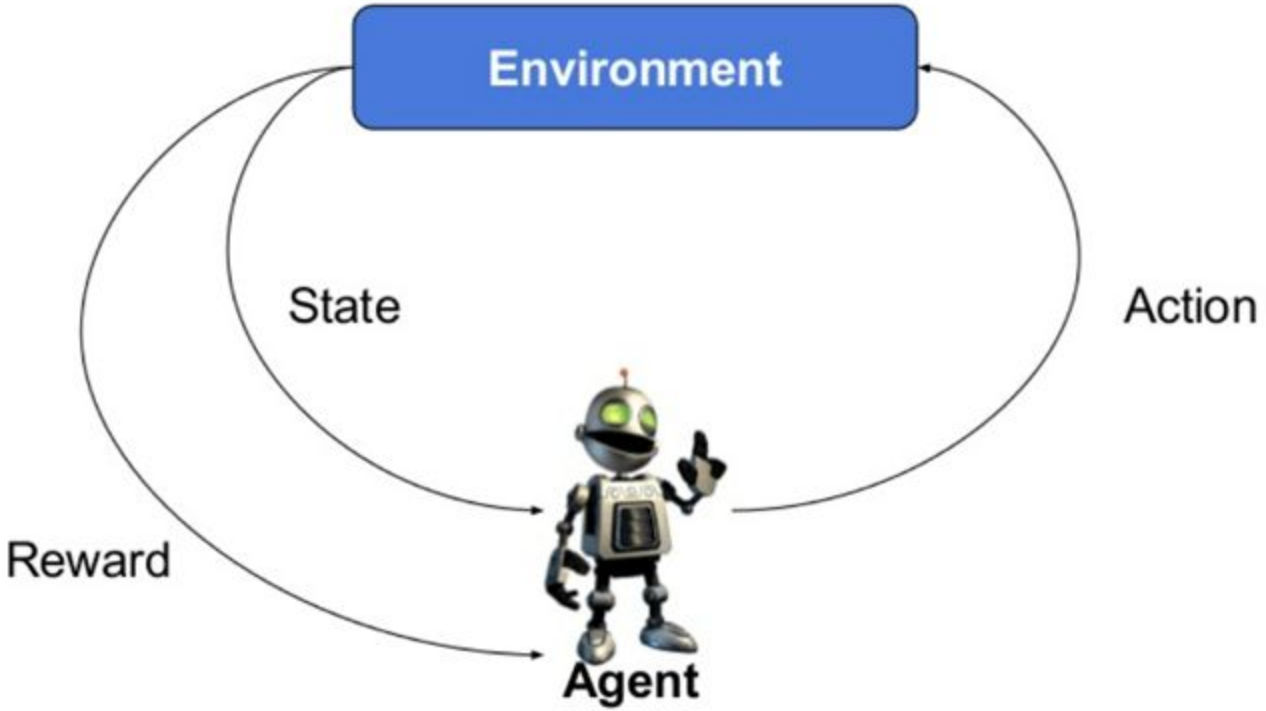
Review on XRL

Explainable Reinforcement Learning

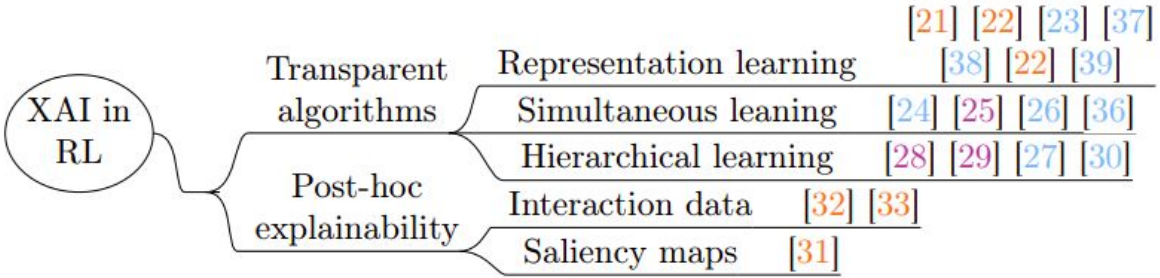
Typical RL scenario



Explainability in Deep Reinforcement Learning

Alexandre Heuillet[a,1], Fabien Couthouis[b,1], Natalia Díaz-Rodríguez[c,*]

[a] *ENSEIRB-MATMECA, Bordeaux INP, 1 avenue du Docteur Albert Schweitzer, 33400 Talence, France*
[b] *ENSC, Bordeaux INP, 109 avenue Roul, 33400 Talence, France*
[c] *ENSTA Paris, Institut Polytechnique Paris, Inria Flowers Team, 828 boulevard des Maréchaux, 91762 Palaiseau, France*



