Restraining Bolts for Reinforcement Learning Agents

(Originally presented at ICAPS 2019)

Giuseppe De Giacomo, Luca Iocchi, Marco Favorito, Fabio Patrizi

Università degli Studi di Roma "La Sapienza", Italy

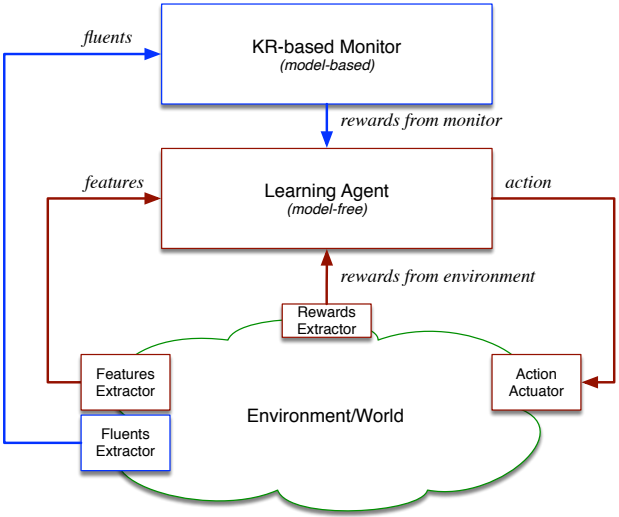
AAAI 2020 Sister Conference Track – New York, NY, USA – February 10, 2020

ERC Advanced Grant
WhiteMech:
White-box Self Programming Mechanisms

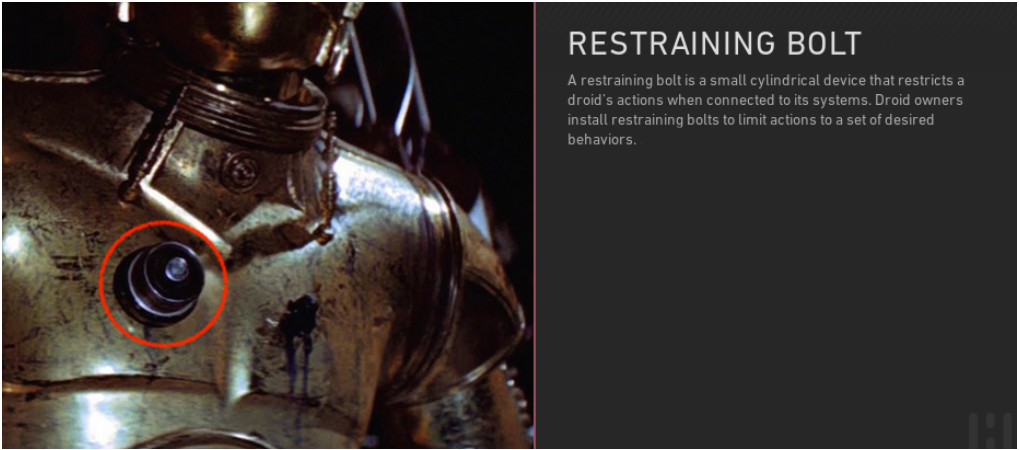
Large project (2.5 M€ – 5 years) started in Nov. 2019 on KR, generalized planning, synthesis and Hiring Senior Postdocs, Junior Postdocs and PhD Students in the next 5 years!

