Counterfactual Explanations for Better Grounding

Reid Simmons
Research Professor, Robotics & CS
reids@cs.cmu.edu

Grounding

- What is Grounding?
  - Connecting abstract knowledge to tangible, real-world data
  - E.g., Explicit examples of agent policies

- Why is Grounding Important?
  - Grounded examples help improve understanding and trust
  - Enables more effective and meaningful interactions

Common Ground Theory

- Theory of Communication Between Individuals
  - Participants in an interaction exchange information in order to come to a **common understanding** of the situation
  - Mutual exchange of **knowledge, beliefs, and assumptions**
  - Herbert H. Clark, *Using Language*, Cambridge University Press, 1996
- Objectives:
  - Communicate what you believe the other does not know, but needs to know for the task at hand
  - **Do not** communicate what you believe the other already knows

Common Ground Conventions

- **Principle of Mutual Responsibility**
  - Endeavor to establish mutual beliefs
  - Akin to *model reconciliation* (Rao & Sreedharan)
- **Presentation/Acceptance Process**
  - Back-and-forth protocol
- **Principle of Least Collaborative Effort**
  - Minimize joint effort in establishing mutual belief

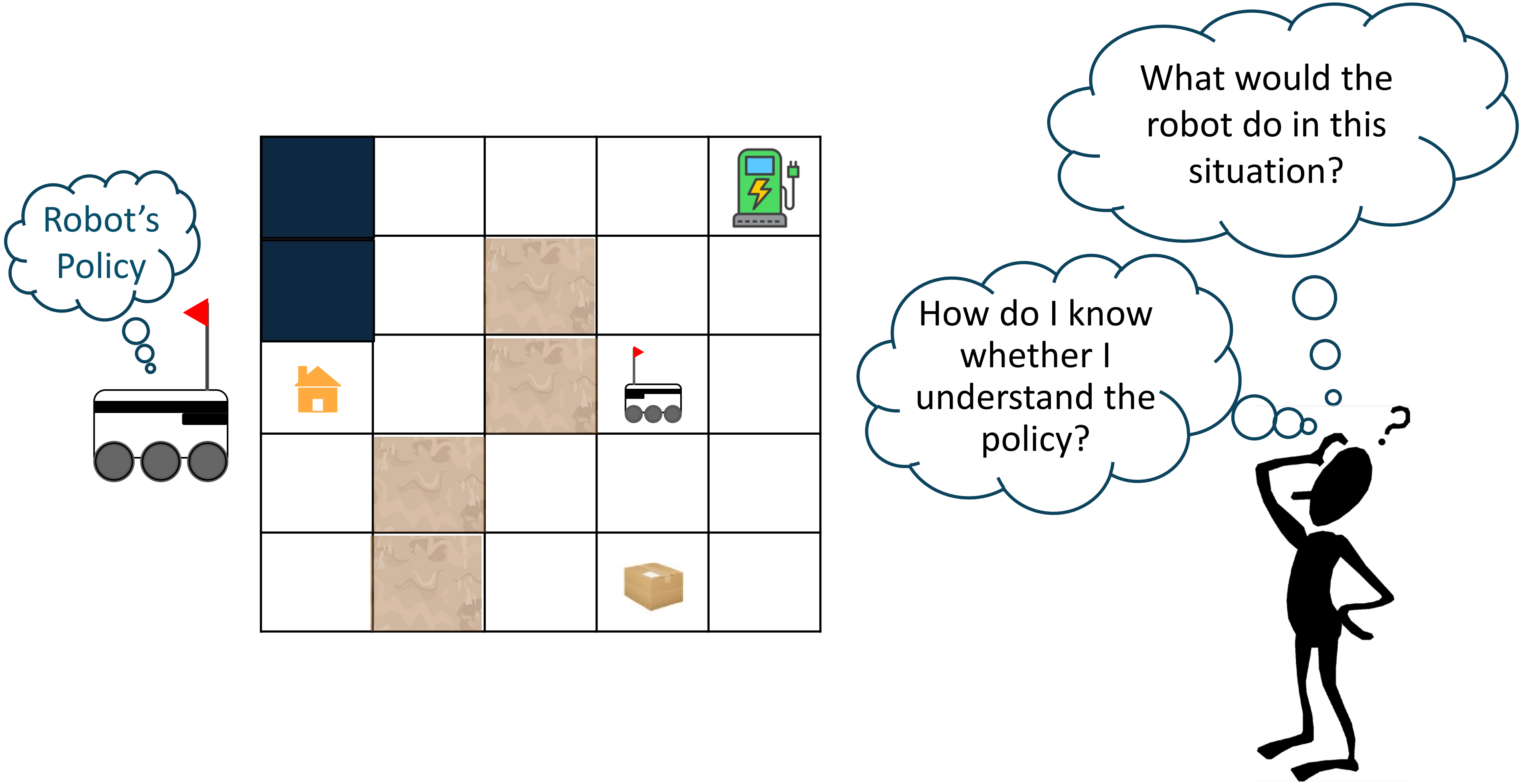
Counterfactuals

- Counterfactual
  - A feature that differs from what occurred
  - E.g., “What if Air Canada had not gone on strike”
- Counterfactual Explanations
  - Explaining a decision in **contrast** to what the person might believe
  - E.g., “I arrived in Montreal on time, since I was not on Air Canada”

Approaches to Counterfactual Reasoning

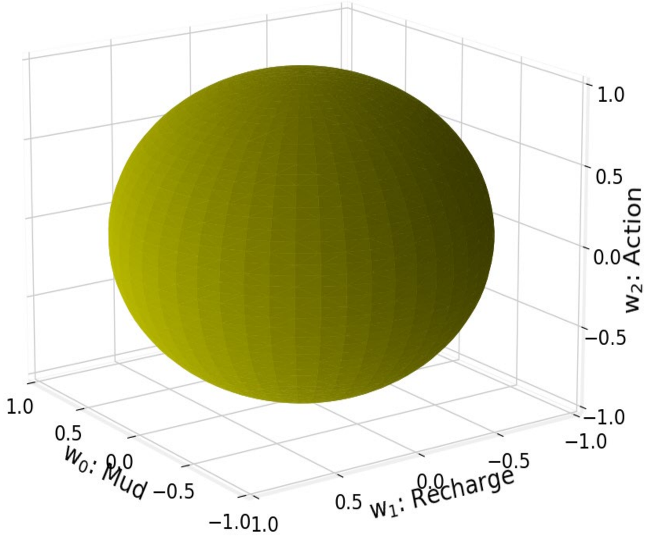
- Counterfactual Demonstrations (Lee, Admoni, Simmons)
- Contrastive Explanations (Sukkerd, Garlan, Simmons)
- Second Order Theory of Mind (Callaghan, Admoni, Simmons)