Transparent Algorithms

Explanation through representation learning

Simultaneous learning of the explanation and the policy

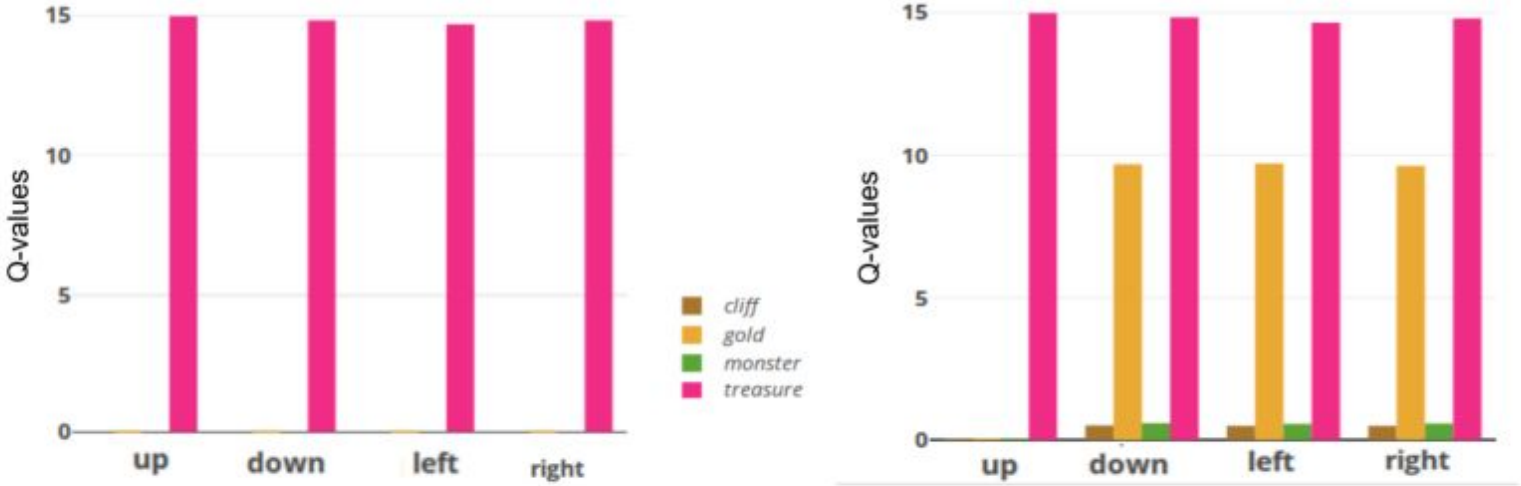


Figure 3: **Left** Reward Decompositions for DQN. **Right** Hybrid Reward Architecture (HRA) at cell (3,4) in Cliffworld. HRA predicts an extra "gold" reward for actions which do not lead to a terminal state. Reproduced with permission of Zoe Juozapaitis [24].

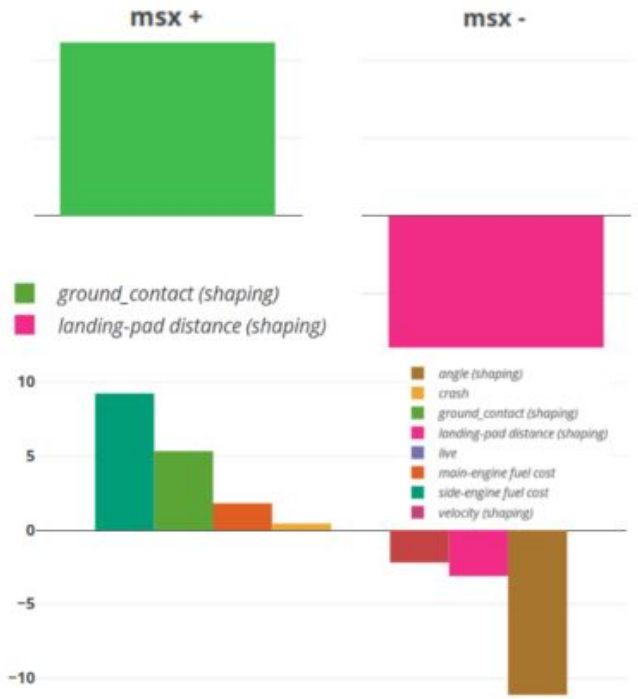


Figure 4: **Top** Minimal Sufficient Explanations (MSX) (fire down engine action vs. *do nothing* action) for decomposed reward DQN in *Lunar Lander* environment near landing site. The shaping rewards dominate decisions. **Bottom** RDX (noop vs. fire-main-engine) for HRA in Lunar Lander before a crash. The RDX shows that noop is preferred to avoid penalties such as fuel cost. Reproduced with permission of Zoe Juozapaitis [24].

