Cooperative Inverse Reinforcement Learning

Presenter: Davin Lawrence

October 13, 2022

"If we use, to achieve our purposes, a mechanical agency with whose operations we cannot interfere effectively…we had better be quite sure that the purpose put into the machine is the purpose which we really desire."

– Norbert Wiener (1960)

"Ultimately, we are in the business of building AI systems that integrate well with humans and human society. And if we don't take that as a fundamental tenet of the field, I think we are potentially in trouble and that is a perspective I wish was more pervasive throughout artificial intelligence, generally."

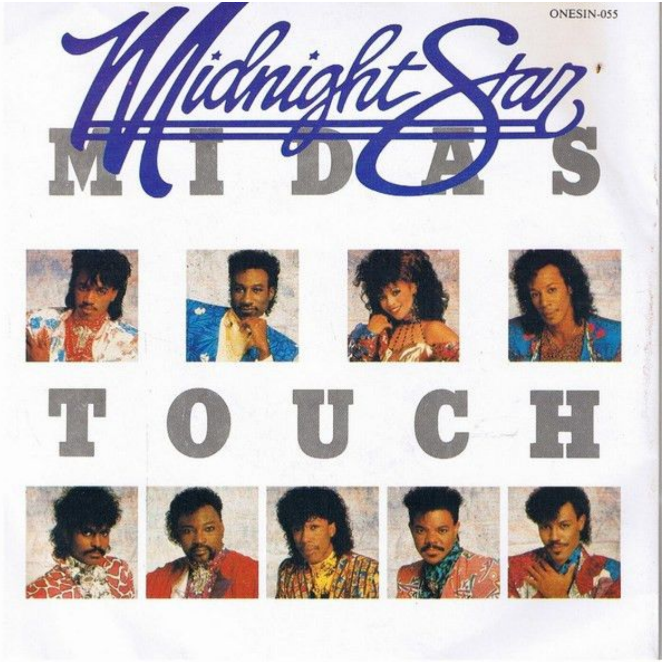
– Dylan Hadfield-Menell (2019)

Reward Engineering is Difficult

❖ Humans are exceptionally good at mis-stating their goals

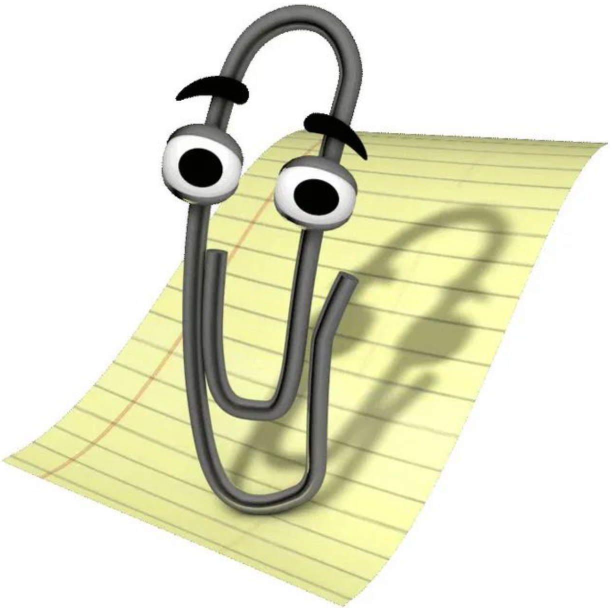
Reward Engineering is Difficult

- ❖ Humans are exceptionally good at mis-stating their goals
- ❖ Humans have a hard time understanding the full implications of their goals



Reward Engineering is Difficult

- Humans are exceptionally good at mis-stating their goals
- Humans have a hard time understanding the full implications of their goals
- Given success in the previous two, there is no guarantee a robot optimizing on a reward shares the same value as a human being



Value Alignment Problem

- These examples highlight the value alignment problem
- Each example focuses on specific rewards while missing the true value
  - Maximizing points rather than winning the race
  - Maximizing personal wealth rather than happiness
  - Maximizing amount of paper clips rather than the overall welfare of everyone
- Reinforcement Learning problems are not made in vacuum.
  - They are a part of a human-robot relationship
  - Simply encoding rewards leads to these errors.
  - Instead, we should train robots to learn the underlying value and desire of their human counterparts

Problem Setting

- Cooperative Inverse Reinforcement Learning is an attempt to formalize the value alignment problem within AI.
- CIRL formalizes this problem as a two player game between a Human (**H**) and a Robot (**R**).
- This game is a partial information game in which one player, the human, knows the reward function, while the robot, does not know the reward.
- The robots payoff is the human's actual reward.
- The optimal solution to this problem maximizes the human's reward
- This solution involves teaching by the human and learning by the robot.

