

Robot Learning

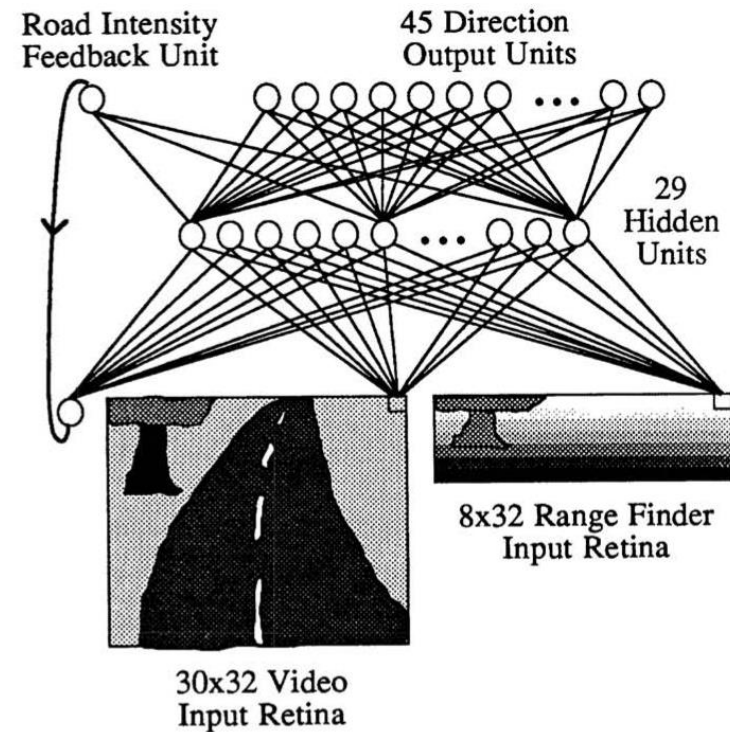
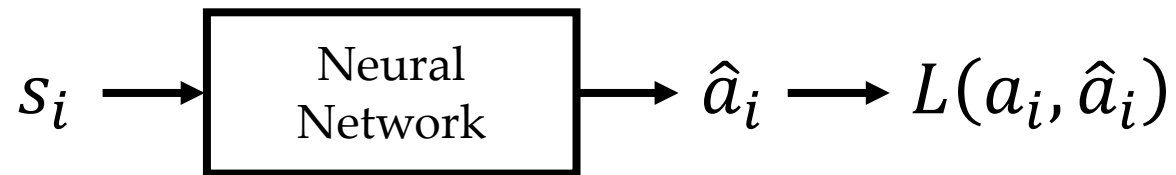
Learning from human feedback

Last time...

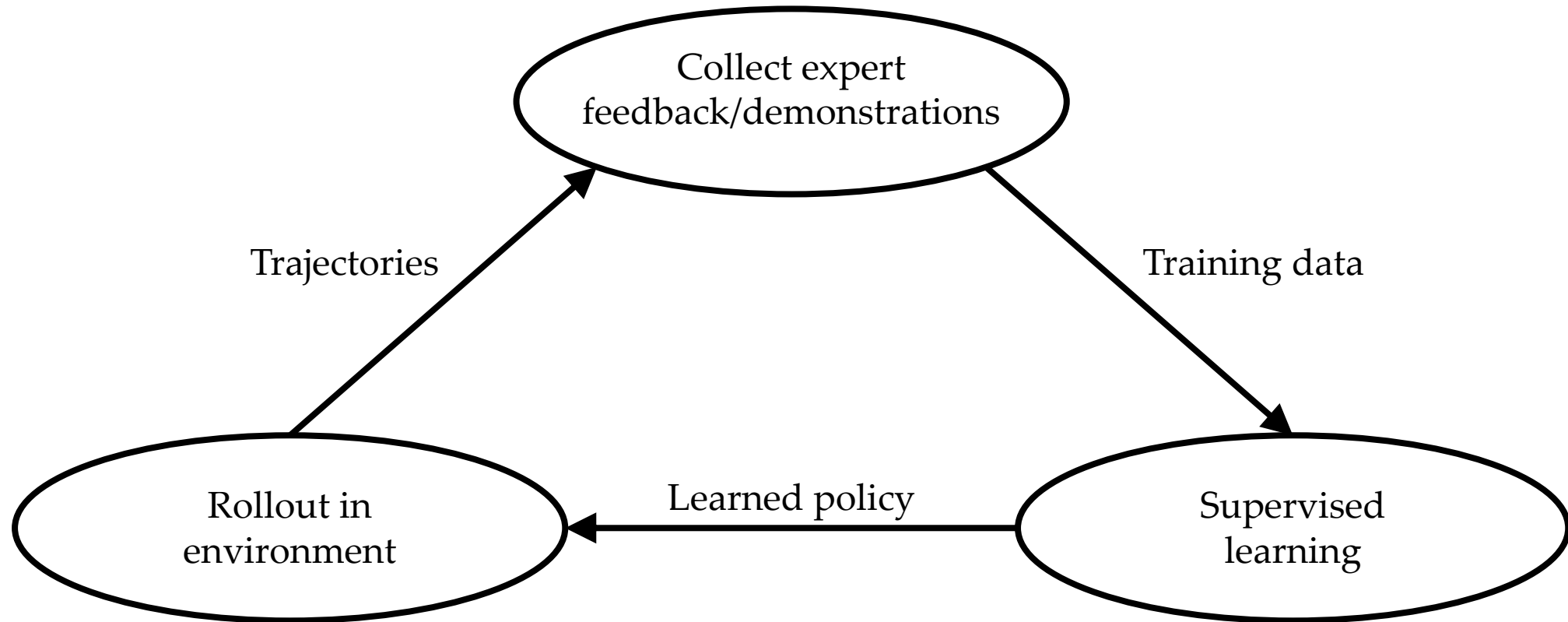
- Imitation learning
- Inverse reinforcement learning (IRL)
 - Apprenticeship learning
 - Maximum margin planning
 - Max-Ent IRL

Behavioral cloning

Train a neural network to map states into expert actions.



Direct policy learning



Apprenticeship learning

We iteratively improve the learned w and policy.

Compute the optimal features $\phi(\pi^*)$

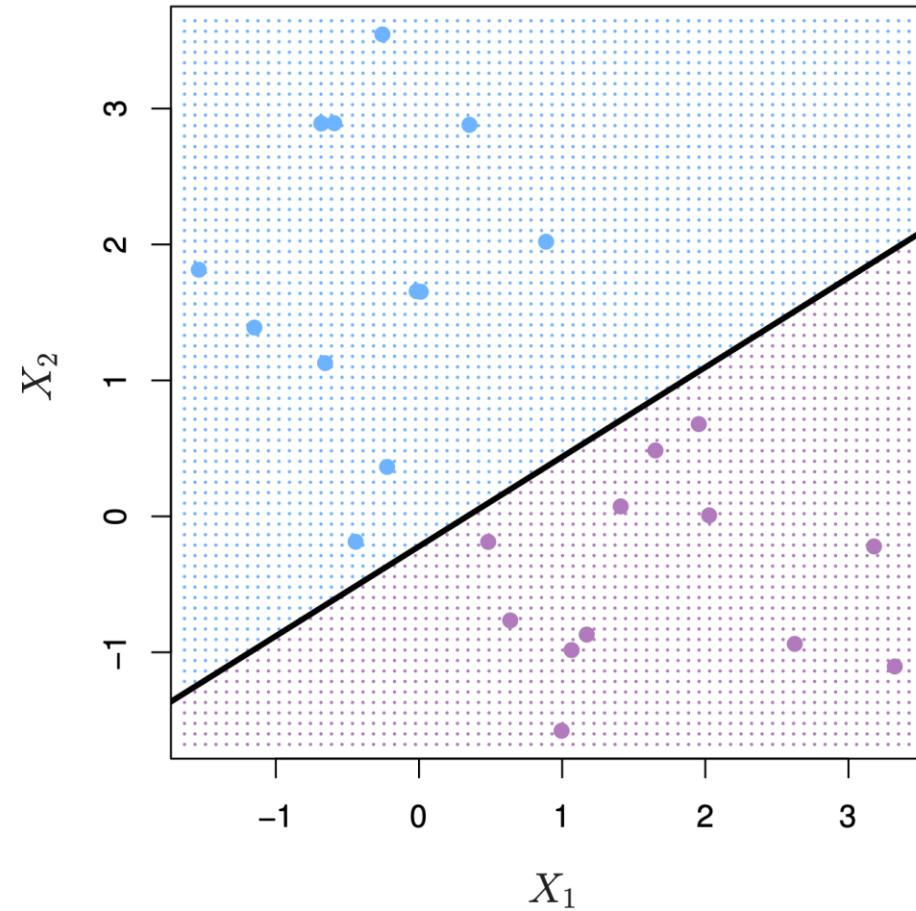
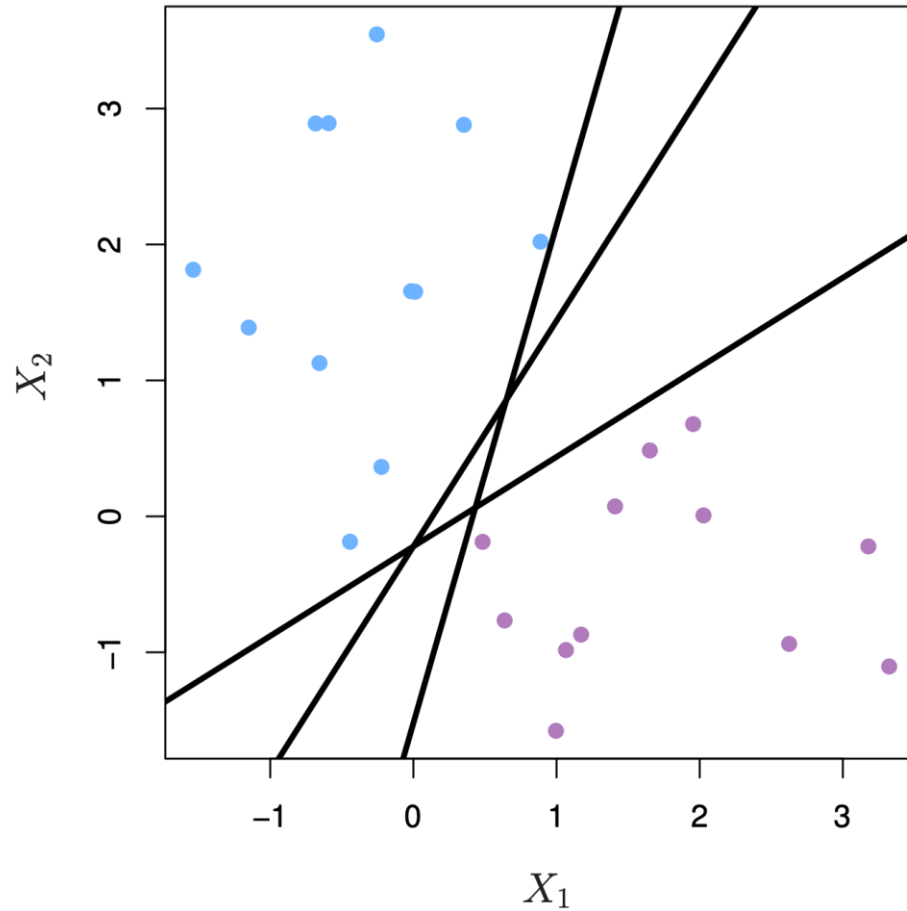
Initialize a policy π_0

Loop $i = 0, 1, \dots$:

Find w_i that best separates π^* from π_i

Assuming w_i is true weights, learn π_{i+1} optimizing the reward

Maximal margin classifiers



Maximum margin planning (MMP)

Let's allow the expert to be suboptimal by adding a slack variable.

We could also be more tolerant to the policies that are similar to π^* .

$$\underset{w, v}{\text{minimize}} \quad \|w\|_2 + Cv$$

$$\text{subject to} \quad w^\top \phi(\pi^*) - w^\top \phi(\pi) \geq 1 - v + d(\pi^*, \pi) \quad \text{for all } \pi \neq \pi^*$$

Max-Ent IRL

1. Initialize w
2. Perform RL to learn a policy that optimizes the reward with w
3. Roll out the learned policy to compute:

$$w \leftarrow w - \mathbb{E}_{\xi \sim P(\xi|w)}[\phi(\xi)] - \phi(\pi^*)$$

4. Repeat from step 2

Max-Ent IRL

Assumption: Experts are noisily optimal, i.e., the probability that they demonstrate trajectory ξ is:

$$P(\xi \mid w) = \frac{\exp(w^\top \phi(\xi))}{\int \exp(w^\top \phi(\xi')) d\xi'}$$

where $\phi(\xi)$ is the cumulative discounted features of trajectory ξ .

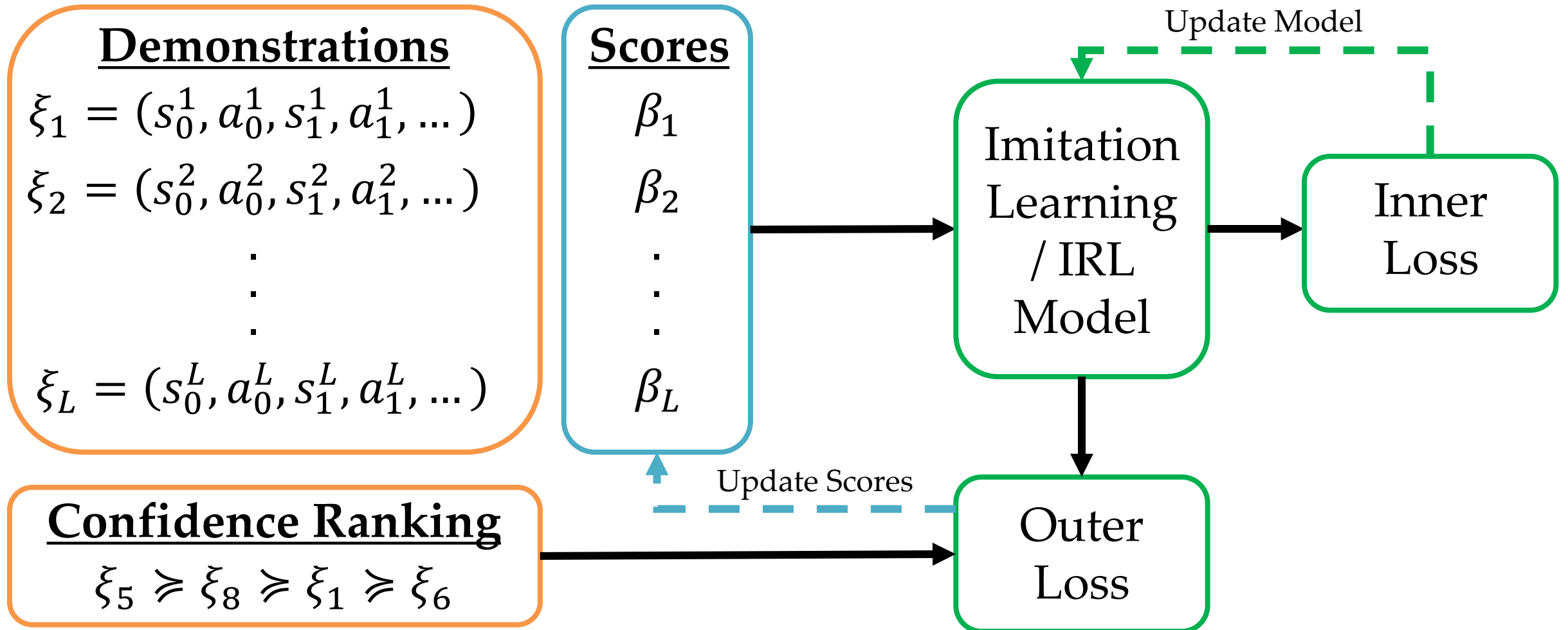
Today...

- Learning from human feedback
 - Suboptimal demonstrations
 - Pairwise comparisons
 - Reinforcement learning from human feedback (RLHF)

Confidence-aware imitation learning

Assume we have some (partial) ranking over expert demonstrations. How would that help?

