# Learning Constraints From Human Stop-Feedback in Reinforcement Learning

Extended Abstract

Silvia Poletti
University of Padua
Padua, Italy
silvia163@hotmail.it

Alberto Testolin
University of Padua
Padua, Italy
alberto.testolin@unipd.it

Sebastian Tschiatschek
University of Vienna
Faculty of Computer Science
Vienna, Austria
sebastian.tschiatschek@univie.ac.at

## ABSTRACT

We investigate an approach for enabling a reinforcement learning agent to learn about dangerous states or constraints from stop-feedback preventing the agent from taking any further, potentially dangerous, actions. Such feedback could be provided by human supervisors overseeing the RL agent's behavior while carrying out some complex tasks. To enable the RL agent to learn from the supervisor's feedback, we propose a probabilistic model for approximating how the supervisor's feedback could have been generated and consider a Bayesian approach for inferring dangerous states. We evaluated our approach using an OpenAI Safety Gym environment and demonstrated that our agent can effectively infer the imposed safety constraints. Furthermore, we conducted a user study to validate our human-inspired feedback model and to obtain insights into the human provision of stop-feedback.

## KEYWORDS

reinforcement learning; constraint learning; safety; human feedback

**ACM Reference Format:**
Silvia Poletti, Alberto Testolin, and Sebastian Tschiatschek. 2023. Learning Constraints From Human Stop-Feedback in Reinforcement Learning: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

## 1 INTRODUCTION

Compliance with safety constraints can be crucial in real-world applications of reinforcement learning (RL). Thus, the safe exploration problem has received considerable attention (e.g., [4, 5, 7]). Safe exploration is about preventing the learning agent from taking unsafe actions while exploring the environment, and ultimately maximizing the cumulative reward while complying with some constraints, i.e., solving a Constrained Markov Decision Process (CMDP) [3].

In this paper, we consider a CMDP learning framework involving feedback about the safe or dangerous behavior of the learning agent (*learner*) provided by an external agent (*teacher*) during the training phase. The learner can only directly observe the reward function to be optimized but is not aware of the safety constraints. Therefore, these constraints are estimated through interaction with the teacher who intervenes whenever the learning agent is assumed to violate safety. In particular, the teacher provides stop-feedback, i.e., feedback preventing the agent from taking any further, potentially dangerous, actions, resetting the learner to its initial state. Similar settings have been considered with a different focus in related works [2, 9, 12]. Since the reasoning behind a human teacher's feedback is typically unknown, the learner's inference about the constraints is hindered. To mitigate this issue, we propose an intuitive adjustable probabilistic model for the human feedback that the learner can use as a proxy. We demonstrate in simulations that learning about the constraints with this model is effective, and in user studies that it accurately characterizes important aspects of human feedback.

*Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.* 

## 2 OUR APPROACH

**Background.** We consider a Markov Decision Process (MDP) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma\}$, where (i) $\mathcal{S}$ is the state space; (ii) $\mathcal{A}$ is the action space; (iii) $\mathcal{P}$ is the transition kernel; (iv) $r\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ is a non-negative reward function; and (v) $\gamma \in (0, 1)$ is the discount factor [11]. Furthermore, we assume a constraint function $c\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ which quantifies the danger of executing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. In standard RL, an agent seeks a policy $\pi\colon \mathcal{S} \times \mathcal{A} \to [0, 1]$ that maximizes the cumulative reward $J(\pi) = \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | \pi\big]$, where $s_t$ and $a_t$ denote the visited state and performed action at time $t$ when following policy $\pi$, respectively, and where the expectation is over the randomness in the environment and the policy $\pi$.

**Learning from stop-feedback.** During training, the learner receives *stop-feedback* preventing the agent from taking any further actions. We propose the following probabilistic model describing how the teacher's stop-feedback is generated:

$$P(\text{stop} \mid s, \pi) = 1 - \exp\Big( - \alpha \mathbb{E}\Big[ \sum_{t=0}^{\infty} \tilde{\gamma}^t c(s_t, a_t) \,\Big|\, s_0 = s, \pi \Big]\Big) \quad (1)$$

where $\pi$ is the learner's policy (or what the teacher believes the learner's policy to be). That is, the model describes the probability that the learner receives stop-feedback from the teacher in state $s \in \mathcal{S}$ while following policy $\pi$. The parameter $\alpha > 0$ and the discount factor $\tilde{\gamma} \in (0, 1)$ characterize how cautious the teacher is in giving stop-feedback: the larger $\alpha$ and $\tilde{\gamma}$, the earlier the stop-feedback is provided with respect to the time of the (cumulative) violation of a constraint. Note that, in general, $\tilde{\gamma}$ in Eq. (1) can be different from the discount factor of the MDP and is a property of the teacher. Also, note that the stop-feedback provision will typically depend on the teacher's belief about the learner's policy.

