How to Explain Reinforcement Learning with Shapley Values

Daniel Beechey
Bath Reinforcement Learning Lab
CDT in Accountable, Responsible and Transparent AI (ART-AI)

Collaborators

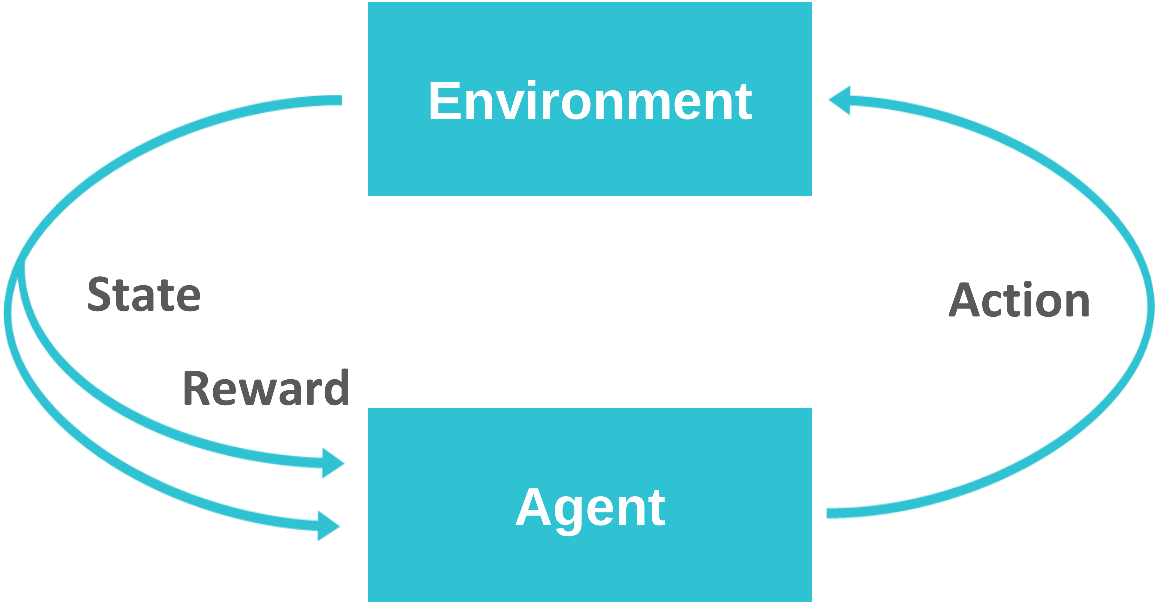
Thomas Smith

tmss20@bath.ac.uk

Özgür Şimşek

os435@bath.ac.uk

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.



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?

art-ai

UNIVERSITY OF BATH

Atari [4]

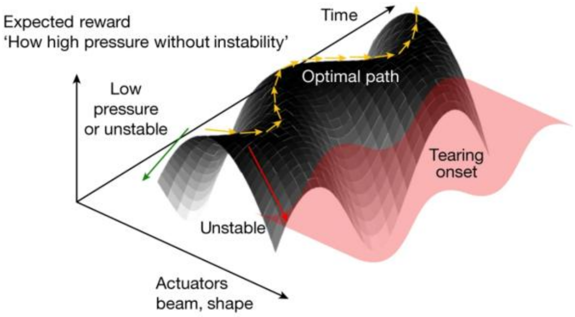
AlphaGo [6][9][16]

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.

