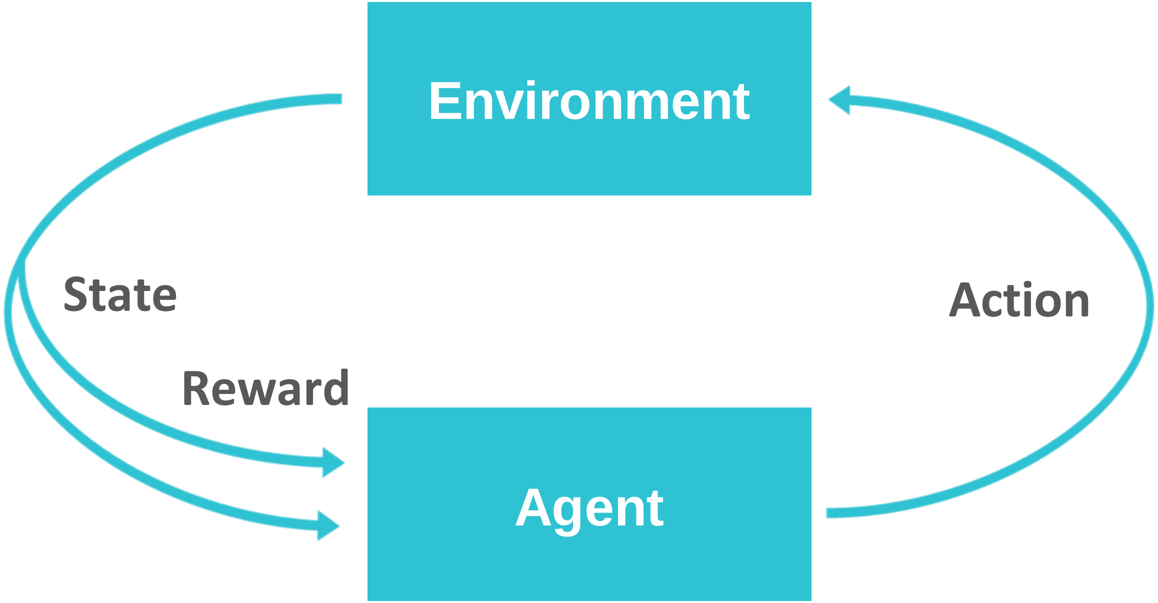
Explaining Reinforcement Learning with Shapley Values: Theory and Algorithms

Daniel Beechey
Bath Reinforcement Learning Lab (BRLL)



Learn a policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$ that maps each state to a probability distribution over actions, **maximising the expected return**:

$$\mathbb{E}[G_t] = \mathbb{E}[\sum_{k=0} \gamma^k R_{t+k+1}]$$

[11]

What Can Reinforcement Learning Do?

Atari [4]

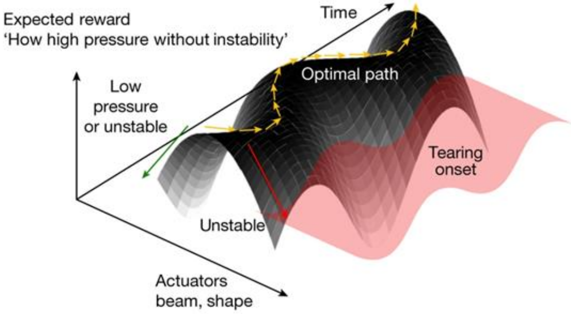
AlphaGo [6][9][16]

StarCraft II [14]

Gran Turismo [20]

Matrix Multiplication [19]


Stratospheric Balloons [15]



Nuclear Fusion Reactor Control [18][21]

