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

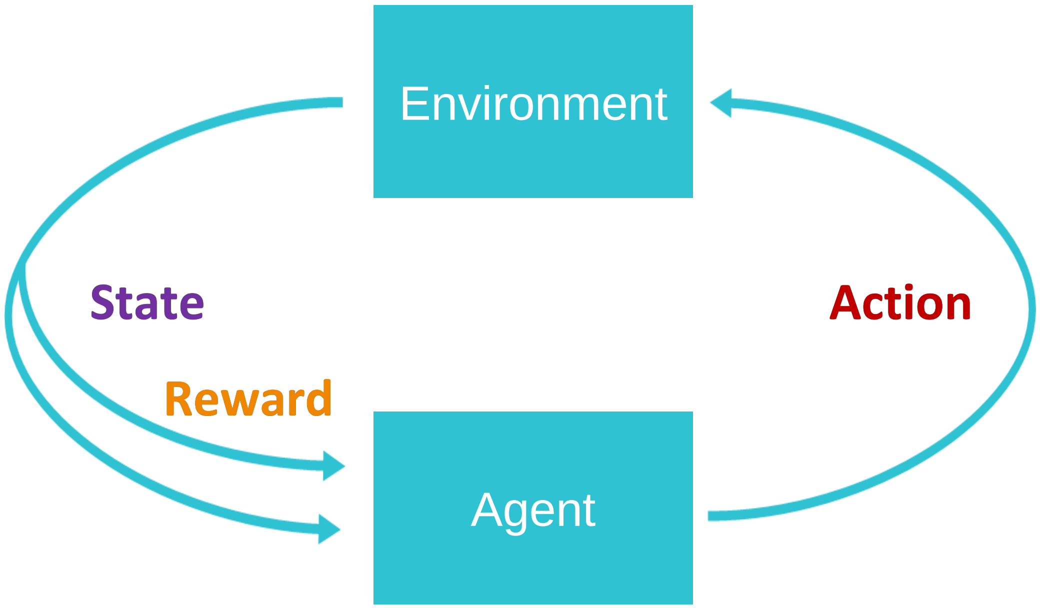
tmss20@bath.ac.uk

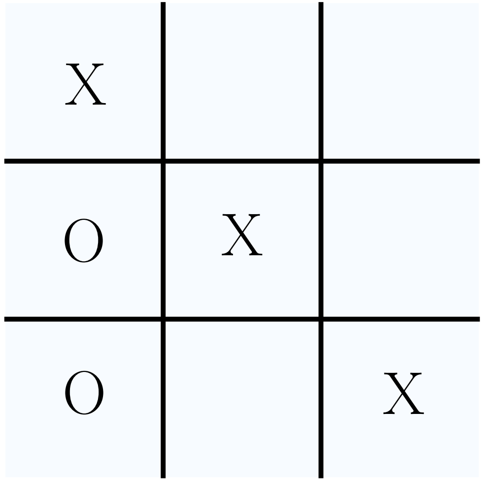
Özgür Şimşek

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.

What is reinforcement learning (RL)?





AI plays as X

RL is **<u>learning how to act</u>** through **trial and error** interaction with the world.

How to **map states to actions** to **maximise long-term reward**.

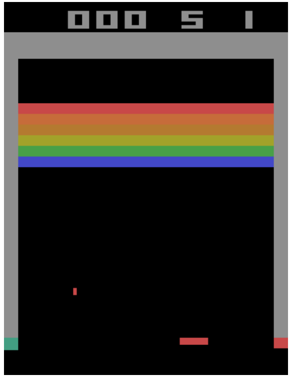
We call this mapping a **policy** π.

What can reinforcement learning do?

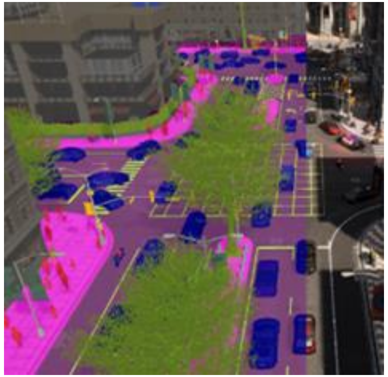
TD-Gammon (Tesauro, 1992)

AlphaGo (Silver et al., 2016)

