Self-Explaining Reinforcement Learning for Mobile Network Resource Allocation

Anonymous Authors[1]

Abstract

Deep reinforcement learning (DRL) methods, though powerful, often lack transparency, which limits their adoption in critical domains. We apply Self-Explaining Neural Networks (SENNs) to RL by parametrizing the policy of a PPO agent with a SENN, producing intrinsic local explanations, and propose a method for aggregating them into global explanations. We evaluate our approach on a mobile network resource allocation problem, our approach performs within a small margin of the state-of-the-art deep learning method and significantly outperforms the best deployed heuristic, while the extracted global explanations correlate strongly with DeepLift and InputXGradient, making SENNs a promising candidate for high-stakes RL.

1. Introduction

Although powerful, Artificial Intelligence (AI) models often operate as black-boxes, making it difficult to interpret their decisions, leading to a lack of trust among stakeholders and consequently hindering their applicability. This lack of transparency of black-box models, such as Deep Neural Networks (DNNs), limits their applicability in high-stakes domains (Linardatos et al.). This limited applicability facilitates rapid growth in the interest of Explainable Artificial Intelligence (XAI). XAI is divided into post-hoc interpretability methods and intrinsic interpretability methods. Post-hoc interpretability methods attempt to explain trained black-box models, by analysing relations between model's inputs and outputs. For instance, LIME or SHAP, are post-hoc perturbation-based methods that compute feature attributions representing the impact of each feature on the output of the model (Watson, 2022; Lundberg & Lee, 2017; Ribeiro et al., 2016). On the other hand, intrinsic interpretability means that the AI method is interpretable by design, for instance, linear regression, where the contribution of each feature towards the final decision is expressed by the corresponding coefficient. Moreover, intrinsic interpretability methods are considered more robust as they provide explanations derived from the model's internal mechanism (Rudin, 2019). XAI methods can be further divided into local and global methods. Local providing explanations to individual predictions and global that provide explanations of the whole model. Extending the prior example, linear regression models provide global intrinsic explanations.

The intuitions that ground XAI also extend to Reinforcement Learning (RL) through Explainable Reinforcement Learning (XRL). To date, many XRL specific techniques have been proposed, including temporal policy decomposition and hierarchical skill acquisition (Ruggeri et al., b; Shu et al.). However, most state-of-the-art RL algorithms are model-free and rely on DNNs for policy learning and modelling; because DNNs lack transparency, these algorithms inherit the same opacity (Qing et al.).

This work focuses on improving the explainability of the used models within the RL formulation, improving the explainability of the whole system and introduces a new approach for XRL. We apply Self-Explaining Neural Networks (SENNs) to RL networks that match DNN expressiveness, while offering adjustable interpretability (Alvarez Melis & Jaakkola, 2018). The main contributions of the paper are following:

- Applying and evaluating SENNs on RL problem and telecom use case.
- Modifying the architecture of SENNs to improve explainability and robustness.
- Proposing a method to extract global explanations from the local SENN explanations.

2. Related work

SENNs are intrinsically interpretable neural networks that decompose their output as a linear combination of learned input concepts weighted by input-dependent relevance scores

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(Alvarez Melis & Jaakkola, 2018). The most recent work on SENNs has extended the original formulation in several directions: C-SENN (Sawada & Nakamura, 2022) introduced contrastive training to improve concept disentanglement, while Q-SENN (Norrenbrock et al., 2024) applied quantization to enforce sparse, binary concept-feature relationships that improved both accuracy and human interpretability on complex vision tasks. These studies did not attempt to extract global explanations from SENNs or or apply them to RL setting.

