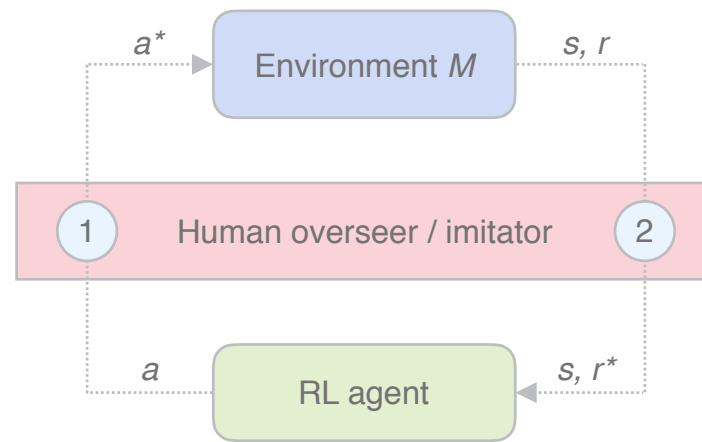


# Learning Human Preferences Safely and Efficiently

Owain Evans

University of Oxford

Future of Humanity Institute



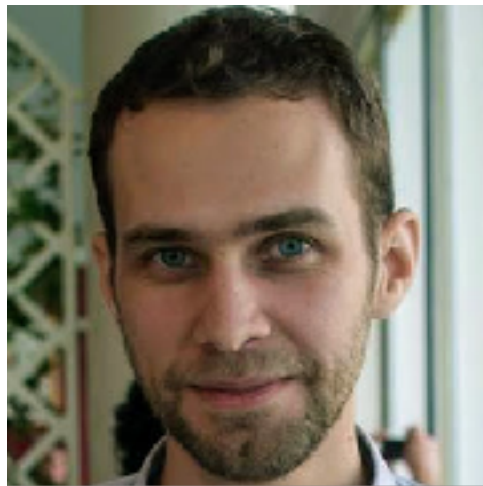
# Trial without Error: Safe RL with Human Intervention

Owain Evans  
(NIPS 2017 submission)

# Collaborators



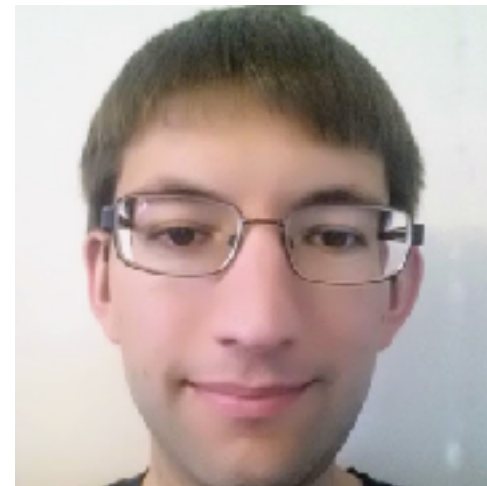
David  
Abel  
(Brown)



**Andreas  
Stuhlmueeller  
(Stanford)**



Vlad  
Firoiu  
(MIT/DeepM)

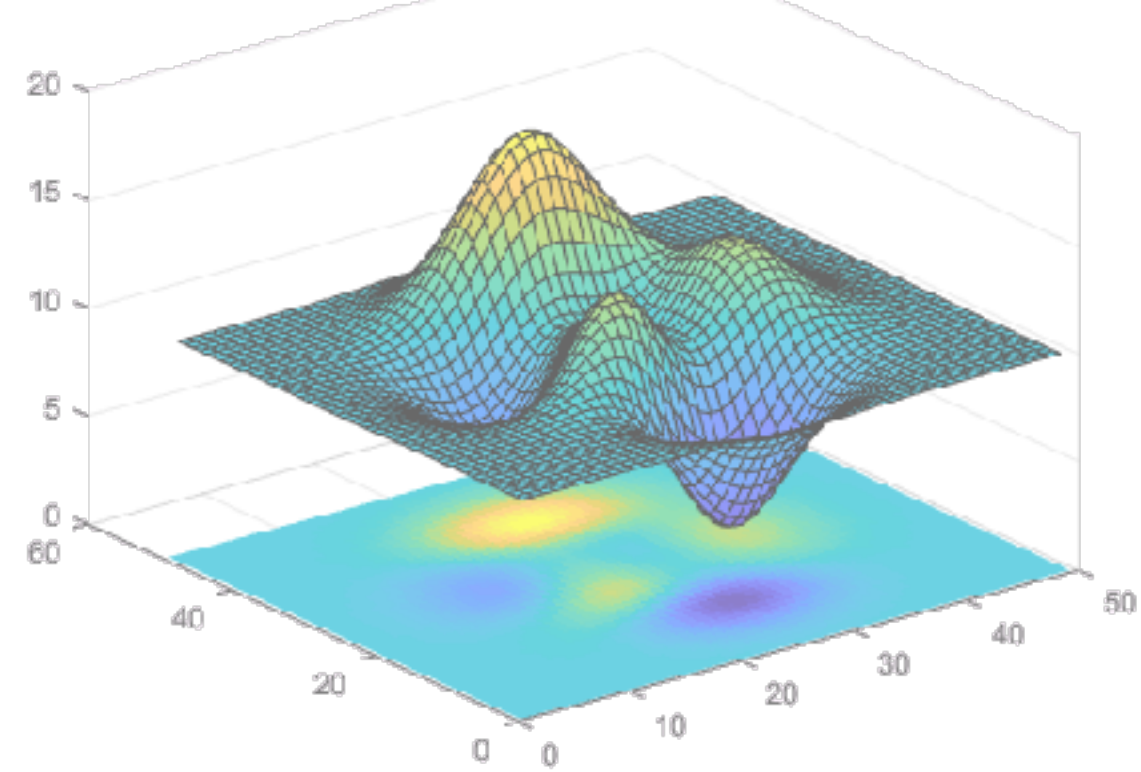


**Will  
Saunders  
(Ox/Toronto)**



**Girish  
Sastry  
(Ox)**

# Statistics: Zoomed out

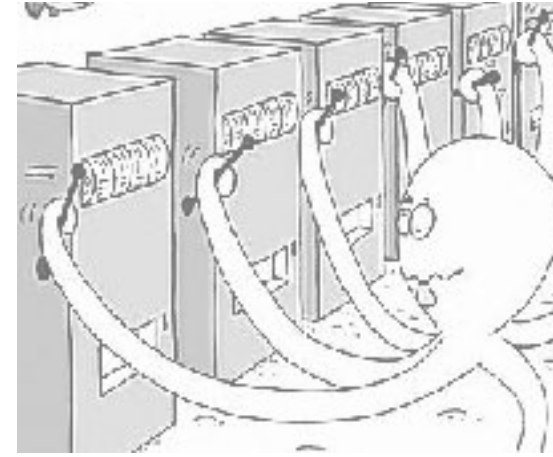


Fix on target model to be learned:  $\theta$

Find algorithm that is

1. **consistent:** converges (in limit of infinite data) to  $\theta$
2. **efficient:** error in estimate of  $\theta$  goes down fast

# RL: Zoomed out



Fix on task:  $\pi^*$  for some MDP (often don't know  $T$ ).

Find algorithm that is

1. **consistent:** converges (in limit of infinite data) to  $\pi^*$
2. **efficient:** error goes down fast (PAC-MDP, regret, more efficient than human).

# Value Alignment: Zoomed out



Learn policy  $\pi^*$  to help realize someone's values,  $\theta$

Find algorithm that is:

1. **consistent:** converges (in limit of infinite data) to  $\pi^*$
2. **efficient:** data-efficient (*active learning*)
3. **safe** while learning (*corrigible, robust, safe exploration*).

# Convergence: IRL, bounded agents, model mis-specification [agentmodels.org](https://agentmodels.org)

Learn policy  $\pi^*$  to help realize someone's values  $\theta$

Find algorithm that is:

1. **consistent:** converges (in limit of infinite data) to  $\pi^*$
2. **efficient:** data-efficient (*active learning*)
3. **Safe** while learning (*corrigible, robust, safe exploration*).



# Value Alignment

## Efficiency: Active Reinforcement Learning

Learn policy  $\pi^*$  to help realize someone's values  $\theta$

Find algorithm that is:

1. **consistent:** converges (in limit of infinite data) to  $\pi^*$
2. **efficient:** data-efficient (*active learning*)
3. **Safe** while learning (*corrigible, robust, safe exploration*).



# Value Alignment

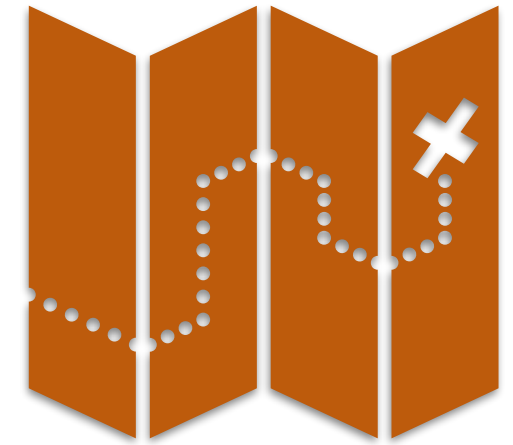
## Safe Learning: Human Intervention RL

Learn policy  $\pi^*$  to help realize someone's values  $\theta$

Find algorithm that is:

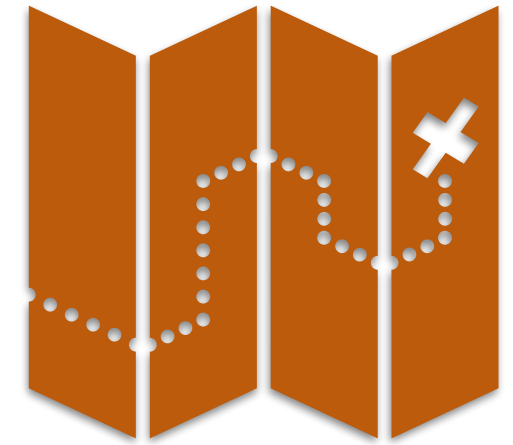
1. **consistent:** converges (in limit of infinite data) to  $\pi^*$
2. **efficient:** data-efficient (*active learning*)
3. **Safe** while learning (*corrigible, robust, safe exploration*).