Inverse Reinforcement Learning

- Attempt to determine the reward function being optimized by observing an actor's behavior in the environment.
- Key assumption is that the observed actor is behaving optimally
  - Hadfield-Mennell dubs this 'Demonstration by Expert' or DBE
- Key difference is in CIRL, the optimal solution includes teaching behaviors

Hidden Goal MDP

- The goal is a hidden part of the state. θ encodes a particular goal state.
- Here $\mathbf{R}$ helps $\mathbf{H}$, but $\mathbf{H}$ is treated as part of the environment rather than a secondary agent.

Optimal Teaching

- The objective is to optimize efficient learning in an agent
- Optimal teaching emerges as a property of CIRL, rather than being the goal

Principal-Agent Models

- Economic framework where a principal specifies incentives to an agent to maximize the principal’s profit.

CIRL Formulation

Definition 1. *A* cooperative inverse reinforcement learning *(CIRL) game M is a two-player Markov game with identical payoffs between a human or principal,* **H***, and a robot or agent,* **R***. The game is described by a tuple, $M = \langle \mathcal{S}, \{\mathcal{A}^{\mathbf{H}}, \mathcal{A}^{\mathbf{R}}\}, T(\cdot|\cdot,\cdot,\cdot), \{\Theta, R(\cdot,\cdot,\cdot;\cdot)\}, P_0(\cdot,\cdot), \gamma\rangle$, with the following definitions:*

$\mathcal{S}$ *a set of world states: $s \in \mathcal{S}$.*
$\mathcal{A}^{\mathbf{H}}$ *a set of actions for* **H***: $a^{\mathbf{H}} \in \mathcal{A}^{\mathbf{H}}$.*
$\mathcal{A}^{\mathbf{R}}$ *a set of actions for* **R***: $a^{\mathbf{R}} \in \mathcal{A}^{\mathbf{R}}$.*
$T(\cdot|\cdot,\cdot,\cdot)$ *a conditional distribution on the next world state, given previous state and action for both agents: $T(s'|s, a^{\mathbf{H}}, a^{\mathbf{R}})$.*
Θ *a set of possible static reward parameters, only observed by* **H***: $\theta \in \Theta$.*
$R(\cdot,\cdot,\cdot;\cdot)$ *a parameterized reward function that maps world states, joint actions, and reward parameters to real numbers. $R : \mathcal{S} \times \mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \Theta \to \mathbb{R}$.*
$P_0(\cdot,\cdot)$ *a distribution over the initial state, represented as tuples: $P_0(s_0, \theta)$*
γ *a discount factor: $\gamma \in [0, 1]$.*

The CIRL Game

- The game begins by sampling the initial state
  - Note $\mathbf{H}$ observes θ, while $\mathbf{R}$ does not.
- At each time step, $\mathbf{H}$ and $\mathbf{R}$ choose their actions
- Both actors receive an award

$$r_t = R(s_t, a_t^{\mathbf{H}}, a_t^{\mathbf{R}}; \theta)$$

- Behavior is defined as a policy pair: $(\pi^{\mathbf{H}}, \pi^{\mathbf{R}})$

$$\pi^{\mathbf{H}} : \left[\mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \mathcal{S}\right]^* \times \Theta \rightarrow \mathcal{A}^{\mathbf{H}} \qquad \pi^{\mathbf{R}} : \left[\mathcal{A}^{\mathbf{H}} \times \mathcal{A}^{\mathbf{R}} \times \mathcal{S}\right]^* \rightarrow \mathcal{A}^{\mathbf{R}}$$

- The optimal joint policy is one that maximizes value, which is the expected sum of discounted rewards

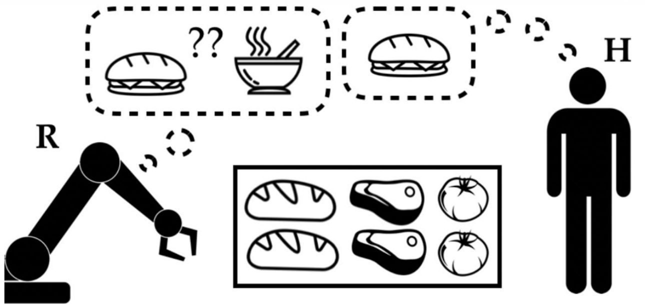


Fig Source: Malik, Palaniappan, Fisac, Hadfield-Mennell, Russell & Dragan, 2018

Computing Optimal Policy Pairs

- The optimal policy pair is representational of $\mathbf{H}$ and $\mathbf{R}$ coordinating perfectly.
- This is an example of a decentralized-partially observed MDP (Dec-POMDP)
  - Dec-POMDPs are NEXP-complete, which is generally regarded as a bad thing
- CIRLs can reduce this complexity
  - The structure of CIRL implies $\mathbf{H}$'s initial observation of θ is private information
  - This allows a reduction from Dec-POMDP to a coordination-POMDP
- Theorem: Let M = CIRL game with state $\mathscr{S}$ and reward space $\boldsymbol{\Theta}$. There exists a POMDP M_c with state space $\mathscr{S}_c$ such that $|\mathscr{S}_c| = |\mathscr{S}| \cdot |\boldsymbol{\Theta}|$ and for any policy pair in M, there is a policy in M_c that achieves the same sum of discounted awards
  - Therefore, there exists an optimal policy pair that only depends on the current state and $\mathbf{R}$'s belief.

