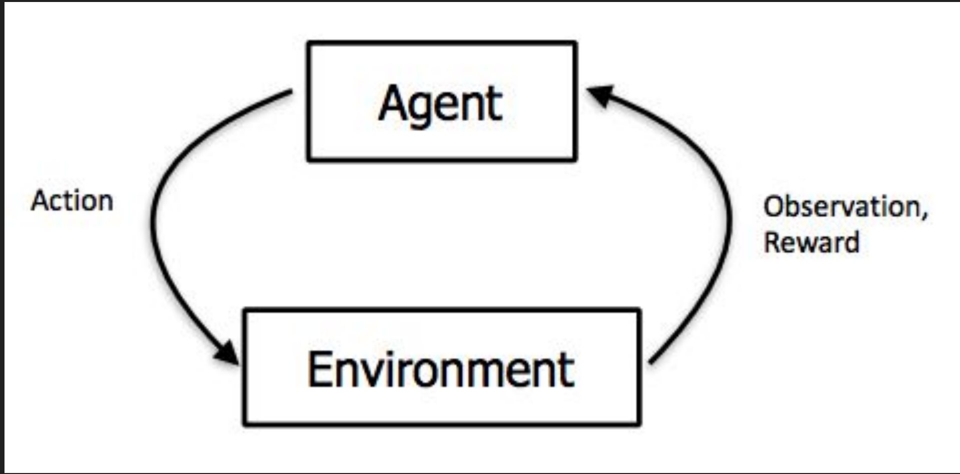
Explainability in RL

Based on the XRL review article of:
Alexandre Heuillet, Fabien Couthouis, Natalia Díaz-Rodríguez

Explainable Machine Learning course,
21.05.2020

1 minute RL primer



Actor

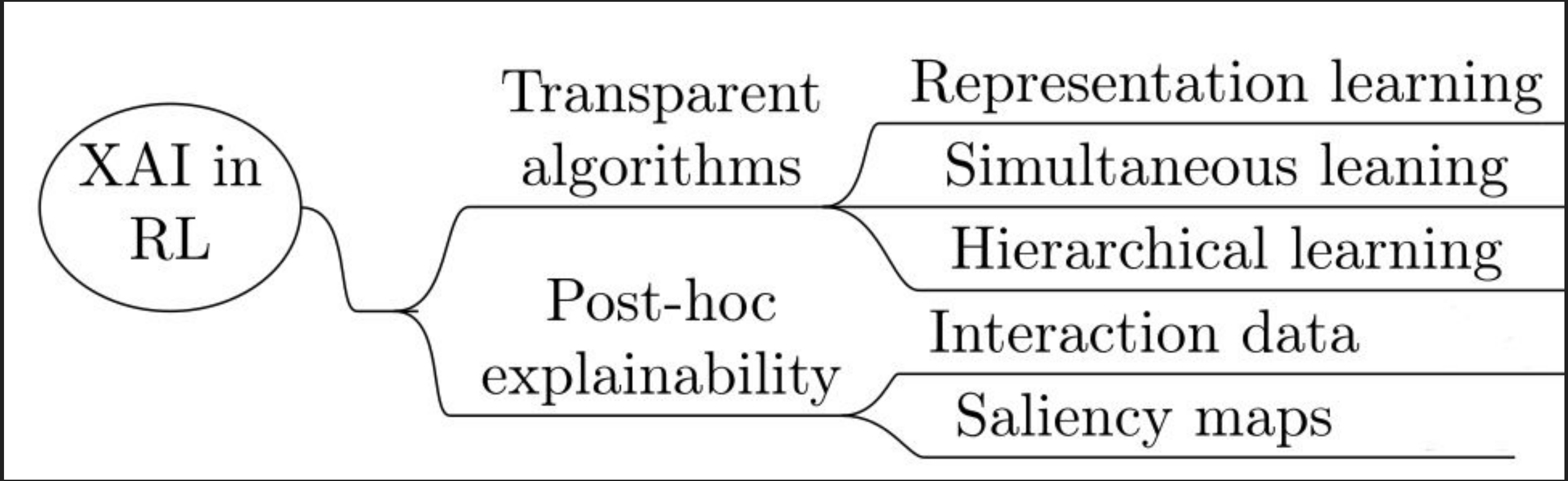
$$a_t \sim \pi(\cdot|s_t)$$

Critic

$$Q^{\pi}(s,a) = \mathop{\mathrm{E}}_{\tau\sim\pi}\left[R(\tau)\,|s_0 = s, a_0 = a\right]$$

