Explaining Reinforcement Learning with Shapley Values

Daniel Beechey

Bath Reinforcement Learning Lab

Collaborators

Thomas Smith

tmss20@bath.ac.uk

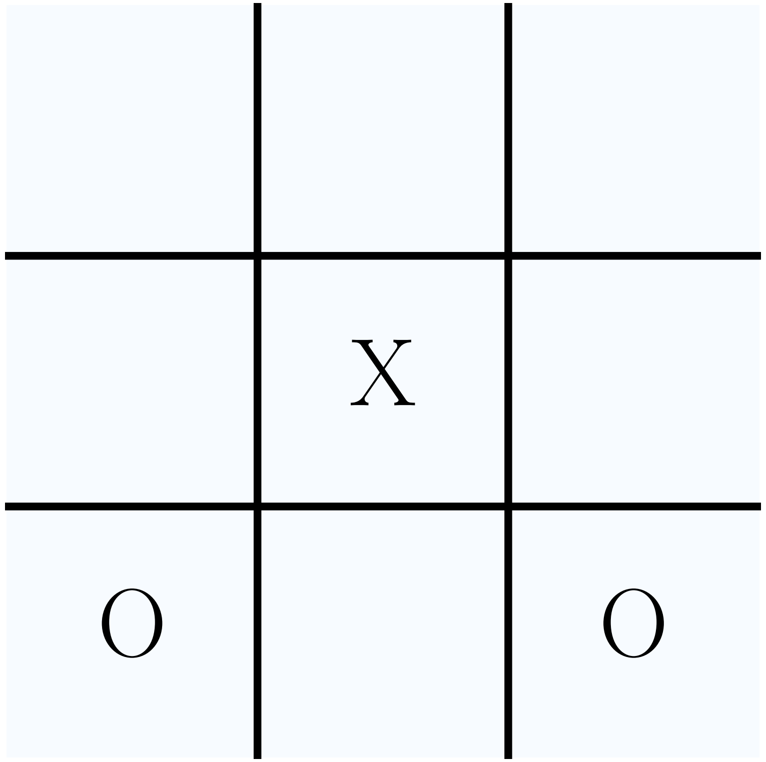
Özgür Şimşek

os435@bath.ac.uk

Beechey, D., Smith, T. M. S., and Şimşek, Ö. Explaining reinforcement learning with Shapley values. In International Conference on Machine Learning, pp. to appear. PMLR, 2023.