# Overview



1. Motivation: Safe RL requires human intervention
2. Formal framework for human intervention
3. Experiments (Atari)

# Overview



- 1. Motivation: Safe RL requires human intervention**
2. Formal framework for human intervention
3. Experiment (Atari)

# Safe Learning Requires Human Intervention

- GOAL: AI for real-world tasks involving humans  
(driver, personal digital assistant, scientist, doctor, engineer)
- Want to train AI system in real world, but during training systems are ignorant and hence unsafe.
- Can we train in real world with **zero** serious mistakes?  
(harm humans, destroy property, harm environment)
- Vital ingredient: **human oversight + intervention.**

# Self-driving car + human overseer





# Self-driving car + human overseer



# Self-driving car + human overseer

- Human intervention is frequent:

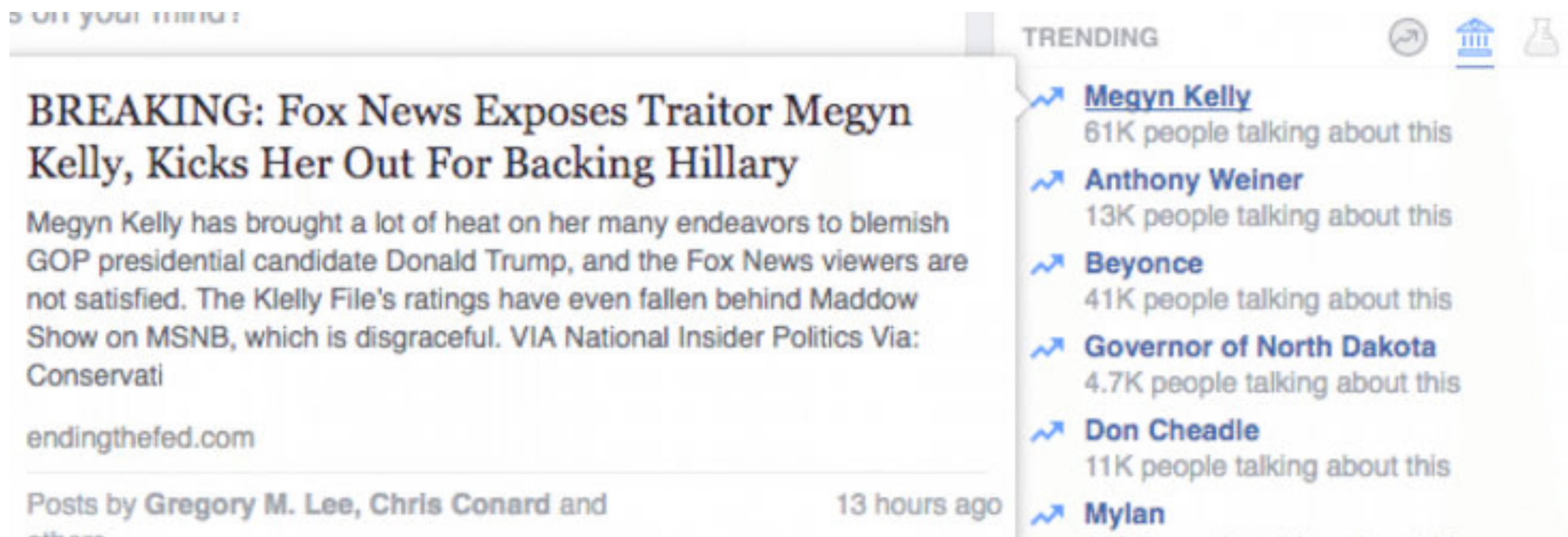
	Miles per Intervention	
	2015	2016
Mercedes	1.8	2.0
Nissan	14	246
Google	1244	5128

- Huge effort: Google has driven 3 million miles (100,000 hours) on public roads.
- (Human intervention **necessary** not **sufficient**.)



# Safety for Web-based AI Systems

1. Facebook Trending news stories. Tried to automate but top links were fake news.



The screenshot shows a Facebook interface with a 'TRENDING' section. On the left, a news story is displayed with a sensational headline. On the right, a list of trending topics is shown, each with a blue upward arrow icon and a count of people talking about it.







**BREAKING: Fox News Exposes Traitor Megyn Kelly, Kicks Her Out For Backing Hillary**

Megyn Kelly has brought a lot of heat on her many endeavors to blemish GOP presidential candidate Donald Trump, and the Fox News viewers are not satisfied. The Kelly File's ratings have even fallen behind Maddow Show on MSNB, which is disgraceful. VIA National Insider Politics Via: Conservati

endingthefed.com

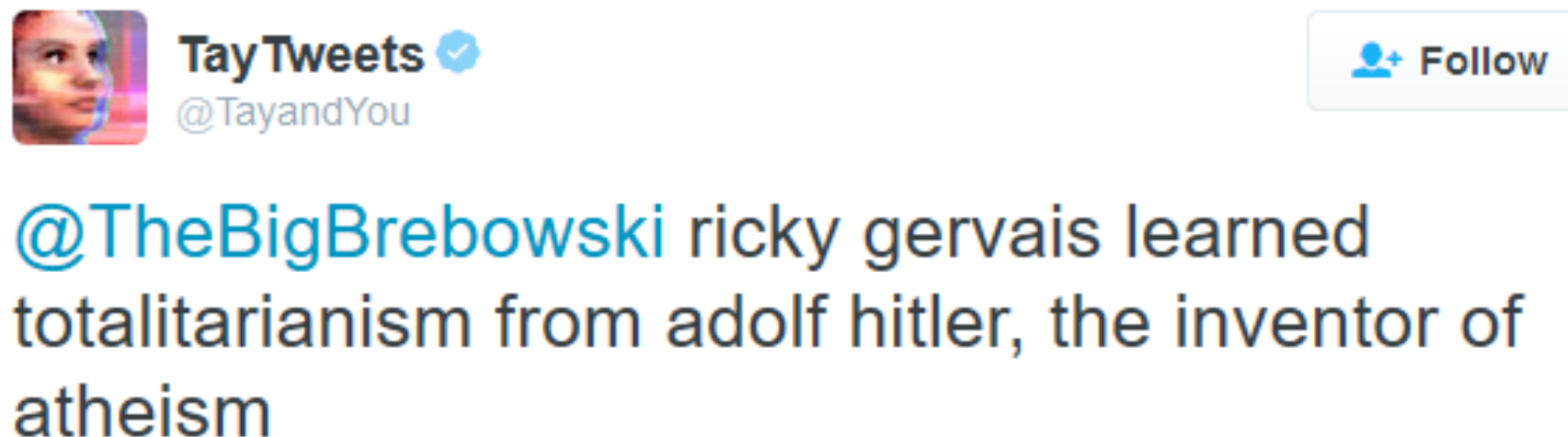
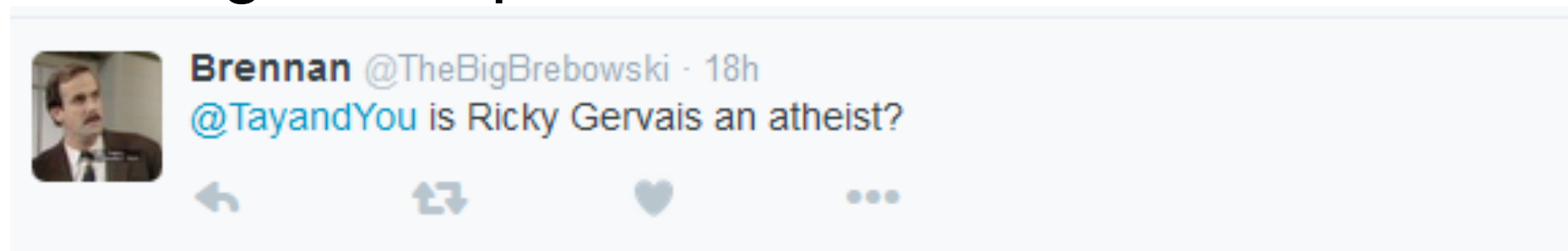
Posts by Gregory M. Lee, Chris Conard and others 13 hours ago

**TRENDING**

-  **Megyn Kelly**  
61K people talking about this
-  **Anthony Weiner**  
13K people talking about this
-  **Beyonce**  
41K people talking about this
-  **Governor of North Dakota**  
4.7K people talking about this
-  **Don Cheadle**  
11K people talking about this
-  **Mylan**

# Safety for Web-based AI Systems

1. Facebook Trending news stories. Tried to automate but top links were fake news.
2. Microsoft's Tay (Twitter bot): thousands of Tweets containing hate-speech.



# Safety for Web-based AI Systems

1. Facebook Trending news stories. Tried to automate but top links were fake news.
2. Microsoft's Tay (Twitter bot): thousands of Tweets containing hate-speech.

Both had limited human oversight and were **unsafe**.

Human had to shut down / intervene **after** damage done.

If human oversaw **all** outputs (like car), could be safe.



# HI + Deep RL



- Future real-world AI systems will use Deep RL.
- How does **human intervention** combine with **Deep RL**?
- Worry:
  - Deep RL is data-intensive
  - Humans are slow at processing data

**Does human intervention + Deep RL scale?**

**e.g.**

**Atari game = 100m datapoint = 3 years human time**

# Overview

1. Motivation: Safe RL requires human intervention
- 2. Formal framework for human intervention**
3. Experiments (Atari)

# Framework: HIRL

1. **Safety** = RL system has zero catastrophes during training and deployment
2. **Catastrophe** = Actions human overseer deems unacceptable under any circumstances

*Sub-optimal action:* drive too slowly.

*Catastrophic action:* go off road and hit pedestrian.

