Human Preferences and Human Control for Reinforcement Learners

Owain Evans
University of Oxford
FHI

Collaborators

Noah Goodman (Stanford, Uber AI)

Andreas Stuhlmueller (Stanford, ought.com)

John Salvatier (Oxford, AI Impacts)

Collaborators

David Abel (Brown)

David Krueger (MILA Montreal)

Jan Leike (DeepMind, Oxford)

Overview

GOAL: agents that (a) learn policies aligned with human preferences (b) via safe learning/exploration ("Safe RL").

Overview

GOAL: agents that (a) learn policies aligned with human preferences (b) via safe learning/exploration ("Safe RL").

Ways to specify optimal policy for RL agent:

1. Hand-code reward function before learning.

2. Learn rewards or policy from demonstration (IRL or imitation learning)

3. Human provides rewards online (TAMER, Active Reward Learning).

Overview

GOAL: agents that (a) learn policies aligned with human preferences (b) via safe learning/exploration ("Safe RL").

Ways to specify optimal policy for RL agent:

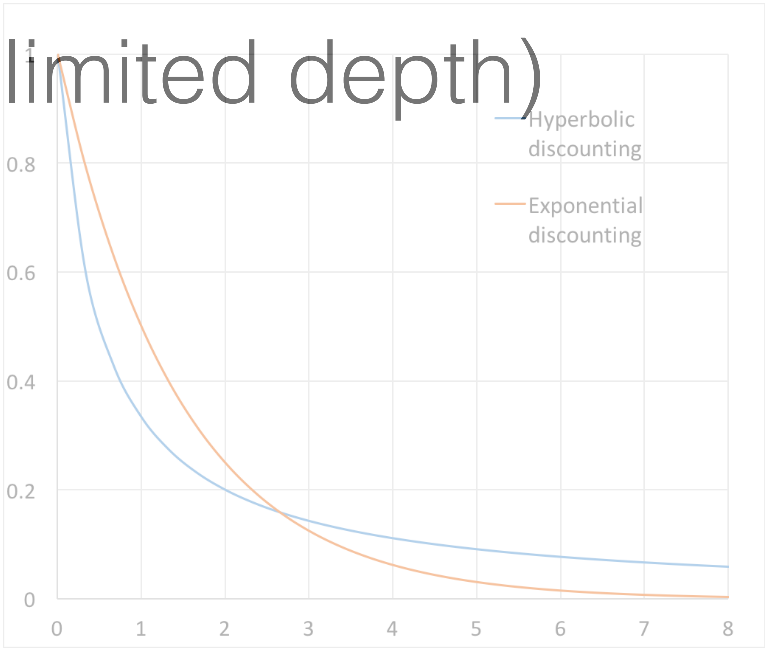
1. Hand-code reward function before learning.
2. Learn rewards or policy from demonstration (IRL or imitation learning)
3. Human provides rewards online (TAMER, Active Reward Learning).

IRL with Bounded, Biased Agents

IRL assumes human demonstrator is optimal up to **random** noise (softmax/Boltzmann)

Humans deviate **systematically** from optimal:

- Biases: hyperbolic discounting, prospect theory.
- Cognitive bounds: forgetting, myopic (limited depth) planning.



IRL with Bounded, Biased Agents

IRL assumes human demonstrator is optimal up to **random** noise (softmax/Boltzmann)

Humans deviate **systematically** from optimal

e.g. Person smokes every week but wishes to quit.