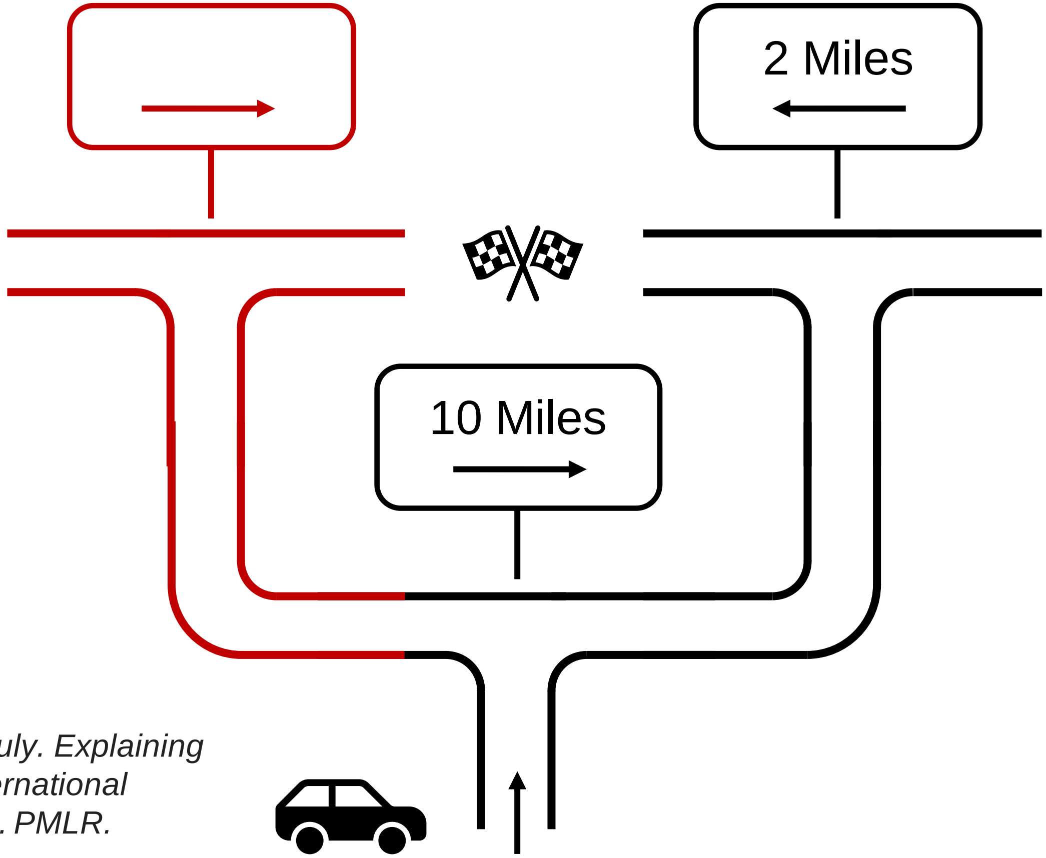
Reinforcement learning agents do not explain their actions.

Understanding Agent-Environment Interactions

Certain features of an agent's observations influence how they interact with their environment.

Contribution: A theoretical and computational framework for explaining agent-environment interactions using the influence of features.

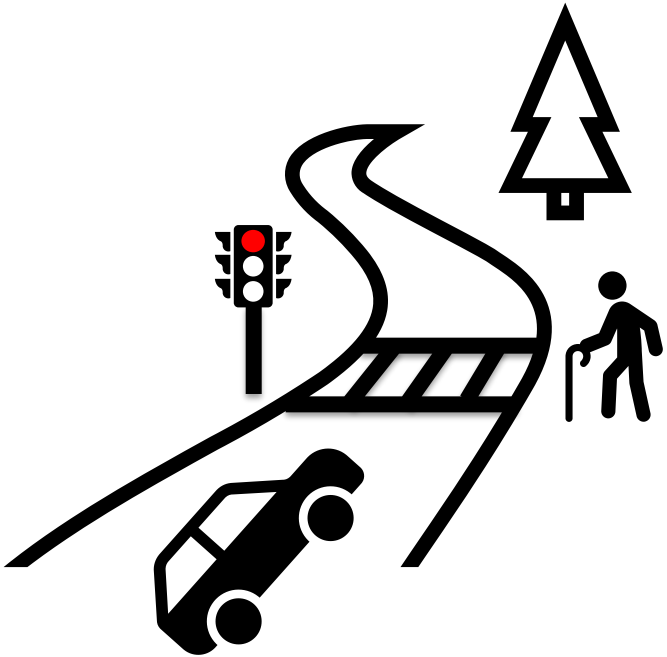
- **Behaviour**
- **Outcome**
- **Prediction**