Counterfactual Demonstrations



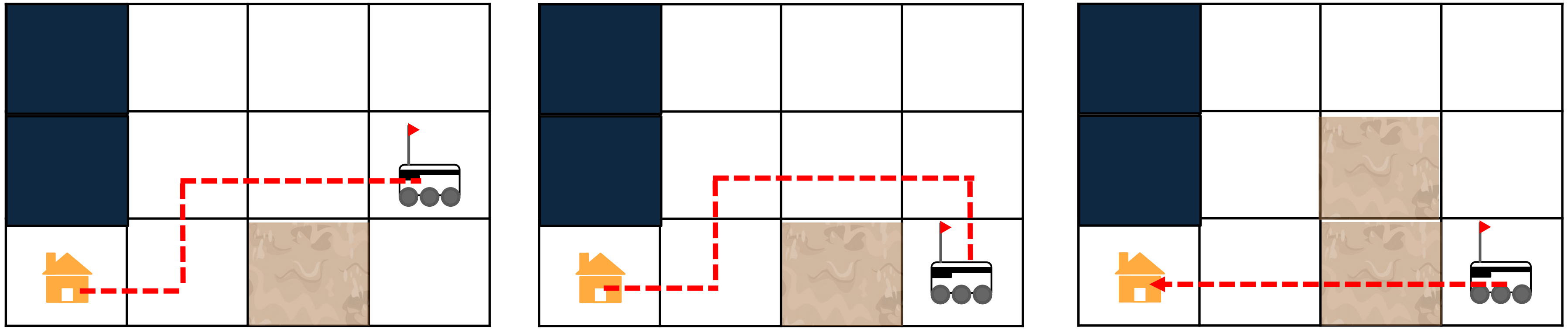
Counterfactual Demonstrations

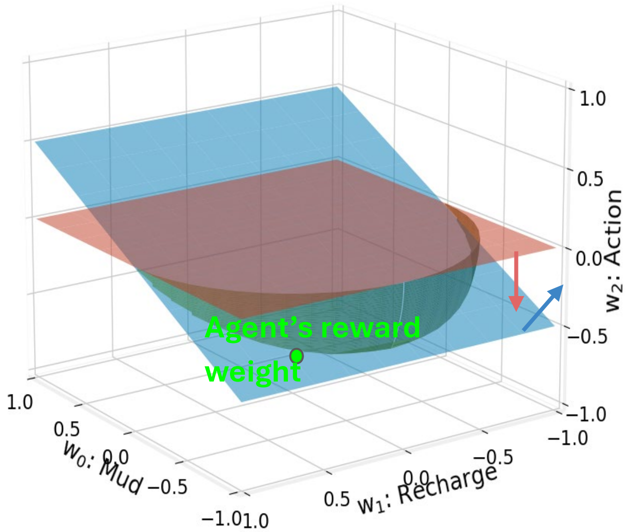
- Goal is for robot to teach its policy by providing **informative demonstrations**
  - Assumes **learning objective** is to understand teacher's policy by determining **reward feature weights**
  - Assumes person is an **imperfect** IRL learner
  - Explicitly model what the human learner is expected to know after each demonstration



Counterfactual Demonstrations

- Model how the potential understanding of the person changes based on demos seen and their responses to tests
  - Choose demonstrations that are **counter** to what a person would choose, given what they are **currently** expected to know

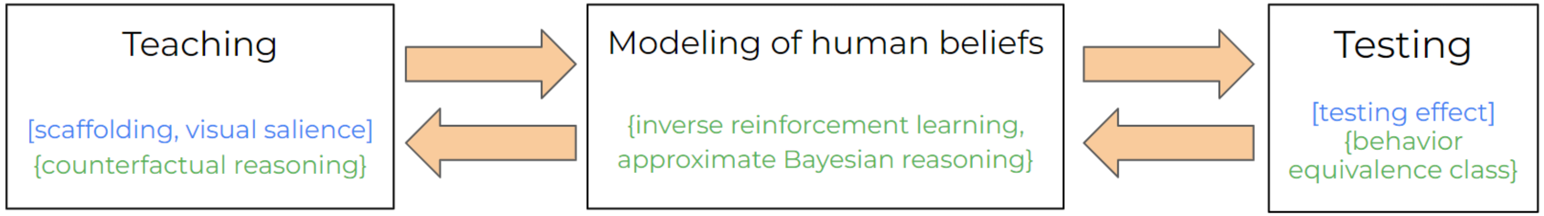




Lee, Admoni, Simmons; Machine Teaching for Human Inverse Reinforcement Learning, 2021