IRL with Bounded, Biased Agents

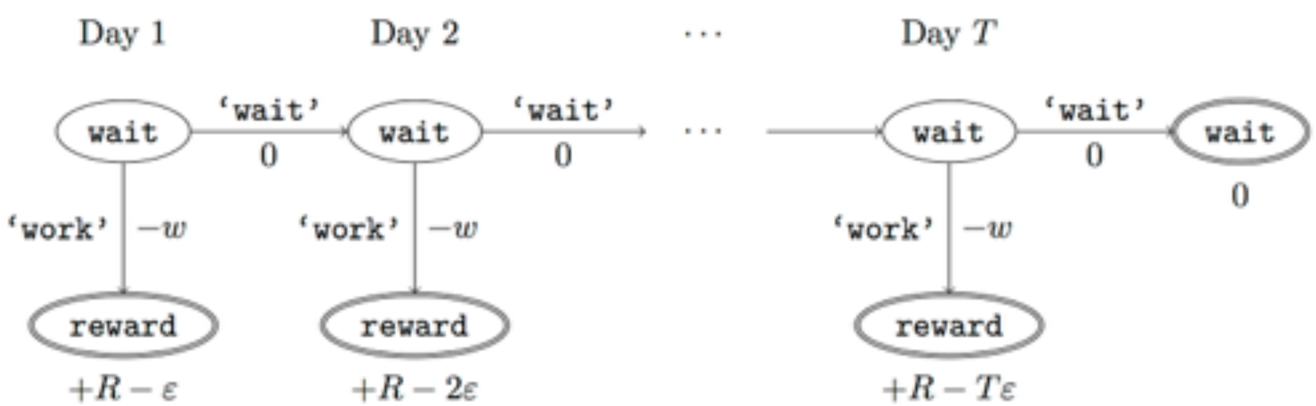
There are decision problems s.t.

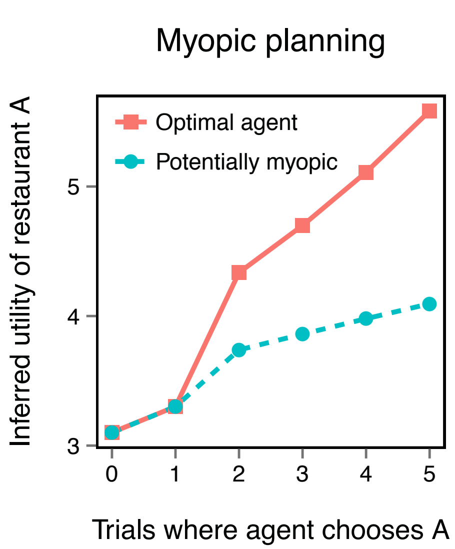
- IRL on biased agents can lead to arbitrarily mistaken inferences ….
- … but true preferences can be recovered (by modifying IRL)
- Problems are simple, uncontrived: Procrastination, Temptation, Bandits (explore/exploit).

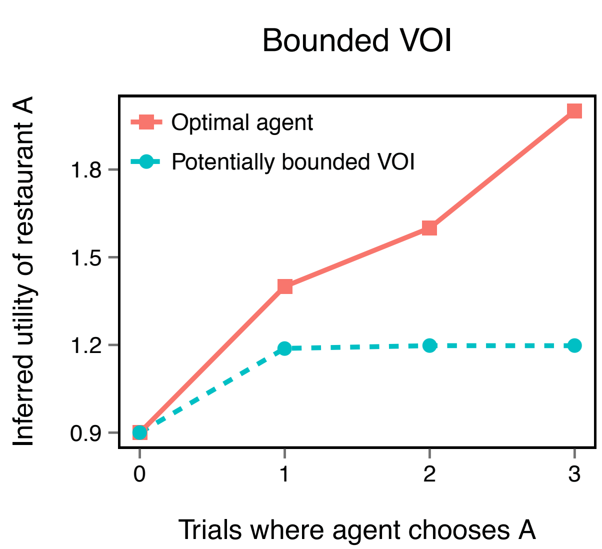
IRL with Bounded, Biased Agents

More info:

"Learning the Preferences of Ignorant, Inconsistent Agents" AAAI 2016.
"Learning the Preferences of Bounded Agents" NIPS workshop 2015.
http://www.agentmodels.org








```
var agent = function(state, delay, timeLeft){
  return Marginal(function(){
    var action = uniformDraw(actions)
    var eu = expUtility(state, action, delay, timeLeft)
    factor(alpha * eu)
    return action
  })
}

var expUtility = function(state, action, delay, timeLeft){
  var u = discountedUtility(state, action, delay, K)
  if (timeLeft == 1){
    return u
  } else {
    return u + expectation(INFER_EU(function(){
      var nextState = transition(state, action)
      var nextAction = sample(agent(nextState, delay+1, timeLeft-
      return expUtility(nextState, nextAction, delay+1, timeLeft-
    }))
  }
}
```

Overview

GOAL: agents that (a) learn policies aligned with human preferences (b) via safe learning/exploration ("Safe RL").