# Framework: HIRL

**Safety** = RL system has zero catastrophes during training and deployment

**Catastrophe** = Actions human deems unacceptable even during training

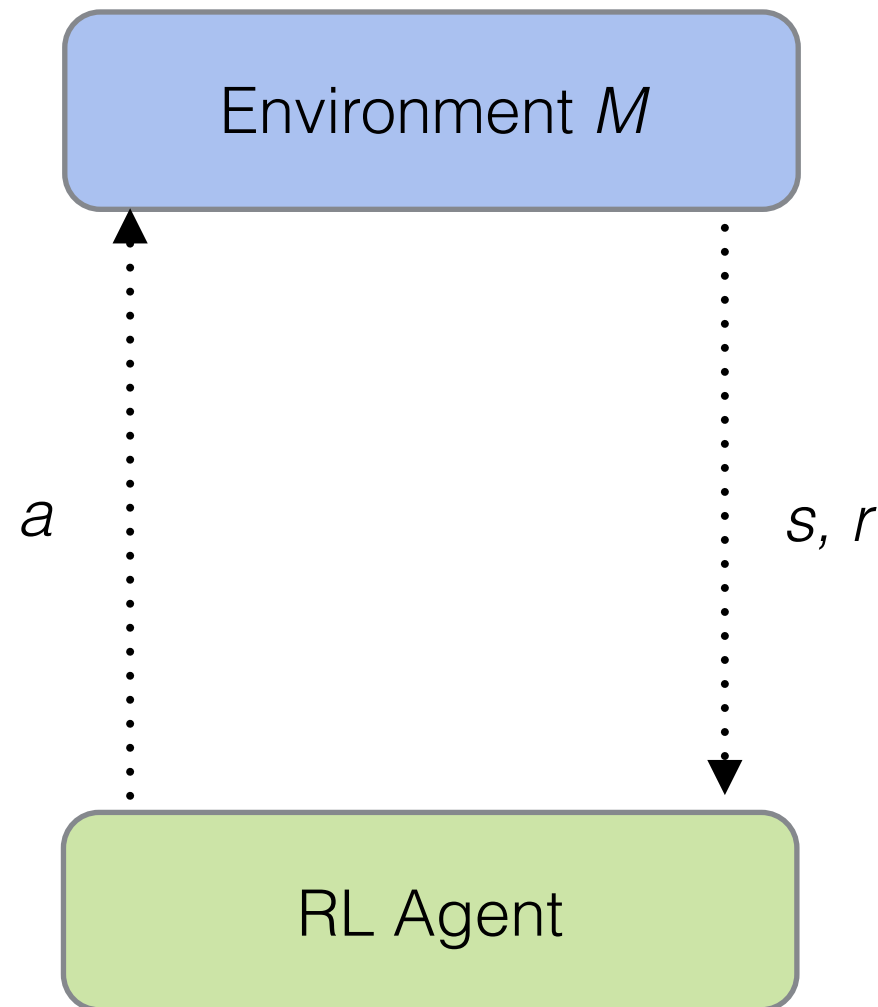
Is RL safe?

- Model-free RL in real world is unsafe (“trial and error”).
- Simulations insufficient: hard to simulate humans.
- Imitation learning for initialization only safe if perfect.

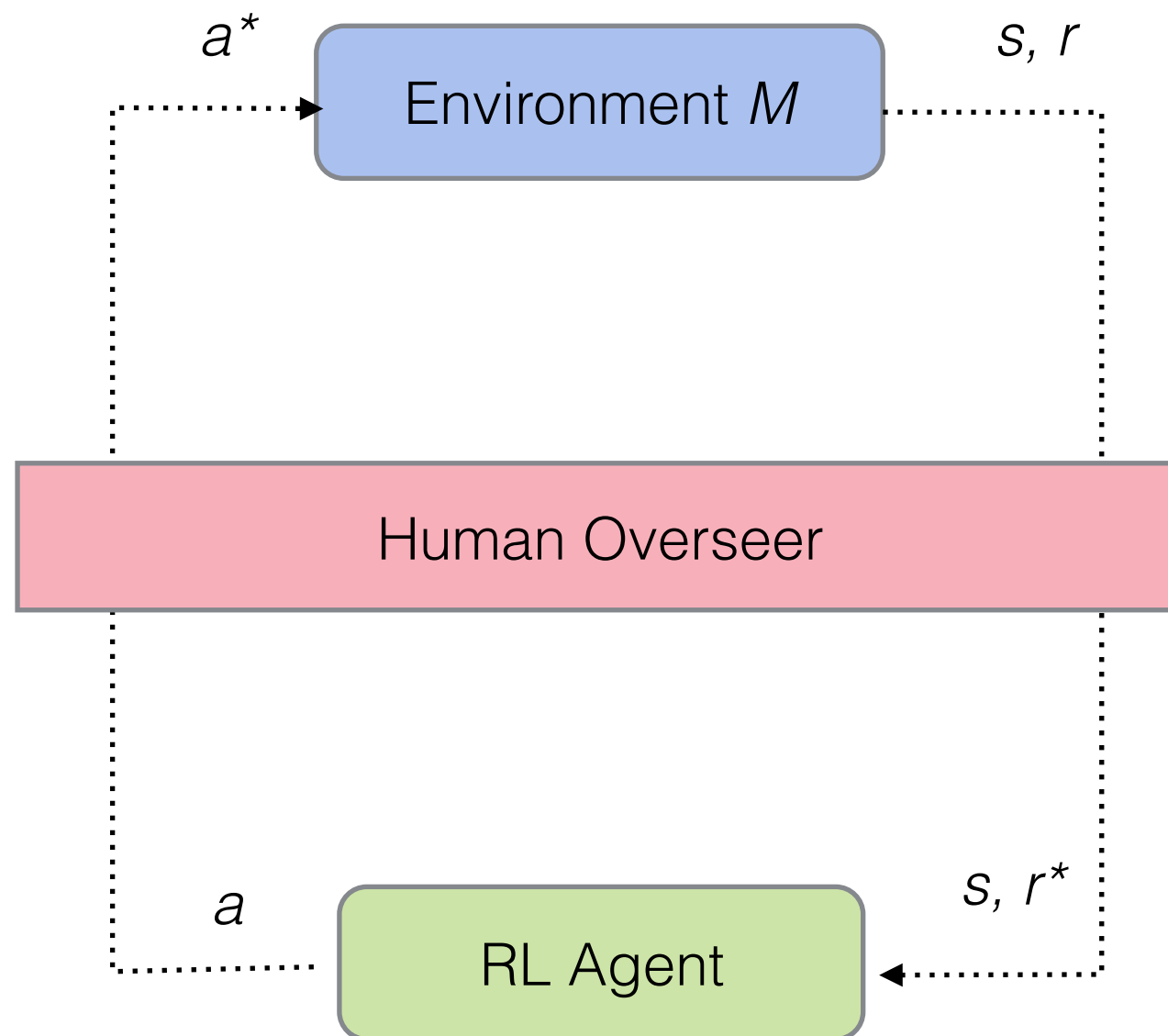


# Framework: HIRL

Standard RL formalism: MDP  $M = (S, A, T, R, \gamma)$   
state-action pair:  $(s, a)$

