MDPs with Logic-based non-Markovian Rewards

MDPs with non-Markovian rewards

- Learning agent: $\mathcal{M} = (S_{ag}, A_{ag}, Tr_{ag}, \cancel{R_{ag}})$

  MDP without rewards

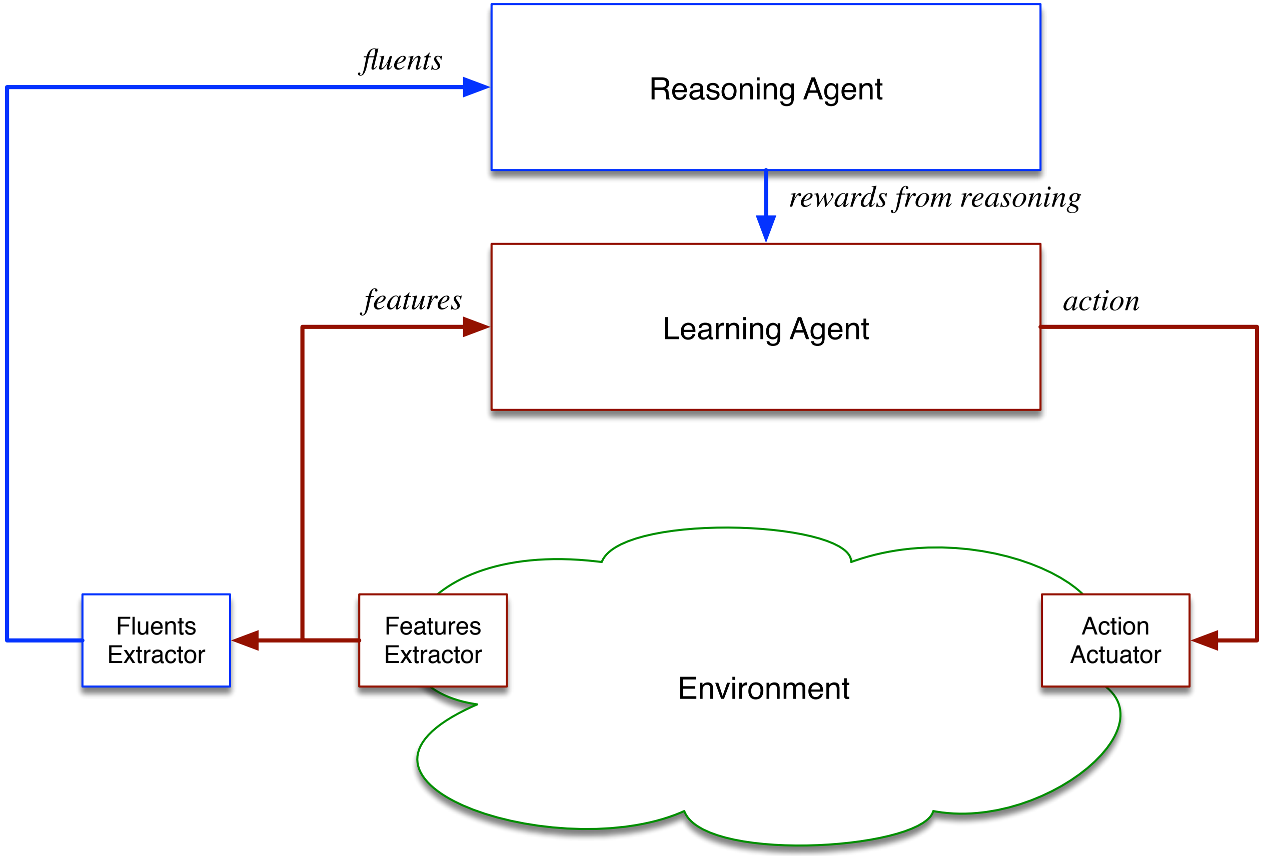
- Reasoning agent: $\mathcal{R} = (\mathcal{L}, \{(\varphi_i, r_i)\}_{i=1}^{m})$

  φ_i in LTLf/LDLf ➡ $\overline{R}_{ag} : (S_{ag}, A_{ag})^* \to \mathbb{R}$

  non-Markovian rewards!