Apprenticeship Learning

- Apprenticeship CIRL is a subclass of CIRL which adds the concept of turns and phases to the general CIRL problem.
  - Learning phase - $\mathbf{H}$ demonstrates the task to teach $\mathbf{R}$
  - Deployment phase - $\mathbf{R}$ becomes the only actor, working on it's belief of θ.
- In the deployment phase, the optimal policy for $\mathbf{R}$ to maximizes the reward in the MDP induced by the mean θ from $\mathbf{R}$'s belief.
- This formulation is used to reason about DBE
  - There exists ACIRL games where the best response for $\mathbf{H}$ to $\pi^{\mathbf{R}}$ violates the expert demonstrator assumption.
  - If **br**(π) is the best response to π, then **br**(**br**($\pi^{\mathbf{E}}$)) ≠ $\pi^{\mathbf{E}}$
- We should expect users to present optimizations for fast learning rather than demonstrations that maximize reward

Generating Instructive Demonstrations

- The expert demonstration assumption is broken, so how should $\mathbf{H}$ act?
- IRL combined with the mean θ from $\mathbf{R}$'s belief, the optimal $\pi^{\mathbf{R}}$ computes a policy that matches the observed feature counts from the learning phase.
  - Note this is under the DBE assumption
- This implies we can compute a demonstration trajectory $\tau^{\mathbf{H}}$.
- We begin by calculating feature counts $\mathbf{R}$ would observe in expectation of θ.
- If ϕ_θ is the expected feature counts, then

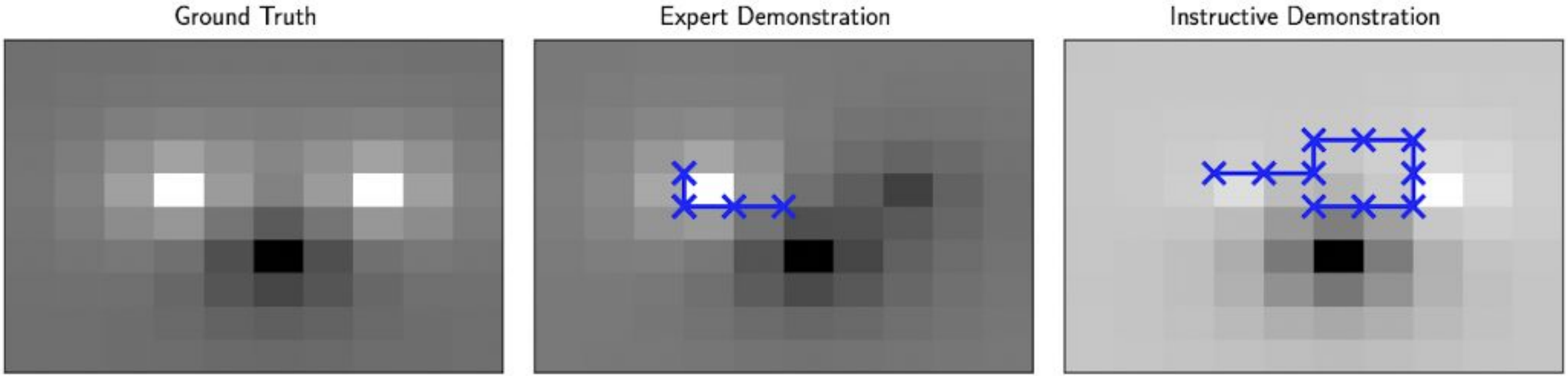
$$\tau^{\mathbf{H}} \leftarrow \underset{\tau}{\operatorname{argmax}}\, \phi(\tau)^\top \theta - \eta||\phi_\theta - \phi(\tau)||^2$$

- This difference is termed as regret

Experiment One Setup

- Experimental setup was simple for this task due to the complexity of calculation
- Simple 2D navigation on a small, discrete grid.
- **H** performs a trajectory while **R** observes in the learning phase
- **R** placed randomly on the grid and given control
- The set of actions consists of only the cardinal directions and nop.

