Impact of AI Decal:
Human-Compatible AI

Gokul Swamy & Brenton Chu

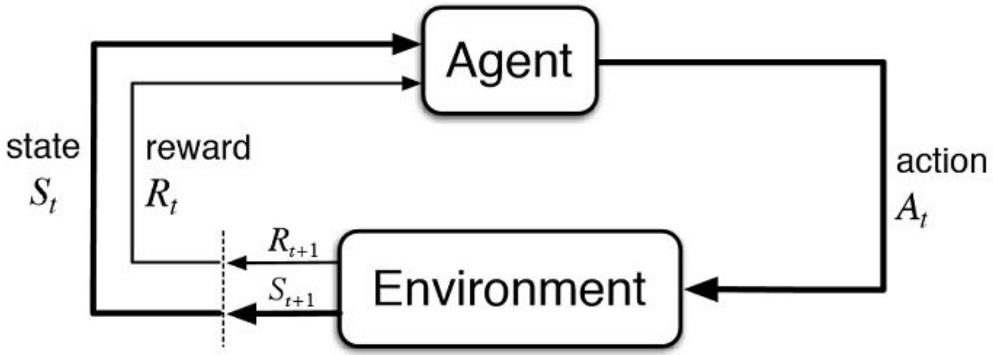
Quiz:
https://tinyurl.com/impactsp19q10

Announcements

- ➔ Guest Lectures will begin next week
  - ◆ We won’t have readings for the last few weeks and quizzes will not cover any guest lecture material
  - ◆ You’re still required to attend!
  - ◆ In fact, we want you to attend more than usual!

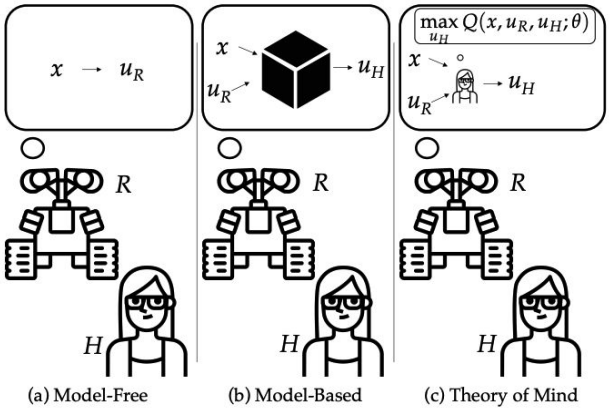
Revisit: Reinforcement Learning

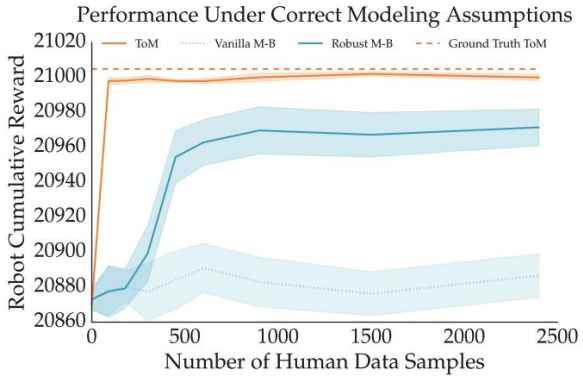
- ➔ RL assumes you control an agent acting in some sort of environment
- ➔ When the agent takes an action, the state of the world moves forward and the agent receives some sort of reward based on how optimal the action it took was

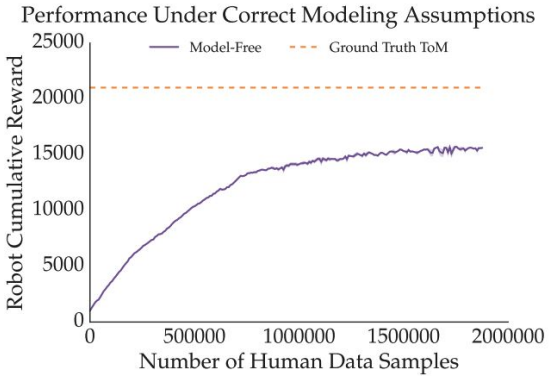


HCAI

- ➔ Remember that our goal here is “**value alignment**”
- ➔ One general purpose technique to get this to work is to model people
- ➔ This has some cool advantages:







5 Problems in AI Safety

- ➔ These are long-term - we talked about the short term ones earlier (bias, adversarial examples, …)
- ➔ P1: Avoiding Negative Side Effects
- ➔ P2: Avoiding Reward Hacking
- ➔ P3: Scalable Oversight
- ➔ P4: Safe Exploration
- ➔ P5: Robustness to Distributional Shift

P1&2: What can go wrong with mis-specified objectives? (1)

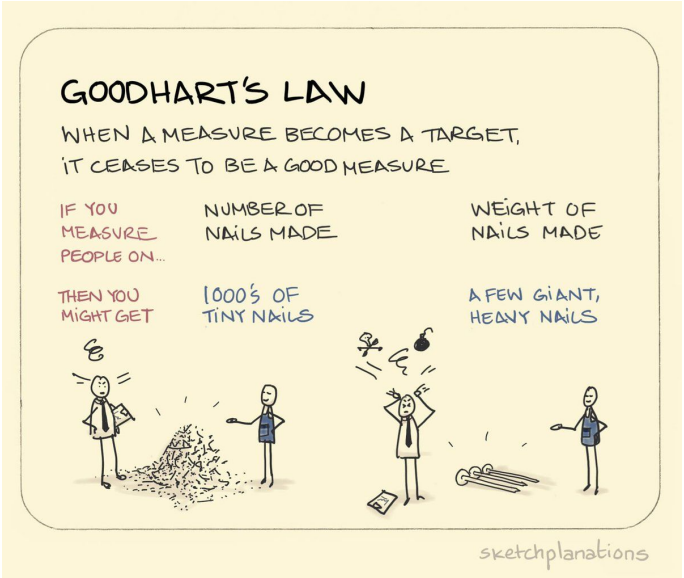
- ➔ Briefly mentioned in the previous lecture
- ➔ Folklore: King Midas, Sorcerer's Apprentice
- ➔ Stuart Russell: Cleaning robot that is told to maximize the amount of dust it picks up
  - ◆ How could this go wrong?

P1&2: What can go wrong with mis-specified objectives? (2)

P1&2: What can go wrong with mis-specified objectives? (2)

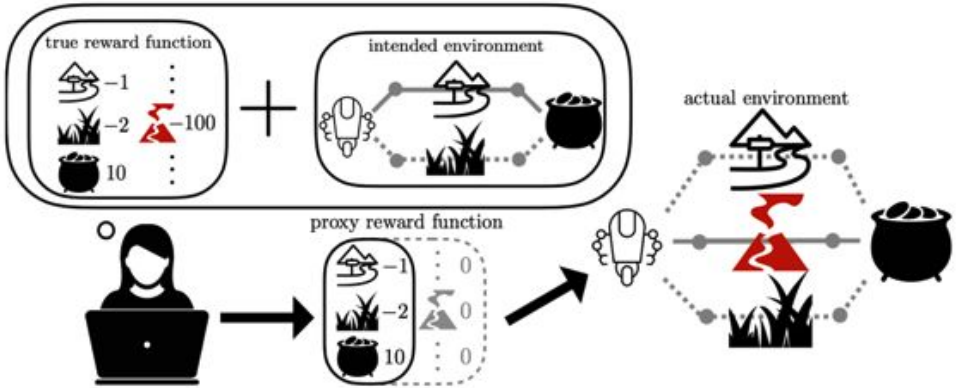
➔ What was the problem here?

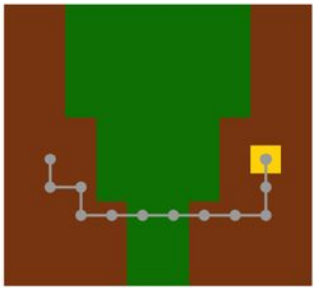
- ◆ Goodhart's Law: "*When a measure becomes a target, it ceases to be a good measure.*"

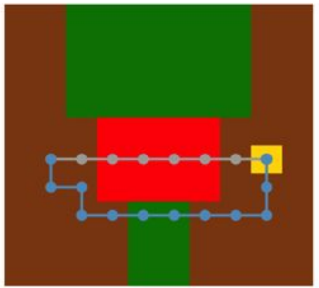


Inverse Reward Design

➔ Treat given reward function as observation about true reward function in designer's head.

