasonable. For instance, placing many resources in Area 1 should result in a high Q-value derived by the state-wise critic responsible for Type 1 accidents.
2. The relative magnitude of Q-values for different actions is consistent between the ordinal model and the average output of all state-wise critics. This is crucial for assessing the accuracy of the contrastive explanation, which highlights the differences from the action suggested by users.

5.2. Experimental Result

First, we evaluated the Q-value of critics when Actor outputs resource allocation of $(c_1^1 = 15, c_1^2 = 15, c_1^3 = 15)$, which is derived by the action with the highest Q-value at $t = 0$. Figure 5 shows the expected Q-value for all states estimated by the ordinal critic and sets of state transitions obtained by state-wise critics. The learning of state-wise critics has converged enough since the difference between the Q-value of ordinal critics and average of state-wise critics is only 3.5%. In Fig. 5, The highest Q-value was associated with accident Type 3, where only Area 3 was damaged. Comparing the allocation of Actor and the damage of accident Type 3 from Table 1, allocating 15 power supply cars in area 3 can fully prevent Type 3 accidents. Similarly, accidents of Type 2 and Type 4 are prevented completely by allocating 15 resources, and therefore, the corresponding critic's Q-values are also large. Conversely, for Type 1 accidents, the Q-value is small since there is a possibility of accidents that cannot be handled with only 15 resources (damage from 151 to 200 (MW)). By analysis based on these comparisons with Table 1, it can be concluded that the condition of the explanation validity 1 is satisfied.

Next, we evaluated the usefulness of contrastive explanation when the user provides an alternative action into critics at $t = 0$. Assuming that users are interested in understanding why Actor prioritized the action shown in Fig. 5 rather than the user-suggested allocations in Fig. 6(a) and (b): $(c_1^1 = 20, c_1^2 = 15, c_1^3 = 10)$ and $(c_1^1 = 20, c_1^2 = 10, c_1^3 = 15)$, respectively. The comparison of the expected Q-values by the ordinal critic and the average of state-wise critics in Fig. 6(a) and (b) demonstrate that the proposed critics can provide a consistent explanation that satisfies validity 2. Then, the agent concluded that the action shown in Fig. 5 is more effective

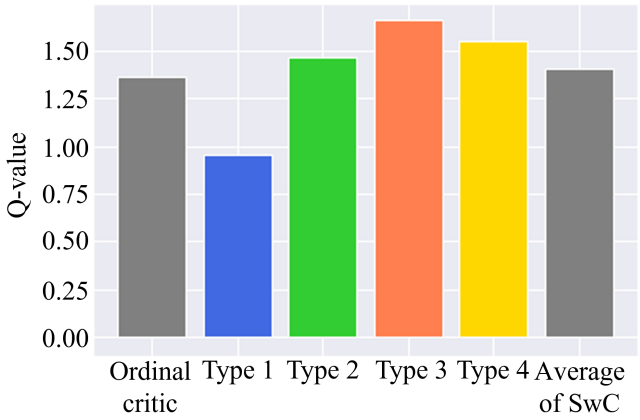
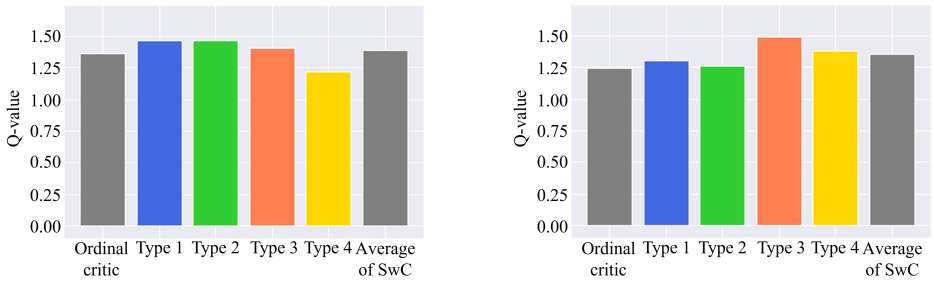


Figure 5: Q-values of Actor's resource allocation $(c_1^1 = 15, c_1^2 = 15, c_1^3 = 15)$. The Q-value of ordinal critics was 1.3614, while the average of state-wise critics (SwC) was 1.4098.



(a) The allocation $(c_1^1 = 20, c_1^2 = 15, c_1^3 = 10)$. **(b)** The allocation $(c_1^1 = 20, c_1^2 = 10, c_1^3 = 15)$.