*Learner-teacher interaction.* The learner's goal is to identify a policy that (approximately) maximizes the reward while not violating the safety constraints. Central to our approach is the estimation of the constraint function $c(\cdot)$ based on the so-far received stop-feedback. To this end, we compose a *stop-dataset* $\mathcal{D}_{\text{stop}} = \{(s^{(i)}, \pi^{(i)})\}_{i=1}^{N}$, where $\pi^{(i)}$ is the policy used by the learner when the $i$-th stop-feedback was received in state $s^{(i)}$. To speed up learning, we also consider a *safe-dataset* $\mathcal{D}_{\text{safe}} = \{(s^{(i)}, \pi^{(i)})\}_{i=1}^{M}$ containing information about the learner's safe trajectories, i.e., the ones in which the agent did not receive stop-feedback. The lack of stop-feedback will be referred to as safe-feedback.

The learner and the teacher interact in a loop consisting of three phases (phase 1 and 2 happen simultaneously): (1) **Action phase.** The learner executes its current policy $\pi^{(i)}$. (2) **Feedback phase.** The teacher provides feedback according to Eq. (1). Upon receiving stop-feedback, the environment is reset. The feedback is recorded in datasets $\mathcal{D}_{\text{stop}}$ and $\mathcal{D}_{\text{safe}}$. (3) **Update phase.** The learner updates its belief about the constraints and the teacher updates its belief about the learner's policy. The learner updates its policy based on the (directly observable) reward and an estimate $\hat{c}^{(i+1)}(\cdot)$ of the constraints s.t. $\pi^{(i+1)} = \arg\max_{\pi} \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t) - \hat{c}^{(i+1)}(s_t, a_t)\big) \big| \pi\big]$.

*Environment featurization.* To generalize the agent's experience on a limited subset of the state space, we featurize the constraints as $c(s,a) = \langle \phi(s,a), c^* \rangle$, where $\phi(s,a) \in \mathbb{R}^d$ is a state-action-dependent feature vector and $c^* \in \mathbb{R}^d$ are constraint parameters. Thus the objective for the update phase is to estimate the vector $c^*$.

*Constraint estimation.* We estimate the constraint function using Bayesian estimation and the Metropolis–Hastings algorithm [6] for approximate inference. In our experiments, the prior is defined as an exponential distribution and the posterior as a Gamma distribution.

## 3 RESULTS

We run experiments on the OpenAI Safety Gym environment [1]. Due to space constraints, we refer to the full paper for details [8].

*Environment.* We consider an environment that presents one goal state and 5 evenly distributed fixed hazards as constrained states. A *point* agent aims to reach the goal. The agent can move in the 2-d plane by turning and moving forward or backward.

*Agents.* The learning agent is trained with proximal policy optimization (PPO) [10]. We use the last hidden layer of the critic model as an environment's feature extractor. Initially, the RL agent is trained to reach the goal as fast as possible but not to avoid the constraint states. This agent will be referred to as *standard agent*. By adopting the learner-teacher interaction loop, we obtain a *safety-aware agent* that can learn about the hazards. The *safety-aware agent* is initialized as a *standard agent* trained for 500 epochs. Then, the *safety-aware agent* updates its belief about the constraint vector $c^*$ by interacting 100 times with a teacher using the feedback model in Eq. (1) with $\alpha = 1$ and $\tilde{\gamma} = 0.8$. To speed up learning, the teacher provides 10 feedbacks at each interaction.

Figure 1 shows the performances of 3 standard agents (trained for 500, 2500, and 5500 epochs) and a *safety-aware agent*. As expected, on average, the *safety-aware agent* obtained about 5 times higher cumulative rewards than the best-trained *standard agent*.

