Learning from Human Feedback: From Demonstrations to Comparative Language

Erdem Bıyık

Assistant Professor
Thomas Lord Department of Computer Science
Ming Hsieh Department of Electrical and Computer Engineering

USC University of Southern California

LiraLab

Machine Learning Solution to Robotics

Tons of text

(OpenAI 2023)

Language model that produces human-like texts

14,000,000 images

(Deng et al. 2009)

Image recognition models at human-level proficiency

44,000,000 chess games

(Silver et al. 2017)

Super-human chess engines

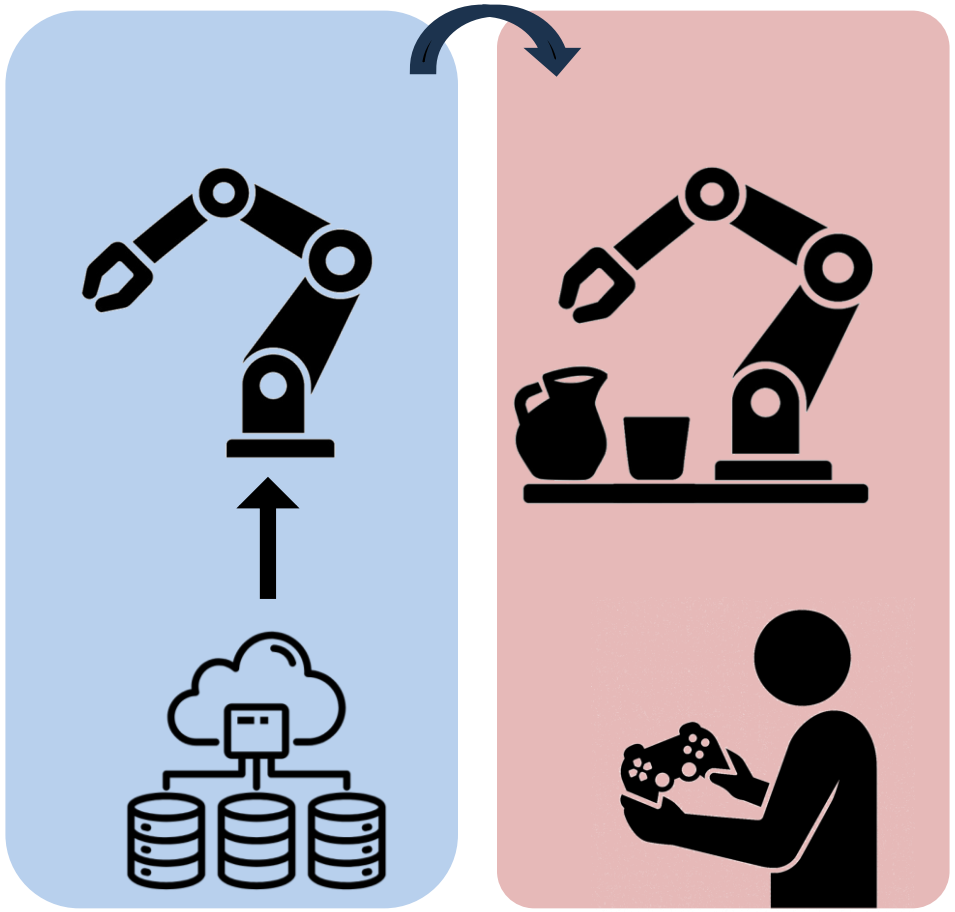
We do not have large datasets in robotics

Pastor, 2016

Comparing datasets

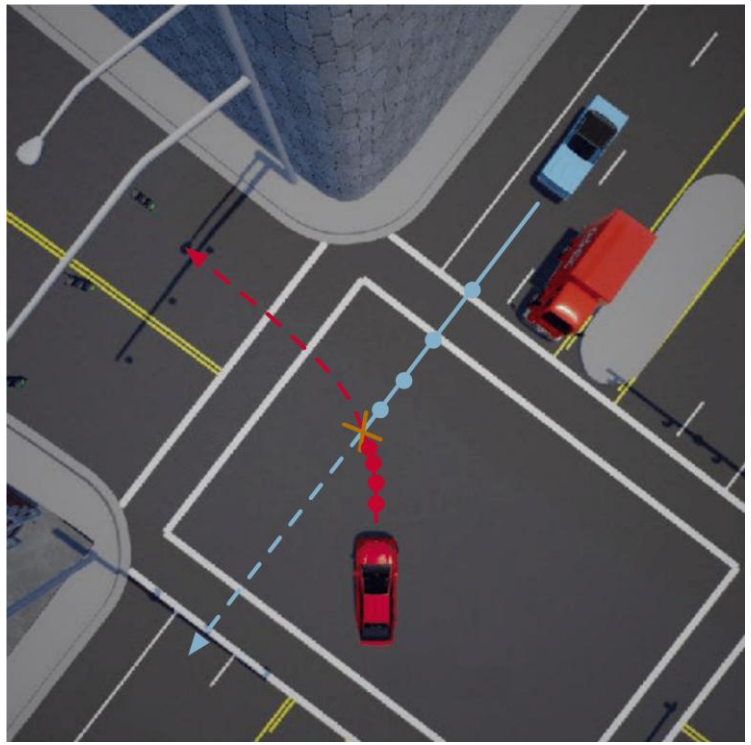
Assuming 238 words/minute, 1.33 tokens/word

Image Courtesy: Kevin Black

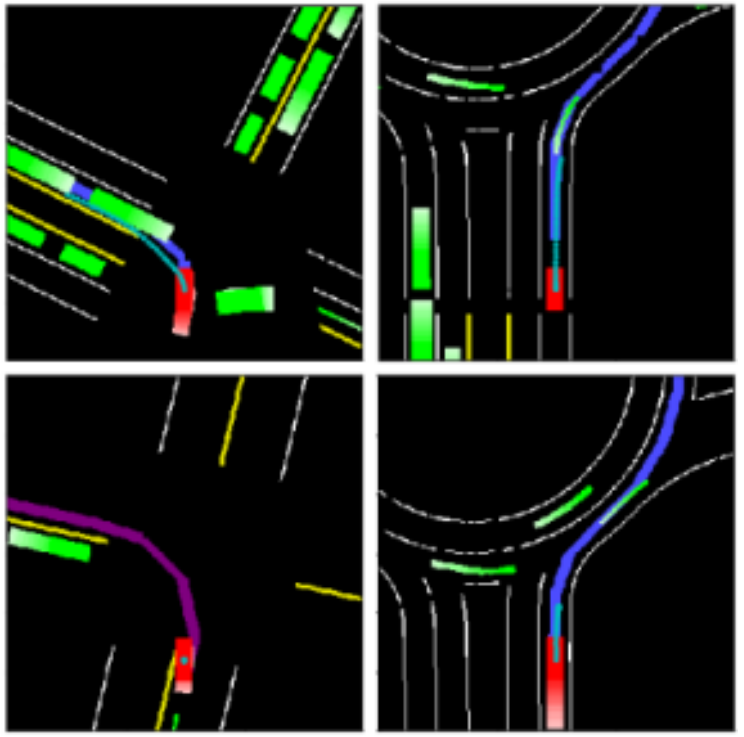


Learning from demonstrations (LfD)

Codevilla et al. ICRA‘18



Cao et al. RSS‘20



Chen et al. IROS‘19

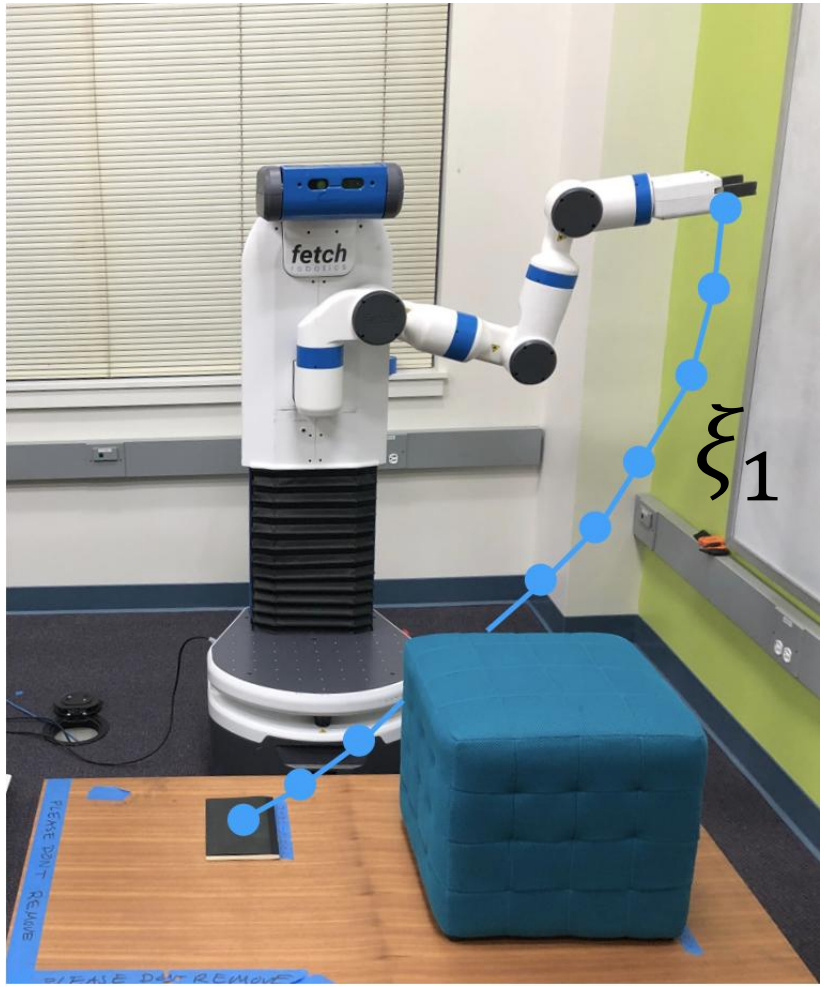
Why does LfD fail?

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

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

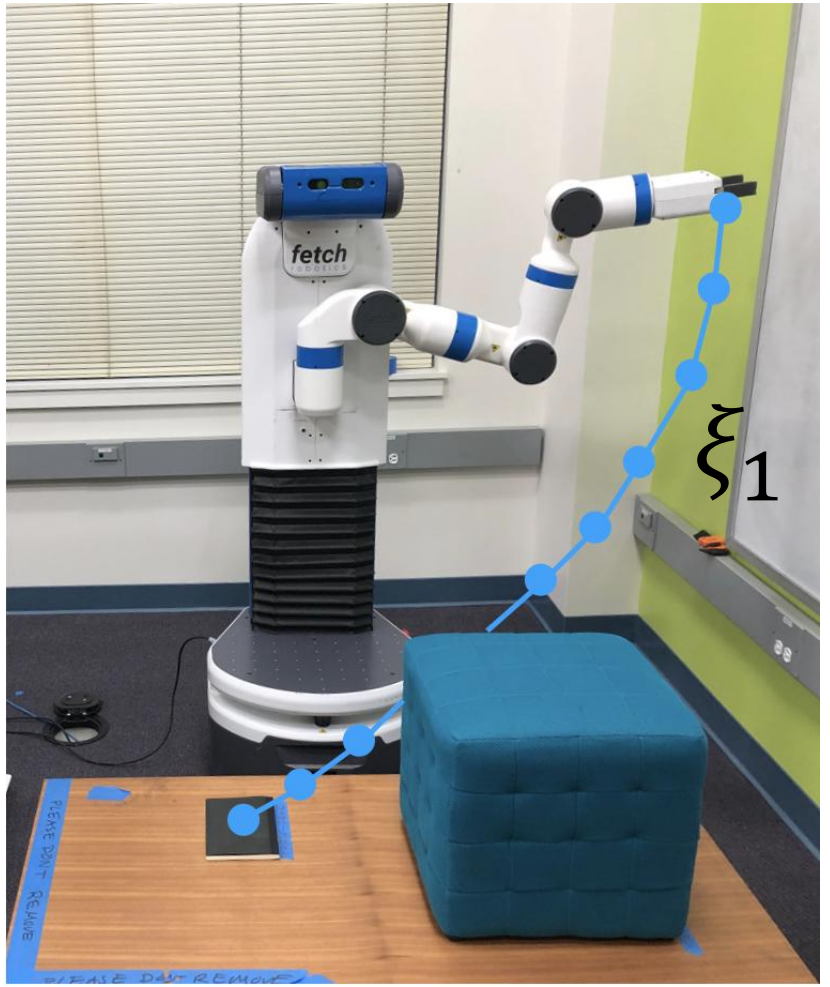
$$\underset{w}{\operatorname{argmax}}\, P(w \mid \mathcal{D})$$

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

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

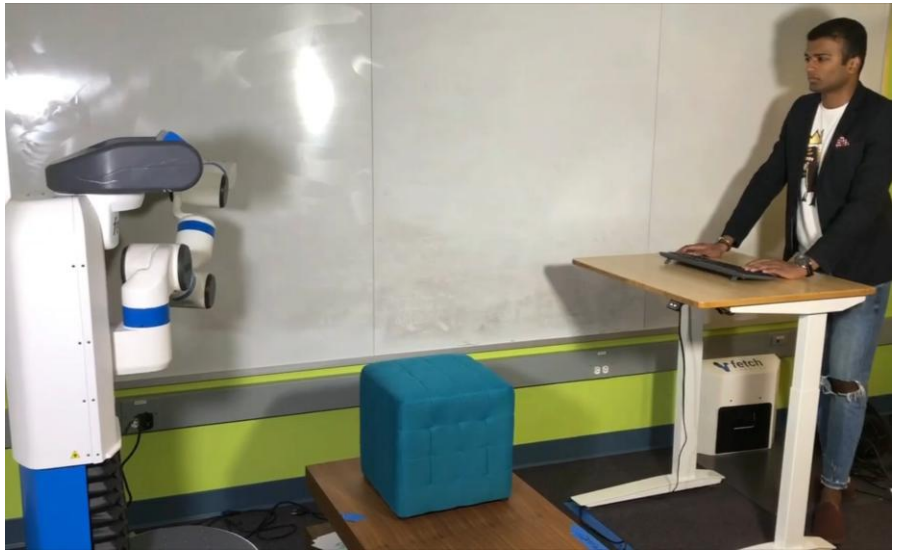
$$\propto P(w) \prod_{i=1}^{L} \exp f_w(\xi_i)$$

(Noisy humans)



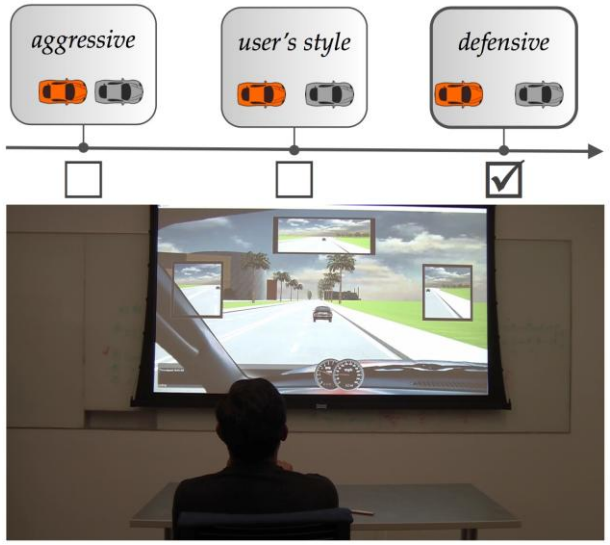
Humans are Suboptimal

Robots with high degrees of freedom are hard to teleoperate.