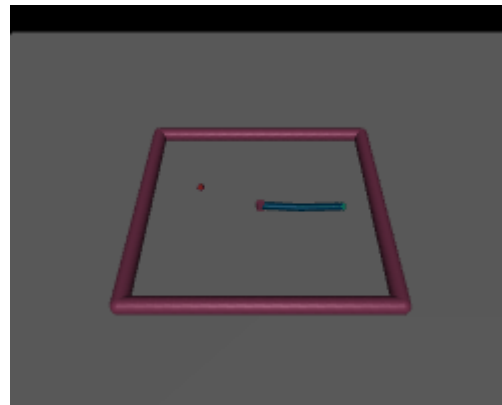
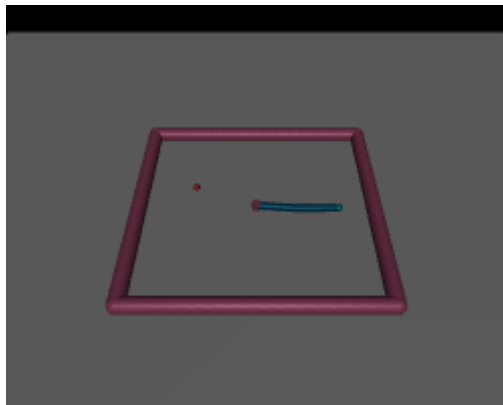
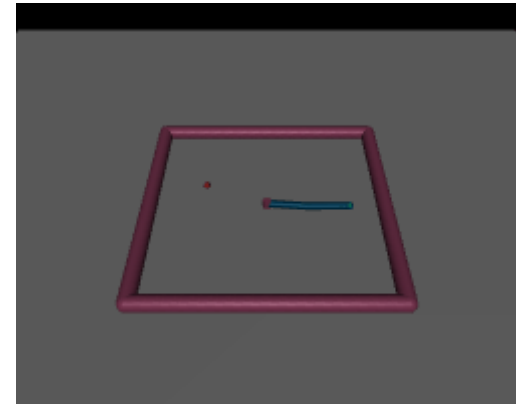
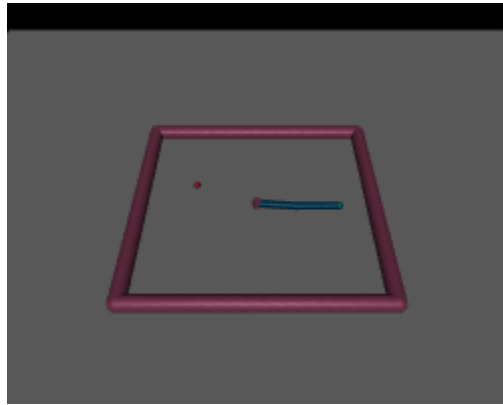
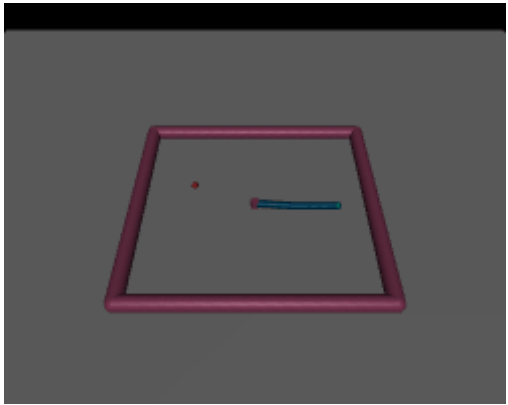
Confidence-aware imitation learning



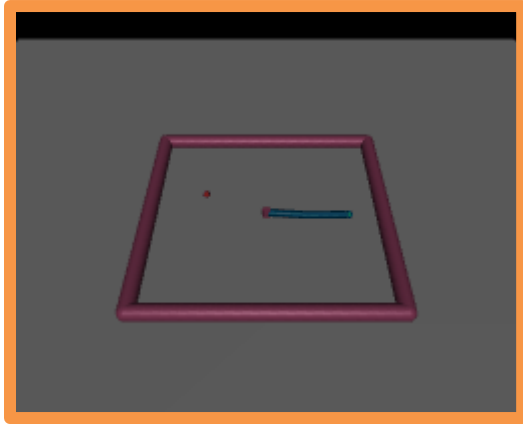
Confidence-aware imitation learning

- Inner loss to learn a policy / reward
 - Uses the demonstrations ξ_i weighted with their confidence scores β_i
 - Learns a policy (or a reward function)
- Outer loss to learn confidence scores
 - Evaluates how well the demonstrations match the given (partial) ranking under the learned policy (or reward) to update β_i 's.

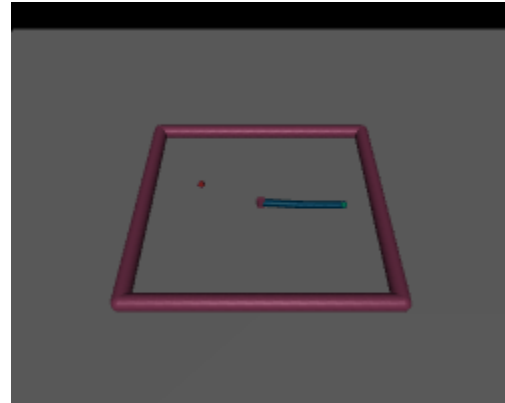
Confidence-aware imitation learning



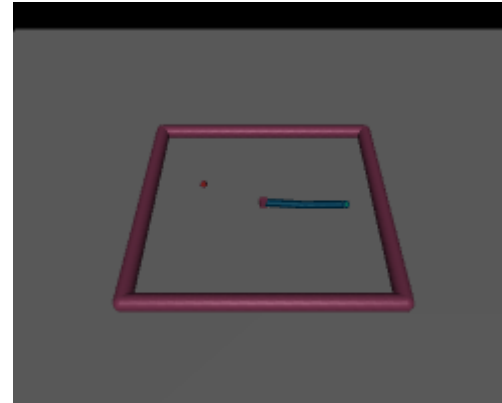
Confidence-aware imitation learning



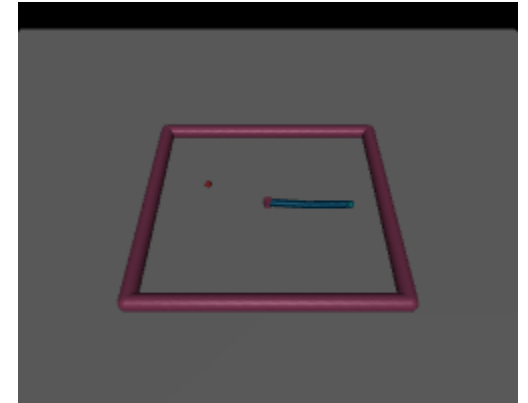
CAIL



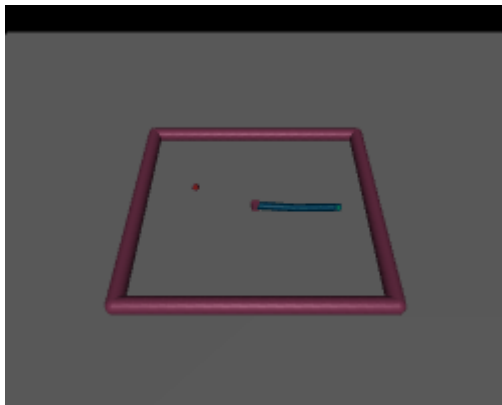
2IWIL



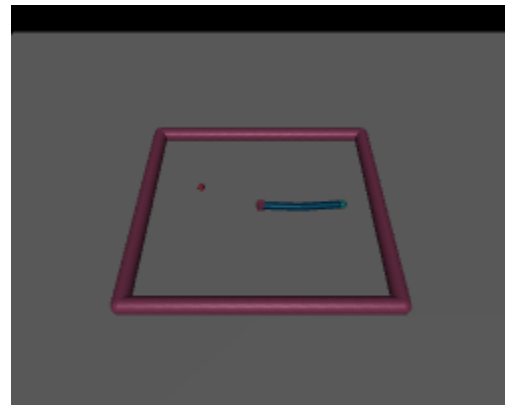
IC-GAIL



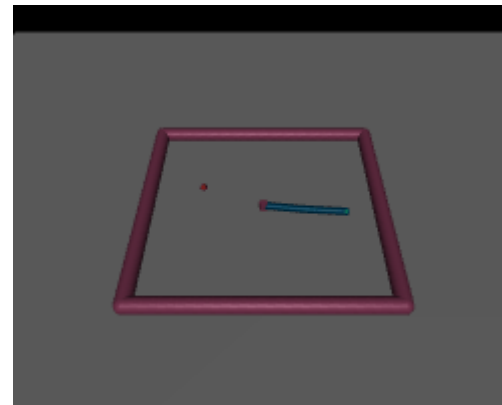
AIRL



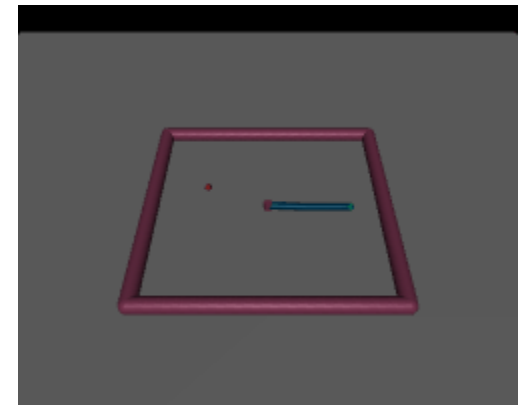
GAIL



T-REX



D-REX



SSRR

Further questions to think about

- What if the demonstrators' suboptimality is context-dependent?
 - Example: I can teleoperate a robot well in coarse actions, but I am not good at precise manipulation motions.
- How do we choose the expert to query if we have that ability?

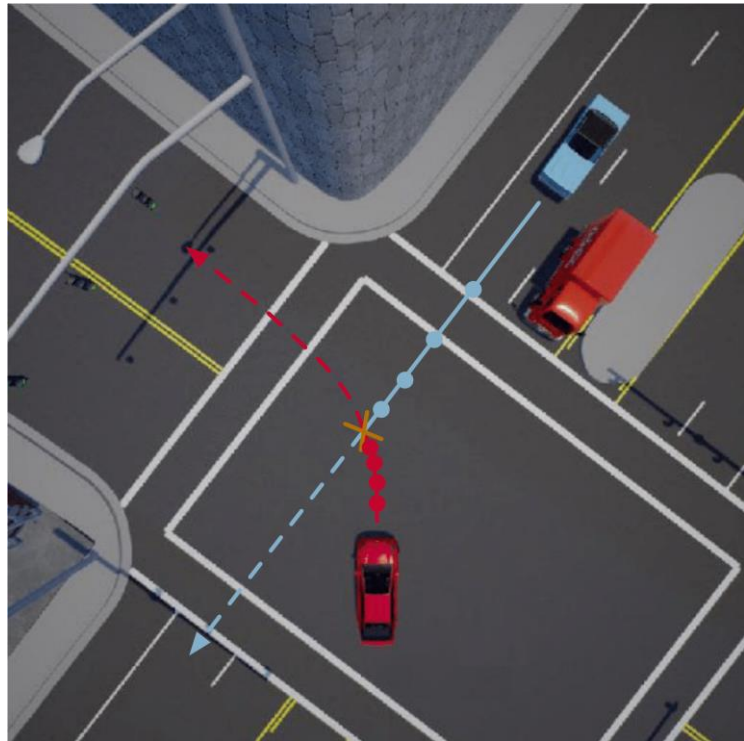
Today...

- Learning from human feedback
 - Suboptimal demonstrations
 - Pairwise comparisons
 - Reinforcement learning from human feedback (RLHF)

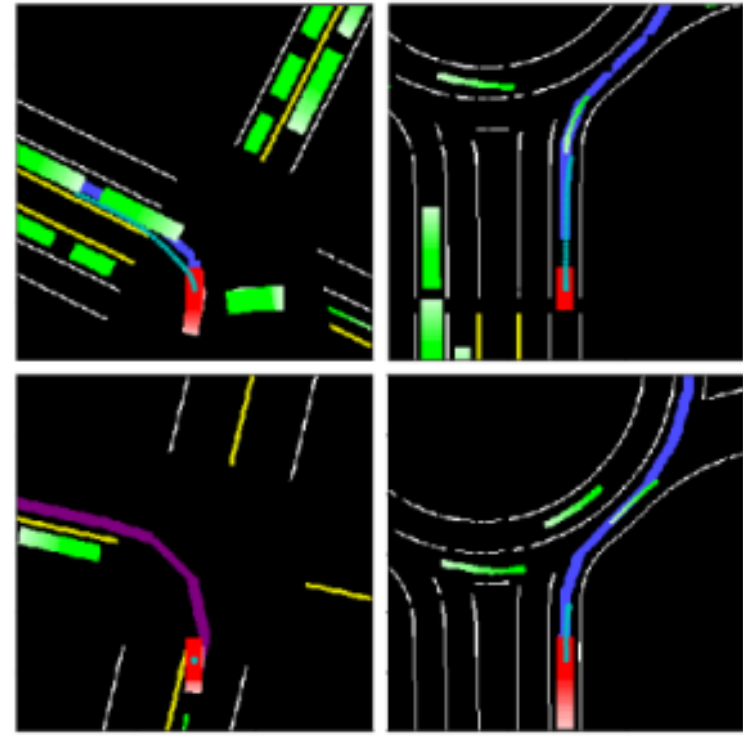
Learning from demonstrations (LfD)



Codevilla et al. ICRA'18



Cao et al. RSS'20



Chen et al. IROS'19

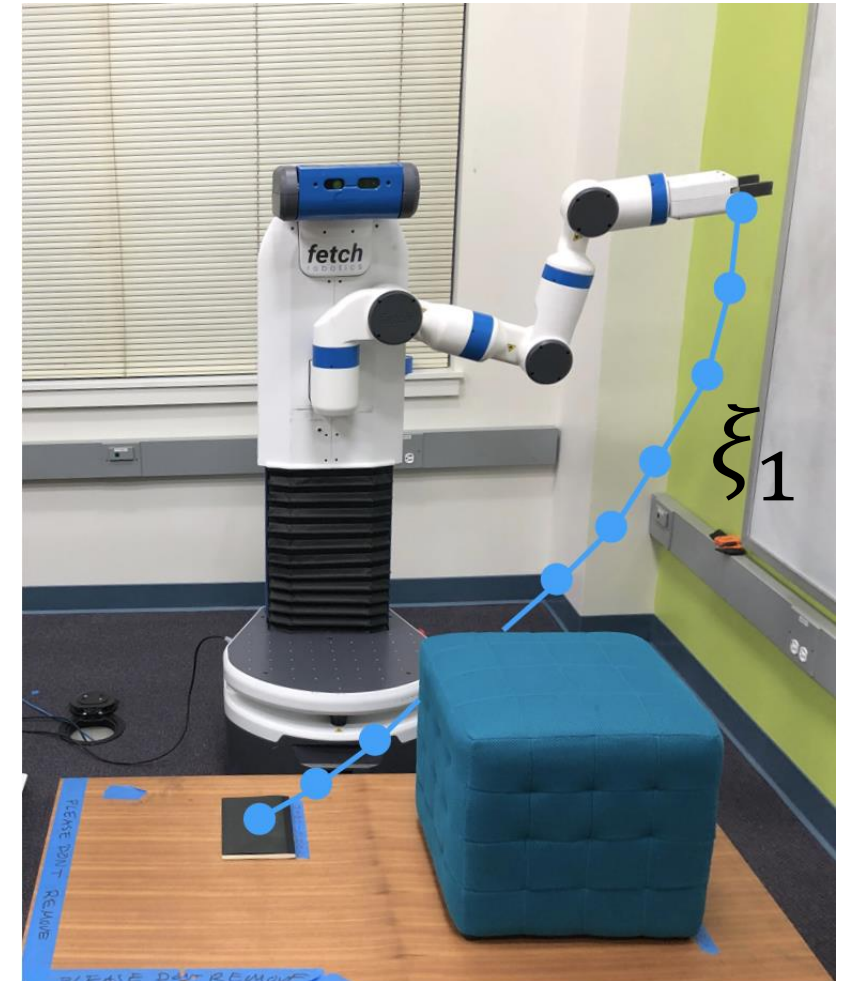
Why does LfD fail?

Demonstrations: $\mathcal{D} = \{\xi_1, \xi_2, \dots, \xi_L\}$

Trajectory features: $\phi(\xi_i) = \phi_i \in \mathbb{R}^d$

- Final distance to the notebook
- Minimum distance to the obstacle
- Average speed
- ...

Reward function: $R(\xi_i) = \underline{f_w}(\phi_i)$

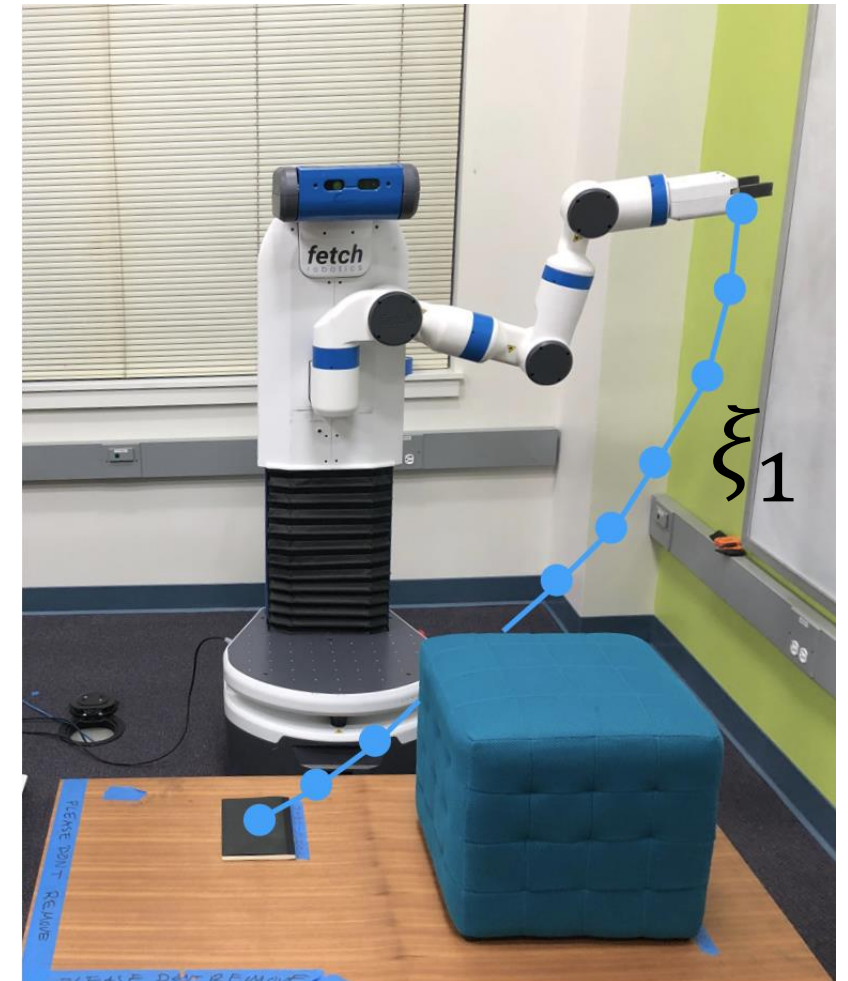


Bayesian inverse reinforcement learning

$$\begin{aligned} & \operatorname{argmax}_w P(w \mid \mathcal{D}) \\ P(w \mid \mathcal{D}) & \propto P(w) \underbrace{P(\mathcal{D} \mid w)} \\ & = P(w) \prod_{i=1}^L P(\xi_i \mid w) \end{aligned}$$

$$\times \quad \propto P(w) \prod_{i=1}^L \exp f_w(\xi_i)$$

(Noisy humans)



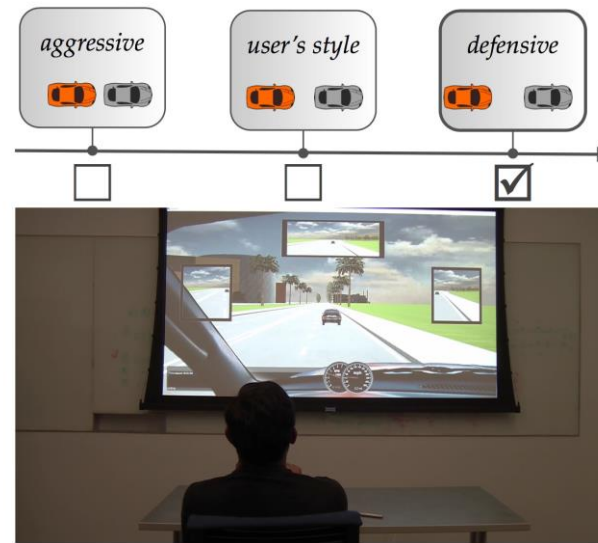
Humans are Suboptimal

Robots with high degrees of freedom are hard to teleoperate.