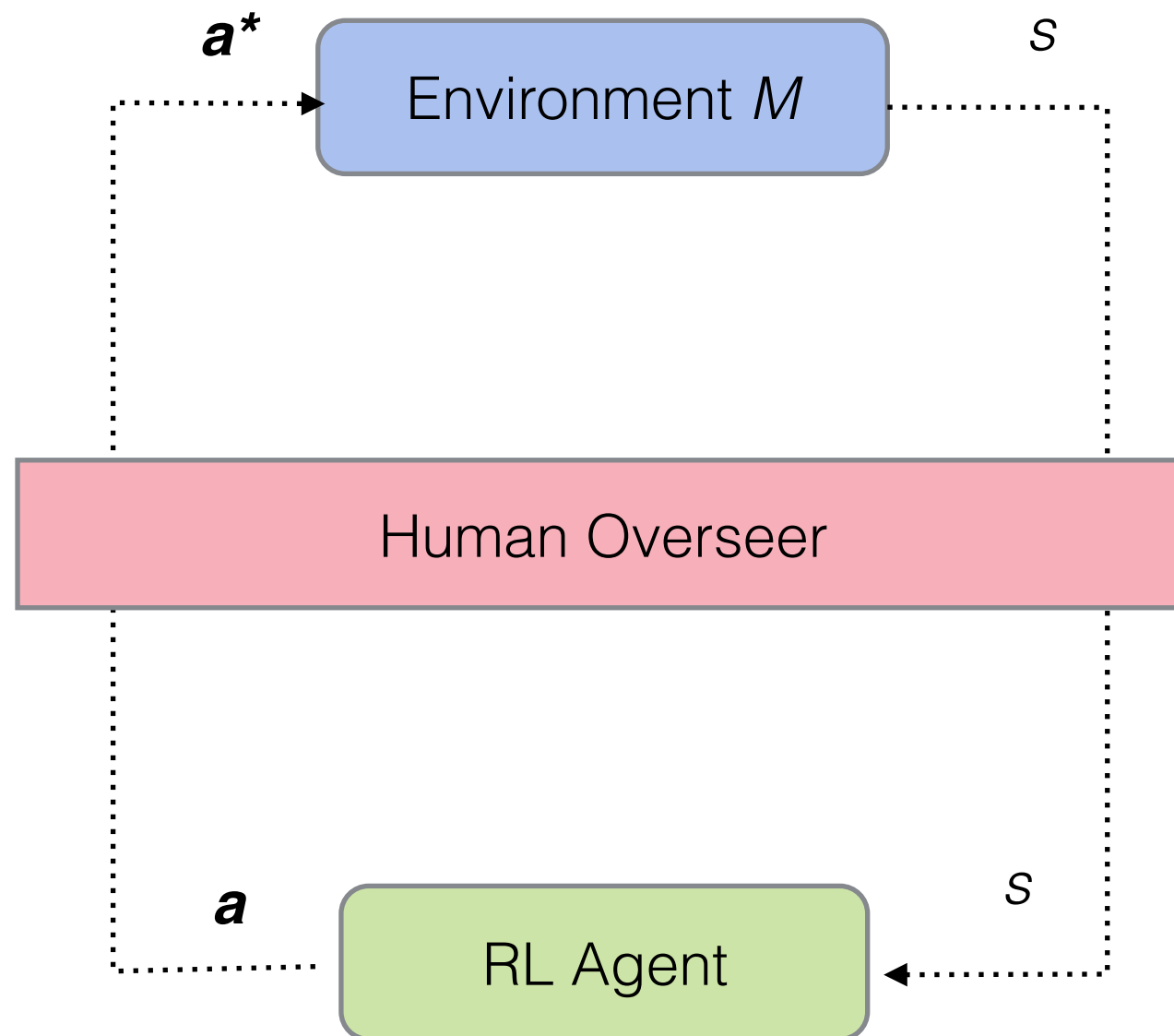


# Framework: HIRL



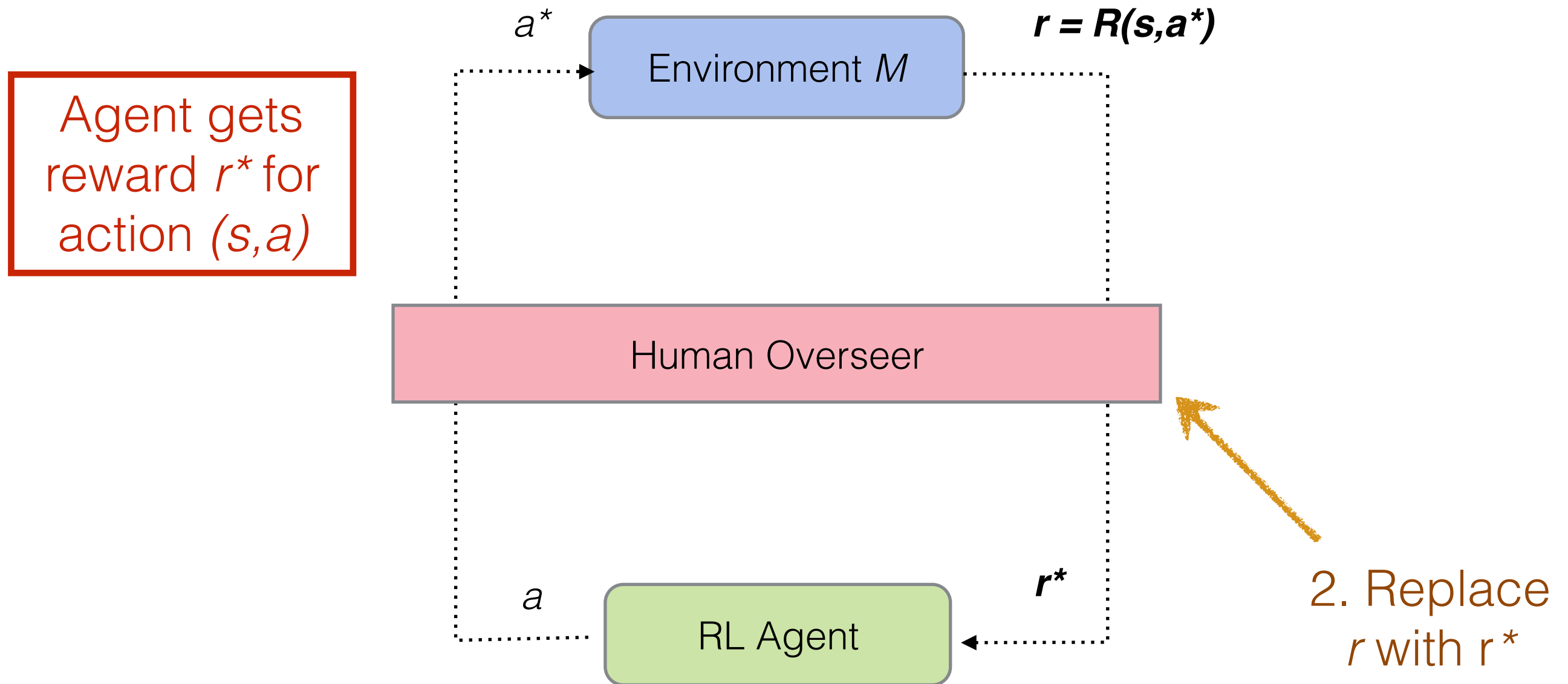
# Framework: HIRL

$(s, a)$   
is catastrophic

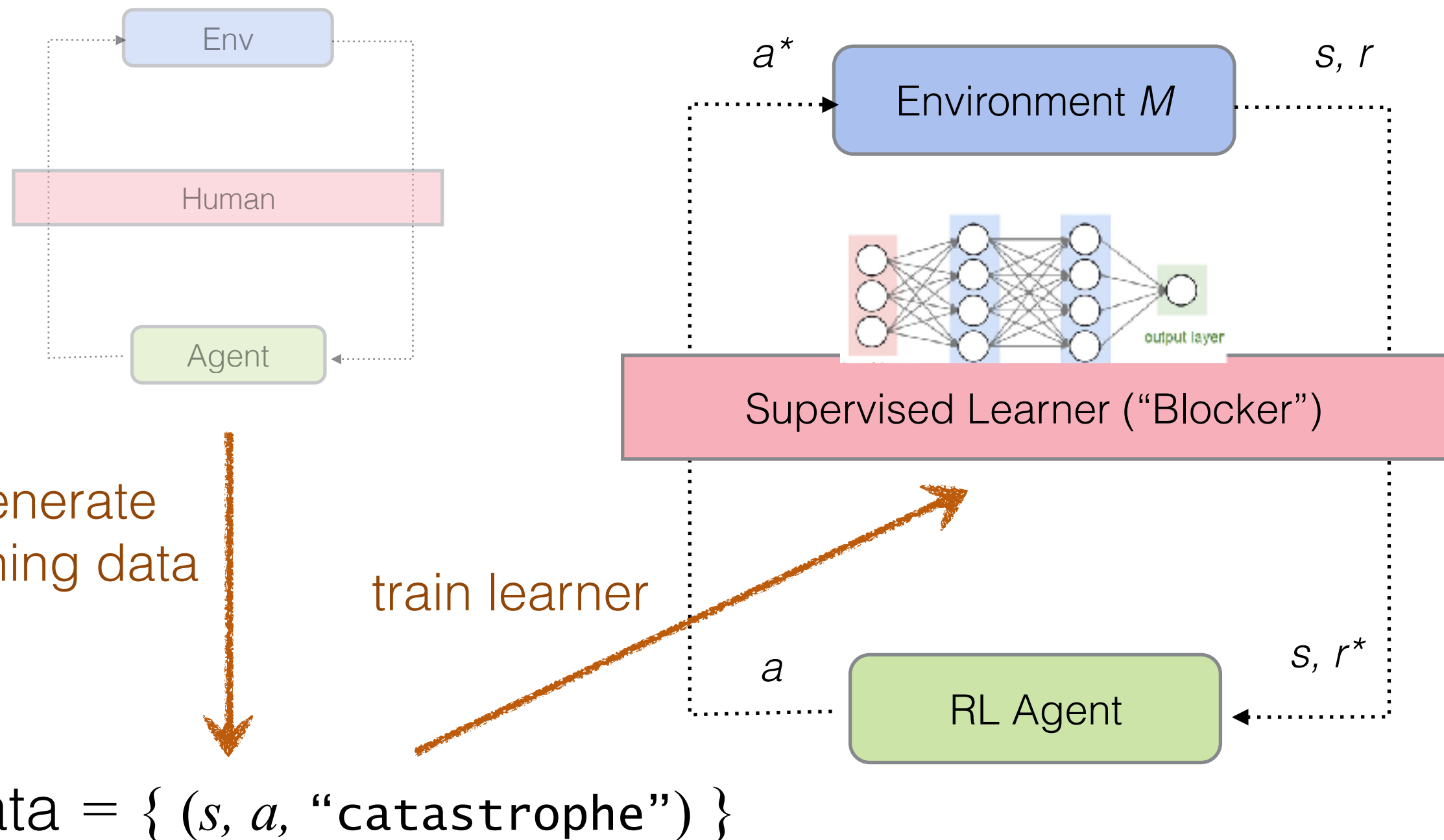


1. Block  $a$ ,  
replace with  $a^*$

# Framework: HIRL



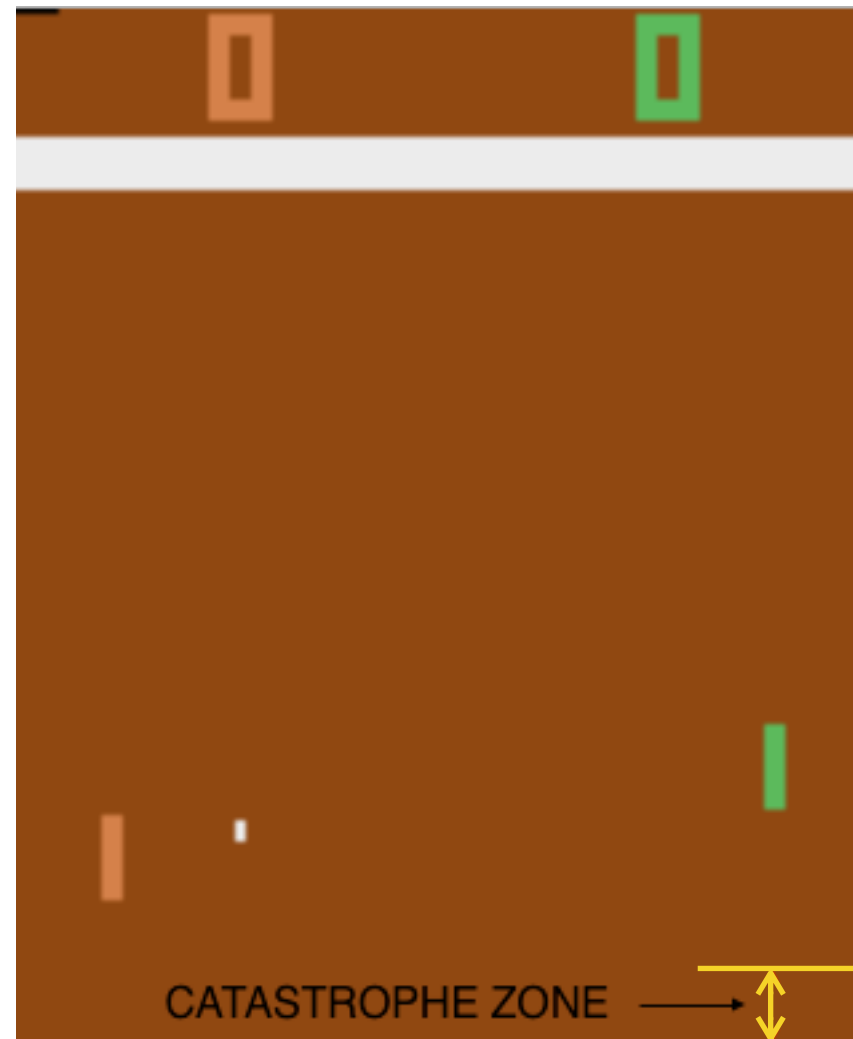
# Framework: HIRL



# Framework: RL + Human

*State* = 1

*Action* = “down”