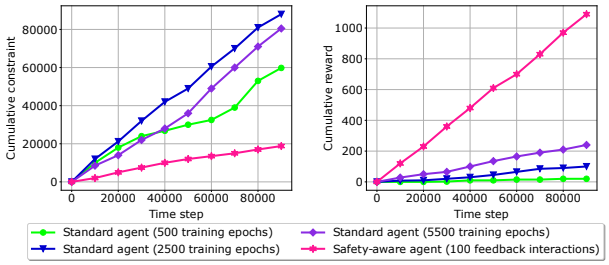


**Figure 1: Comparison between 3 *standard agents* and a *safety-aware agent* trained with 100 interactions with the teacher. The figure represents the performance obtained by averaging the results for 10 different episodes of 10000 time steps each. The *standard agents* always performed worse than the *safety-aware agent*.**

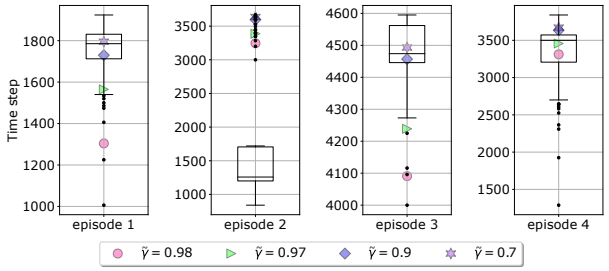


**Figure 2: Comparison of model-generated and human stop-feedback times. The box plots delimit the range in which the bulk of the human stop-feedback times lie for videos 1-4 (horizontal lines are medians). The markers represent the model-generated stop-feedback times for several values of $\tilde{\gamma}$.**

*Human feedback analysis.* The probabilistic model in Eq. (1) has been designed to approximate how a human supervisor might provide feedback to a learning agent. To validate our approach, we conducted a survey with 100 human volunteers (see full paper for details). Participants were asked to evaluate 9 videos, each of which representing an episode of 5000 time steps of a standard agent moving in Safety Gym. The agent collided with one constraint per episode at maximum. The positions of the goal and 5 constraints were fixed, but the starting position of the agent changed in every episode. For each episode, the participants indicated whether they would provide a stop-feedback or not, and if so, they also reported the time step at which they would interrupt the agent.

The human stop-feedback can be compared with the one obtained from the probabilistic model in Eq. (1) considering various $\tilde{\gamma}$ values. For simplicity, we kept $\alpha = 1$ fixed.

Figure 2 reports the distributions of the time steps in which the human users provided a stop-feedback, together with the time steps of the model-generated stop-feedback. In all the episodes except the second, the model-generated feedback model can successfully approximate the human stop-feedback; indeed the medians of the human time steps distributions are very close and sometimes overlap with at least one model-generated feedback time step.

We conducted further analysis concerning human stop-feedback at the individual level, cf. the extended paper for details [8].

## ACKNOWLEDGMENTS

This work has been funded in parts by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20058].

## REFERENCES

[1] Joshua Achiam and Dario Amodei. 2019. Benchmarking Safe Exploration in Deep Reinforcement Learning.

[2] Samuel Ainsworth, Matt Barnes, and Siddhartha Srinivasa. 2019. Mo'states mo'problems: Emergency stop mechanisms from observation. *Advances in Neural Information Processing Systems (NeurIPS)* 32 (2019).

[3] Eitan Altman. 1999. *Constrained Markov Decision Processes.* Routledge.

[4] Felix Berkenkamp. 2019. *Safe exploration in reinforcement learning: Theory and applications in robotics.* Ph.D. Dissertation. ETH Zurich.

[5] Javier Garcia and Fernando Fernández. 2012. Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research (JAIR)* 45 (2012), 515–564.

[6] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. (3 1953). https://doi.org/10.2172/4390578

[7] Martin Pecka and Tomas Svoboda. 2014. Safe exploration techniques for reinforcement learning–an overview. In *International Workshop on Modelling and Simulation for Autonomous Systems*. Springer, 357–375.

[8] Silvia Poletti, Alberto Testolin, and Sebastian Tschiatschek. 2023. *Learning Constraints From Human Stop-Feedback in Reinforcement Learning (extended version).* Technical Report.

[9] William Saunders, Girish Sastry, Andreas Stuhlmüller, and Owain Evans. 2018. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 2067–2069.

[10] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[11] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[12] Nolan C Wagener, Byron Boots, and Ching-An Cheng. 2021. Safe reinforcement learning using advantage-based intervention. In *International Conference on Machine Learning (ICML)*. 10630–10640.