Beechey, D., Smith, T.M. and Şimşek, Ö., 2023, July. Explaining reinforcement learning with Shapley values. In International Conference on Machine Learning (pp. 2003-2014). PMLR.

Compute the influence of features by observing the effect of their removal.

Features are interdependent; removing one feature does not properly capture its contribution.



Shapley Values

A **cooperative game** is a set of players $\mathcal{F}$ and a characteristic function $v(\mathcal{C}) : 2^{|\mathcal{F}|} \rightarrow \mathbb{R}$.

How to assign the contribution $\phi_i(v)$ of player i to the outcome of the game $(\mathcal{F}, v)$?

$$\phi_i(v) = \sum_{\mathcal{C} \subseteq \mathcal{F} \setminus i} \frac{|\mathcal{C}|! \cdot (|\mathcal{F}| - |\mathcal{C}| - 1)!}{|\mathcal{F}|!} \left[v(\mathcal{C} \cup i) - v(\mathcal{C}) \right]$$

[1]

Shapley Value Axioms

$$\phi_i(v) = \sum_{\mathcal{C} \subseteq \mathcal{F} \setminus i} \frac{|\mathcal{C}|! \cdot (|\mathcal{F}| - |\mathcal{C}| - 1)!}{|\mathcal{F}|!} [v(\mathcal{C} \cup i) - v(\mathcal{C})]$$

Shapley values $\phi_i(v)$ are the **unique solution** to the **contribution assignment problem** that satisfies the **four axioms** of fair contribution.

Efficiency: $v(\mathcal{F}) = v(\emptyset) + \sum_{i \in \mathcal{F}} \phi_i(v).$

Symmetry: $\phi_i(v) = \phi_j(v) \quad \text{if} \quad v(\mathcal{C} \cup \{i\}) = v(\mathcal{C} \cup \{j\}) \quad \forall \mathcal{C} \subseteq \mathcal{F} \setminus \{i, j\}.$

Nullity: $\phi_i(v) = 0 \quad \text{if} \quad v(\mathcal{C} \cup \{i\}) = v(\mathcal{C}) \quad \forall \mathcal{C} \subseteq \mathcal{F} \setminus \{i\}.$

Linearity: $\phi_i(\alpha u + \beta v) = \alpha \phi_i(u) + \beta \phi_i(v).$

[1]

Shapley Values for Explaining Reinforcement Learning (SVERL)

$$\phi_i(v) = \sum_{\mathcal{C} \subseteq \mathcal{F} \setminus i} \frac{|\mathcal{C}|! \cdot (|\mathcal{F}| - |\mathcal{C}| - 1)!}{|\mathcal{F}|!} \left[v(\mathcal{C} \cup i) - v(\mathcal{C}) \right]$$

Beechey, D., Smith, T. and Şimşek, Ö., 2025. A Theoretical Framework for Explaining Reinforcement Learning with Shapley Values. arXiv preprint arXiv:2505.07797.

Explaining Behaviour. The contribution of feature values to the probability of selecting action a in state s.