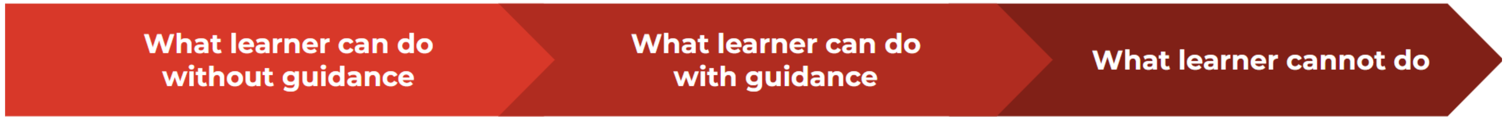
Counterfactual Demonstrations

- Interactive Teaching
  - Scaffold demonstrations to build up knowledge incrementally
  - Use testing to update user model



Key:
[Educational principles]
[Algorithmic principles]

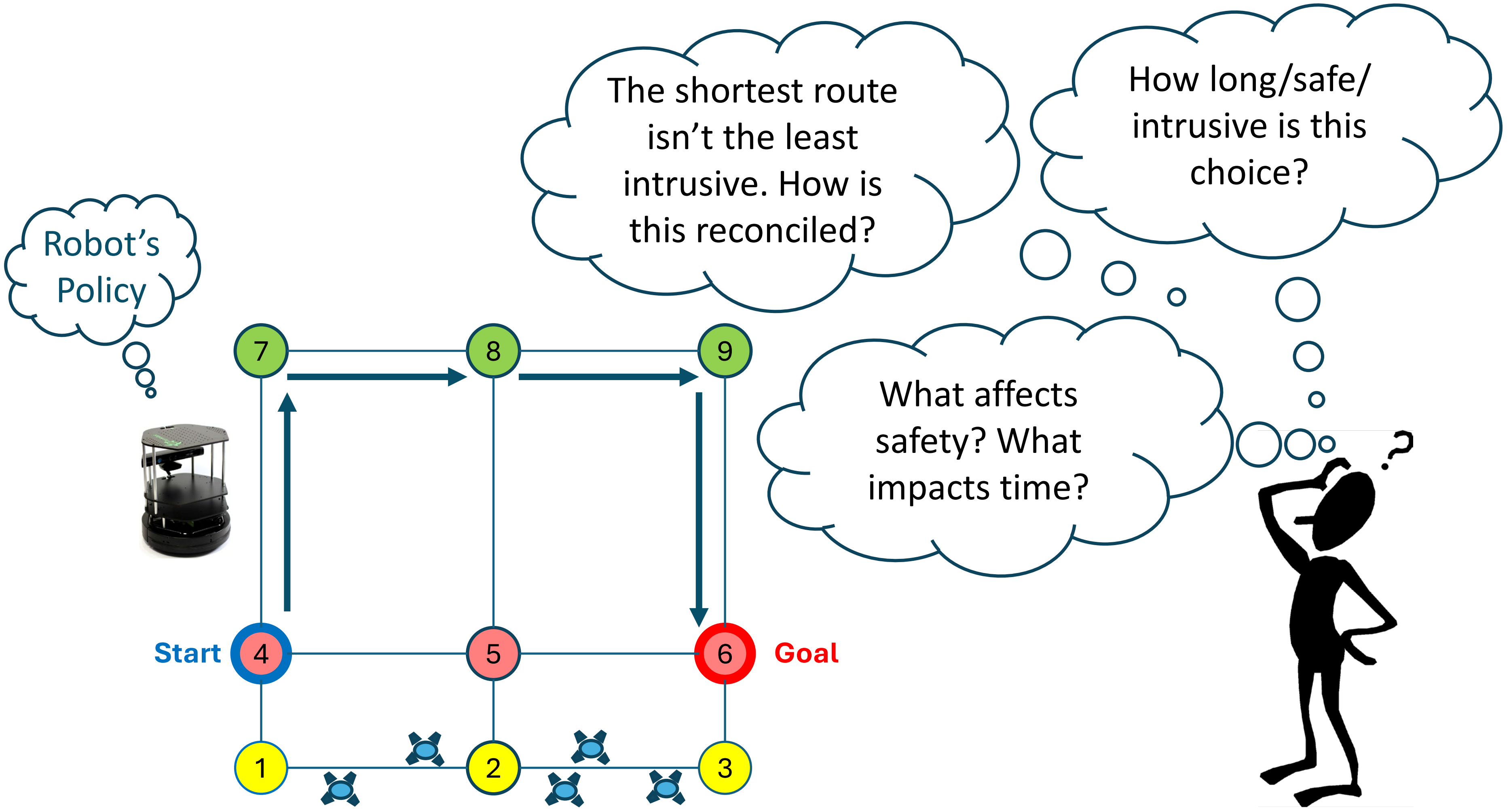
  - Operate within the **Zone of Proximal Development**



