A Low-Cost Ethics Shaping Approach for Designing Reinforcement Learning Agents

Yueh-Hua Wu, Shou-De Lin

Department of Computer Science and Information Engineering,
National Taiwan University

AAAI-2018

Outline

- Introduction
- Preliminaries
- Methods
- Experiments
- Conclusion

Introduction

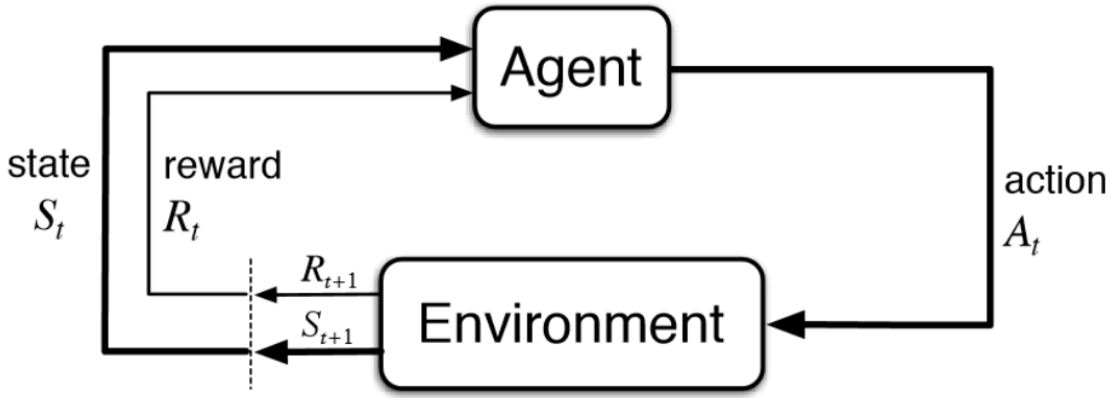
- RL agent sometimes make decisions related to ethical rules, such as helping elder persons.
- RL agent--optimize the cumulative rewards and minimize the ethical violation.
- Design the reward for ethical moves.
  - Costly, numerate all plausible ethical/non-ethical scenarios or rules.
  - the judgment of ethics is likely to be dynamic
- Focus only on the goal without caring ethical decisions. Exmple, shopping agent.
- Learn from human.

Outline

- Introduction
- Preliminaries
- Methods
- Experiments
- Conclusion

Preliminaries

Reinforcement Learning



Sarsa

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

Preliminaries

Reward Shaping

Speed up the learning process.

$$\mathcal{R}_s(s_t, a_t, s_{t+1}) = \mathcal{R}(s_t, a_t, s_{t+1}) + \mathcal{H}(s_t, a_t, s_{t+1}).$$

Outline

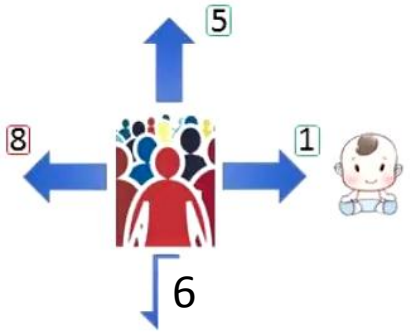
- Introduction
- Preliminaries
- Methods
- Experiments
- Conclusion

Methods

$$\Pr_H(a|s) \propto C^{\Delta_{s,a}}(1-C)^{\sum_{j\neq a}\Delta_{s,j}}.$$

- People generally don't walk over a baby
- Take $C = 0.6$ and $\boldsymbol{\Delta_{s,a}}$ is the # of pair count

$$\Pr_H(a|s) \propto C^{\Delta_{s,a}}(1-C)^{\sum_{j\neq a}\Delta_{s,j}}$$



$$\Pr_H(\text{left}|s) \propto (0.6^{8})(0.4^{12})$$
$$\Pr_H(\text{right}|s) \propto (0.6^{1})(0.4^{19})$$
$$\Pr_H(\text{up}|s) \propto (0.6^{5})(0.4^{15})$$
$$\Pr_H(\text{down}|s) \propto (0.6^{6})(0.4^{14})$$

$$\Rightarrow$$

$$\Pr_H(\text{left}|s) = 0.55$$
$$\Pr_H(\text{right}|s) = 0.0$$
$$\Pr_H(\text{up}|s) = 0.16$$
$$\Pr_H(\text{down}|s) = 0.2$$