Ways to specify optimal policy for RL agent:

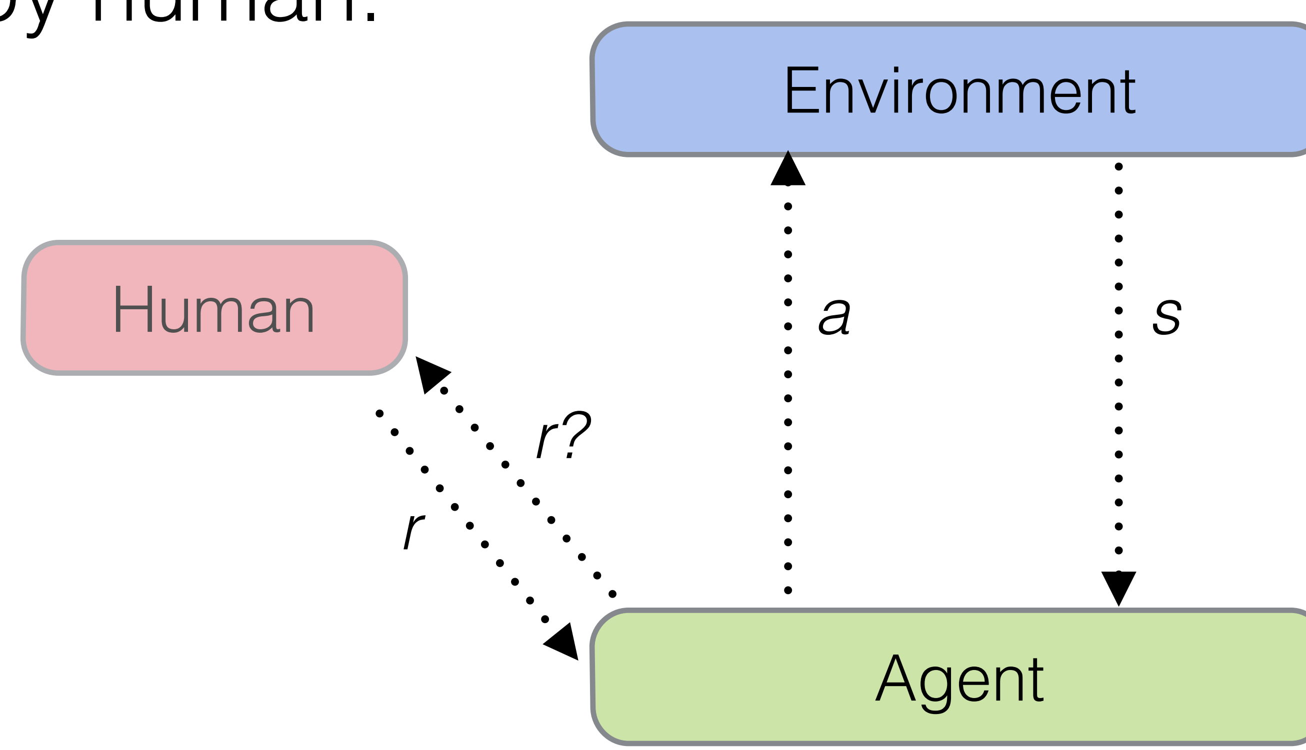
1. Hand-code reward function before learning.