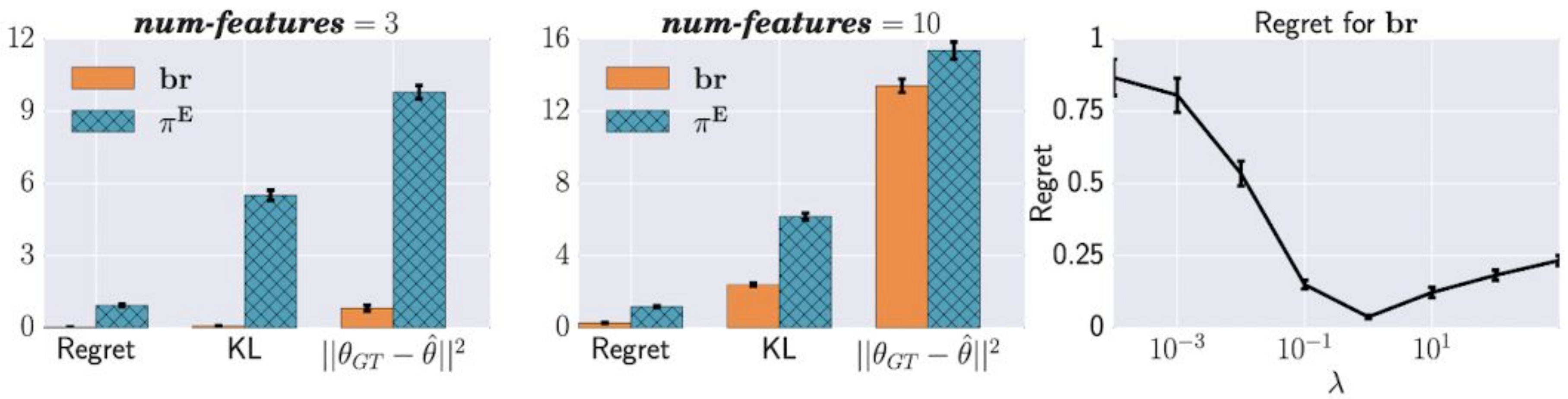
Experiment One Results



Experiment Two/Three Setup

- Both experiments use Maximum Entropy IRL to implement $\mathbf{R}$'s policy.
- Experiment Two: Compare DBE vs Approximate Best Response
  - Human agent can choose either best response or DBE
  - Robot uses IRL to compute its estimate of theta during deployment
  - Run with number of features = 3 and 10
- Experiment Three: Applying CIRL to Maximum Entropy IRL
  - Exploits the free parameter λ which controls how optimal $\mathbf{R}$ believes $\mathbf{H}$ is acting.
  - This experiment the effects of modifying $\mathbf{R}$'s belief on $\mathbf{H}$'s action

Experimental Results



Critique / Limitations / Open Issues

- The main problem with CIRL is the complexity of the space
  - This limits to simple experiments
- The complexity of the reduced coordination POMP is ambiguous
- The first experiment does not really make clear what is happening

Future Work For Paper

- Formalize complexity space for CIRL as a coordinated POMP
  - The ambiguity does not elicit faith in the paper's findings
- Show CIRL or ACIRL can be used in realistic, complex domains rather than simple toy examples
  - The framework and ideology behind this paper is important, but theory without practice is dead
  - Malik, et al followed up with a modified Bellman update in service of CIRL, which shows promise
- More concrete, clear experimentation

Extended Readings

- Supplementary and Review Material
  - https://papers.nips.cc/
- Algorithms for Inverse Reinforcement Learning (Ng and Russell, 2000)
  - https://ai.stanford.edu/~ang/papers/icml00-irl.pdf
- An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement Learning (Malik, Palaniappan, Fisac, et al, 2018)
  - https://arxiv.org/abs/1806.03820
- Apprenticeship learning via inverse reinforcement learning (Abbeel and Ng, 2004)
  - https://ai.stanford.edu/~ang/papers/icml04-apprentice.pdf
- ELI5: Cooperatively Learning Human Values
  - https://bair.berkeley.edu/blog/2017/08/17/cooperatively-learning-human-values/
- Podcast Interview with Dylan Hadfield-Menell
  - https://futureoflife.org/2019/01/17/cooperative-inverse-reinforcement-learning-with-dylan-hadfield-menell/

Summary

- Reward Engineering is difficult for a number of reasons
- This is difficult because humans have a hard time communicating what they want
- AI, in general, focuses on specific rewards without consideration of true goals
- Key takeaways
  - CIRL provides a formalization of the value-alignment problem
  - DBE is not the optimal policy for training a robot in IRL
  - The regret metric gives us a way to compute the optimal human trajectory
- CIRL gives future research a framework for analyzing and working with the value-alignment problem.