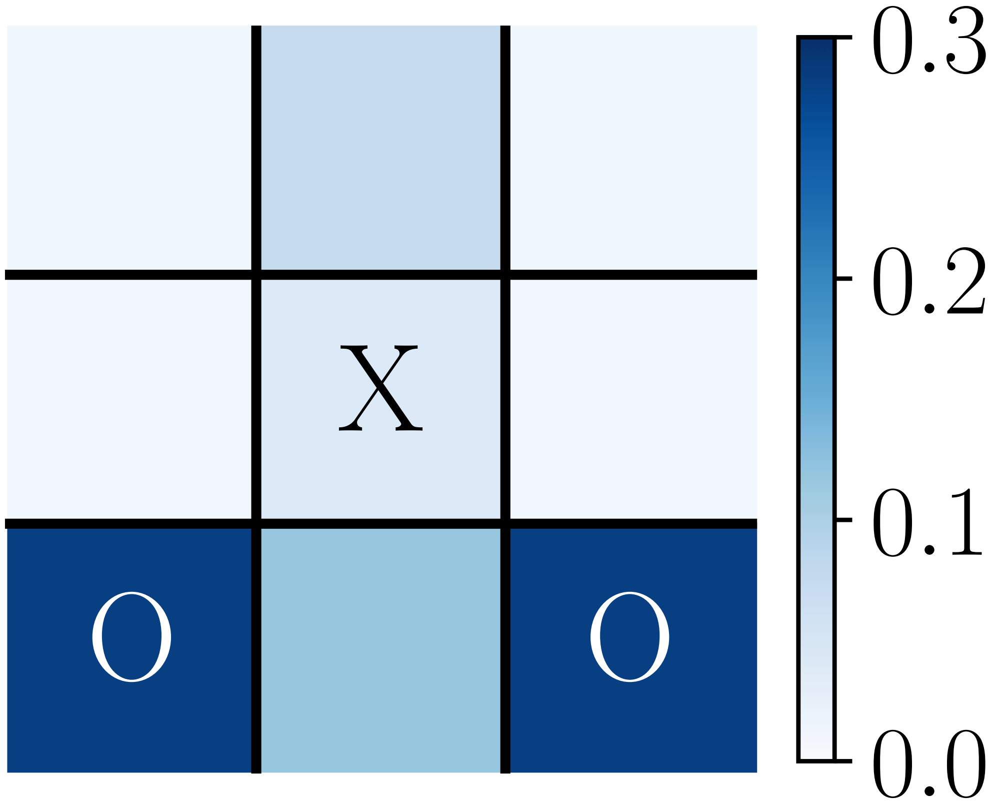
Motivation

AI playing as X in Tic-Tac-Toe.

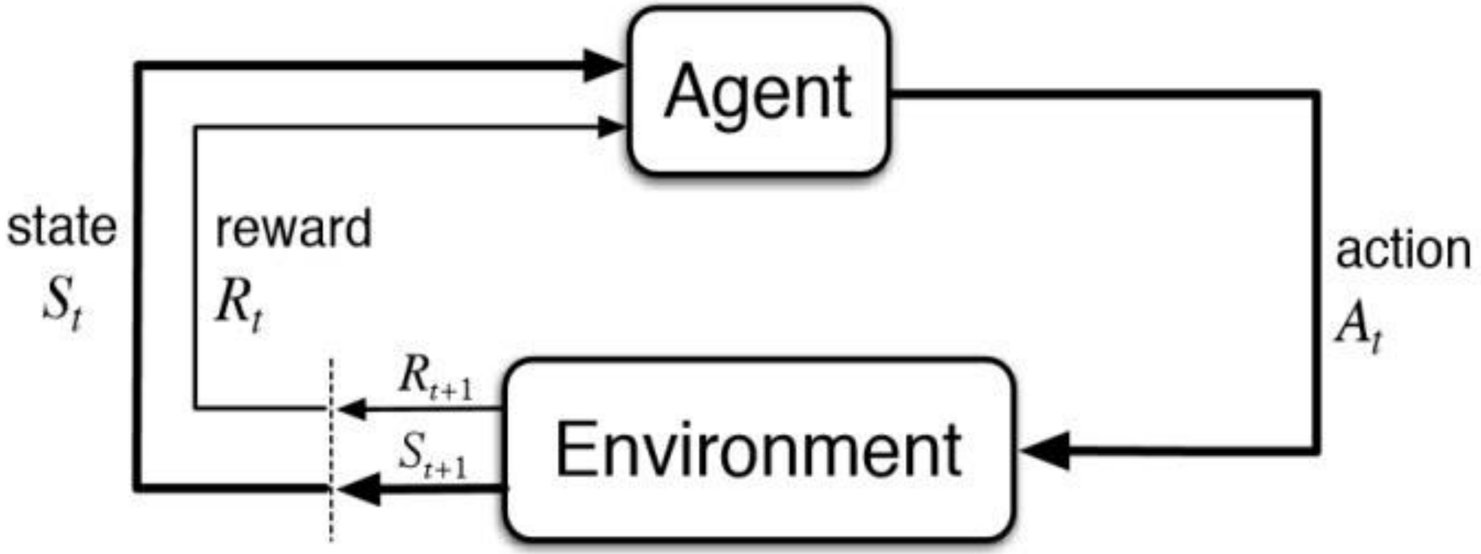


Motivation

AI playing as X in Tic-Tac-Toe.