Low information gain	Moderate information gain	High information gain
Low difficulty	Moderate difficulty	High difficulty

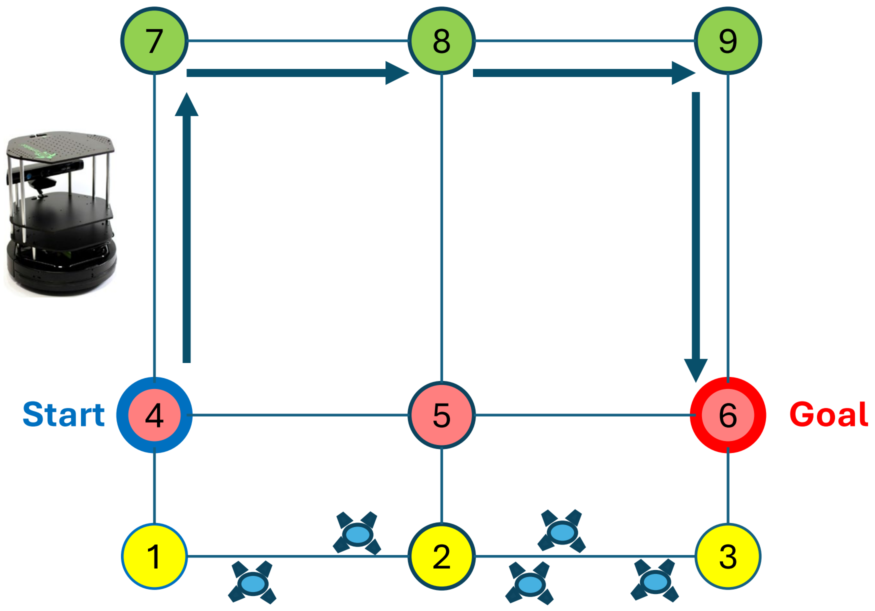
Lee, Simmons, Admoni; Improving the Transparency of Robot Policies Using Demonstrations and Reward Communication, THRI, 2025

Contrastive Explanations



Contrastive Explanations

I'm planning to go to *L6* via *route L4-L7-L8-L9-L6*. It is expected to take **10 minutes**, have **0 collision**, and be **non-intrusive**.



Instead, by **going through** *route L4-L1-L2-L3-L6* I could **reduce *time* to 7 minutes**, but at the expense of **increasing *collisions* to 0.4** and **increasing *intrusiveness* to moderate**.

However, I decided not to do that because the **decrease in *time*** is not worth the **increase in *collision* and *intrusiveness***.