Palan et al. RSS‘19

Humans do not like their own demonstrations.

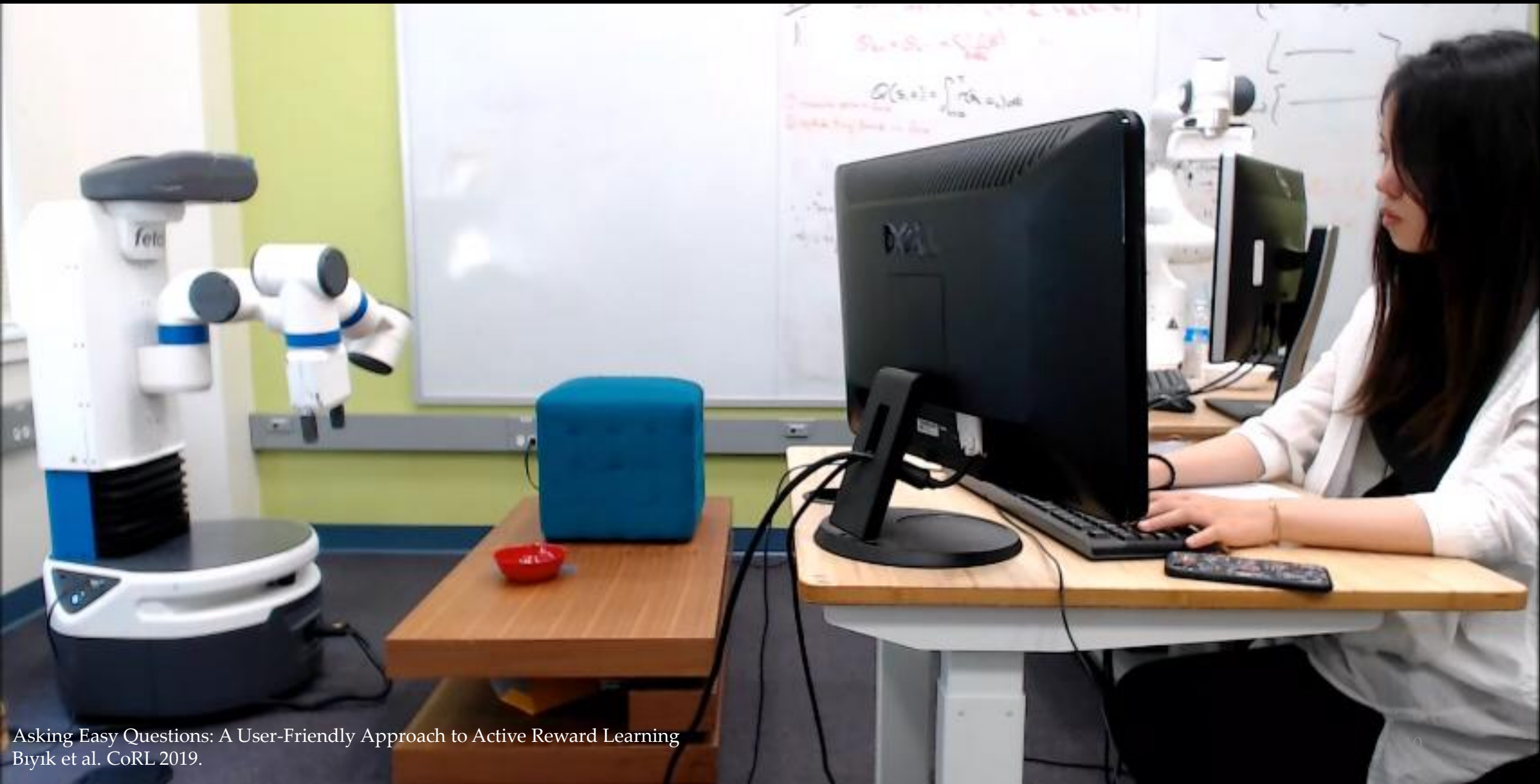


Basu et al. HRI‘17

Humans take suboptimal actions in risky situations.

Kwon et al. HRI‘20

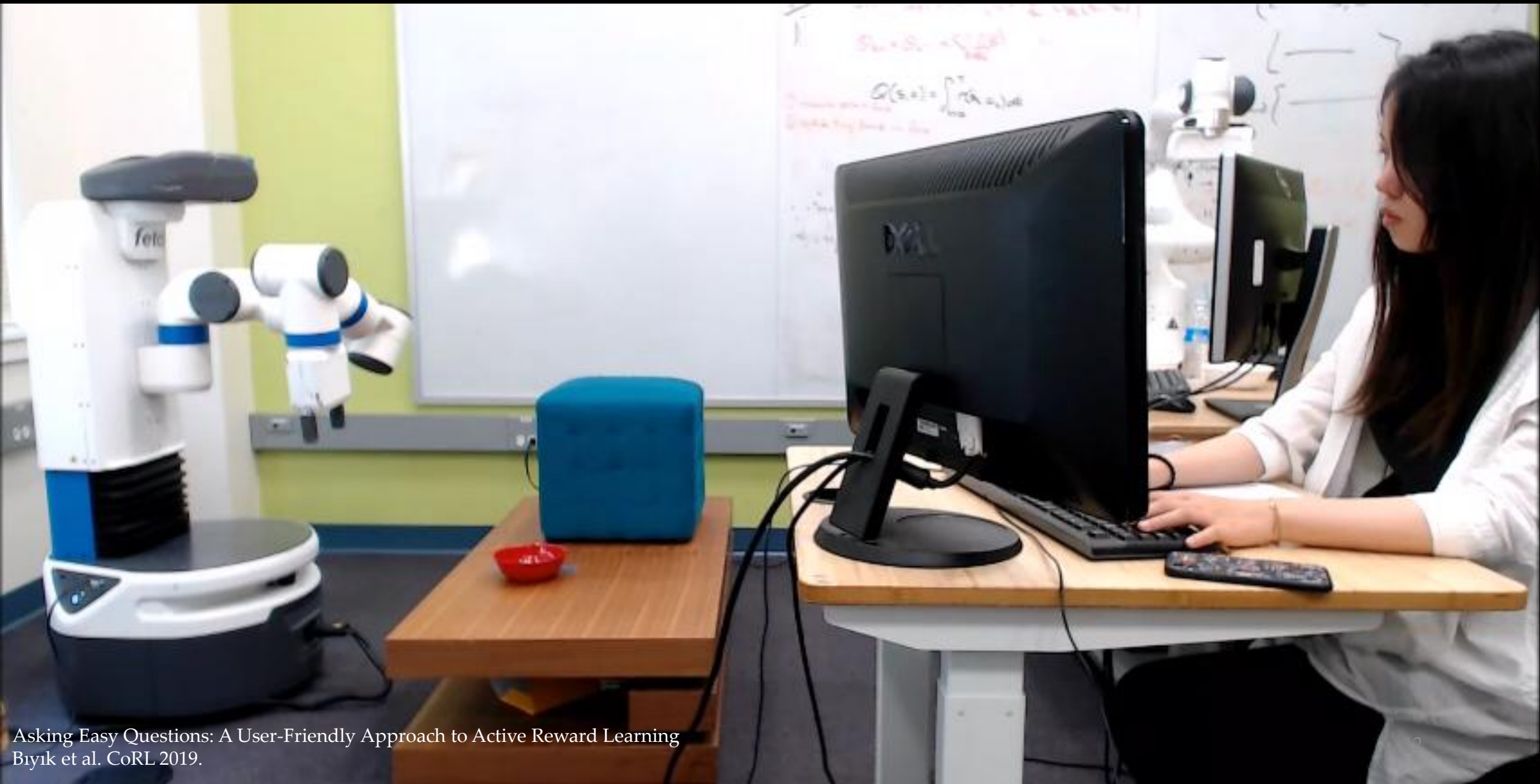
We can let the human evaluate a robot demonstration



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

How dark is this blue?

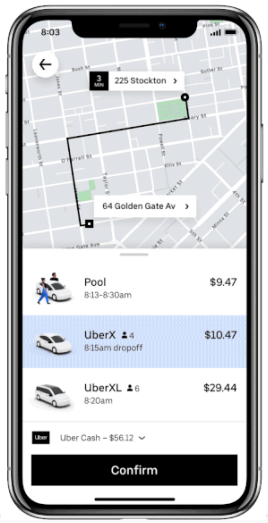
Human evaluations are often unreliable

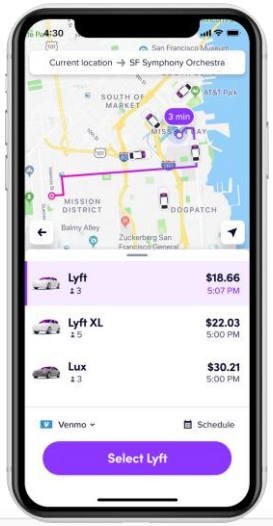


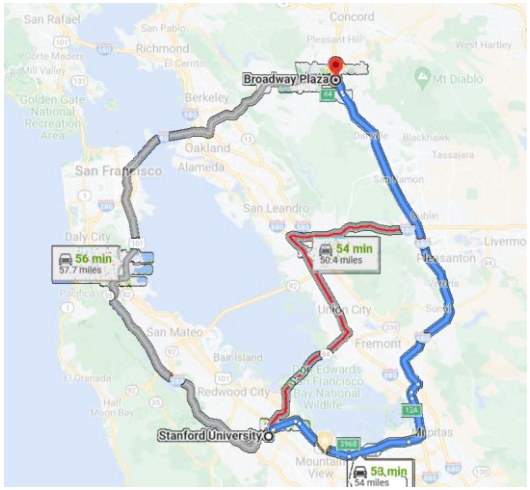
Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019.

Which blue is darker?

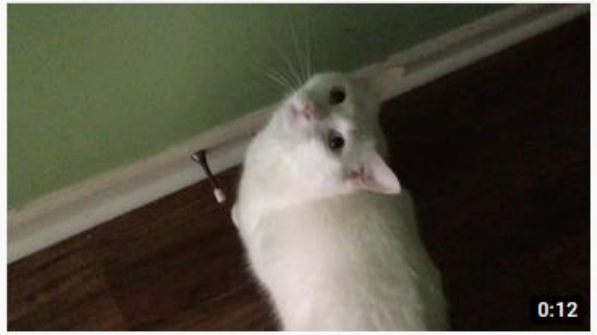
Comparison data







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

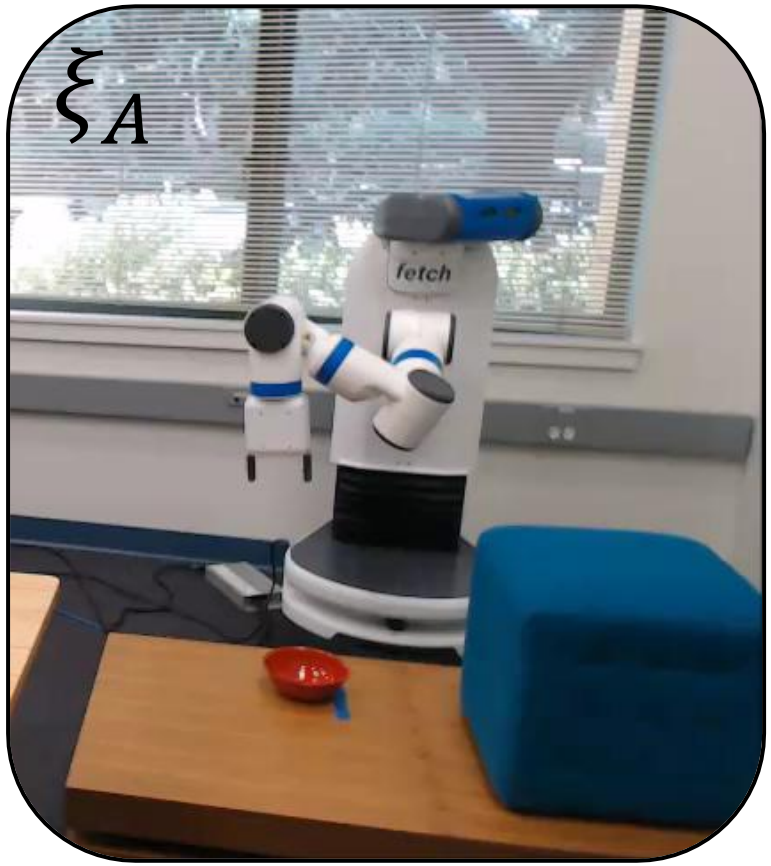
John Lowe 9-dart finish FIRST EVER ON TV
Unicorn Darts
6.6M views • 4 years ago

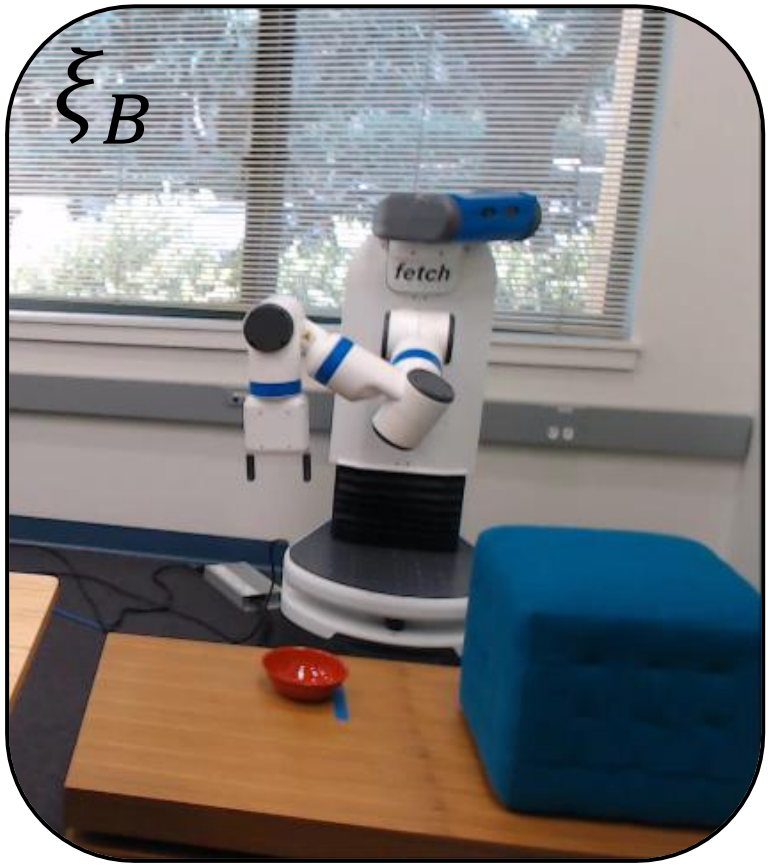
Carlsen - Nepomniachtchi | Game 8 | World Chess Championship | Howell,...
chess24
24K watching

Today…

- Learning from human feedback
  - Pairwise comparisons
  - Reinforcement learning from human feedback (RLHF)
  - Comparative language feedback

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

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

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

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

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

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

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

Incorporating Comparisons

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

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

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

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

Incorporating Comparisons

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

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

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

How do we compute this?