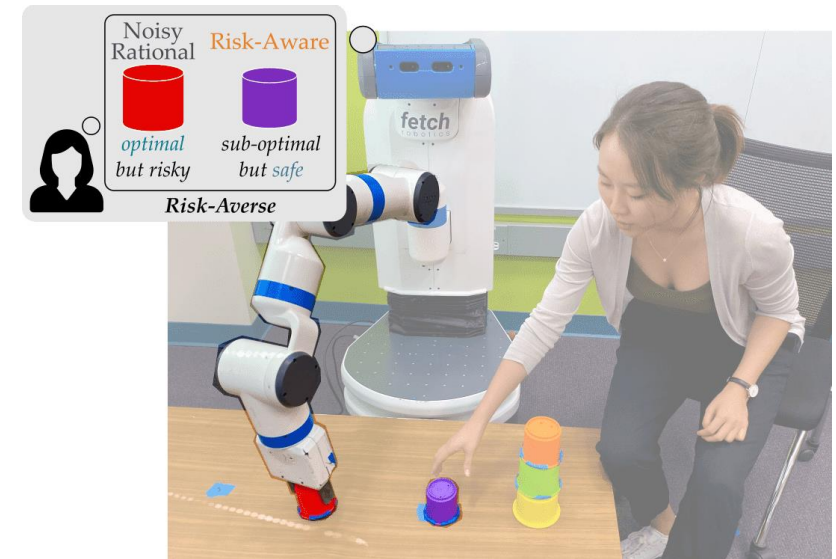
Palan et al. RSS'19

Humans do not like their own demonstrations.



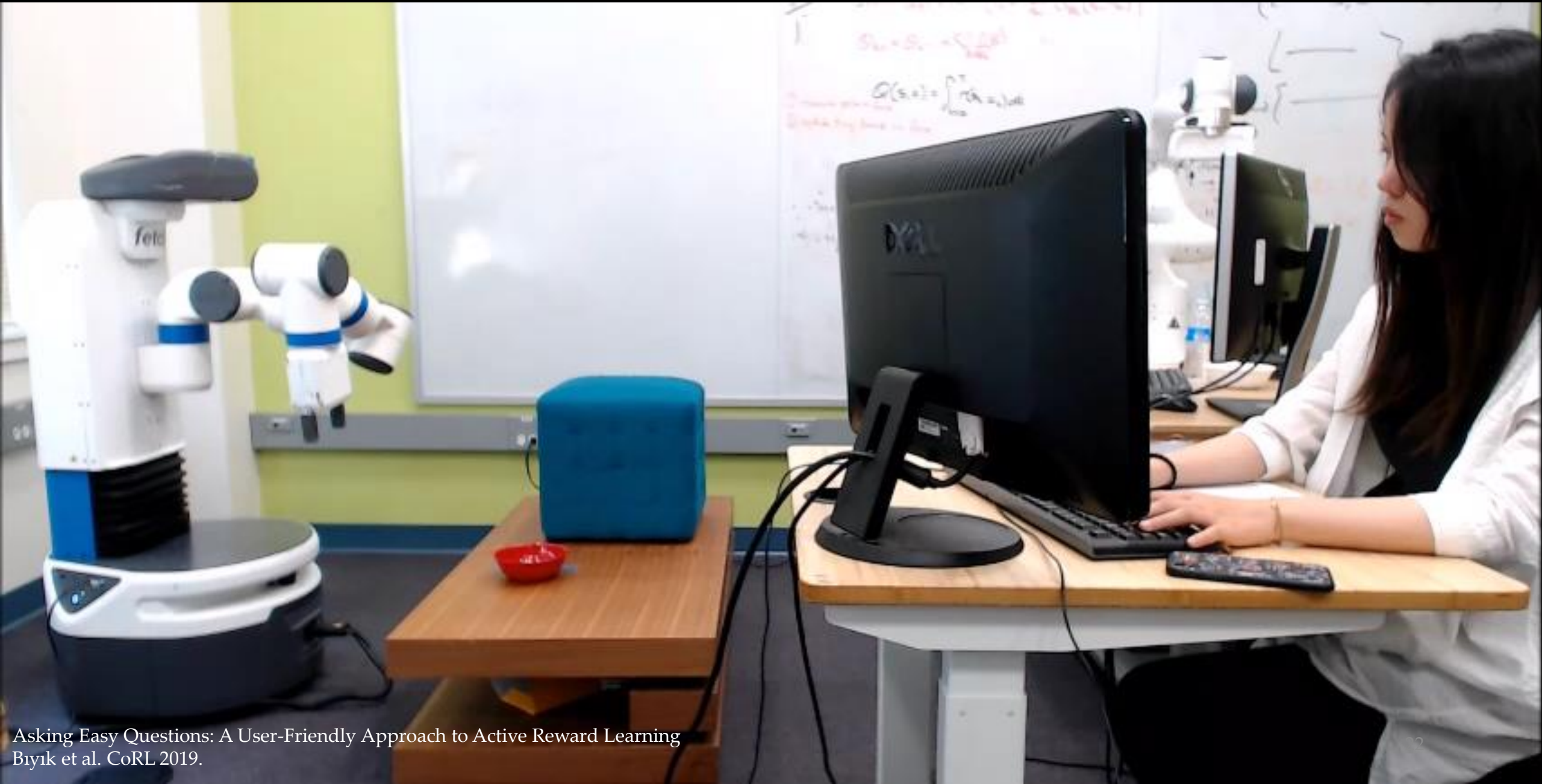
Basu et al. HRI'17

Humans take suboptimal actions in risky situations.



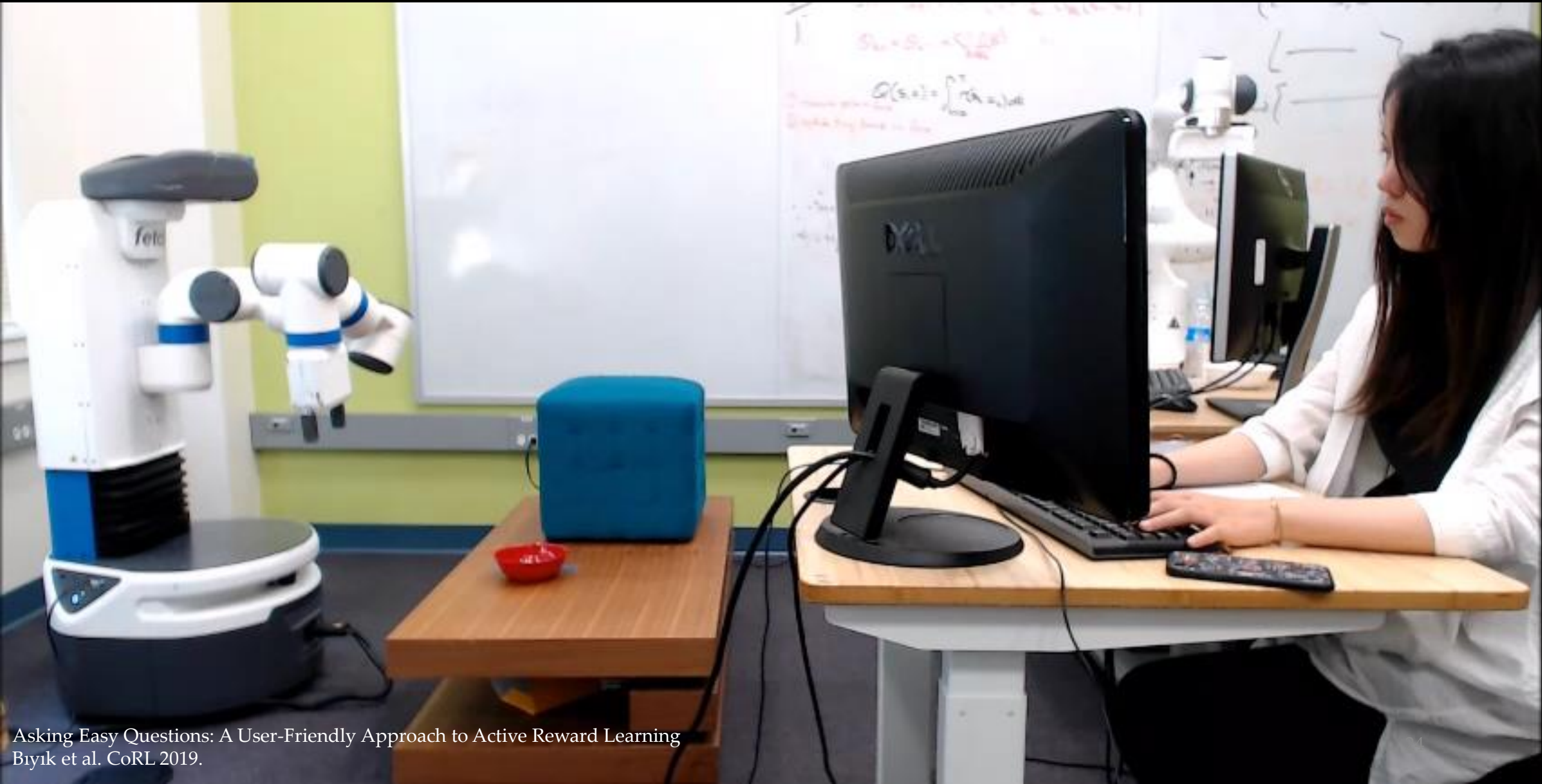
Kwon et al. HRI'20

We can let the human evaluate a robot demonstration



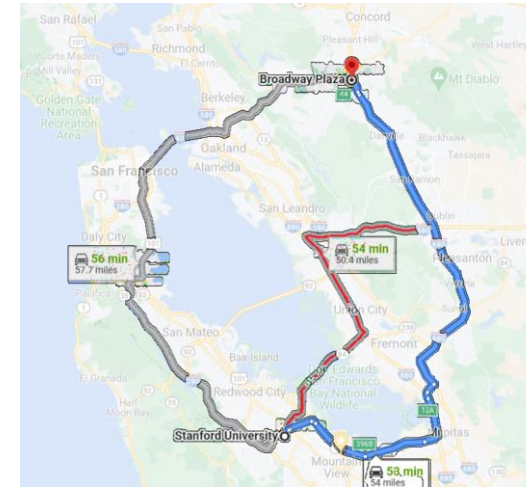
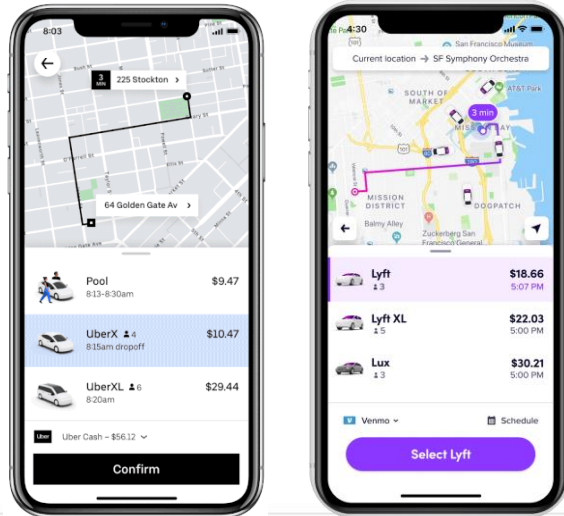
How dark is this blue?

Human evaluations are often unreliable

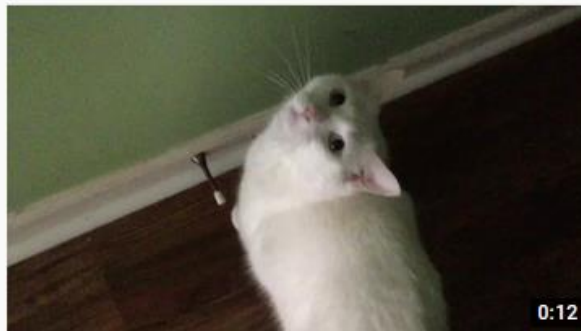


Which blue is darker?

Comparison data



Mohsen Namjoo & Nederlands Blazers Ensemble - Nobahaari" @ Theaters...
Café Nim
71K views • 5 years ago



how my deaf cat meows
Thundy
7.6M views • 11 months ago

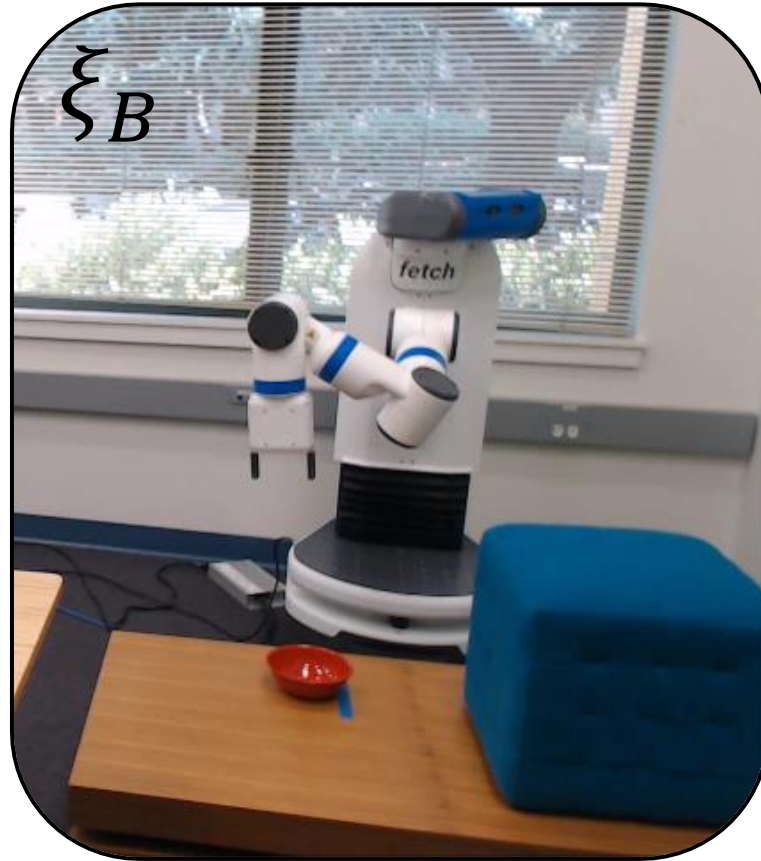
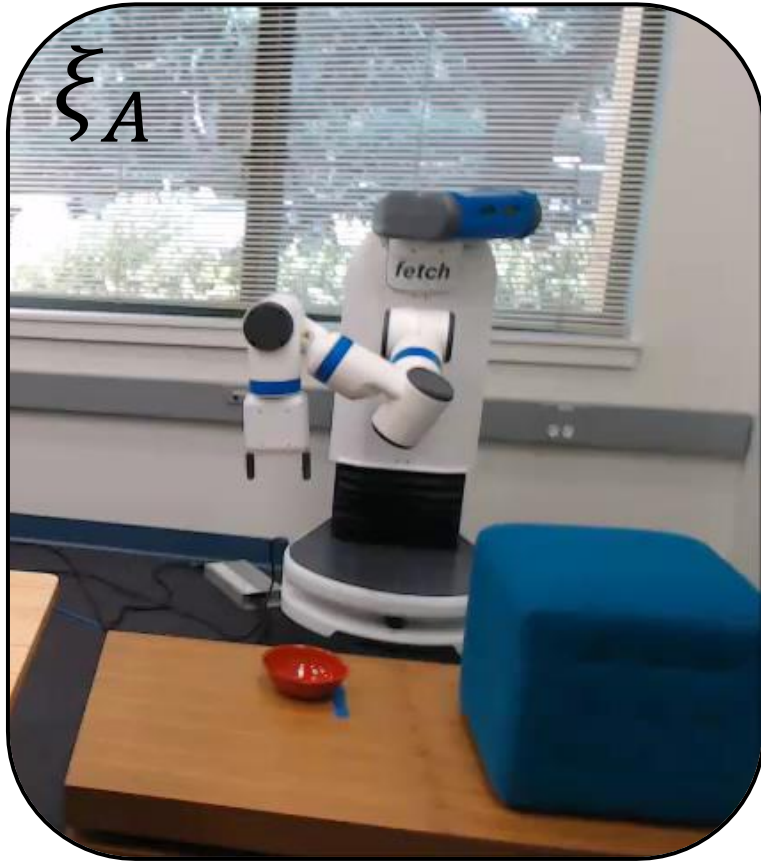


John Lowe 9-dart finish FIRST EVER ON TV
Unicorn Darts
6.6M views • 4 years ago



Carlsen - Nepomniachtchi | Game 8 | World Chess Championship | Howell,...
chess24
24K watching

Incorporating Comparisons



ξ_A or ξ_B ?



Incorporating Comparisons

Demonstrations: $\mathcal{D} = \{\xi_1, \xi_2, \dots, \xi_L\}$

Comparisons: $\mathcal{C} = \left\{ \left(\xi_A^{(1)}, \xi_B^{(1)}, q^{(1)} \right), \dots, \left(\xi_A^{(N)}, \xi_B^{(N)}, q^{(N)} \right) \right\}$

Trajectory features: $\phi(\xi_i) = \phi_i \in \mathbb{R}^d$

- Final distance to the notebook
- Minimum distance to the obstacle
- Average speed
- ...