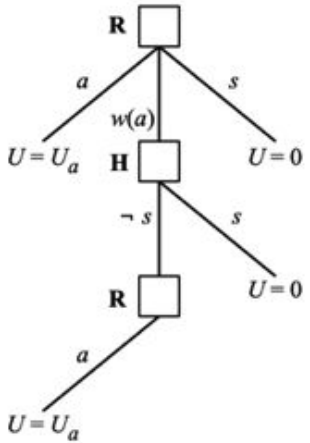






The Off-Switch Game

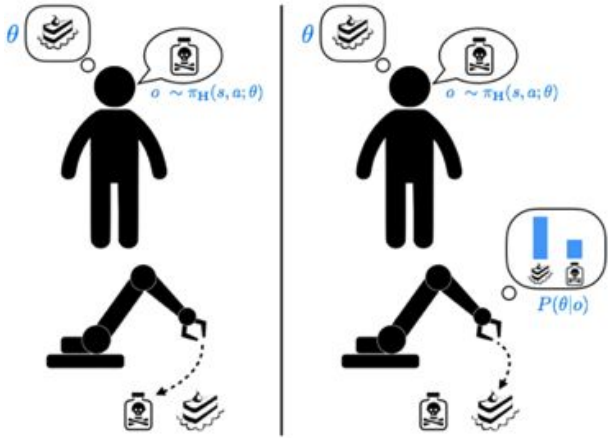
- ➔ Some human error will always slip through - how can we deal with this?
  - ◆ Corrigibility: we can correct misbehaving systems.
- ➔ For systems to be corrigible, they need to have some uncertainty about their utility functions



$$\pi^{\mathbf{H}}(U_a) = \begin{cases} 1 & U_a \geq 0 \\ 0 & o.w. \end{cases}$$

Should robots be obedient?

- ➔ Should robots listen to all people?
  - ◆ What about a child telling a self-driving car to stop?
- ➔ Why don't we listen at first to determine preferences and then act accordingly?



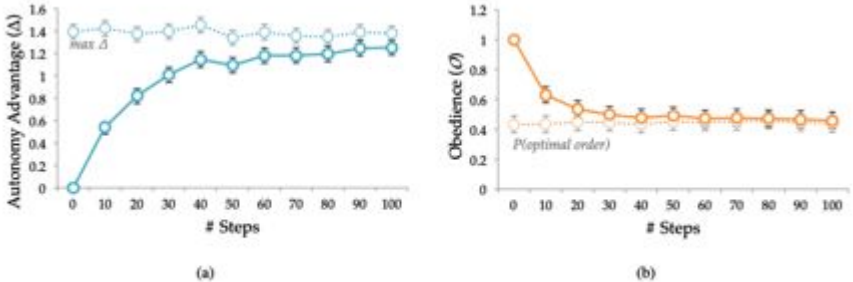
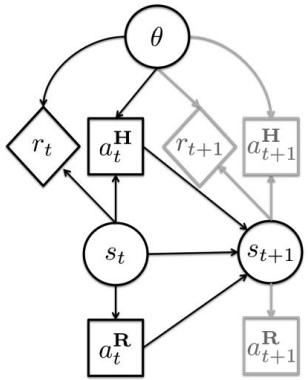


Figure 2: Autonomy advantage Δ (left) and obedience $\mathcal{O}$ (right) over time.