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

Luce's choice axiom

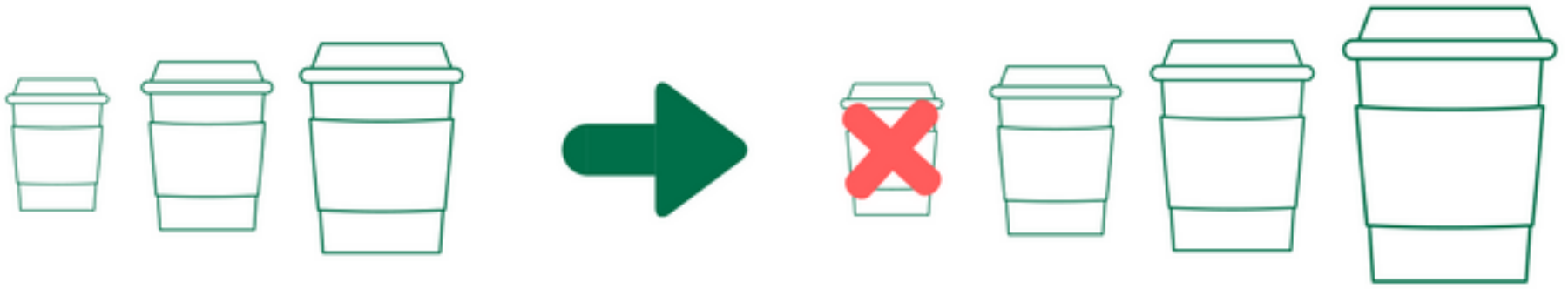
The probability of selecting one item over another from a pool of many items is not affected by the presence or absence of other items in the pool.

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"*



From: Kent Hendricks

Counterexamples for fun

- Coca Cola vs. Pepsi

1985 (Spring)

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

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

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

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

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

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

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

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

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

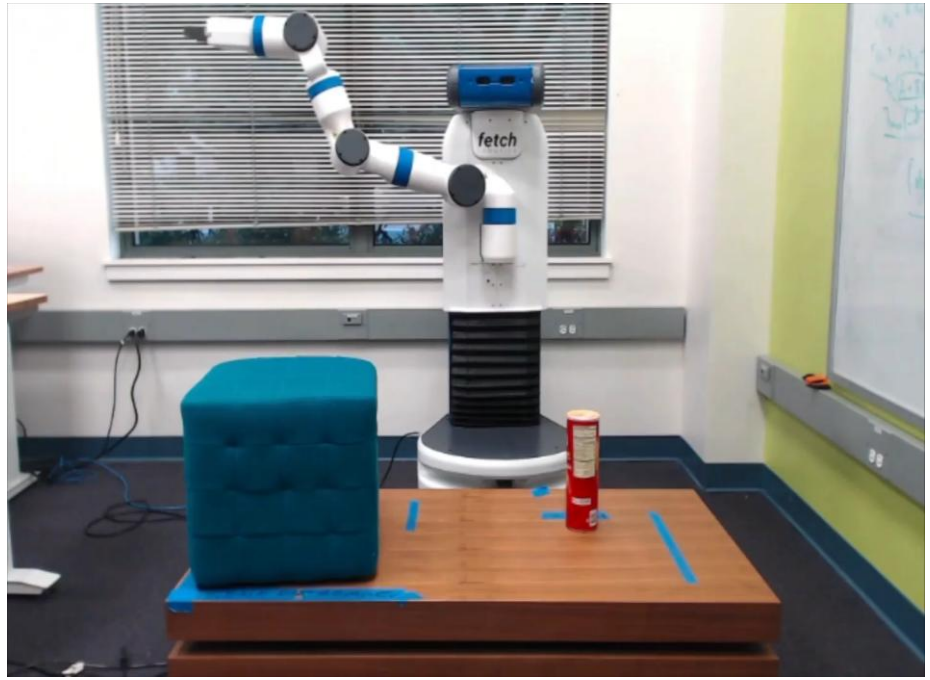
$$= P(w) \prod_{i=1}^{L} P(\xi_i \mid w) \prod_{i=1}^{N} P\left(q^{(i)} \mid w, \xi_A^{(i)}, \xi_B^{(i)}\right)$$

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

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

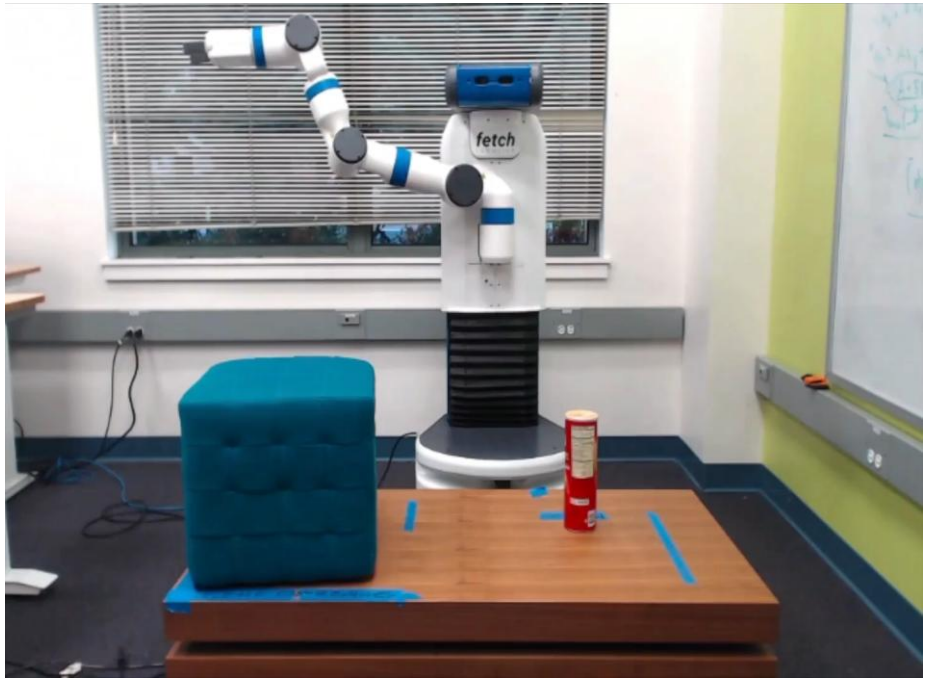
Benefit of comparisons

Bayesian IRL



5 demonstrations

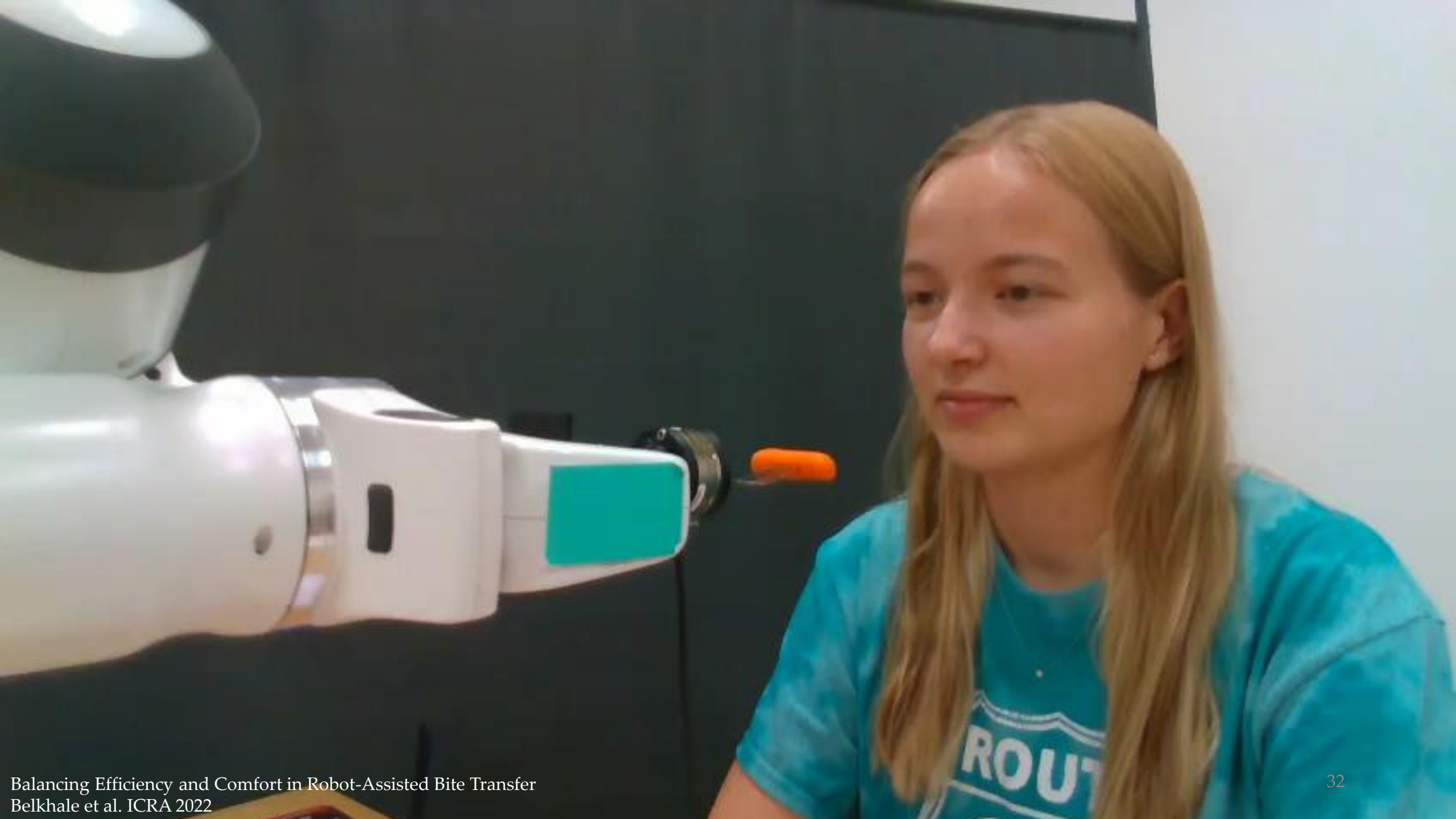
Ours



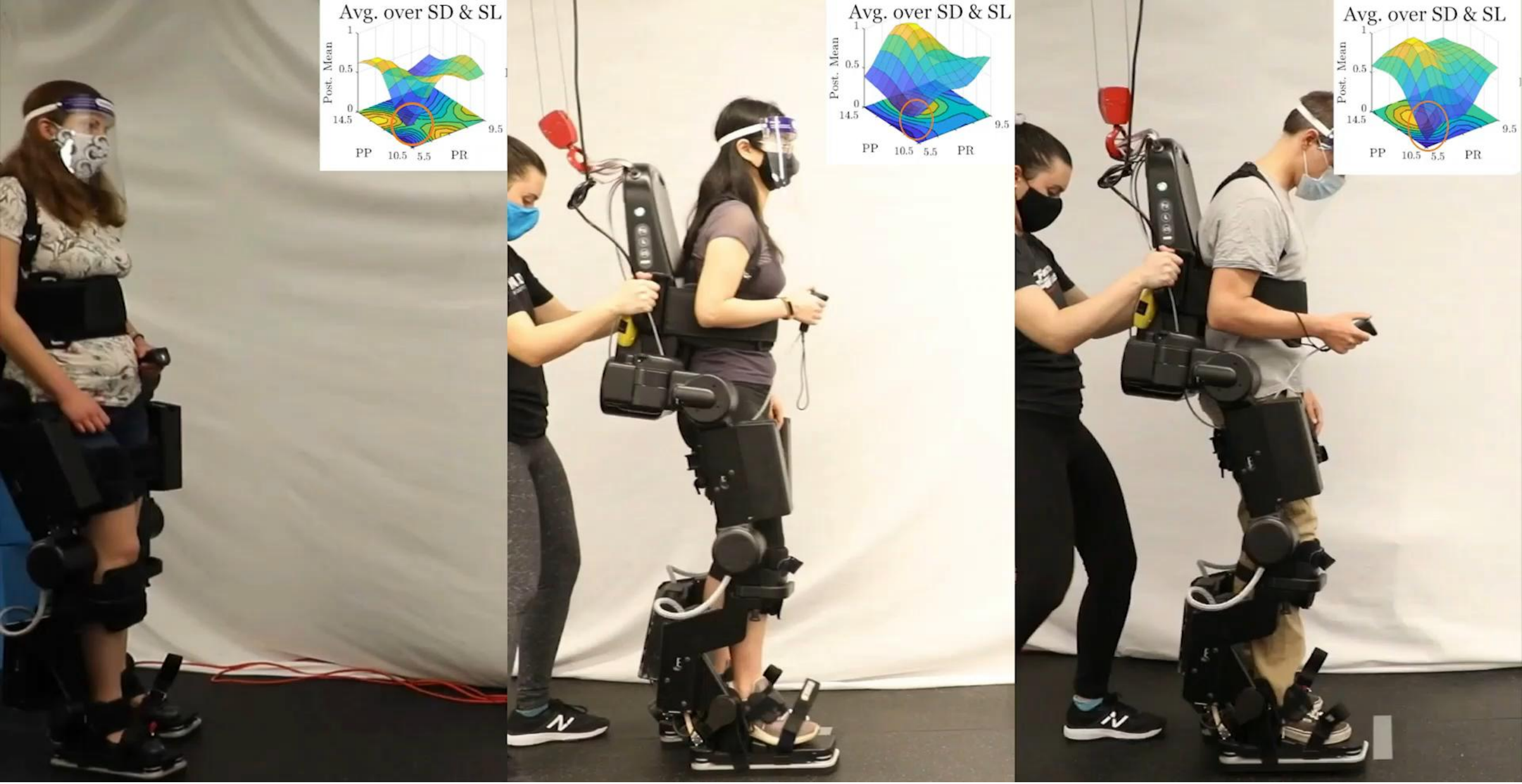
1 demonstration + 15 comparisons

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.



