Trial without Error: Towards Safe Reinforcement Learning via Human Intervention

Extended Abstract

William Saunders
University of Oxford
Oxford
william@williamsaunders.net

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

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

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:

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.

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 \( M = (S, A, T, R, \gamma) \), where \( S \) is the state space, \( A \) is the action space, \( T : S \times A \times S \mapsto [0, 1] \) is the transition function, \( R : S \times A \mapsto \mathbb{R} \) is the reward function, and \( \gamma \) 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 \( r = R(s, a^*) \) 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.
We then trained a Blocker consisting of a convolutional neural network (CNN) on the training set of human interventions to minimize 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. \(^1\)

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.

\section{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.

\section{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.

\section{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