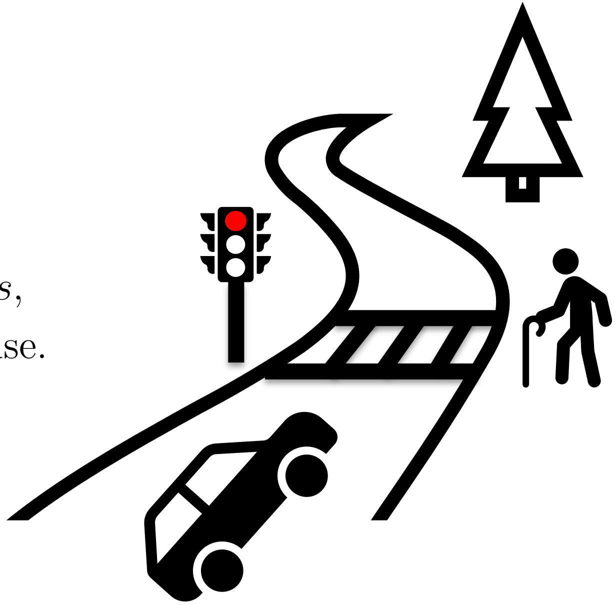
$$\tilde{\pi}_s^a(\mathcal{C}) = \mathbb{E}[\pi(S,a) \mid S_{\mathcal{C}} = s_{\mathcal{C}}] = \sum_{s' \in \mathcal{S}^+} p^\pi(s' \mid s_{\mathcal{C}})\pi(s',a)$$

Explaining Outcome. The contribution of feature values to the expected return $v^\pi(s)$.

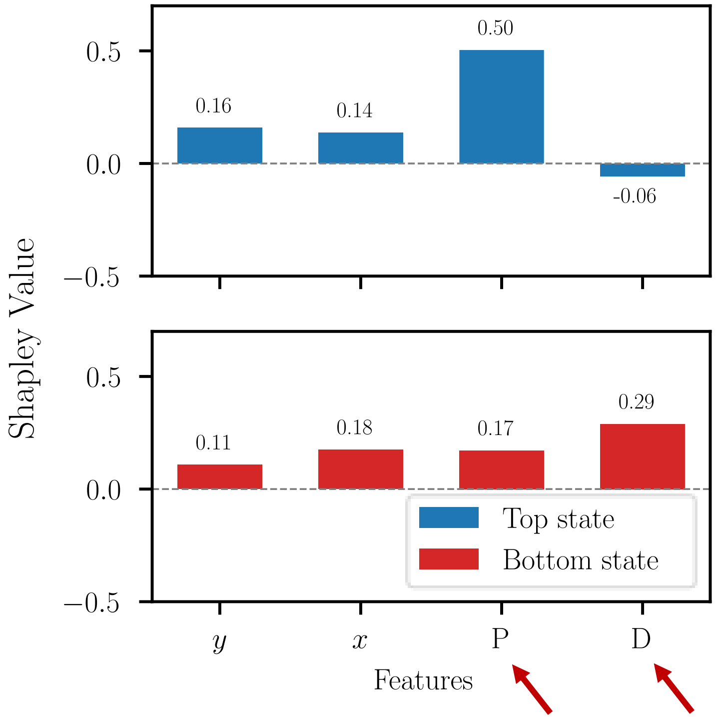
$$\tilde{v}_s^\pi(\mathcal{C}) = \mathbb{E}_\mu\left[G_t \mid S_t = s\right], \text{ where } \mu(s_t, a_t) = \begin{cases} \pi_{s_t}^{a_t}(\mathcal{C}) & \text{if } s_t = s, \\ \pi(s_t, a_t) & \text{otherwise.} \end{cases}$$

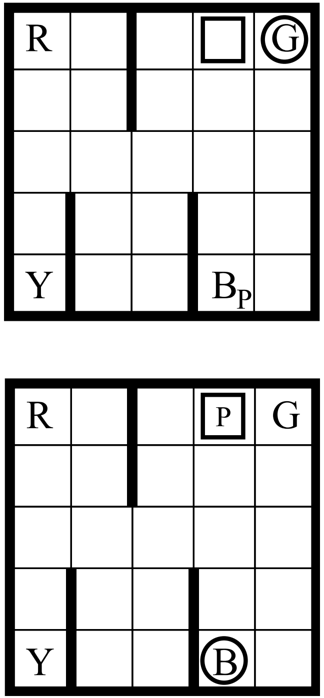
Explaining Prediction. The contribution of feature values to the predicted expected return $\hat{v}^\pi(s)$.