Reinforcement Learning

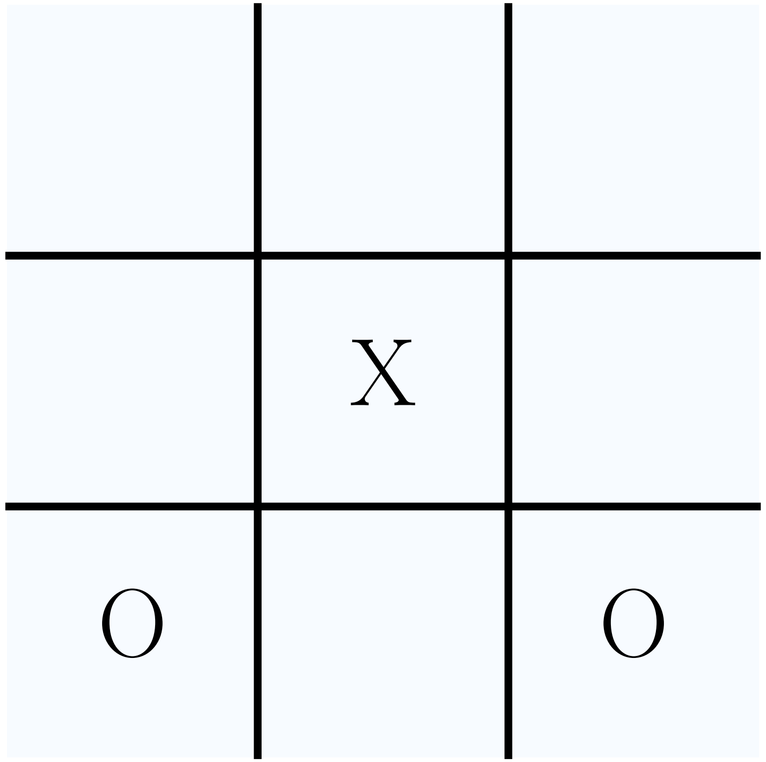


Sutton, R.S. and Barto, A.G., 2018. *Reinforcement learning: An introduction*. MIT Press, chapter 1 pp.1-13.