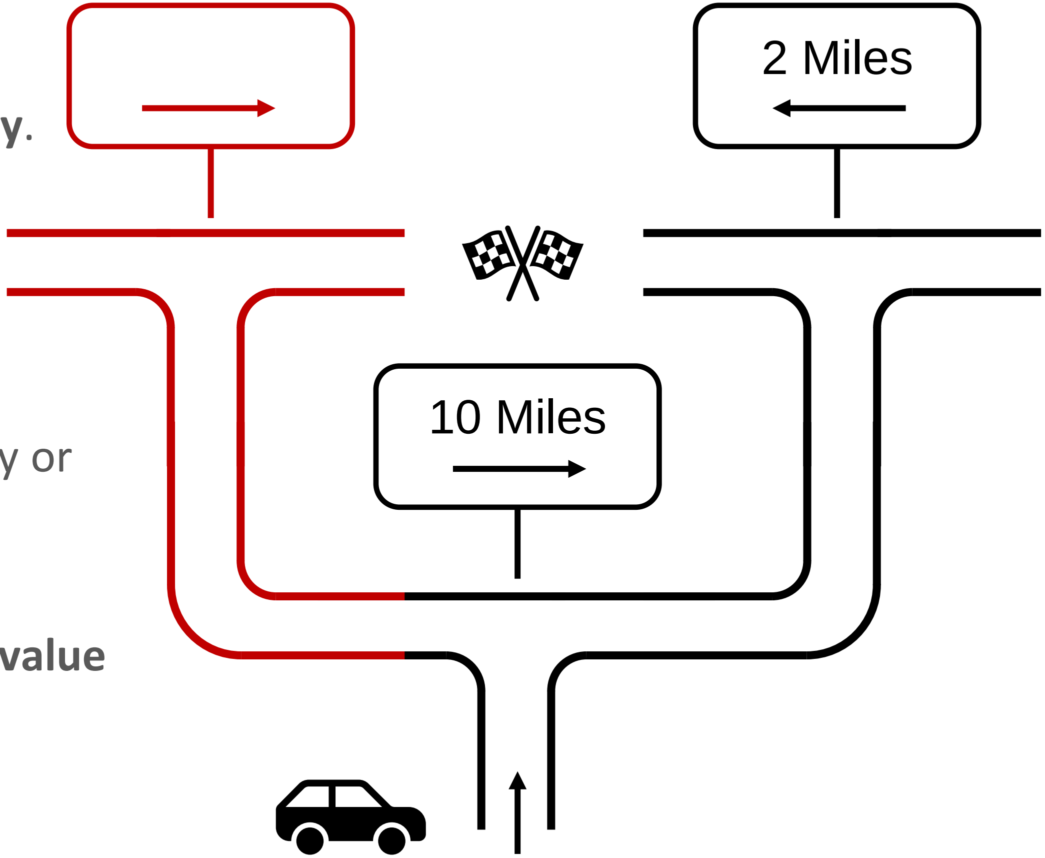
Understanding Agent-Environment Interactions

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

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

What Needs Explaining?

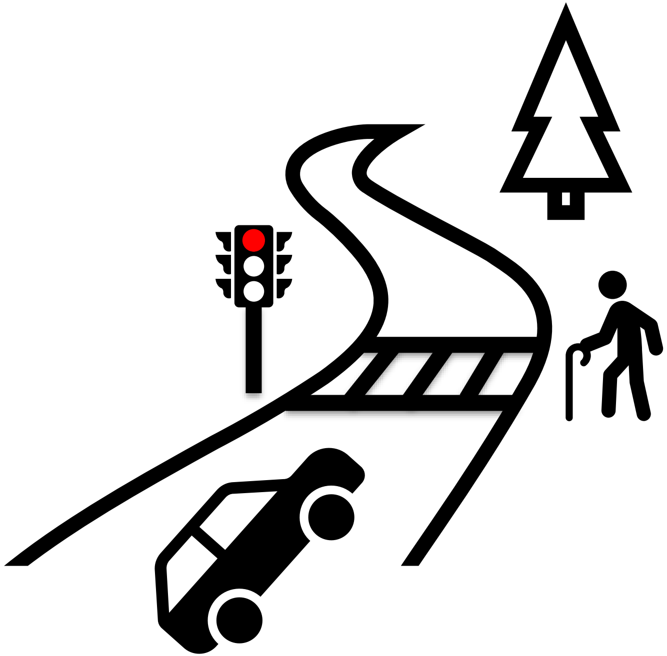
- Arrow directions influence **policy**.
- Arrow directions influence **performance**.
- Distances do not influence policy or performance.
- Destination distances influence **value prediction**.