Existing XRL methods approach explainability from two angles (Milani et al.). The first angle focuses on explaining RL components such as states, rewards, or trajectories. Some of the well-known methods within this approach include reward decomposition and Shapley values calculated on states (Beechey et al., 2023; Juozapaitis et al., 2019; Ruggeri et al., a). The second angle instead addresses the dominant source of opacity in state-of-the-art RL algorithms that stems the deep neural networks — such as those used in DQN or PPO — that parametrize policy or value functions (Qing et al.; Schulman et al., 2017; Mnih et al.).Within this approach the main line of research use more explainable models or explain model's decisions through popular model agnostic methods such as SHAP, DeepLift or InputxGradient, which estimate input feature attribution towards model's final decision (Shrikumar et al., 2017a;b; Lundberg & Lee, 2017; Atrey et al.; Liu et al.; Ruggeri et al., c). In this work, we follow the line of research on intrinsic explainability by applying SENNs to RL policies. Unlike previous work, which focuses on replacing deep models with less expressive tree-based and linear models, SENNs have the potential to be as performant as ordinary deep models while bringing enhanced explainability.

Proposed methods are evaluated using *mobile-env* a simulation environment introduced in (Schneider et al., 2022) that simulates a resource allocation problem in mobile networks. Currently the optimal solution for mobile-env is based on PPO algorithm where actor and critic are parametrized using DNNs. RL, despite its proven superior performance, to heuristics-based solutions, is not yet widely used in the Coordinated Multipoint problem and, to the best of our knowledge most infrastructure still relies on heuristics (Schneider et al., 2023b). Most likely due to their reliability and transparency, despite their poor scaling and poorer performance in longer time-horizons (Schneider et al., 2023b).

3. Self-explaining reinforcement learning

SENNs consist of three modules: conceptizer, parametrizer, and aggregator. The conceptizer extracts features, providing the concept vector $\mathbf{h}$, the parametrizer produces a relevance score vector $\mathbf{R}$. Lastly, the aggregator aggregates the concept vector and relevance score matrix, and is implemented through dot product. Final output, is expressed as:

$$f(x) = \theta(x)^T h(x) = \mathbf{R}_{C\times m} \cdot \mathbf{h}_{m\times 1} \tag{1}$$

where, $\mathbf{R}$ are relevance scores matrix, $\mathbf{h}$ is concept vector, C and m are number of classes and number of concepts respectively.

The explainability of SENNs originates from the relevance scores $\mathbf{R}$ and local stability of its relevance scores with regard to the concept vector. Each concept is weighted by the corresponding relevance score, providing us with the attribution of each feature towards the final output. Local stability forces similar attributions for samples with similar concept values. To measure the stability of explanations, Lipschitz continuity is leveraged with Lipschitz constant, defined as:

$$L = \max_{x\in X} \frac{\|f(x) - f(x_0)\|}{\|h(x) - h(x_0)\|}, \tag{2}$$

where $f(x)$ is a SENN model, $h(x)$ is a conceptizer.

Due to the computational complexity of calculating the exact value of L, authors of (Alvarez Melis & Jaakkola, 2018), use stochastic gradient descent to estimate it. To ensure explainability, SENNs are trained with a loss function, $\mathcal{L}_y(f(x), y) + \lambda\mathcal{L}_\theta(f) + \xi\mathcal{L}_h(x, \hat{x})$, that allows to influence the local stability of the model through adjustment of λ. $\mathcal{L}_\theta(f)$ is a robustness loss, $\mathcal{L}_y(f(x), y)$ is a classification loss, the last term is conceptizer's reconstruction loss (Alvarez Melis & Jaakkola, 2018).

3.1. Modifications

In this work, we consider low-dimensional, non-visual domain where the input dimensions are interpretable. Motivated by this setting, we adopt an identity conceptizer that uses the input features directly as concepts for the explanations. Beyond this, we introduce a trainable bias vector in the aggregator. We hypothesize that the bias term can absorb the state-independent component of the policy's action preferences, allowing the relevance scores $\theta(s)$ to represent state-specific deviations from the baseline rather than encoding both signals jointly, which could result in noisier explanations.

