Robot Learning

Learning from human feedback

USC University of Southern California

LiraLab

Last time…

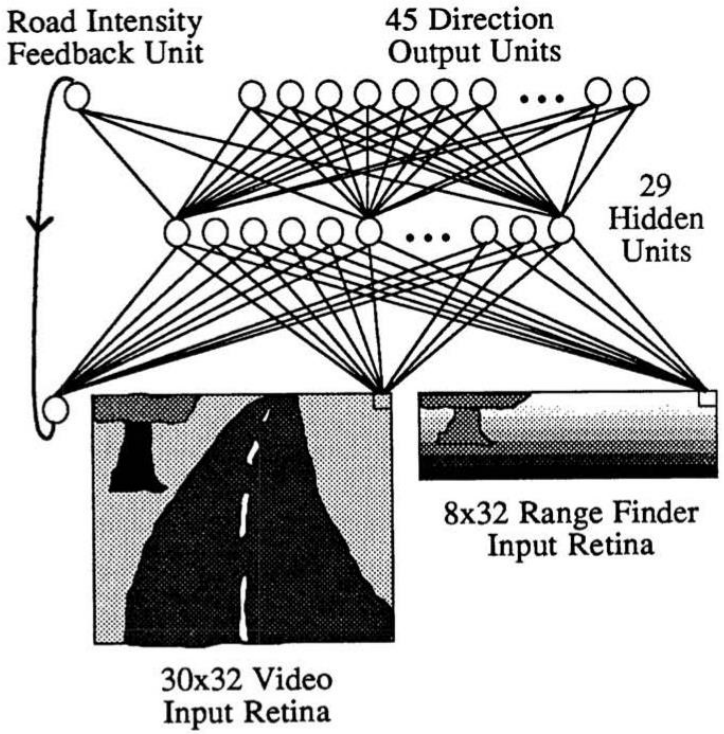
- Imitation learning

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

Behavioral cloning

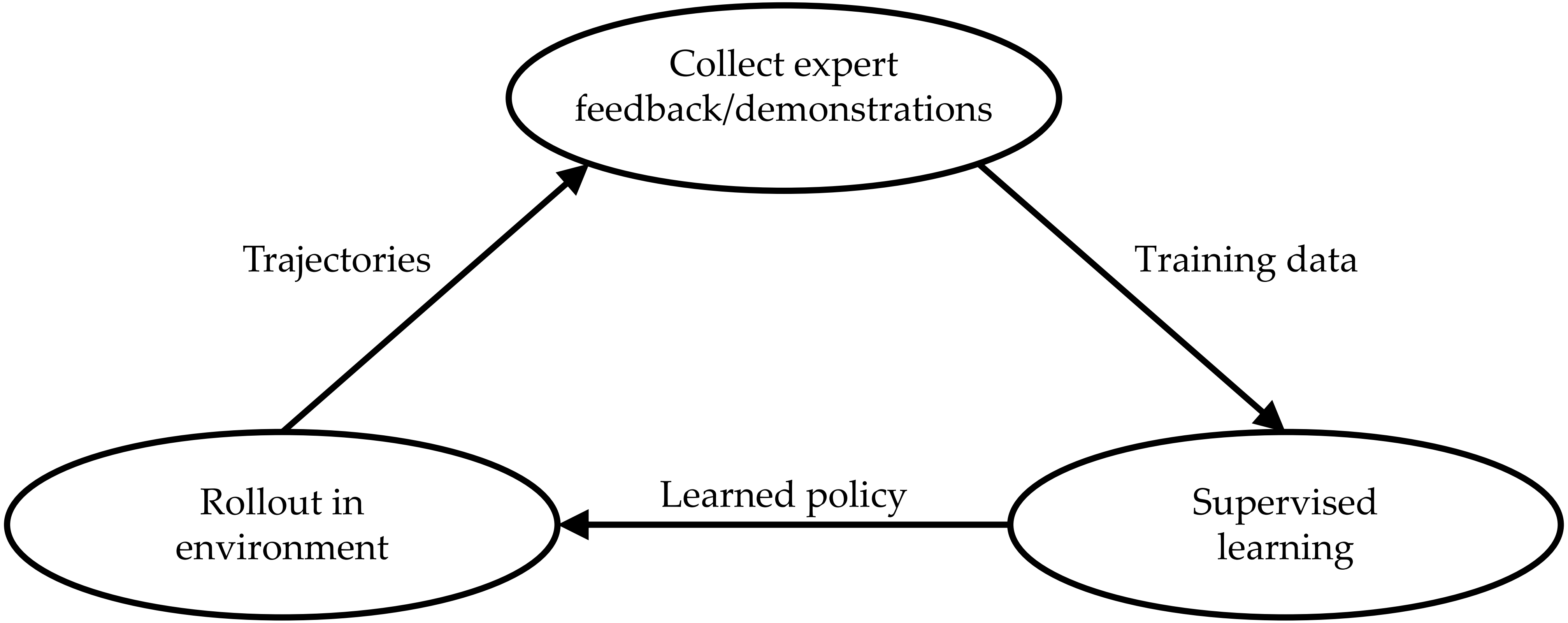
Train a neural network to map states into expert actions.

$s_i \longrightarrow$ Neural Network $\longrightarrow \hat{a}_i \longrightarrow L(a_i, \hat{a}_i)$



ALVINN: An Autonomous Land Vehicle in a Neural Network
Pomerleau, NeurIPS 1988

Direct policy learning



Apprenticeship learning

We iteratively improve the learned w and policy.

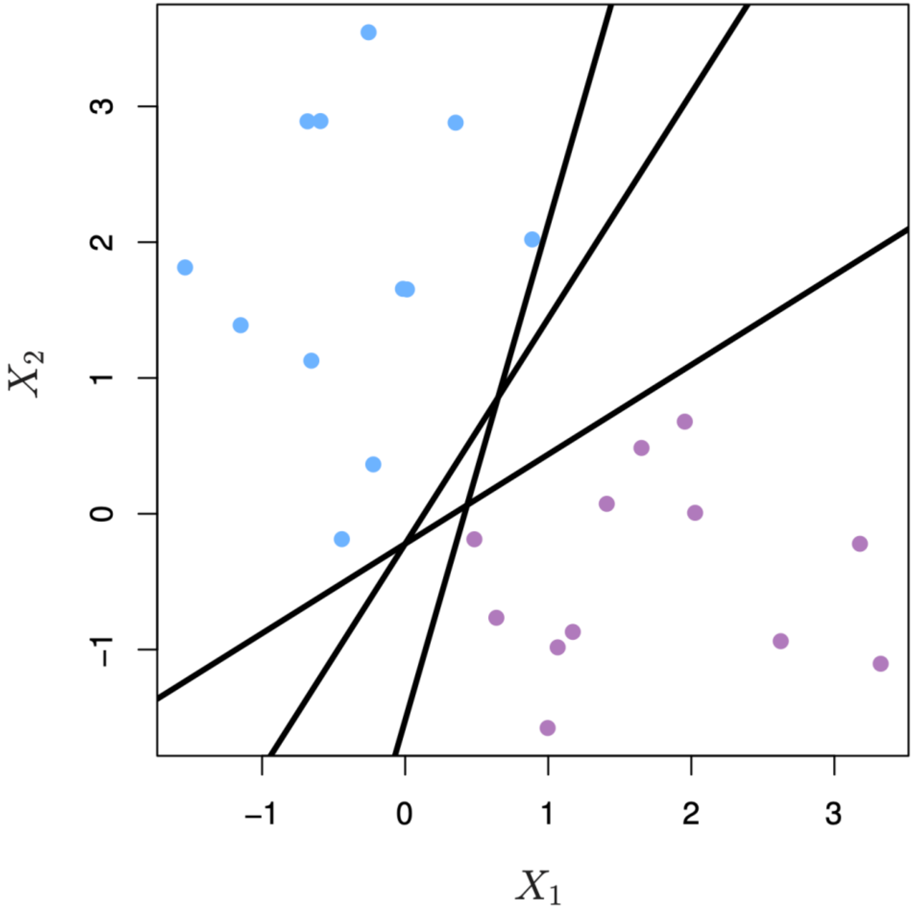
Compute the optimal features $\phi(\pi^*)$

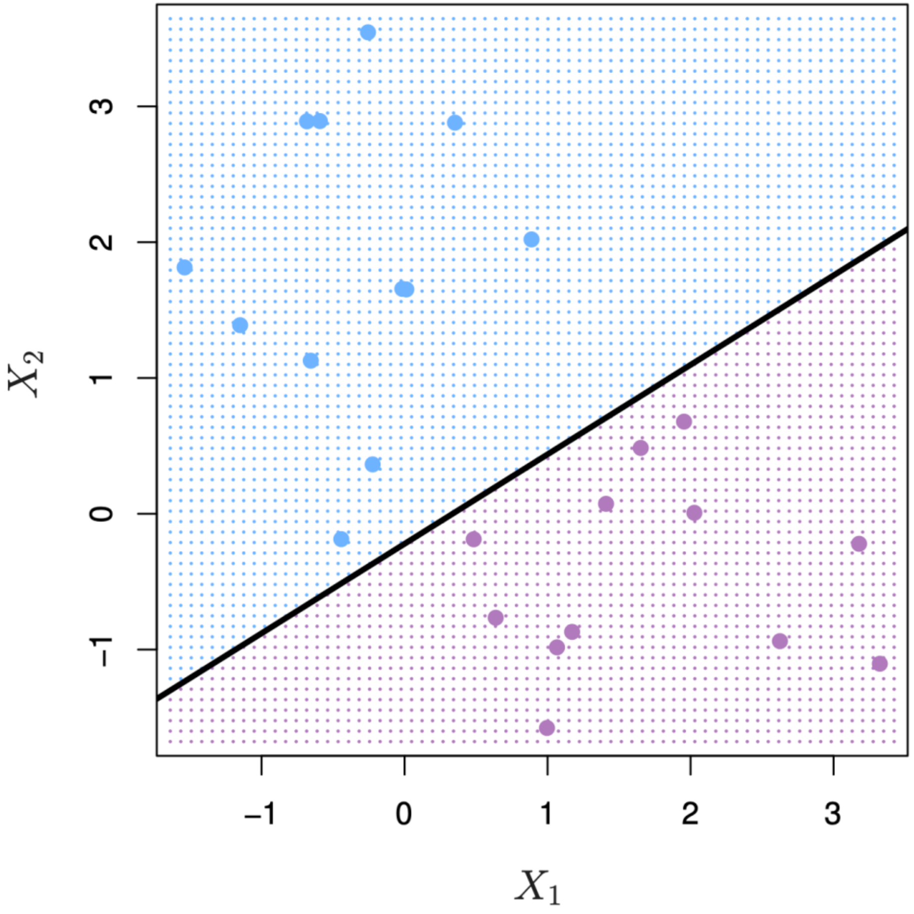
Initialize a policy π_0

Loop $i = 0,1,\ldots$**:**

- 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





From: Stanford STATS202

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:

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

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

Today…

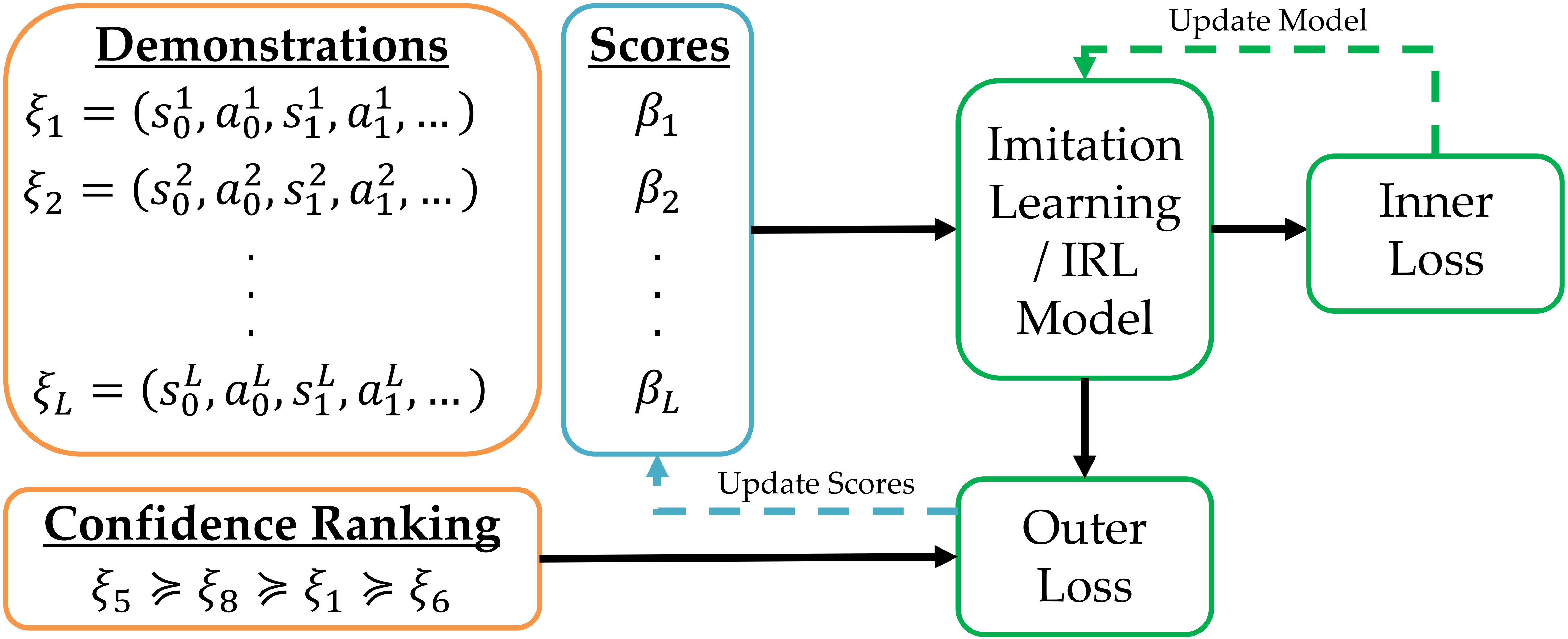
- Learning from human feedback
  - Suboptimal demonstrations
  - Pairwise comparisons
  - Reinforcement learning from human feedback (RLHF)

Confidence-aware imitation learning

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

Confidence-Aware Imitation Learning from Demonstrations with Varying Optimality
Zhang et al. NeurIPS 2021

Confidence-aware imitation learning



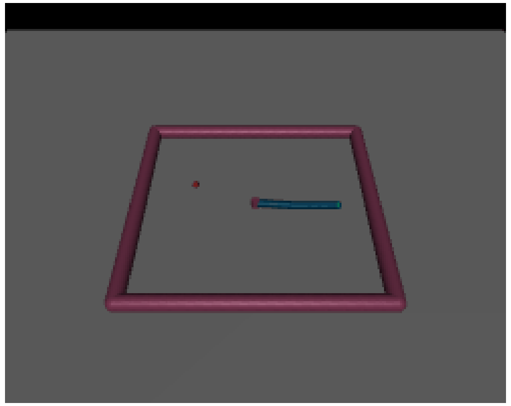
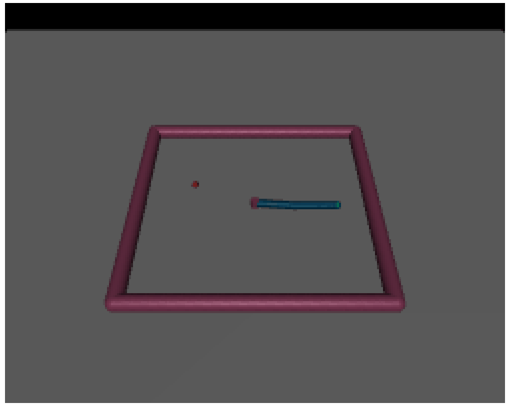
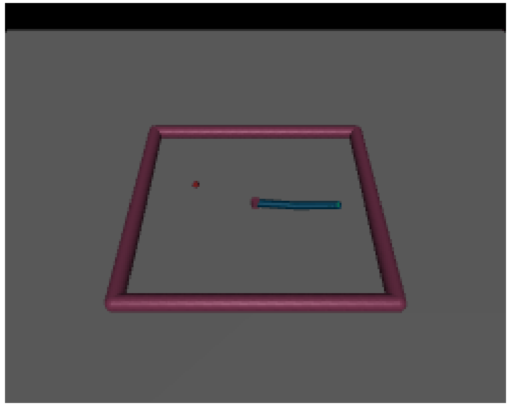
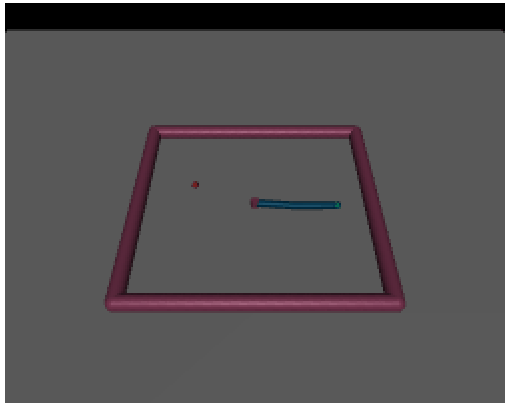
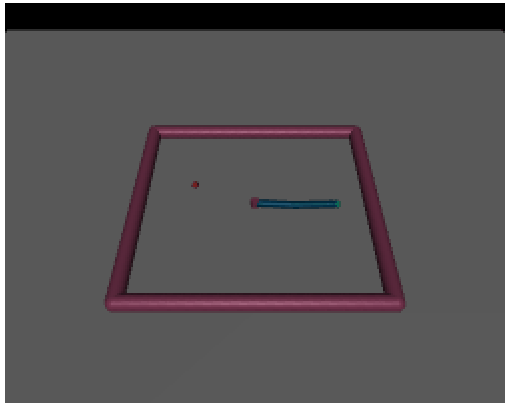
Confidence-Aware Imitation Learning from Demonstrations with Varying Optimality
Zhang et al. NeurIPS 2021

