# Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes (Kenny, E. M., Tucker, M., and Shah, J., 2023)

Thorben Klamt - 27.06.2024

Advanced Topics in Reinforcement Learning, Theresa Eimer, Prof. Dr. rer. nat. Marius Lindauer,
Gottfried Wilhelm Leibniz University Hannover

# Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes (2023)

## Prototype-Wrapper Network

- enhances interpretable-by-design deep RL models (and Meta deep RL models)
- uses human-friendly prototypes to make decisions, maintaining performance of black-box models
- clearer, more human-understandable reasoning process

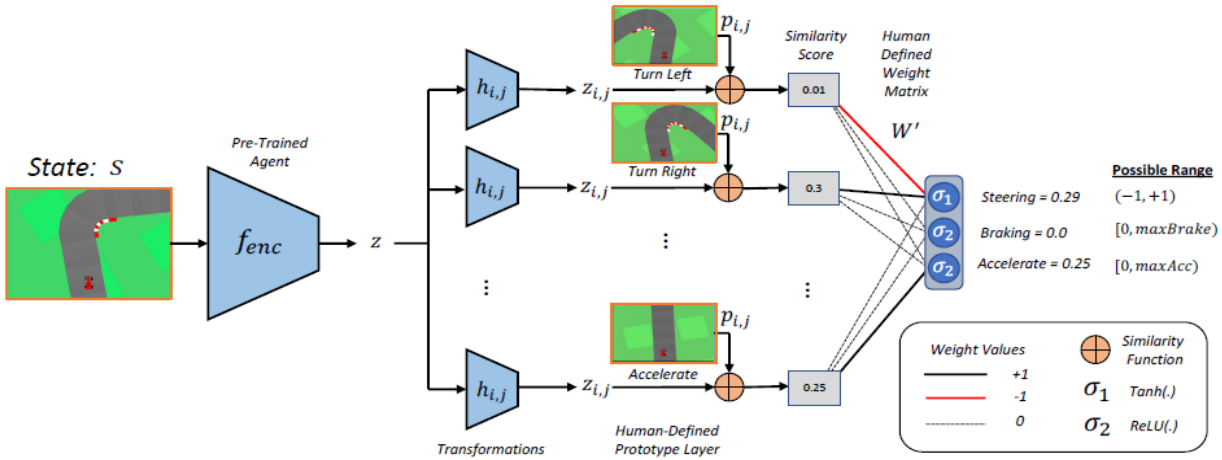


Figure 1: Prototype-Wrapper Network in Car Racing from OpenAI's gym. A state is encoded as z, transformed, and compared to human prototypes, influencing each output action.

# Towards Interpretable Deep Reinforcement Learning with Human-Friendly Prototypes (2023)

- performance comparable to black-box models across multiple domains (e.g., car racing, Atari Pong).
- maintains reward and accuracy metrics
- improved user ability to predict model failures
- enhanced trust calibration

- may not scale well
- lacks extensive testing
- user study might be quite specific
- lacks analysis on performance trade-offs

| | Atari Pong | | Lunar Lander | |
|---|---|---|---|---|
| Method | Reward | Accuracy | Reward | Accuracy |
| PW-Net | **10.72 ± 0.26** | **88.93 ± 0.00** | **216.94 ± 16.92** | **97.63 ± 0.00** |
| VIPER | N/A | N/A | -408.81 ± 60.98 | 59.26 ± 1.01 |
| PW-Net* | 8.85 ± 1.69 | 84.84 ± 0.76 | 124.54 ± 120.53 | 88.67 ± 0.01 |
| k-means | -21.00 ± 0.00 | 11.79 ± 4.15 | -419.46 ± 119.08 | 10.10 ± 5.87 |
| Black-Box | 11.94 ± 0.16 | N/A | 212.94 ± 2.63 | N/A |

Table 1: Continuous Action Spaces: PW-Net matches black-box performance, unlike other baselines.

| | Atari Pong | | Lunar Lander | |
|---|---|---|---|---|
| Method | Reward | Accuracy | Reward | Accuracy |
| PW-Net | **10.72 ± 0.26** | **88.93 ± 0.00** | **216.94 ± 16.92** | **97.63 ± 0.00** |
| VIPER | N/A | N/A | -408.81 ± 60.98 | 59.26 ± 1.01 |
| PW-Net* | 8.85 ± 1.69 | 84.84 ± 0.76 | 124.54 ± 120.53 | 88.67 ± 0.01 |
| k-means | -21.00 ± 0.00 | 11.79 ± 4.15 | -419.46 ± 119.08 | 10.10 ± 5.87 |
| Black-Box | 11.94 ± 0.16 | N/A | 212.94 ± 2.63 | N/A |

Table 2: Baseline comparisons in discrete action spaces show PW-Net consistently performs well.

Kenny, Eoin M., Mycal Tucker, and Julie Shah. "Towards interpretable deep reinforcement learning with human-friendly prototypes." *The Eleventh International Conference on Learning Representations*. 2023