- Actor and critic functions usually approximated by some parametric models
- Usually trained by:
  - regressing towards critic consistency equations (Belman backups) or
  - directly optimizing for expected sum of rewards of actor (policy gradients)

Taxonomy of explanation methods in RL



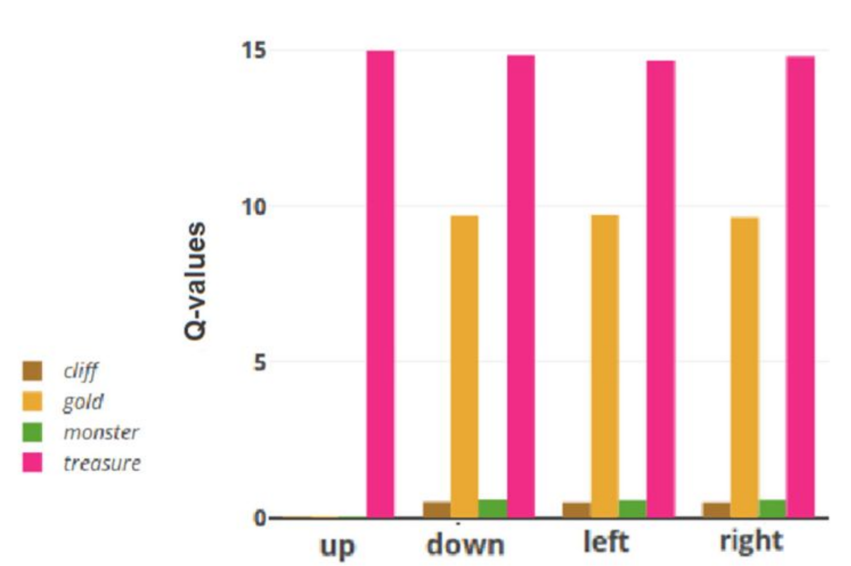
Explainable RL via Reward Decomposition

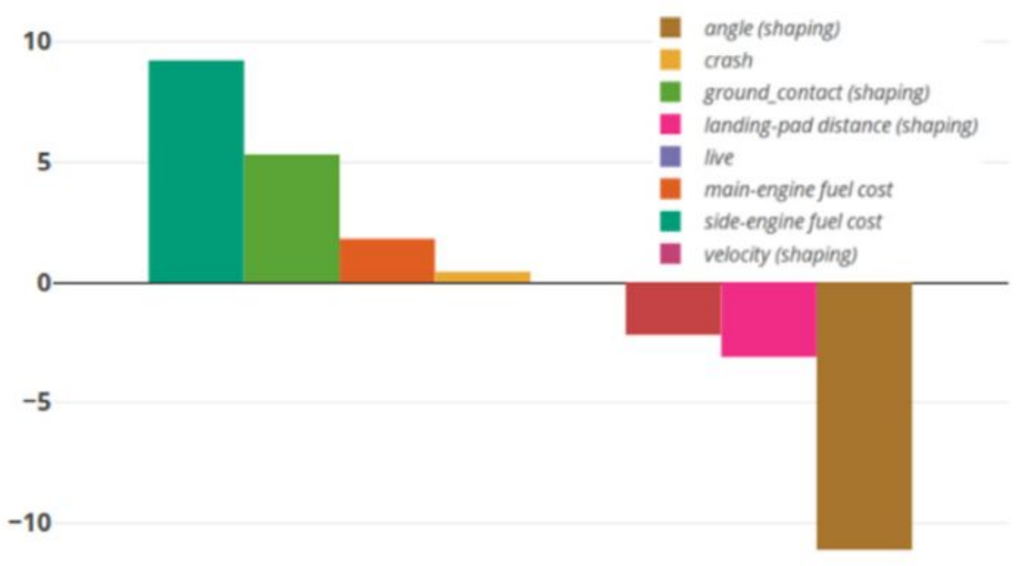
- Manually divide rewards gained into c interesting categories
- Separately learn critic for each category
- local explanations as expected future reward for each category
- can be useful for debugging reward shaping

$$L(\theta_c) = \sum_{i=1}^{k} (y_{c,i} - Q_c(s_i, a_i; \theta_c))^2$$

$$y_{c,i} = \begin{cases} r_c, & \text{for terminal } s'_i \\ r_c + \gamma Q_c(s'_i, a_i^+; \theta'_c), & \text{for non-terminal } s'_i \end{cases}$$

$$a_i^+ = \arg\max_{a'} \sum_{c \in C} Q_c(s'_i, a', \theta'_c)$$


15
10
5
0
Q-values
cliff
gold
monster
treasure
up
down
left
right


10
5
0
−5
−10
angle (shaping)
crash
ground_contact (shaping)
landing-pad distance (shaping)
live
main-engine fuel cost
side-engine fuel cost
velocity (shaping)