$$\hat{v}_s^\pi(\mathcal{C}) \overset{\text{def}}{=} \hat{u}^\pi(s_{\mathcal{C}}) = \mathbb{E}[\hat{v}^\pi(S) \mid S_{\mathcal{C}} = s_{\mathcal{C}}] = \sum_{s' \in \mathcal{S}^+} p^\pi(s' \mid s_{\mathcal{C}})\, \hat{v}^\pi(s').$$



Explaining Behaviour in Taxi



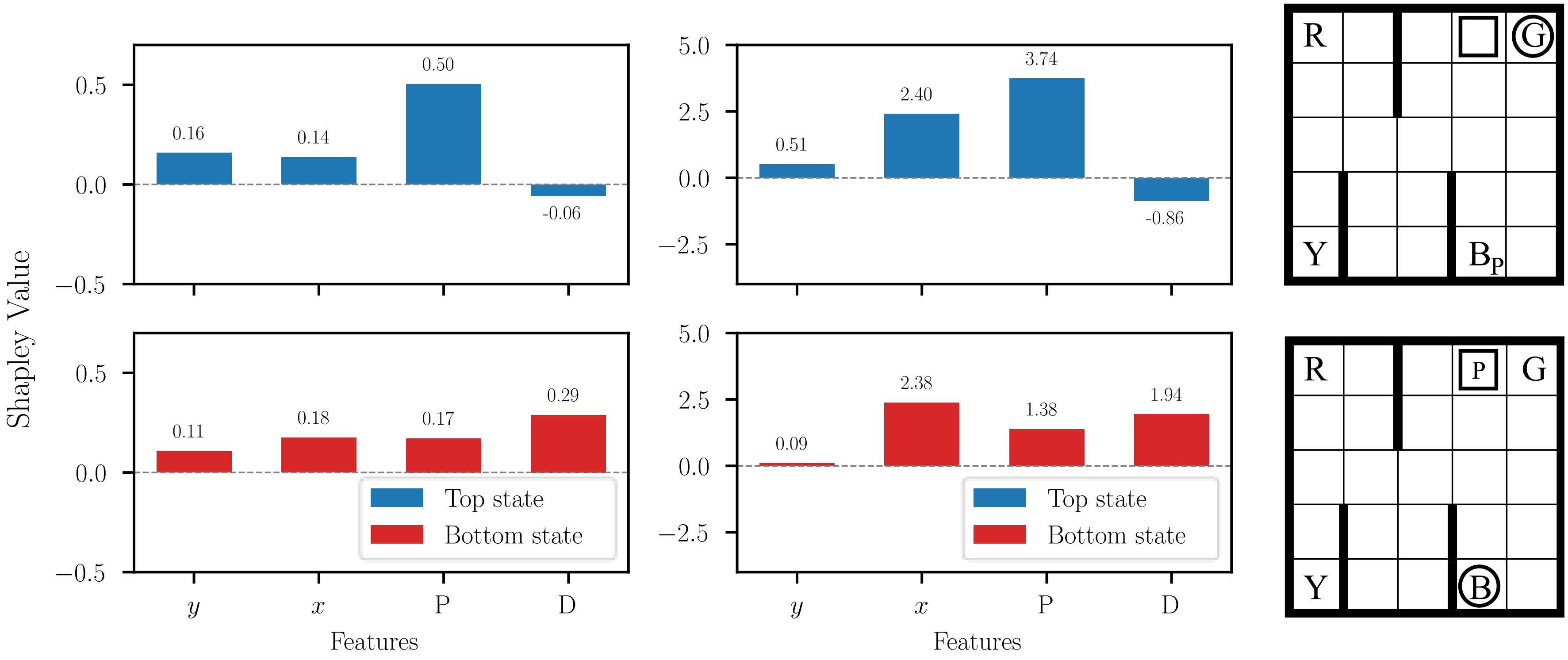


[2]

