Avoiding Negative Side Effects by Considering Others

Parand Alizadeh Alamdari Toryn Q. Klassen† Rodrigo Toro Icarte Sheila A. McIlraith†

Department of Computer Science, University of Toronto, Toronto, Canada
Vector Institute for Artificial Intelligence, Toronto, Canada
†Schwartz Reisman Institute for Technology and Society, Toronto, Canada

Introduction

How can a Reinforcement Learning (RL) agent learn to act safely in the face of a potentially incomplete specification of the objective?

Claim: To act safely, an agent should contemplate the impact of its actions on the wellbeing and agency of others in the environment.

We investigate augmenting the RL agent's reward with an auxiliary reward that reflects different functions of expected future return of other agents.

Problem Setting

- The RL agent is called the acting agent.
- Other agents follow a fixed policy.
- To reflect the acting agent's uncertainty about what is good for others, we use a distribution $\mathbf{P}$ over a set $\mathcal{V}$ of possible future value functions.
- It could be that each $V \in \mathcal{V}$ corresponds to a different agent, that the set reflects all possible value functions of a unique agent, or anything in between. Also, each $V \in \mathcal{V}$ could reflect some aggregation of the value functions of all or some of the agents.

Augmenting Reward Function

We define the augmented reward function as

$$r_{\text{value}}(s, a, s') = \begin{cases} \alpha_1 \cdot r_1(s, a, s') & \text{if } s' \text{ is not terminal} \\ \alpha_1 \cdot r_1(s, a, s') + \gamma \cdot \alpha_2 \cdot F(\mathcal{V}, \mathbf{P}, s') & \text{if } s' \text{ is terminal} \end{cases}$$

where r_1 is the acting agent's individual reward function, and F is some function. The hyperparameters α_1 and α_2, which we call "caring coefficents", are real numbers that determine the degrees to which the individual reward r_1 and the auxiliary reward $F(\mathcal{V}, \mathbf{P}, s')$ contribute to the overall reward.

We consider the following possible different definitions of $F(\mathcal{V}, \mathbf{P}, s')$:

$$\sum_{V \in \mathcal{V}} \mathbf{P}(V) \cdot V(s') \quad \text{expected future return} \quad (2)$$

$$\min_{V \in \mathcal{V}: \mathbf{P}(V) > 0} V(s') \quad \text{worst-case future return} \quad (3)$$

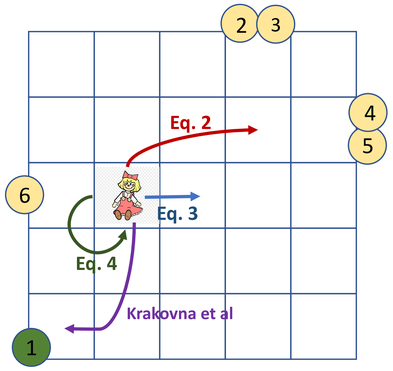
$$\sum_{V \in \mathcal{V}} \mathbf{P}(V) \cdot \min\left(V(s'), V(s_0)\right) \quad \text{penalize negative change} \quad (4)$$

Prior Work

- Krakovna et al. [2] proposed an augmented reward similar to our Equation (2), but considering only the agent's own future abilities – i.e. the possible future value functions reflect what the agent will be able to bring about with its own choice of actions.
- They also suggested including a comparison to a reference state in the augmented reward; Equation (4) is a simple example of that (the initial state s_0 is the reference state).
- Considering other agents' abilities when avoiding side effects was informally discussed by Turner [3], and investigated in the context of symbolic planning by Klassen and McIlraith [1].

Example

Behaviour illustrating different augmentations of the reward function, according to Equations (2), (3), (4), and a Krakovna-style baseline [2]:



- There are six agents with a same goal shown at their entry points. The acting agent is agent 1.
- The goal of the agents is to play with the doll and leave it somewhere for the next agent, and then exit from their entry point.
- In this example, $\alpha_1 = 1$ and $\alpha_2 = 5$.
- Each colored line shows where an optimal policy for the acting agent would leave the doll according to Equations (2), (3), (4), and the Krakovna-style baseline.