Figure 6: Q-values by user-suggested resource allocations. The ordinal critic yielded average Q-values of 1.3610 and 1.2453 for Fig. 6(a) and (b), respectively, while the state-wise critics produced averages of 1.3890 and 1.3562, respectively.

in terms of expected performance than those in Fig. 6. However, it is difficult for the ordinary critic alone to analyze the trade-offs between the different actions.

Based on the analysis of state-wise critics, the user-suggested allocations in Fig. 6(a) and (b) have large Q-values for Type 1 accidents. The result indicates that the users' actions can prevent Type 1 accident completely by placing 20 resources in Area 1. On the other hand, they are weak against cases where Area 2 or 3 suffer extensive damage. For the action in Fig. 6(a), Area 3 lacks measures to address accidents of Type 3 (101 to 120 (MW)) and Type 4 (101 to 150 (MW)), and Fig. 6(b) has a similar trend for Area 2. As a result, the Q-values associated with those accident types are smaller than Agent's action. Therefore, the agent concluded that the negative effects of these actions are significant, and the action in Fig. 5 is more effective in terms of expected

Table 3
The 5-time average of the computation time required to train each method.

	Conventional (w/o critics)	Proposed (w/ critics)
Calculation time (sec)	19.5	107.1

performance. Overall, users could compare the trade-off relationship between Agent's assumed actions and their own actions based on specific state transitions.

To evaluate the impact of state-wise critics on computation time, we present the five-time averages of learning times for 1,500 episodes in Table 3. The proposed method trains both critics and the RL model simultaneously, whereas the conventional method only trains the RL model. The experiments were conducted on an Intel Core i7-1065G7 CPU @1.30 GHz with 16 GB RAM running Windows10 64-bit. As shown in Table 3, the learning time of the proposed method increased by approximately five times than the conventional. This is due to the addition of four critics, with each critic taking 20 seconds to train, similar to the RL model. Moreover, the sampling time of the data is also provided. From the above results, we confirmed that the calculation time increases linearly with the number of each critic, as described in Section 4.2.

Finally, we evaluated the relationship between the number of learning episodes and the optimal action. Fig. 7 shows the learning curve of the RL model, which seems to converge after around 500 iterations; however, the action $(c_1^1 = 15, c_1^2 = 15, c_1^3 = 15)$ was not optimal. To investigate the cause using critics, Fig. 8(a), (b), (c) show the Q-value of the five-time average and standard deviation as error bar at 500, 1,000, and 1,500 episodes, respectively. Comparing the Q-values, there is no substantial difference between the ordinal critic and the average of state-wise critics. Thus, even if the learning is inexperienced, the consistency with the RL model is maintained robustly. Next, Agent evaluated the Q-value of Type 1 to be significantly low at 500 episodes; thus, $(15, 15, 15)$ action was not optimal in the early stage because complete prevention from Type 1 accident was crucial for Agent. However, the error bars for Type 3 accidents are larger at 500 and 1,000, and for Type 4 at 1,000 episodes. These results indicate that the evaluation of $(15, 15, 15)$ action is fluctuated as the learning progresses due to the occurrence of accidents with low transition probabilities. After the 1,500 episodes, as shown in Fig. 8 (c), the evaluation of Types 3 and 4 increased significantly, and Agent concluded that $(15, 15, 15)$ action was optimal. Thus, we were able to analyze the evaluation of the action based on specific state transitions using the critics.

6. Conclusion and Future Work

In this paper, we introduced a method for decomposing the Q-value of each state transition to interpret the agent's action intentions concerning future states. By incorporating users' interest, we can obtain sufficient explanations, even for complex problems. Additionally, the estimated Q-values align with the original by utilizing the RL model as the target network within the critics' loss function. We evaluated the efficacy of the generated explanations in resource pre-allocation problems with numerous state transitions, and the proposed critics were able to provide users with insights into Agent's actions from the perspective of trade-offs in