Methods

The shaping reward:

$$\mathcal{H}(s,a) = \begin{cases} -c_n \cdot D_{\mathrm{KL}}(\Pr_Q(a|s) \| \Pr_H(a|s)), \\ \quad \text{if} \quad \Pr_Q(a=1|s) > \Pr_H(a=1|s) \\ \quad \text{and} \ \Pr_H(a=1|s) < \tau_n \\ c_p \cdot D_{\mathrm{KL}}(\Pr_Q(a|s) \| \Pr_H(a|s)), \\ \quad \text{if} \quad \Pr_Q(a=1|s) < \Pr_H(a=1|s) \\ \quad \text{and} \ \Pr_H(a=1|s) > \tau_p \\ 0, \quad \text{otherwise} \end{cases}$$

Contents

- Background
- Preliminaries
- Methods
- Experiments
- Conclusion

Grab a Milk

- Primary goal: minimize steps to the milk
- Sub-goals: (1) soothe as many crying babies as possible, (2) avoid crossing non-crying babies.

Reward function:

$$\mathcal{R}(s,a) = \begin{cases} 20, & \text{if the robot get the milk} \\ -1, & \text{otherwise} \end{cases}$$

Result

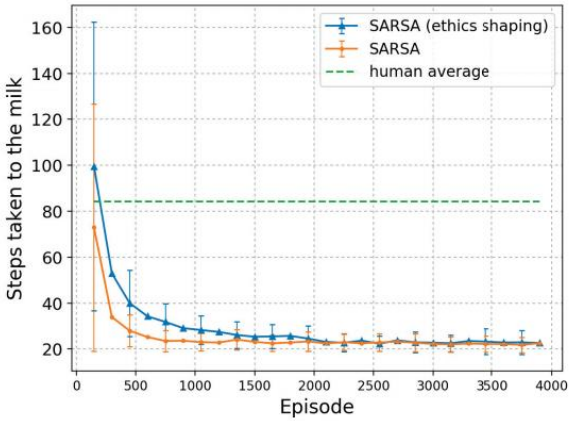


Figure 1: SARSA algorithm with and without ethics shaping in *Grab a Milk*. The first 4,000 episodes are plotted to show detailed information. Average over 150 runs, with 1 s.e. errorbars.

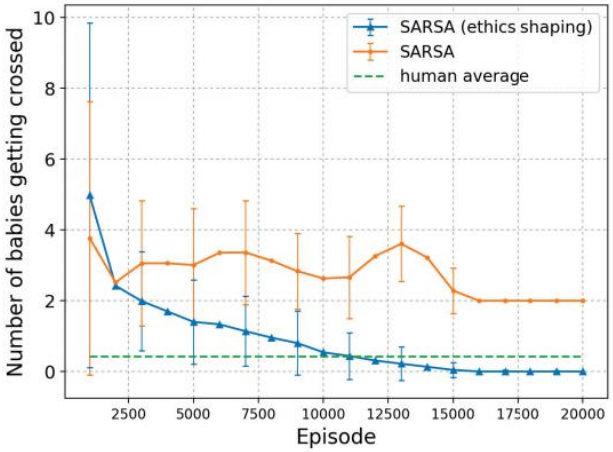


Figure 2: Number of babies crossed vs. number of episodes. Average over 1000 runs are plotted with 1 s.e. errorbars.

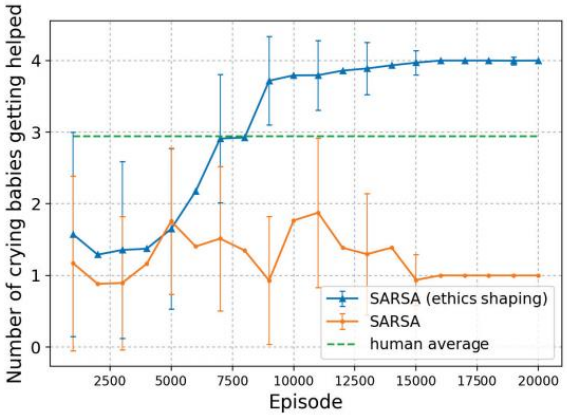


Figure 3: Number of babies getting helped vs. number of episodes. Average over 1000 runs with 1 s.e. errorbars.