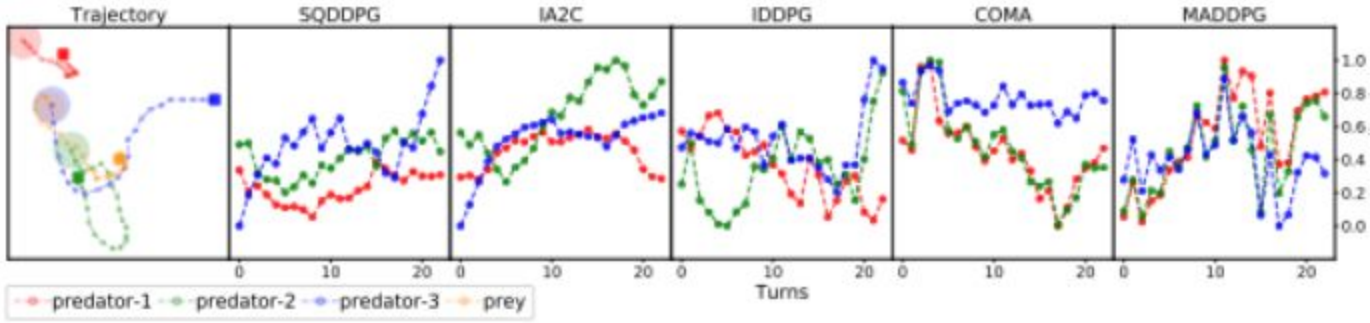


Figure 6: Credit assignment to each predator for a fixed trajectory in prey and predator task (Multiagent Particles environment [70]). **Left figure**: Trajectory sampled by an expert policy. The square represents the initial position whereas the circle indicates the final position of each agent. The dots on the trajectory indicate each agent's temporary positions. **Right figures**: normalized credit assignments generated by different multiagent RL algorithms according to this trajectory. SQDDPG presents fairer credit assignments in comparison with other methods. Reproduced with permission of Jianhong Wang [26].

Explanation through hierarchical goals

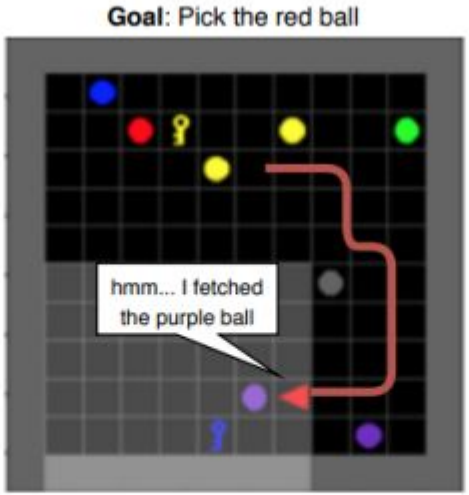


Figure 8: MiniGrid environment [74], where the agent is instructed through a textual string to pick up an object and place it next to another one. The model learns to represent the achieved goal (e.g. "Pick the purple ball") via language. As this achieved goal differs from the initial goal ("Pick the red ball"), the goal mapper relabels the episode, and both trajectories are appended to the replay buffer. Reproduced with permission of M. Seurin [28].

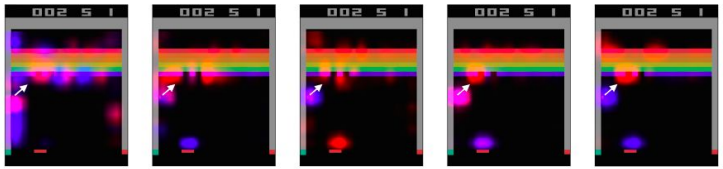
Post-hoc explainability

EXPLANATION THROUGH SALIENCY MAPS

Visualizing and Understanding Atari Agents

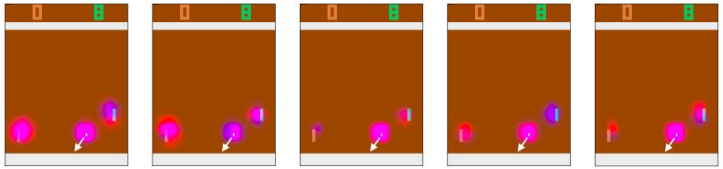
Sam Greydanus [1] Anurag Koul [1] Jonathan Dodge [1] Alan Fern [1]

Visualizing and Understanding Atari Agents



(a) Breakout: learning what features are important.