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)

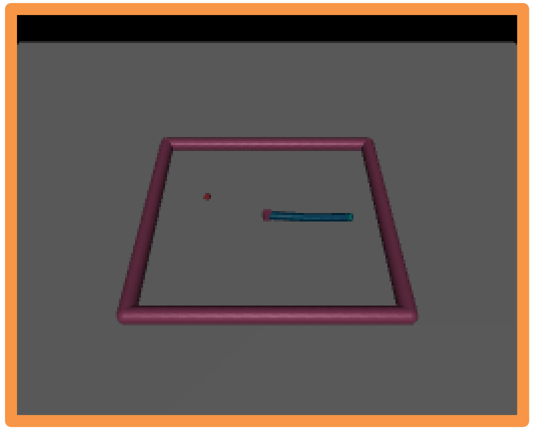
- Outer loss to learn confidence scores
  - Evaluates how well the demonstrations match the given (partial) ranking under the learned policy (or reward) to update β_i's.

Confidence-aware imitation learning

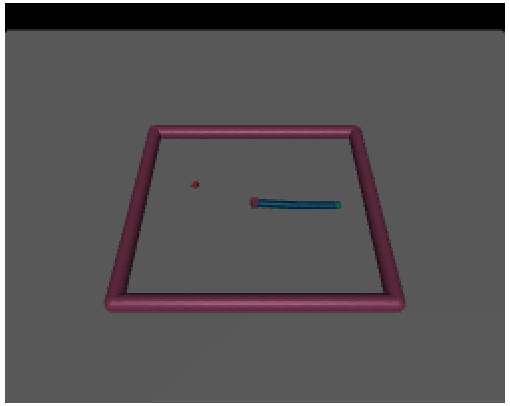






Confidence-Aware Imitation Learning from Demonstrations
with Varying Optimality
Zhang et al. NeurIPS 2021

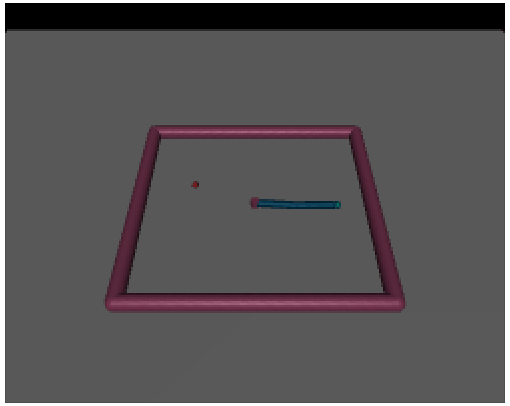
Confidence-aware imitation learning



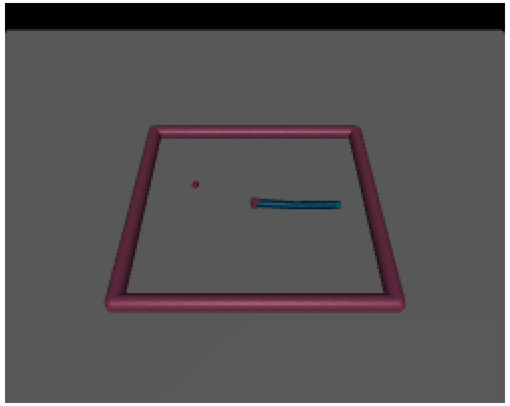
CAIL



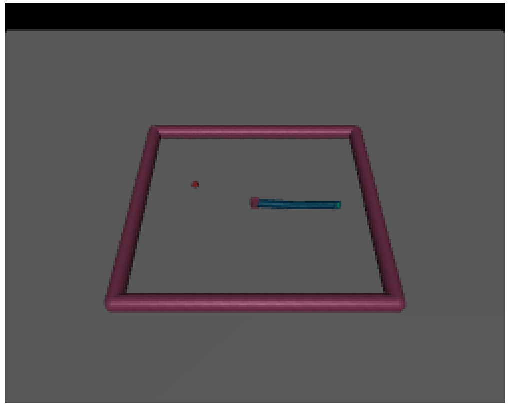
2IWIL



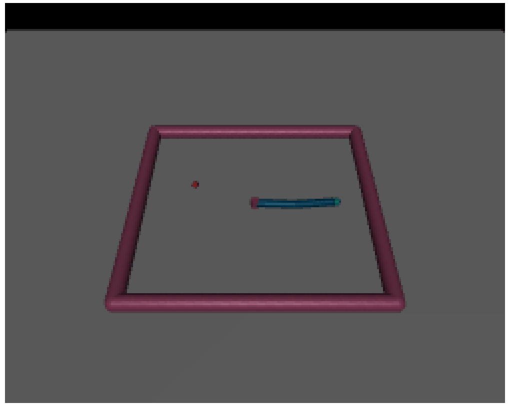
IC-GAIL



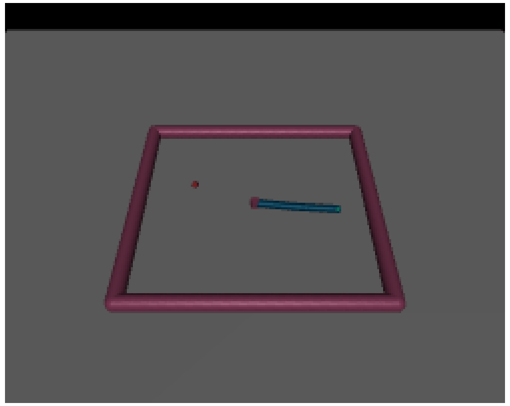
AIRL



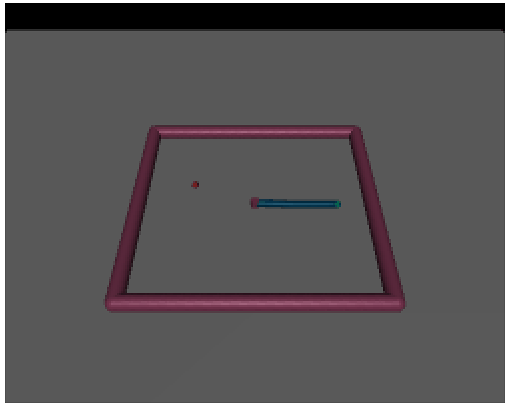
GAIL



T-REX



D-REX



SSRR

Confidence-Aware Imitation Learning from Demonstrations with Varying Optimality
Zhang et al. NeurIPS 2021

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.

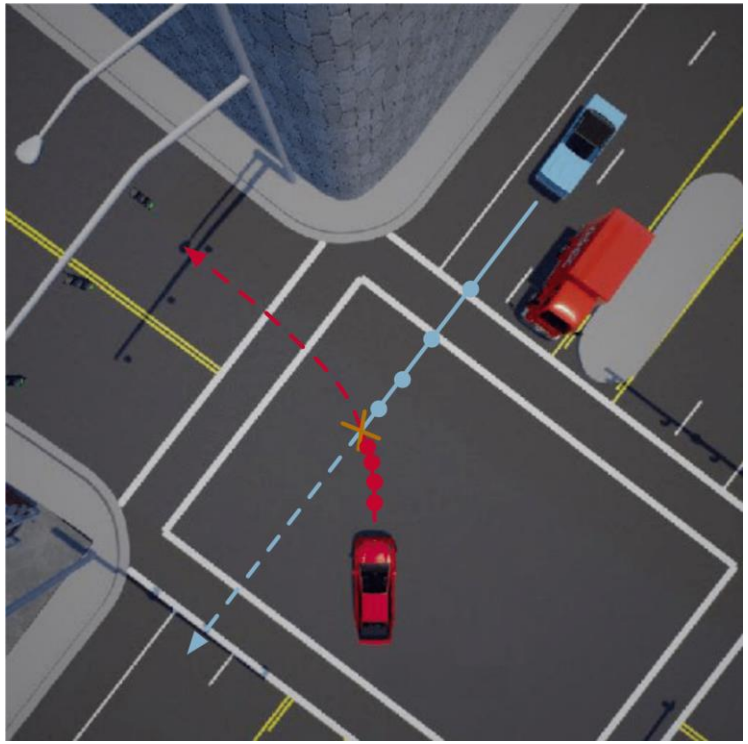
- How do we choose the expert to query if we have that ability?

Today…

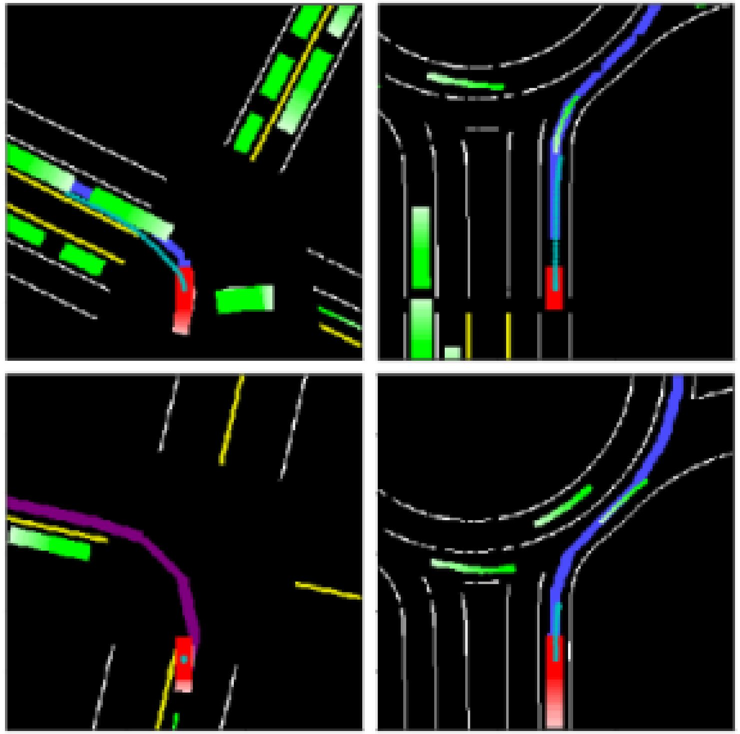
- Learning from human feedback
  - Suboptimal demonstrations
  - Pairwise comparisons
  - Reinforcement learning from human feedback (RLHF)

Learning from demonstrations (LfD)

Codevilla et al. ICRA'18



Cao et al. RSS'20



Chen et al. IROS'19

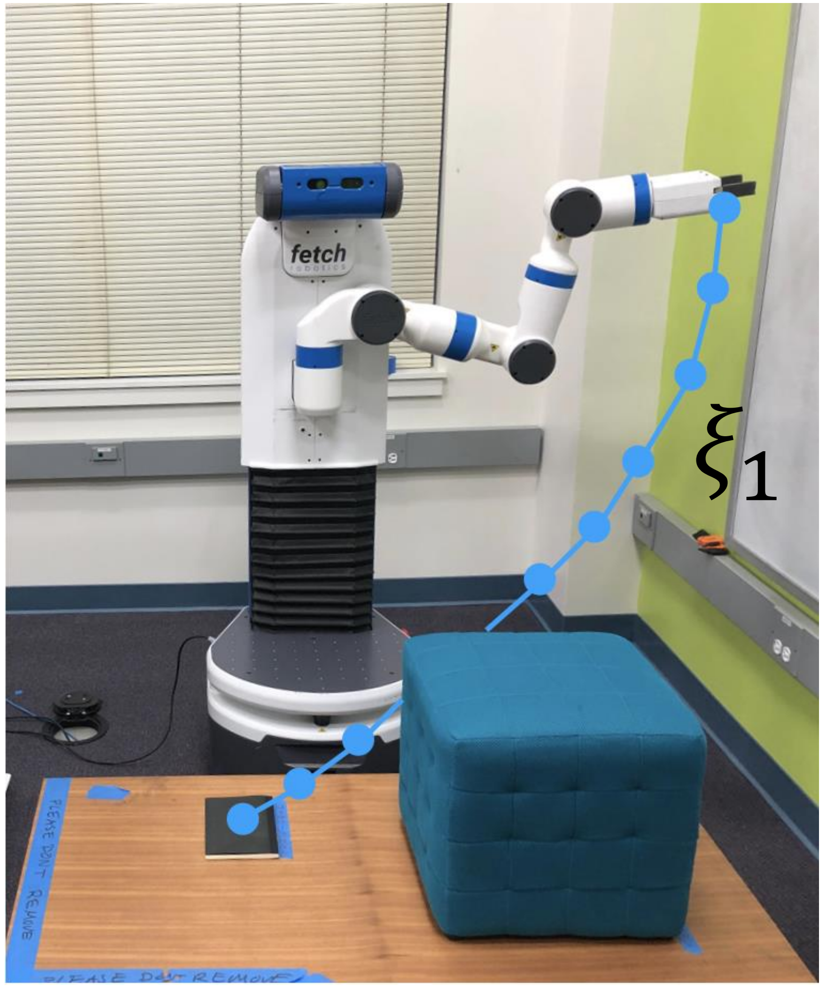
Why does LfD fail?