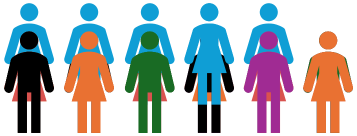
Computing Feature Influence

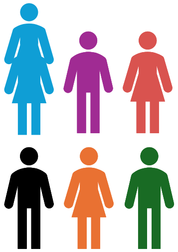
Compute the influence of features by observing the behaviour change caused by their removal.

Features are interdependent, removing one feature does not properly capture its influence.



Shapley Values

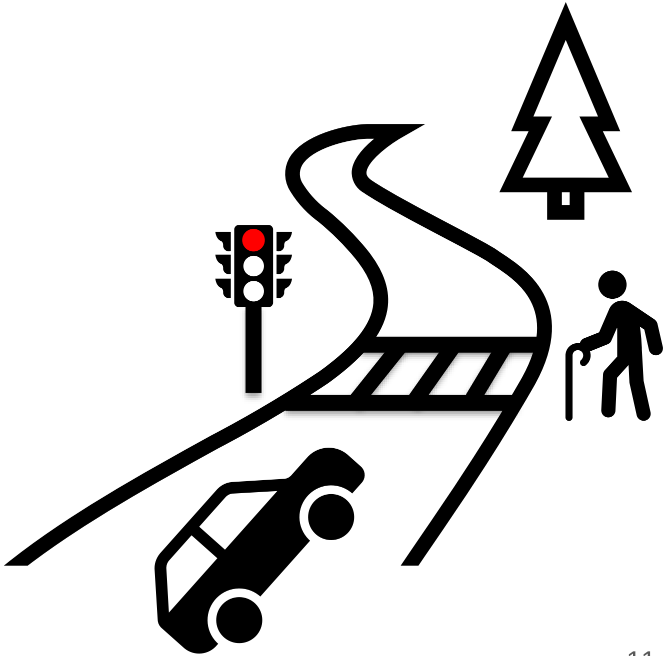




[1]

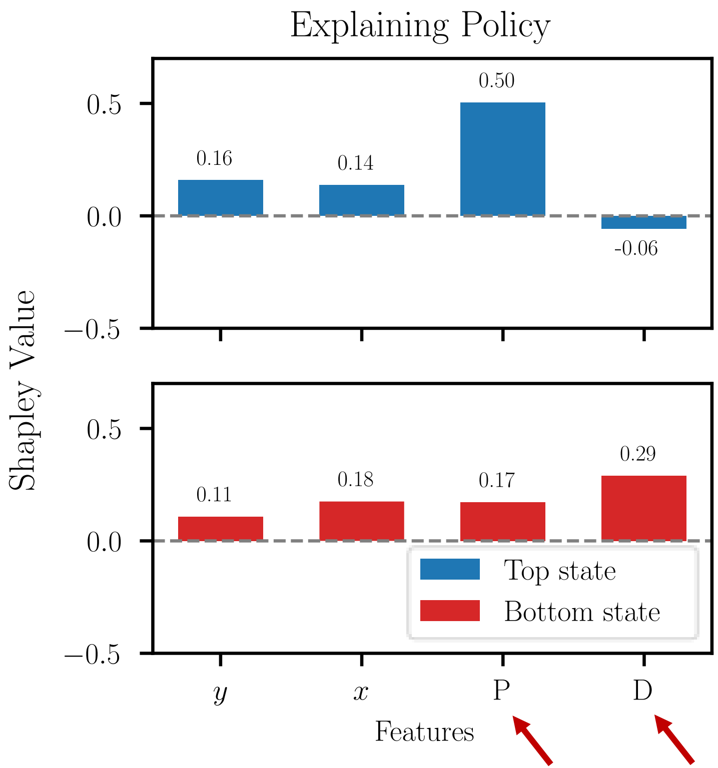
Shapley Values for Explaining Reinforcement Learning (SVERL)

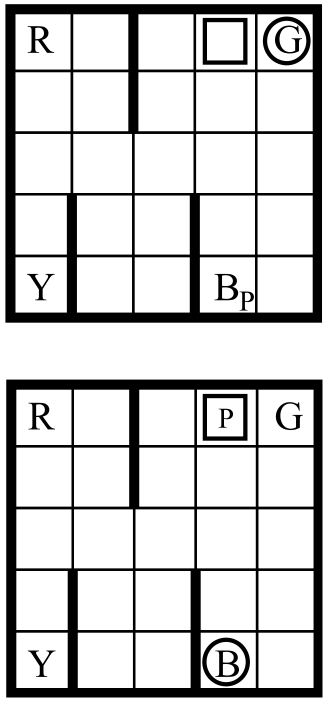
A collection of cooperative games played by features of an agent's observations whose outcomes are different aspects of agent-environment interactions.