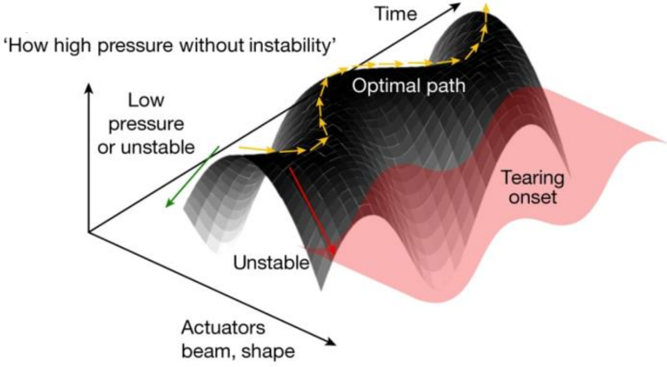
RoboCup (Riedmiller & Gabel, 2007)



Atari (Minh et al. 2015)



Autonomous Driving (ML4AD@NeurIPS)



Nuclear Fusion Reactor Control (Seo et al. 2024)

Reinforcement learning agents do not explain their decisions.

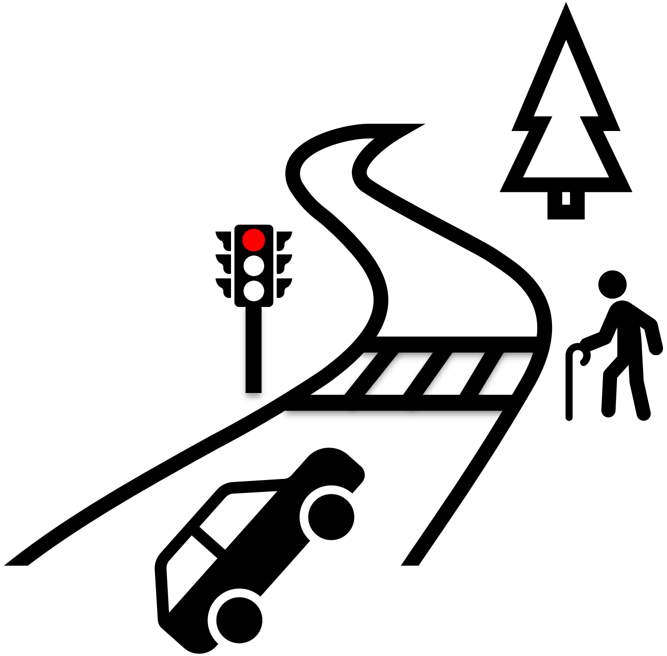
Certain features of their observations influence the behaviour of reinforcement learning agents.

Contribution: A mathematical framework for explaining agent behaviour using the influence of features.

How to compute the influence of features

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.

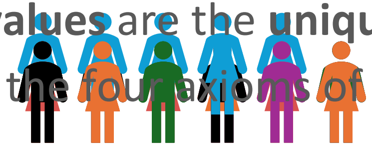


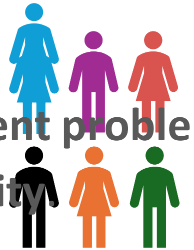
Shapley values

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

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

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





Shapley values are the **unique solution** to the **contribution assignment problem** satisfying the four axioms of **efficiency, symmetry, nullity and linearity.**

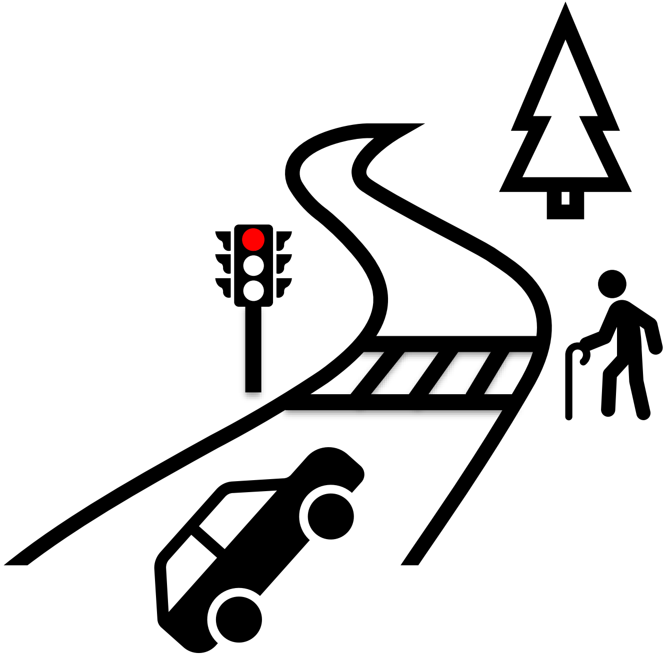
A collection of cooperative games played by features of an agent’s observation whose outcomes are different aspects of the agent-environment interaction.