Reinforcement Learning

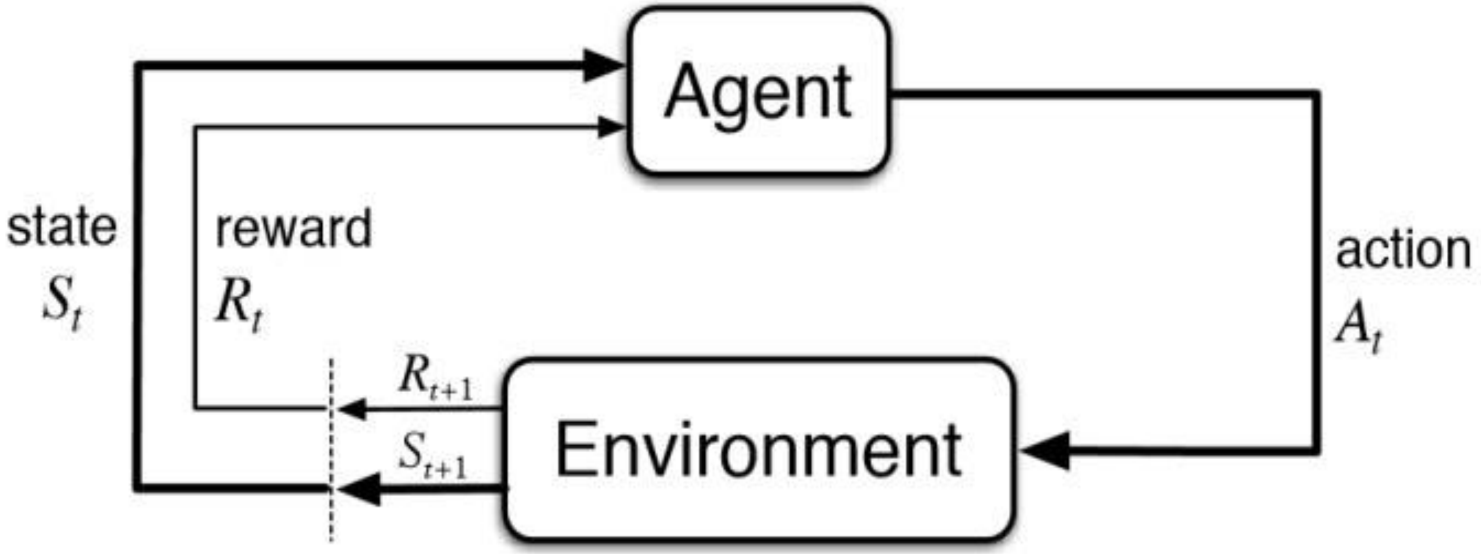
Rewards

- 1 for winning
- 0 for drawing
- -1 for losing



Sutton, R.S. and Barto, A.G., 2018. *Reinforcement learning: An introduction*. MIT Press, chapter 1 pp.1-13.

Reinforcement Learning

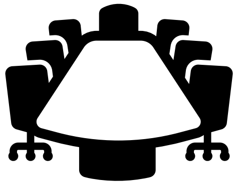


An agent selects actions according to policy π.

This policy is often defined using the value function $V^{\pi}(s) = \mathbb{E}[\sum_{t=1}^{\infty} r_t | s_0 = s]$.

Sutton, R.S. and Barto, A.G., 2018. *Reinforcement learning: An introduction*. MIT Press, chapter 1 pp.1-13.

Shapley Values

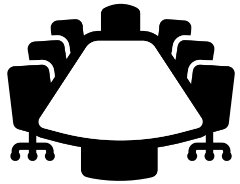


A **cooperative game** is a set of players $\mathcal{F}$ and a characteristic value function $v: 2^{|\mathcal{F}|} \rightarrow \mathbb{R}$.

Shapley values are the **unique** solution to a set of four mathematical axioms that specify the fair contributions of players to the outcome of a cooperative game.

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

Shapley Values



- **Axiom 1 (Efficiency)** $\sum_{i \in \mathcal{F}} \phi_i(v) = v(\mathcal{F})$
- **Axiom 2 (Nullity)** $\phi_i(v) = 0 \quad \text{if} \quad v(C \cup \{i\}) = v(C) \qquad \forall C \subseteq \mathcal{F} \setminus \{i\}$
- **Axiom 3 (Symmetry)** $\phi_i(v) = \phi_j(v) \quad \text{if} \quad v(C \cup \{i\}) = v(C \cup \{j\}) \qquad \forall C \subseteq \mathcal{F} \setminus \{i, j\}$

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

Shapley Values for Explaining Reinforcement Learning (SVERL)

Explaining the Value Function

Characteristic value function: $$v^{\hat{V}}(\mathcal{C}) := \hat{V}^{\pi}_{\mathcal{C}}(s) = \sum_{s' \in \mathcal{S}} p^{\pi}(s'|s_{\mathcal{C}}) \hat{V}^{\pi}(s')$$

Explaining the Policy

Characteristic value function: $$v^{\pi}(\mathcal{C}) := \pi_{\mathcal{C}}(a|s) = \sum_{s' \in \mathcal{S}} p^{\pi}(s'|s_{\mathcal{C}}) \pi(a|s')$$

Explaining Performance (SVERL-Performance)

Local characteristic value function: $$v^{\text{local}}(\mathcal{C}) := \mathbb{E}_{\hat{\pi}}\left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s\right] \quad \text{where} \quad \hat{\pi}(a_t|s_t) = \begin{cases} \pi_{\mathcal{C}}(a_t|s_t) & \text{if } s_t = s \\ \pi(a_t|s_t) & \text{otherwise} \end{cases}$$

Shapley Values for Explaining Reinforcement Learning (SVERL)

Explaining the Value Function

Explains the predictions of the value function under the assumption that all features will be observed by the agent when acting in the environment.