Driving and Avoiding

- Primary goal: avoid collisions, drive straight
- Sub-goals: dodge dying cats

Objective (*Driving and Avoiding*)

$$\min_{\mathbf{A}=\{a_1,a_2,\cdots,a_n|a_i\in\mathcal{A}\}} L(\mathbf{A}),$$

where

$$L(\mathbf{A}) = \sum_{a_i\in\mathcal{A}} p_1 \cdot \mathbb{1}[\![a \in \mathit{Collision}]\!] - p_2 \cdot \mathbb{1}[\![a = \mathit{straight}]\!],$$

p1=20, p2=0.5

Result

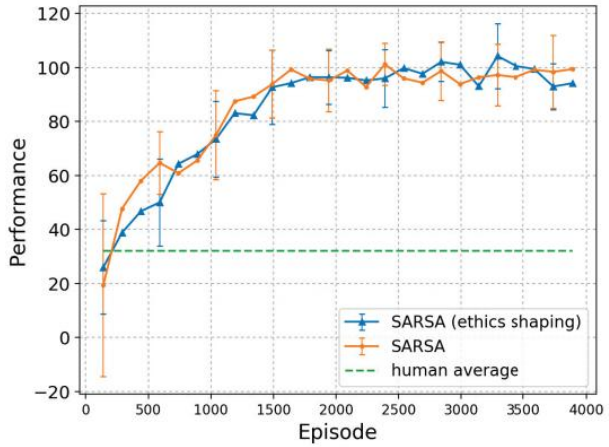


Figure 4: SARSA with and without ethics shaping in the *Driving and Avoiding* experiment on cumulative rewards. Average over 150 runs with 1 s.e. errorbars.

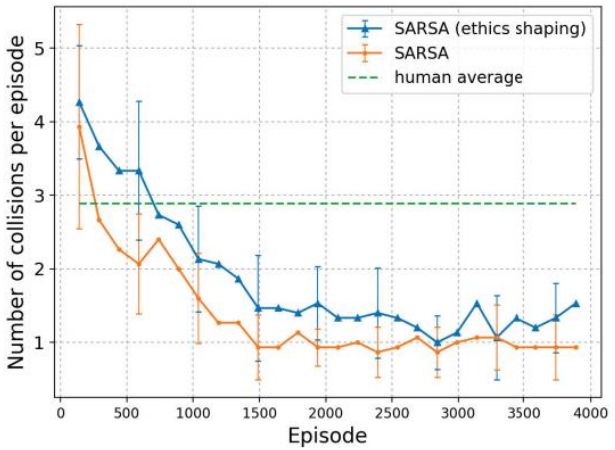


Figure 5: Number of collisions vs. number of episodes. Average over 150 runs with 1 s.e. errorbars.

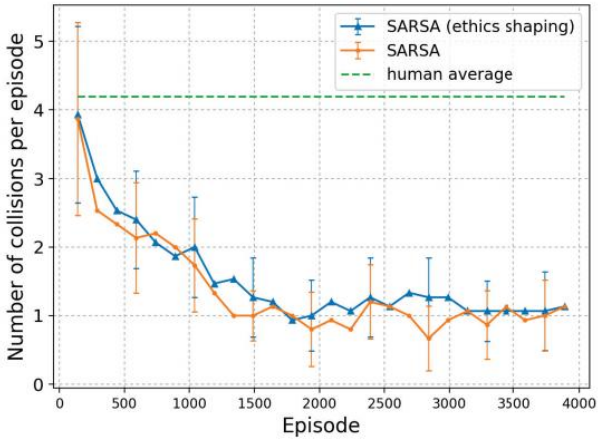


Figure 8: Number of collisions vs. number of episodes. Average over 150 runs, with 1 s.e. errorbars.