2. Learn rewards or policy from demonstration (IRL or imitation learning)

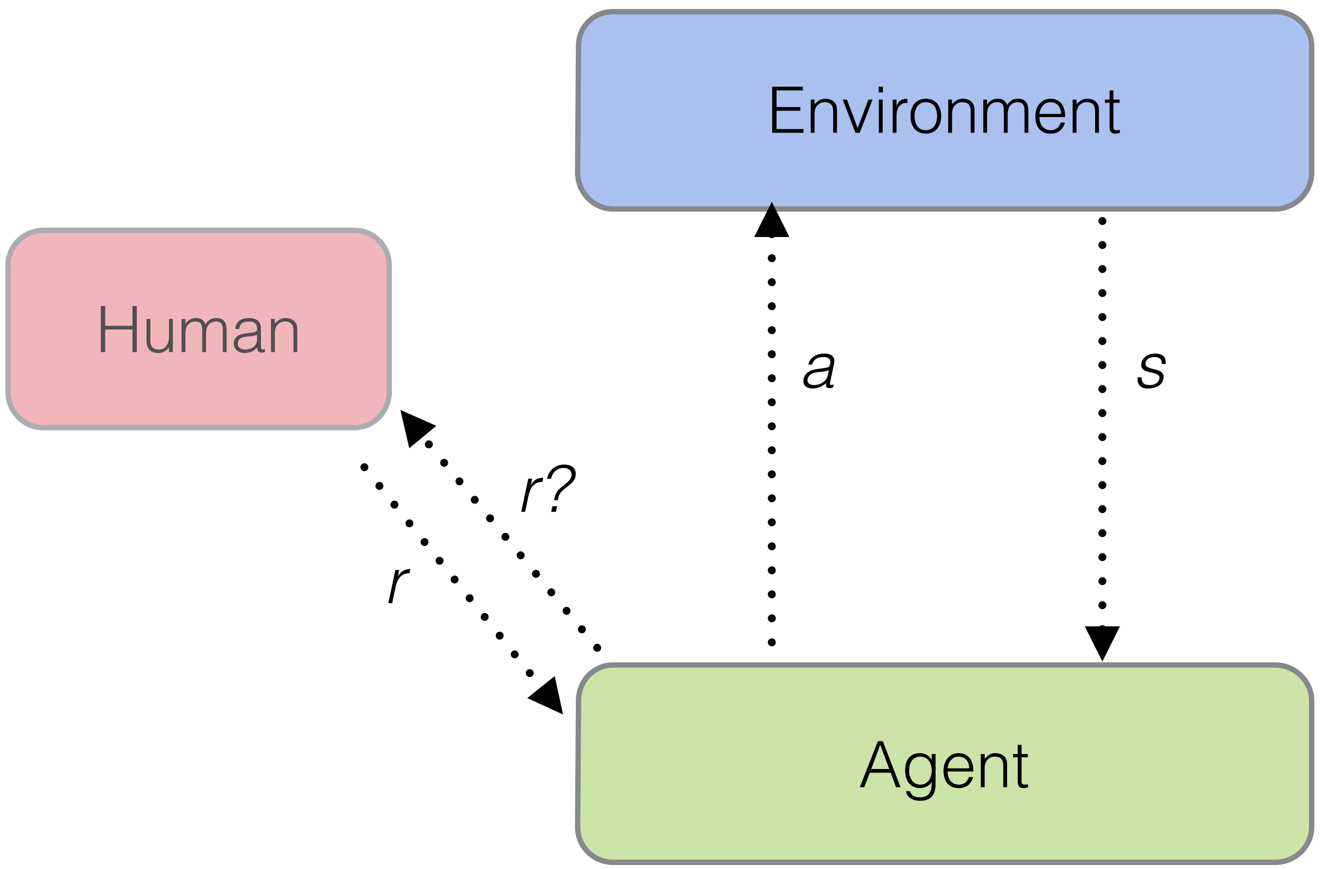
3. Human provides rewards online (TAMER, Active Reward Learning).

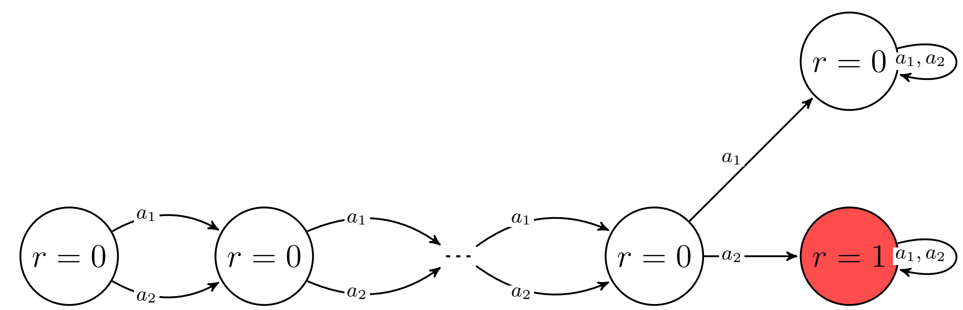
Active Reinforcement Learning

- Human provides rewards **online**
- **Label** the state-actions that actually occur
- Problem: how to reduce burden on human?
- *Active Reinforcement Learning*: agent selects which state-actions are labeled by human.