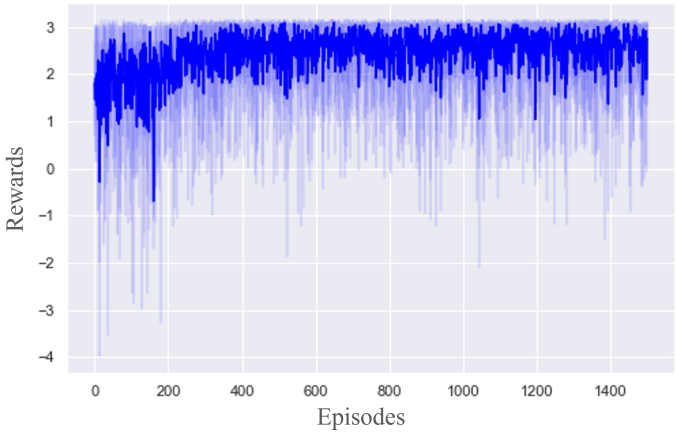
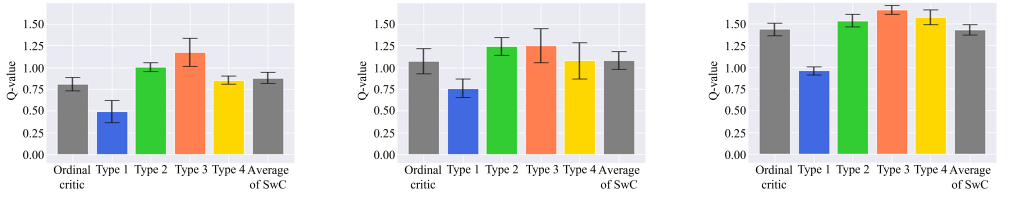


Figure 7: The learning curve of the RL model. The strong blue line shows the average curve of 5 runs.



(a) Q-values at 500 episodes. **(b)** Q-values at 1,000 episodes. **(c)** Q-values at 1,500 episodes.

Figure 8: Q-values at each learning episode.

Q-values for concrete state transitions.

Our method decomposes the Q-value for the state at the next time step, and it may not provide useful information for actions that expect states beyond a single time step. One possible solution to handle such cases is to visualize the Q-value of multiple steps by applying our method repeatedly while making assumptions about future state transitions. However, this approach leads to an increase in computational cost and needs to be addressed in future studies. It is also crucial for the method's wide applicability to determine how to efficiently train each critic independently after the RL model. Furthermore, obtaining feedback from real users on the generated explanations and conducting more quantitative evaluations are crucial for future research.

References

[1] V. Mnih, et al., Playing atari with deep reinforcement learning (2013). arXiv:1312.5602.

[2] D. Silver, et al., Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484–489. doi:10.1038/nature16961.

[3] A. P. Badia, et al., Agent57: Outperforming the Atari human benchmark, in: Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 507–517.
[4] M. M. Hosseini, M. Parvania, Resilient operation of distribution grids using deep reinforcement learning, IEEE Transactions on Industrial Informatics 18 (2022) 2100–2109.
[5] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (xai) program, AI Magazine 40 (2019) 44–58. doi:10.1609/aimag.v40i2.2850.
[6] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.
[7] S. Greydanus, A. Koul, J. Dodge, A. Fern, Visualizing and understanding Atari agents, in: Proceedings of the 35th International Conference on Machine Learning, volume 80, PMLR, 2018, pp. 1792–1801.
[8] J. van der Waa, J. van Diggelen, K. van den Bosch, M. A. Neerincx, Contrastive explanations for reinforcement learning in terms of expected consequences (2018). arXiv:1807.08706.
[9] H. Yau, C. Russell, S. Hadfield, What did you think would happen? explaining agent behaviour through intended outcomes, in: Workshop on Extending Explainable AI Beyond Deep Models and Classifiers, ICML, Vienna, Austria, 2020.
[10] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning (2015). arXiv:1509.02971.
[11] E. Puiutta, E. M. S. P. Veith, Explainable reinforcement learning: A survey (2020). arXiv:2005.06247.
[12] A. Verma, V. Murali, R. Singh, P. Kohli, S. Chaudhuri, Programmatically interpretable reinforcement learning, in: Proceedings of the 35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 5045–5054.
[13] G. Liu, O. Schulte, W. Zhu, Q. Li, Toward Interpretable Deep Reinforcement Learning with Linear Model U-Trees: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part II, 2019, pp. 414–429. doi:10.1007/978-3-030-10928-8_25.
[14] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, Explainable reinforcement learning through a causal lens, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 2493–2500.
[15] Z. Juozapaitis, A. Koul, A. Fern, M. Erwig, F. Doshi-Velez, Explainable reinforcement learning via reward decomposition, in: in proceedings at the International Joint Conference on Artificial Intelligence. A Workshop on Explainable Artificial Intelligence., 2019.
[16] V. Konda, J. Tsitsiklis, Actor-critic algorithms, in: S. Solla, T. Leen, K. Müller (Eds.), Advances in Neural Information Processing Systems, volume 12, MIT Press, 1999.
[17] B. F. Malle, Folk explanations of intentional action, in: B. Malle, L. J. Moses, D. Baldwin (Eds.), Intentions and Intentionality: Foundations of Social Cognition, MIT Press, 2001, pp. 265–286.