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

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 : $R(\xi_i) = f_w(\phi_i)$



Bayesian inverse reinforcement learning

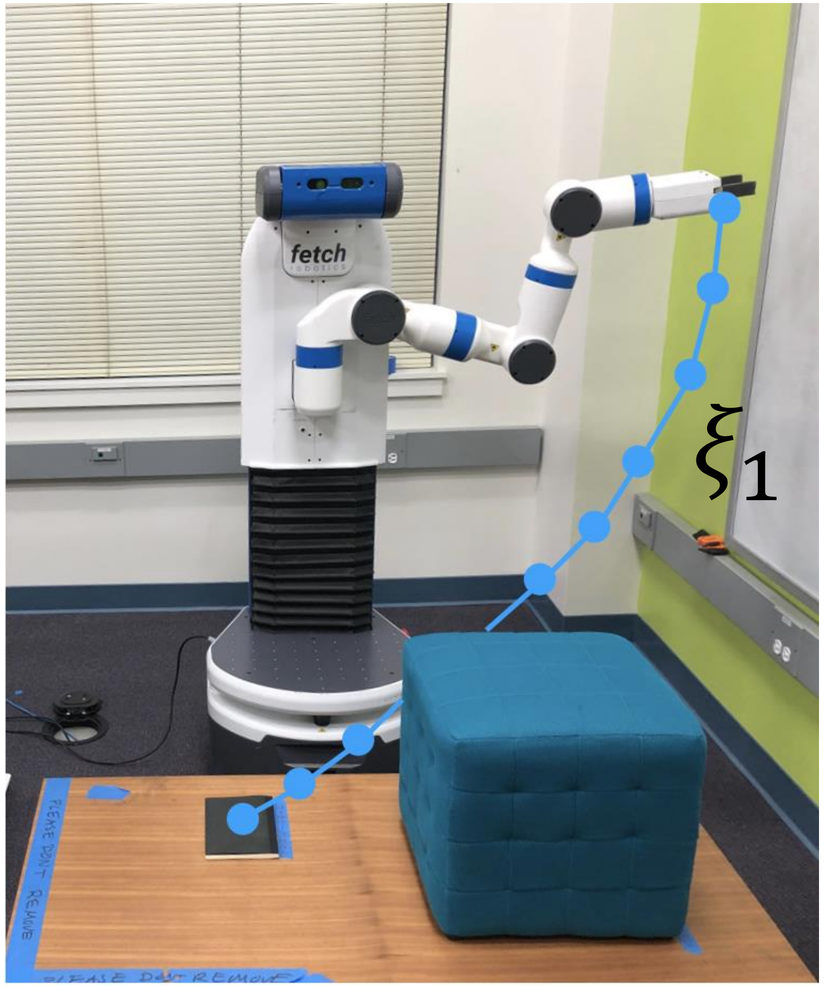
$$\underset{w}{\operatorname{argmax}}\, P(\, w \mid \mathcal{D}\,)$$

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

$$= P(w) \prod_{i=1}^{L} P(\xi_i \mid w)$$

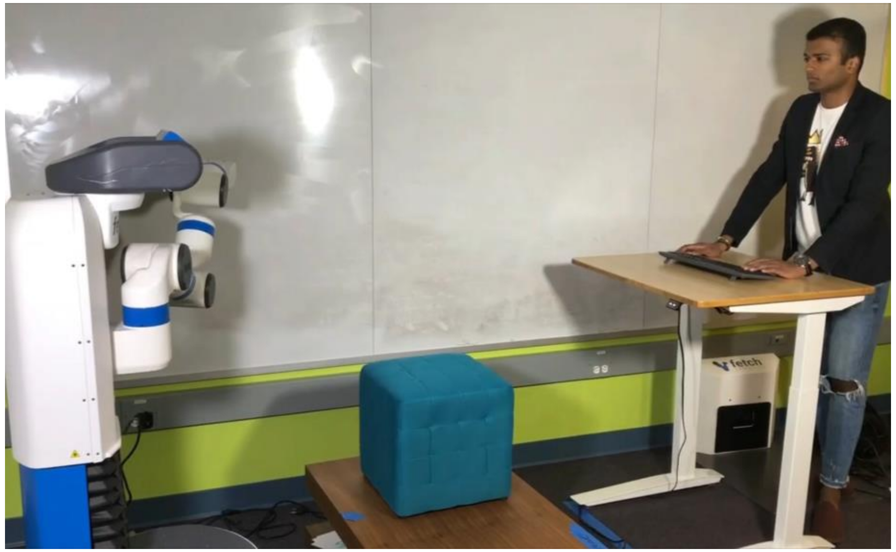
$$\boxed{\times \quad \propto P(w) \prod_{i=1}^{L} \exp f_w(\xi_i)}$$

(Noisy humans)



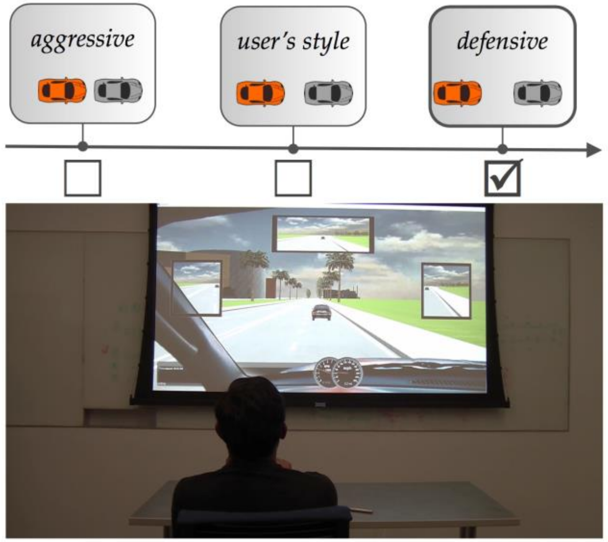
Humans are Suboptimal

Robots with high degrees of freedom are hard to teleoperate.



Palan et al. RSS'19

Humans do not like their own demonstrations.

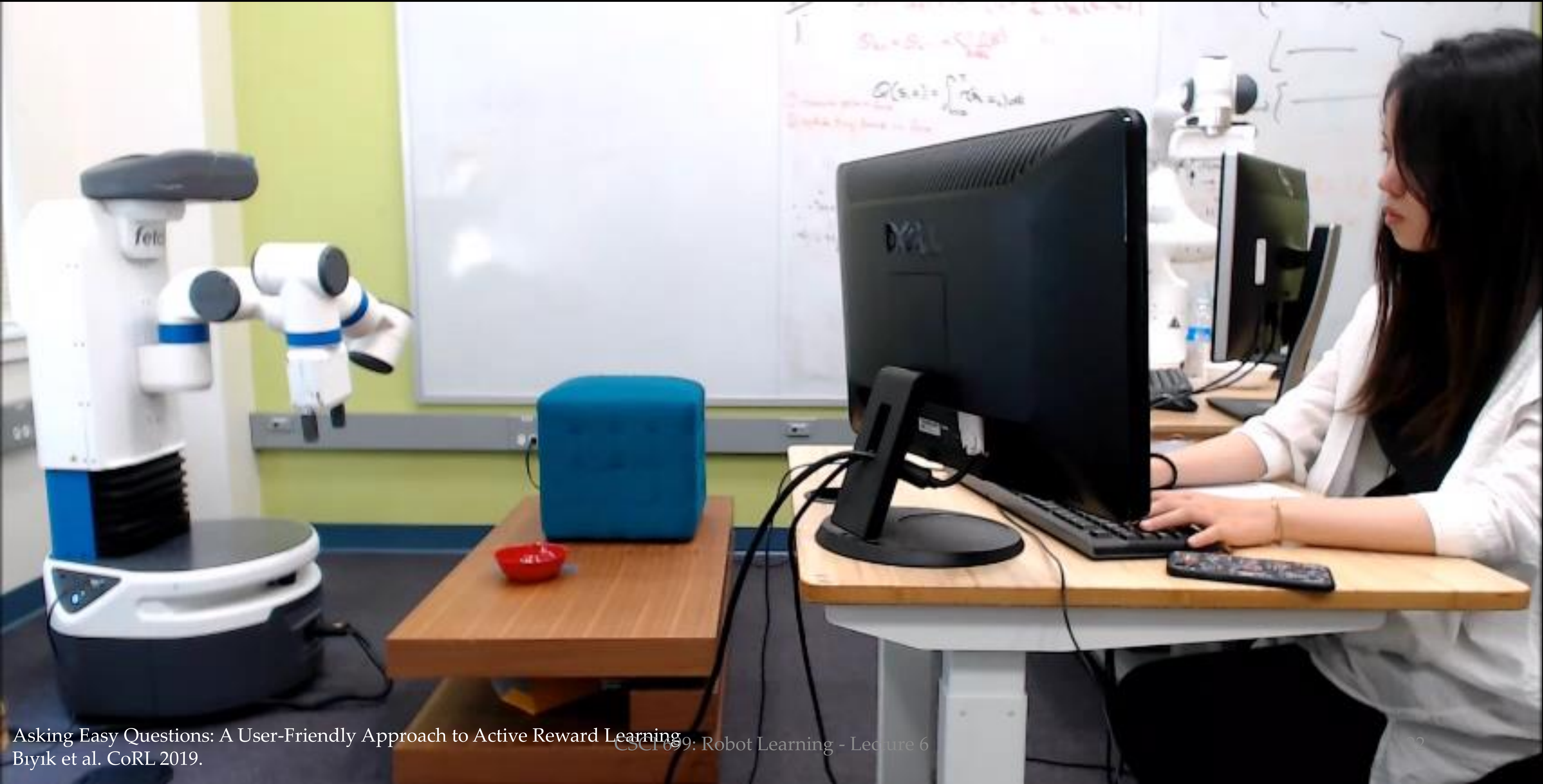


Basu et al. HRI'17

Humans take suboptimal actions in risky situations.

Kwon et al. HRI'20

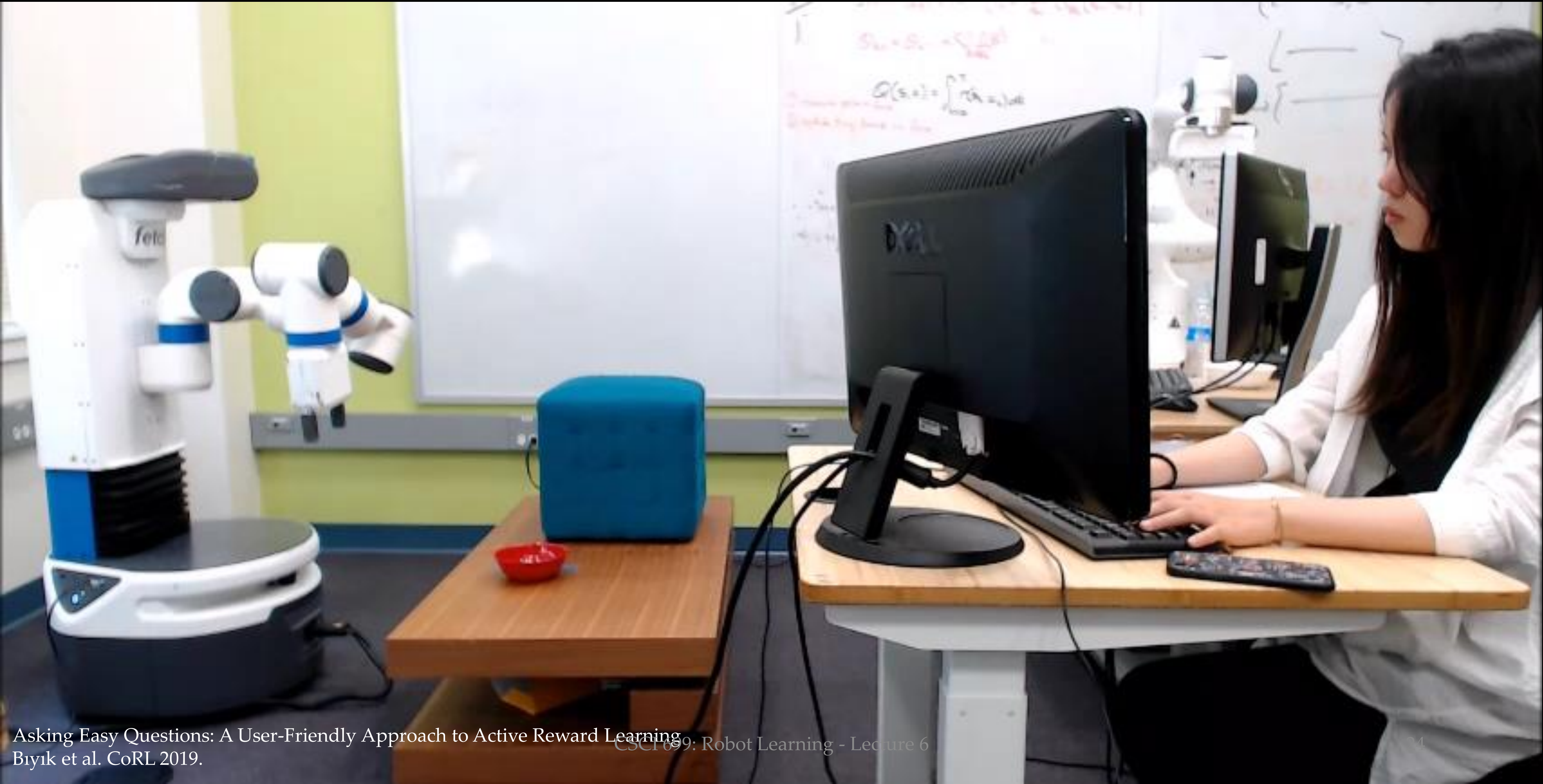
We can let the human evaluate a robot demonstration



Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019.

How dark is this blue?

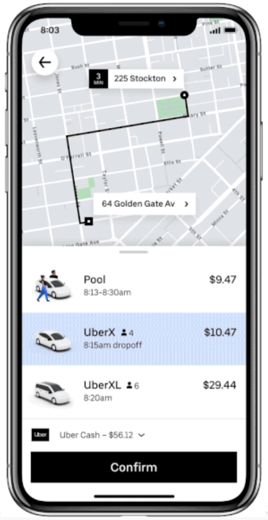
Human evaluations are often unreliable

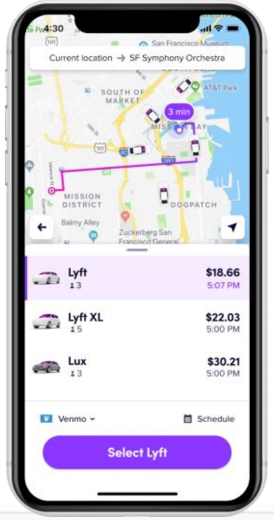


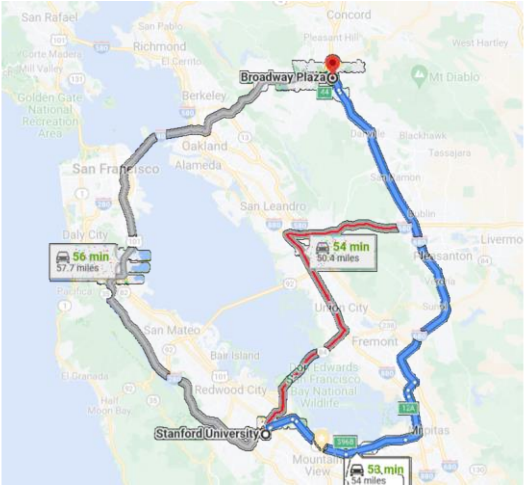
Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019.

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

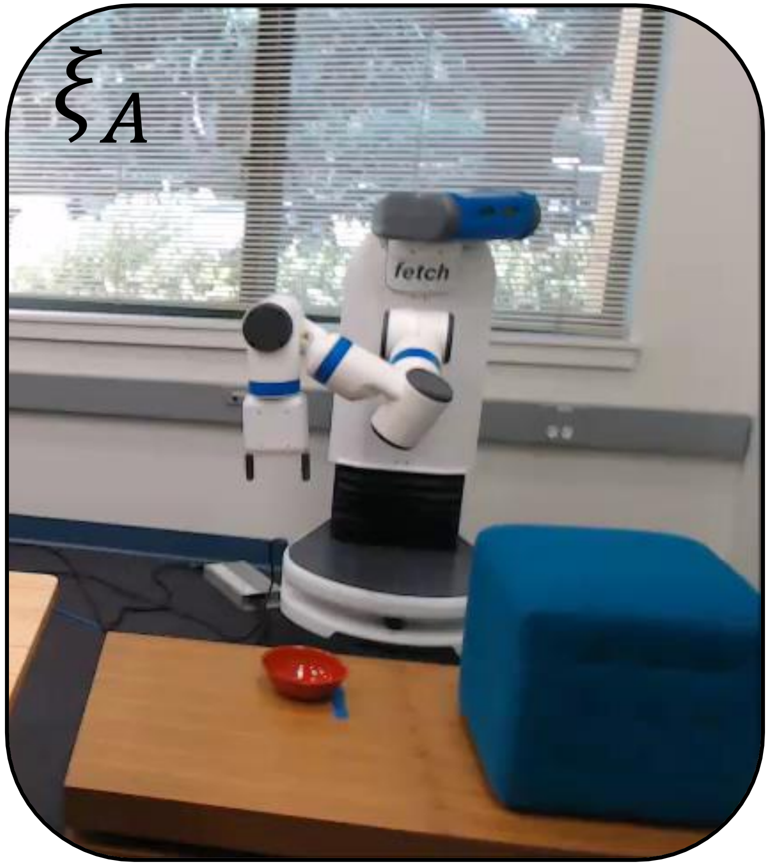
6.6M views • 4 years ago

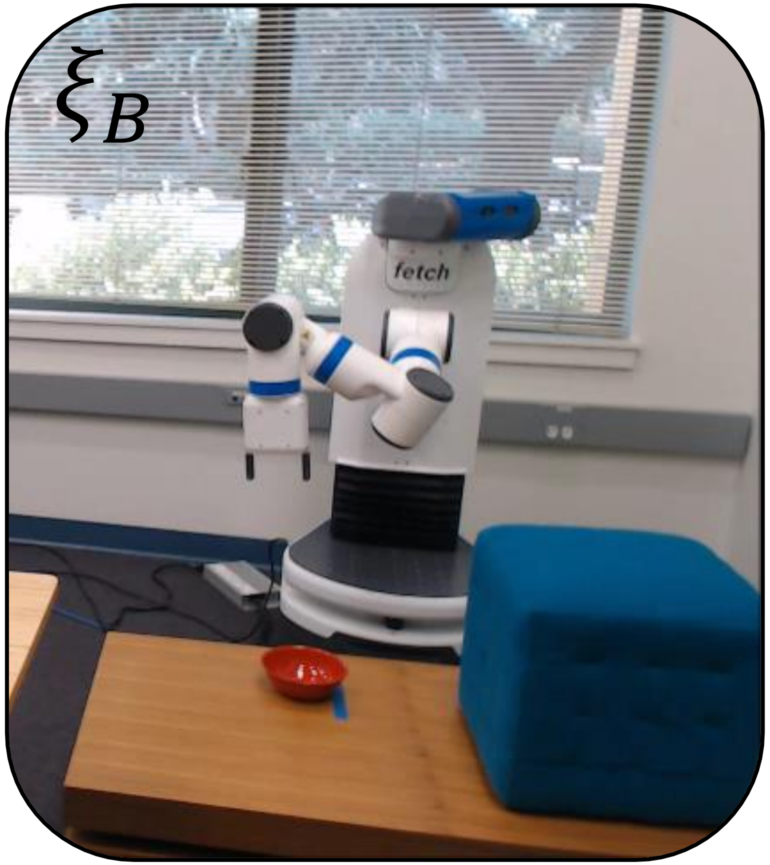
Carlsen - Nepomniachtchi | Game 8 | World Chess Championship | Howell,...

chess24

24K watching

Incorporating Comparisons





ξ_A or ξ_B?

Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences
Bıyık et al. IJRR 2021.

Incorporating Comparisons

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

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

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 : $R(\xi_i) = f_w(\phi_i)$

Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences
Bıyık et al. IJRR 2021.

Incorporating Comparisons

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

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

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

How do we compute this?

Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences
Bıyık et al. IJRR 2021.

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

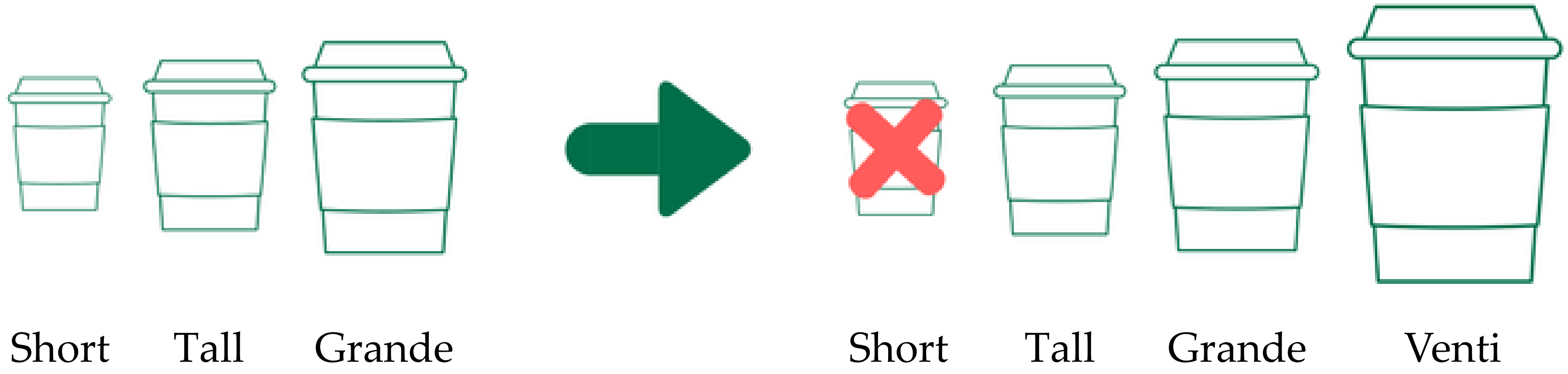
The probability of selecting one item over another from a pool of many items is not affected by the presence or absence of other items in the pool.

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

The choice axiom after twenty years
Luce. Journal of Mathematical Psychology 1977

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*.

The choice axiom after twenty years
Luce. Journal of Mathematical Psychology 1977

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

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

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

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

How do we compute this?

Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences
Bıyık et al. IJRR 2021.

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

$$\begin{aligned} P(q = A) &= P(u_A > u_B) \\ &= P\big(f_w(\phi(\xi_A)) + z_A > f_w(\phi(\xi_B)) + z_B\big) \\ &= P\left(z_A - z_B > f_w(\phi(\xi_B)) - f_w(\phi(\xi_A))\right) \end{aligned}$$

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

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

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

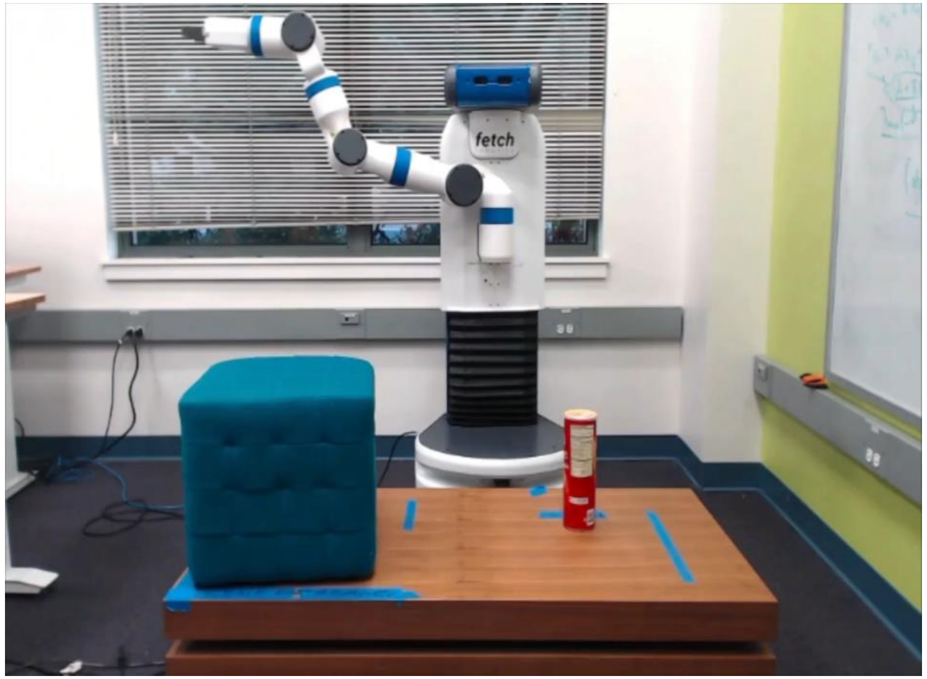
$$= P(w) \prod_{i=1}^{L} P(\xi_i \mid w) \prod_{i=1}^{N} P\left(q^{(i)} \mid w, \xi_A^{(i)}, \xi_B^{(i)}\right)$$

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

Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences
Bıyık et al. IJRR 2021.