Contrastive Explanations

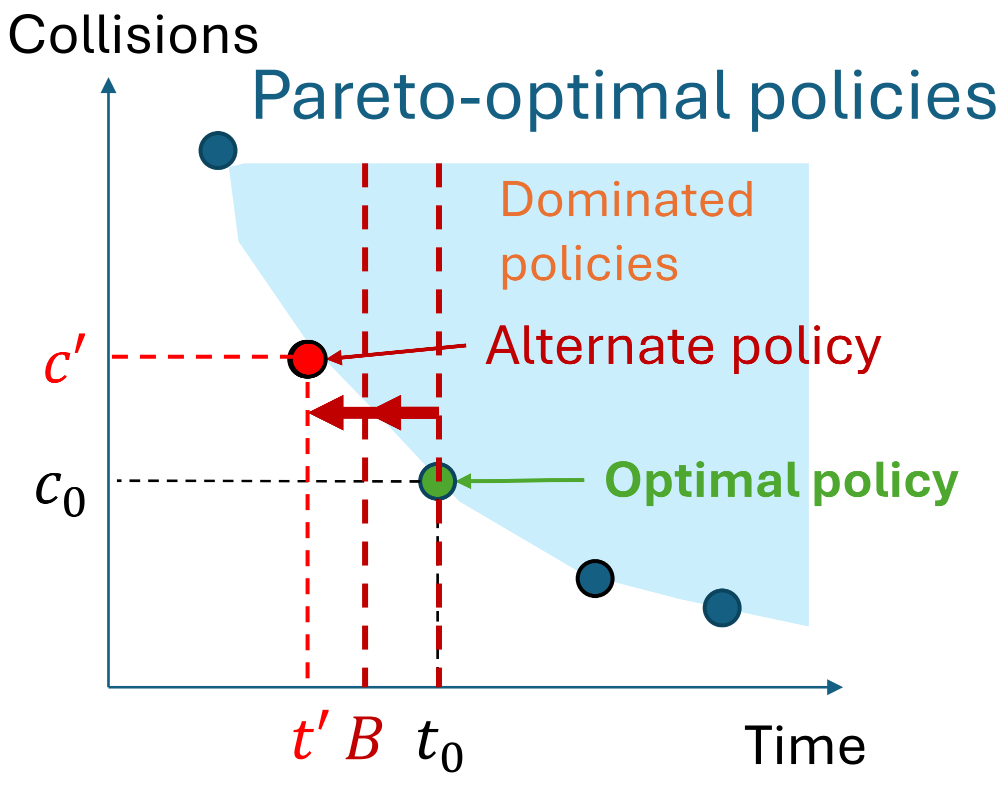
- Goal is to for robot to teach its policy by providing both positive and negative examples
  - “Positive” examples show how the robot would act optimally
  - “Negative” examples, in contrast, show how the robot would act under a different reward function
  - The contrastive reward function should be something the person might think **would have been** optimal

Contrastive Explanations

- **Approach**:
  - Model reward function as linear combination of features
  - Analyze optimal trajectory wrt features (akin to IRL)
  - Generate explanation for that trajectory
  - Select an **alternate reward function** and generate explanation for trajectory under that function
  - Describe provide the reason why the alternate trajectory is not inferior under the optimal reward function

Sukkerd, Simmons, Garlan, Tradeoff-Focused Contrastive Explanation for MDP Planning, *RO-MAN* 2020.