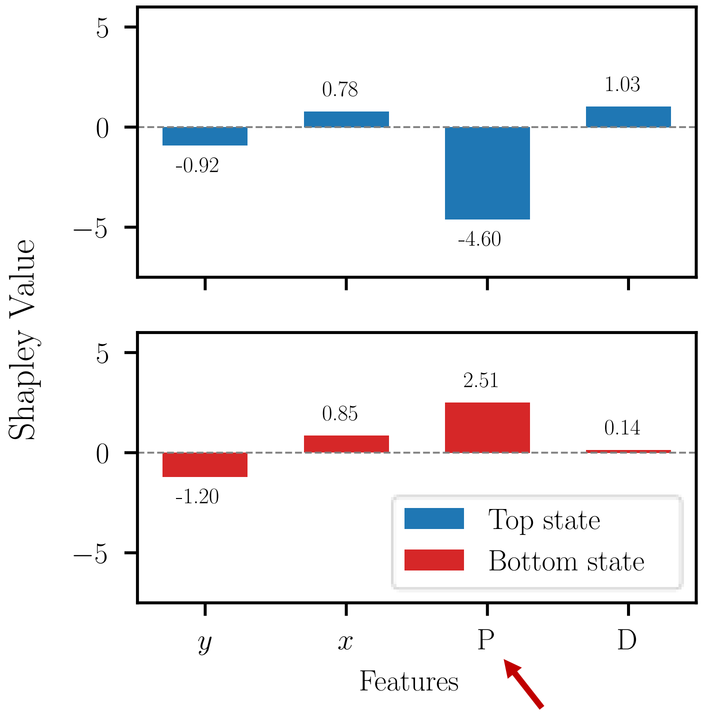
Explaining Outcomes in Taxi

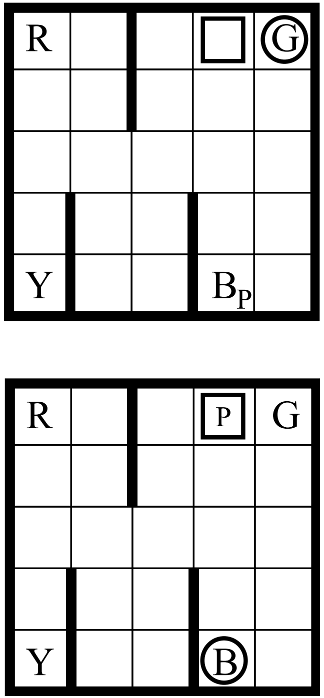
art-ai

UNIVERSITY OF BATH



Explaining Prediction in Taxi





How to Approximate SVERL for Real-World Applications

Beechey, D. and Şimşek, Ö., 2025. Approximating Shapley explanations in reinforcement learning. In Advances in Neural Information Processing Systems.

Characteristics average over states and the distribution $p^{\pi}(s|s_{\mathcal{C}})$:

$$\pi_s^a(\mathcal{C}) \stackrel{\text{def}}{=} \mathbb{E}\left[\pi(S, a) \,|\, S_{\mathcal{C}} = s_{\mathcal{C}}\right] = \sum_{s' \in \mathcal{S}} p^{\pi}(s' \,|\, s_{\mathcal{C}}) \pi(s', a)$$

Approximate $\pi_s^a(\mathcal{C})$ with a parametric function, $\hat{\pi}(s, a \,|\, \mathcal{C}; \beta)$

Shapley values sum over the powerset of features, $2^{|\mathcal{F}|-1}$:

$$\phi_i(\pi_s^a) = \sum_{\mathcal{C} \subseteq \mathcal{F} \setminus i} \frac{|\mathcal{C}|! \cdot (|\mathcal{F}| - |\mathcal{C}| - 1)!}{|\mathcal{F}|!} \left[\pi_s^a(\mathcal{C} \cup i) - \pi_s^a(\mathcal{C})\right]$$