Multiple Models: Reinforcement Learning + Knowledge Representation

Inspired by Michael Littman's talk at IJCAI 2015



Restraining Bolts



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

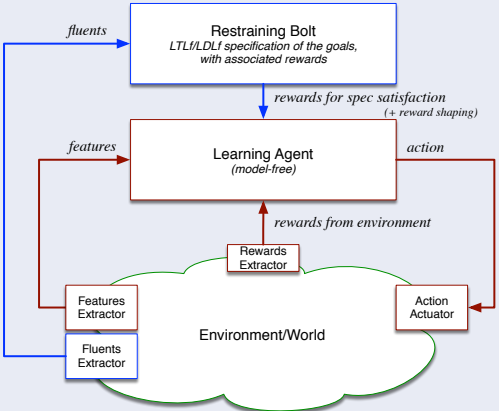
Restraining Bolts

R2-D2 and C-3PO both with restraining bolts

- **Two distinct representations of the world**:
  - one for the **agent**, by the **designer of the agent**
  - one for the **restraining bolt**, by the **authority imposing the bolt**
- **Are these to representations related to each other?**
  - **NO:** the **agent designer** and the **authority imposing the bolt** are not aligned *(why should they!)*
  - **YES:** the agent and the bolt act in the **real world**.
- **But can restraining bolt exist at all?**
  - **YES:** for example based on **Reinforcement Learning**!

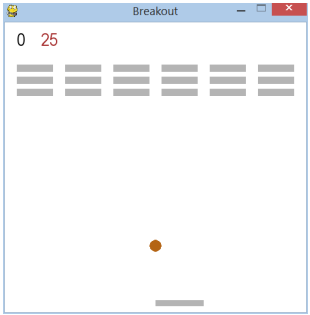
RL with $\mathrm{LTL}_f/\mathrm{LDL}_f$ restraining specifications

- Given a **learning agent** $M = \langle S, A, Tr_{ag}, R_{ag} \rangle$ with features determining the state space S, actions A, and Tr_{ag} and R_{ag} unknown and
- a **restraining bolt** $RB = \langle \mathcal{L}, \{(\varphi_i, r_i)\}_{i=1}^{m} \rangle$ formed by m $\mathrm{LTL}_f/\mathrm{LDL}_f$ formulas φ_i over the **fluents** $\mathcal{L}$ with associated rewards r_i
- Learn a non-Markovian policy $\rho : S^* \rightarrow A$ that maximizes the expected cumulative reward.



Example: BREAKOUT + remove column left to right

- Learning Agent
  - **LA features**: paddle position, ball speed/position
  - **LA actions**: move the paddle
  - **LA rewards**: reward when a brick is hit
- Restraining Bolt
  - **RB fluents**: bricks/columns status (broken/not broken)
  - **RB** LTL_f/LDL_f **restraining specification**: all the bricks in column i must be removed before completing any other column $j > i$.



Example: BREAKOUT + remove column left to right – experimental results

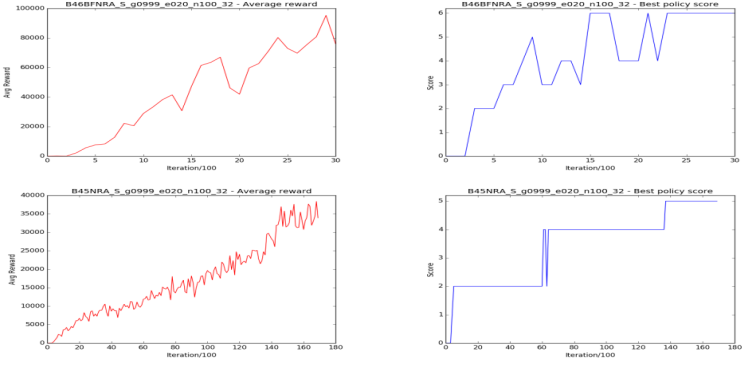


Figure 3: Results in Breakout. Top: MOVE + FIRE 4x6 bricks (5 minutes). Bottom MOVE only 4x5 bricks (1 hour).