Cooperative Inverse Reinforcement Learning

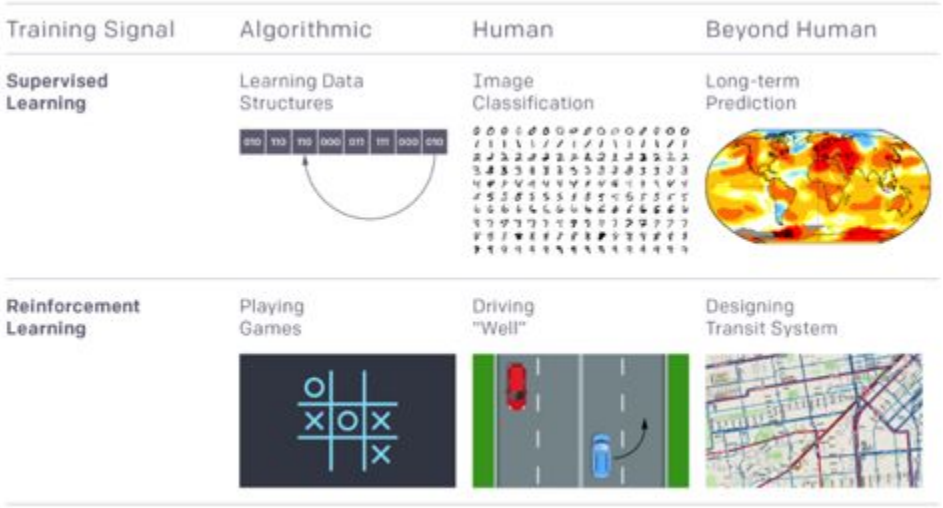
- ➔ Consider a 2-player game with a human and a robot
- ➔ It is cooperative, so they both receive the same reward
- ➔ However, only the person has access to the reward parameters
  - ◆ So, by observing the person, the robot can learn the reward parameters
- ➔ “Optimal CIRL solutions produce behaviors such as **active teaching, active learning**, and communicative actions that are more effective in achieving value alignment” - Dylan



P3: How do we make oversight scalable?

➔ Having people judge tasks constantly is annoying

➔ But what about tasks we can't judge?



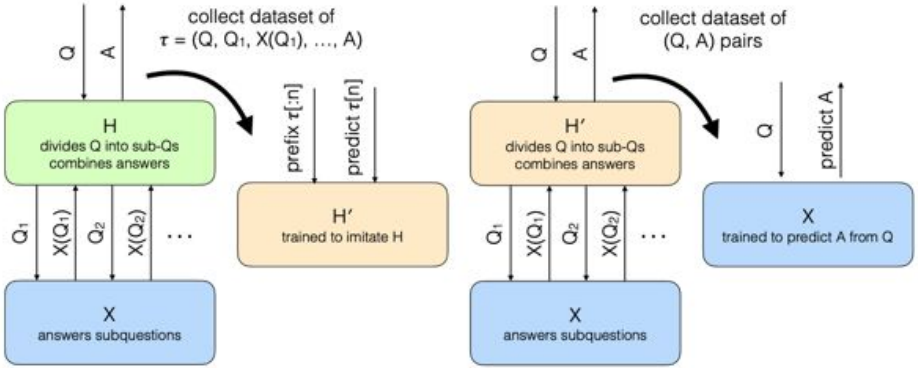
AI Safety via Debate

➔ By having AI systems explain what they are doing, we can both perform better and better judge them (more explainable AI)

- Think of a regular debate - having to justify your arguments makes them stronger

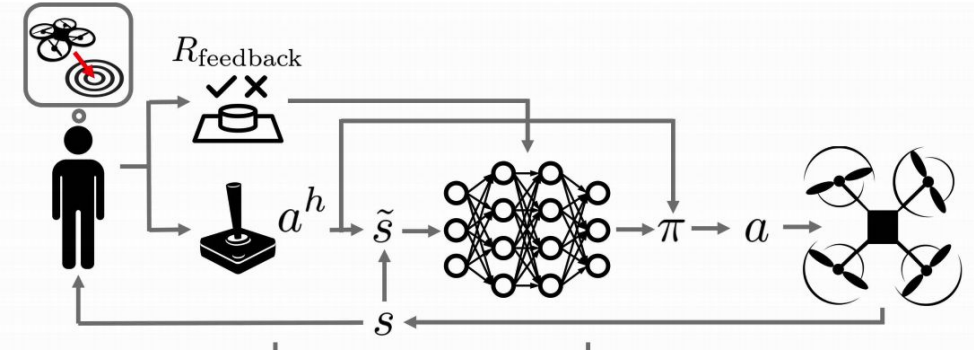
Supervising strong learners by amplifying weak experts