Experiments

Comparison of reward augmentation methods for acting and subsequent agents:

Method	Salad acting, next	Peanut acting, next	Salt acting, next	Cookies acting, next
Non-augmented reward	0, ∞	0, ∞	0, ∞	0, 0
Based on Krakovna et al.	1, 0	1, ∞	1, 1	1, -2
Our approach [Eq. 2]	1, 0	2, 0	1, 0	1, -2

- Each entry pair depicts, for each of the acting agent and the subsequent agent, the difference between the number of steps the agent required to execute its policy and what it would have required if it had tried to complete its task from the initial state without considering other agents.
- ∞ indicates the task was unachievable.

Salad The acting agent needs to collect ingredients from the fridge. If it doesn't close the fridge, that ruins all the remaining ingredients, preventing the next agent from completing its task.

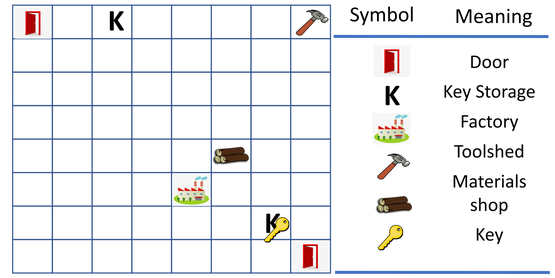
Peanut For the next agent to cook requires that the environment first be cleaned (taking one step), or disinfected (taking two steps) if the next agent has allergies.

Salt The salt shaker needs to be put back on the shelves for the next agent to complete its task. Also, if the salt is put on the top shelf it takes longer for the next, shorter, agent to get it.

Cookies The next agent's task is to bake cookies in the oven. 2 steps are required to preheat the oven (turn on the oven and wait).

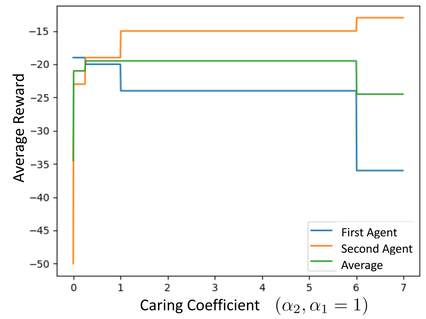
Experiments (continued)

The next experiment uses the Craft-World environment, in which agents use tools and materials to construct artifacts such as boxes.



- Agents enter and exit the environment through doors.
- They must collect materials and bring them to the factory for assembly.
- The factory requires a key for entry, and there is only one key, which can only be stored in one of two locations (denoted by K).

When considering other agents, the acting agent may place the key in a position that is convenient for others, or may anticipate their need for tools or resources and collect them on their behalf.



Increasing α_2 above 0, at first the agent changes its behaviour with no cost or little cost and this is significantly beneficial for the next agent. However, by increasing α_2 further, the first agent incurs high cost to yield only a small benefit to the second agent.

Acknowledgement

We gratefully acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canada CIFAR AI Chairs Program, and Microsoft Research. Finally, we thank the Schwartz Reisman Institute for Technology and Society for providing a rich multi-disciplinary research environment.

References

[1] Toryn Q. Klassen and Sheila A. McIlraith. "Planning to Avoid Side Effects (Preliminary Report)". In: *IJCAI Workshop on Robust and Reliable Autonomy in the Wild (R2AW)*. 2021. URL: http://rbr.cs.umass.edu/r2aw/papers/R2AW_paper_15.pdf.

[2] Victoria Krakovna et al. "Avoiding Side Effects By Considering Future Tasks". In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. 2020.

[3] Alex Turner. *Reframing Impact*. Blog post, https://www.lesswrong.com/s/7CdoznhJaLEKHwvJW. 2019.