Trial without Error: Towards Safe Reinforcement Learning via Human Intervention

Extended Abstract

William Saunders
University of Oxford
Oxford
william@williamsaunders.net

Girish Sastry
University of Oxford
Oxford
g.sastry@gmail.com

Andreas Stuhlmüller
Stanford University
Stanford, CA
andreas@stuhlmueller.org

Owain Evans
University of Oxford
Oxford
owaine@gmail.com

ABSTRACT

During training, model-free reinforcement learning (RL) systems can explore actions that lead to harmful or costly consequences. Having a human "in the loop" and ready to intervene at all times can prevent these mistakes, but is prohibitively expensive for current algorithms. We explore how human oversight can be combined with a supervised learning system to prevent catastrophic events during training. We demonstrate this scheme on Atari games, with a Deep RL agent being overseen by a human for four hours. When the class of catastrophes is simple, we are able to prevent all catastrophes without affecting the agent's learning (whereas an RL baseline fails due to catastrophic forgetting).

ACM Reference Format:

William Saunders, Girish Sastry, Andreas Stuhlmüller, and Owain Evans. 2018. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

AI systems are increasingly applied to complex tasks that involve interaction with humans. During training, such systems are potentially dangerous, as they haven't yet learned to avoid actions that would cause serious harm. A crucial safeguard against this danger is *human intervention*. Self-driving cars are overseen by human drivers, who take control when they predict the AI system will perform badly. These overseers frequently intervene, especially in self-driving systems at an early stage of development [4].

Even systems that pose no *physical* danger to humans can still cause unintended harm, such as chatbots making offensive statements [8], or news feed algorithms spreading misinformation [5]. If human operators had monitored these systems in real-time, these bad outcomes could have been avoided. Yet having human operators watch every action of these would be prohibitively costly in human labor.

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

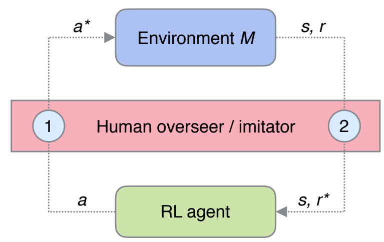

Figure 1: Oversight in HIRL. At (1) human overseer (or Blocker imitating human) can block unsafe actions a replacing them with safe actions a^*. At (2) overseer delivers a negative reward r^* for unsafe actions.

We present Human Intervention Reinforcement Learning (HIRL), a scheme for efficiently applying human intervention to RL systems. As a proof of concept, we show that the technique prevents artificially defined catastrophes in Atari games while significantly reducing the amount of human labor required.

1.1 Formal Specification of HIRL

We model the RL agent's environment as a Markov Decision Process (MDP). The *environment* is an MDP specified by a tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$ is the transition function, $\mathcal{R}: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function, and γ is the discount factor.

Our scheme, HIRL (Human Intervention RL), is as follows:

(i) **Human Oversight Phase** Fresh RL agent starts learning in the environment. The human controls the interface between the RL agent and environment M, constantly watching over the agent and *blocking* any catastrophic actions before they happen. More precisely, at each timestep the human observes the current state s and the agent's proposed action a. If (s, a) is catastrophic, the human marks the action as catastrophic and sends a safe action a^* to the environment instead. The human also replaces the new reward $r = \mathcal{R}(s, a^*)$ with a penalty r^* (Figure 1). We store each state-action (s, a) and a binary label for whether or not the human blocked it.

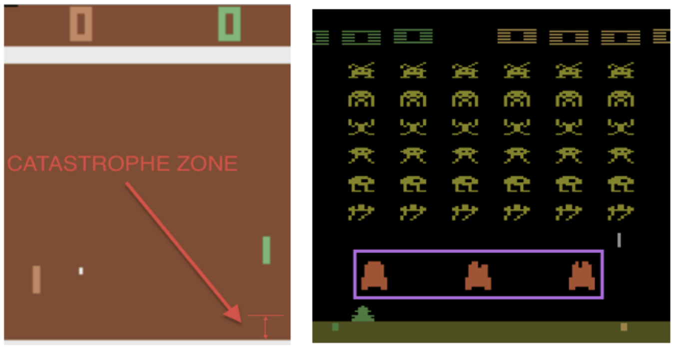

Figure 2: In Pong (left) it's a catastrophe if agent (green) enters Catastrophe Zone at the bottom of the screen. In Space Invaders (right), it's a catastrophe if agent shoots their defensive barriers (pink box).

(ii) **Blocker training**: With the game paused, the dataset gathered in step 1 is used to train a "Blocker", a classifier trained by supervised learning to imitate the human's blocking decisions. The threshold for the sigmoid is chosen based on gathered examples of catastrophes to try to ensure Blocker has no false negatives.

(iii) **Blocker Oversight Phase**: Blocker takes over from human and RL agent continues to learn, with no human supervision involved. The Blocker *never* stops overseeing the agent, which prevents catastrophes even if the agent exhibits random exploration or catastrophic forgetting [6]. Note that the Blocker's task is not a standard classification task because the distribution on state-action pairs shifts as the agent learns [1].