Reward function : $R(\xi_i) = \underline{f_w}(\phi_i)$

Incorporating Comparisons

$$\operatorname{argmax}_w P(w \mid \mathcal{D}, \mathcal{C})$$

$$P(w \mid \mathcal{D}, \mathcal{C}) \propto P(w)P(\mathcal{D} \mid w)P(\mathcal{C} \mid w)$$

$$= P(w) \prod_{i=1}^L P(\xi_i \mid w) \prod_{i=1}^N P(q^{(i)} \mid w, \xi_A^{(i)}, \xi_B^{(i)})$$

How do we compute this?

Confidence-aware imitation learning

- Inner loss to learn a policy / reward
 - Uses the demonstrations ξ_i weighted with their confidence scores β_i
 - Learns a policy (or a reward function)

- Outer loss to learn confidence scores
 - Evaluates how well the demonstrations match the given (partial) ranking under the learned policy (or reward) to update β_i 's.

How does this exactly work?

Given a reward w , what's the probability that the human gives that ranking?

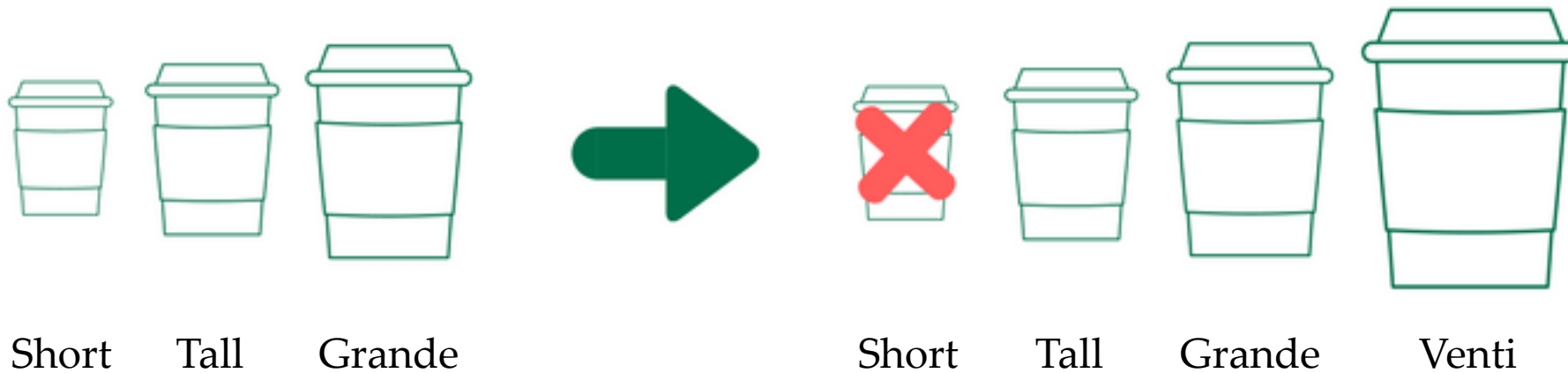
Luce's choice axiom

The probability of selecting one item over another from a pool of many items is not affected by the presence or absence of other items in the pool.

Selection of this kind is said to have *independence from irrelevant alternatives*.

Counterexamples for fun

- Starbucks: *“Compromise Effect”*



Counterexamples for fun

- Coca Cola vs. Pepsi



1985 (Spring)
(was successful only in LA)



1985 (Summer)

Regardless...

The probability of selecting one item over another from a pool of many items is not affected by the presence or absence of other items in the pool.

Selection of this kind is said to have *independence from irrelevant alternatives*.

Corollary

$$P(\xi_i \succcurlyeq \xi_j \succcurlyeq \xi_k) = P(\xi_i \succcurlyeq \xi_j, \xi_k)P(\xi_j \succcurlyeq \xi_k)$$

We only need to model the probability that the human chooses trajectory ξ over a pool of many trajectories.

Incorporating comparisons

$$\operatorname{argmax}_w P(w \mid \mathcal{D}, \mathcal{C})$$

$$P(w \mid \mathcal{D}, \mathcal{C}) \propto P(w)P(\mathcal{D} \mid w)P(\mathcal{C} \mid w)$$

$$= P(w) \prod_{i=1}^L P(\xi_i \mid w) \prod_{i=1}^N P(q^{(i)} \mid w, \xi_A^{(i)}, \xi_B^{(i)})$$

How do we compute this?

Models from discrete choice theory

$$P(q \mid w, \xi_A, \xi_B)$$

Thurstonian Model:

- Add Gaussian noise to the rewards:

- $u_A = f_w(\phi(\xi_A)) + z_A$

- $u_B = f_w(\phi(\xi_B)) + z_B$

where $z_A, z_B \sim \mathcal{N}(0, \sigma^2)$.

- The human choice is the noisy winner:

- $q = \begin{cases} A, & \text{if } u_A > u_B \\ B, & \text{otherwise} \end{cases}$

$$P(q = A) = P(u_A > u_B)$$

$$= P(f_w(\phi(\xi_A)) + z_A > f_w(\phi(\xi_B)) + z_B)$$

$$= P(z_A - z_B > f_w(\phi(\xi_B)) - f_w(\phi(\xi_A)))$$

Models from discrete choice theory

$$P(q \mid w, \xi_A, \xi_B)$$

Bradley-Terry Model:

- The probability that the user chooses an option is proportional to the exponentials of the rewards:

$$P(q = A) = \frac{e^{\beta f_w(\phi(\xi_A))}}{e^{\beta f_w(\phi(\xi_A))} + e^{\beta f_w(\phi(\xi_B))}}$$

Incorporating comparisons

$$\operatorname{argmax}_w P(w \mid \mathcal{D}, \mathcal{C})$$

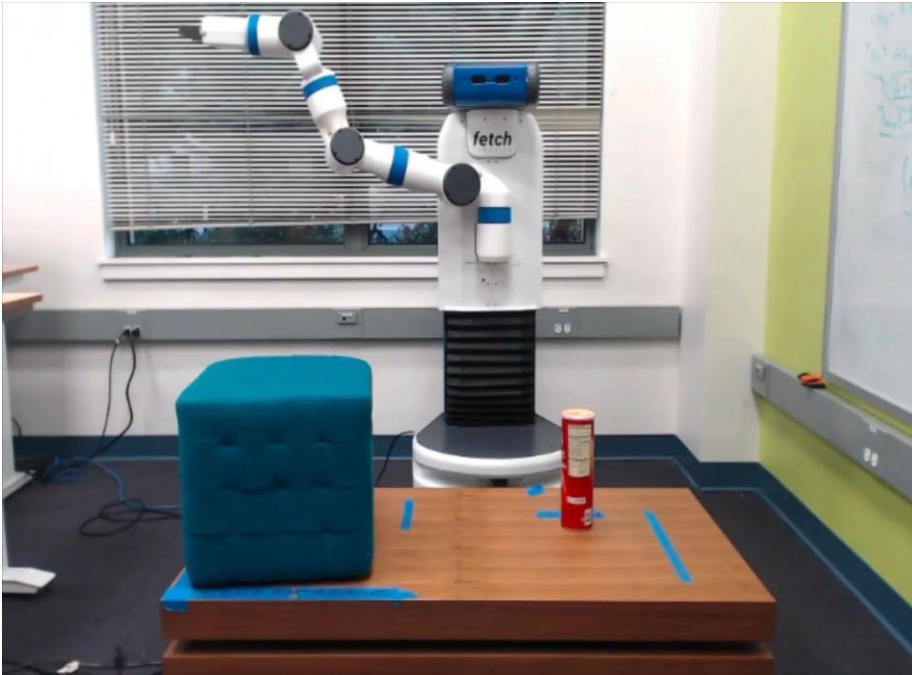
$$P(w \mid \mathcal{D}, \mathcal{C}) \propto P(w)P(\mathcal{D} \mid w)P(\mathcal{C} \mid w)$$

$$= P(w) \prod_{i=1}^L P(\xi_i \mid w) \prod_{i=1}^N P(q^{(i)} \mid w, \xi_A^{(i)}, \xi_B^{(i)})$$

$$\propto P(w) \prod_{i=1}^L \exp f_w(\xi_i) \prod_{i=1}^N \frac{\exp f_w(\xi_{q^{(i)}}^{(i)})}{\exp f_w(\xi_{q^{(i)}}^{(i)}) + \exp f_w(\xi_{\neg q^{(i)}}^{(i)})}$$

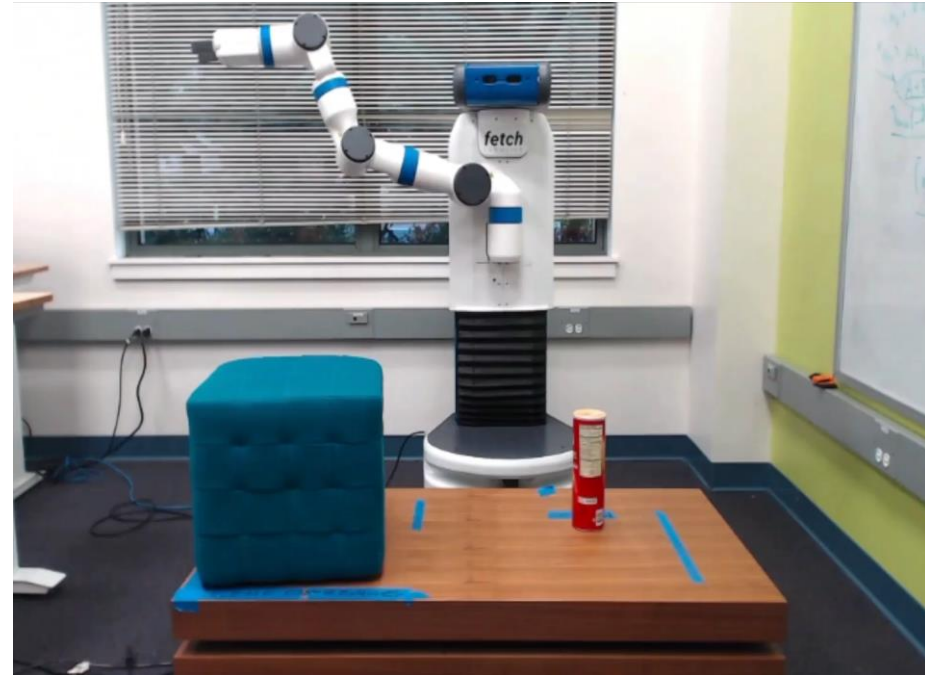
Benefit of comparisons

Bayesian IRL



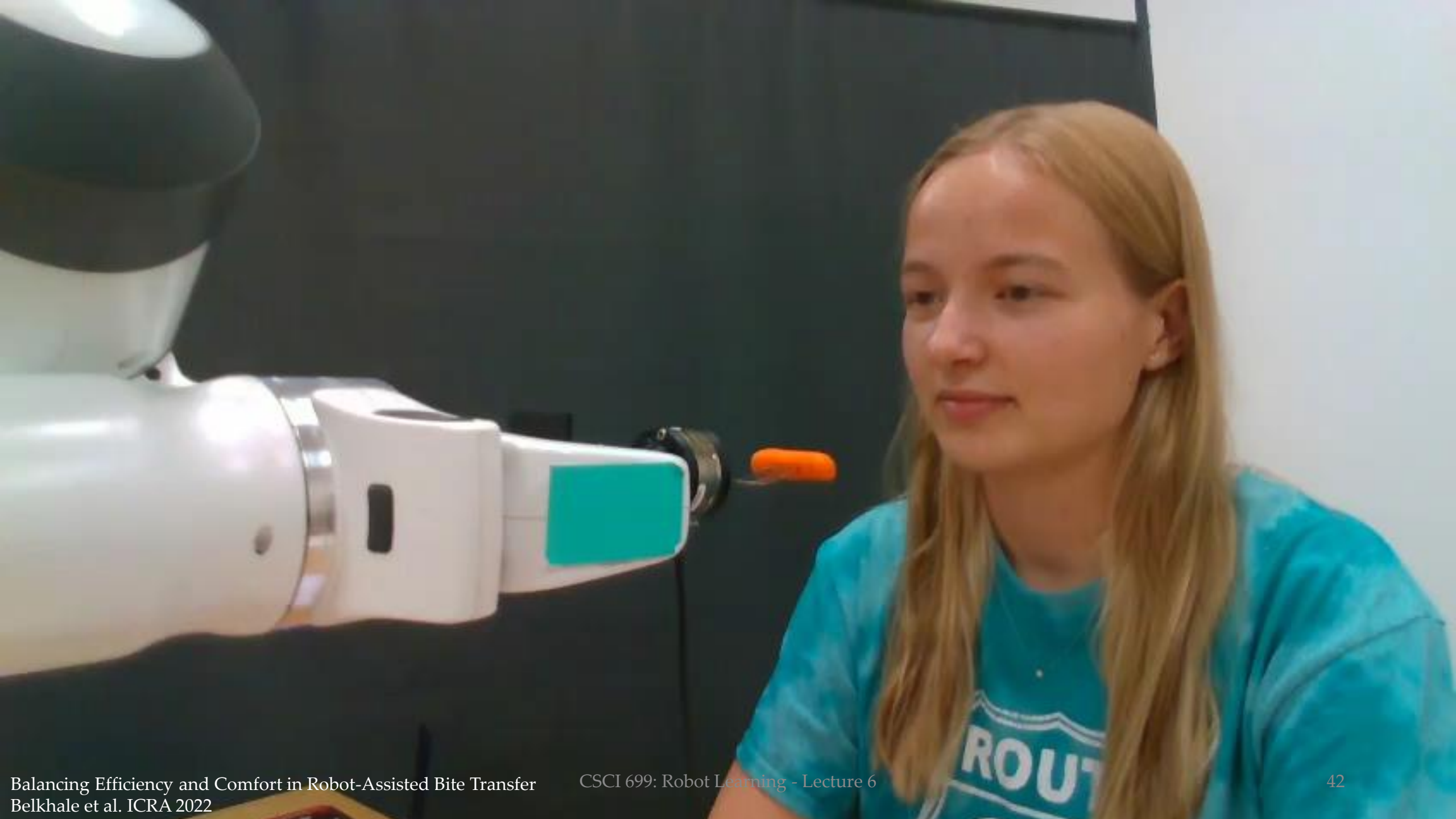
5 demonstrations

Ours



1 demonstration + 15 comparisons

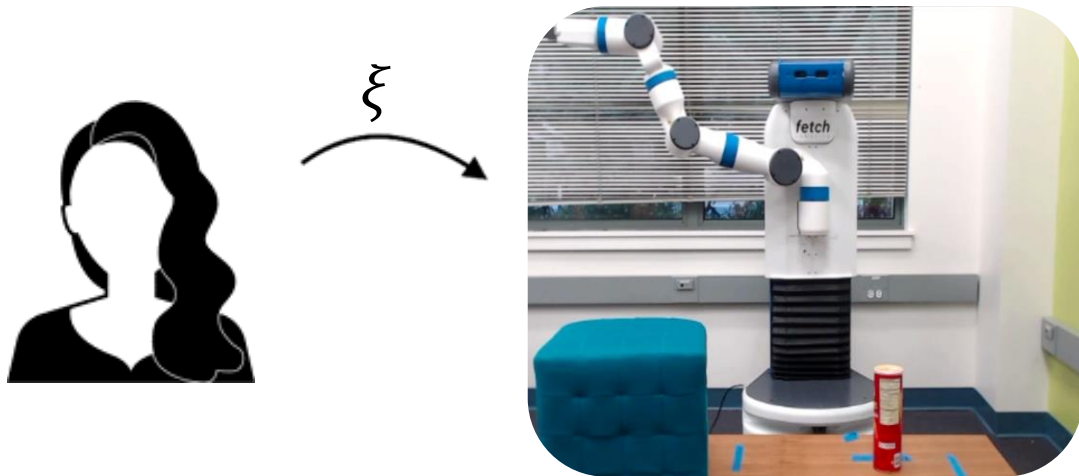




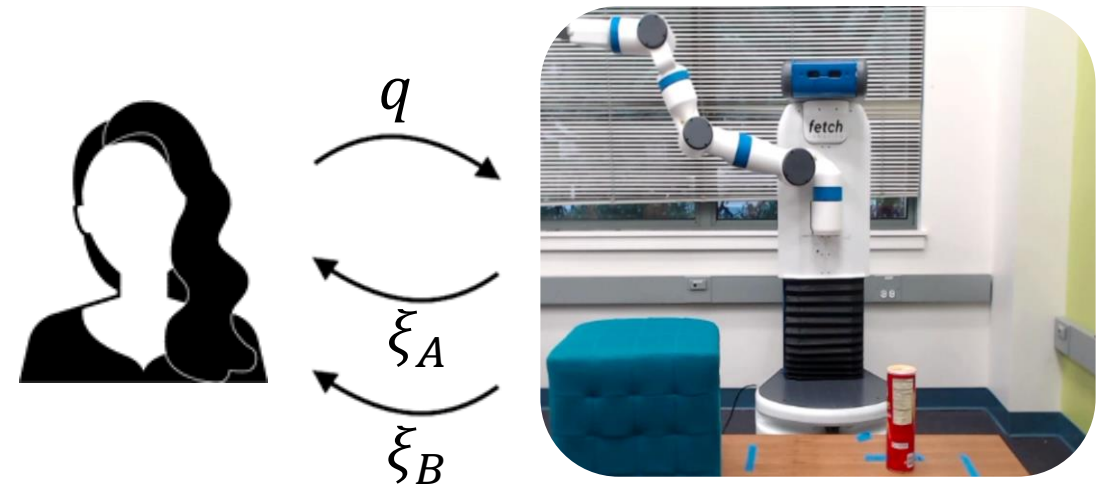


Choosing Queries

Demonstrations



Comparisons



How do we quantify information?

