# EXPLAINING CONTINUAL REINFORCEMENT LEARNING AGENTS WITH SHAPLEY VALUES

Daniel Beechey and Özgür Şimşek

UNIVERSITY OF BATH · BATH RL LAB · art-ai

## Why explain decision-making?

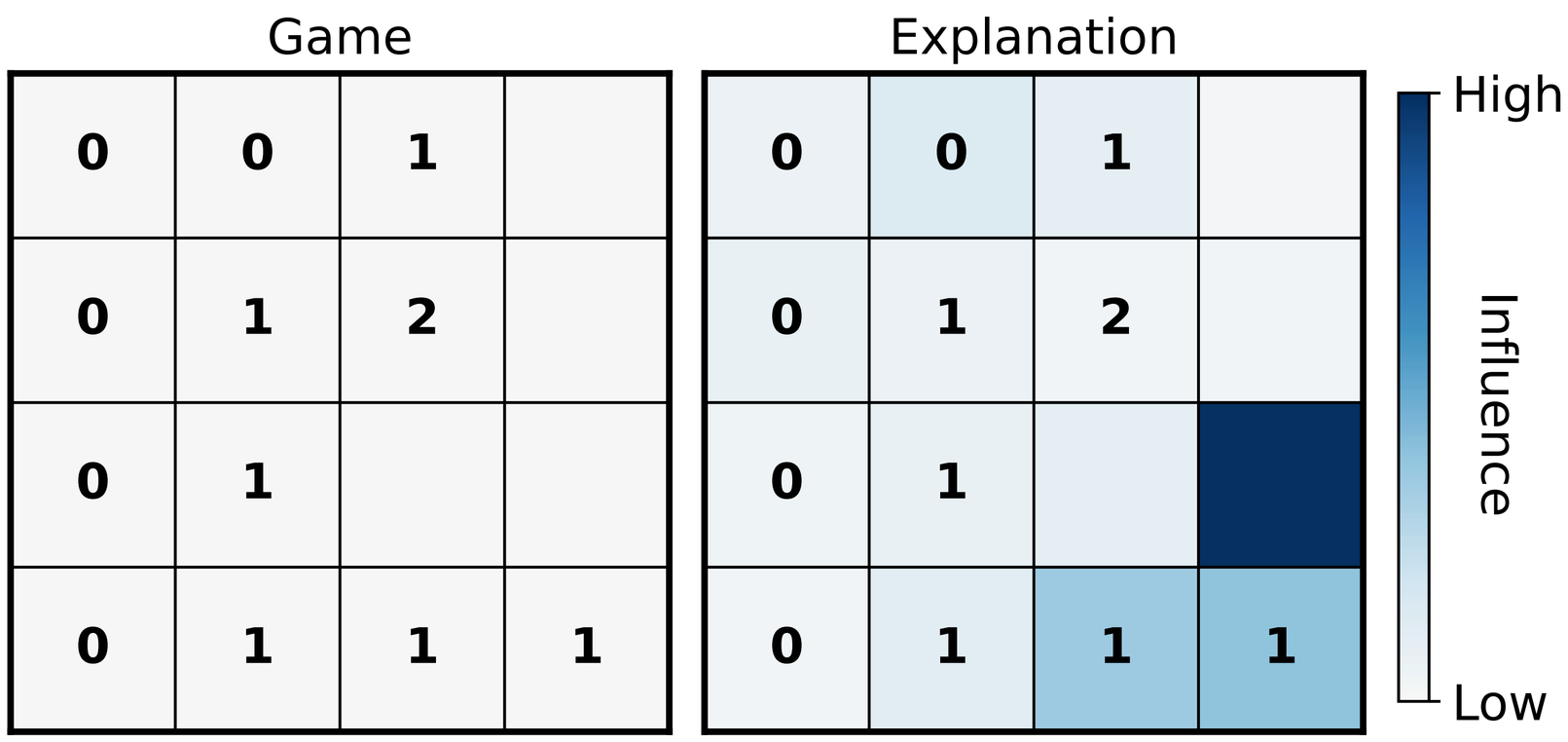
- Reinforcement learning agents typically learn to act but not to explain themselves.
- This hinders deployment in settings where accountability and trust are essential.