Active Reinforcement Learning





- Agent chooses whether to observe reward R_t on time-step t
- Observing R_t has cost c
- Goal: maximize $\sum_t R_t - c\, q_t, \qquad q_t = \begin{cases} 1 & \text{if } R_t \text{is observed} \\ 0 & \text{else} \end{cases}$

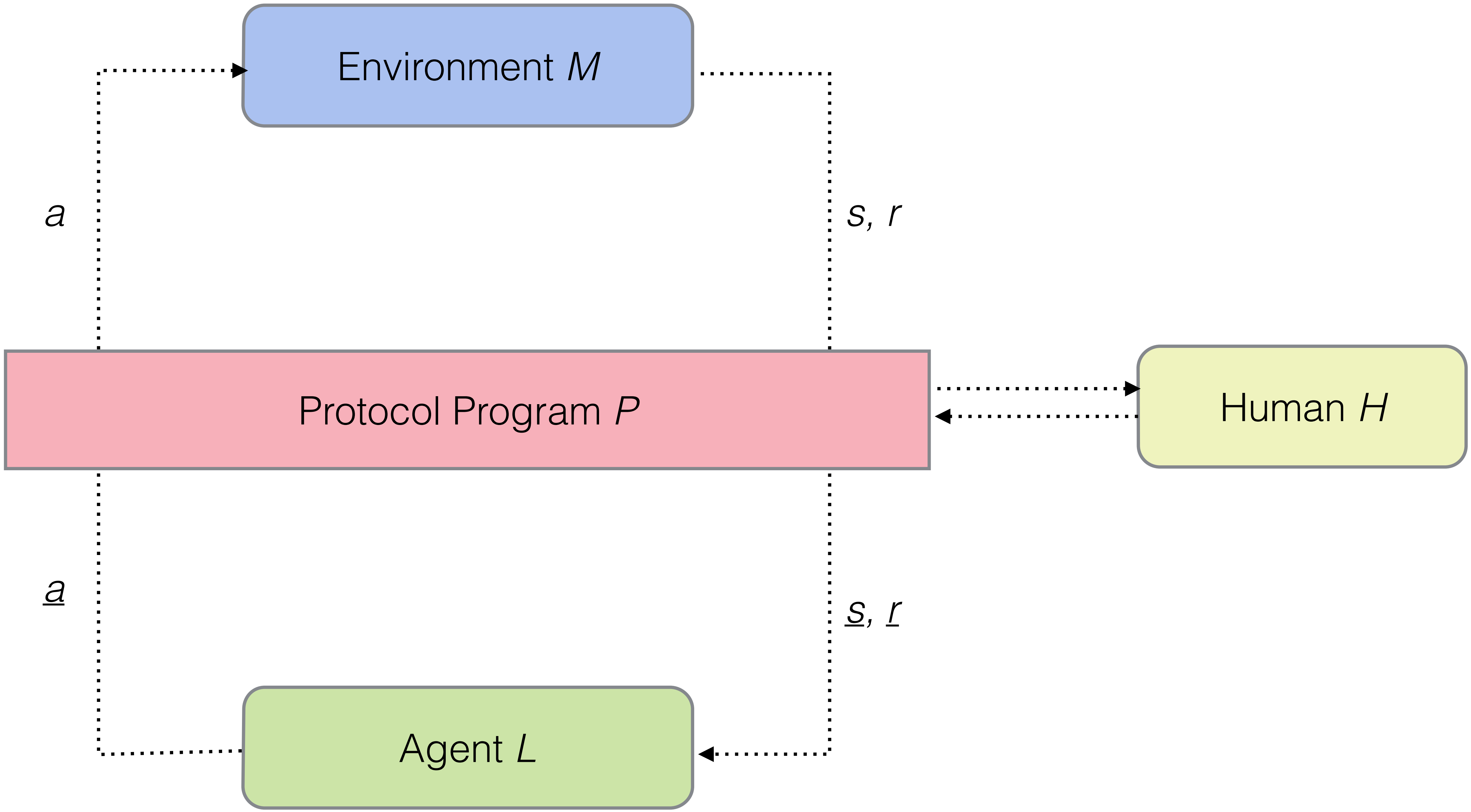
Overview

GOAL: agents that (a) learn policies aligned with human preferences (b) via safe learning/exploration (“Safe RL”).