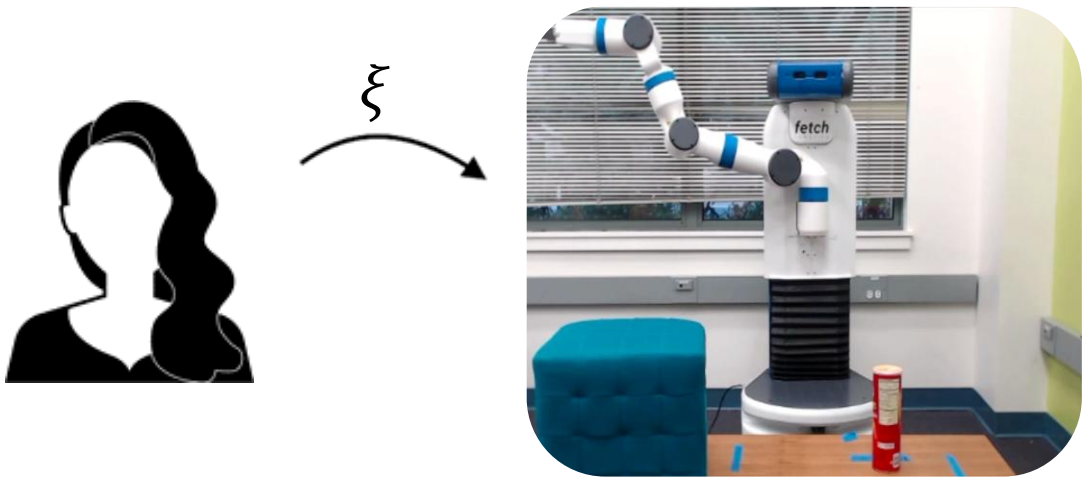
Balancing Efficiency and Comfort in Robot-Assisted Bite Transfer
Belkhale et al. ICRA 2022



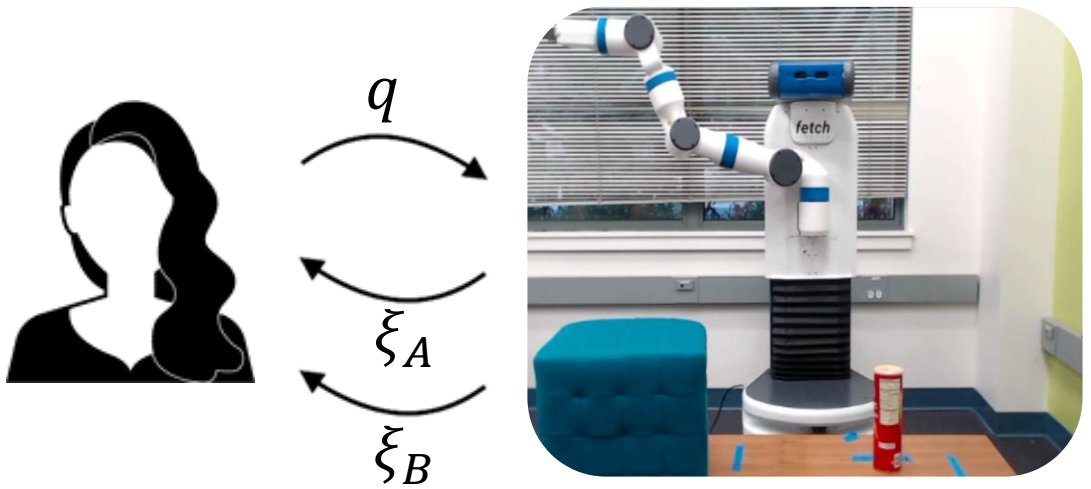
ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes
Li et al., ICRA 2021

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

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

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

Entropy (a measure of uncertainty)

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

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

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

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

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

Another example

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

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

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

Mutual information

Uncertainty

$H(X) = 1$

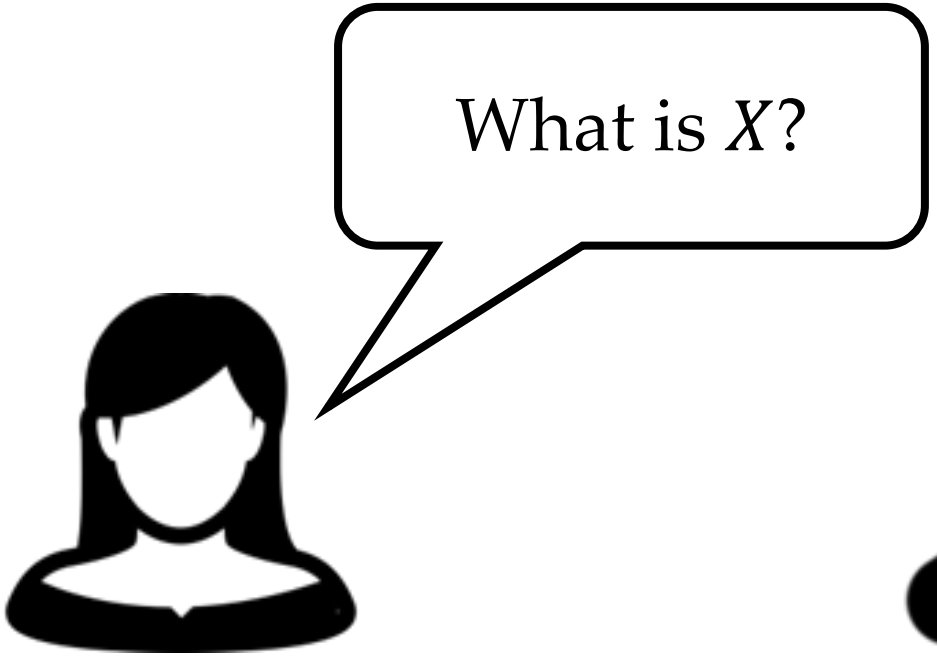
Alice

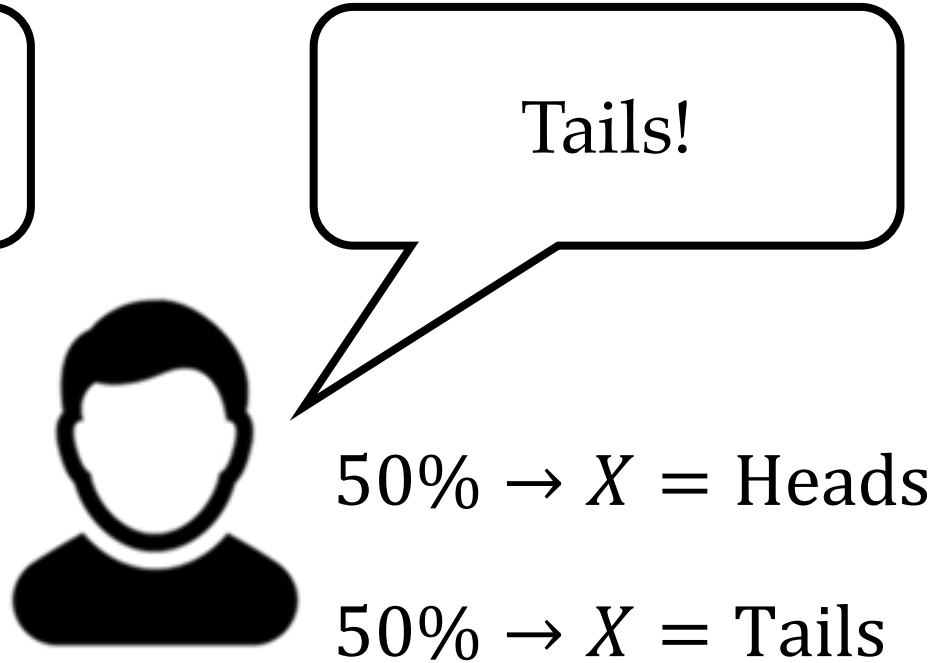
Bob

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

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

Mutual information





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

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

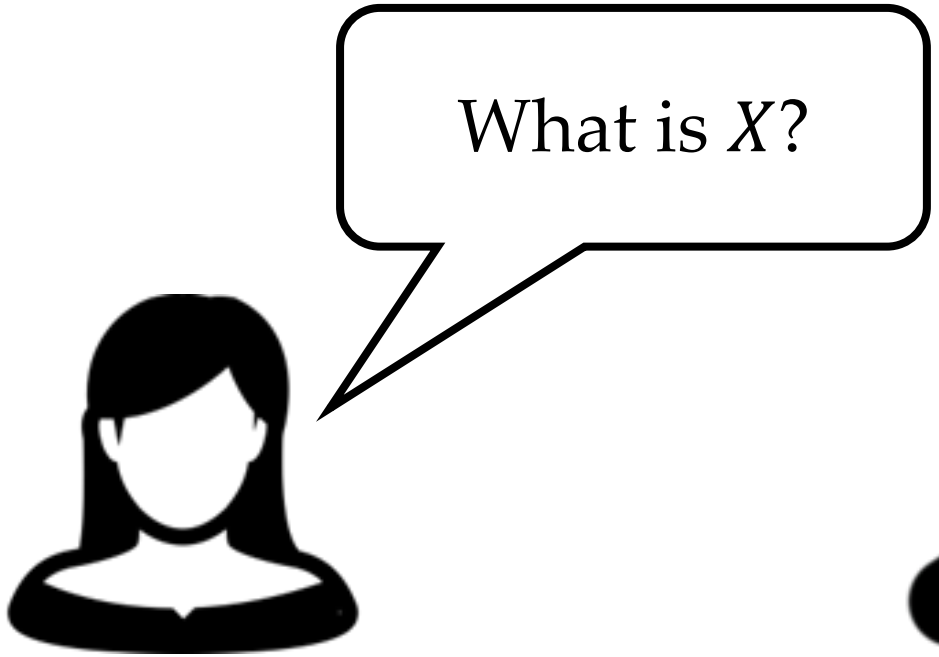
Uncertainty

$H(X \mid X) = 0$

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

This is 0 in information theory.

Mutual information



Uncertainty

$H(X \mid X) = 0$

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

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

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

$= 1 - 0 = 1$

Mutual information

Uncertainty

$H(X) = 1$

Alice

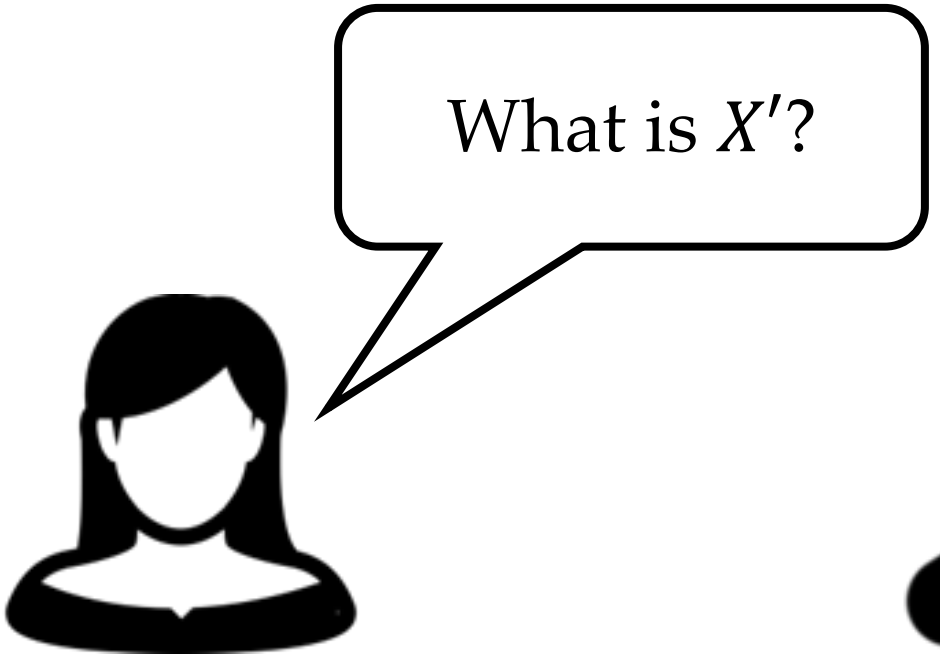
$50\% \rightarrow X = \text{Heads}$

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

Noisy Bob

(lies with 5% probability)

Mutual information



Alice

Noisy Bob
(lies with 5% probability)

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

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

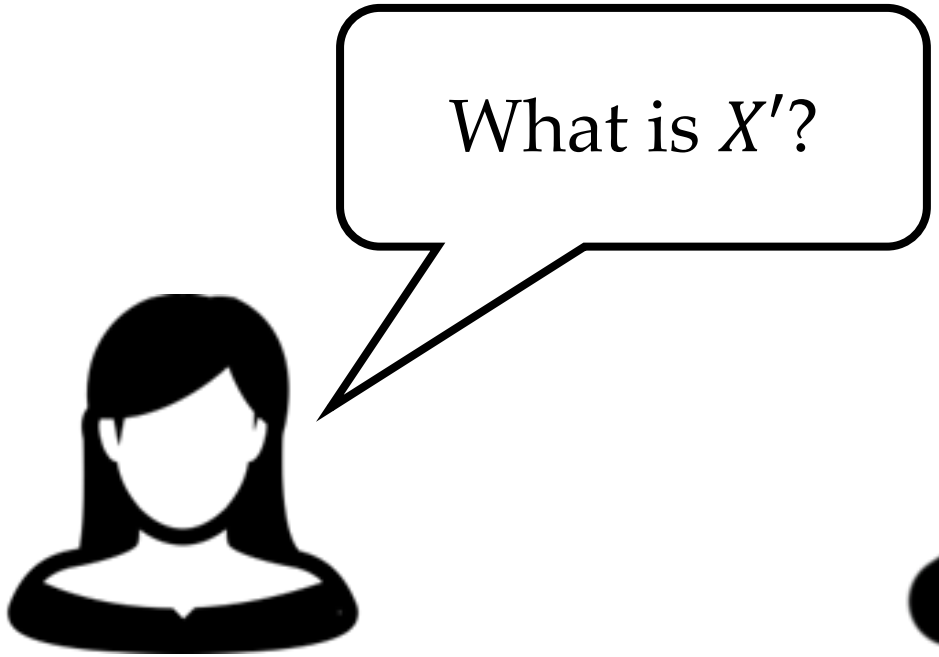
Uncertainty

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

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

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

Mutual information



Uncertainty

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

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

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

Alice

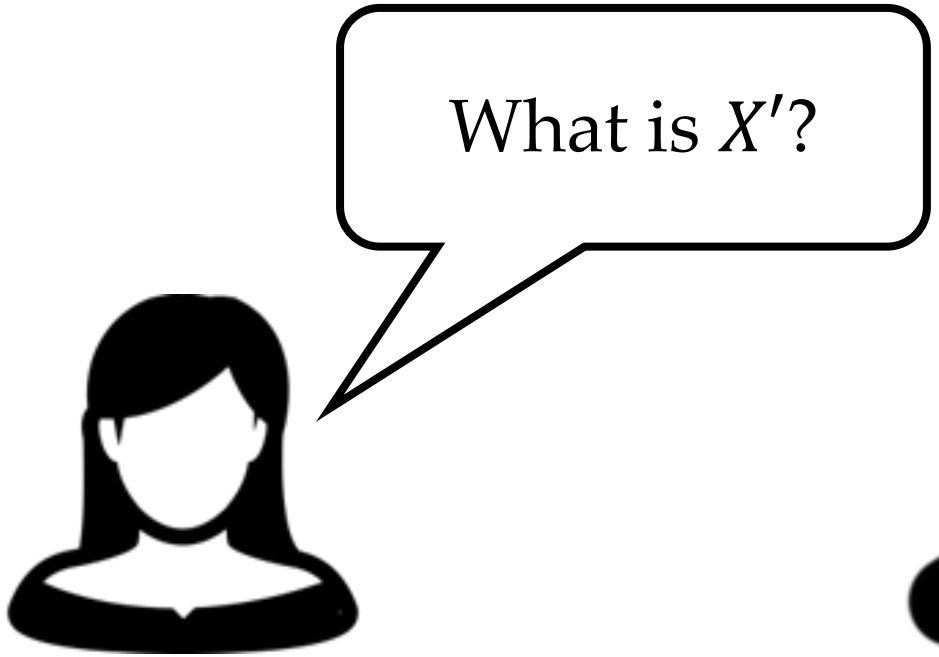
Noisy Bob

(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



Uncertainty

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

Alice

Noisy Bob
(lies with 5% probability)