With these modifications we can rewrite Equation (1) as follows:

$$f'(x) = \theta(x)^T x + b, \tag{3}$$

where b stands for the bias term. We aim to obtain an explainable policy learned by the actor, leaving value prediction uninterpretable as it is not used during inference. Thus, the actor, policy is implemented via SENN and critic, the value function via DNN. Fig. 1 presents the SENNs' architecture and the Proximal Policy Optimization (PPO)'s diagram.

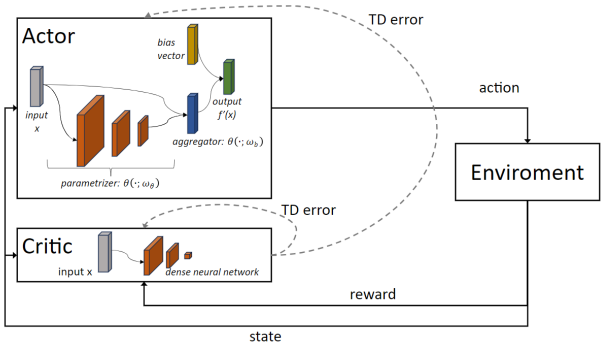

Figure 1. Overview of the method, the SENN is an actor and models the policy. Critic, evaluating actor decision is implemented via the DNN (Schulman et al., 2017). The environment takes in the actor's action and returns the state and reward.

3.2. Local explanations

SENNs provides local intrinsic explanations, which are sourced from the parametrizer's relevance score. Relevance scores weigh the input features, providing a numerical measure of the feature's contribution towards the final decision, similar to linear models. For each decision, the local explanation consists of the relevance scores $\theta(x)$ and the effect scores. The effect scores are defined as follows, inspired by linear models' explanations (Molnar, 2025): $E_i = \theta(x_i) \odot x, \quad \forall i \in \{1, \ldots, m\}$, where $\theta \in \mathcal{R}^{m \times n}$ and $x \in \mathcal{R}^{n \times 1}$, n is input size and m is number of output classes and $\odot$ is the Hadamard product. The effects are relevance scores scaled by the feature values incorporating information of direct feature's contribution. Relevance scores reflect feature importance, while effects capture their direct contribution to the predictions, together they offer a more complete explanation of model's decision. Furthermore, the lack of black-box conceptizer improves the clarity of the explanations, discarding the need for often dubious concept explanations (Sawada & Nakamura, 2022).

3.3. Global explanations

While local explanations can clarify individual decisions made by the model, they do not capture the model's general behaviour like global explanations do. We therefore propose a novel clustering-based method for extrapolating global explanations based on the local explanations naturally provided by SENNs. To create global explanations we first collect set of decisions for which we extract local explanations, then we aggregate those explanations creating more general, global explanations. Then, we calculate *sway scores*, defined as follows:

$$s_i(x) = \theta_i^{a^*} \cdot h_i(x) \; - \; \max_{a \neq a^*} \left(\theta_i^a \cdot h_i(x) \right) \qquad (4)$$

A sway score expresses a decision margin of the i-th concept i.e., how much concept i-th argues for action a^* in comparison to the highest i-th concept value from the remaining actions. Unlike relevance scores, sway scores aim to capture the impact on the decision itself, relative to other actions, rather than the magnitude of concept activation. Finally, we cluster sway scores via k-means and average within each cluster, this preserves the structure that would be lost when averaging across all samples of an action. Additionally, we can aggregate the clustering results to create attribution scores (see section 4.3). This allows us to validate our method by comparing it against state-of-the-art post-hoc techniques.

4. Experimental setup

4.1. Mobile-env

We apply our method to mobile-env, a simulator that implements a resource allocation problem in mobile networks (Schneider et al., 2022). In this environment, base stations (BSs) are distributed over an area and provide mobile connectivity to, randomly moving, user equipment (UE), which rely on the BSs to maintain network access. The goal is to determine a sequence of BS–UE assignments that maximizes the overall utility, defined as the cumulative Quality of Experience (QoE) across all UEs. Finding the exact solution to this problem is a challenge even for the small instance; thus, existing solutions rely on simple heuristics or deep reinforcement learning (Schneider et al., 2023a). Following centralized training with decentralized execution, we train a single model on data aggregated from all UEs, then deploy a copy to each UE independently. We approach a small instance of the problem with three UEs and three BSs. Each UE observes a 13-element vector comprising connection statuses, SNR values, its own utility, and the load and utility of each BS. The reward is QoE, a logarithmic score in $(-10, 10)$, and the single action available to each UE is to connect to or disconnect from a chosen BS.