Surprise

$$\begin{array}{ll} 95\% \rightarrow X = \text{Heads} & \longrightarrow \text{ Surprise: } \log_2 \frac{1}{0.95} \cong 0.074 \\ 5\% \rightarrow X = \text{Tails} & \longrightarrow \text{ Surprise: } \log_2 \frac{1}{0.05} \cong 4.322 \end{array}$$

Entropy (a measure of uncertainty)

$$95\% \rightarrow X = \text{Heads} \quad \longrightarrow \quad \text{Surprise: } \log_2 \frac{1}{0.95} \cong 0.074$$

$$5\% \rightarrow X = \text{Tails} \quad \longrightarrow \quad \text{Surprise: } \log_2 \frac{1}{0.05} \cong 4.322$$

Entropy is the expected surprise.

$$\text{Entropy: } H(X) = 0.95 \times \log_2 \frac{1}{0.95} + 0.05 \times \log_2 \frac{1}{0.05} \cong 0.286$$

Another example

50% $\rightarrow X = \text{Heads}$

50% $\rightarrow X = \text{Tails}$

Another example

50% $\rightarrow X = \text{Heads}$ \longrightarrow Surprise: $\log_2 \frac{1}{0.50} = 1$

50% $\rightarrow X = \text{Tails}$ \longrightarrow Surprise: $\log_2 \frac{1}{0.50} = 1$

Another example

50% $\rightarrow X = \text{Heads}$ \longrightarrow Surprise: $\log_2 \frac{1}{0.50} = 1$

50% $\rightarrow X = \text{Tails}$ \longrightarrow Surprise: $\log_2 \frac{1}{0.50} = 1$

$$\text{Entropy: } H(X) = 0.50 \times \log_2 \frac{1}{0.50} + 0.50 \times \log_2 \frac{1}{0.50} \cong 1$$

Mutual information

Uncertainty

$$H(X) = 1$$



Alice



Bob

50% $\rightarrow X = \text{Heads}$

50% $\rightarrow X = \text{Tails}$

Mutual information

Uncertainty

$$H(X | X) = 0$$

$$0 \times \log \frac{1}{0} + 1 \times \log \frac{1}{1} = 0$$

This is 0 in
information theory.



Alice

What is X ?



Bob

Tails!

50% $\rightarrow X = \text{Heads}$

50% $\rightarrow X = \text{Tails}$

Mutual information

Uncertainty

$$H(X | X) = 0$$



Alice

What is X ?



Bob

Tails!

50% $\rightarrow X = \text{Heads}$

50% $\rightarrow X = \text{Tails}$

$$\begin{aligned} \text{Mutual Information} = \text{Reduction in Entropy: } I(X; X) &= H(X) - H(X | X) \\ &= 1 - 0 = 1 \end{aligned}$$

Mutual information

Uncertainty

$$H(X) = 1$$



Alice

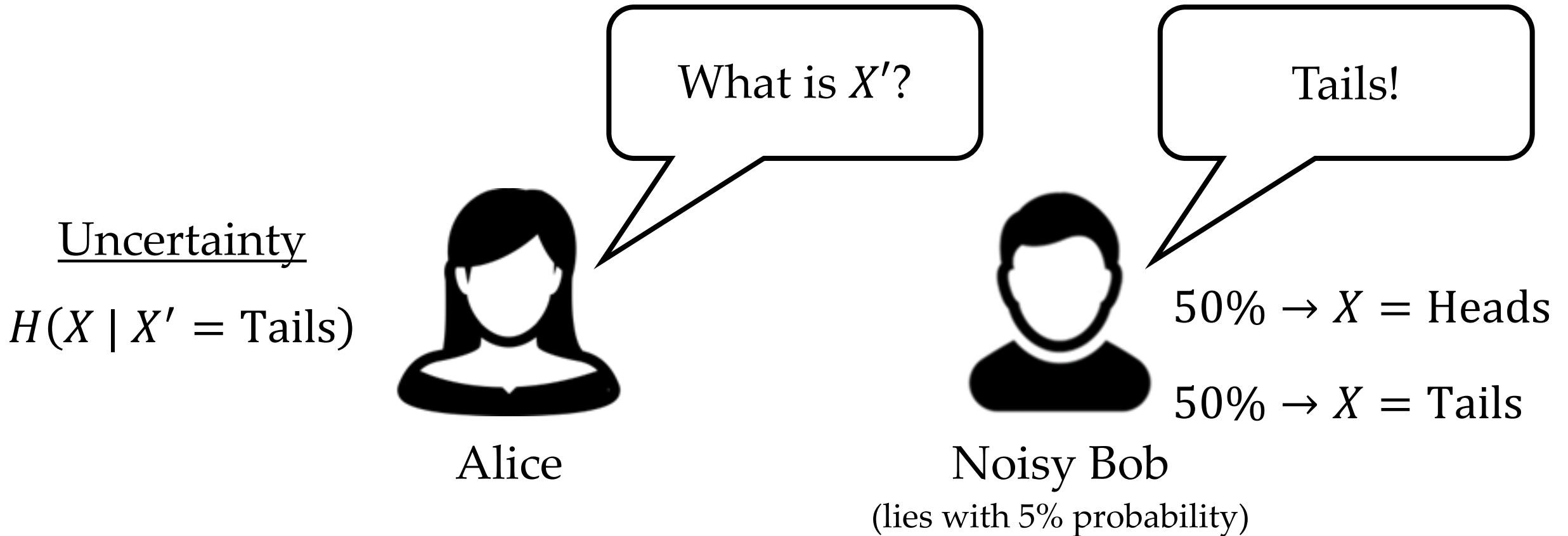


50% $\rightarrow X = \text{Heads}$

50% $\rightarrow X = \text{Tails}$

Noisy Bob
(lies with 5% probability)

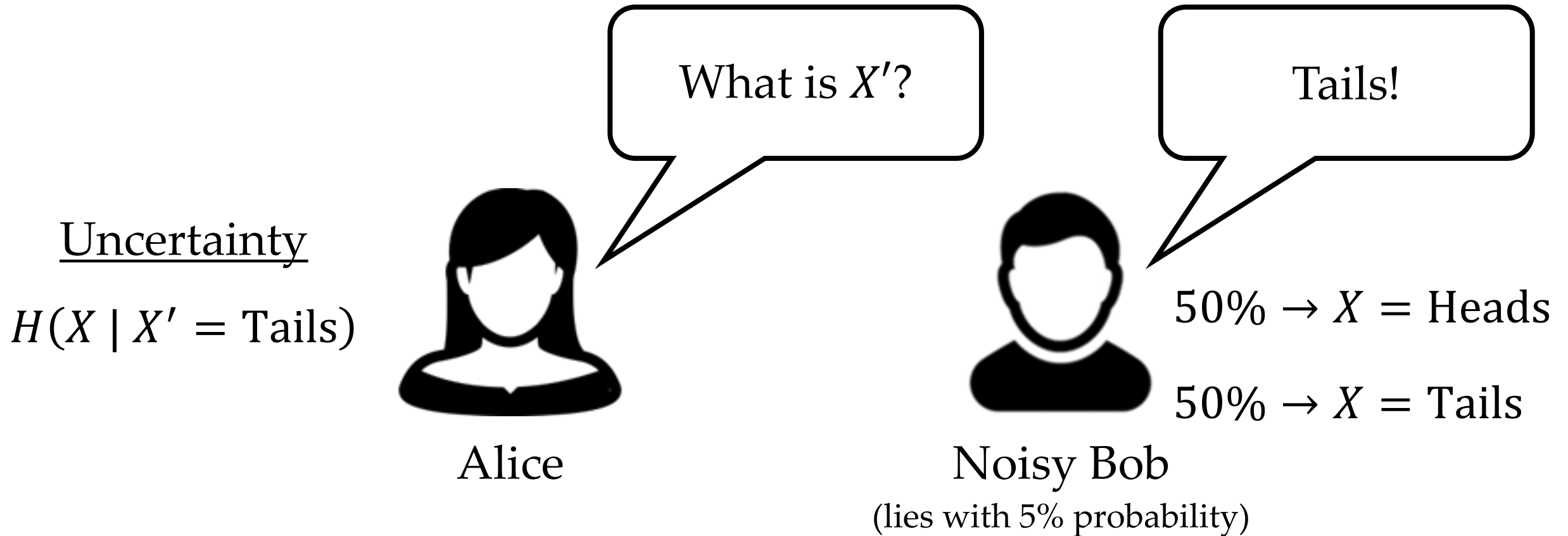
Mutual information



$$P(X = \text{Tails} | X' = \text{Tails}) \propto P(X' = \text{Tails} | X = \text{Tails})P(X = \text{Tails})$$

$$P(X = \text{Heads} | X' = \text{Tails}) \propto P(X' = \text{Tails} | X = \text{Heads})P(X = \text{Heads})$$

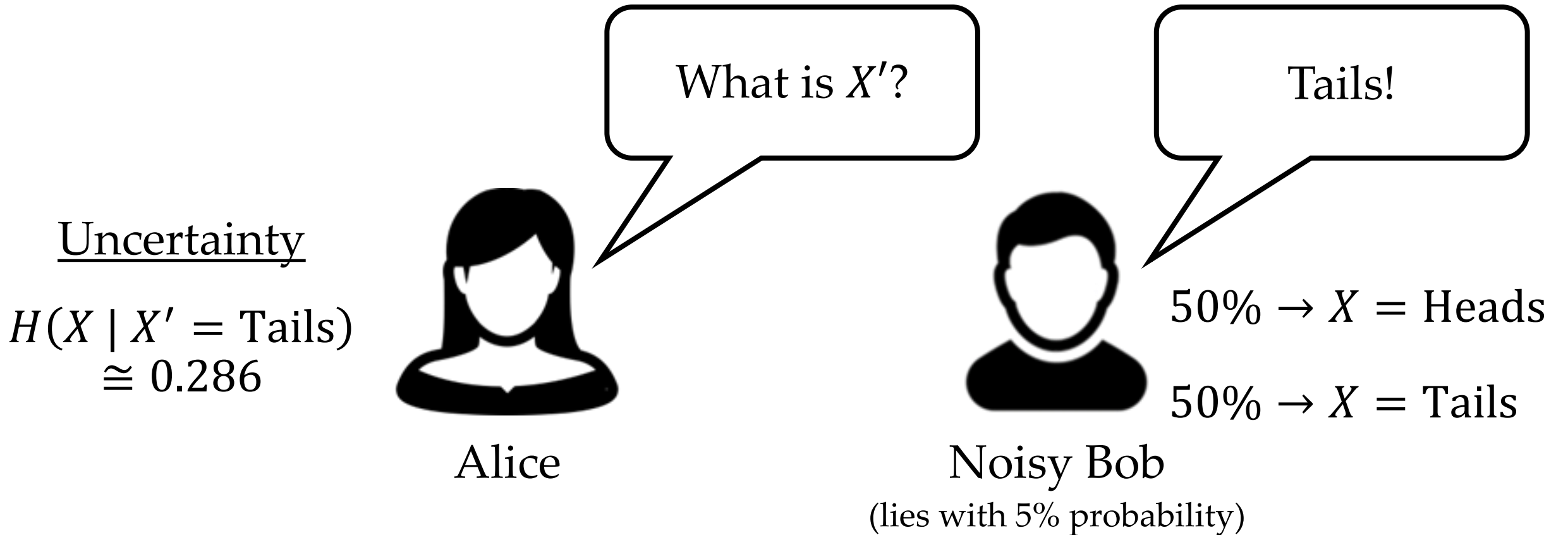
Mutual information



$$P(X = \text{Tails} | X' = \text{Tails}) = 0.95$$

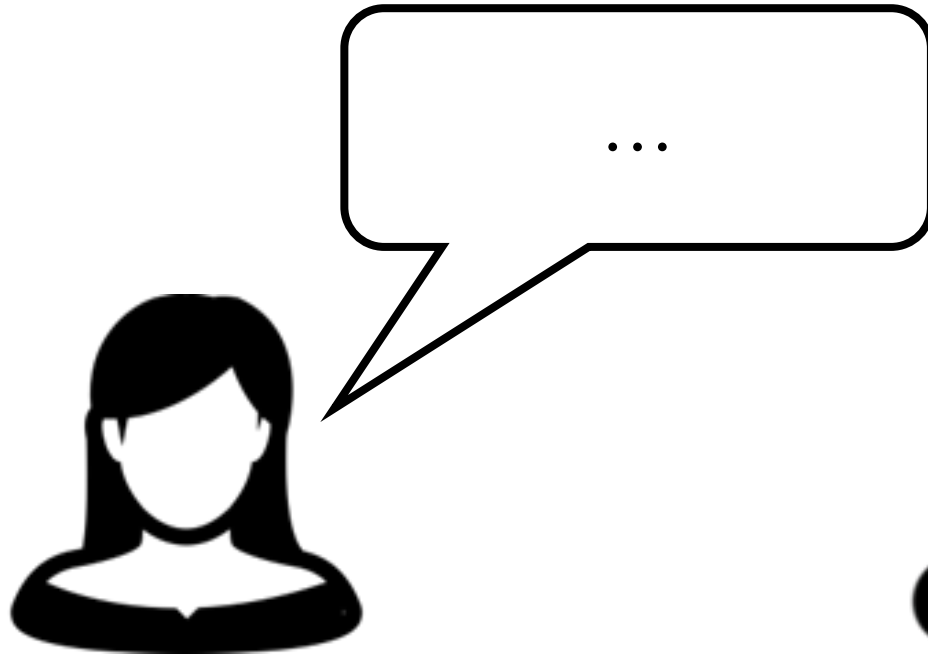
$$P(X = \text{Heads} | X' = \text{Tails}) = 0.05$$

Mutual information



Mutual Information = Reduction in Entropy: $I(X; X') = H(X) - H(X | X') \cong 1 - 0.286 = 0.714$

Mutual information: what do you ask?



Alice

Noisy Bob

(tells the truth for X_1 and X_2 ,
lies with 5% probability for X_3)

20% $\rightarrow X = (H, H, H)$

20% $\rightarrow X = (H, H, T)$

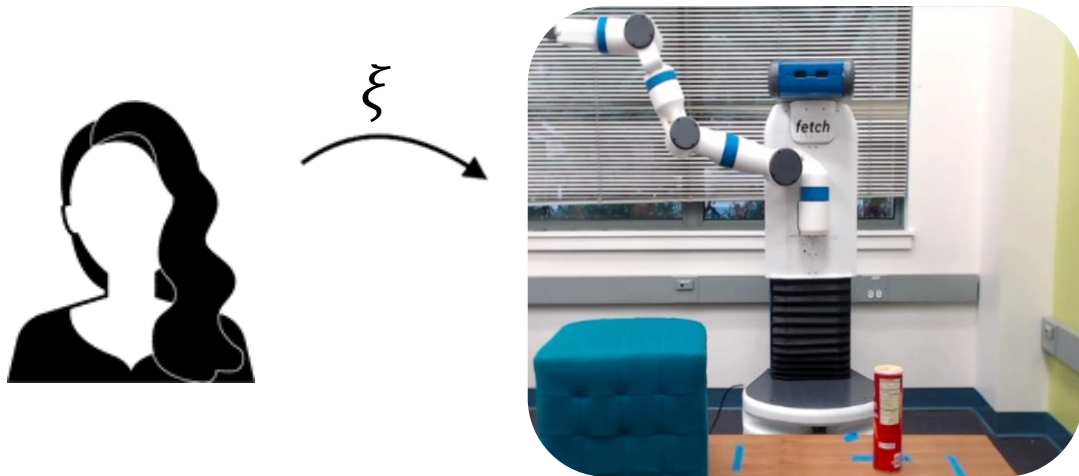
20% $\rightarrow X = (H, T, H)$

20% $\rightarrow X = (H, T, T)$

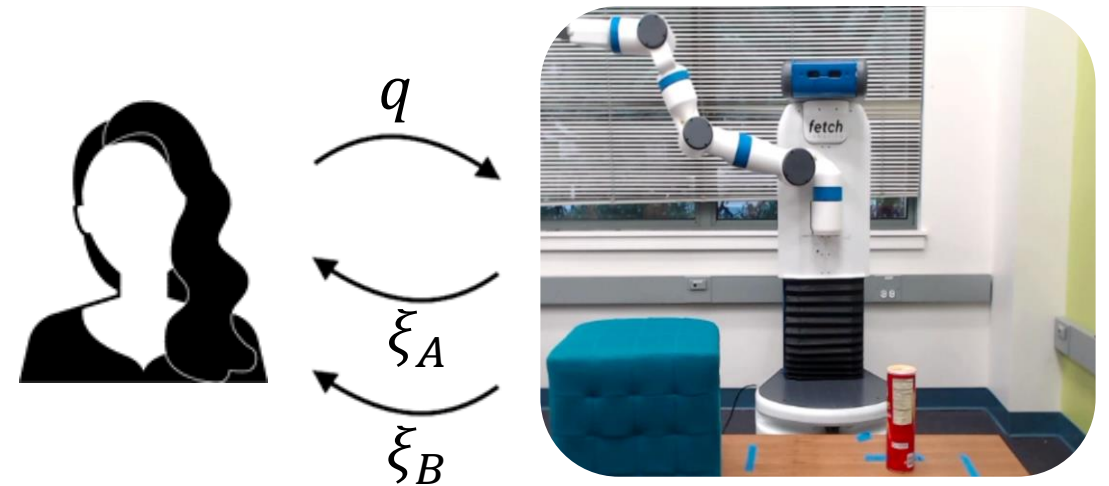
20% $\rightarrow X = (T, T, T)$

Choosing queries

Demonstrations



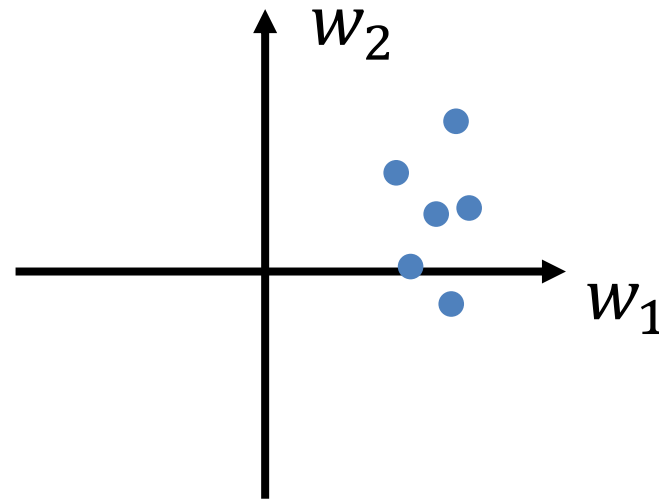
Comparisons



The robot can query the user with the query that will give the **most information**.