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

Explaining Policy in Taxi





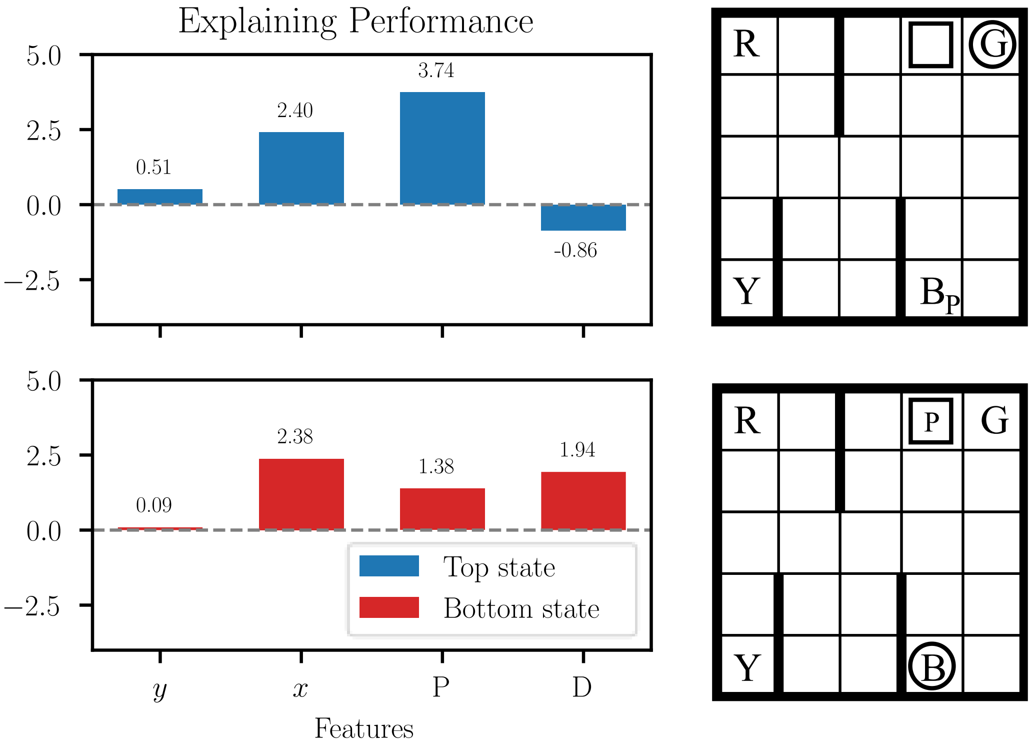
A collection of cooperative games played by features of an agent's observations whose outcomes are different aspects of agent-environment interactions.

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

Explaining Performance.

Explaining Value Prediction.

Explaining Performance in Taxi



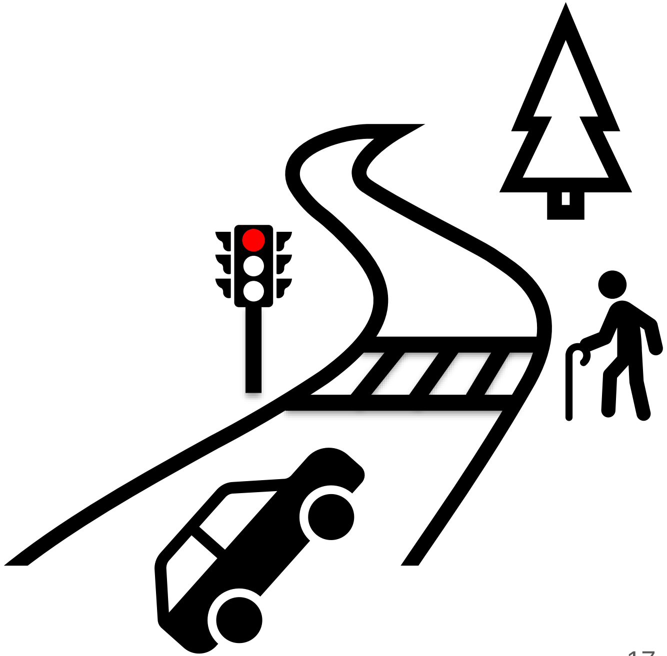
[2]