4.2. Experiments

We trained a biased SENN model, a SENN model, and a DNN model, using the PPO algorithm with training of 400,000 timesteps across 2 parallel, randomly seeded simulations each consisting of 3 UEs and 3 BSs we decided on episode length of 1200 timesteps. We trained models using PPO and hyperparameters reported by (Schneider et al., 2023b) except: entropy coefficient of 0.001 and GAE-λ of 0.99. DNN-based policy was modeled by 3-layer, 256 hidden units each, neural network, SENN-based policies had parametrizer network of the same size and identity conceptizer. Value network for all cases was modeled by 4-layer, 256 hidden units each DNN.

Table 1. Comparison of mean and median episodic return from 15 episodes each 1200 timesteps.

MODEL	MEAN RETURN	MEDIAN RETURN
DNN	1.803 ± 0.247	1.890
SENN	1.805 ± 0.200	1.807
BIASED SENN	1.780 ± 0.239	1.799
DYNAMIC SELECTION	1.428 ± 0.219	1.456

4.3. Evaluation

To evaluate performance of trained models on the resource allocation task we evaluated it over 18,000 timesteps. This evaluation was run on fifteen differently seeded episodes each of 1200 timesteps. Furthermore, to compare effect of local stability and robustness factor on the predictive performance we ran training with different λ values and estimated Lipschitz constant for each model using gradient descent, in accordance with (Alvarez Melis & Jaakkola, 2018). Trajectories collected during performance evaluation were first filtered, we removed toggle actions that were not applied to the environment due to User Equipment (UE) being out of range of the Base Station (BS) range, and used them to generate explanations. Local explanations were taken from randomly chosen timestep. To generate *global explanations* we calculated sway scores for each decision in filtered trajectories and clustered sway scores using K-means algorithm on the arbitrarily chosen range of $k = \{2, 3, 4, 5, 6\}$. For each action we selected number of clusters with the highest silhouette score. After clustering to have more quantifiable measure of explanations faithfulness we decided to compare explanations with well-established existing post-hoc methods: GradSHAP, DeepLift and InputXgradient (Lundberg & Lee, 2017; Shrikumar et al., 2017a;b). We generated clustering-based attributions by averaging sway scores over clusters for each action.

5. Results and discussion

In this section we present the performance results of SENN and biased SENN against baselines, the stability–accuracy trade-off, and exemplary local and global explanations.

5.1. Performance comparison

We first benchmark SENNs against DNN and heuristic baselines to confirm predictive performance worth explaining. We trained all deep models with PPO. For a faithful comparison we first tried to match the episodic return reported in (Schneider et al., 2023b), using DNNs. Then with the same hyperparameters we trained SENNs variants. Lastly, we also compared obtained results with currently used and best performing heuristic, Dynamic Selection (Schneider et al., 2023b).

Tab. 1 shows results of the performance evaluation over 15 episodes. DNN model is performing the best, all deep models significantly outperform Dynamic Selection($\epsilon = 0.5$) and biased SENN and SENN have similar performance. Small performance gap between DNN and SENNs, is most likely a result of the stability imposed on the SENNs through the robustness loss. For this experiment robustness loss factor was set to $\lambda = 0.001$.