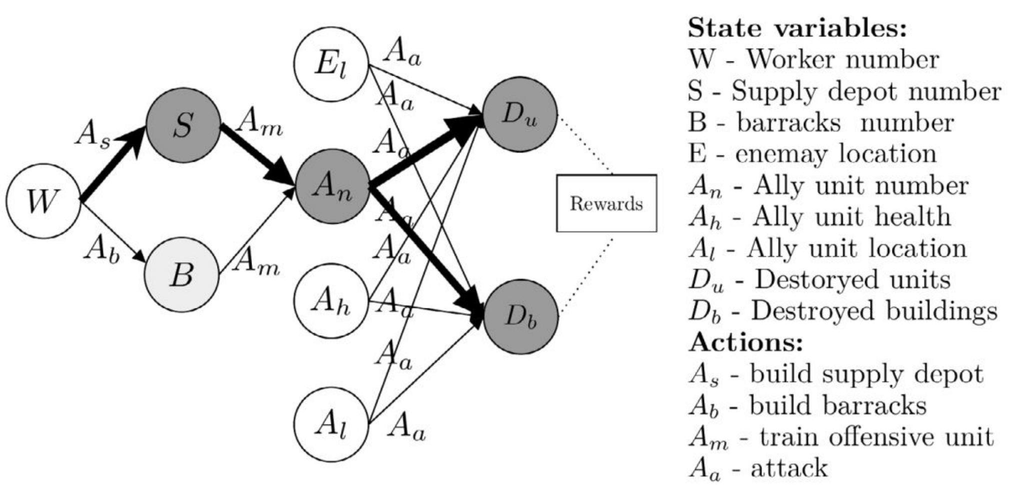
Explainable Reinforcement Learning Through a Causal Lens

- Learns jointly specific model of the environment (casual model):
  - given state – how action and features influence themselves
- To generate explanation:
  - traverse graph backward from reward nodes

Formally, a *signature* S is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is the set of exogenous variables, $\mathcal{V}$ the set of endogenous variables, and $\mathcal{R}$ is a function that denotes the range of values for every variable $\mathcal{Y} \in \mathcal{U} \cup \mathcal{V}$.

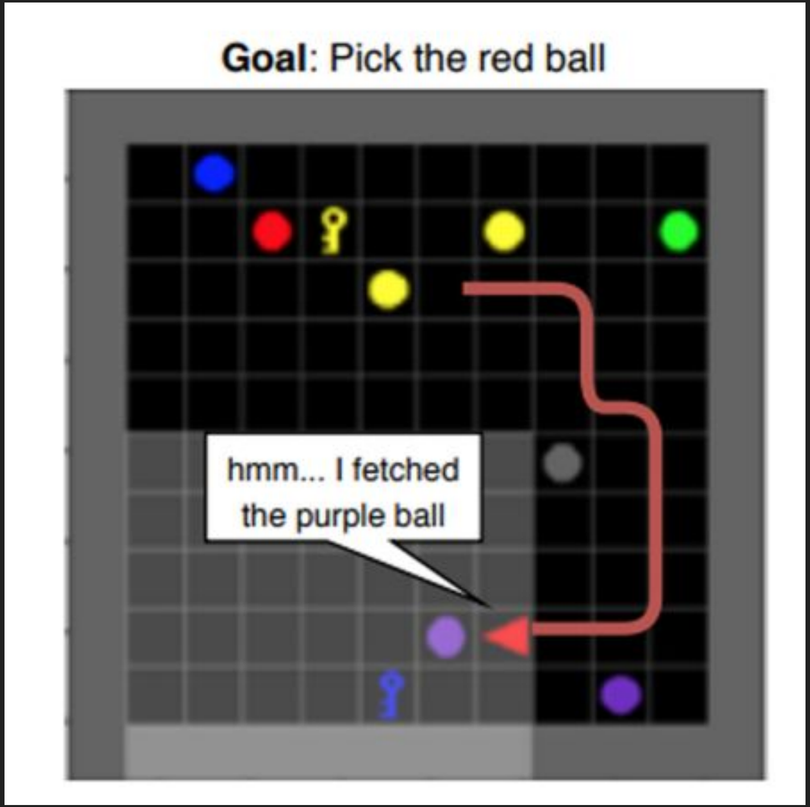
Definition 1. A *structural causal model* is a tuple $M = (\mathcal{S}, \mathcal{F})$, where $\mathcal{F}$ denotes a set of structural equations, one for each $X \in \mathcal{V}$, such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$ give the value of X based on other variables in $\mathcal{U} \cup \mathcal{V}$. That is, the equation F_X defines the value of X based on some other variables in the model.