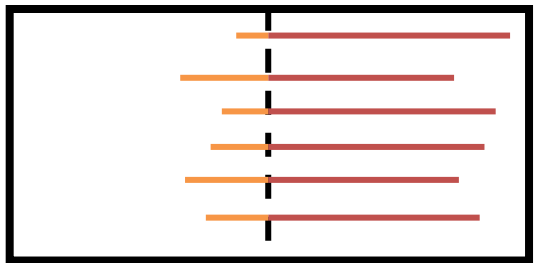
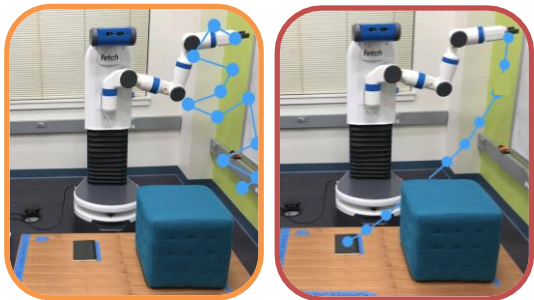
Maximum volume removal

Posterior $P(w | \mathcal{C})$

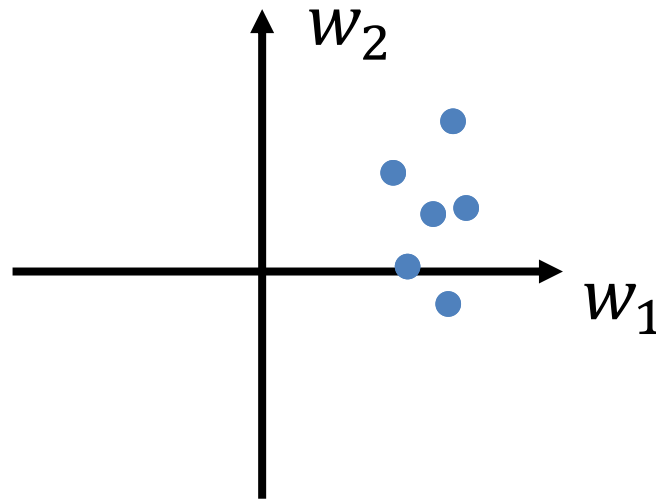


Maximum volume removal

Posterior $P(w | \mathcal{C})$

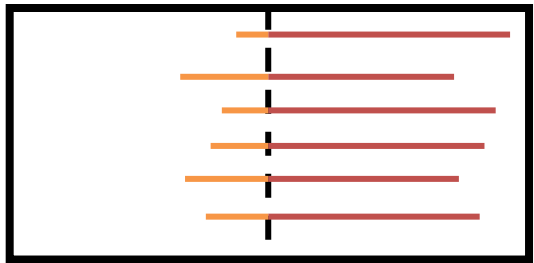
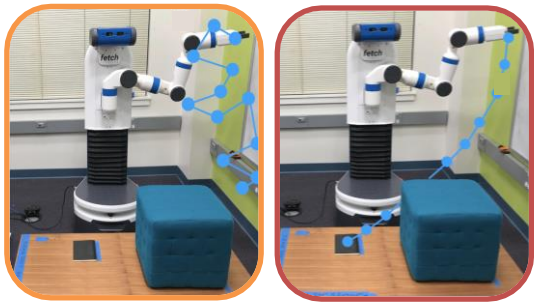


User Choice

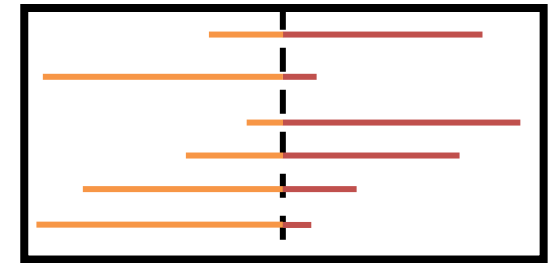
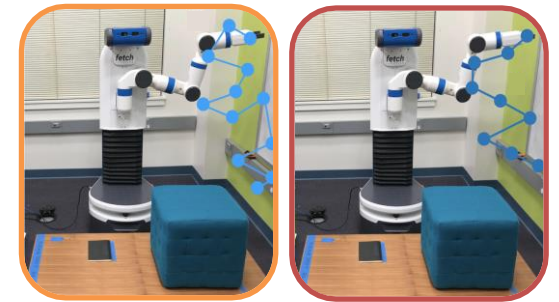
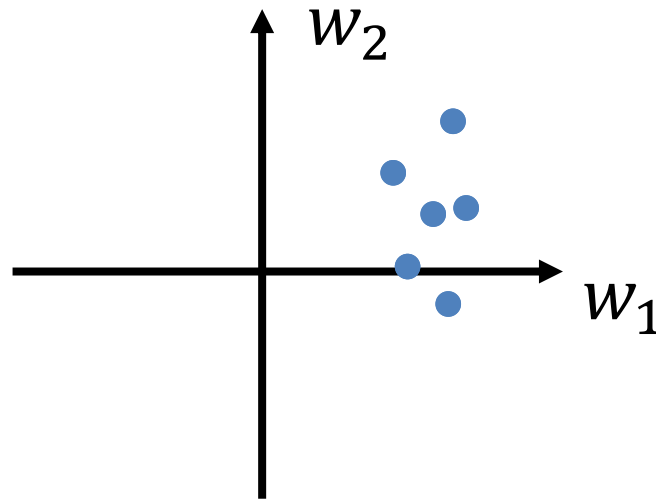


Maximum volume removal

Posterior $P(w | \mathcal{C})$



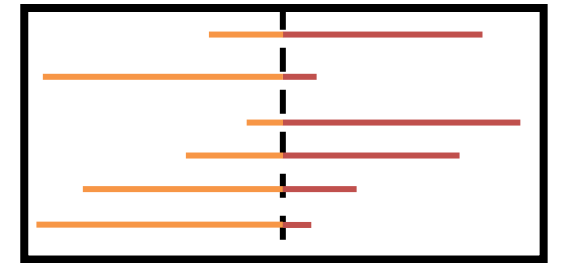
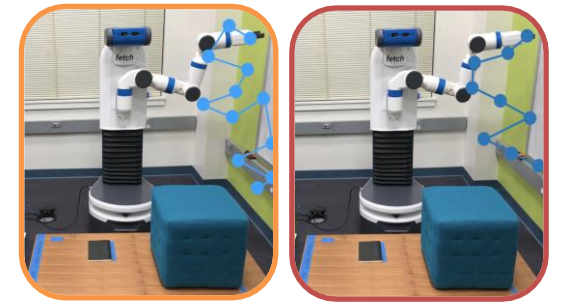
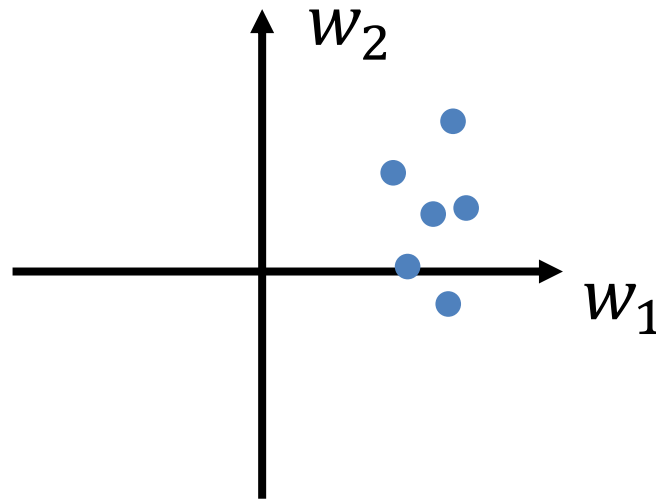
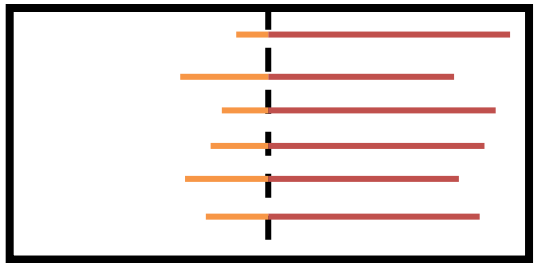
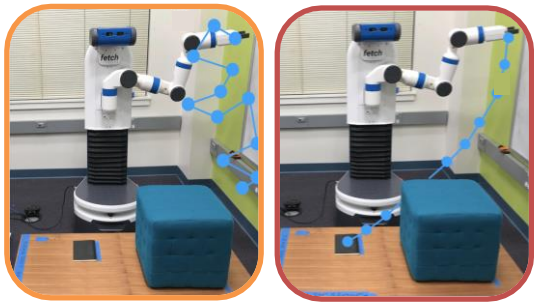
User Choice



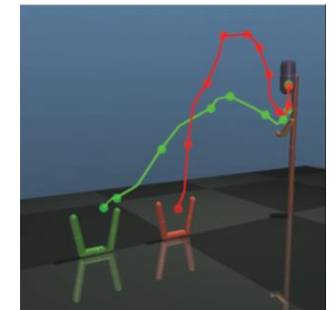
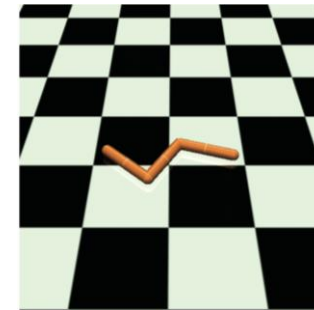
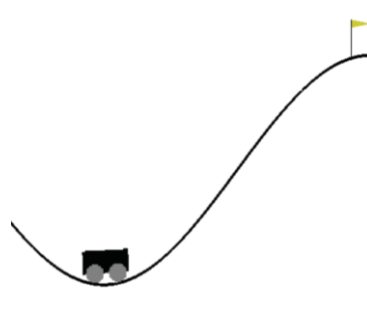
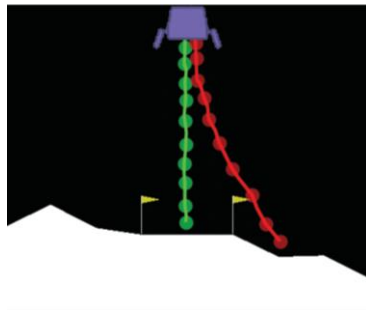
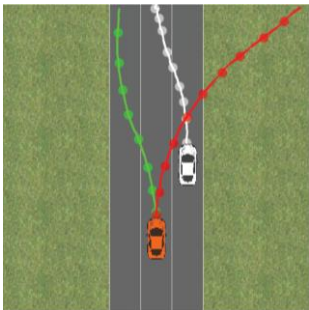
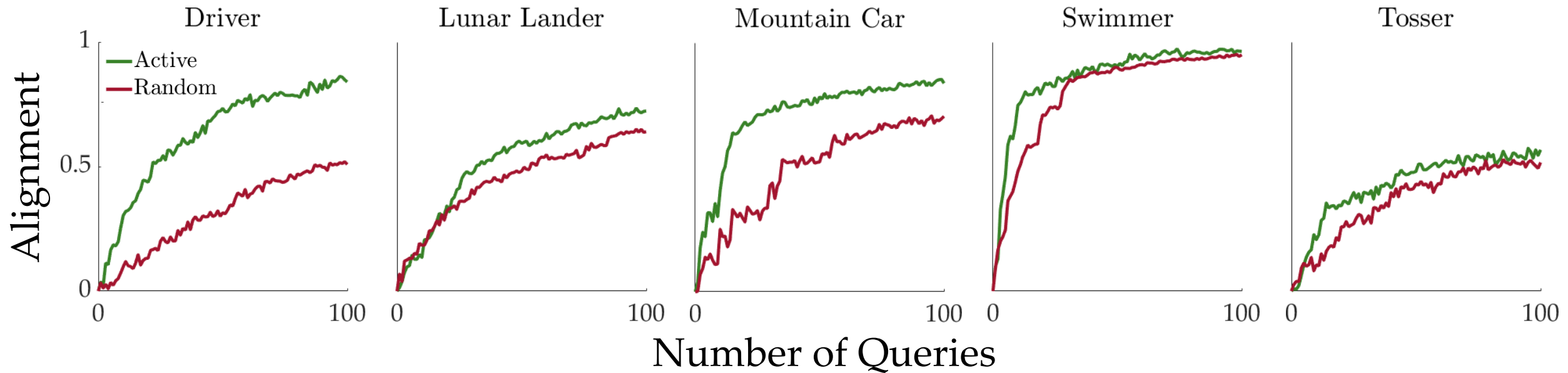
User Choice

Maximum volume removal

Posterior $P(w | \mathcal{C})$

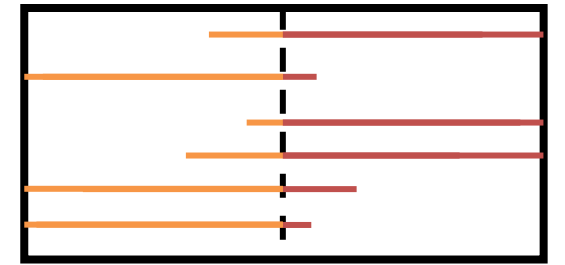
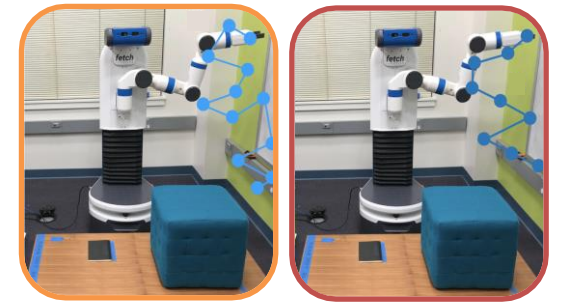
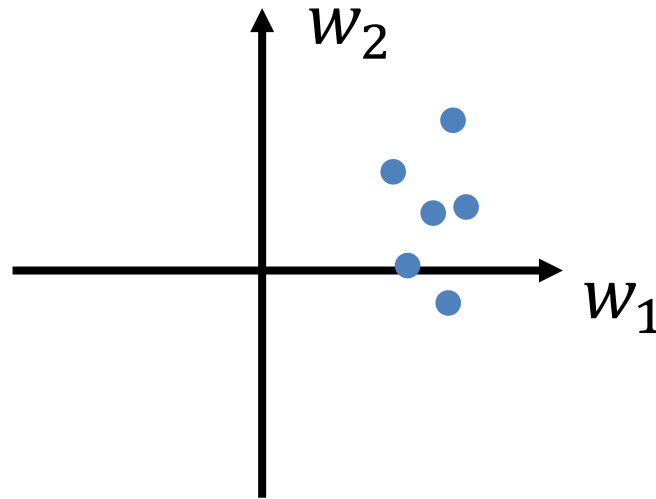
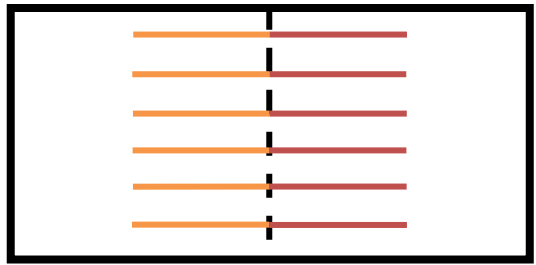
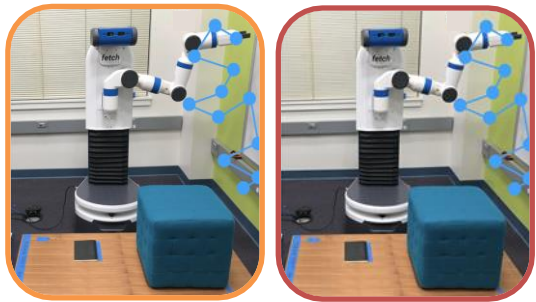


Active vs. random querying

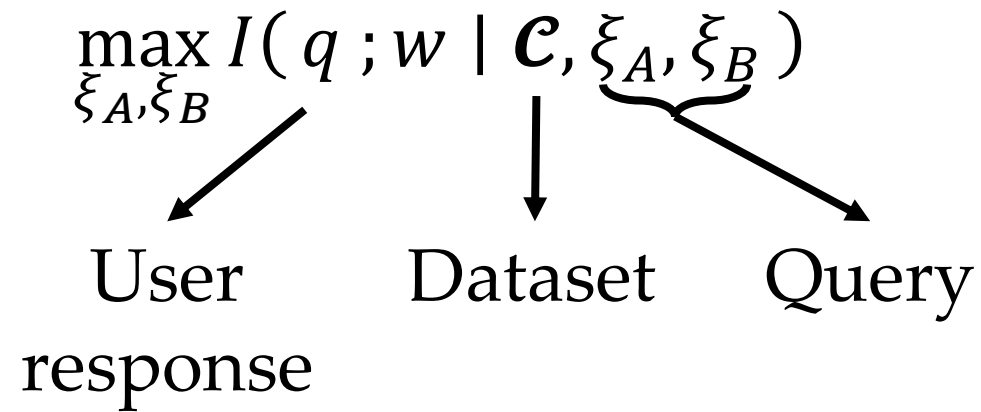


Maximum volume removal

Posterior $P(w | \mathcal{C})$



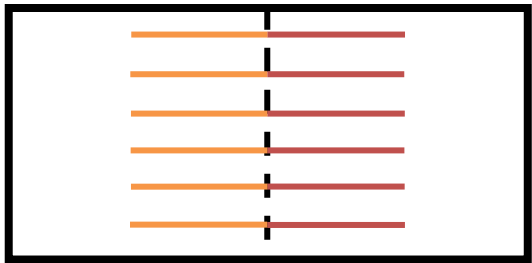
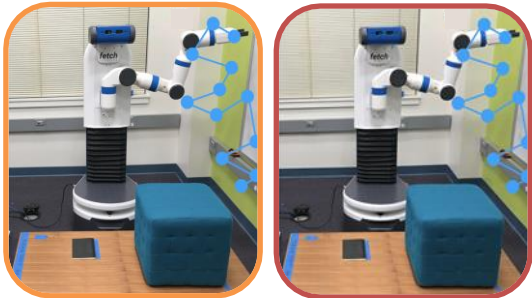
Mutual information maximization



Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

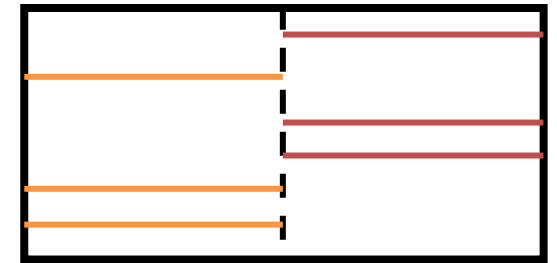
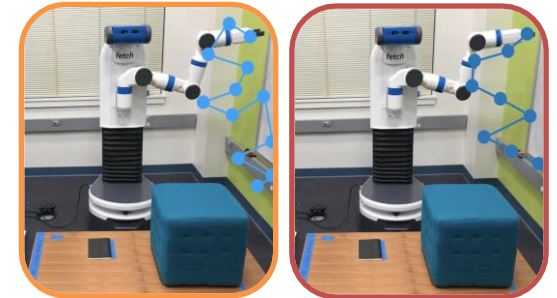
$$\max_{\xi_A, \xi_B} \underbrace{H(q \mid \mathcal{C}, \xi_A, \xi_B)}_{\text{Model Uncertainty}} - \underbrace{H(q \mid \mathcal{C}, \xi_A, \xi_B, w)}_{\text{User Uncertainty}}$$



User Choice

Model
Uncertainty

User
Uncertainty



User Choice

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} -\mathbb{E}_{q \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \mathcal{C}, \xi_A, \xi_B)] + \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \mathcal{C}, \xi_A, \xi_B, w)]$$

No w here!