Contrastive Explanations

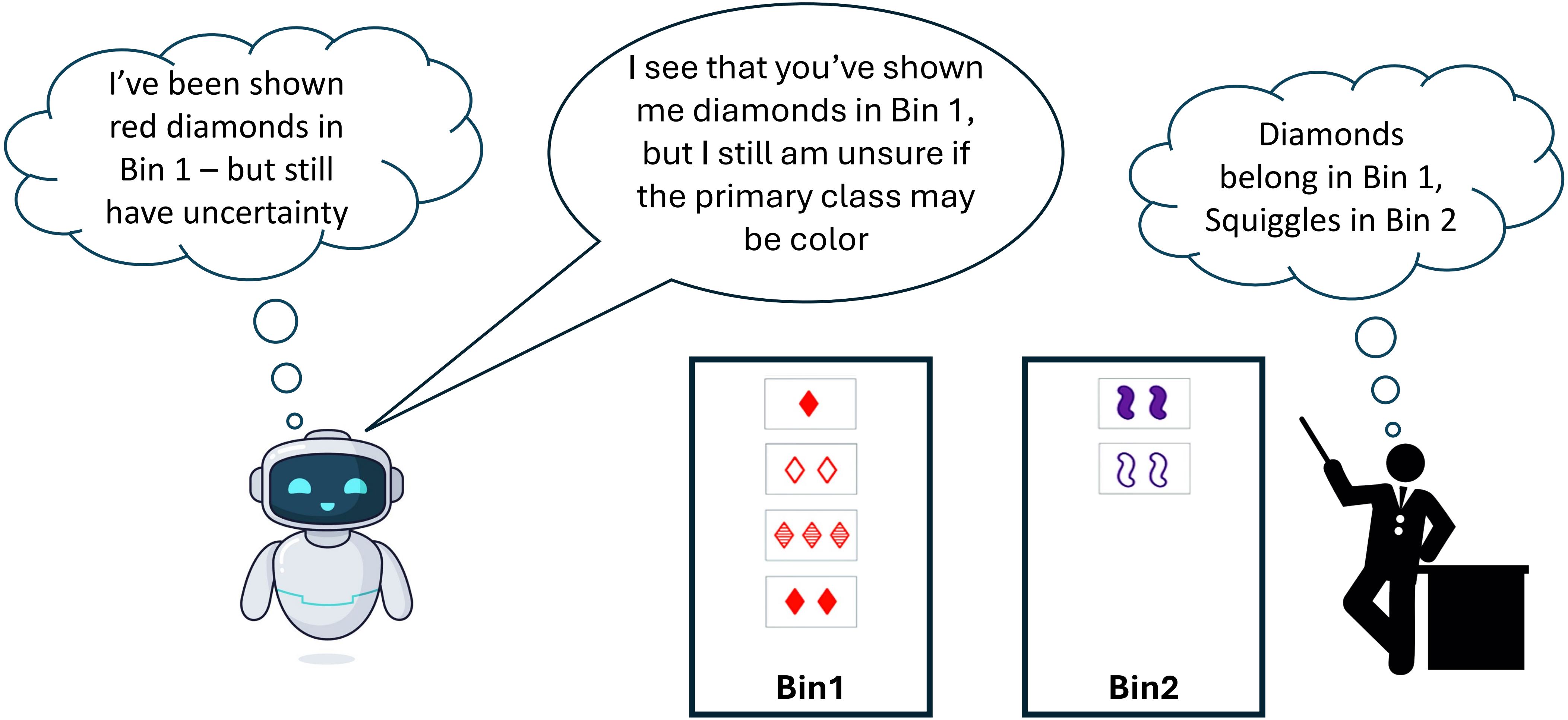


Fix "*Time*" feature value

$$J(s) = \min_{a \in A_s} \left[\boldsymbol{C'(s,a)} + \sum_{s' \in S} Pr(s'|s,a)\, J(s') \right]$$

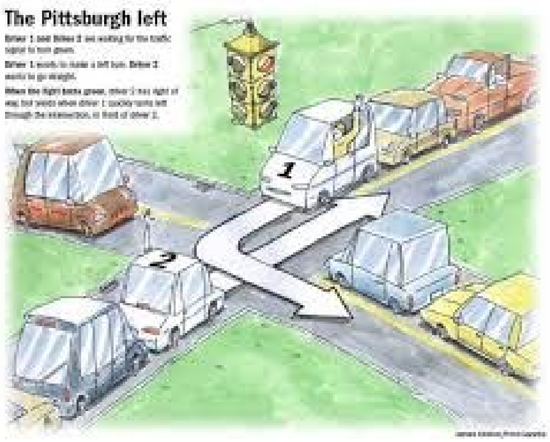
subject to: $J^{\pi^*}_{time}(s) \leq B$