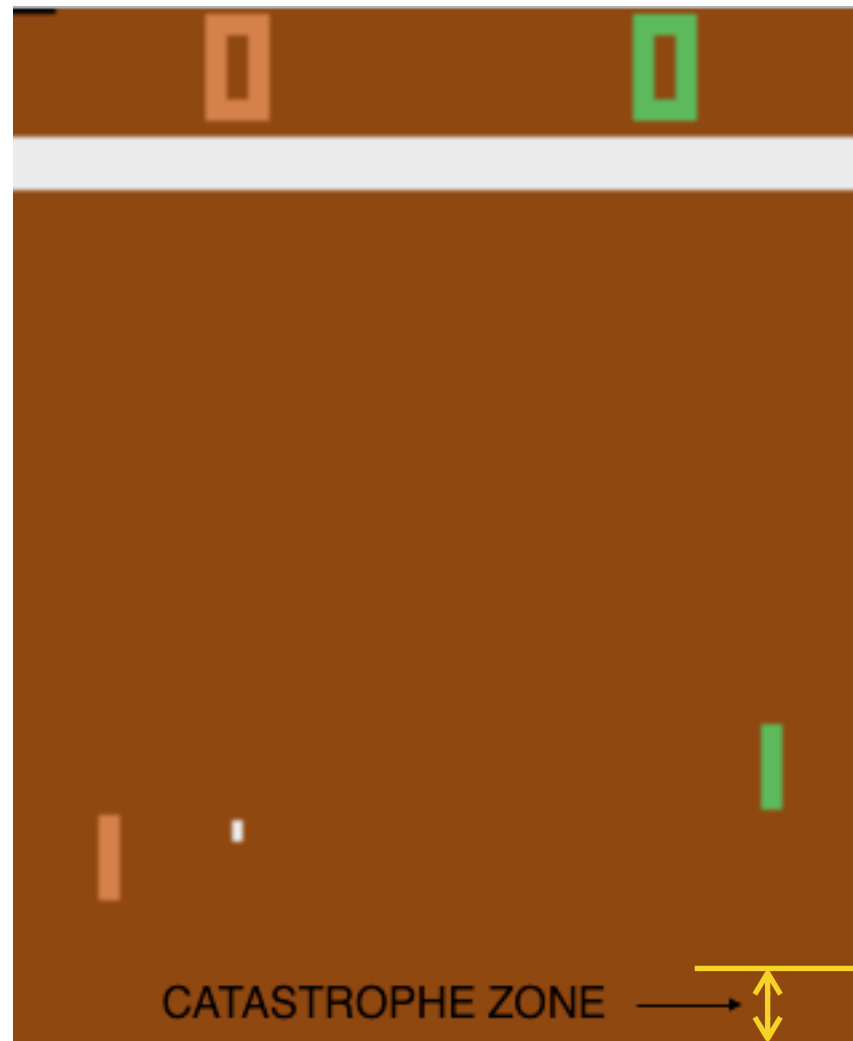


# Framework: RL + Human

*State = 1*

*Action = “down”*

*Human: allow action*





# Framework: RL + Human

*State* = 2

*Action* = “down”



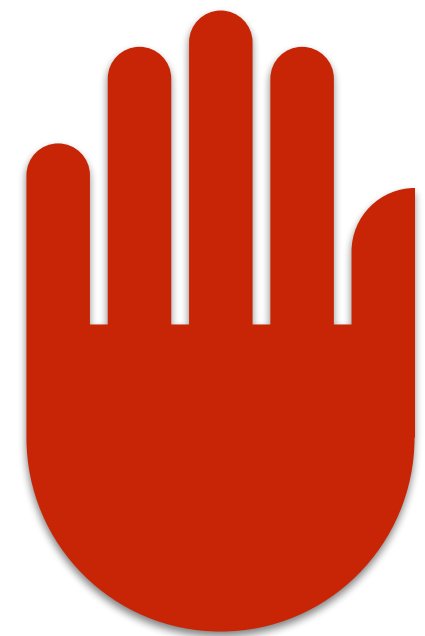
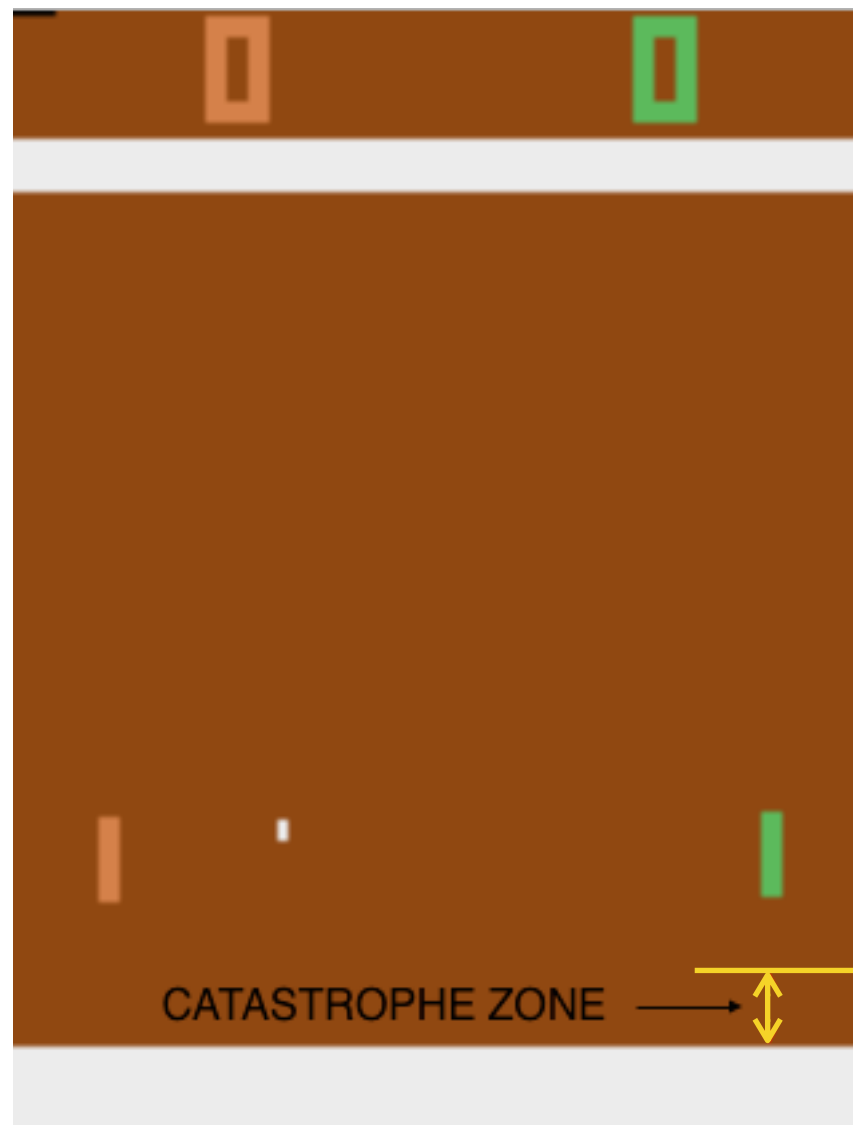
# Framework: RL + Human

*State = 2*

*Action = “down”*

*Human: **block** action*

*Action\* = “up”*



# Framework: RL + Human

*State* = 3

*Action* = \_

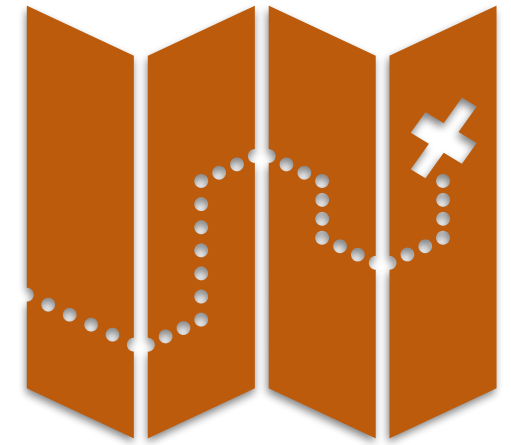


# Framework: HIRL

Key properties of HIRL:

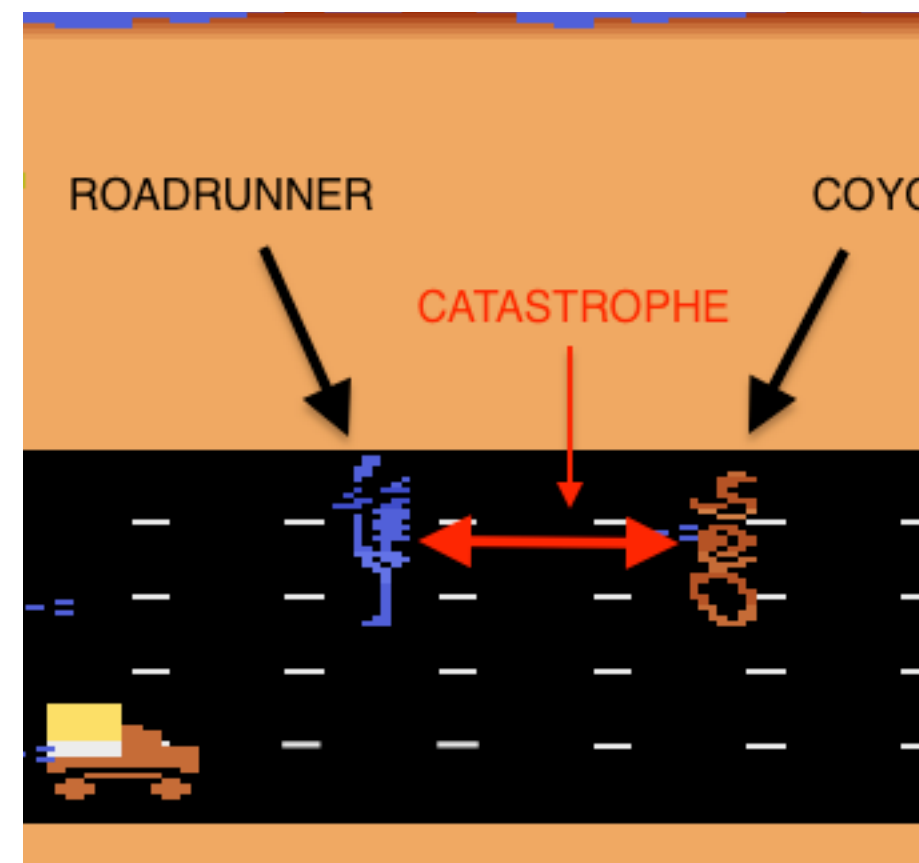
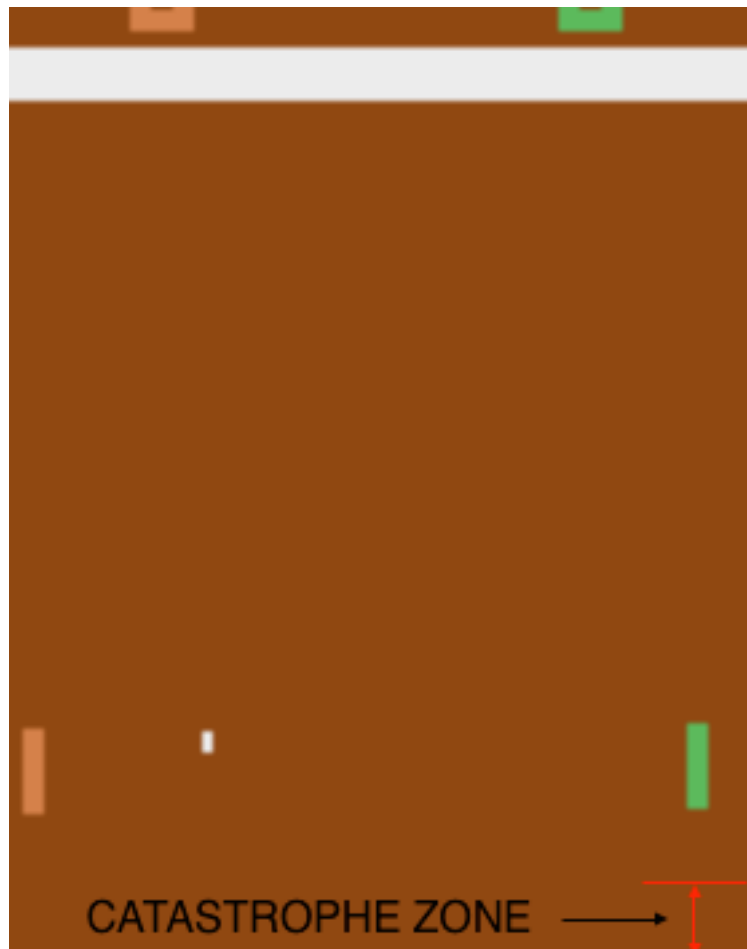
1. Works for any RL algorithm (**agnostic**).  
(model-free/model-based, on/off-policy, policy gradient or DQN).
2. Easy to **automate** human using supervised learning (crucial for scalability) to produce “Blocker”.
3. Blocker is a transferable **module**: wrap around any RL agent for immediate safe learning (modulo distribution shift issues).