Approximate SVERL with a parametric $\hat{\phi}(s, a; \theta) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{F}|}$

art-ai

UNIVERSITY OF BATH

	Clue 1	Pos 1	Pos 2	Pos 3	Pos 4	Clue 2
Guess 6						
Guess 5		C	C	C	B	
Guess 4	2	B	C	C	C	2
Guess 3	2	C	B	C	C	2
Guess 2	0	C	C	C	C	3
Guess 1	0	A	A	C	A	1

Thank you for listening!

1. **Shapley Values for Explaining Reinforcement Learning (SVERL)**
    - Explaining behaviour
    - Explaining outcomes
    - Explaining prediction

2. **How to approximate SVERL in large-scale domains.**
    - Parametric approximations of explanations
    - Learnt off-policy for online learning
    - Continually adapt to evolving agent behaviour

Future Work

- *A real-world application of SVERL*
- *User studies*

Collaborators

Thomas Smith

tmss20@bath.ac.uk

Özgür Şimşek

os435@bath.ac.uk

References

[1] Shapley, L.S. A value for n-person games. (1953).
[2] Dietterich, G.T. The MAXQ method for hierarchical reinforcement learning. *ICML* **98**, 118–126 (1998).
[3] Strumbelj, E., Kononenko, I. An efficient explanation of individual classifications using game theory. *The Journal of Machine Learning Research*, 11, 1-18 (2010).
[4] Mnih, V., Kavukcuoglu, K., Silver, D. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
[5] Ribeiro, M.T., Singh, S., 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,* 1135-1144 (2016).
[6] Silver, D., Huang, A., Maddison, C. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
[7] Wang, Z., Schaul, T., Hessel, M. et al. Dueling network architectures for deep reinforcement learning. *ICML*, 1995–2003 (2016).
[8] Lundberg, S.M., Lee, S.-L. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30 (2017).
[9] Silver, D., Schrittwieser, J., Simonyan, K. *et al.* Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
[10] Greydanus, S., Koul, A., Dodge, J. et al. Visualizing and understanding atari agents. ICML 1792–1801 (2018).
[11] Sutton, R.S., Barto, A.G. Reinforcement learning: An introduction. *MITpress*, (2018).
[12] Mott, A., Zoran, D., Chrzanowski, M. et al. Towards interpretable reinforcement learning using attention augmented agents. *Advances in neural information processing systems* 32, (2019).
[13] Rizzo, S.G., Vantini, G., Chawla, S. Reinforcement learning with explainability for traffic signal control. *In 2019 IEEE intelligent transportation systems conference* 3567-3572. (2019).
[14] Vinyals, O., Babuschkin, I., Czarnecki, W.M. *et al.* Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* **575**, 350–354 (2019).
[15] Bellemare, M.G., Candido, S., Castro, P.S. *et al.* Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature* **588**, 77–82 (2020).
[16] Schrittwieser, J., Antonoglou, I., Hubert, T. *et al.* Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* **588**, 604–609 (2020).
[17] Zhang, K., Xu, P., Zhang, J. Explainable AI in deep reinforcement learning models: A shap method applied in power system emergency control. *In 2020 IEEE 4th conference on energy internet and energy system integration* (EI2), 711-716. (2020).
[18] Degrave, J., Felici, F., Buchli, J. *et al.* Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature* **602**, 414–419 (2022).
[19] Fawzi, A., Balog, M., Huang, A. *et al.* Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610**, 47–53 (2022).
[20] Wurman, P.R., Barrett, S., Kawamoto, K. *et al.* Outracing champion Gran Turismo drivers with deep reinforcement learning. *Nature* **602**, 223–228 (2022).
[21] Seo, J., Kim, S., Jalalvand, A. *et al.* Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature* **626**, 746–751 (2024).