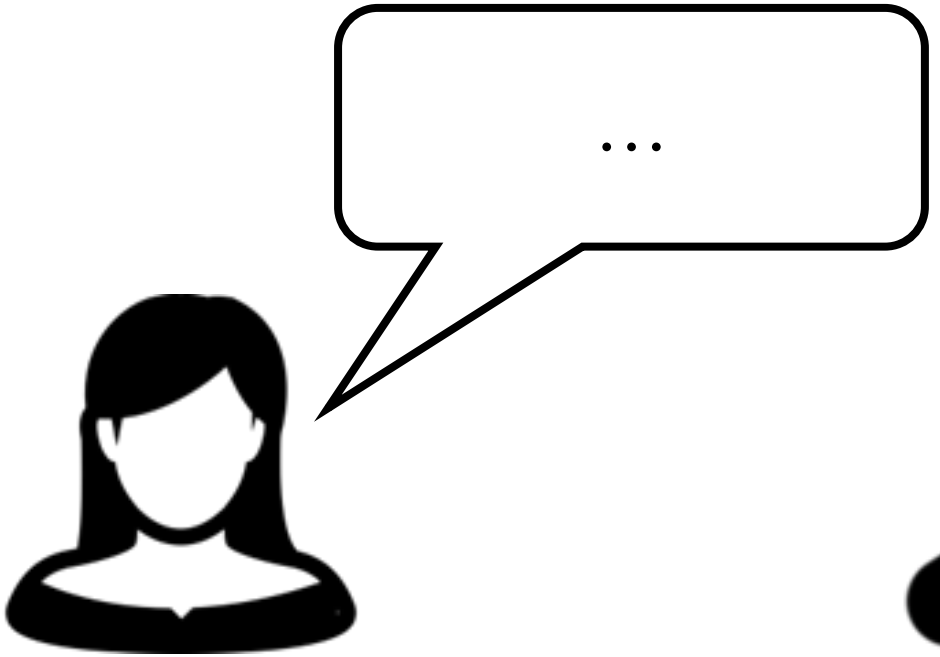
$50\% \rightarrow X = \text{Heads}$

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

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)

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

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

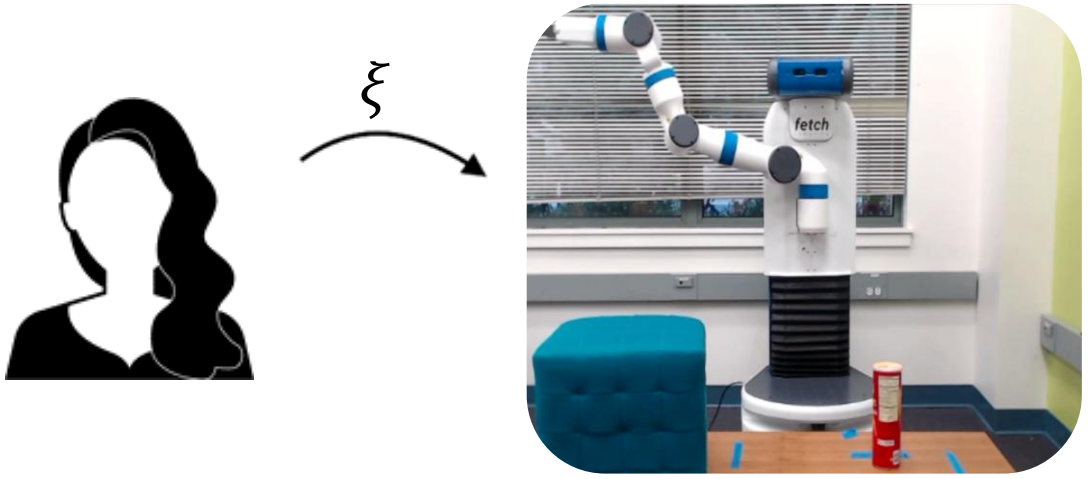
$20\% \rightarrow X = (\mathrm{H}, \mathrm{T}, \mathrm{H})$

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

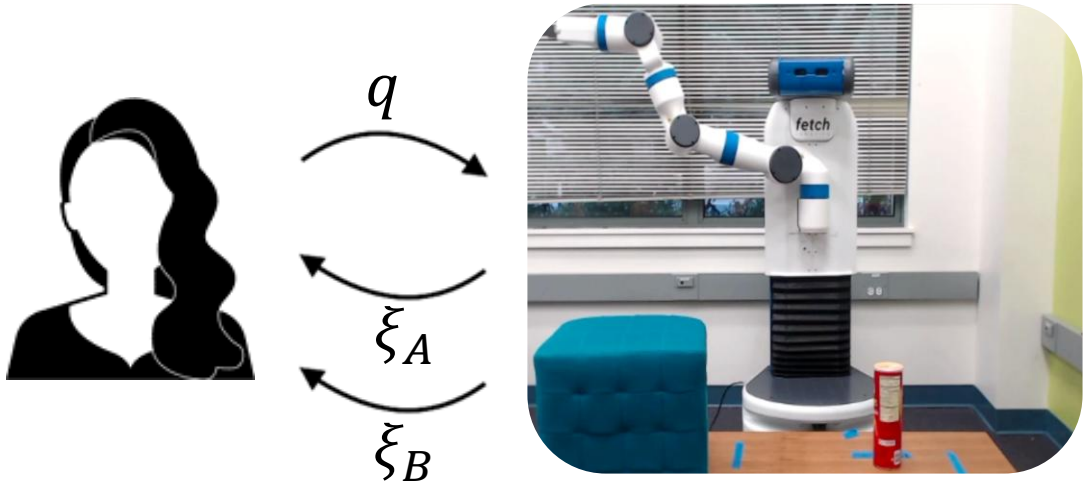
$20\% \rightarrow X = (\mathrm{T}, \mathrm{T}, \mathrm{T})$

Choosing queries

Demonstrations



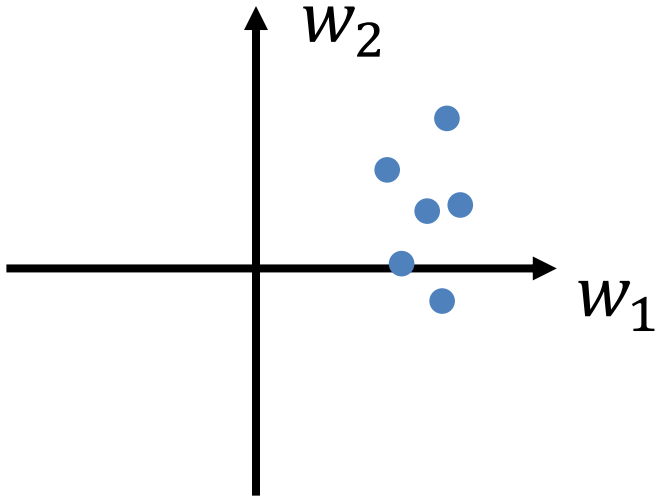
Comparisons



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

Maximum volume removal

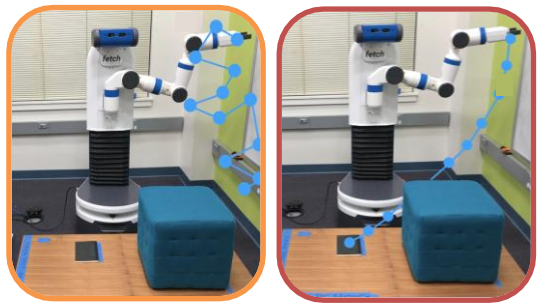
Posterior $P(w \mid \boldsymbol{\mathcal{C}})$

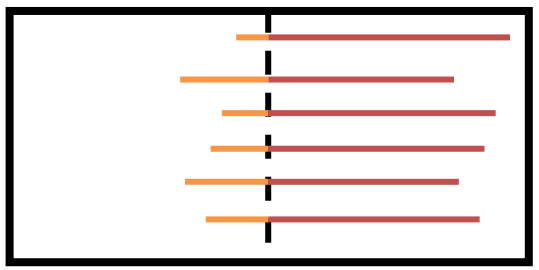


Active Preference-Based Learning of Reward Functions
Sadigh et al. RSS 2017

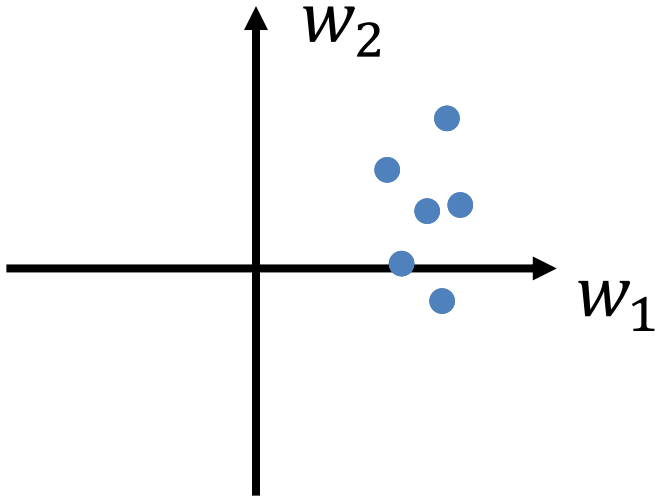
Maximum volume removal

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





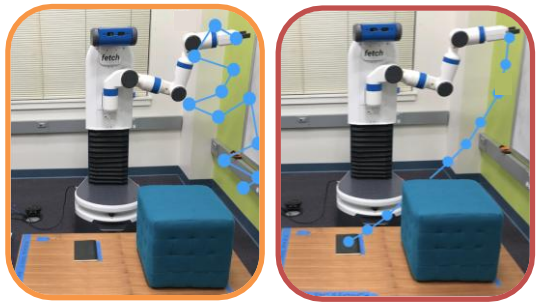
User Choice

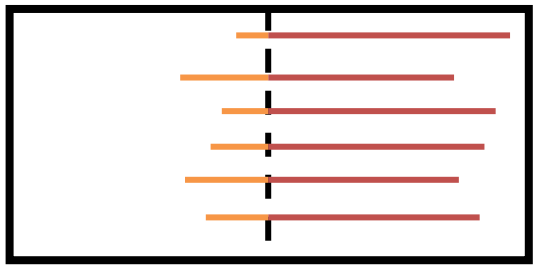


Active Preference-Based Learning of Reward Functions
Sadigh et al. RSS 2017

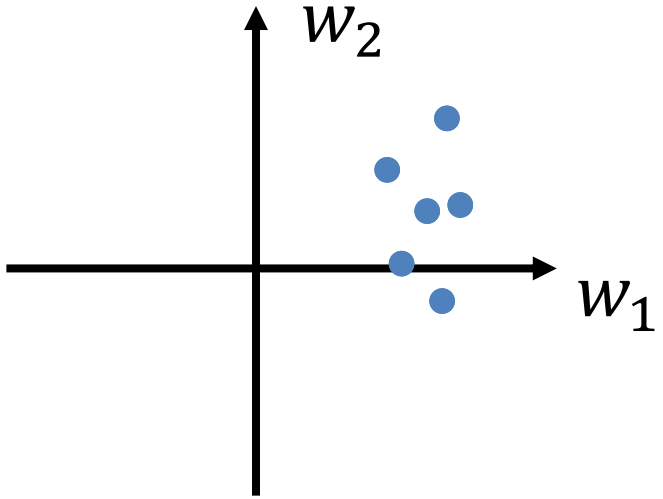
Maximum volume removal

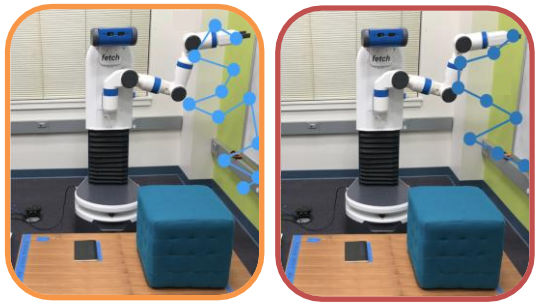
Posterior $P(w \mid \mathcal{C})$

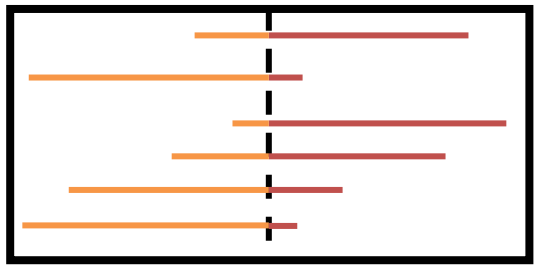




User Choice





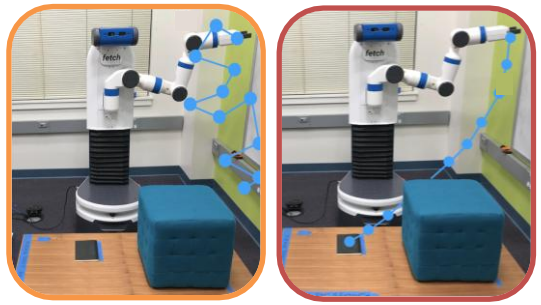


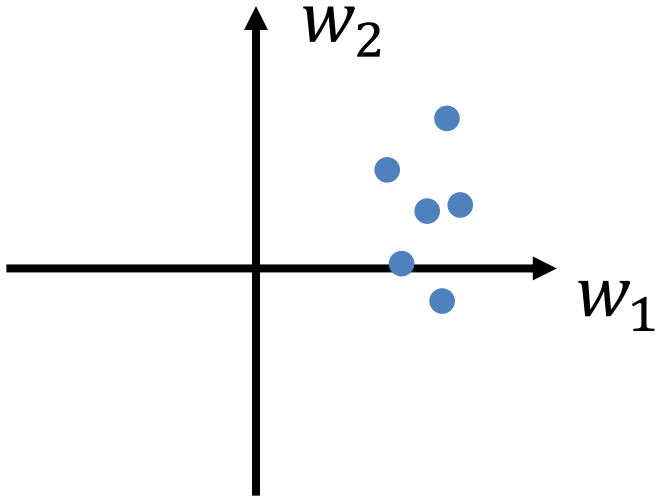
User Choice

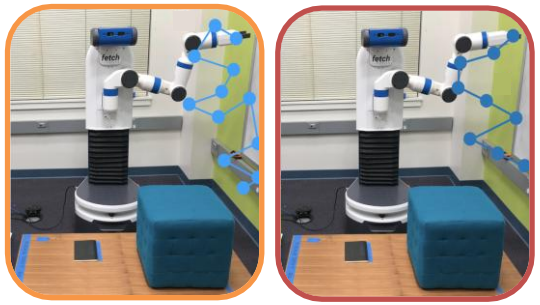
Active Preference-Based Learning of Reward Functions
Sadigh et al. RSS 2017