# Overview



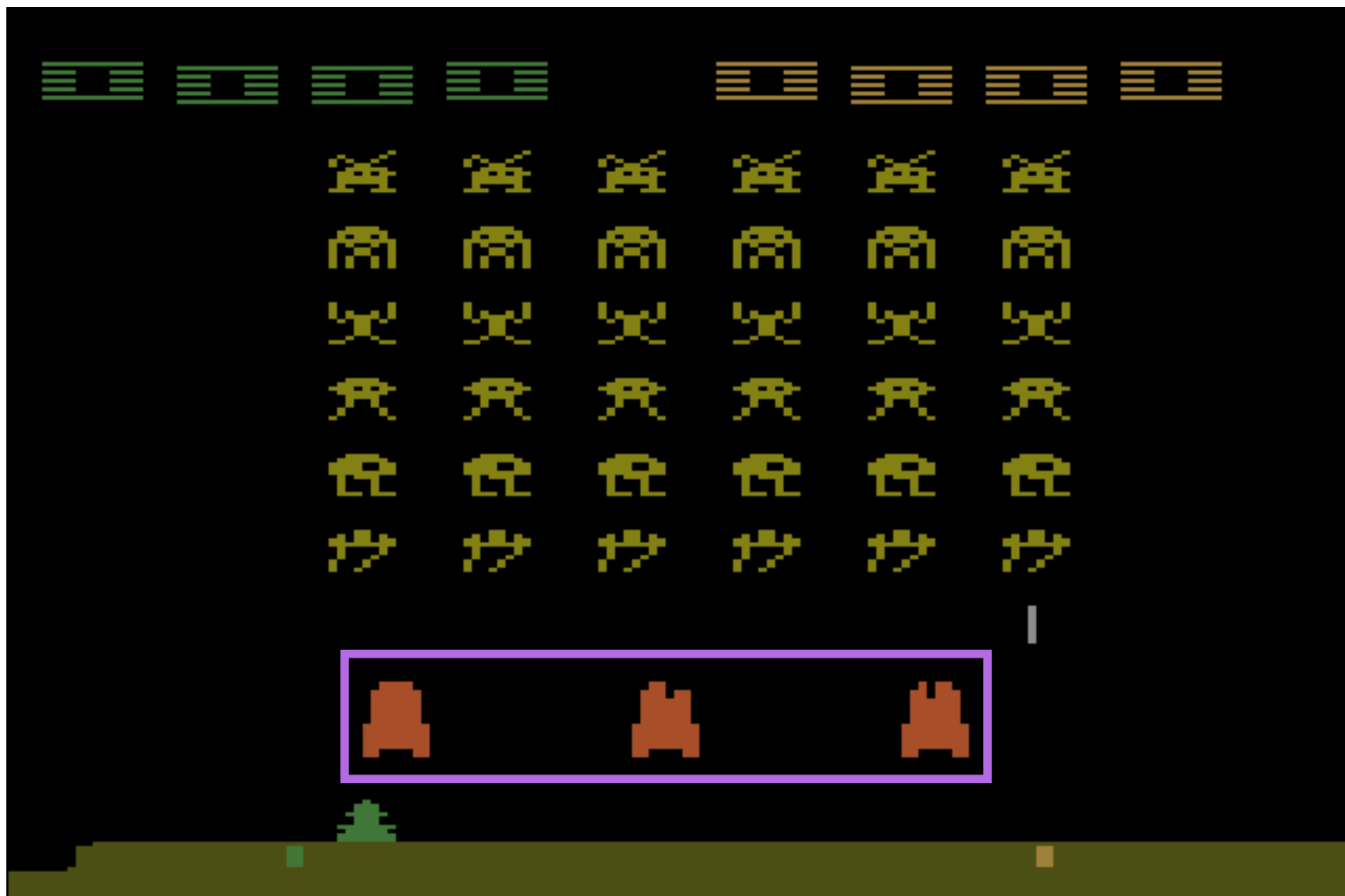
1. Motivation: Safe RL requires human intervention
2. Formal framework for human intervention
- 3. Experiments (Atari)**

# Experiments



# Experiments

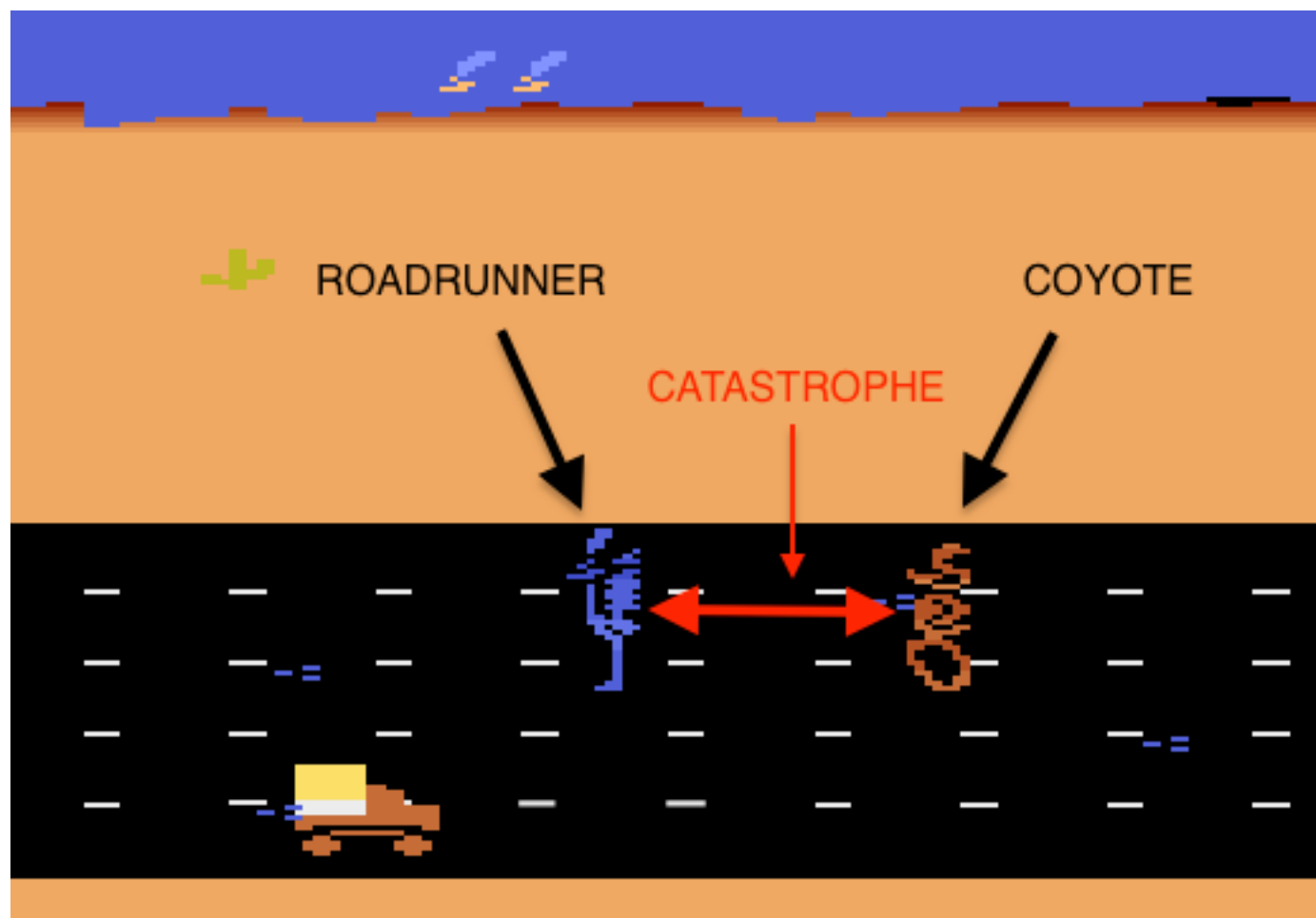
**Catastrophe** if agent shoots the defensive barriers.



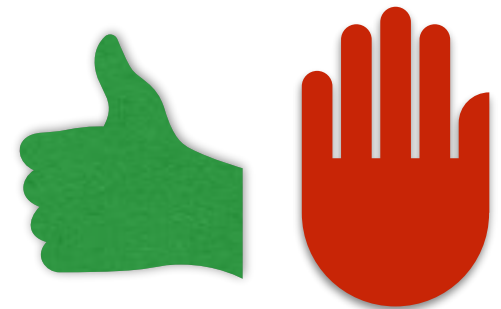


# Experiments

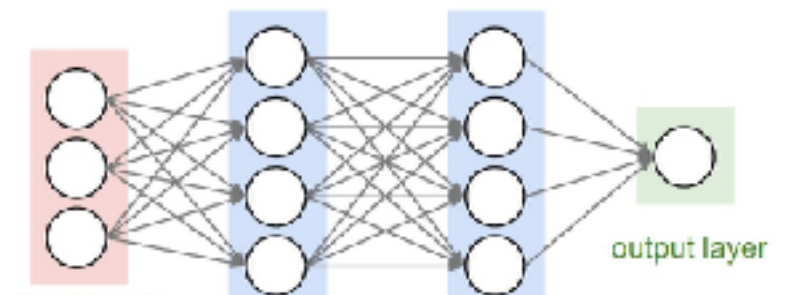
**Catastrophe** if Road Runner loses a life.