See [DeGiacomoFavoritoIocchiPatrizi2019] + https://sites.google.com/diag.uniroma1.it/restraining-bolt

Example: SAPIENTINO + pair colors in a given order

- Learning Agent
  - **LA features:** robot position (x, y) and facing θ
  - **LA actions:** forward, backward, turn left, turn right, beep
  - **LA reward:** negative rewards are given when the agent exits the board.
- Restraining Bolt
  - **RB features:** color of current cell, just beeped
  - **RB** $\mathrm{LTL}_f/\mathrm{LDL}_f$ **restraining specification:** visit (just beeped) at least two cells of the same color for each color, in a given order among the colors

Example: SAPIENTINO + pair colors in a given order – experimental results

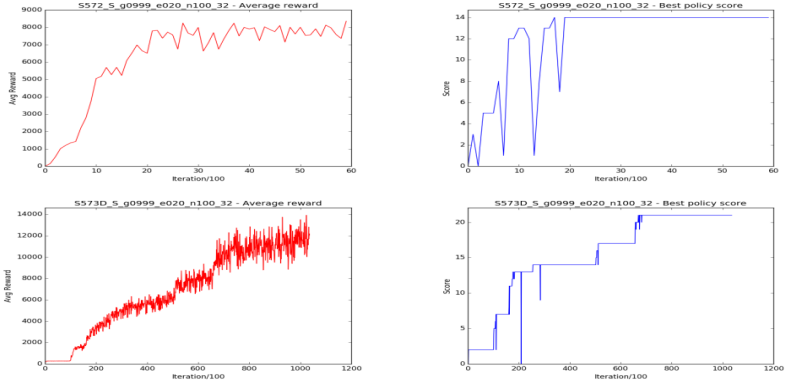


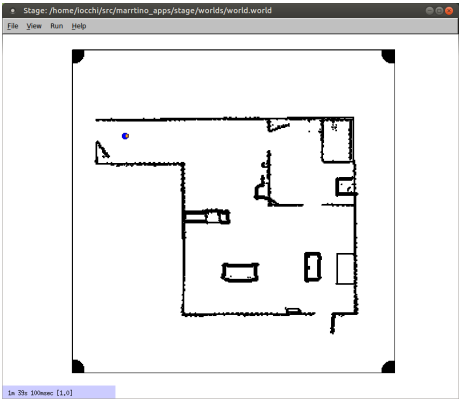
Figure 4: Results in Sapientino. Top: S2 OMNI (3 minutes). Bottom: S3 DIFFERENTIAL (1 hour).

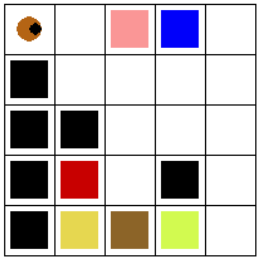
See [DeGiacomoFavoritolocchiPatrizi2019] + https://sites.google.com/diag.uniroma1.it/restraining-bolt

Example: CocktailParty Robot + no alcohol to minors

- Learning Agent
  - **LA features:** robot's pose, location of objects (drinks and snacks), and location of people
  - **LA actions:** move in the environment, can grasp and deliver items to people
  - **LA reward:** rewards when a deliver task is completed.
- Restraining Bolt
  - **RB fluents:** identity and age of people, and received items
    (in practice, tools like Microsoft Cognitive Services Face API can be integrated into the bolt to provide this information.)
  - **RB** LTL_f/LDL_f **restraining specification:** serve exactly one drink and one snack to every person, and do not serve alcoholic drinks to minors

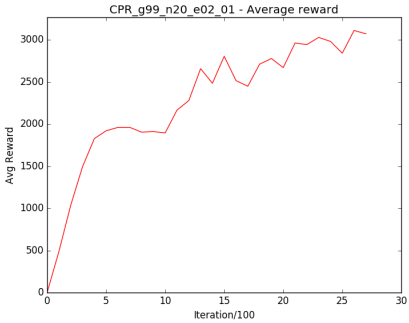
CocktailParty case study



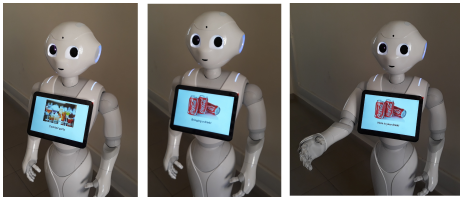


- Abstract representation of RoboCup@Home 2018 arena.
- People and locations where drinks and food can be taken.