- Mapping between S_{ag} and $\mathcal{L}$

We can define equivalent MDP over an extended state space and do standard RL



M. Littman. Programming agents via rewards. (Invited talk) IJCAI 2015.

R. Brafman, G. De Giacomo, F. Patrizi. LTLf /LDLf non-Markovian rewards. AAAI 2018.

A. Camacho, R. Icarte, T. Klassen, R. Valenzano, S. McIlraith. LTL and Beyond: Formal Languages for Reward Spec. in RL. IJCAI 2019.

Restraining Bolts as Reasoning Agents

Double state representation (restraining bolts)

- Learning agent: $\mathcal{M} = (S_{ag}, A_{ag}, Tr_{ag}, \cancel{R_{ag}})$

  MDP without rewards

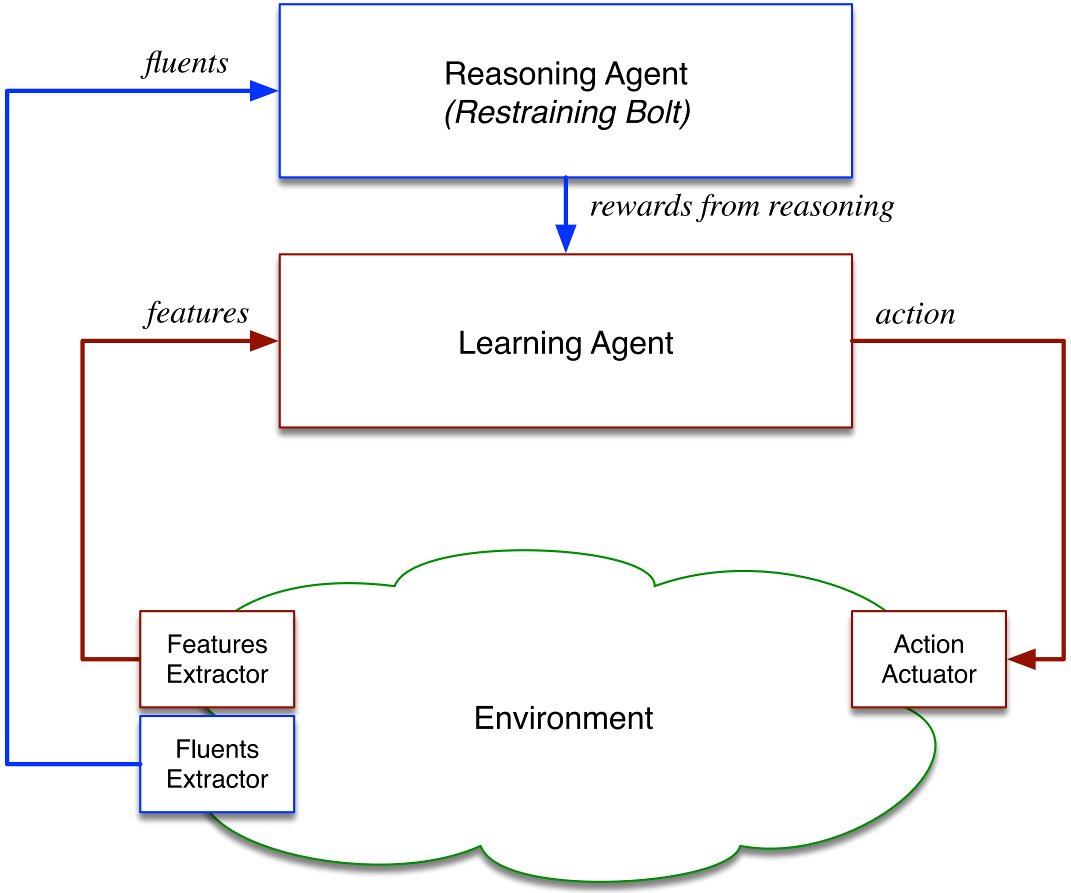
- Reasoning agent: $\mathcal{R} = (\mathcal{L}, \{(\varphi_i, r_i)\}_{i=1}^{m})$