Local stability forces SENNs to learn relevance scores consistent with concept values, but enforcing it introduces the robustness loss, a strong regularizer. Understanding the relation between episodic reward and local stability (estimated via L, Eq. (2)) is a key to selecting an appropriate λ and assessing the cost of explainability on predictive performance. Fig. 2 presents, the relation between stability and episodic rewards. Each, biased SENN model was trained with different λ and evaluated on 3600 timesteps, estimation of L is calculated based on the trajectories collected during the evaluation. It shows trade-off between local stability and performance, episodic returns drop with an increasing Lipschitz constant and λ. The relation appears to be linear in nature; however, its character can be problem dependent as hinted by (Alvarez Melis & Jaakkola, 2018).

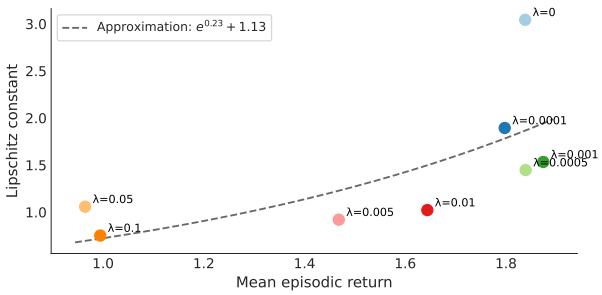

Figure 2. Relation between, different λ values, Lipschitz constant and episodic return (averaged over 3600 timesteps).

Results, in line with (Alvarez Melis & Jaakkola, 2018), confirm that SENN-based models almost match the performance of similarly sized DNNs, and thus (Schneider et al., 2023b), demonstrating successful application to resource allocation in mobile networks and RL more broadly. They additionally offer adjustable explainability via local stability and λ. The bias term, in biased SENNs, has negligible effect on the performance. Overall, SENNs-based models seem to be a good alternative to DNNs, even though they perform slightly worse than DNNs, they offer intrinsic explainability.

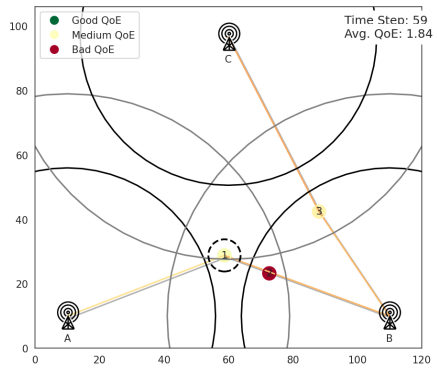

Figure 3. Visualization of the environment, timestep 59.

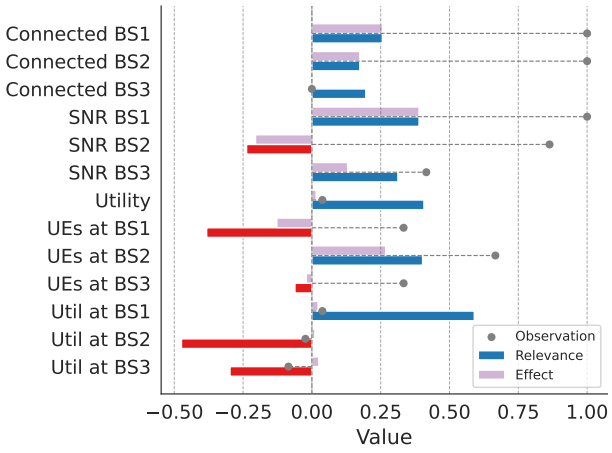

Figure 4. Relevance scores, effect scores and observations corresponding to the timestep 59, UE1 and action toggle BS1, see Fig. 3.

5.2. Local explanations

SENNs-based models' explainability mainly stems from exposing the relevance scores, a set of input-dependent coefficients that weigh concepts. Visualizing relevance scores shows direct contribution of concepts for each individual decision.