## How to explain decision-making

- Prior work [1, 2] attributes decisions to features of an agent's observations using Shapley values [3], a theory-driven approach to fairly assigning credit.
- But computing Shapley values exactly is infeasible in real-world settings.

## OUR CONTRIBUTION

**FastSVERL: A scalable method for explaining decision-making by attributing actions to features of an agent's observations.**




You're playing Minesweeper—what's your next move?

FastSVERL handles off-policy data and adapts to changing behaviour, enabling Shapley-based interpretability in practical reinforcement learning settings.

## How it works (if you're interested)

### What are Shapley-based explanations?

We explain an agent's decision by attributing how each feature influences the probability of taking an action, $\pi(s, a)$.

To do so, we consider how the action probability changes when different features are known or unknown. This is captured by a characteristic function $\tilde{\pi}_s^a(\mathcal{C})$, which measures the expected probability of action $a$ when only features in $\mathcal{C} \subseteq \mathcal{F}$ are known:

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

**Shapley values** assign credit to each feature based on its **average marginal contribution** across all feature subsets:

$$\phi^i(\tilde{\pi}_s^a) = \sum_{\mathcal{C} \subseteq \mathcal{F} \setminus \{i\}} \frac{|\mathcal{C}|! \cdot (|\mathcal{F}| - |\mathcal{C}| - 1)!}{|\mathcal{F}|!} [\tilde{\pi}_s^a(\mathcal{C} \cup \{i\}) - \tilde{\pi}_s^a(\mathcal{C})]$$

These values uniquely satisfy axioms formalising **fair credit assignment**.

But exact computation is infeasible in complex settings: the total cost per explanation is $\mathcal{O}(2^{|\mathcal{F}|} \cdot |\mathcal{S}|)$, $2^{|\mathcal{F}|}$ expectations over the state space $\mathcal{S}$.

**Both characteristic functions and Shapley values must be approximated.**

### Approximating the characteristic function

We approximate the characteristic function $\tilde{\pi}_s^a(\mathcal{C})$ with a model $\hat{\pi}(s, a \,|\, \mathcal{C}; \beta)$ trained to minimise prediction error:

$$\mathcal{L}(\beta) = \mathop{\mathbb{E}}_{p^\pi(s)} \mathop{\mathbb{E}}_{\text{Unif}(a)} \mathop{\mathbb{E}}_{\text{Unif}(\mathcal{C})} \left| \pi(s, a) - \hat{\pi}(s, a \,|\, \mathcal{C}; \beta) \right|^2$$

### Approximating Shapley values

We approximate the Shapley value summation with a second model $\hat{\phi}(s, a; \theta) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{F}|}$, trained to minimise a least-squares objective:

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{p^\pi(s)} \mathop{\mathbb{E}}_{\text{Unif}(a)} \mathop{\mathbb{E}}_{p(\mathcal{C})} \left| \tilde{\pi}_s^a(\mathcal{C}) - \tilde{\pi}_s^a(\emptyset) - \sum_{i \in \mathcal{C}} \hat{\phi}^i(s, a; \theta) \right|^2$$

where $\quad p(\mathcal{C}) \propto \dfrac{n - 1}{\binom{n}{|\mathcal{C}|} \cdot |\mathcal{C}| \cdot (n - |\mathcal{C}|)}$

**These models amortise the cost of Shapley value approximation across all states and actions.**

**Email:** djeb20@bath.ac.uk

**Website:** djeb20.github.io

[1] 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.
[2] Beechey, D., Smith, T. and Şimşek, Ö., 2025. A Theoretical Framework for Explaining Reinforcement Learning with Shapley Values. arXiv preprint arXiv:2505.07797.
[3] Lloyd S Shapley. A value for n-person games. Contributions to the Theory of Games, 2(28):307–317, 1953.