A collection of cooperative games played by features of an agent's observations whose outcomes are different aspects of agent-environment interactions.

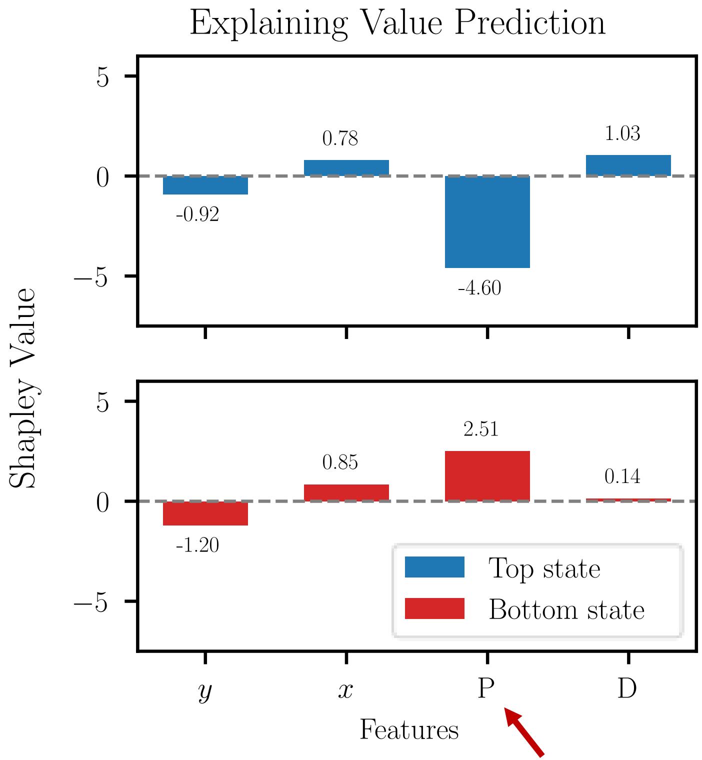
Explaining Policy. The contribution of feature values to the probability of selecting action a in state s.

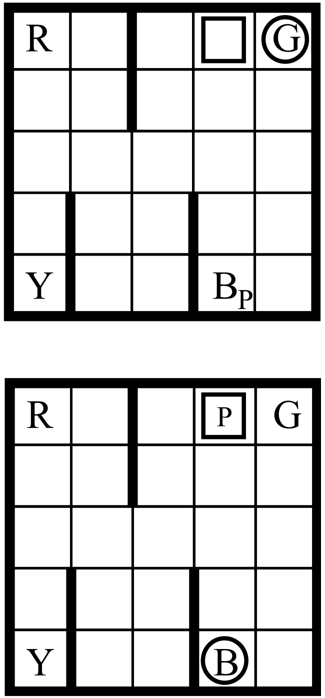
Explaining Performance. The contribution of feature values to expected return from state s.

Explaining Value Prediction.



Explaining Value Prediction in Taxi





Related Work

Feature Importance Methods

- Gradient [7]
- Perturbation [10]
- Attention [12]
- LIME [5]

Shapley Values in Supervised Learning

- SHAP [3][8]

Shapley Values in Reinforcement Learning

- SHAP applied to RL [13][17]

Shapley Values for Explaining Reinforcement Learning (SVERL)

- Explaining policy
- Explaining performance
- Explaining value prediction

Active Research

- *How to approximate SVERL in large and complicated domains.*
- *A participant-based study on using SVERL.*

Thank you for listening!

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).