Explaining Policy

Explaining Performance

Explaining Performance Prediction

A cooperative game played by the values of the features at state s, whose outcome $\pi_s^a : 2^{|\mathcal{F}|} \rightarrow \mathbb{R}$ is the probability of selecting action a at state s when only the value of features $\mathcal{C}$ are known.

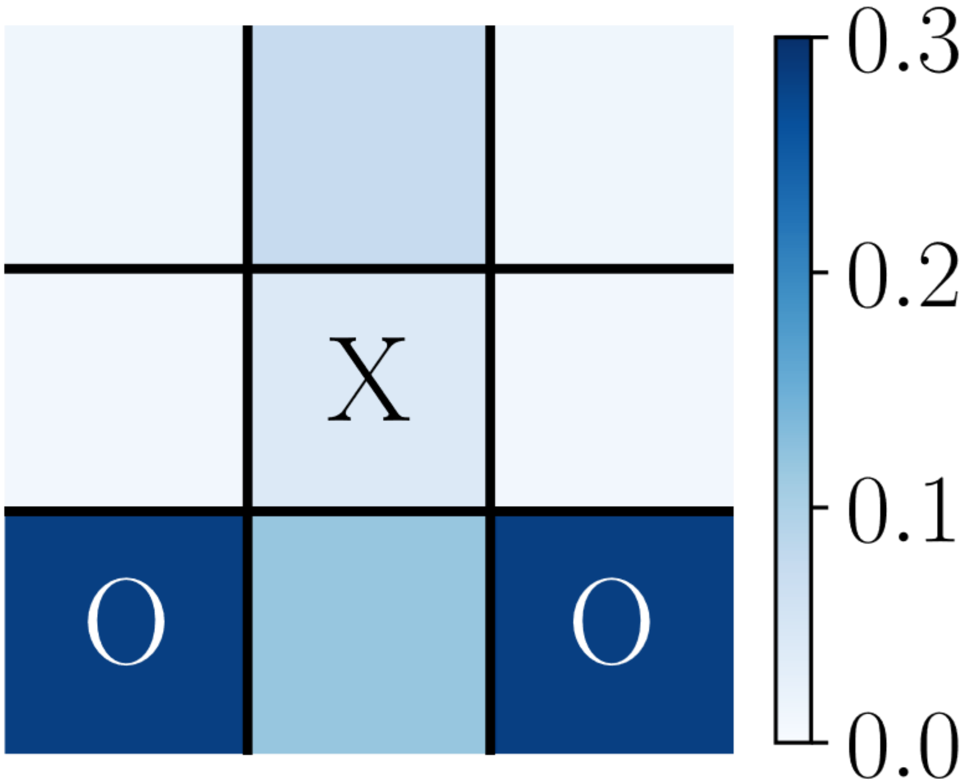


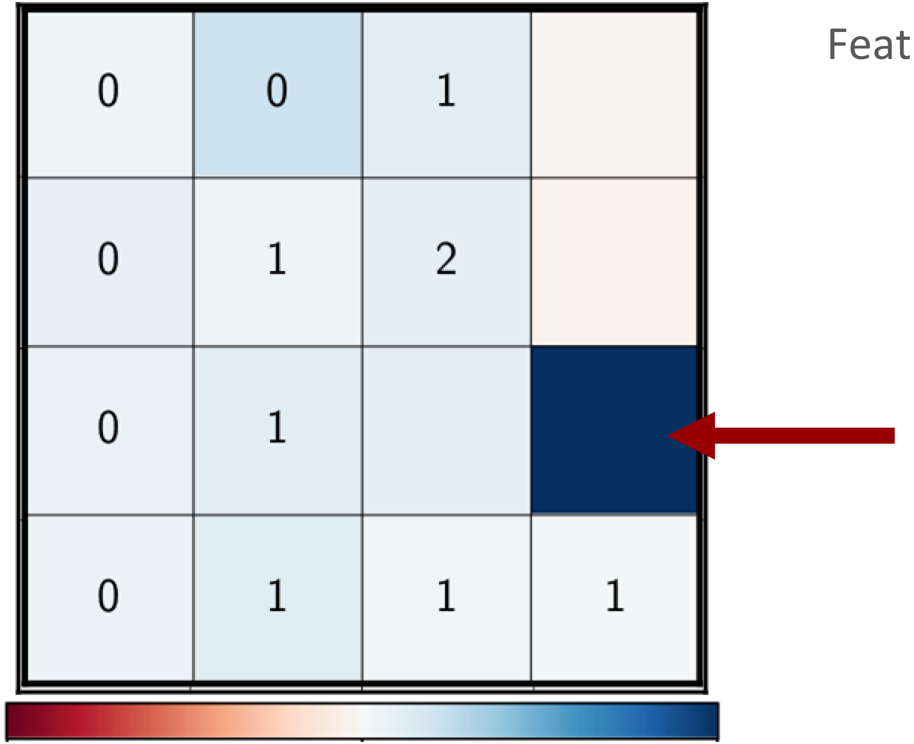
The contribution of feature values to the probability of selecting action a in state s.