Second Order Theory of Mind



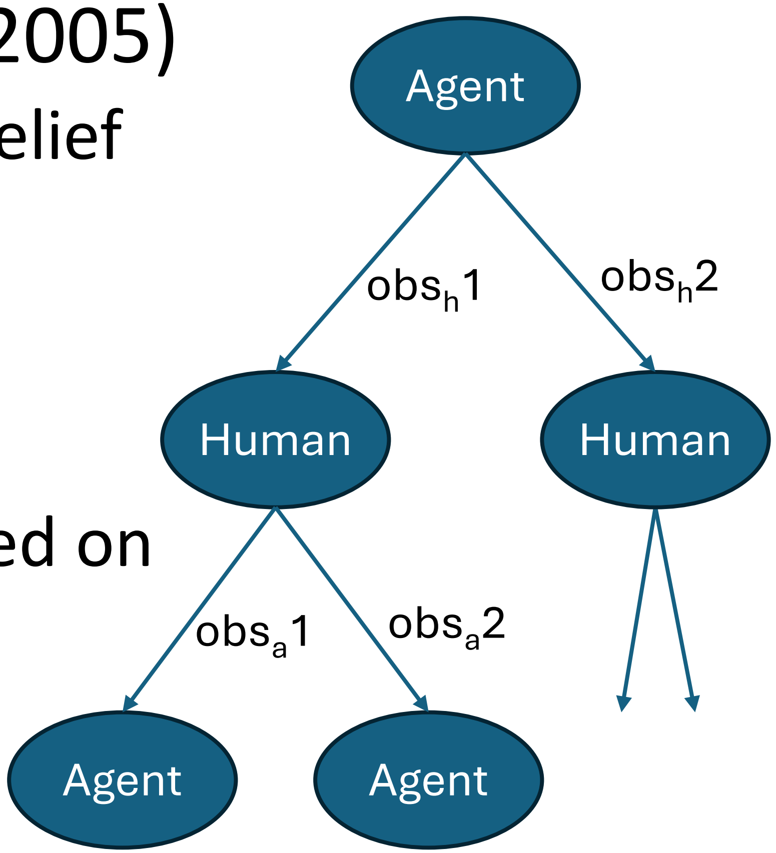
Second Order Theory of Mind

- If robot and human have conflicting beliefs about what each will do, can lead to highly negative behaviors
- Use Second Order Theory of Mind for the robot to model what it believes the person believes about the robot's intentions
    - Robot provides feedback reconcile the beliefs



Second Order Theory of Mind

- Adapting the I-POMDP framework (Doshi, 2005)
  - "Interactive" belief states – own beliefs plus belief over the other agent's beliefs
- Modeling human's observation function in terms of "confirmation bias"
  - Compute using **counterfactual inference**, based on similarity of cards played to possible rules
  - Agent feedback based on difference between its beliefs and inferred human beliefs



Callaghan, Simmons, Admoni; Using Second-order ToM to Account for Human Teacher and Robot Learner Misunderstandings of One Another, Workshop on ToM4AI, AAAI 2025

Second Order Theory of Mind

- Feedback is confidence statement if no confirmation bias detected:
  - “Sure”, “believe”, “unsure”
  - “I’m unsure if the primary class is color”
- Feedback incorporates counterfactual statement if confirmation bias:
  - “I understand that you are trying to …”
  - “I understand that you are trying to show me that diamonds belong in Bin 1, but I’m still unsure whether the primary class could be color”

Summary

- **Counterfactuals** are very useful for establishing common ground
  - They efficiently establish the differences between what agents believe
- **Counterfactual demonstrations** utilize models of user beliefs to guide understanding, in line with educational principles
- **Contrastive explanations** provide users understanding of why alternate solutions are not, in fact, optimal, based on the agent's actual reward function
- **Theory of Mind** enables agents to provide feedback that corrects for user confirmation bias, enabling more effective learning