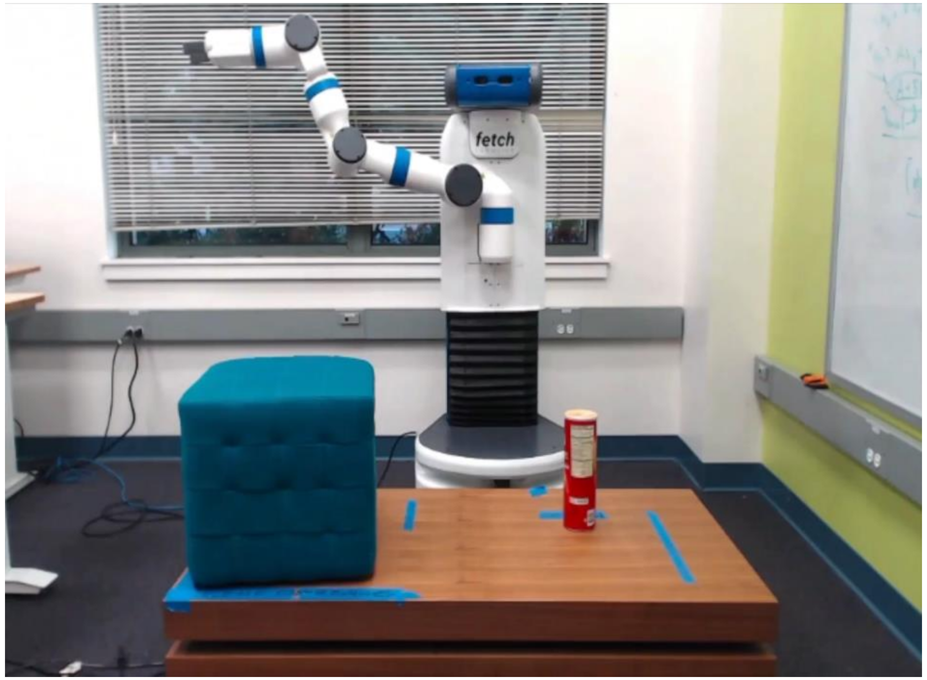
Benefit of comparisons

Bayesian IRL



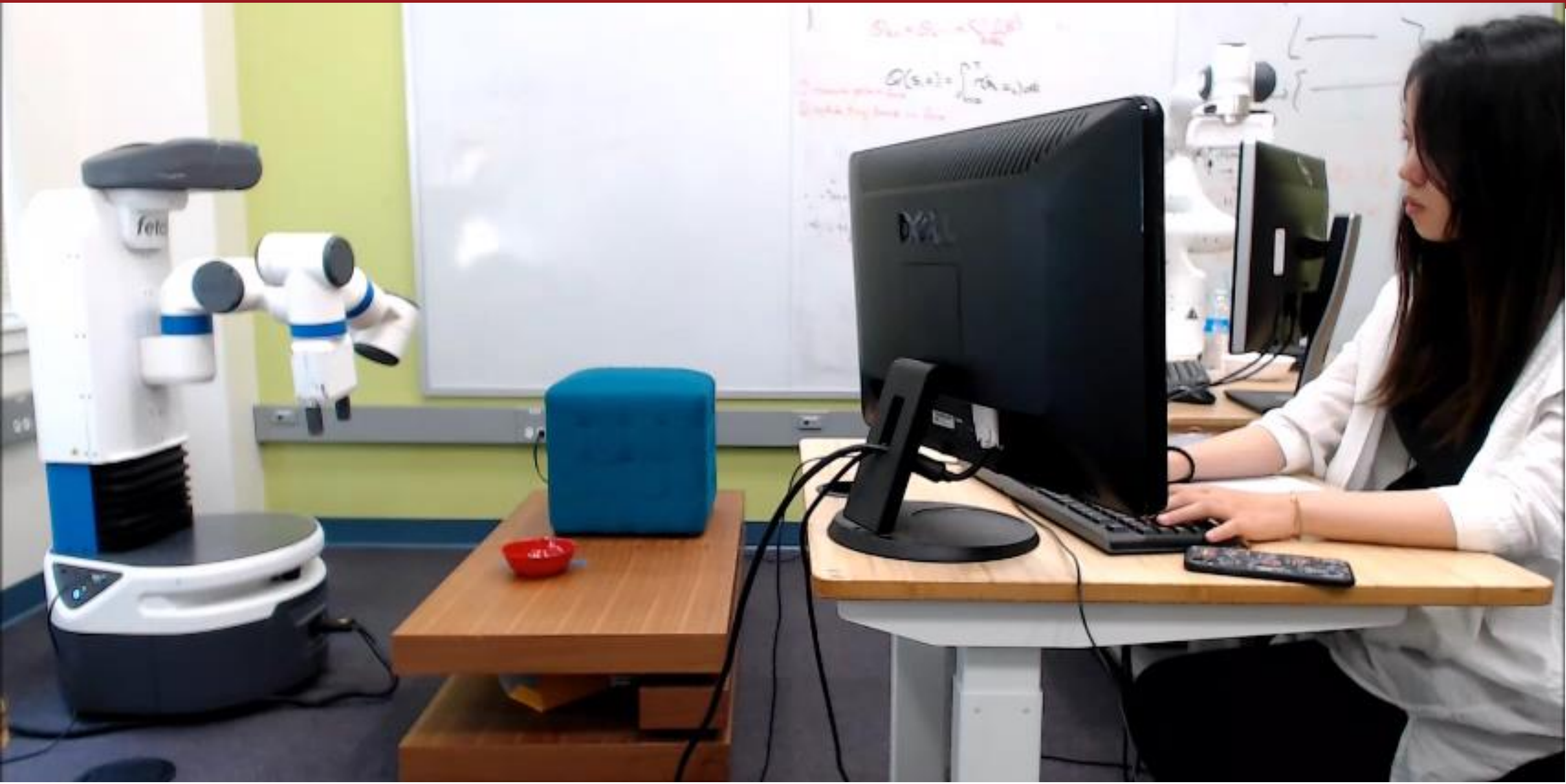
5 demonstrations

Ours

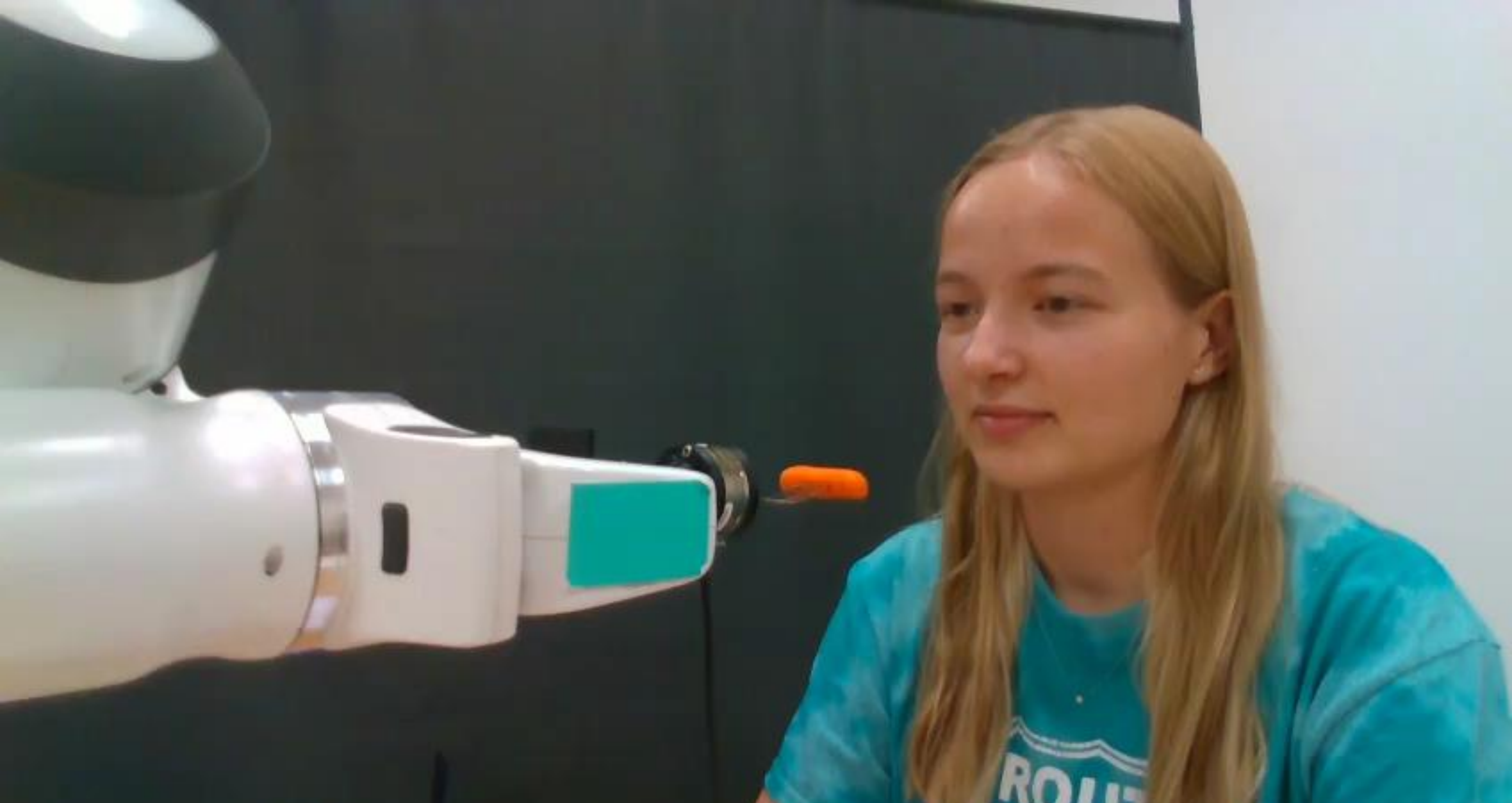


1 demonstration + 15 comparisons

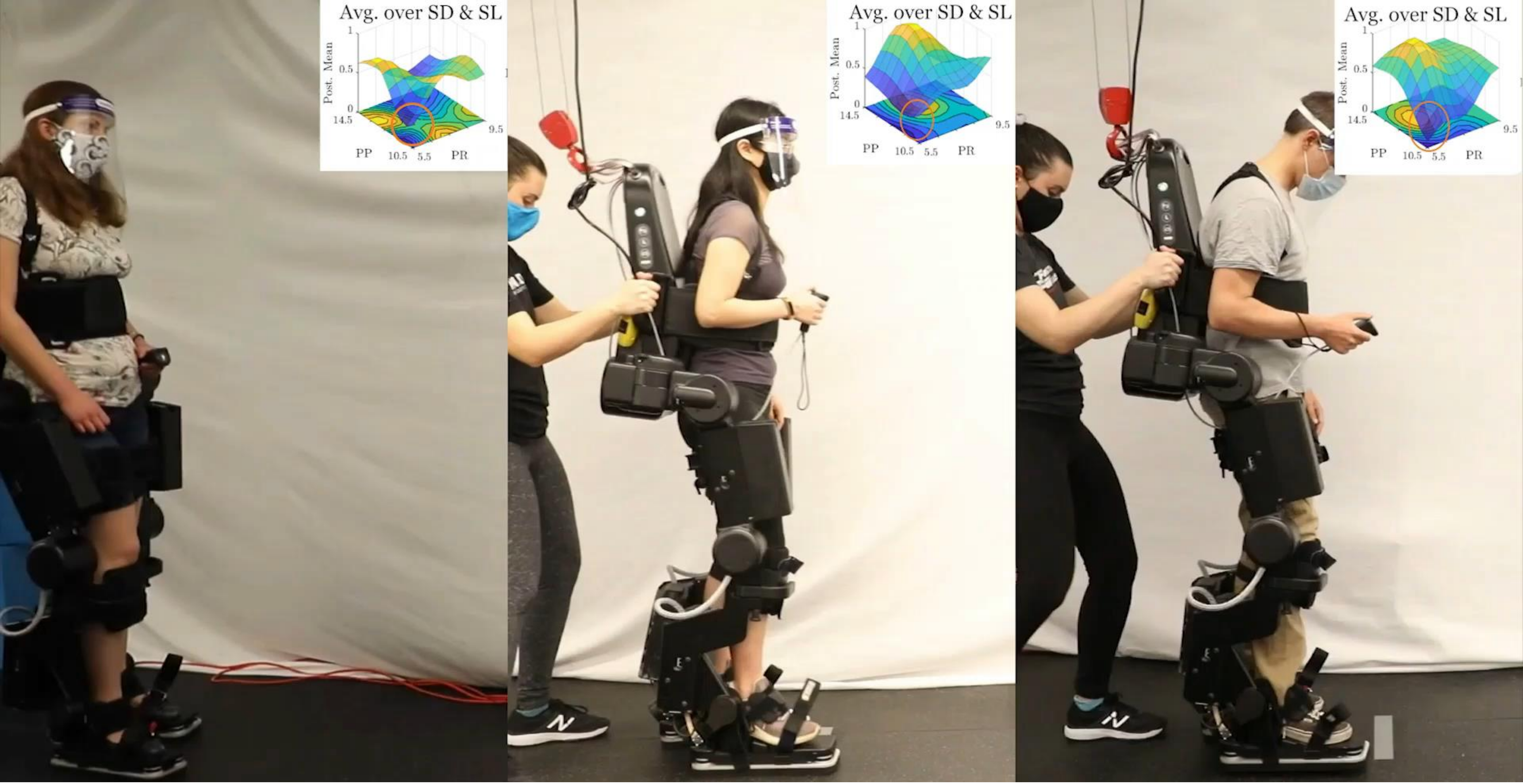
Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences
Bıyık et al. IJRR 2021.



Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019.



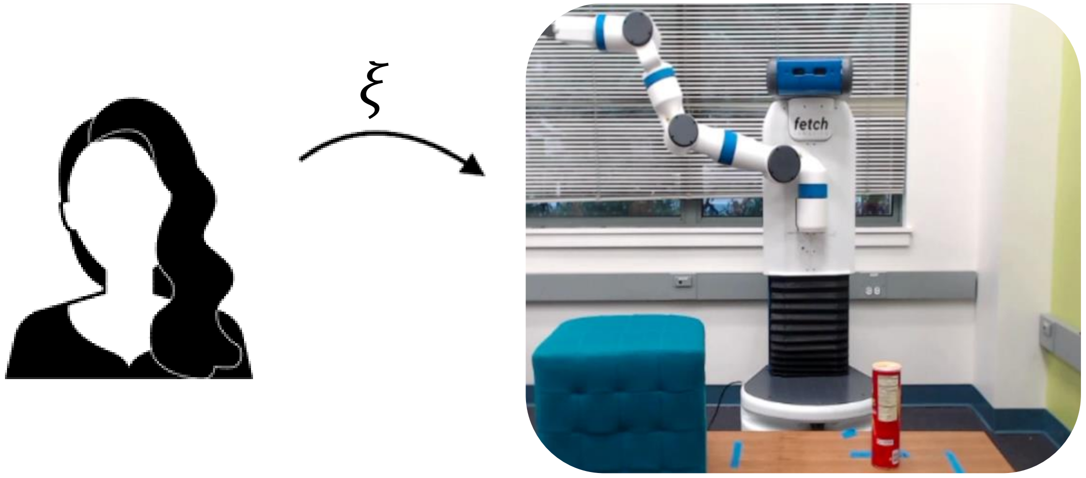
Balancing Efficiency and Comfort in Robot-Assisted Bite Transfer
Belkhale et al. ICRA 2022



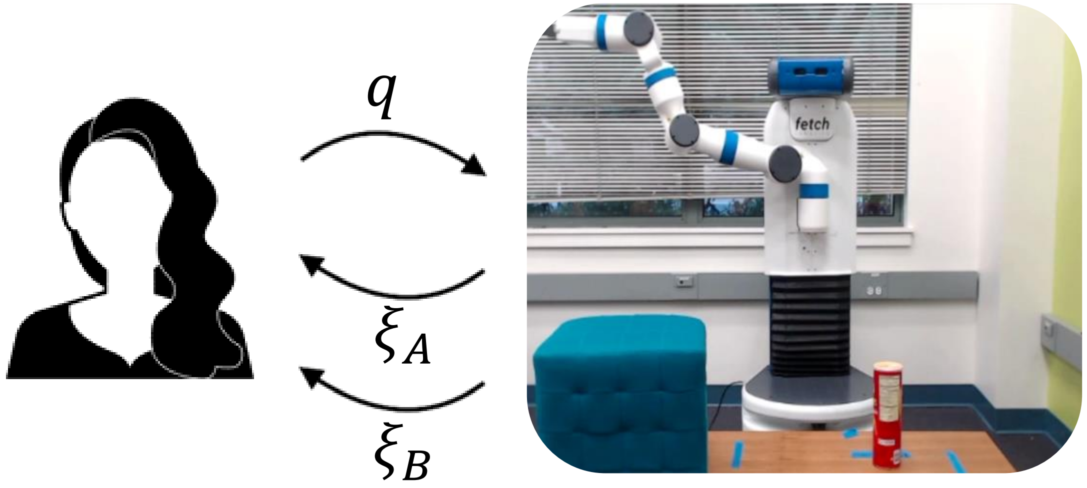
ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes
Li et al., ICRA 2021

Choosing Queries

Demonstrations



Comparisons



How do we quantify information?

Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences
Bıyık et al. IJRR 2021.

Surprise

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

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

Entropy (a measure of uncertainty)

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

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

Entropy is the expected surprise.

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

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

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

Another example

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

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

$H(X) = 1$

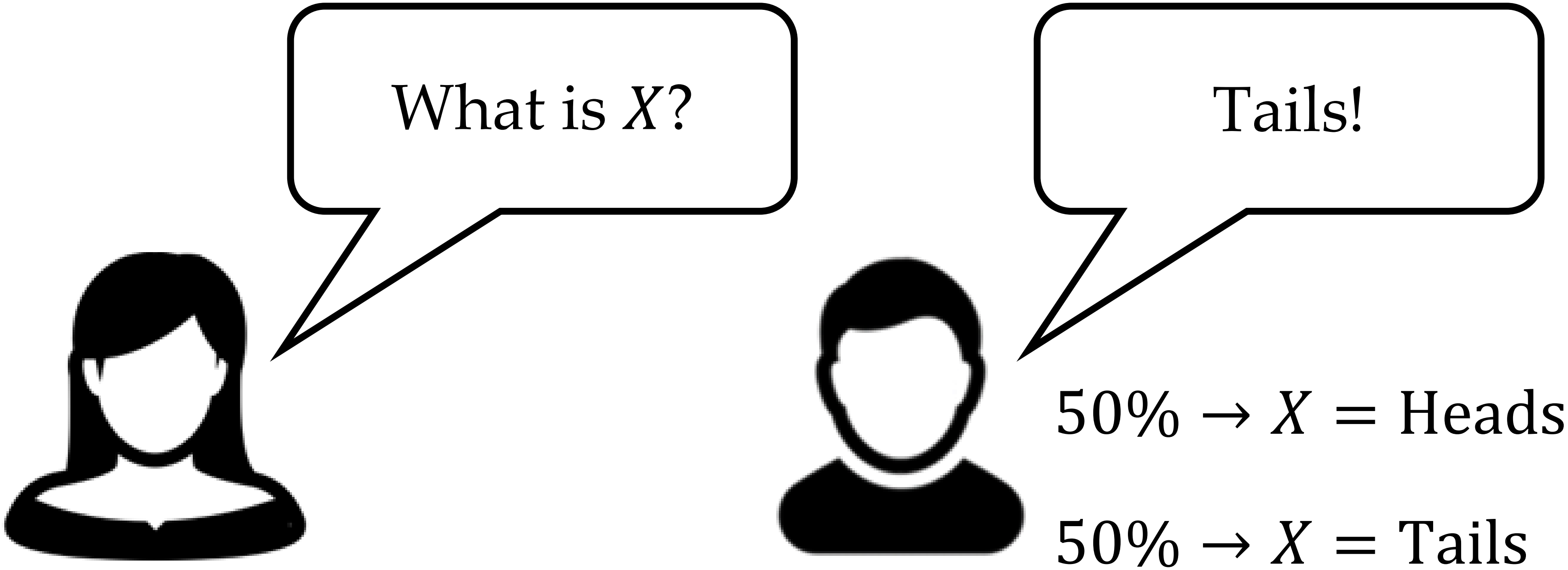
Alice

Bob

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

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

Mutual information



Alice

Bob

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

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

<u>Uncertainty</u>

$$H(X \mid X) = 0$$

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

This is 0 in information theory.

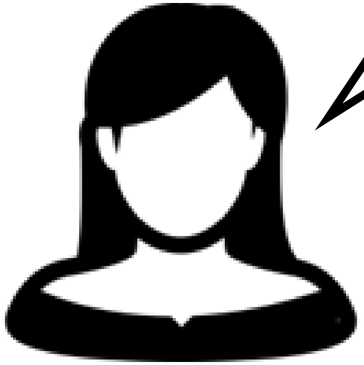
Mutual information

What is X?

Tails!

Uncertainty

$H(X \mid X) = 0$



Alice

Bob

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

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

Mutual Information = Reduction in Entropy: $I(X;X) = H(X) - H(X \mid X)$

$= 1 - 0 = 1$

Mutual information

<u>Uncertainty</u>

$H(X) = 1$

Alice

Noisy Bob

(lies with 5% probability)

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

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

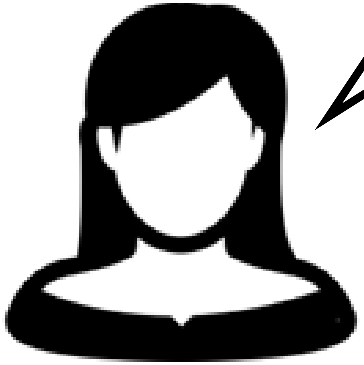
Mutual information

What is X'?

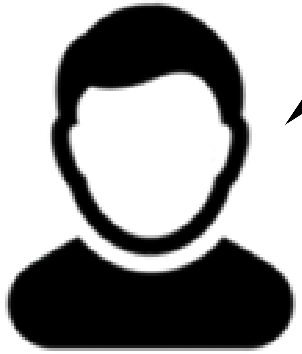
Tails!

<u>Uncertainty</u>

$H(X \mid X' = \text{Tails})$



Alice



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

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

Noisy Bob

(lies with 5% probability)

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

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

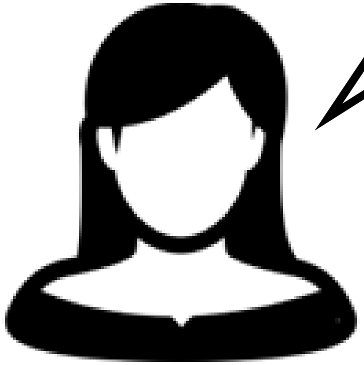
Mutual information

What is X'?

Tails!

<u>Uncertainty</u>

$H(X \mid X' = \text{Tails})$



Alice

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

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

Noisy Bob

(lies with 5% probability)

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

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

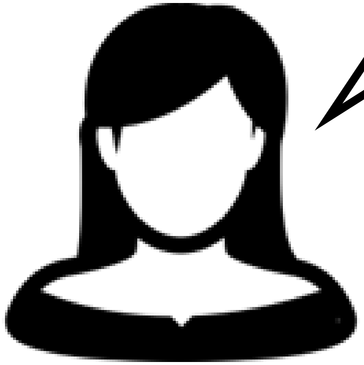
Mutual information

What is X'?

Tails!