Explaining Tic Tac Toe

Agent plays as X.

Features are the grid squares.





Features are the 16 grid squares.

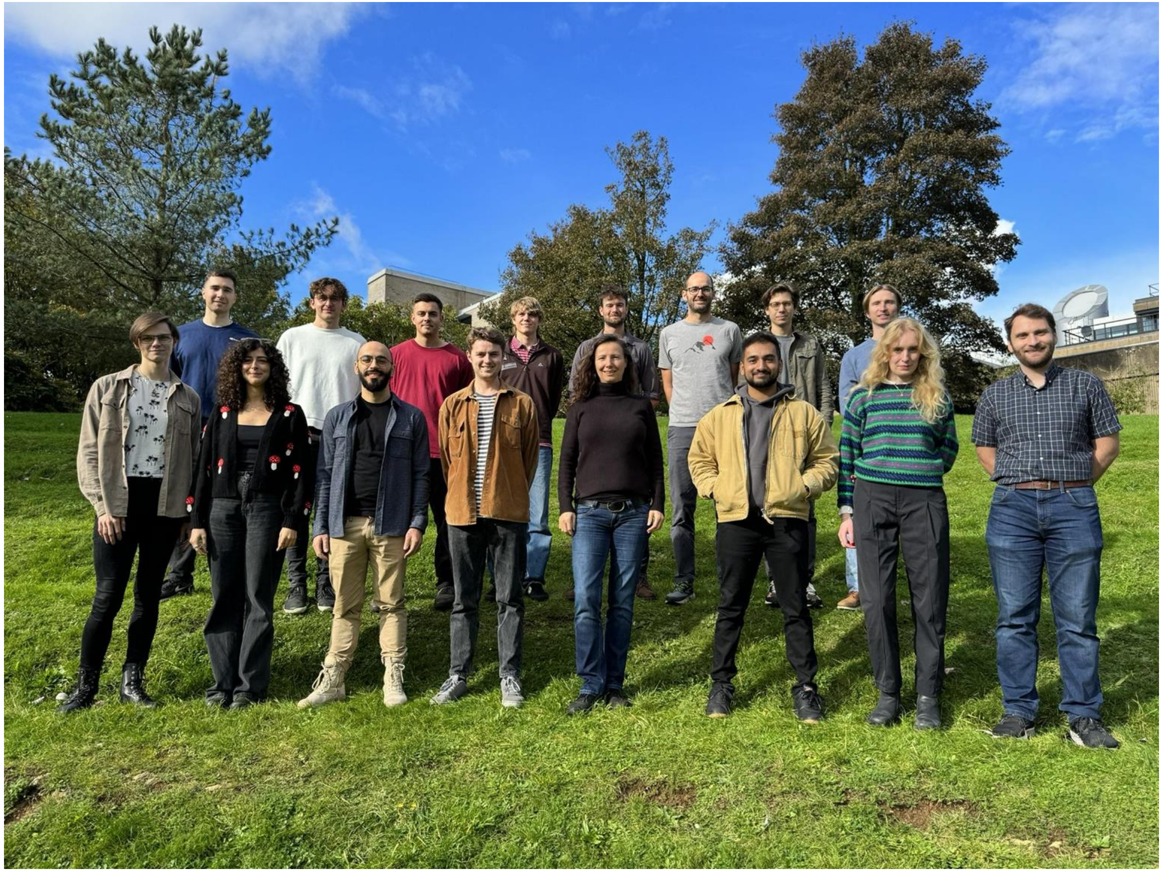
Current and future work

art-ai

UNIVERSITY OF BATH

- SVERL: The complete framework.
- How to approximate SVERL in large and complicated domains.
- Real-world applications of SVERL.

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.



Lead by Prof Özgür Şimşek

My email: djeb20@bath.ac.uk

Thanks for listening!