We present observations for a randomly chosen timestep and UE in Fig. 4 and their visualization in Fig. 3. Chosen UE 1 is connected to base BSs A (1) and B (2), in that timestep it made a decision to disconnect from BS B. Relevance scores and effect scores for that particular decision are presented in Fig. 4, effects are concepts scaled by the relevance scores, see Eq. (3.2). Fig. 4, shows concepts that contributed strongly positive (connection with BS1, connection with BS2, SNR of BS1, SNR of BS3, UE's utility, number of UEs at BS2 , Utility at BS1) and concepts that contributed negatively (SNR of BS2, number of UEs at BS1). These relevance scores provide a fairly clear idea of how concepts participated in making a decision at a given timestep. Lastly, contribution of each concept does not seem to be surprising and corresponding relevance scores seem reasonable, *hypothetical* an interpretation of the reason behind this action is that, UE disconnected from the BS2 due to stronger signal of BS1 and its higher utility as well as low own utility, despite strong signal from BS2.

SENNs-based models provide understandable, intrinsic explanations that can be accessed for every individual decision. Even though SENNs are deep models, due to the local stability, relevance scores are forced to behave linearly locally in respect to concepts. In conclusion SENNs provide intrinsic, per-decision understandable explanations with locally linear relevance scores.

5.3. Global explanations

Local explanations, while highly useful, do not provide insights into the model's general behavior that global explanations offer. In this section, we present results of applying the proposed clustering-based method of aggregating concept sways, see Eq. (4), across multiple runs to explain the general behavior of SENNs.

Figure 5. Mean sway scores per concept for the toggle BS3 action, obtained by clustering sways collected over 18 000 evaluation timesteps and averaging within each cluster.

We used biased SENN trained with $\lambda = 0.001$. After collecting the data we clustered sway scores by action. We then averaged sway scores within the resulting clusters to obtain typical concept sways for each action. Fig. 5 presents two clusters with typical sways for the toggle BS3 action. In Cluster 1, the SNR strongly sways the model toward changing the connection status with BS2, while the SNR of BS2 and the number of UEs at BS3 sway against the action. In cluster 0, however, the average sway scores are considerably more surprising, with no significant contributions except from the SNR of BS1 and the connection status to BS1. Given a Silhouette score of 0.433, this most likely indicates

a weak cluster structure in cluster 0.

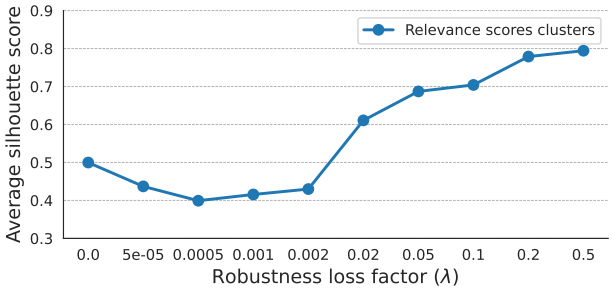

Figure 6. Silhouette score averaged over all actions, plotted against the robustness loss factor.

Figure 7. Spearman correlation per action, of clustering derived attributions and other attributive XAI methods.

Furthermore, we tested how the quality of clustering changes under different local stability constraints, for this purpose 5 different models were trained with different robustness losses. We averaged the silhouette score over actions, for each λ. Results of this experiment are presented in Fig. 6. It appears that for higher λ values the silhouette score significantly improves.

Lastly, to evaluate accuracy of global explanations and validate the approach we decided to extract attributions of each concept and compare it with well-established post-hoc methods: GradientSHAP, DeepLift and InputXGradient (Lundberg & Lee, 2017; Shrikumar et al., 2017a;b). We calculated the attributions by averaging sway scores over concepts, then plot Spearman correlation of obtained attributions scores with attributions from other methods, see Fig. 7. Fig. 7, shows that DeepLift and InputXGradient are strongly correlated with our attributions, while GradSHAP shows weak correlation at best. The key advantage of clustering over simple averaging is that it preserves the diversity of the model's behaviour, rather than collapsing all decisions into a single mean that may average out opposing values, clustering separates them into distinct typical sway profiles. Furthermore, the improvement in silhouette score with higher λ values (Fig. 7) hints at relation between enforcing local stability and preserving cluster structure within the relevance scores space.