Uncertainty

$H(X \mid X' = \text{Tails})$
$\cong 0.286$



Alice

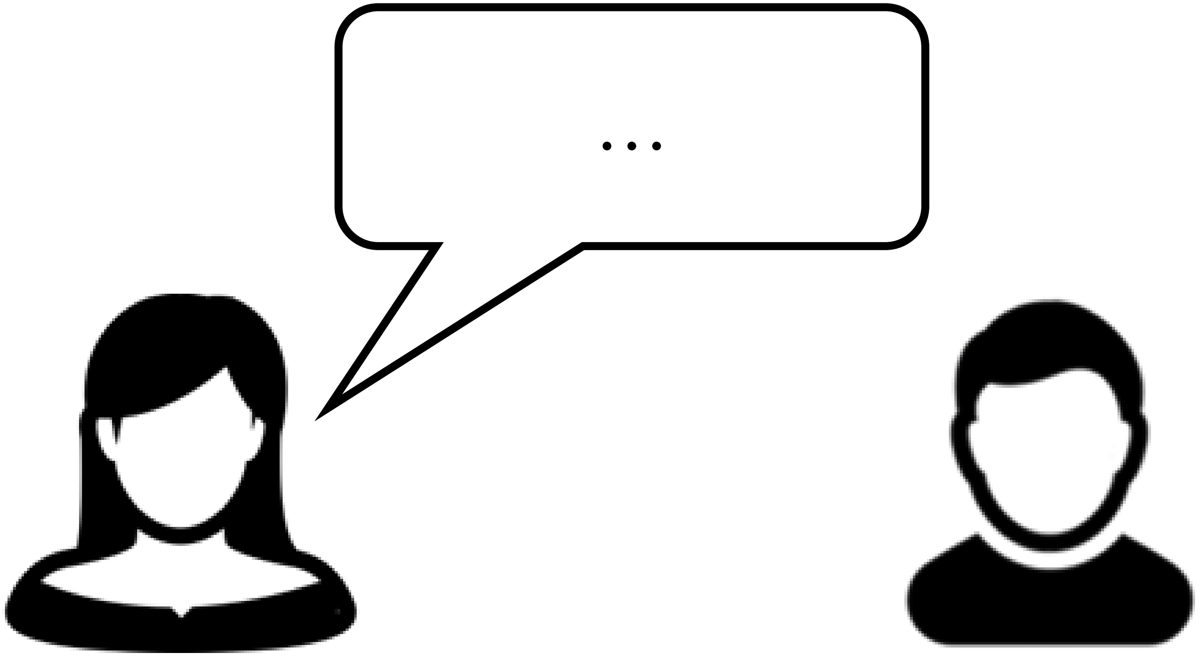
$50\% \rightarrow X = \text{Heads}$

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

Noisy Bob
(lies with 5% probability)

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

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)

$20\% \rightarrow X = (\mathrm{H}, \mathrm{H}, \mathrm{H})$

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

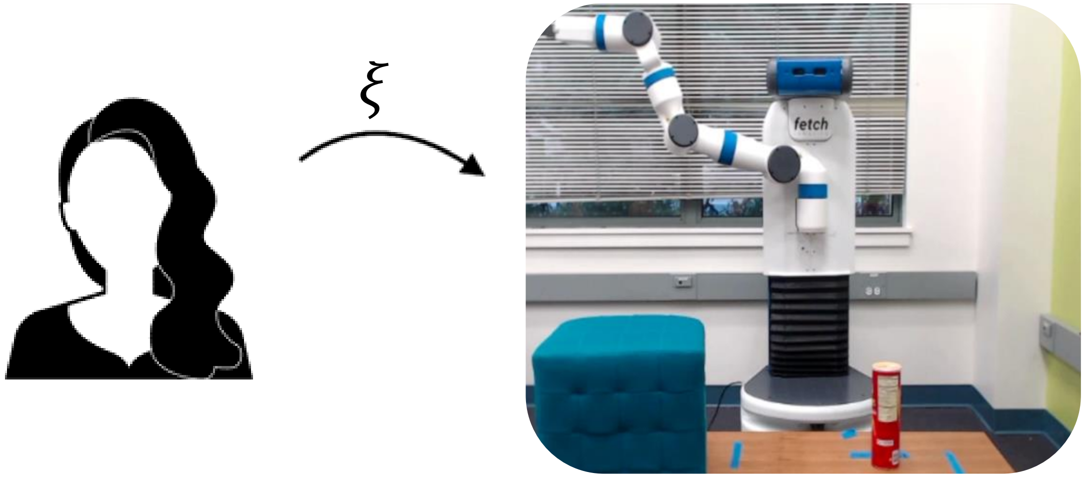
$20\% \rightarrow X = (\mathrm{H}, \mathrm{T}, \mathrm{H})$

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

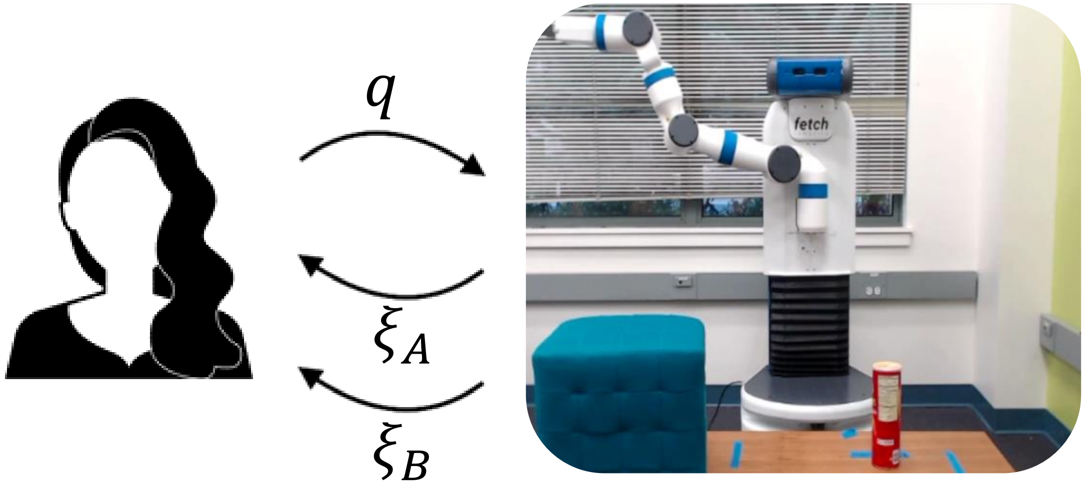
$20\% \rightarrow X = (\mathrm{T}, \mathrm{T}, \mathrm{T})$

Choosing queries

Demonstrations



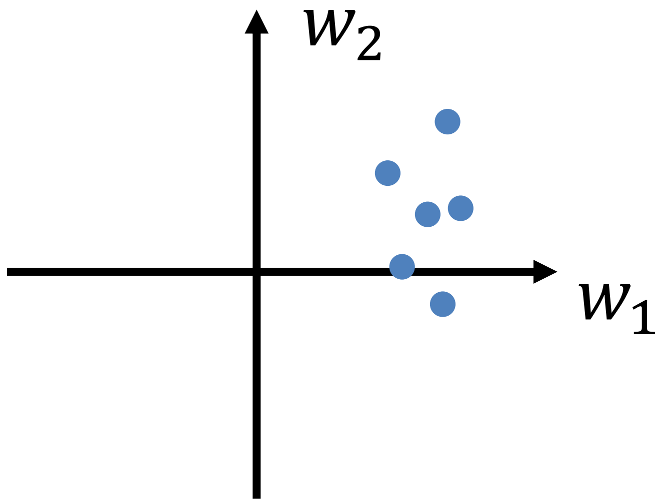
Comparisons



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

Maximum volume removal

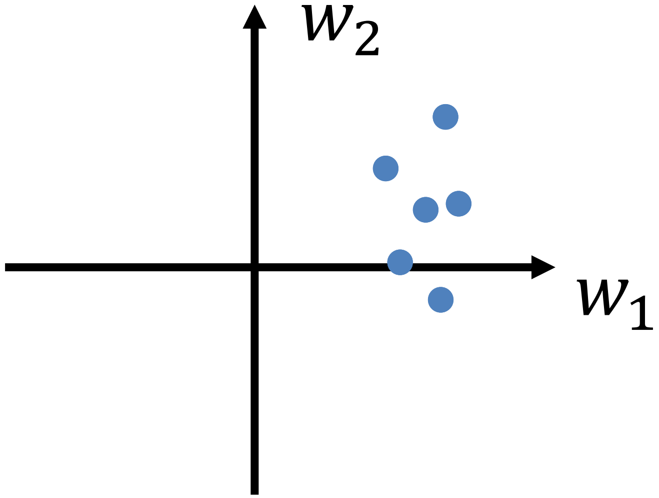
Posterior $P(w \mid \boldsymbol{C})$

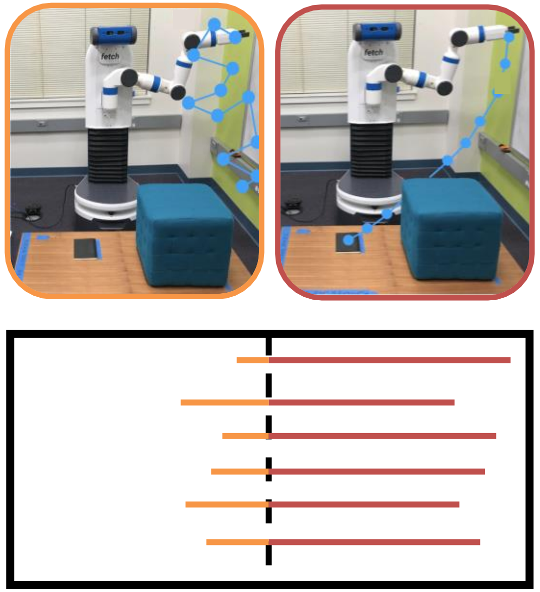


Active Preference-Based Learning of Reward Functions
Sadigh et al. RSS 2017

Maximum volume removal

Posterior $P(w \mid \boldsymbol{c})$

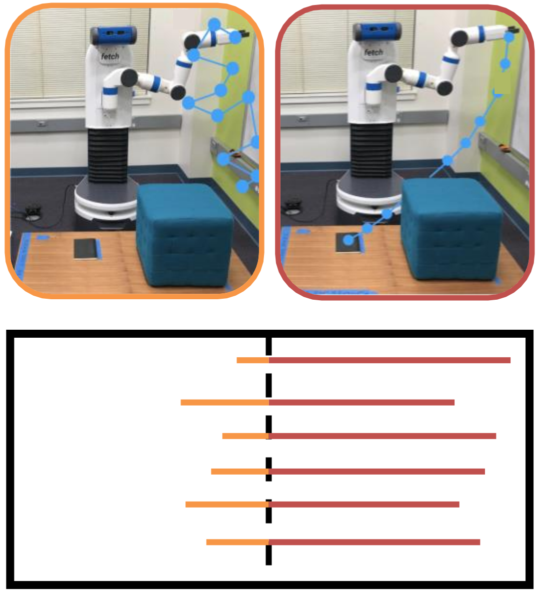


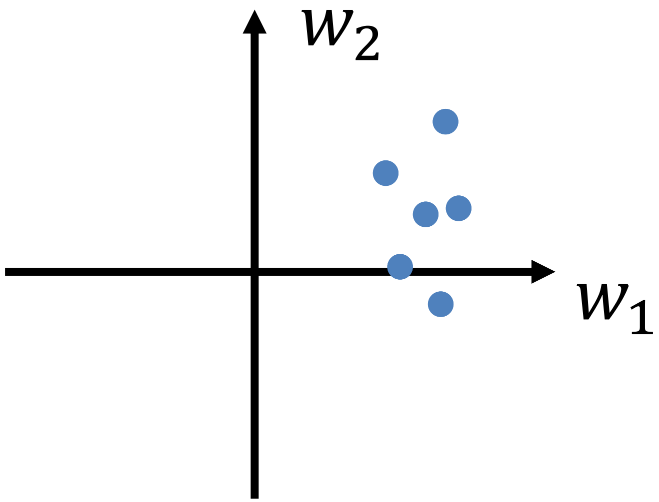


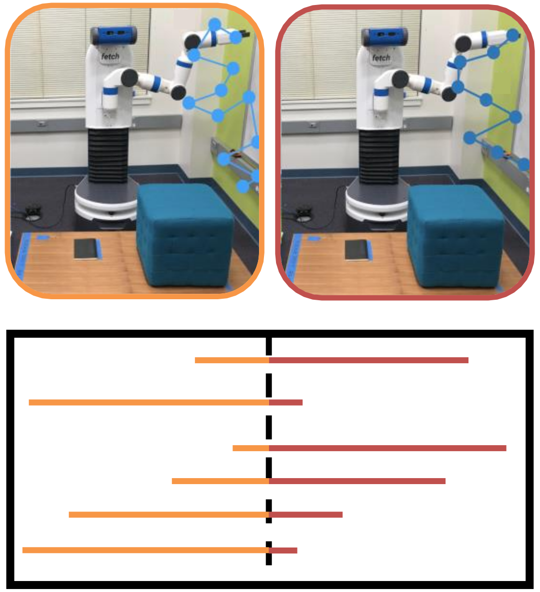
User Choice

Maximum volume removal

Posterior $P(w \mid \boldsymbol{c})$



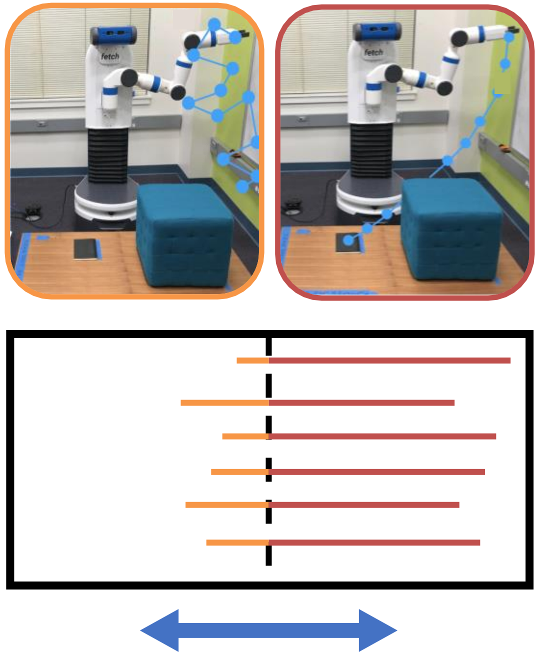


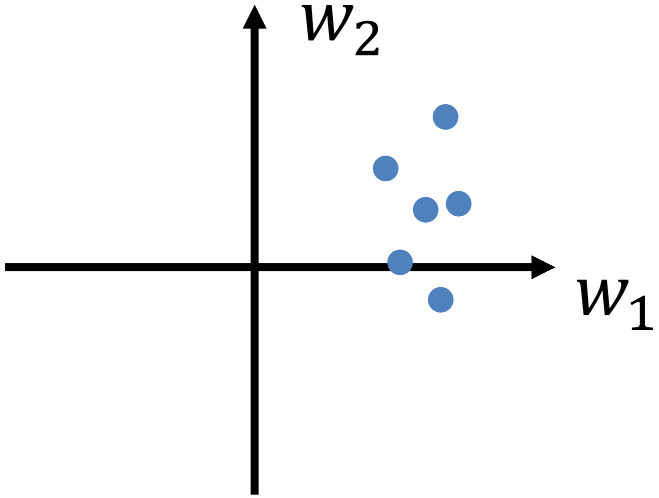


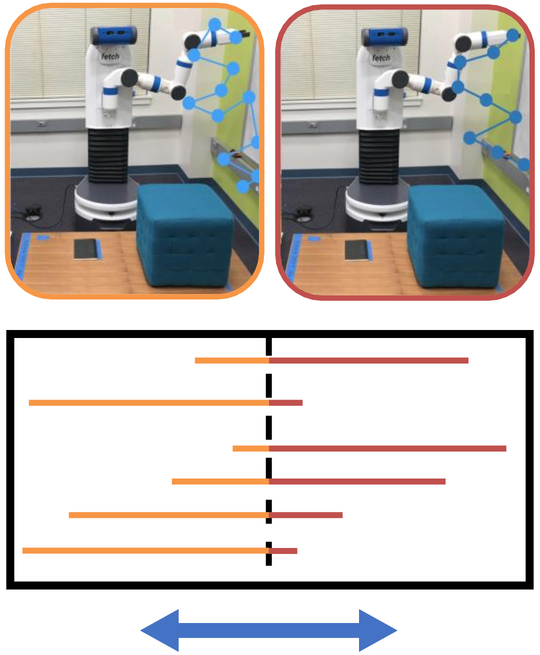
Active Preference-Based Learning of Reward Functions
Sadigh et al. RSS 2017