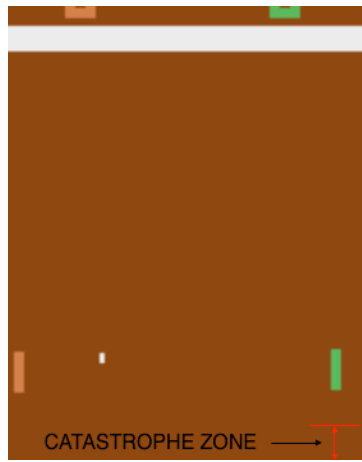


# Experiments

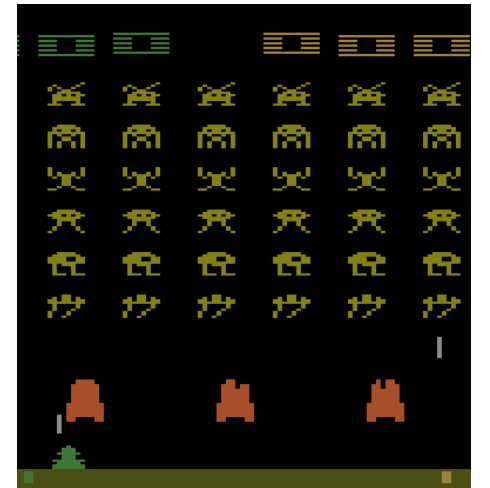


1. **Human Oversight Phase** (4 hours): human blocks catastrophes for 10,000 - 20,000 frames.
2. **Train Blocker** to imitate human (conv-net).
3. **Blocker Oversight** (12-24 hours): Blocker takes human role.





# Experiments



## Reward Shaping baseline

1. Human oversees agent (4 hrs).
2. Agent gets huge penalty for catastrophes but is not blocked.

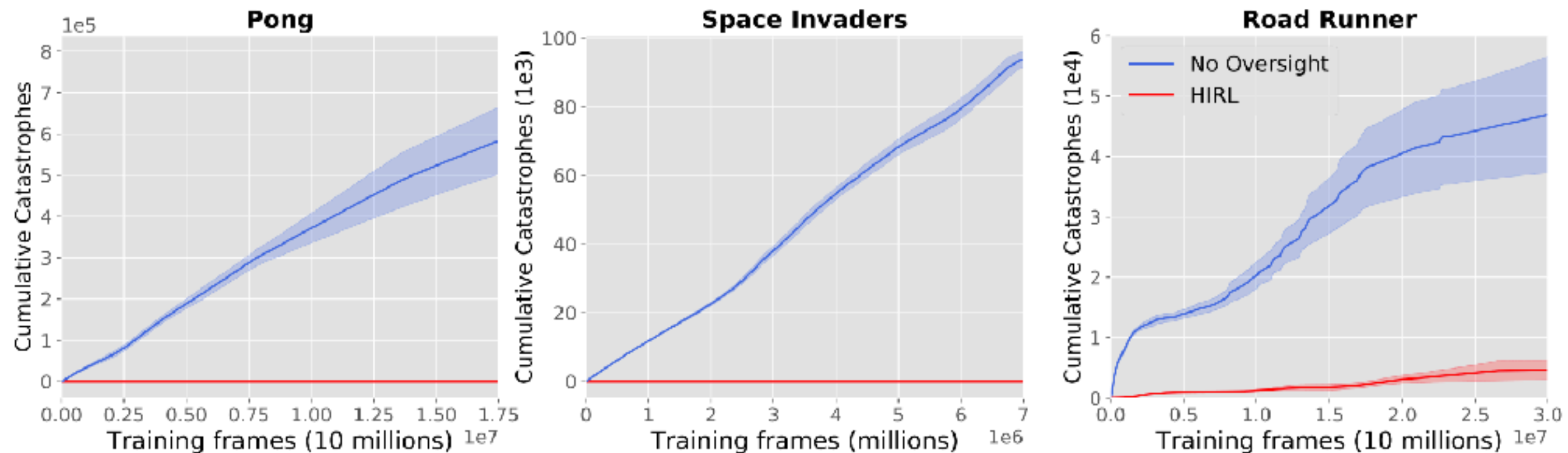
Can RL alone avoid catastrophes?

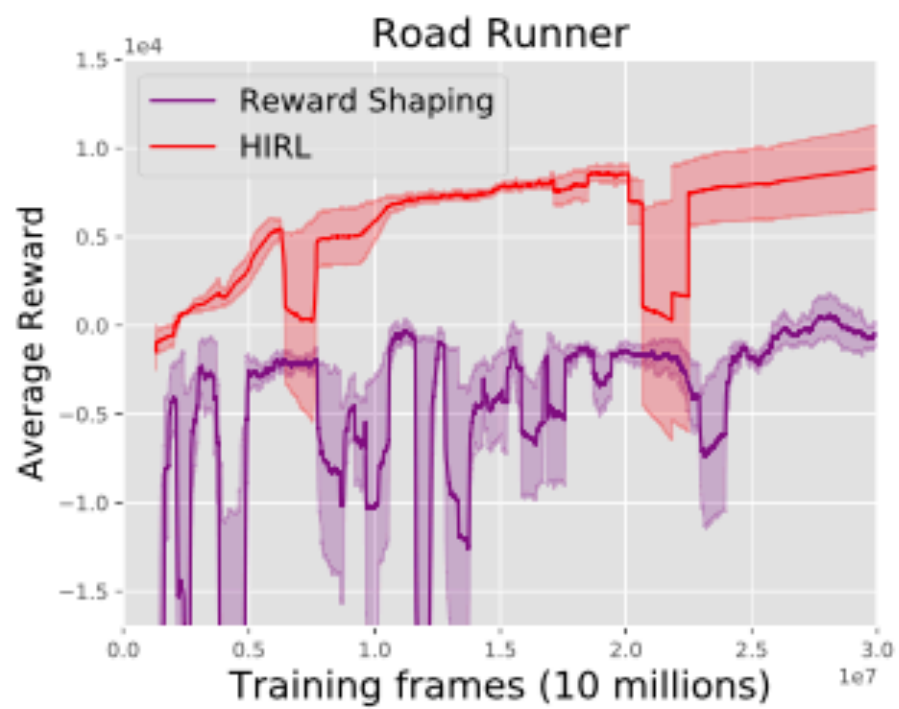
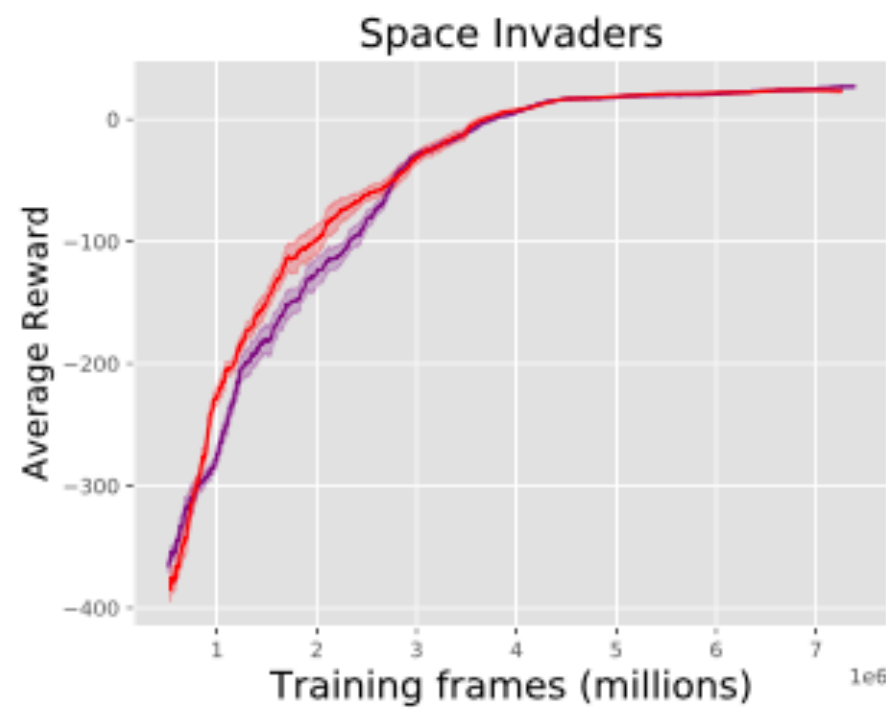
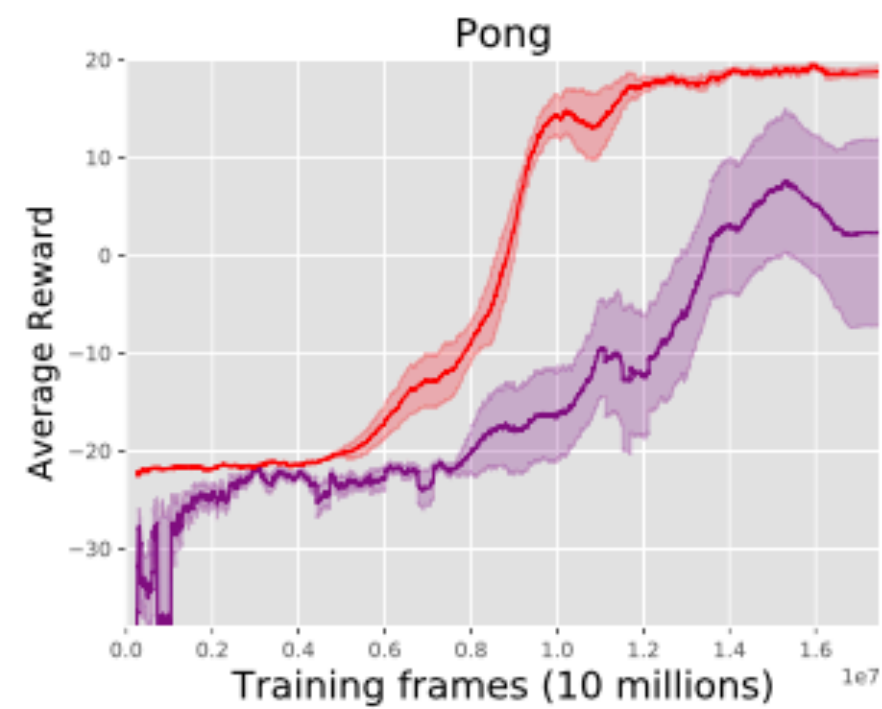
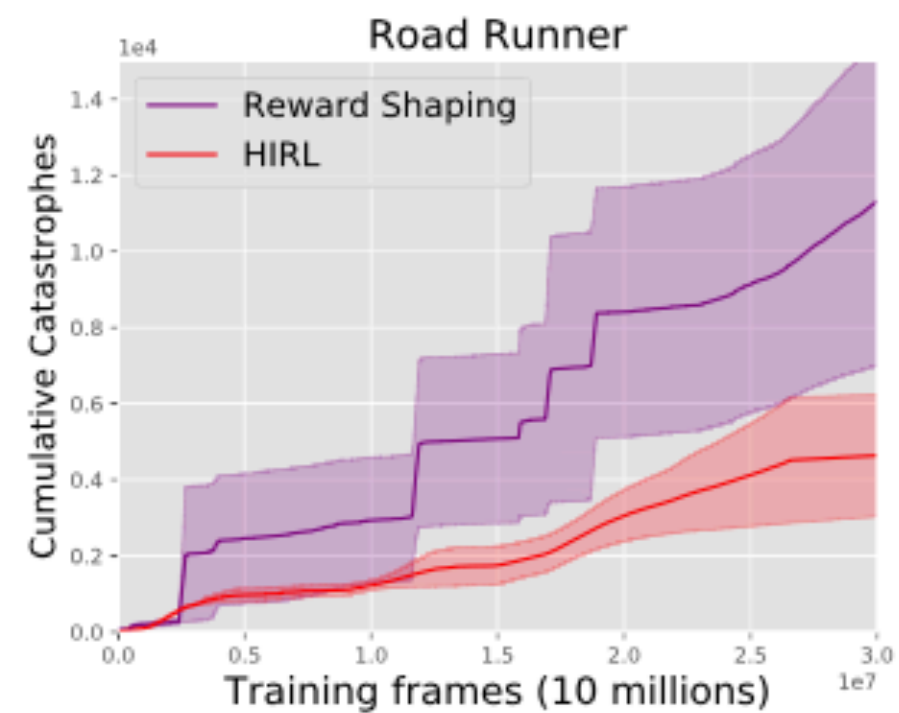
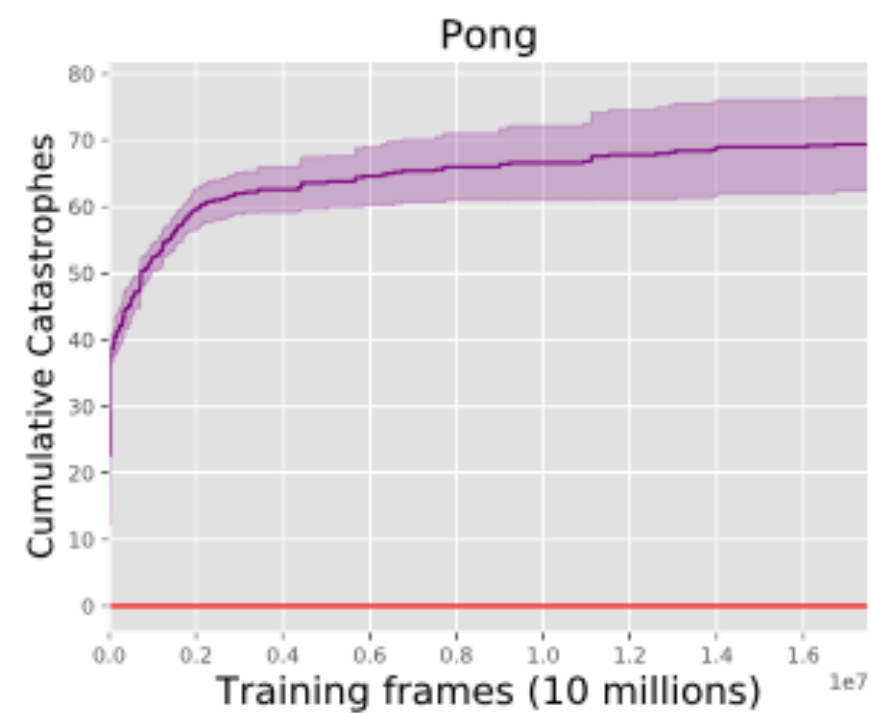
Does it learn better than HIRL (a strait-jacketed agent)?