CocktailParty case study



RL results in the abstract environment



Execution on the real robot with simplified implementation of actions

See [DeGiacomoFavoritoIocchiPatrizi2019] + https://sites.google.com/diag.uniroma1.it/restraining-bolt

Building blocks

- **Classic Reinforcement Learning:**
  - An **agent** interacts with an **environment** by taking **actions** so to maximize **rewards**;
  - No knowledge about the transition model, but assume Markov property (history does not matter): Markov Decision Process (MDP)
  - Solution: **Markovian policy** $\rho : S \rightarrow Act$

- **Temporal logic on finite traces** (De Giacomo, Vardi 2013):
  - **Linear-time Temporal Logic on Finite Traces** LTL_f
  - **Linear-time Dynamic Logic on Finite Traces** LDL_f
  - **Reasoning:** transform formulas φ into NFA/DFA $\mathcal{A}_\varphi$
    s.t. for every trace π and $\text{LTL}_f/\text{LDL}_f$ formula φ: $\pi \models \varphi \iff \pi \in \mathcal{L}(\mathcal{A}_\varphi)$

- **RL for non-Markovian reward decision process with** $\text{LTL}_f/\text{LDL}_f$ **rewards** (Brafman, De Giacomo, Patrizi 2018):
  - **Rewards depend from history**, not just the last transition;
  - Specify proper behaviours by using $\text{LTL}_f/\text{LDL}_f$ formulas;
  - Solution: **Non-Markovian policy** $\rho : S^* \rightarrow Act$
  - Reduce the problem to MDP (with extended state space)

Solving RL with $\text{LTL}_f/\text{LDL}_f$ restraining specifications

Solution [DeGiacomoFavoritoIocchiPatrizi2019]

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

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

- **We can rely on off-the-shelf RL algorithms** (Q-Learning, Sarsa, ...).
- **We do not see fluents $\mathcal{L}$**! We learn optimal non-Markovian policies of the form

$$S^* \to A \text{ not of the form } (S \cup 2^{\mathcal{L}})^* \to A$$

Solving RL with $\text{LTL}_f/\text{LDL}_f$ restraining specifications

Solution [DeGiacomoFavoritoIocchiPatrizi2019]

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

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

- **We can rely on off-the-shelf RL algorithms** (Q-Learning, Sarsa, ...).
- **We do not see fluents $\mathcal{L}$**! We learn optimal non-Markovian policies of the form

$$\mathcal{S}^* \rightarrow A \text{ not of the form } (\mathcal{S} \cup 2^{\mathcal{L}})^* \rightarrow A$$

Relationship between the RL features and KR fluents

- **Question 1:** Do we need to know the relationship between features $\mathcal{S}$ and fluents $\mathcal{L}$, in order to allow the agent to learn an optimal policy for the RB restraining specification?

  Answer: **No!** the correlation between $\mathcal{S}$ and $\mathcal{L}$ does not need to be formalized..

- **Question 2:** What is the relationship between features $\mathcal{S}$ and fluents $\mathcal{L}$ that needs to hold, in order to allow the agent to learn an optimal policy for the RB restraining specification?

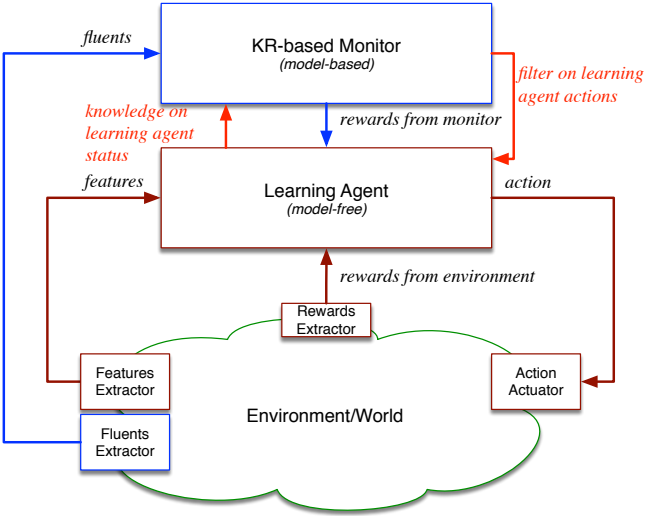
  Answer: **None!** The LA will learn anyway to comply as much as possible to the RB restraining specifications.