Regarding the comparison with gradient-based methods, while clustering attributions correlate with DeepLift and InputXGradient, an important distinction must be noted: relevance scores explain the final decision of the model directly, not the internal computation of the parametrizer. This makes them more immediately interpretable, as they quantify each concept's direct contribution to the output rather than propagating gradients through opaque intermediate layers, however it also introduces an element of uncertainty coming from opaqueness of the parametrizer. Lastly, in our experiments we visualized the learned bias, it provides a useful indication of which actions the policy will take more frequently; however, its effects on decontamination of the relevance scores are insignificant. In conclusion the proposed methods enable SENNs to provide both local and global intrinsic explanations, which constitute the main interpretability advantage of the approach.

6. Conclusions

This work is the first to apply SENN-based models to an RL problem, demonstrating that SENNs trained with PPO achieve performance close to state-of-the-art DNN-based methods while significantly outperforming the best heuristics. Beyond predictive performance, SENNs provide intrinsic local explanations for individual decisions, and the proposed clustering-based method successfully aggregates these into global explanations, validated through strong correlation with established gradient-based post-hoc methods. This combination of competitive performance with both local and global explanations grounded in the model's actual decision mechanism makes SENNs a promising candidate for high-stakes domains where transparency is a prerequisite and input features are understandable.

The primary limitation is that explanation quality is sensitive to the choice of λ, which currently requires empirical tuning. For future work, one natural direction is an adaptive local stability mechanism, where the Lipschitz constraint varies with concept space density, reducing the performance cost of explainability. Additionally, further investigation into how local stability shapes the relevance score space could deepen theoretical understanding and guide deployment across broader RL settings.

7. Impact Statement

Our results suggest that SENNs can be applied to a broad range of RL problems, offering strong predictive performance alongside intrinsic explainability grounded in the model's internal mechanism. In practice, SENN-based RL systems could provide reliable per-timestep explanations of agent decisions and help diagnose biases in learned policies, contributing positively to safety and fairness - particularly in sensitive applications.

However, the approach carries risks that should be acknowledged. The choice of robustness loss factor strongly influences explanation quality, and no universal value guarantees consistent explanations across problems. Additionally, SENNs-based models provide per timestep decisions and do not capture the temporal aspect, which may be important for understanding long-term strategies. Lastly, our solution, in general, depends on interpretable input features or concepts that could be extracted from the input data. Lack of such could be detrimental towards understandability of provided explanations.

Overall, SENN-based RL methods represent a meaningful step toward more transparent and trustworthy reinforcement learning, but their responsible deployment requires careful tuning and awareness.

References

Alvarez Melis, D. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Atrey, A., Clary, K., and Jensen, D. Exploratory not explanatory: Counterfactual analysis of saliency maps for deep reinforcement learning. URL https://iclr.cc/virtual_2020/poster_rkl3m1BFDB.html.

Beechey, D., Smith, T. M. S., and Şimşek, Ö. Explaining reinforcement learning with shapley values. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 2003–2014. PMLR, 2023. URL https://proceedings.mlr.press/v202/beechey23a.html.

Juozapaitis, Z., Koul, A., Fern, A., Erwig, M., and Doshi-Velez, F. Explainable reinforcement learning via reward decomposition. In *in proceedings at the International Joint Conference on Artificial Intelligence. A Workshop on Explainable Artificial Intelligence.*, 2019.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. Explainable AI: A review of machine learning interpretability methods. 23(1):18. ISSN 1099-4300. doi: 10.3390/e23010018. URL https://www.mdpi.com/1099-4300/23/1/18.