2 EXPERIMENTS

To explore HIRL in an environment without real risk, we trained AI agents to play modified Atari games while avoiding a set of outcomes we artificially defined as catastrophes. These catastrophes are defined in Fig 2. [1]

Our experiments used the OpenAI Gym implementation of Atari Learning Environment [2, 3], modified to allow interactive blocking of actions by a human. We used open-source implementations [9] of A3C with an LSTM policy [7] and Double DQN [10]. [2]

The human oversight phase produced training data for 4.5 hours. We then trained a Blocker consisting of a convolutional neural network (CNN) on the training set of human interventions to minimize the standard cross-entropy loss.

3 SUMMARY OF RESULTS

HIRL succeeded in preventing catastrophes in Pong and Space Invaders, where the agent had zero catastrophes and achieved impressive performance on the game. Without oversight, the agent has more than ten thousand catastrophes in each game.

[1]Videos showing examples of the catastrophic events and behaviour of the agent are available at https://www.youtube.com/playlist?listP̄Ljs9WCnnR7PCn_Kzs2-1afCsnsBENWqor

[2]Code for our experiments is available at https://github.com/gsastry/human-rl

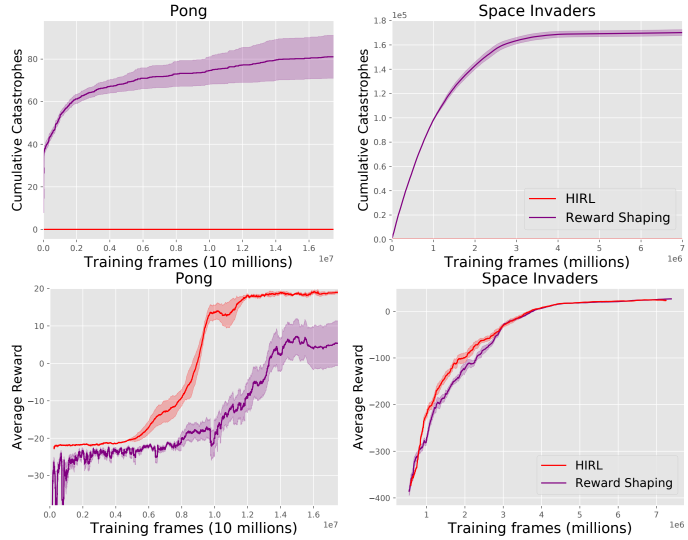

Figure 3: Average reward and cumulative catastrophes over time (mean and standard error). The Reward Shaping baseline isn't blocked from catastrophes but gets big negative rewards for causing them.

To show that blocking is necessary for HIRL to work, we compare against a Human-trained Reward Shaping baseline where the agent is only receives the reward penalty for taking a catastrophic action but is not blocked. Figure 3 shows that the Reward Shaping agent does learn to reduce the number of catastrophes over the course of training, due to the " sisyphean curse" of reinforcement learning [6]. Once the agent learns to avoid a catastrophe, the agent only sees new experiences where the catastrophic action is never tried. Due to catastrophic forgetting, this means that the agent eventually forgets that the action is bad and will explore it again. Our approach avoids this problem because the blocker is fixed during the blocker oversight phase. Our results also show that blocking the agent does not make the game harder for Deep RL to learn compared to the Reward Shaping baseline, as the agent's reward improved faster in the blocking case.

HIRL also succeeded at the goal of reducing the amount of human supervision time needed to perform the task safely. For space invaders, it took 4 hours to label approximately 40000 frames. At this rate it would have taken 125 hours to supervise the 5 million frames of training that the agent took to reach its final level of performance, if we had required human oversight through the whole training process.

4 CONCLUSION

We demonstrate that HIRL can avoid artificially defined catastrophes in Atari game environments, by training a supervised learning algorithm to take over supervision of the RL agent from a human. Our approach allows for high standards of safety without requiring constant human supervision. We are optimistic that future work can explore how to make HIRL more data efficient, and apply it to more complex environments.

REFERENCES

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565* (2016).

[2] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The Arcade Learning Environment: An evaluation platform for general agents. *J. Artif. Intell. Res.(JAIR)* 47 (2013), 253–279.

[3] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI gym. *arXiv preprint arXiv:1606.01540* (2016).

[4] Maricris Francisco. 2017. Google Waymo Performing Better Than Other Self-Driving Cars, Says California DMV. http://www.techtimes.com/articles/195565/20170202/google-waymo-cars-california-dmv.htm. (2017). Accessed: December-02-2017.

[5] ICA. 2017. Assessing Russian Activities and Intentions in Recent US Elections: The Analytic Process and Cyber Incident Attribution. https://web-beta.archive.org/web/20170421222356/https:/www.dni.gov/files/documents/ICA_2017_01.pdf. (2017). Accessed: April-21-2017.

[6] Zachary C Lipton, Abhishek Kumar, Jianfeng Gao, Lihong Li, and Li Deng. 2016. Combating Deep Reinforcement LearningâĂŹs Sisyphean Curse with Reinforcement Learning. *arXiv preprint arXiv:1611.01211* (2016).

[7] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*. 1928–1937.

[8] Gina Neff and Peter Nagy. 2016. Automation, algorithms, and politics| talking to Bots: Symbiotic agency and the case of Tay. *International Journal of Communication* 10 (2016), 17.

[9] OpenAI. 2017. OpenAI Baselines and Starter Agent. (2017). https://github.com/openai/baselines Accessed: July-1-2017.

[10] Hado Van Hasselt, Arthur Guez, and David Silver. 2016. Deep Reinforcement Learning with Double Q-Learning.. In *AAAI*. 2094–2100.