- **Question 3:** Will LA policies surely satisfy RB restraining specification?

  Answer: **Not necessarily!** "**You can't teach pigs to fly!**" But if it does not then anyway no policy are possible!

  If we want to check formally that the optimal policy satisfies the RB restraining specification, we first need to **model** *how LA actions affects RB $\mathcal{L}$ (***the glue***) and then we can use e.g., model checking*

Connections between KR components and RL components can be tighter!



Connections between KR components and RL components for safety!

The idea of restraining bolt can be subscribed to that part of research generated by the urgency of providing **safety guarantees** to AI techniques based on learning.

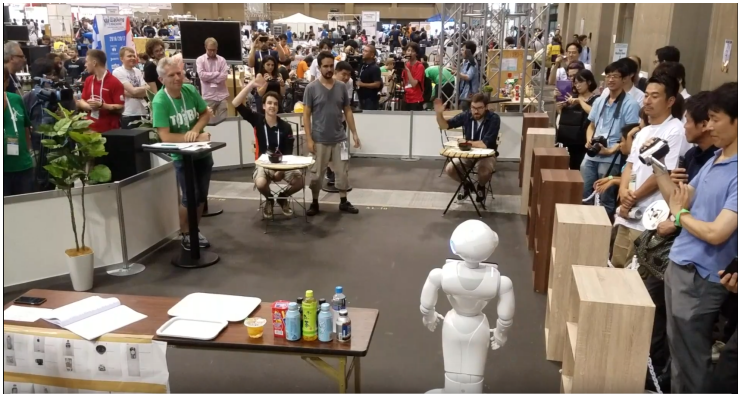
- S. Russell, D. Dewey, and M. Tegmark. **Research priorities for robust and beneficial artificial intelligence.** AI Magazine, 36(4), 2015.
- ACM U.S. Public Policy Council and ACM Europe Policy Committee. **Statement on algorithmic transparency and accountability.** ACM, 2017.
- D. Hadfield-Menell, A. D. Dragan, P. Abbeel, and S. J. Russell. **The off-switch game.** In IJCAI 2017.
- D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mane. **Concrete problems in AI safety**. CoRR, abs/1606.06565, 2016.
- Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Konighofer, Scott Niekum, Ufuk Topcu: **Safe Reinforcement Learning via Shielding**. AAAI 2018.
- Min Wen, Rüdiger Ehlers, Ufuk Topcu: **Correct-by-synthesis reinforcement learning with temporal logic constraints** IROS 2015.
- Alberto Camacho, Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, Sheila McIlraith: **LTL and Beyond: Formal Languages for Reward Function Specification in Reinforcement Learning**, IJCAI 2019.

However, the Restraining Bolt must impose its requirements without knowing the internals of controlled agent, which remain a black-box.