- ➔ We can decompose difficult questions into things people can answer and then reconstruct the answer to the more complex query
  - ◆ And then learn to imitate the people
- ➔ Goal is to decompose complex tasks



Shared Autonomy via Deep Reinforcement Learning

- ➔ Shared autonomy: person and robot control a single system together
- ➔ “From the agent’s perspective, the user acts like a **prior policy** that can be fine-tuned, and an additional sensor generating observations from which the agent can implicitly decode the user’s private information. From the user’s perspective, the agent behaves like an **adaptive interface** that learns a **personalized mapping** from user commands to actions that maximizes task reward.” - Sid



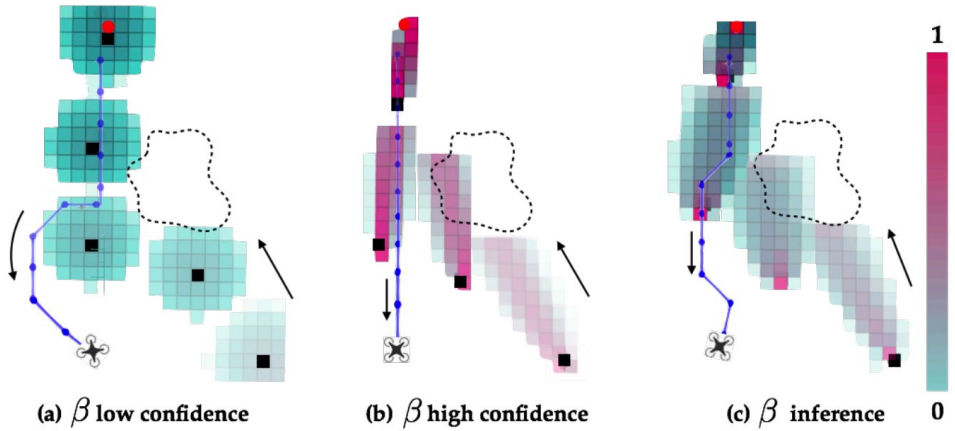
P4/5: How can we explore safely?

- ➔ P4: General problem somewhat dealt with by Inverse Reward Design + risk-averse planning
  - ◆ However, what if there are other agents in the environment? How do we not interfere with them then?
- ➔ P5: How do we not learn bad things when we see bad data?

Probabilistically Safe Robot Planning with Confidence-Based Human Predictions

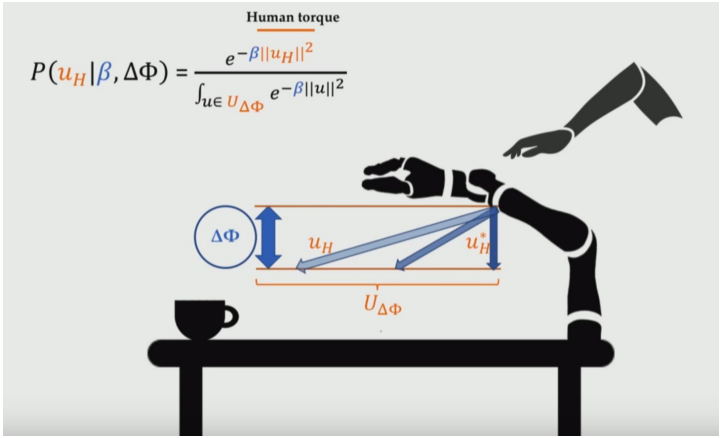
➔ We can plan around humans by continually estimating how rational they are.

➔ Anca's Explanation: https://youtu.be/_VceNn8ZWAg?t=18694



Learning under Misspecified Objective Spaces

- ➔ If a person looks very irrational under our model, we should realize that we probably aren't considering something they are
  - ◆ Outlier detection
- ➔ Andreea's Explanation: https://youtu.be/FSsEqEJKo8A?t=6353 (if time)



Q:

Do you think the approaches previously discussed are sufficient to create safe AI?

Impact of AI Decal: *Activity*

Activity: Q&A

➔ AMA!

Impact of AI Decal:

Next: Guest Lecture