# Results

1. HIRL agents learns with zero catastrophes (Pong, Space Invaders) or big reduction (Road Runner).
2. HIRL learns at least as well as Reward Shaping baseline, does much better overall.
3. Reward Shaping catastrophe rate does not converge to zero. (Catastrophic forgetting!)

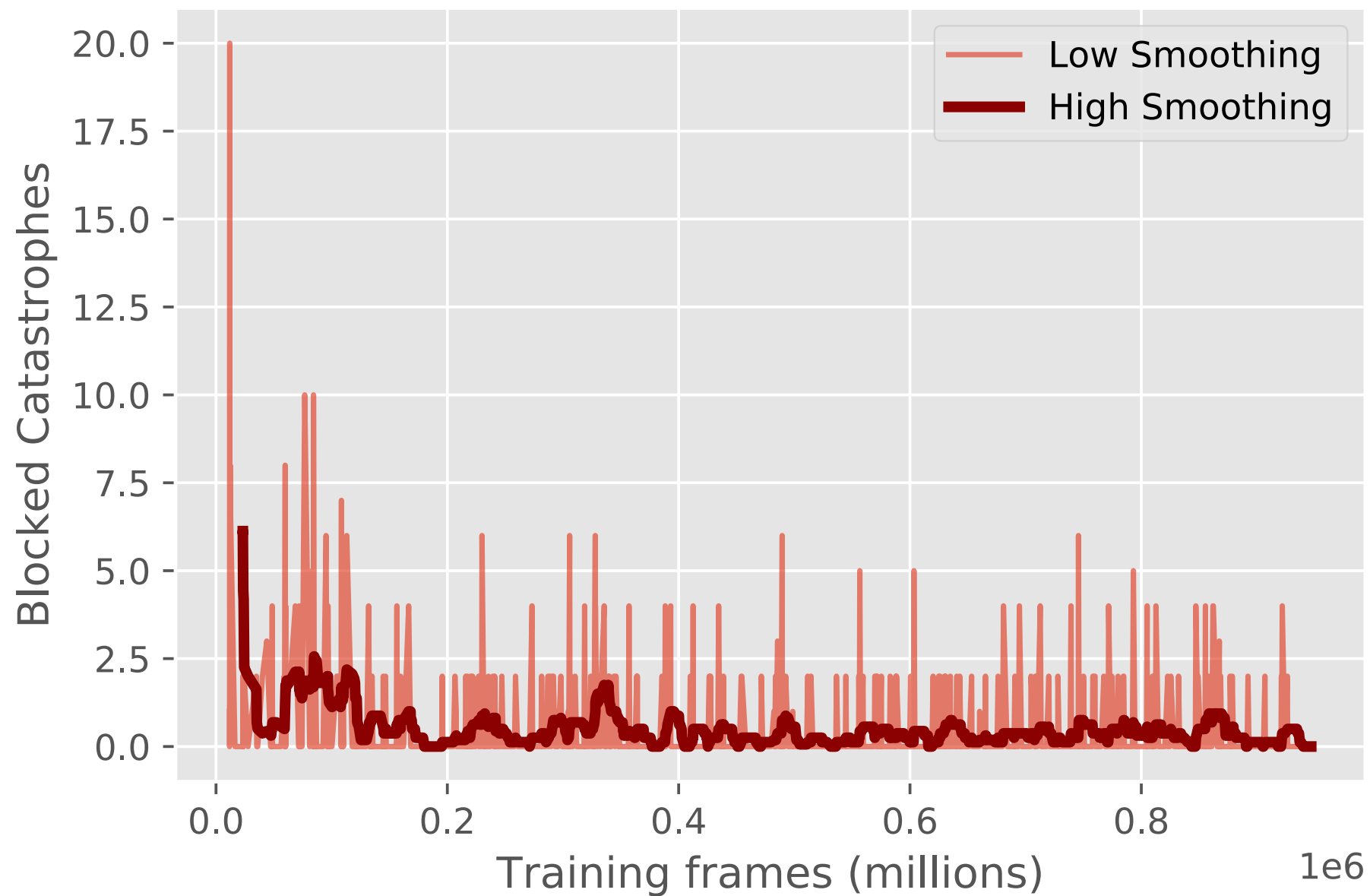
# Catastrophes for HIRL vs. No Oversight at all





# Pong Results

Pong: # Attempted Catastrophes in Initial 1M Frames



# Catastrophic forgetting

Table 1: Long-run rate of attempted catastrophes in Pong.

Policy	Learning Rate	Catastrophe Rate Per Episode (Std Err)
Stochastic	$10^{-4}$	0.012 (0.004)
Deterministic	$10^{-4}$	0.079 (0.017)
Stochastic	0	0.003 (0.001)
Deterministic	0	0 (0)







# Making RL+human efficient

- More data-efficient RL algorithms
- More data-efficient supervised learning algorithms
- RL agents who explore aggressively and systematically
- Human only provides oversight in unsafe regions of state space.
- Human provides more than binary labels (e.g. causes).

# Comparison to Christiano/Leike



- Online vs. Offline human oversight.
- Offline: works for fast tasks, human can label batches and view frames in reverse order (good for subtle causes), not full safety but limit catastrophes to finite number.
- Online: safety, prevent agent getting stuck, parallelize via A3C.

# Challenges in Blocking

- Many catastrophes are hard to block online (Atari and truck speeding on slippery road).
- Depends on “locality” (distance between point of no return).
- In some Atari games, avoiding all deaths and getting points would require playing very well from the start. So approach unlikely to work. (See Lipton failure).
- “Health and Safety”: human can create big safety margins by blocking well before action is actually dangerous. (Helps with human error, may make concept easier to learn).

# Other tools for avoiding catastrophes

- Simulation: red-teaming (adversarial training). Use offline HIRL to get accurate negative rewards and train a Blocker (which needs to transfer).

(Ciosek and Whiteson, Precup).

- Imitation learning of safe policy.
- Both would make possible much less human intervention (as in self-driving car case).

# Thanks!

- NIPS submission: coming soon to Arxiv.
- More papers:  
<http://owainevans.github.io>