(b) Breakout: learning a tunneling strategy.



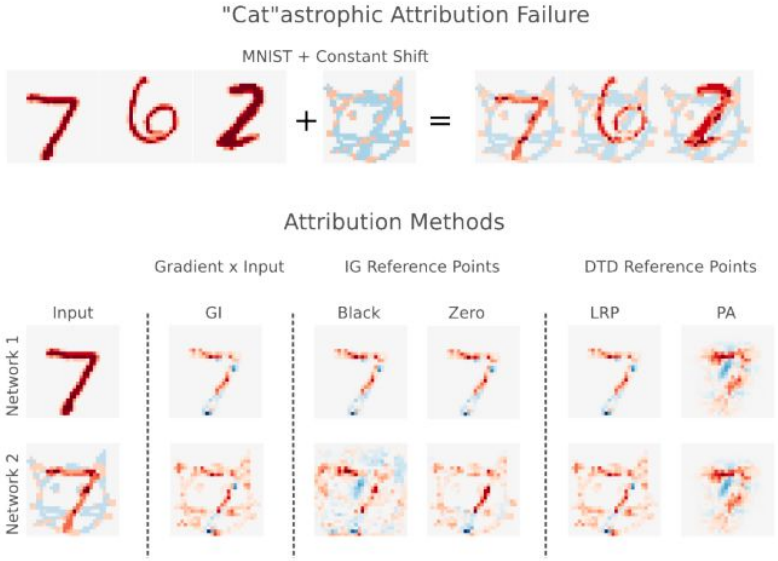
(c) Pong: learning a kill shot.

THE (UN)RELIABILITY OF SALIENCY METHODS

Pieter-Jan Kindermans[†], **Sara Hooker**[†], **Julius Adebayo**
Google Brain*
{pikinder, shooker}@google.com

Maximilian Alber, Kristof T. Schütt, Sven Dähne
TU-Berlin

Dumitru Erhan, Been Kim
Google Brain

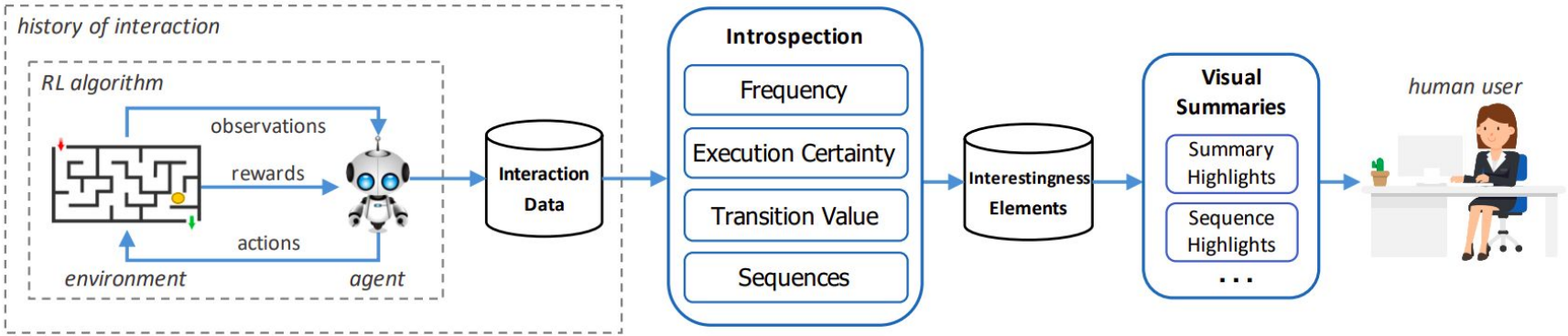


EXPLANATION THROUGH INTERACTION DATA

Interestingness Elements for Explainable Reinforcement Learning: Understanding Agents' Capabilities and Limitations*

Pedro Sequeira and Melinda Gervasio
SRI International
333 Ravenswood Avenue, Menlo Park, CA 94025, United States
pedro.sequeira@sri.com, melinda.gervasio@sri.com

August 20, 2020



OTHER METHODS

Explainability of DNNs

Compositionality as a proxy tool to improve understandability

Improving trust via imitation learning

Transparency-oriented explanation building

Conclusions and further research