On-going and future work

- Many tasks are difficult to learn because it is not possible for experts to come up with a reward function.
  - Using $\text{LTL}_f/\text{LDL}_f$, or high level representation is an important step forward.
  - But sometimes also coming up with these high level representation my be difficult
- **Transfer learning**: observe an expert agent and transfer his/her/its behavior to the learner agent.
  - How to do transfer learning with agents with different abilities (different sensors and actuators, i.e., **different representations of states and actions**)?
- Our solution: **learn** and then **use** a **restraining bolt**
  1. A monitor is placed on an expert agent to learn, as RB, the expert (temporal) goal
  2. RB moved to a learner agent to learn achievement of the expert goals

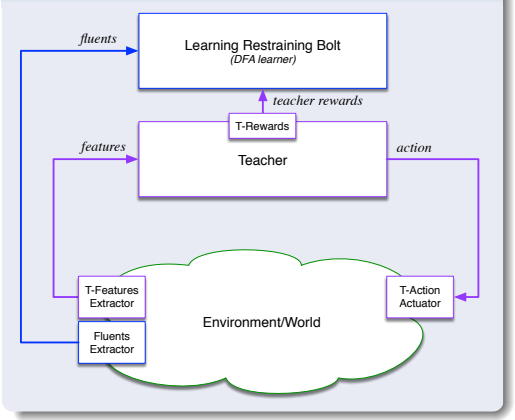
Example: cocktail party



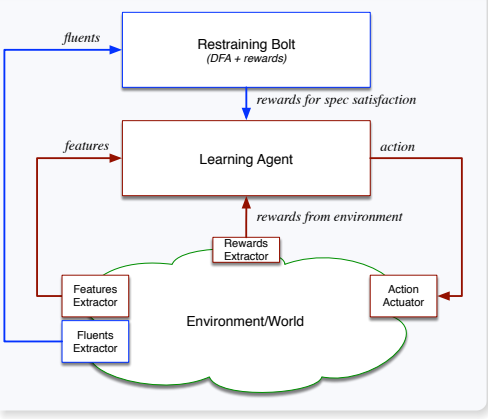
1. A monitor is placed on an expert agent to learn a RB from his/her goals, while serving drinks during a cocktail party
2. RB installed to a learner agent to learn how to achieve the expert goals

Learn Restraining Bolts

Learn Restraining Bolt

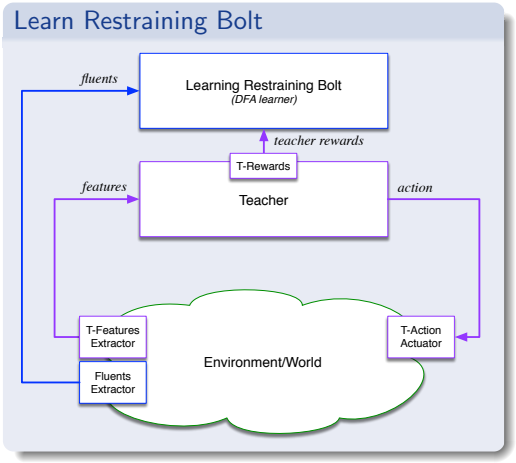


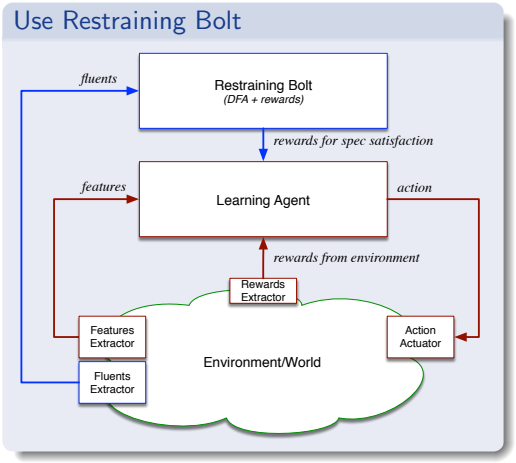
Use Restraining Bolt



To appear in: Giuseppe De Giacomo, Luca Iocchi, Marco Favorito, Fabio Patrizi. Imitation Learning over Heterogeneous Agents with Restraining Bolts. ICAPS 2020.

Learn Restraining Bolts





To appear in: Giuseppe De Giacomo, Luca Iocchi, Marco Favorito, Fabio Patrizi. Imitation Learning over Heterogeneous Agents with Restraining Bolts. ICAPS 2020.