Maximum volume removal

Posterior $P(w \mid \boldsymbol{C})$

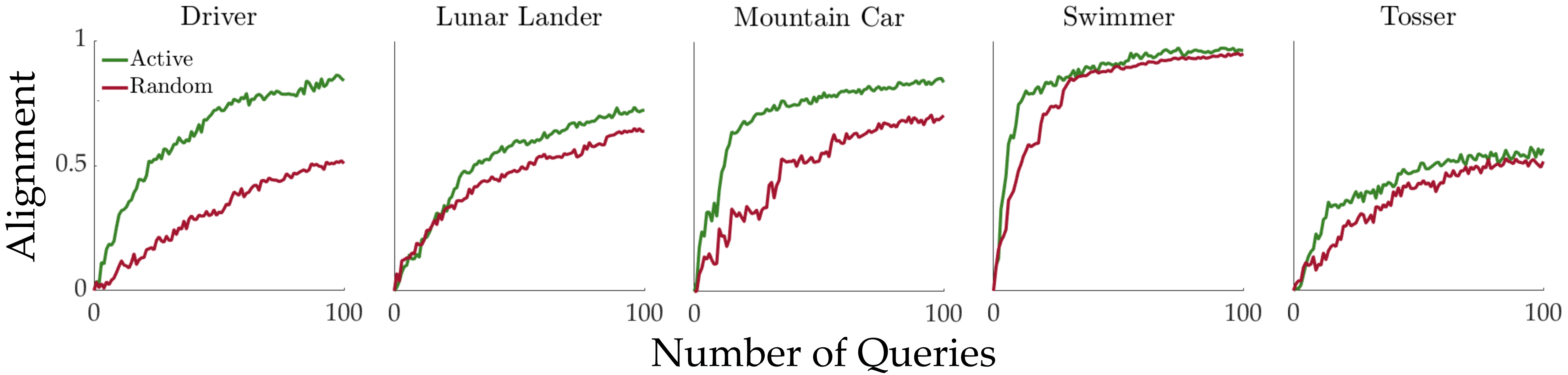


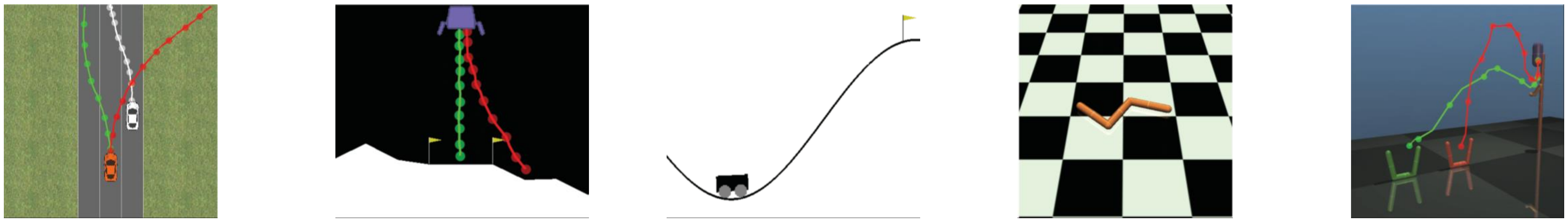




Active Preference-Based Learning of Reward Functions
Sadigh et al. RSS 2017

Active vs. random querying

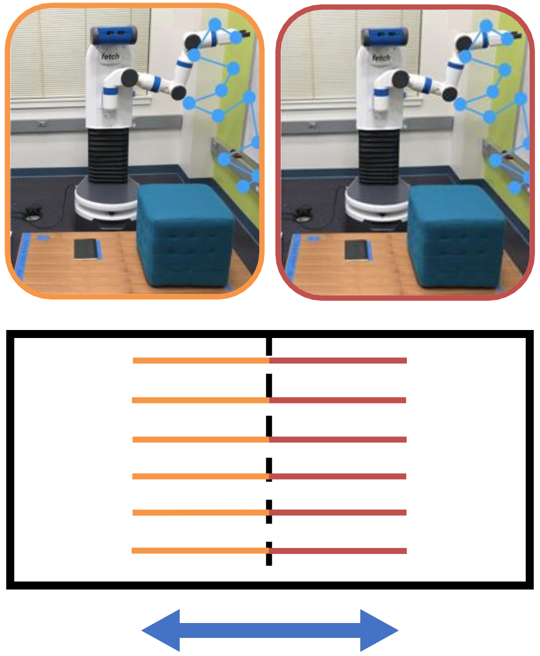


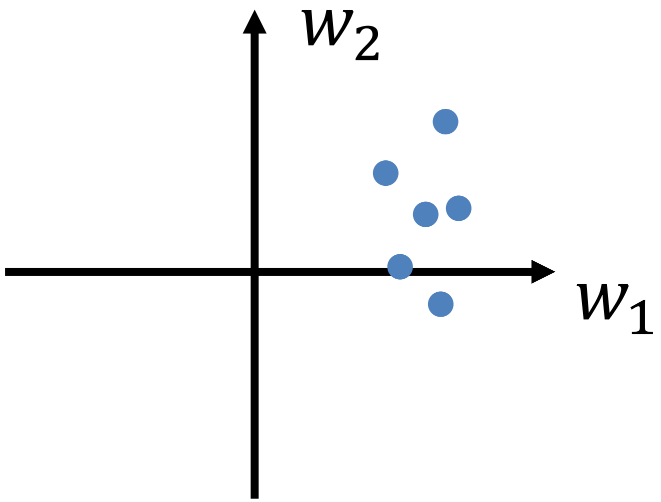


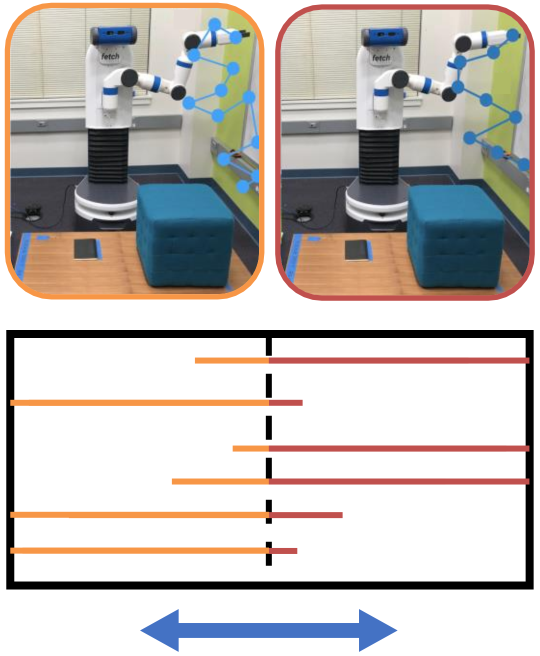
Batch Active Preference-Based Learning of Reward Functions
Bıyık and Sadigh, CoRL 2018

Maximum volume removal

Posterior $P(w \mid \boldsymbol{c})$







Active Preference-Based Learning of Reward Functions
Sadigh et al. RSS 2017

Mutual information maximization

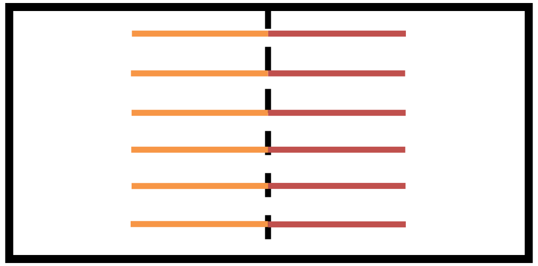
$$\max_{\xi_A,\xi_B} I(\underbrace{q}_{\text{User response}} ; w \mid \underbrace{\boldsymbol{C}}_{\text{Dataset}}, \underbrace{\xi_A, \xi_B}_{\text{Query}})$$

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

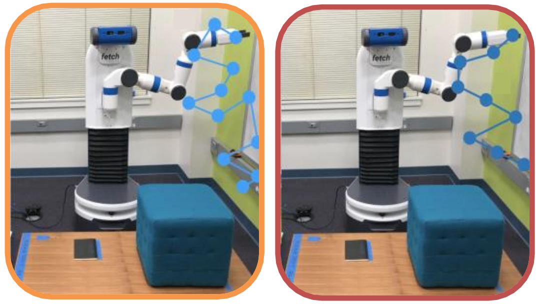
Mutual information maximization

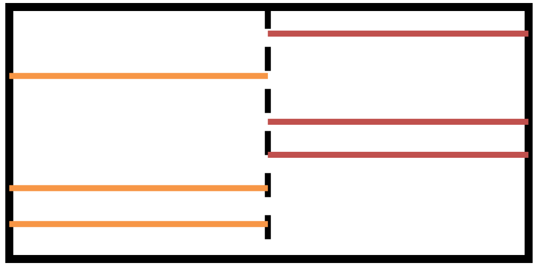
$$\max_{\xi_A,\xi_B} I(\,q\,;w \mid \boldsymbol{\mathcal{C}}, \xi_A, \xi_B\,)$$

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



User Choice





User Choice

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization

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

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

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

No *w* here!

Mutual information maximization

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

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

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

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization

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

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

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

Mutual information maximization

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

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

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

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

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

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization

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

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

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

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

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

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

Mutual information maximization

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

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

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

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

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization

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

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

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

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

Mutual information maximization

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

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

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

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

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

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

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization

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

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

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

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

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

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization

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

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

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

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

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

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization

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

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

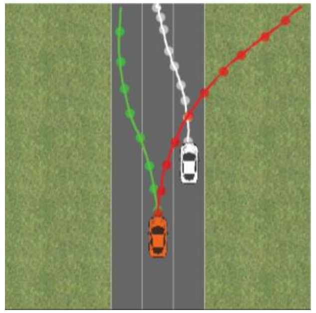
$$\max_{\xi_A,\xi_B} \mathbb{E}_{q,w \mid \boldsymbol{\mathcal{C}},\xi_A,\xi_B}[\log P(q \mid \xi_A, \xi_B, w) - \log P(q \mid \boldsymbol{\mathcal{C}}, \xi_A, \xi_B)]$$

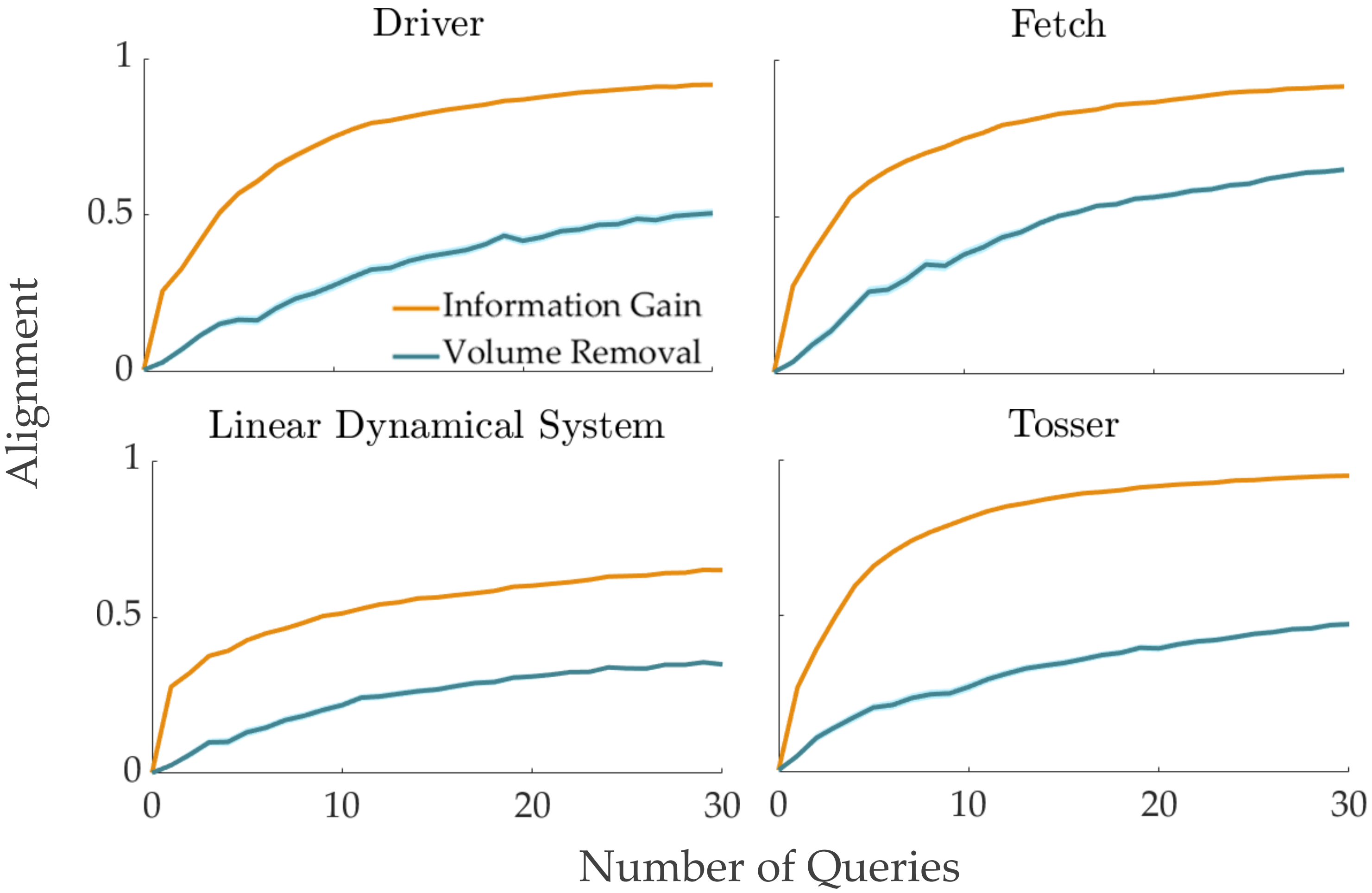
$$\max_{\xi_A,\xi_B} \mathbb{E}_{q,w \mid \xi_A,\xi_B}\left[\log \frac{P(q \mid \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w')}\right]$$

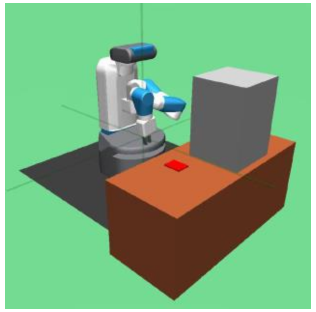
$$\max_{\xi_A,\xi_B} \sum_{w \in \Omega} \sum_{q} P(q \mid \xi_A, \xi_B, w)\left[\log \frac{P(q \mid \xi_A, \xi_B, w)}{\sum_{w' \in \Omega} P(q \mid \xi_A, \xi_B, w')}\right]$$

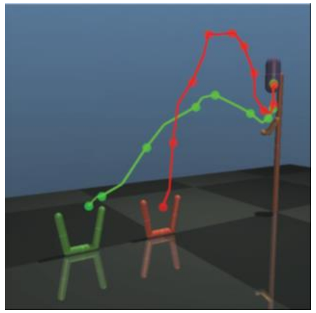
Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization





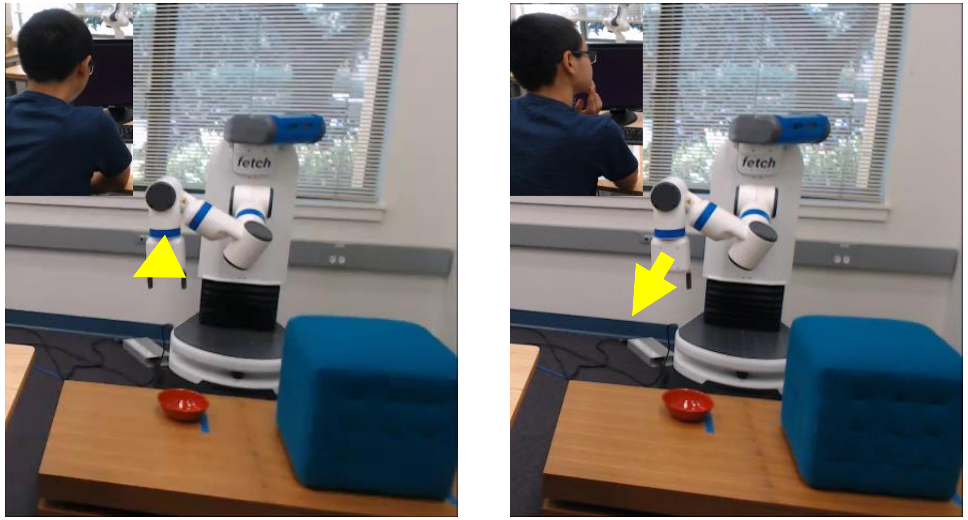




Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

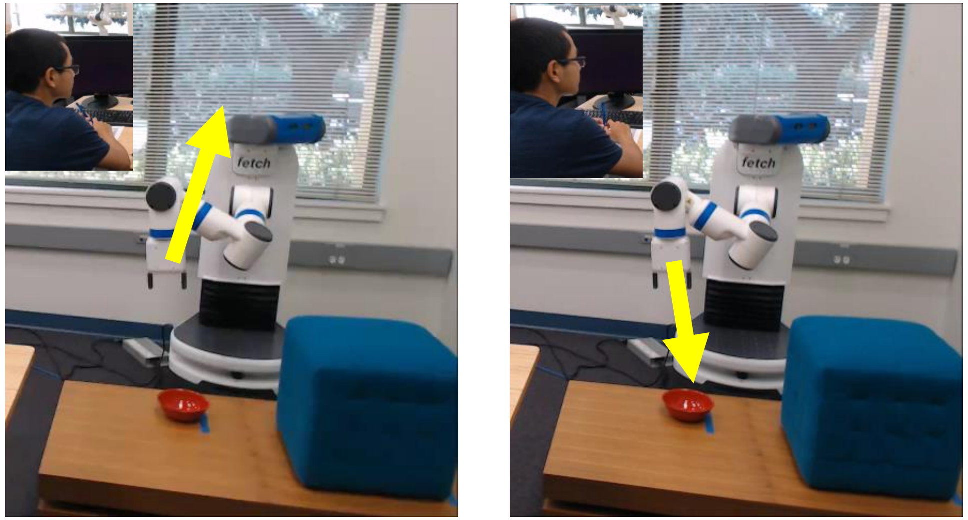
Volume Removal

Similar Trajectories



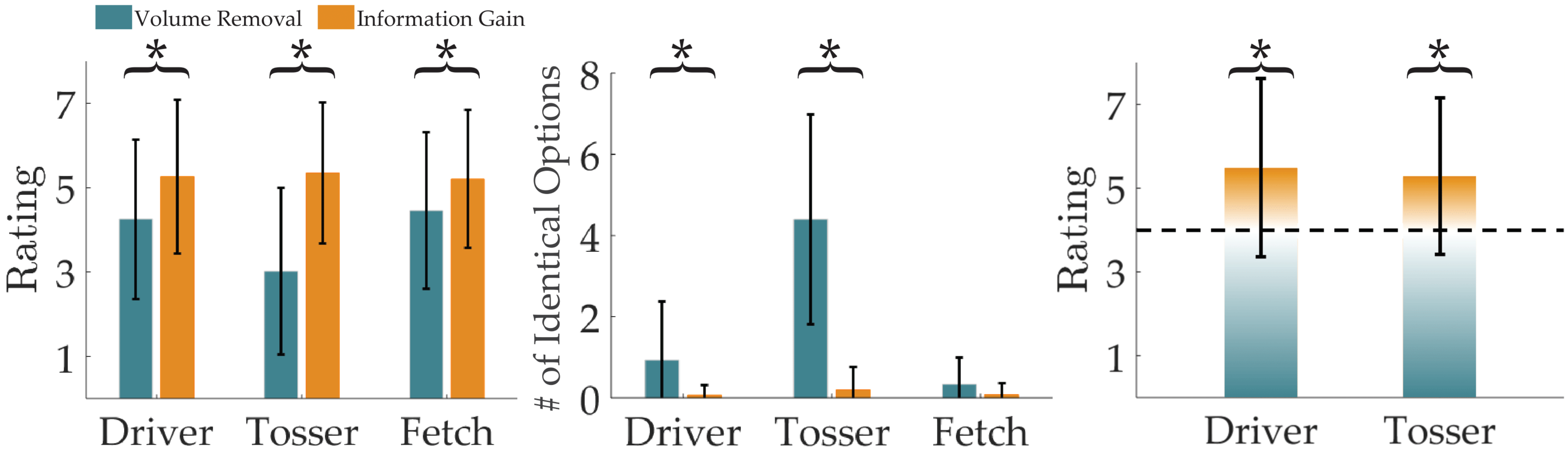
Information Gain

More Distinguishable Query



Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization



Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Other types of human feedback

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

Reward-rational (implicit) choice: A unifying formalism for reward learning
Jeon et al. NeurIPS 2020

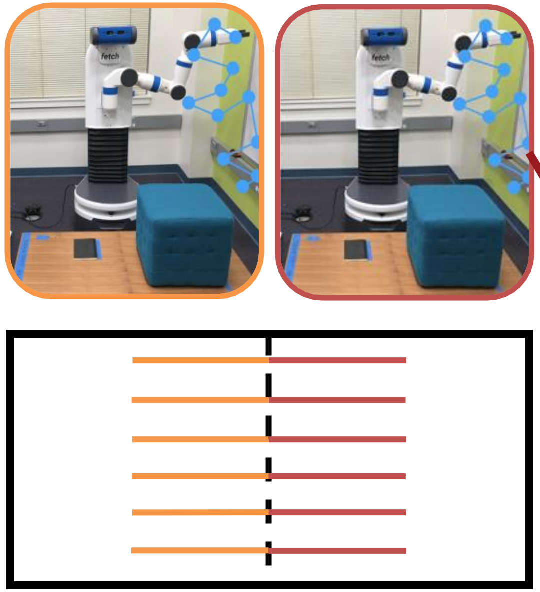
Today…

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

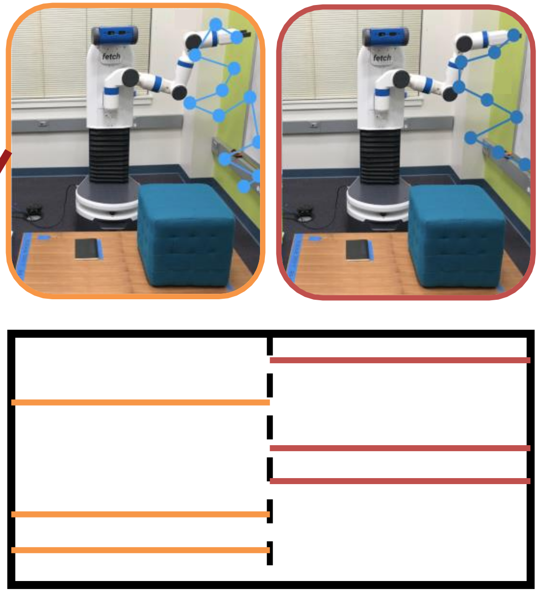
Mutual information maximization

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

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



User Choice



User Choice

Where do these trajectories come in the first place?

Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Incorporating comparisons

How do we solve this optimization problem?

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

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

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

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

Learning Reward Functions from Diverse Sources of Human Feedback: Optimally Integrating Demonstrations and Preferences
Bıyık et al. IJRR 2021.

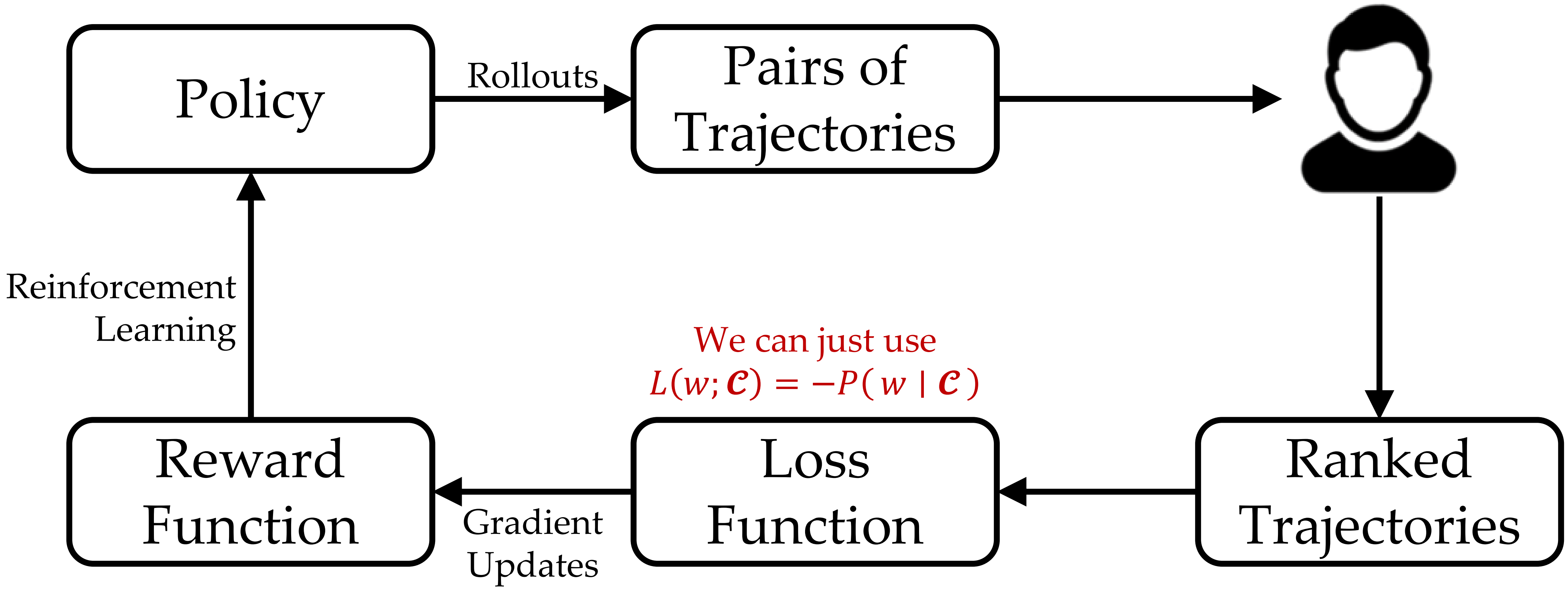
RLHF

Two major changes to preference-based reward learning:

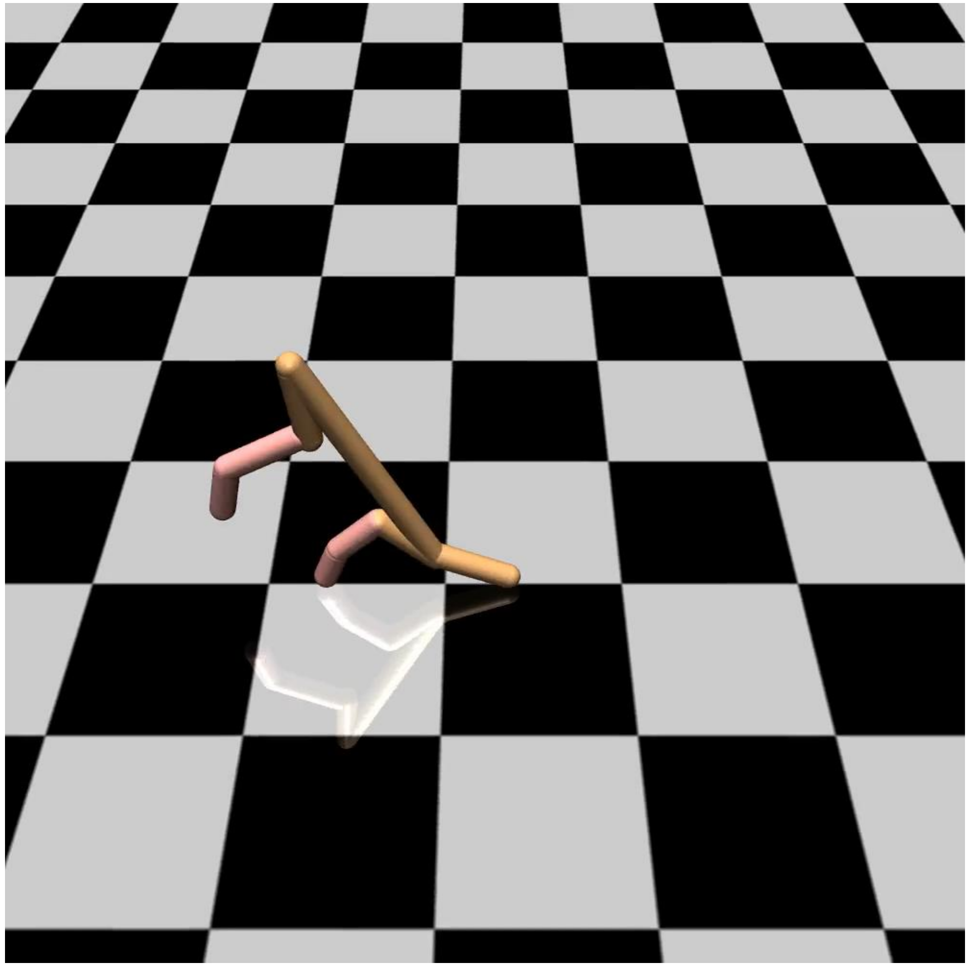
1. Instead of Bayesian learning, write a loss function and learn with gradient updates

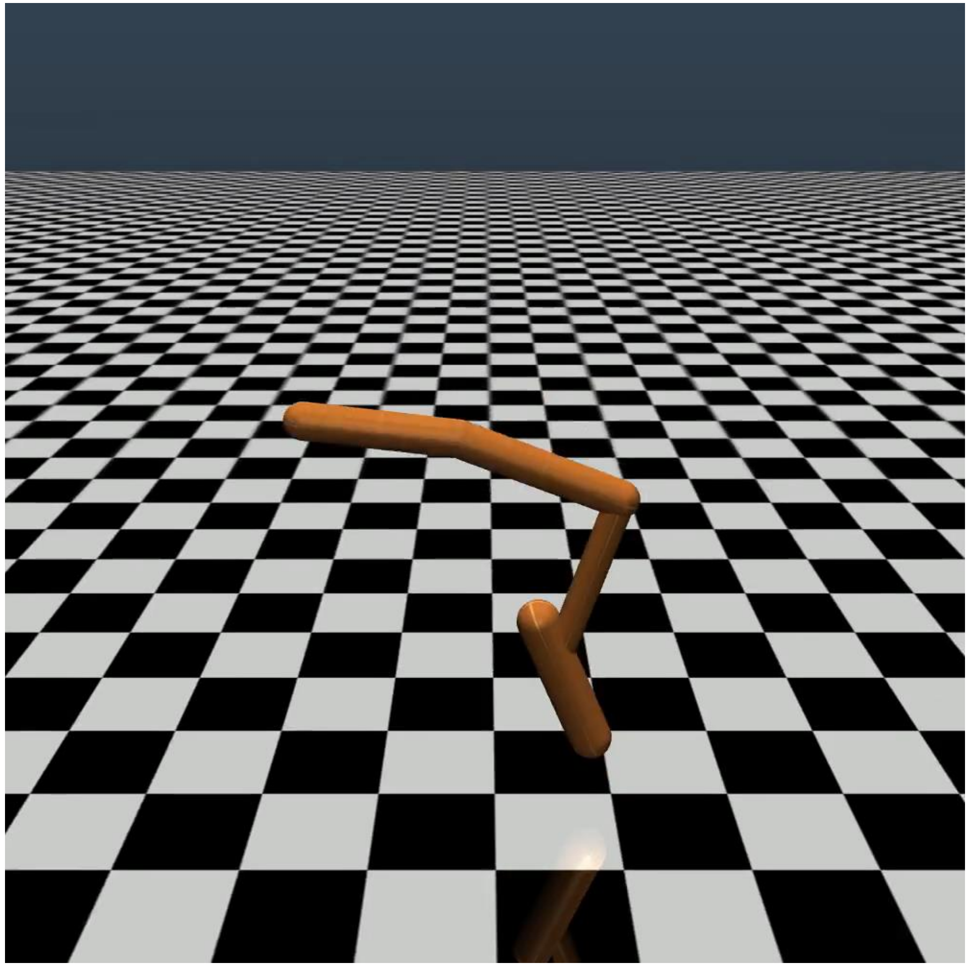
2. After learning a reward, train a policy to generate new trajectories for the next iteration of reward learning

RLHF


Policy
Rollouts
Pairs of Trajectories
Reinforcement Learning
We can just use
$L(w; \boldsymbol{C}) = -P(w \mid \boldsymbol{C})$
Reward Function
Gradient Updates
Loss Function
Ranked Trajectories

RLHF

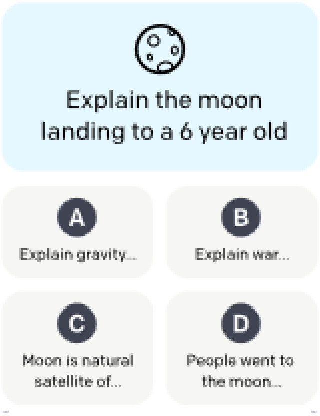




Deep reinforcement learning from human preferences
Christiano et al., NeurIPS 2017

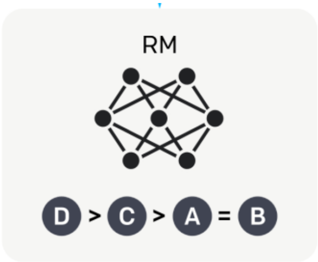
InstructGPT

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.



Training language models to follow instructions with human feedback
Ouyang et al., NeurIPS 2022

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).