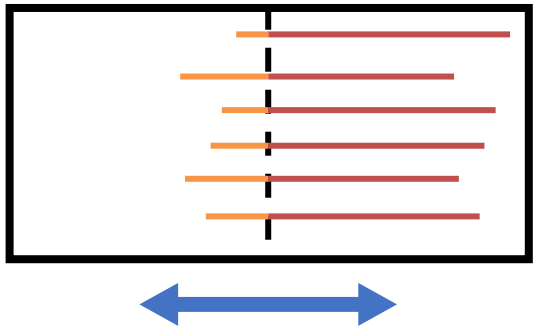
Maximum volume removal

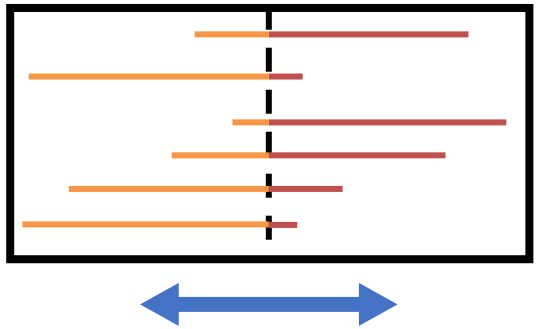
Posterior $P(w \mid \mathcal{C})$





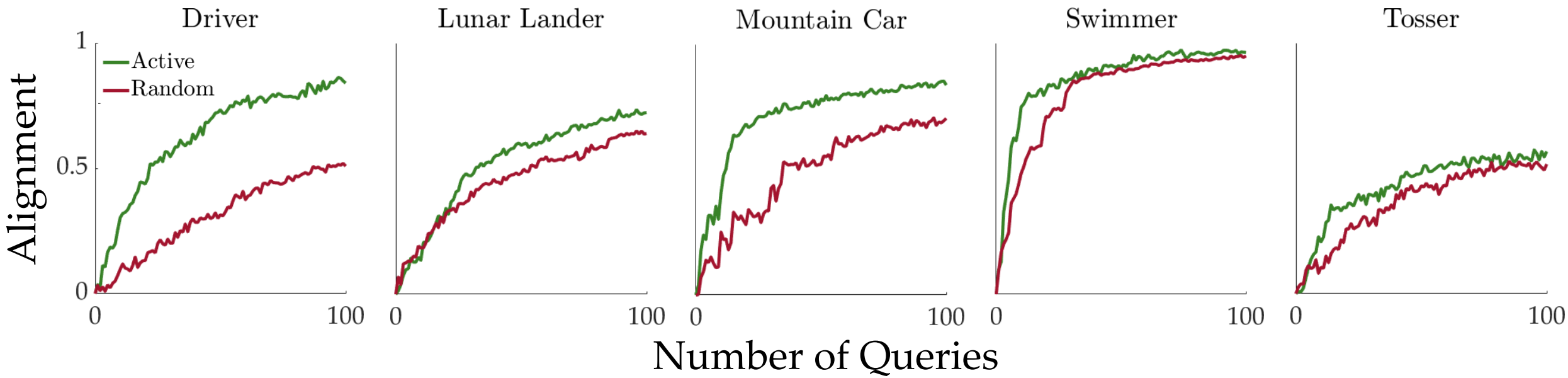


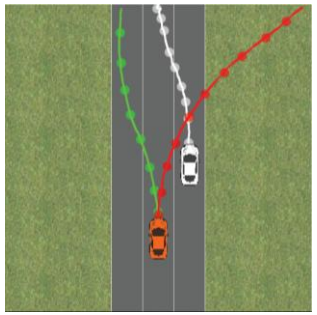
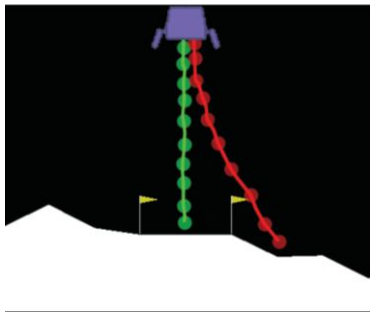
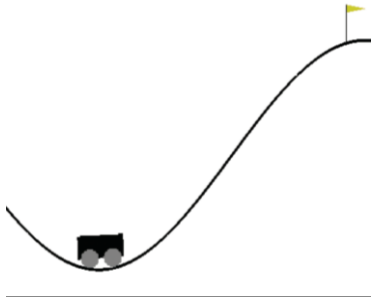
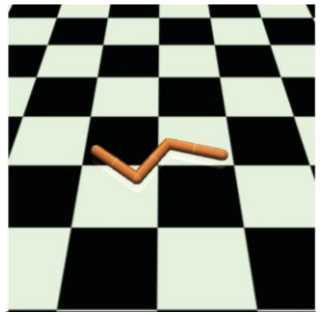
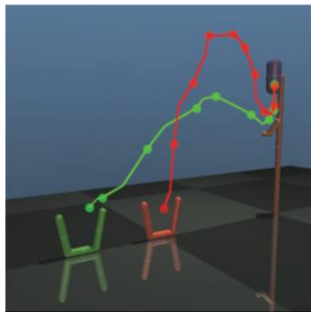




Active Preference-Based Learning of Reward Functions
Sadigh et al. RSS 2017

Active vs. random querying

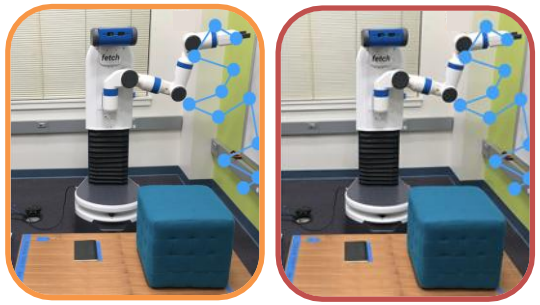


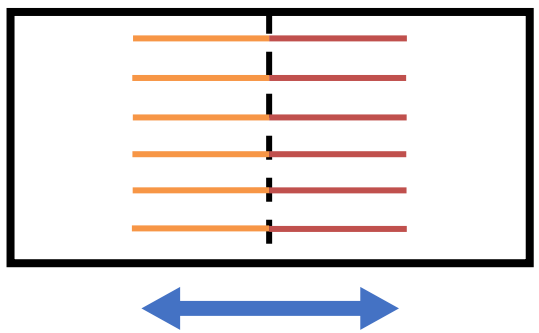






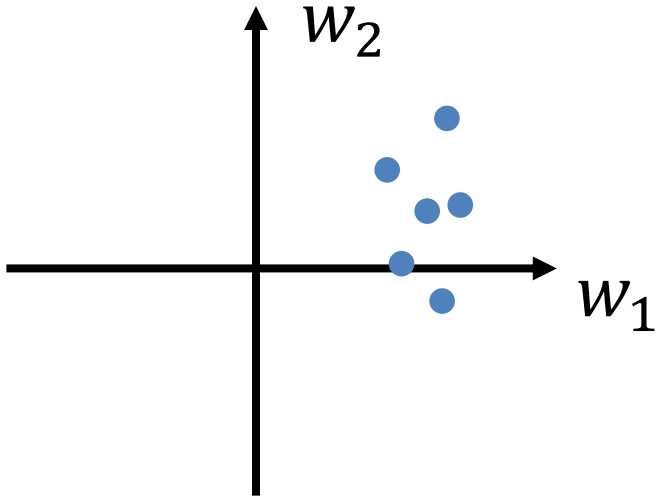
Batch Active Preference-Based Learning of Reward Functions
Bıyık and Sadigh, CoRL 2018

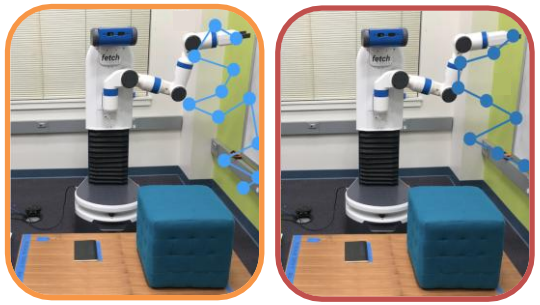
Maximum volume removal

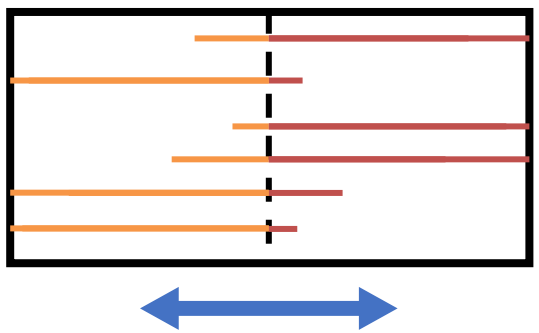
Posterior $P(w \mid \mathcal{C})$











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

Mutual information maximization

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

User response

Dataset

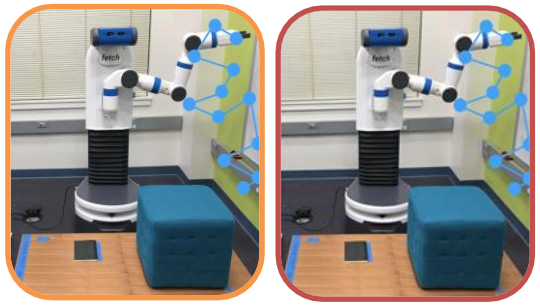
Query

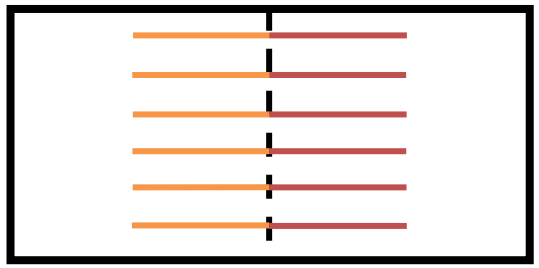
Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization

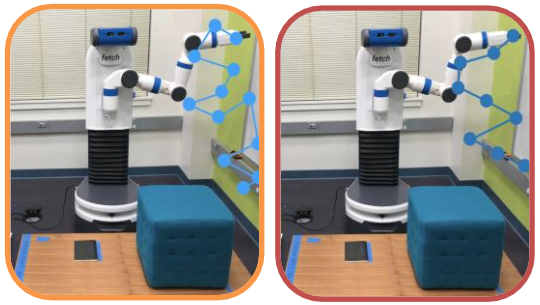
$$\max_{\xi_A, \xi_B} I(q; w \mid \boldsymbol{\mathcal{C}}, \xi_A, \xi_B)$$

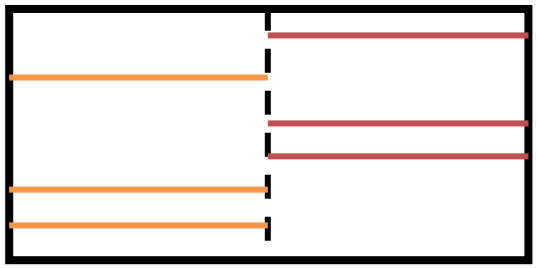
$$\max_{\xi_A, \xi_B} \underbrace{H(q \mid \boldsymbol{\mathcal{C}}, \xi_A, \xi_B)}_{\text{Model Uncertainty}} - \underbrace{H(q \mid \boldsymbol{\mathcal{C}}, \xi_A, \xi_B, w)}_{\text{User Uncertainty}}$$





User Choice





User Choice

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

Mutual information maximization

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

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

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

No w here!

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

Mutual information maximization

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

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

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

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

Mutual information maximization

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

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

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

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

Mutual information maximization

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

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

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

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

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

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

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

Mutual information maximization

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

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

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

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

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

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

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

Mutual information maximization

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

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

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

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

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

Mutual information maximization

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

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

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

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

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

Mutual information maximization

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

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

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

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

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

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

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

Mutual information maximization

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

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

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

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

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

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

Mutual information maximization

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

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

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

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

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

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

Mutual information maximization

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

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

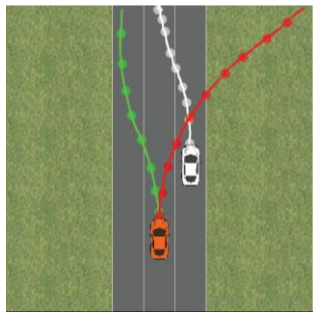
$$\max_{\xi_A,\xi_B} \mathbb{E}_{q,w|\boldsymbol{\mathcal{C}},\xi_A,\xi_B}[\log P(\, q \mid \xi_A, \xi_B, w\,) - \log P(\, q \mid \boldsymbol{\mathcal{C}}, \xi_A, \xi_B\,)]$$

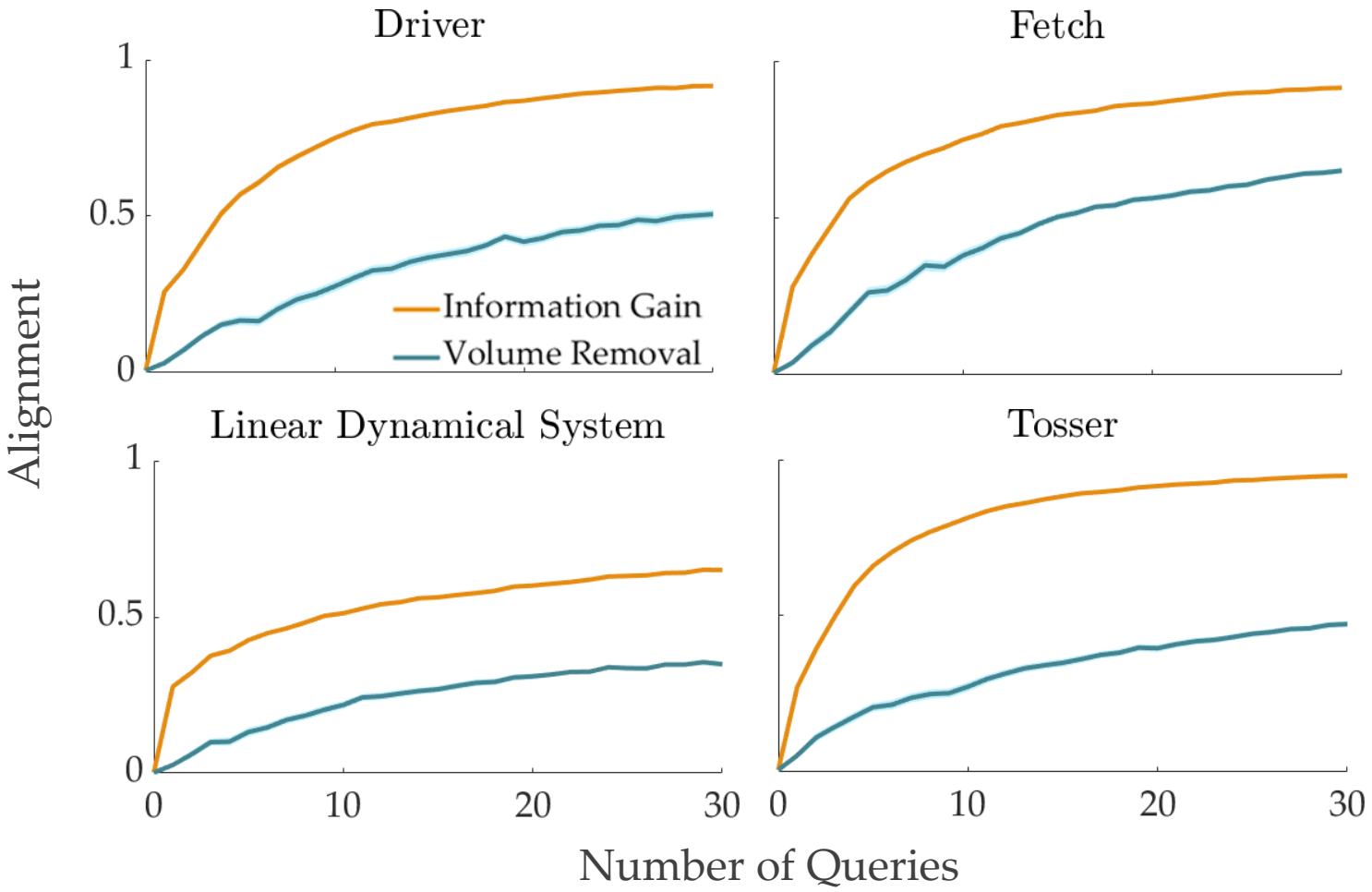
$$\max_{\xi_A,\xi_B} \mathbb{E}_{q,w|\xi_A,\xi_B}\left[\log \frac{P(\, q \mid \xi_A, \xi_B, w\,)}{\sum_{w'\in\Omega} P(\, q \mid \xi_A, \xi_B, w'\,)}\right]$$