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} -\mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \mathcal{C}, \xi_A, \xi_B)] + \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \mathcal{C}, \xi_A, \xi_B, w)]$$

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \mathcal{C}, \xi_A, \xi_B, w) - \log P(q \mid \mathcal{C}, \xi_A, \xi_B)]$$

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log P(q \mid \mathcal{C}, \xi_A, \xi_B)]$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log \int \underbrace{P(q, w' \mid \mathcal{C}, \xi_A, \xi_B)}_{dw'}]$$

$$P(w' \mid \mathcal{C}, \xi_A, \xi_B) P(q \mid \mathcal{C}, \xi_A, \xi_B, w')$$

$$= P(w' \mid \mathcal{C}) P(q \mid \xi_A, \xi_B, w')$$

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log P(q \mid \mathcal{C}, \xi_A, \xi_B)]$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log \int P(w' \mid \mathcal{C}) P(q \mid \xi_A, \xi_B, w') dw']$$

This is an expectation over $w' \mid \mathcal{C}$

Take samples from $w' \mid \mathcal{C}$ to compute.

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log P(q \mid \mathcal{C}, \xi_A, \xi_B)]$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} \left[\log P(q \mid \xi_A, \xi_B, w) - \log \frac{1}{|\Omega|} \sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w') \right]$$

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log P(q \mid \mathcal{C}, \xi_A, \xi_B)]$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} \left[\log P(q \mid \xi_A, \xi_B, w) - \log \sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w') \right]$$

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w | \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q | \mathcal{C}, \xi_A, \xi_B) - H(q | \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w | \mathcal{C}, \xi_A, \xi_B} [\log P(q | \xi_A, \xi_B, w) - \log P(q | \mathcal{C}, \xi_A, \xi_B)]$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w | \mathcal{C}, \xi_A, \xi_B} \left[\log \frac{P(q | \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q | \xi_A, \xi_B, w')} \right]$$

$$\begin{aligned} P(q, w | \mathcal{C}, \xi_A, \xi_B) &= P(w | \mathcal{C}, \xi_A, \xi_B) P(q | \mathcal{C}, \xi_A, \xi_B, w) \\ &= P(w | \mathcal{C}) P(q | \xi_A, \xi_B, w) \end{aligned}$$

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log P(q \mid \mathcal{C}, \xi_A, \xi_B)]$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \xi_A, \xi_B} \left[\log \frac{P(q \mid \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w')} \right]$$

$$\max_{\xi_A, \xi_B} \frac{1}{|\Omega|} \sum_{w \in \Omega} \mathbb{E}_{q \mid \xi_A, \xi_B, w} \left[\log \frac{P(q \mid \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w')} \right]$$

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log P(q \mid \mathcal{C}, \xi_A, \xi_B)]$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \xi_A, \xi_B} \left[\log \frac{P(q \mid \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w')} \right]$$

$$\max_{\xi_A, \xi_B} \sum_{w \in \Omega} \mathbb{E}_{q \mid \xi_A, \xi_B, w} \left[\log \frac{P(q \mid \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w')} \right]$$

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

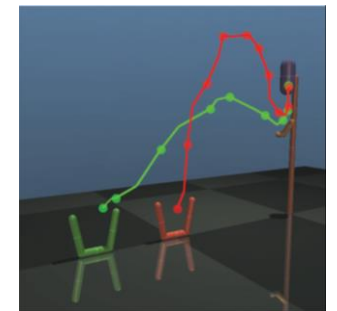
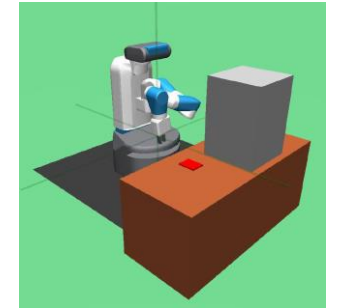
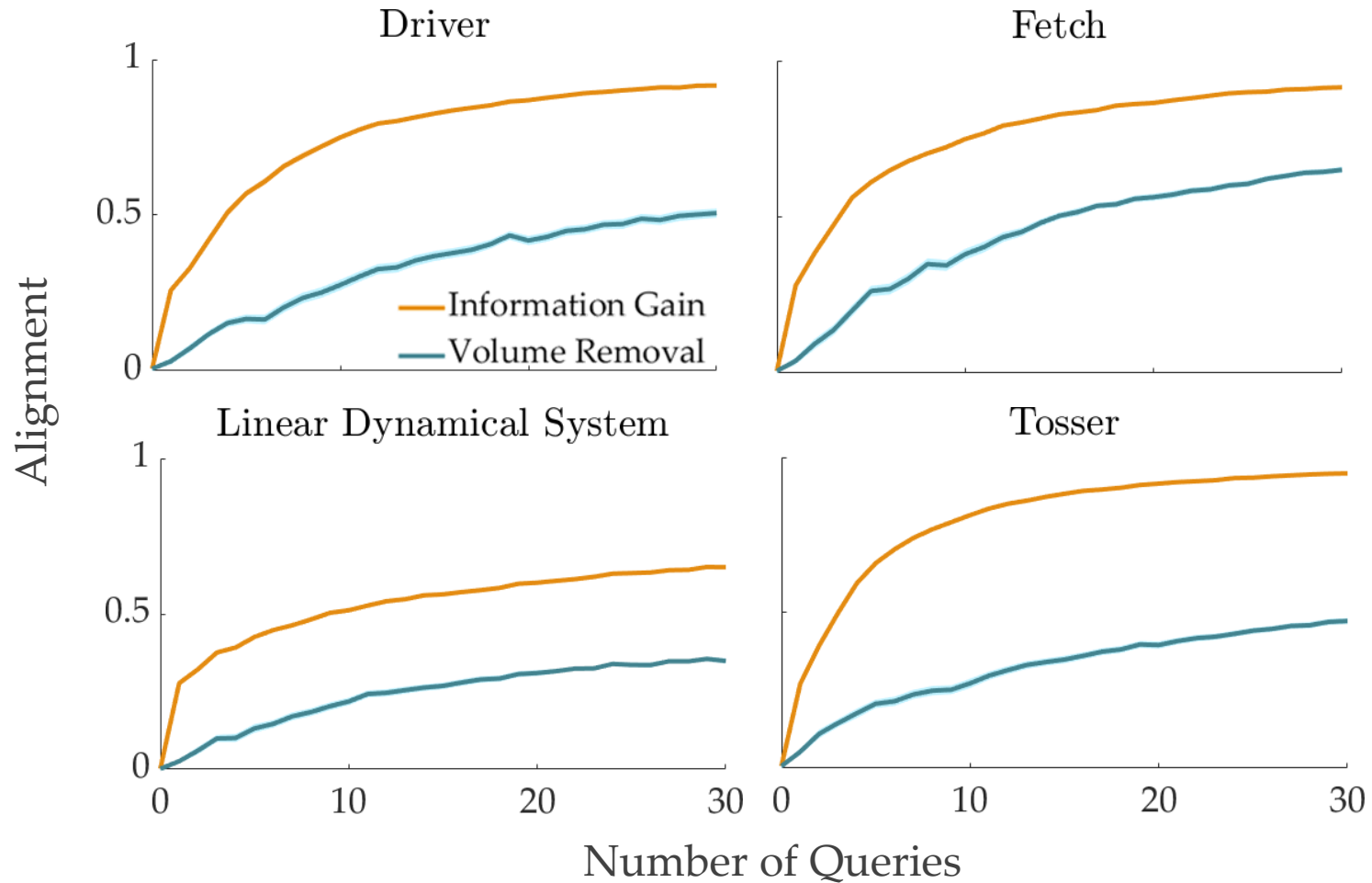
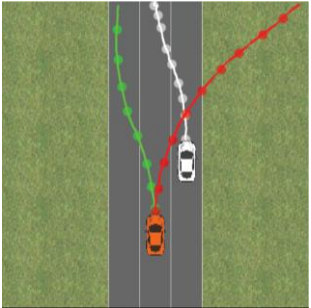
$$\max_{\xi_A, \xi_B} H(q \mid \mathcal{C}, \xi_A, \xi_B) - H(q \mid \mathcal{C}, \xi_A, \xi_B, w)$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \mathcal{C}, \xi_A, \xi_B} [\log P(q \mid \xi_A, \xi_B, w) - \log P(q \mid \mathcal{C}, \xi_A, \xi_B)]$$

$$\max_{\xi_A, \xi_B} \mathbb{E}_{q, w \mid \xi_A, \xi_B} \left[\log \frac{P(q \mid \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w')} \right]$$

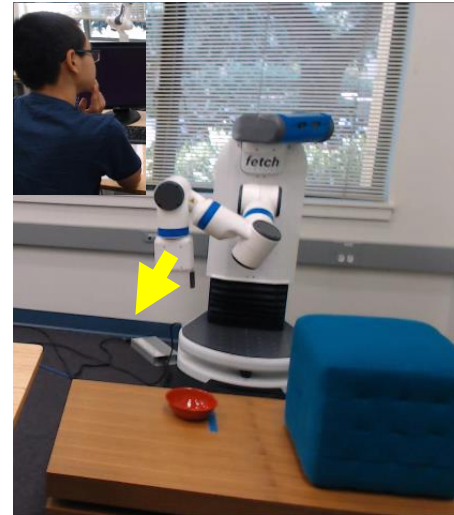
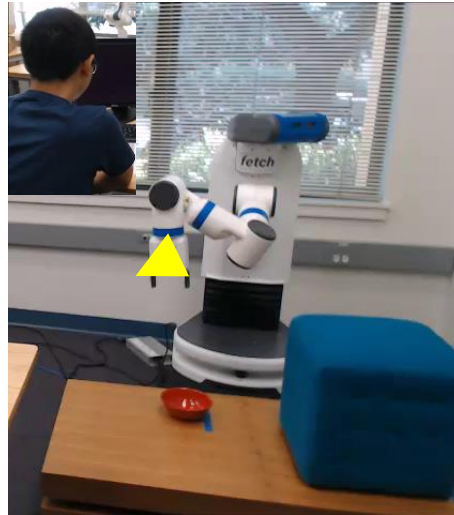
$$\max_{\xi_A, \xi_B} \sum_{w \in \Omega} \sum_q P(q \mid \xi_A, \xi_B, w) \left[\log \frac{P(q \mid \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w')} \right]$$

Mutual information maximization



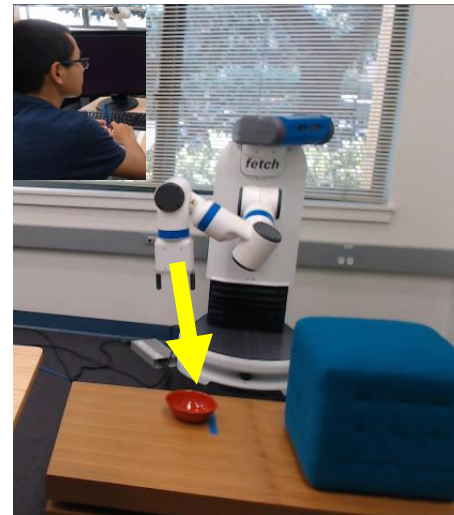
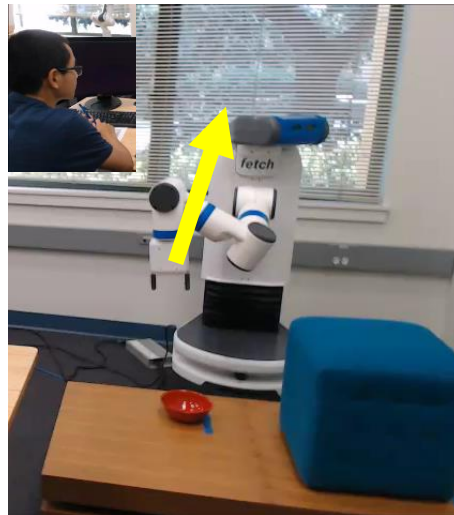
Volume Removal