Explaining the Policy

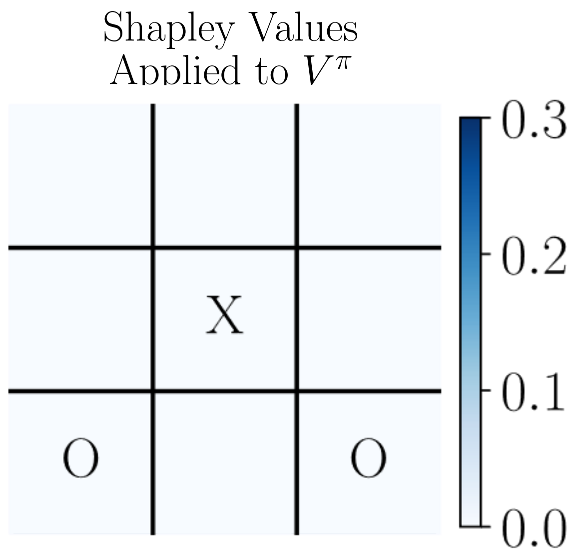
Explains the probability of selecting each action.

Explaining Performance (SVERL-Performance)

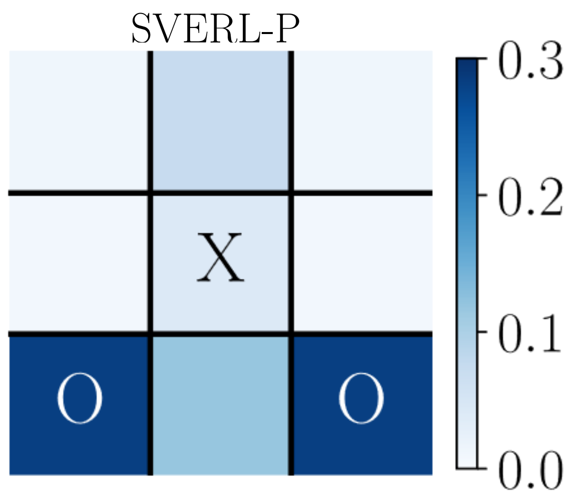
Explains agent performance from state s.

Explaining Tic-Tac-Toe

Shapley values applied to V^{π} show the contributions of features to the value function's predictions.