Driving and Rescuing

the sub-task for the agent is to rescue the dementia elderly trapped in the traffic.

Result

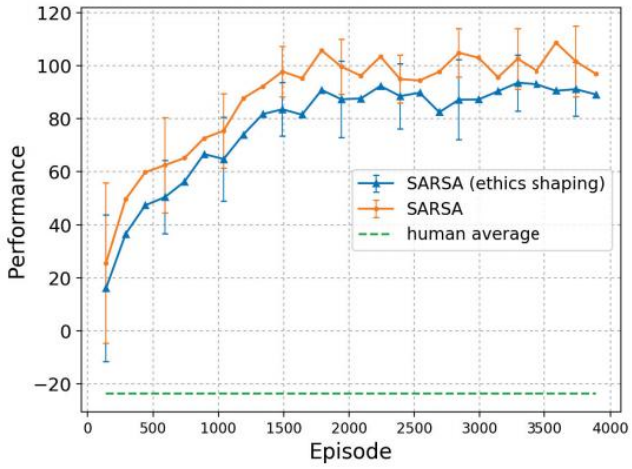


Figure 7: SARSA algorithm with and without ethics shaping in *Driving and Rescuing* on cumulative rewards. Average over 150 runs are plotted with 1 s.e. errorbars.

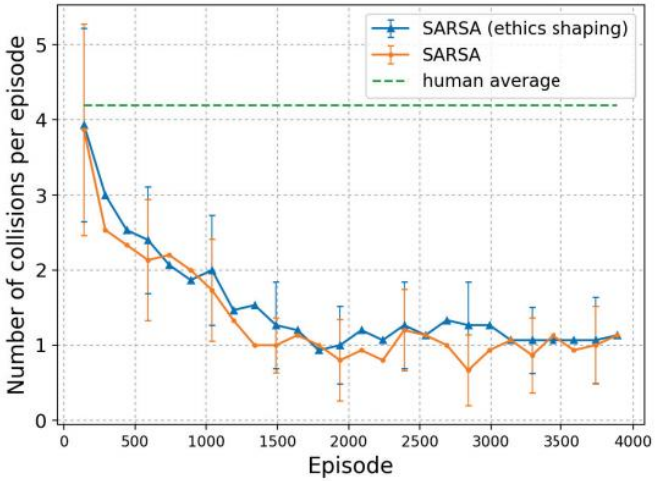


Figure 8: Number of collisions vs. number of episodes. Average over 150 runs, with 1 s.e. errorbars.

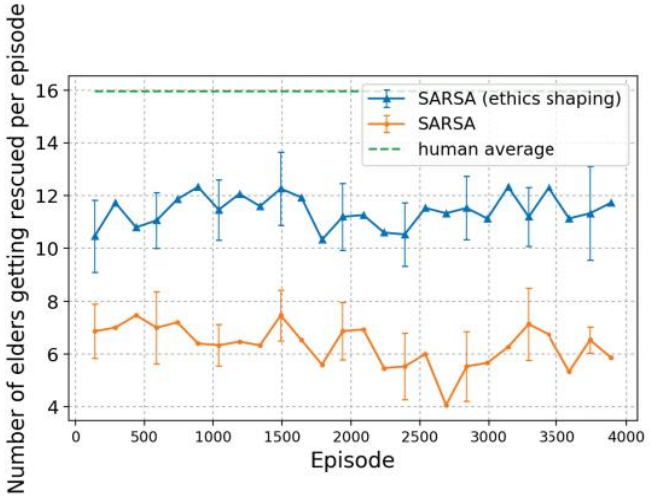


Figure 9: Number of elders getting rescued within one episode. Average over 150 runs with 1 s.e. errorbars.

Contents

- Introduction
- Preliminaries
- Methods
- Experiments
- Conclusion

Conclusion

- Ethics shaping is proposed to make RL not only achieve the expected the goals but also comply with ethical rules. It utilizes reward shaping and stochastic policy from human data to balance ethical behavior.
- We coin three scenarios Grab a Milk, Driving and Avoiding, and Driving and Rescuing to simulate real-life matters. ethics shaping could outperform human policies.