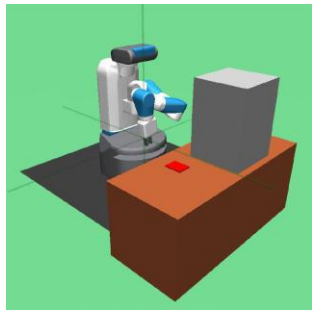
$$\max_{\xi_A,\xi_B} \sum_{w\in\Omega}\sum_{q} P(q \mid \xi_A, \xi_B, w)\left[\log \frac{P(\, q \mid\ \xi_A, \xi_B, w\,)}{\sum_{w'\in\Omega} P(\, q \mid \xi_A, \xi_B, w'\,)}\right]$$

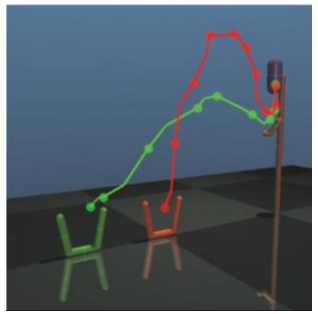
Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization





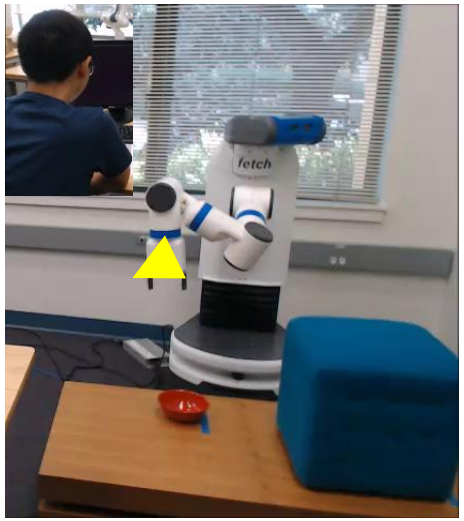


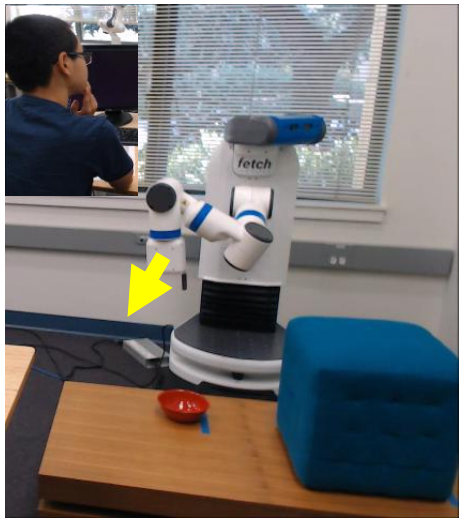


Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Volume Removal

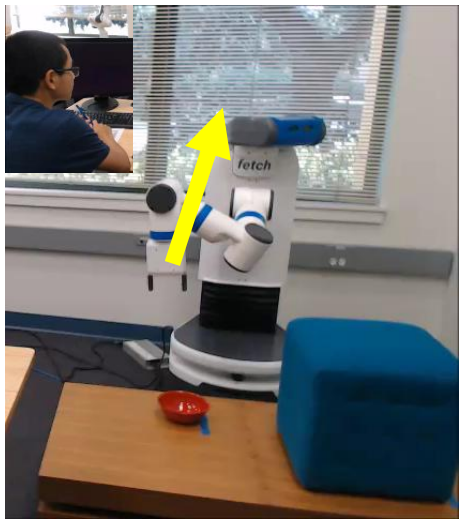
Similar Trajectories

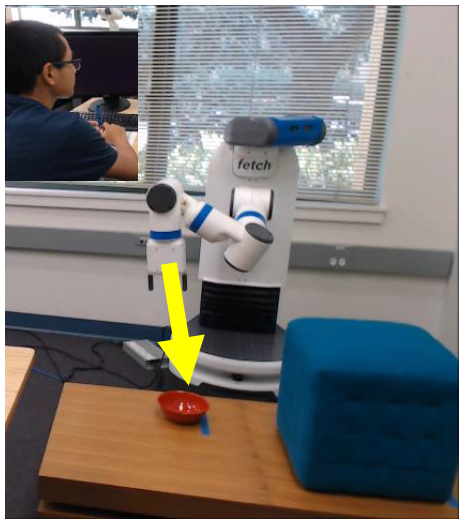




Information Gain

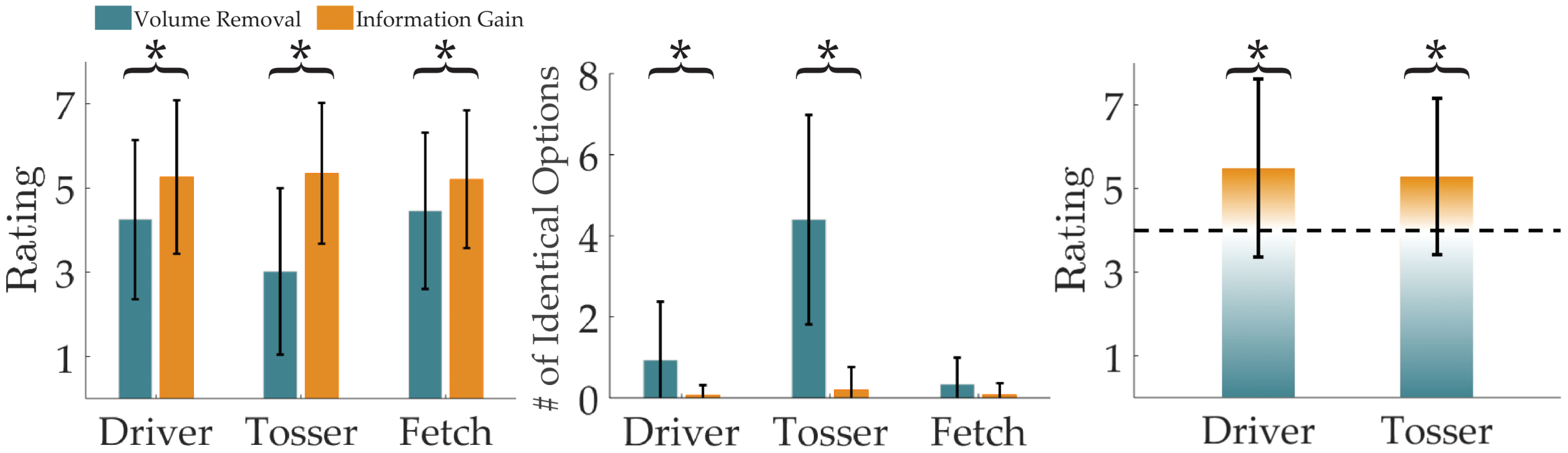
More Distinguishable Query



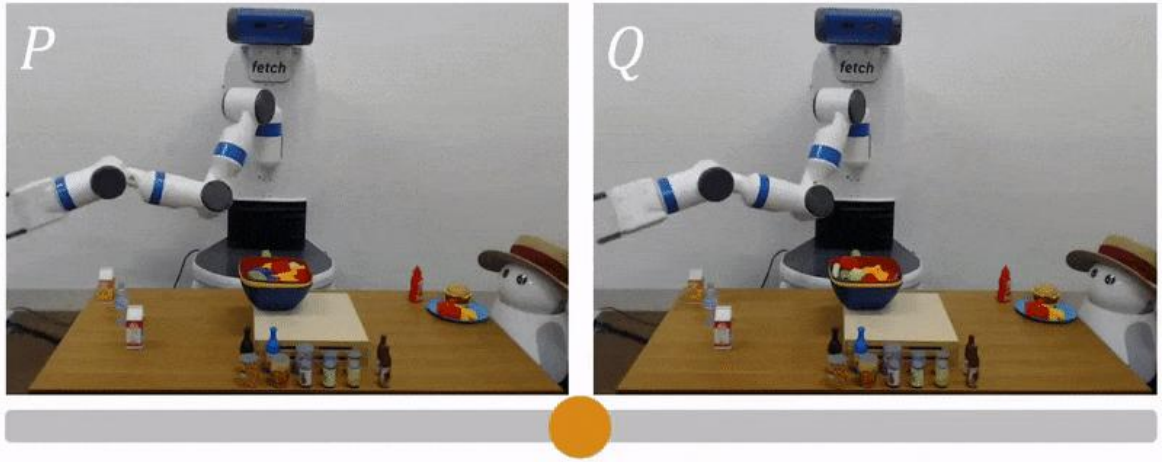


Asking Easy Questions: A User-Friendly Approach to Active Reward Learning
Bıyık et al. CoRL 2019

Mutual information maximization



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



Richer forms of comparative feedback enables learning more general reward functions.

*Basu, **Bıyık**, He, Singhal, Sadigh. IROS'19.*

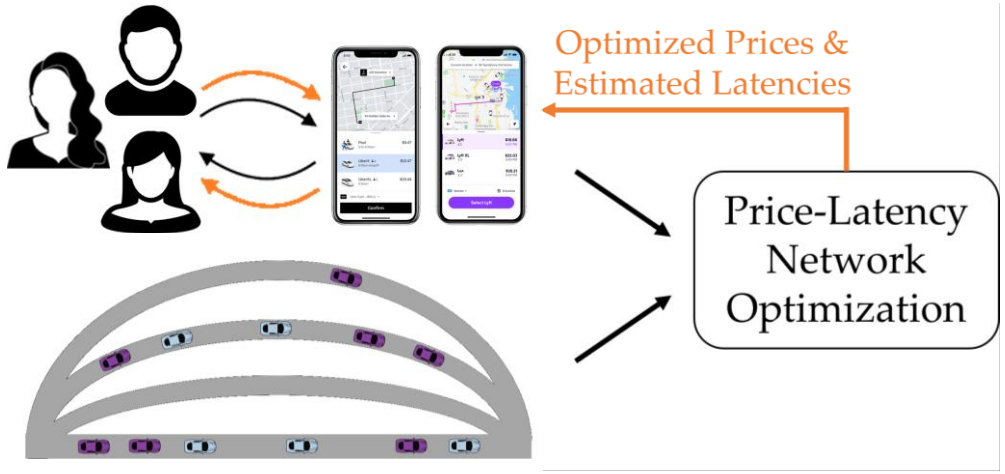
Wilde, **Bıyık***, Sadigh, Smith. CoRL'21.*

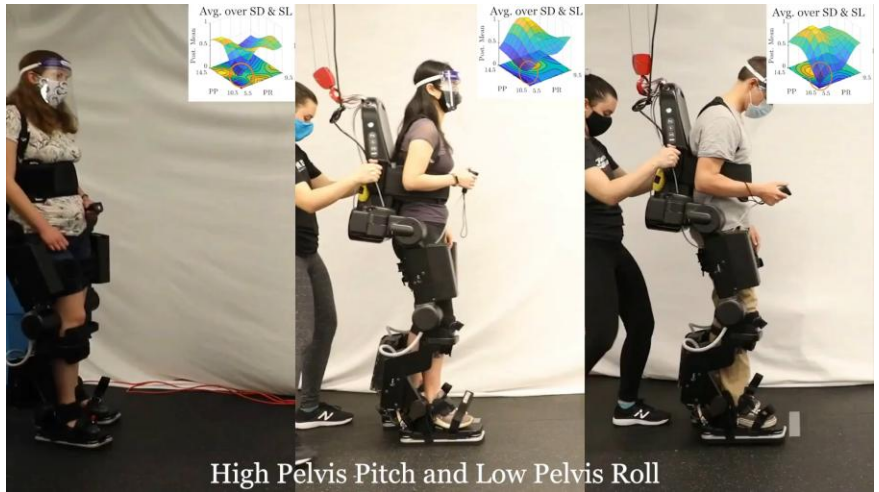
Data from Uber/Lyft are comparative feedback. We can use them to optimize traffic routing.

***Bıyık**, Lazar, Sadigh, Pedarsani. CDC'19.*

Bıyık, Lazar*, Pedarsani, Sadigh. TCNS'21.*

*Beliaev, **Bıyık**, Lazar, Wang, Sadigh, Pedarsani. ICCPS'21.*





Active preference-based Gaussian process regression enables learning complex rewards with small amounts of data.

Bıyık, Huynh*, Kochenderfer, Sadigh. RSS'20.*

*Li, Tucker, **Bıyık**, Novoseller, Burdick, Sui, Sadigh, Yue, Ames. ICRA'21.*

***Bıyık**, Huynh, Kochenderfer, Sadigh. IJRR'23.*

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.