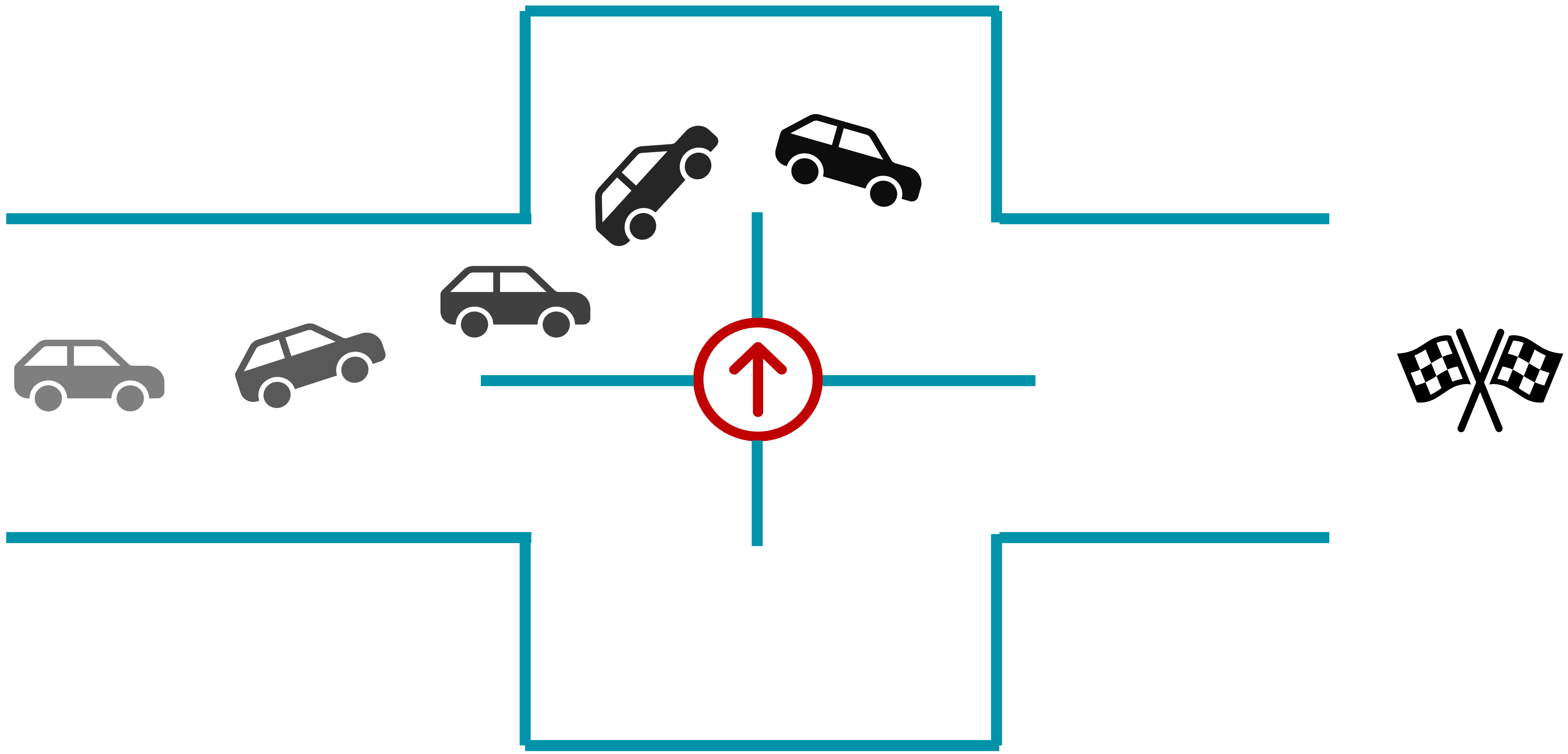


SVERL-Performance shows the contributions of features to agent performance.

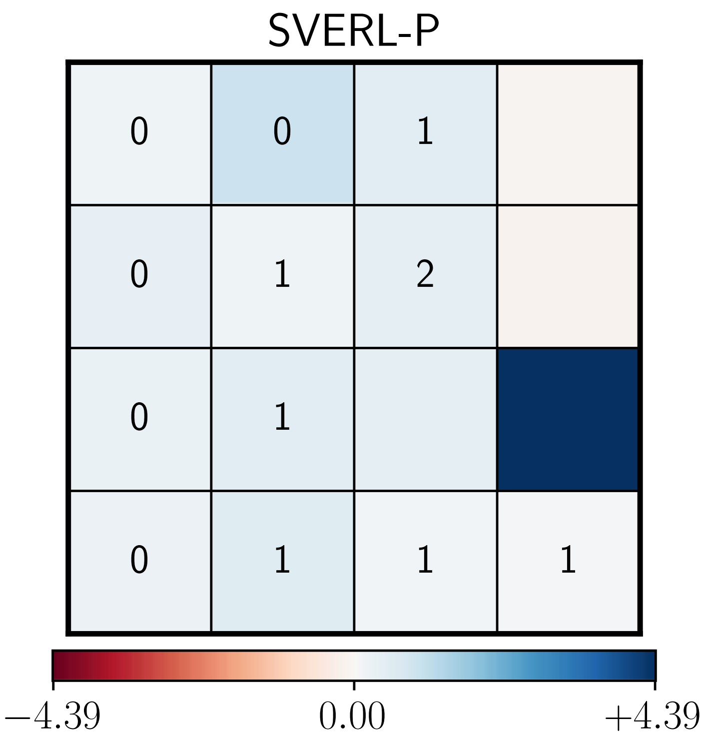


Shapley Values Applied to π

Explains the behaviour of an agent, but more is to be understood about agent performance.



Explaining Performance in Minesweeper



Features are the 16 grid squares.

One square contributes the most to performance.

Thank you for listening!

Beechey, D., Smith, T. M. S., and Şimşek, Ö. Explaining reinforcement learning with Shapley values. In International Conference on Machine Learning, pp. to appear. PMLR, 2023.