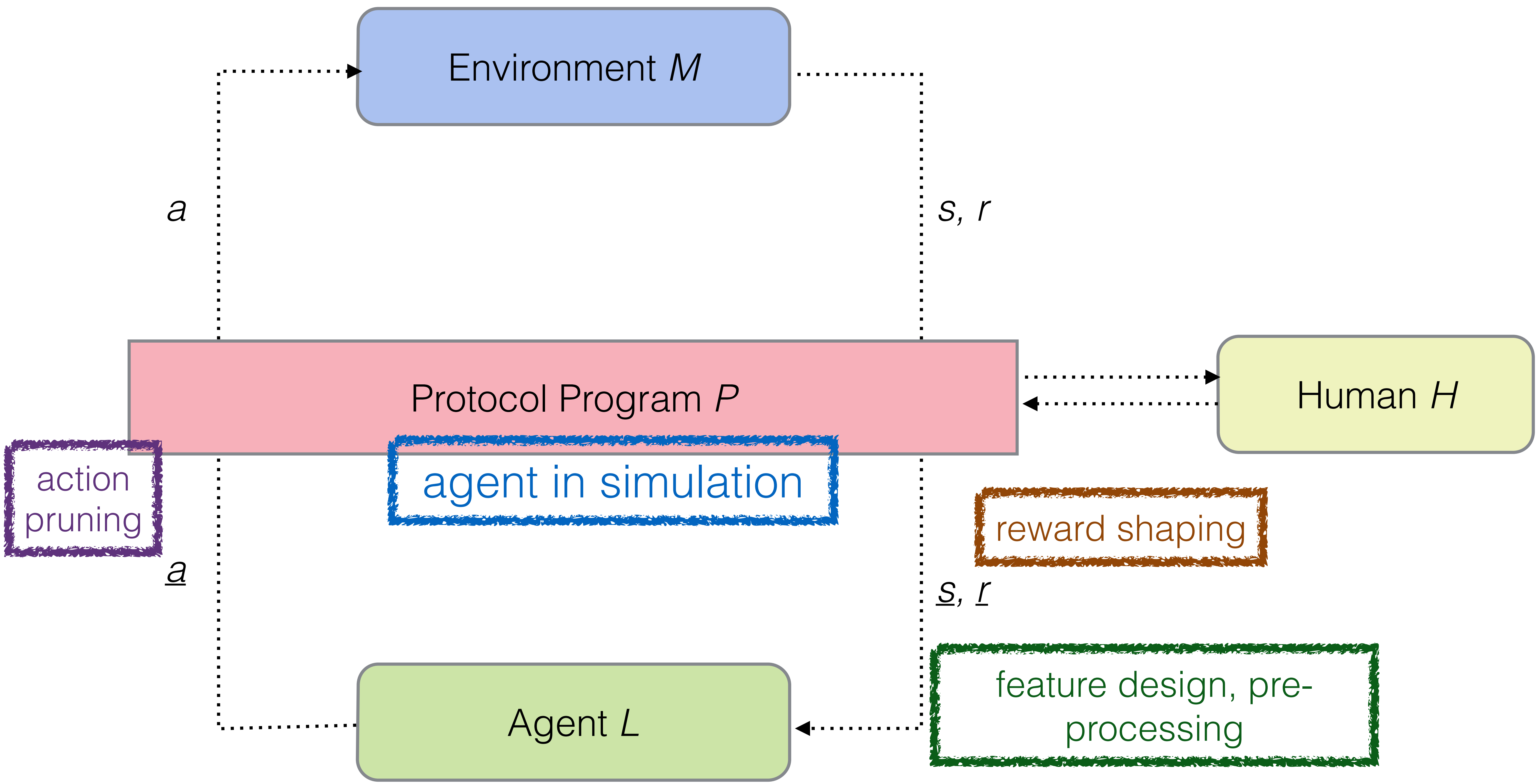
Ways to specify optimal policy for RL agent:

1. Hand-code reward function before learning.
2. Learn rewards or policy from demonstration (IRL or imitation learning)
3. Human provides rewards online (TAMER, Active Reward Learning).