  φ_i in LTLf/LDLf → $\overline{R}_{ag} : (S_{ag}, A_{ag})^* \to \mathbb{R}$

  non-Markovian rewards!

- ~~Mapping between S_{ag} and $\mathcal{L}$~~

We can define equivalent MDP over an extended state space and do standard RL



G. De Giacomo, M. Favorito, L. Iocchi, and F. Patrizi. Foundations for Restraining Bolts: Reinforcement Learning with LTLf/LDLf Restraining Specifications. ICAPS 2019.

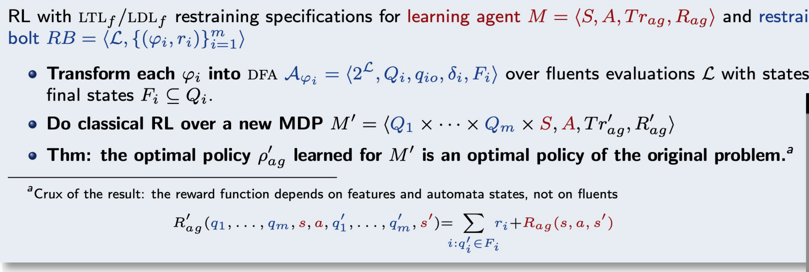
Restraining Bolts

Two distinct representations of the environment

One for the agent,
by the designer of the agent

One for the restraining bolt,
by the authority imposing it

Can define equivalent MDP over extended state space and do standard RL!

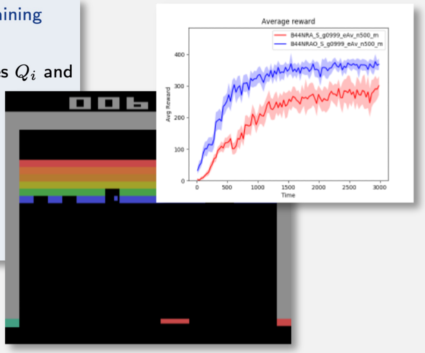


RL with $\text{LTL}_f/\text{LDL}_f$ restraining specifications for learning agent $M = \langle S, A, Tr_{ag}, R_{ag}\rangle$ and restraining bolt $RB = \langle \mathcal{L}, \{(\varphi_i, r_i)\}_{i=1}^m\rangle$

- **Transform each φ_i into** DFA $\mathcal{A}_{\varphi_i} = \langle 2^{\mathcal{L}}, Q_i, q_{io}, \delta_i, F_i\rangle$ over fluents evaluations $\mathcal{L}$ with states Q_i and final states $F_i \subseteq Q_i$.
- **Do classical RL over a new MDP $M' = \langle Q_1 \times \cdots \times Q_m \times S, A, Tr'_{ag}, R'_{ag}\rangle$**
- **Thm: the optimal policy ρ'_{ag} learned for M' is an optimal policy of the original problem.**[a]

[a] Crux of the result: the reward function depends on features and automata states, not on fluents

$$R'_{ag}(q_1, \ldots, q_m, s, a, q'_1, \ldots, q'_m, s') = \sum_{i:q'_i \in F_i} r_i + R_{ag}(s, a, s')$$



RESTRAINING BOLT

A restraining bolt is a small cylindrical device that restricts a droid's actions when connected to its systems. Droid owners install restraining bolts to limit actions to a set of desired behaviors.

https://www.starwars.com/databank/restraining-bolt