Those we will learn!



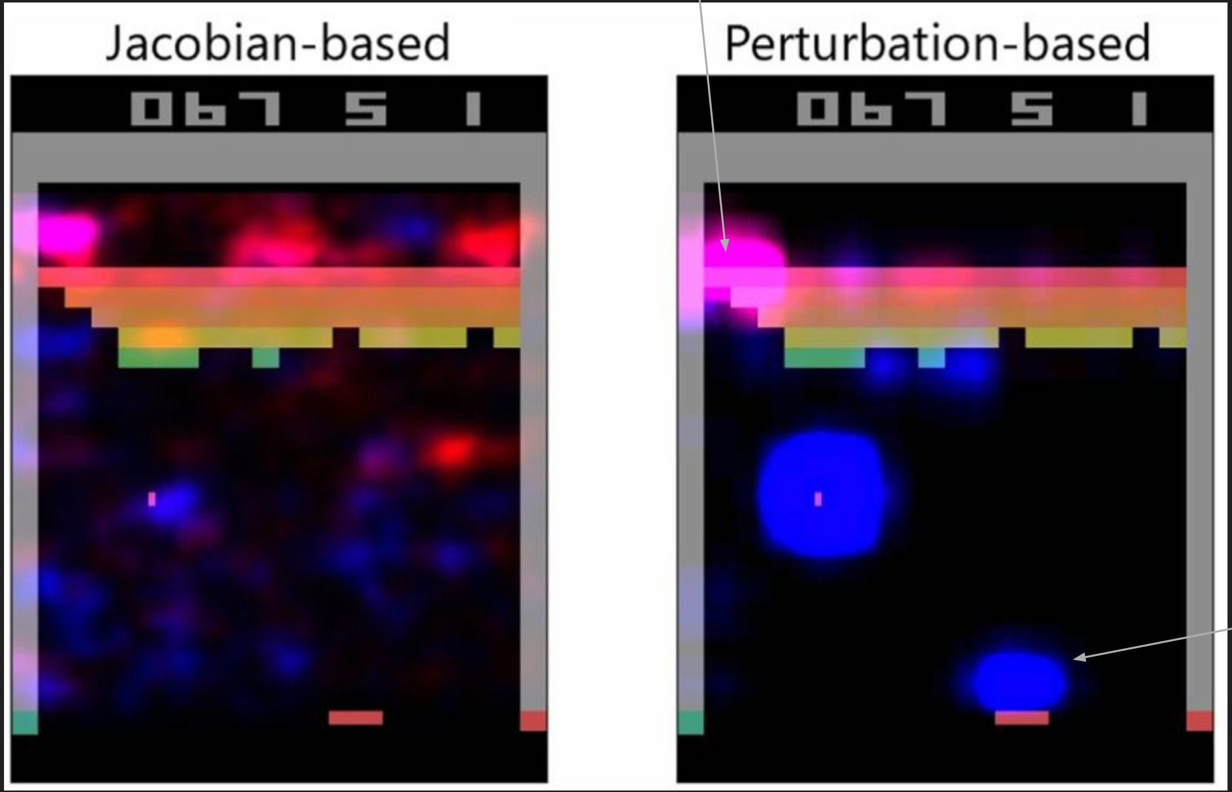
Improving instruction following with Hindsight Generation for Experience Replay

- Text-conditioned grid-world like agent
- For failed episodes generate and learn to predict final state textual description for additional learning signal (as classical in HER)
- Profit? Textual explanation of failed episodes helpful for model debugging!



Explanations through saliency maps

Critic explanation



Actor explanation

References

- Explainability in deep reinforcement learning – Alexandre Heuillet and Fabien Couthouis and Natalia Díaz-Rodríguez
- Explainable Reinforcement Learning via Reward Decomposition – Zoe Juozapaitis, et. al
- Explainable Reinforcement Learning Through a Causal Lens – Prashan Madumal, et. al
- HIGhER : Improving instruction following with Hindsight Generation for Experience Replay – Geoffrey Cideron, et. al
- Visualizing and understanding atari agents – S. Greydanus, A. Koul, J. Dodge, A. Fern
- Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps – Karen Simonyan, et. al

Thanks!

& discussion