Similar Trajectories

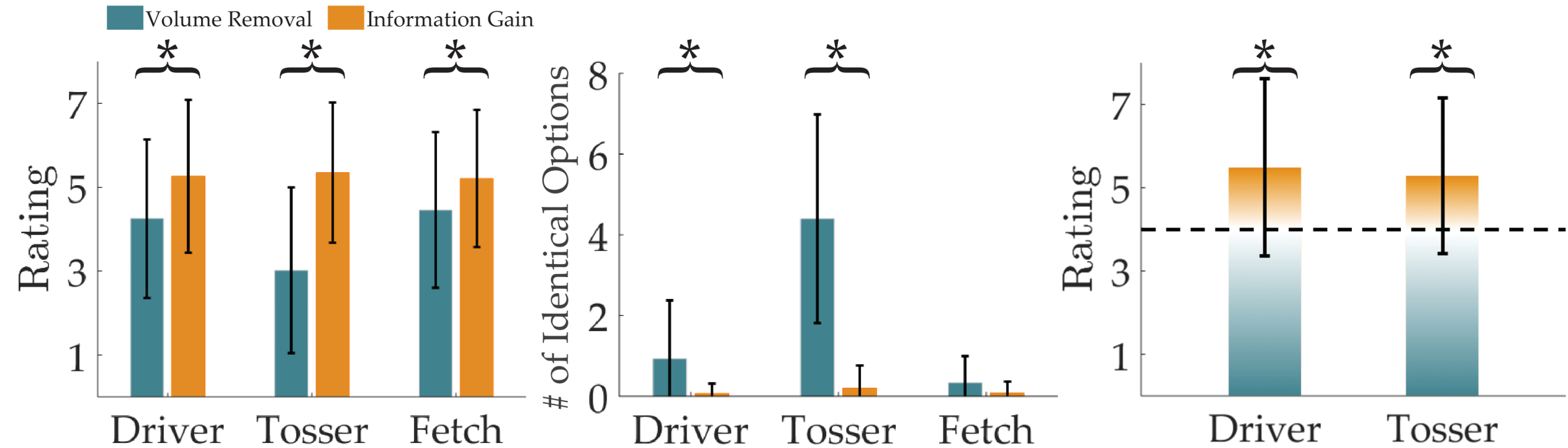


Information Gain

More Distinguishable Query



Mutual information maximization



Other types of human feedback

| Feedback | Constraint | Probabilistic |
|-------------------|--|---|
| Comparisons | $r(\xi_1) \geq r(\xi_2)$ | $\mathbb{P}(\xi_1 r, \mathcal{C}) = \frac{\exp(\beta \cdot r(\xi_1))}{\exp(\beta \cdot r(\xi_1)) + \exp(\beta \cdot r(\xi_2))}$ |
| Demonstrations | $r(\xi_D) \geq r(\xi) \quad \forall \xi \in \Xi$ | $\mathbb{P}(\xi_D r, \Xi) = \frac{\exp(\beta \cdot r(\xi_D))}{\sum_{\xi \in \Xi} \exp(\beta \cdot r(\xi))}$ |
| Corrections | $r(\xi_R + A^{-1}\Delta q) \geq r(\xi_R + A^{-1}\Delta q') \quad \forall \Delta q' \in Q - Q$ | $\mathbb{P}(\Delta q' r, Q - Q) = \frac{\exp(\beta \cdot r(\xi_R + A^{-1}\Delta q))}{\sum_{\Delta q \in Q - Q} \exp(\beta \cdot r(\xi_R + A^{-1}\Delta q))}$ |
| Improvement | $r(\xi_{\text{improved}}) \geq r(\xi_R)$ | $\mathbb{P}(\xi_{\text{improved}} r, \mathcal{C}) = \frac{\exp(\beta \cdot r(\xi_{\text{improved}}))}{\exp(\beta \cdot r(\xi_{\text{improved}})) + \exp(\beta \cdot r(\xi_R))}$ |
| Off | $r(\xi_R^{0:t} \xi^t \dots \xi^t) \geq r(\xi_R)$ | $\mathbb{P}(\text{off} r, \mathcal{C}) = \frac{\exp(\beta \cdot r(\xi_R^{0:t} \xi^t \dots \xi^t))}{\exp(\beta \cdot r(\xi_R^{0:t} \xi^t \dots \xi^t)) + \exp(\beta \cdot r(\xi_R))}$ |
| Language | $\mathbb{E}_{\xi \sim \text{Unif}(G(\lambda^*))} [r(\xi)] \geq \mathbb{E}_{\xi \sim \text{Unif}(G(\lambda))} [r(\xi)] \quad \forall \lambda \in \Lambda$ | $\mathbb{P}(\lambda^* r, \Lambda) = \frac{\exp(\beta \cdot \mathbb{E}_{\xi \sim \text{Unif}(G(\lambda^*))} [r(\xi)])}{\sum_{\lambda \in \Lambda} \exp(\beta \cdot \mathbb{E}_{\xi \sim \text{Unif}(G(\lambda))} [r(\xi)])}$ |
| Proxy Rewards | $\mathbb{E}_{\tilde{\xi} \sim \pi(\tilde{\xi} \tilde{r})} [r(\tilde{\xi})] \geq \mathbb{E}_{\tilde{\xi} \sim \pi(\tilde{\xi} c)} [r(\tilde{\xi})] \quad \forall c \in \tilde{\mathcal{R}}$ | $\mathbb{P}(\tilde{r} r, \tilde{\mathcal{R}}) = \frac{\exp(\beta \cdot \mathbb{E}_{\tilde{\xi} \sim \pi(\tilde{\xi} \tilde{r})} [r(\tilde{\xi})])}{\sum_{c \in \tilde{\mathcal{R}}} \exp(\beta \cdot \mathbb{E}_{\tilde{\xi} \sim \pi(\tilde{\xi} c)} [r(\tilde{\xi})])}$ |
| Reward/Punish | $r(\xi_R) \geq r(\xi_{\text{expected}})$ | $\mathbb{P}(+1 r, \mathcal{C}) = \frac{\exp(\beta \cdot r(\xi_R))}{\exp(\beta \cdot r(\xi_R)) + \exp(\beta \cdot r(\xi_{\text{expected}}))}$ |
| Initial state | $\mathbb{E}_{\xi \sim \psi(s^*)} [r(s^*)] \geq \mathbb{E}_{\xi \sim \psi(s)} [r(s)] \quad \forall s \in \mathcal{S}$ | $\mathbb{P}(s^* r, \mathcal{S}) = \frac{\exp(\beta \cdot \mathbb{E}_{\xi \sim \psi(s^*)} [r(\xi)])}{\sum_{s \in \mathcal{S}} \exp(\beta \cdot \mathbb{E}_{\xi \sim \psi(s)} [r(\xi)])}$ |
| Meta-choice | $\mathbb{E}_{\xi \sim \psi(c_i)} [r(\xi)] \geq \mathbb{E}_{\xi \sim \psi(c_j)} [r(\xi)] \quad \forall j \in [n]$ | $\mathbb{P}(C_i r, \mathcal{C}_0) = \frac{\exp(\beta_0 \cdot \mathbb{E}_{\xi \sim \psi_0(c_i)} [r(\xi)])}{\sum_{j \in [n]} \exp(\beta_0 \cdot \mathbb{E}_{\xi \sim \psi_0(c_j)} [r(\xi)])}$ |
| Credit assignment | $r(\xi^*) \geq r(\xi) \quad \forall \xi \in \mathcal{C}$ | $\mathbb{P}(\xi^* r, \mathcal{C}) = \frac{\exp(\beta \cdot r(\xi^*))}{\sum_{\xi \in \mathcal{C}} \exp(\beta \cdot r(\xi))}$ |

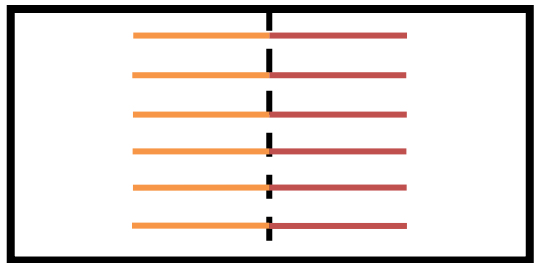
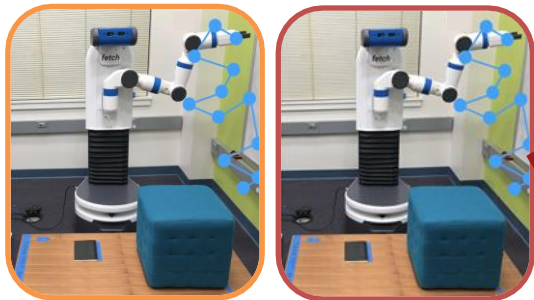
Today...

- Learning from human feedback
 - Suboptimal demonstrations
 - Pairwise comparisons
 - Reinforcement learning from human feedback (RLHF)

Mutual information maximization

$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

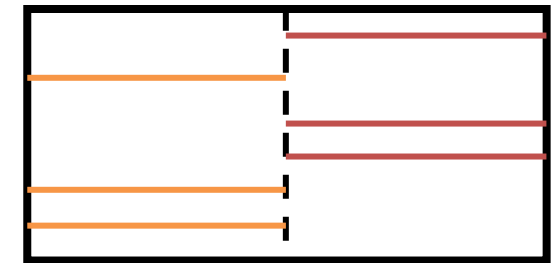
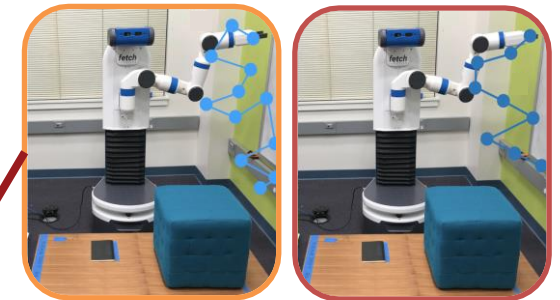
$$\max_{\xi_A, \xi_B} \underbrace{H(q \mid \mathcal{C}, \xi_A, \xi_B)}_{\text{Model Uncertainty}} - \underbrace{H(q \mid \mathcal{C}, \xi_A, \xi_B, w)}_{\text{User Uncertainty}}$$



User Choice

Model
Uncertainty

User
Uncertainty



User Choice

Where do these trajectories
come from in the first place?

Incorporating comparisons

$$\operatorname{argmax}_w P(w \mid \mathcal{D}, \mathcal{C})$$

$$P(w \mid \mathcal{D}, \mathcal{C}) \propto P(w)P(\mathcal{D} \mid w)P(\mathcal{C} \mid w)$$

$$= P(w) \prod_{i=1}^L P(\xi_i \mid w) \prod_{i=1}^N P(q^{(i)} \mid w, \xi_A^{(i)}, \xi_B^{(i)})$$

$$\propto P(w) \prod_{i=1}^L \exp f_w(\xi_i) \prod_{i=1}^N \frac{\exp f_w(\xi_{q^{(i)}}^{(i)})}{\exp f_w(\xi_{q^{(i)}}^{(i)}) + \exp f_w(\xi_{\neg q^{(i)}}^{(i)})}$$

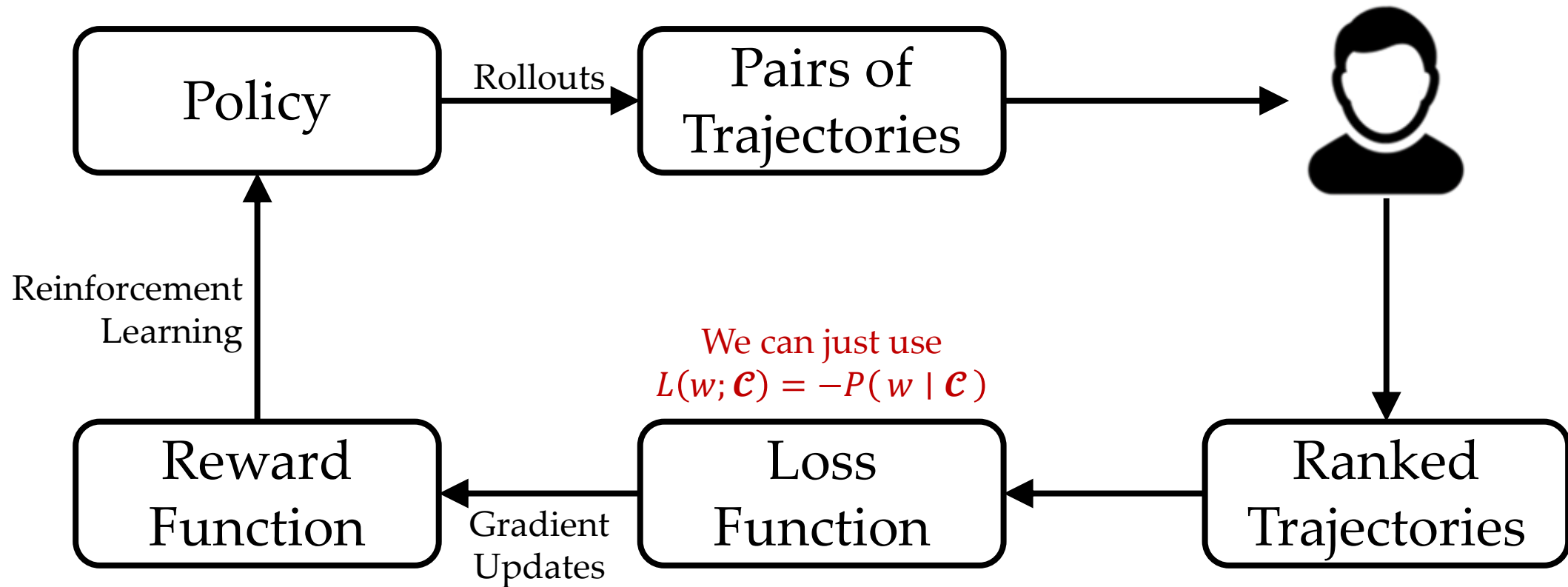
How do we solve this optimization problem?

RLHF

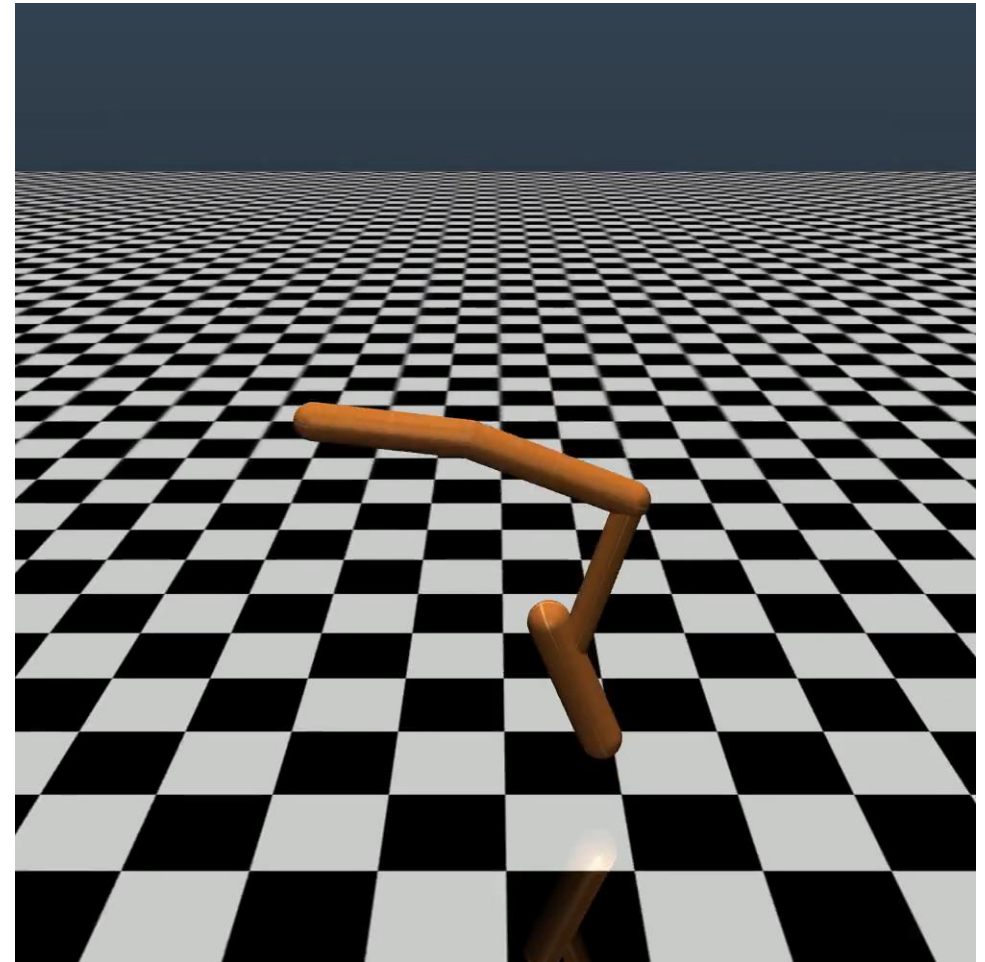
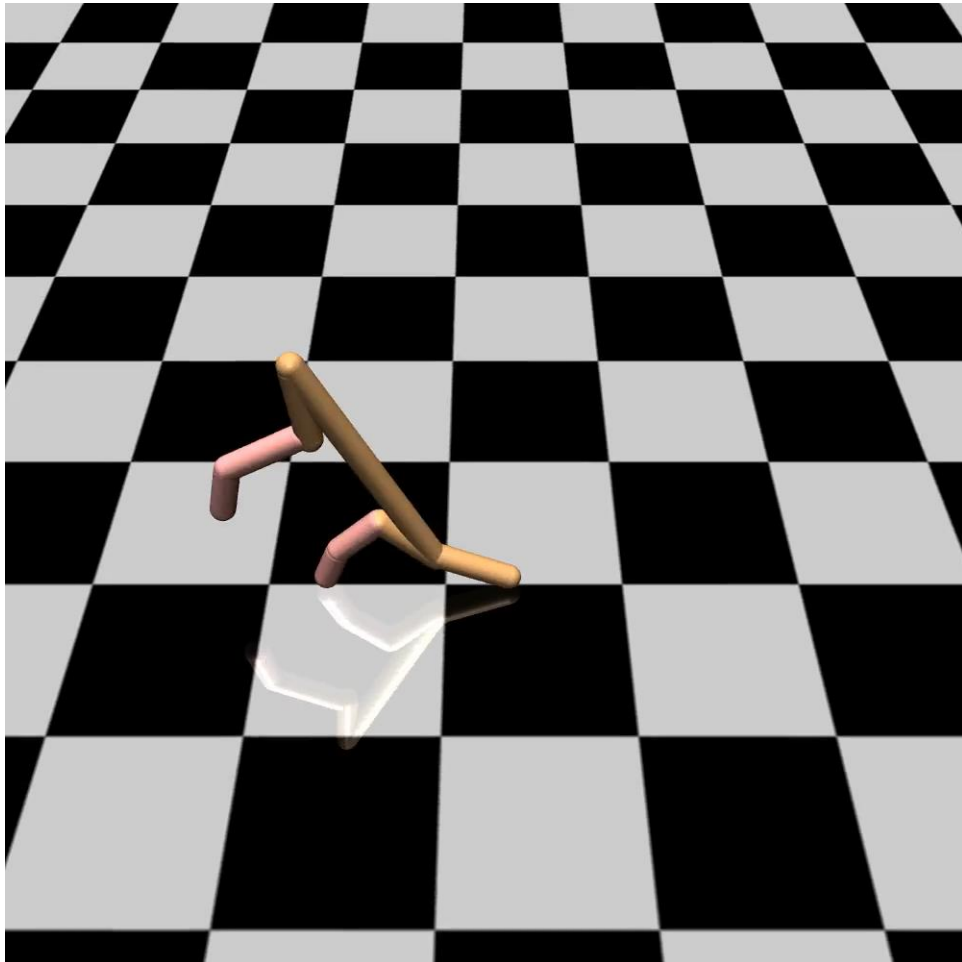
Two major changes to preference-based reward learning:

1. Instead of Bayesian learning, write a loss function and learn with gradient updates
2. After learning a reward, train a policy to generate new trajectories for the next iteration of reward learning

RLHF

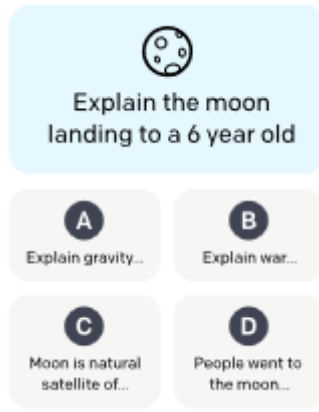


RLHF



InstructGPT

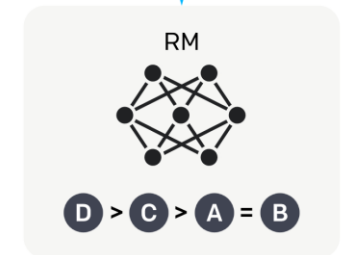
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Today...

- Learning from human feedback
 - Suboptimal demonstrations
 - Pairwise comparisons
 - Reinforcement learning from human feedback (RLHF)

Next time...

A day full of presentations!

- Myers et al., [Active Reward Learning from Online Preferences](#) (2023).
- Bajcsy et al., [Learning Robot Objectives from Physical Human Interaction](#) (2017).
- Bobu et al., [Inducing Structure in Reward Learning by Learning Features](#) (2022).
- Hadfield-Menell et al., [Inverse Reward Design](#) (2017).
- Kwon et al., [When Humans aren't Optimal: Robots that Collaborate with Risk-aware Humans](#) (2020).
- Chan et al., [Human Irrationality: Both Bad and Good for Reward Inference](#) (2021).
- Jeon et al., [Shared Autonomy with Learned Latent Actions](#) (2020).