Liu, G., Schulte, O., Zhu, W., and Li, Q. Toward interpretable deep reinforcement learning with linear model u-trees. In Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., and Ifrim, G. (eds.), *Machine Learning and Knowledge Discovery in Databases*, pp. 414–429. Springer International Publishing. ISBN 978-3-030-10928-8. doi: 10.1007/978-3-030-10928-8_25.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Milani, S., Topin, N., Veloso, M., and Fang, F. Explainable reinforcement learning: A survey and comparative review. 56(7):168:1–168:36. ISSN 0360-0300. doi: 10.1145/3616864. URL https://dl.acm.org/doi/10.1145/3616864.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. 518(7540):529–533. ISSN 1476-4687. doi: 10.1038/nature14236. URL https://www.nature.com/articles/nature14236.

Molnar, C. *Interpretable Machine Learning*. 3 edition, 2025. ISBN 978-3-911578-03-5. URL https://christophm.github.io/interpretable-ml-book.

Norrenbrock, T., Rudolph, M., and Rosenhahn, B. Q-senn: quantized self-explaining neural networks. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i19.30145. URL https://doi.org/10.1609/aaai.v38i19.30145.

Qing, Y., Liu, S., Song, J., Zhou, Y., Chen, K., Wang, H., and Song, M. A survey on explainable reinforcement learning: Concepts, algorithms, challenges. URL http://arxiv.org/abs/2211.06665.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?": Explaining the predictions of

any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL https://doi.org/10.1038/s42256-019-0048-x.

Ruggeri, F., Emanuelsson, W., Terra, A., Inam, R., and Johansson, K. H. Rollout-based shapley values for explainable cooperative multi-agent reinforcement learning. In *2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*, pp. 227–233, a. doi: 10.1109/ICMLCN59089.2024.10624777. URL https://ieeexplore.ieee.org/document/10624777.

Ruggeri, F., Russo, A., Inam, R., and Johansson, K. H. Explainable reinforcement learning via temporal policy decomposition, b. URL http://arxiv.org/abs/2501.03902.

Ruggeri, F., Terra, A., Inam, R., and Johansson, K. H. Evaluation of intrinsic explainable reinforcement learning in remote electrical tilt optimization. In Yang, X.-S., Sherratt, R. S., Dey, N., and Joshi, A. (eds.), *Proceedings of Eighth International Congress on Information and Communication Technology*, volume 696, pp. 835–854. Springer Nature Singapore, c. ISBN 978-981-99-3235-1 978-981-99-3236-8. doi: 10.1007/978-981-99-3236-8_67. URL https://link.springer.com/10.1007/978-981-99-3236-8_67. Series Title: Lecture Notes in Networks and Systems.

Sawada, Y. and Nakamura, K. C-senn: Contrastive self-explaining neural network, 2022. URL https://arxiv.org/abs/2206.09575.

Schneider, S., Werner, S., Khalili, R., Hecker, A., and Karl, H. mobile-env: An open platform for reinforcement learning in wireless mobile networks. In *NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium*, pp. 1–3, 2022. doi: 10.1109/NOMS54207.2022.9789886.

Schneider, S., Karl, H., Khalili, R., and Hecker, A. Multi-agent deep reinforcement learning for coordinated multi-point in mobile networks. *IEEE Transactions on Network and Service Management (TNSM)*, 2023a.

Schneider, S., Karl, H., Khalili, R., and Hecker, A. Multi-agent deep reinforcement learning for coordinated multi-point in mobile networks. *IEEE Transactions on Network and Service Management (TNSM)*, 2023b.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://dblp.uni-trier.de/db/journals/corr/corr1707.htmlSchulmanWDRK17.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153. PMLR, 06–11 Aug 2017a. URL https://proceedings.mlr.press/v70/shrikumar17a.html.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences, 2017b. URL https://arxiv.org/abs/1605.01713.

Shu, T., Xiong, C., and Socher, R. Hierarchical and interpretable skill acquisition in multi-task reinforcement learning. URL https://openreview.net/forum?id=SJJQVZW0b.

Watson, D. S. Conceptual challenges for interpretable machine learning. *Synthese*, 200, 2022. doi: 10.1007/s11229-022-03485-5. URL https://doi.org/10.1007/s11229-022-03485-5.