Human-in-the-loop RL



Human teaching for **any** RL agent



Prevent Catastrophes with Interactive RL

Catastrophic action: action that RL agent should essentially never take, i.e. P(action) $< \epsilon$

Examples:

- breaking laws / moral rules
- physically harm humans
- manipulate or psychologically harm humans

Prevent Catastrophes with Interactive RL

Catastrophic action: action that RL agent should never take, i.e. P(action) $< \epsilon$

Examples:

- breaking laws / moral rules
- physically harm humans
- manipulate or psychologically harm humans

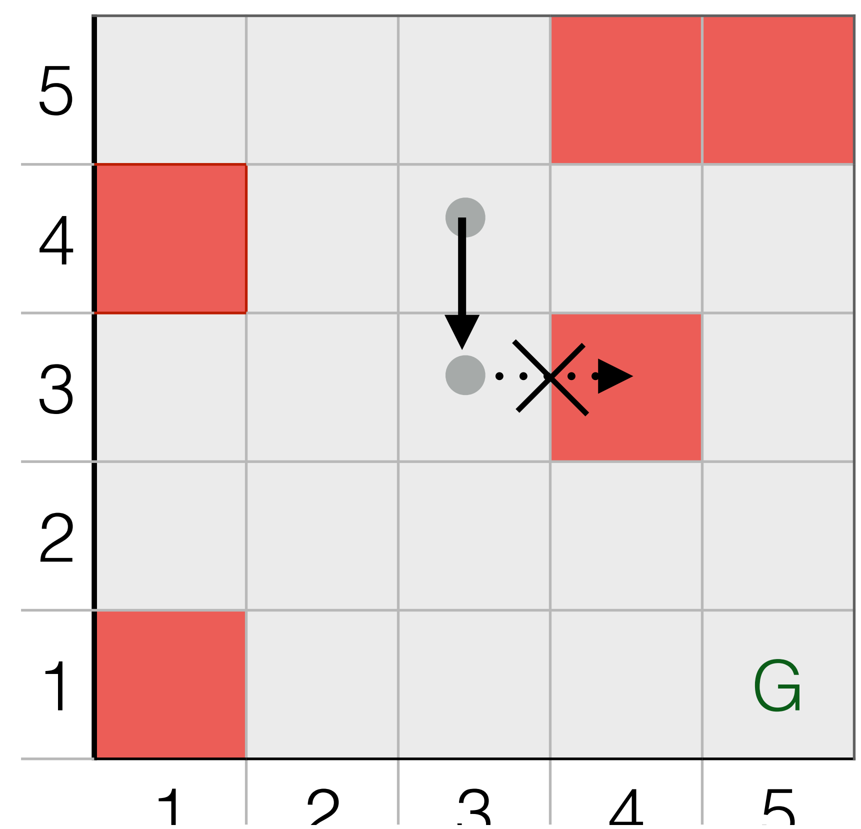
Prevent Catastrophes with human in loop

Related work: Safe RL and avoiding SREs (Moldovan and Abeel, Frank et al., Paul et. al, Lipton et al.)

Challenge:

- Simulation often inadequate (esp. for extreme events)
- RL agents learn by **trial** and **error** (don't know R and T in advance)
- Solution: human blocks catastrophes **before** they happen

Prevent Catastrophes with human in loop



1. Human blocks agent trying bad action,

   gives big negative reward.

2. Classifier learns to recognize bad actions
3. Classifier takes over human role.
4. (Human interactively defines a new MDP).

Problems: efficiency, robust generalization.

Human Preferences and Human Control for Reinforcement Learners

Owain Evans
University of Oxford
FHI

GOAL: agents that (a) learn policies aligned with human preferences (b) via safe learning/ exploration ("Safe RL").

Ways to specify optimal policy for RL agent:

1. Hand-code reward function before learning.

2. Learn rewards or policy from demonstration (IRL or imitation learning)

3. Human provides rewards online (TAMER, Active Reward Learning).

THANKS!