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	$\mathbb{E}_{\xi \sim \text{Unif}(G(\lambda^*))}[r(\xi)] \geq \mathbb{E}_{\xi \sim \text{Unif}(G(\lambda))}[r(\xi)]\ \forall \lambda \in \Lambda$	$\mathbb{P}(\lambda^* \mid r, \Lambda) = \frac{\exp(\beta \cdot \mathbb{E}_{\xi \sim \text{Unif}(G(\lambda^*))}[r(\xi)])}{\sum_{\lambda \in \Lambda} \exp(\beta \cdot \mathbb{E}_{\xi \sim \text{Unif}(G(\lambda))}[r(\xi)])}$
Proxy Rewards	$\mathbb{E}_{\tilde{\xi} \sim \pi(\tilde{\xi} \mid \tilde{r})}[r(\tilde{\xi})] \geq \mathbb{E}_{\tilde{\xi} \sim \pi(\tilde{\xi} \mid c)}[r(\tilde{\xi})] \quad \forall c \in \tilde{\mathcal{R}}$	$\mathbb{P}(\tilde{r} \mid r, \tilde{\mathcal{R}}) = \frac{\exp(\beta \cdot \mathbb{E}_{\tilde{\xi} \sim \pi(\tilde{\xi} \mid \tilde{r})}[r(\tilde{\xi})])}{\sum_{c \in \tilde{\mathcal{R}}} \exp(\beta \cdot \mathbb{E}_{\tilde{\xi} \sim \pi(\tilde{\xi} \mid c)}[r(\tilde{\xi})])}$
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	$\mathbb{E}_{\xi \sim \psi(s^*)}[r(s^*)] \geq \mathbb{E}_{\xi \sim \psi(s)}[r(s)] \quad \forall s \in \mathcal{S}$	$\mathbb{P}(s^* \mid r, \mathcal{S}) = \frac{\exp(\beta \cdot \mathbb{E}_{\xi \sim \psi(s^*)}[r(\xi)])}{\sum_{s \in S} \exp(\beta \cdot \mathbb{E}_{\xi \sim \psi(s)}[r(\xi)])}$
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
  - Pairwise comparisons
  - Reinforcement learning from human feedback (RLHF)
  - Comparative language feedback

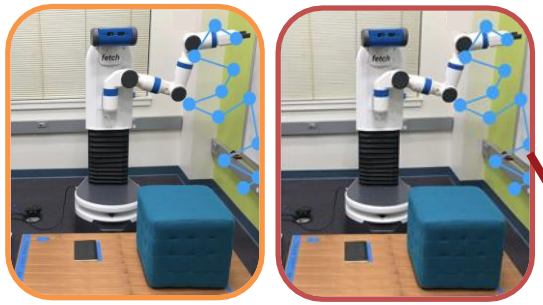
Mutual information maximization

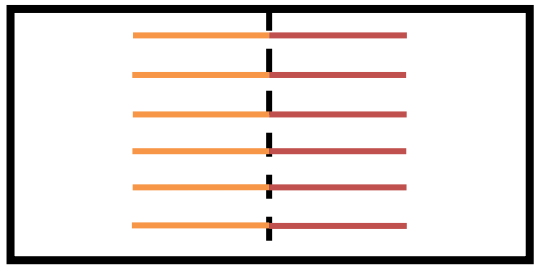
$$\max_{\xi_A, \xi_B} I(q; w \mid \mathcal{C}, \xi_A, \xi_B)$$

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

Model Uncertainty

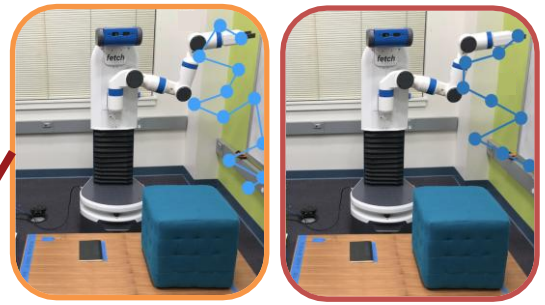
User Uncertainty

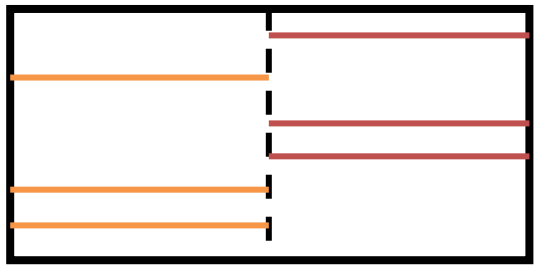




User Choice

Where do these trajectories come from in the first place?





User Choice

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

Incorporating comparisons

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

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

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

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

How do we solve this optimization problem?

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

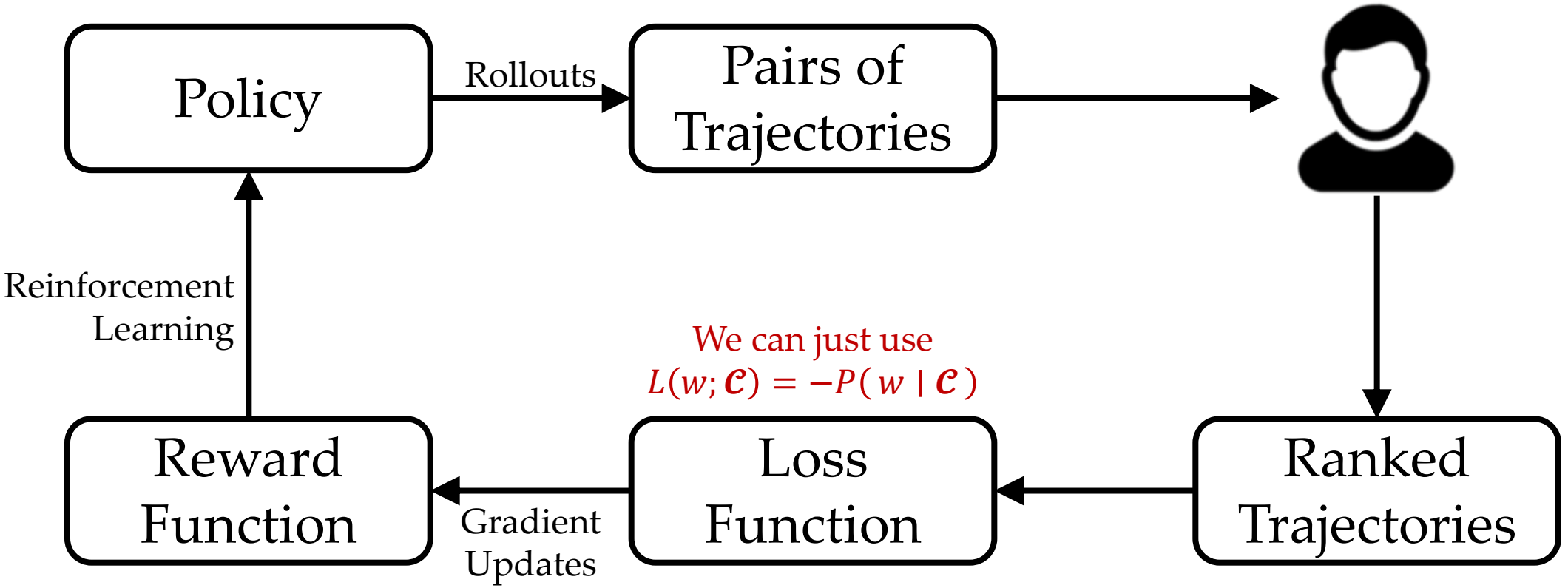
RLHF

Two major changes to preference-based reward learning:

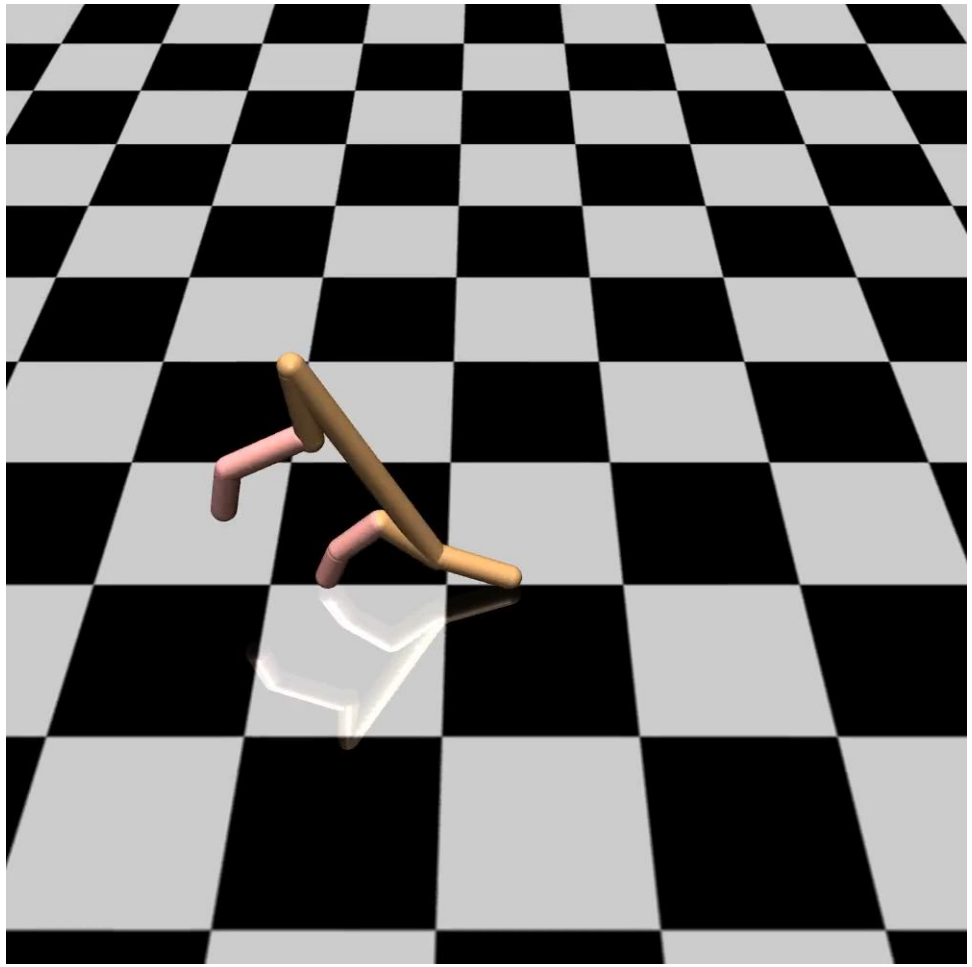
1. Instead of Bayesian learning, write a loss function and learn with gradient updates

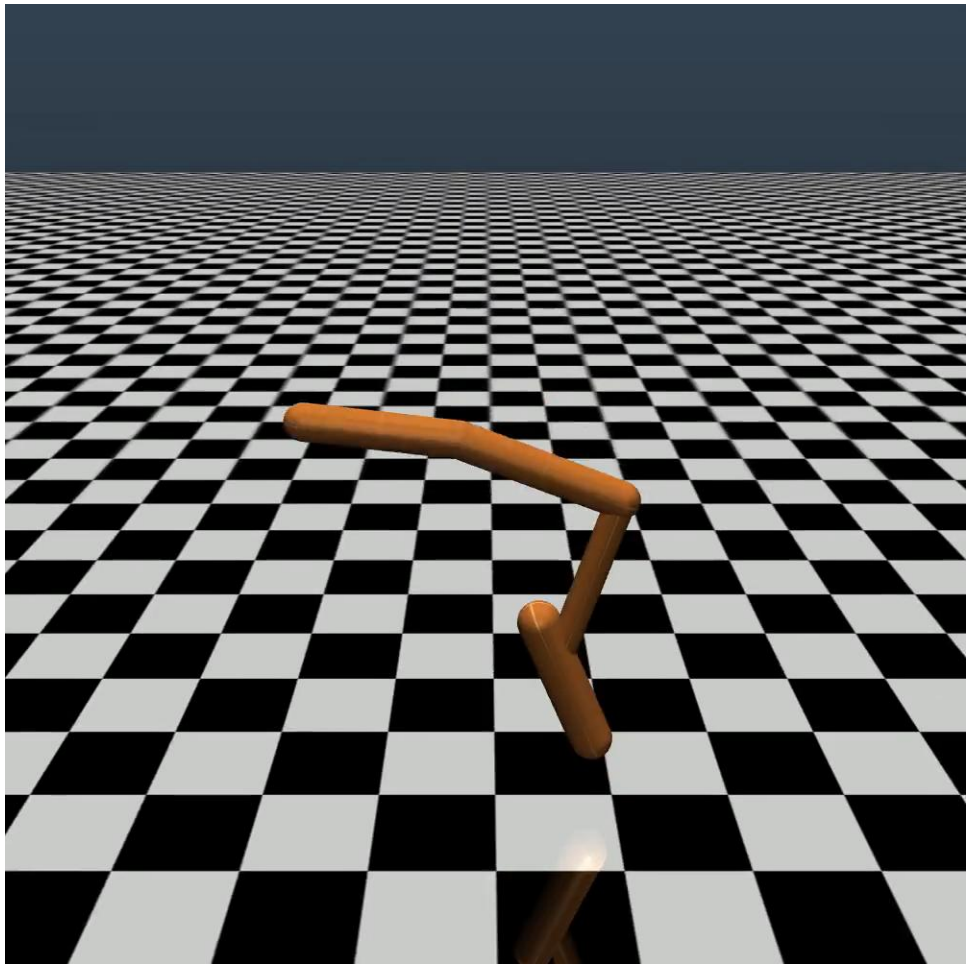
2. After learning a reward, train a policy to generate new trajectories for the next iteration of reward learning

RLHF


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

RLHF



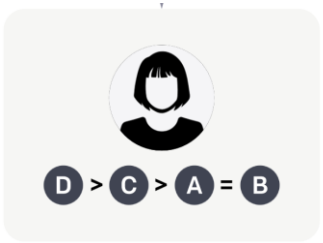


Deep reinforcement learning from human preferences
Christiano et al., NeurIPS 2017

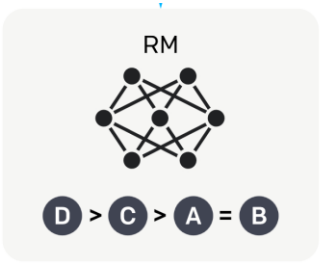
InstructGPT

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.

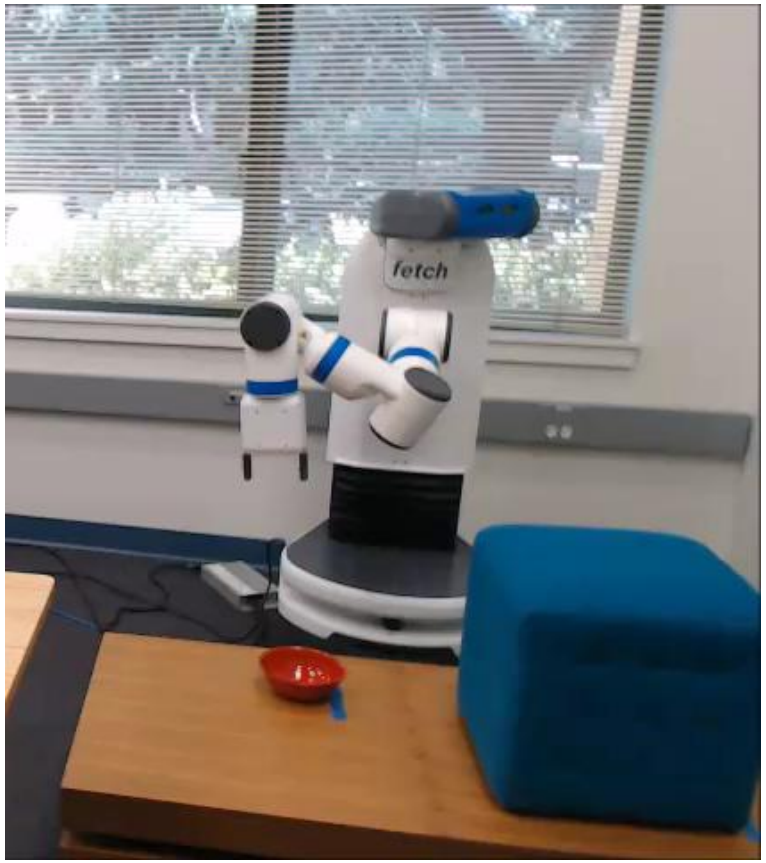


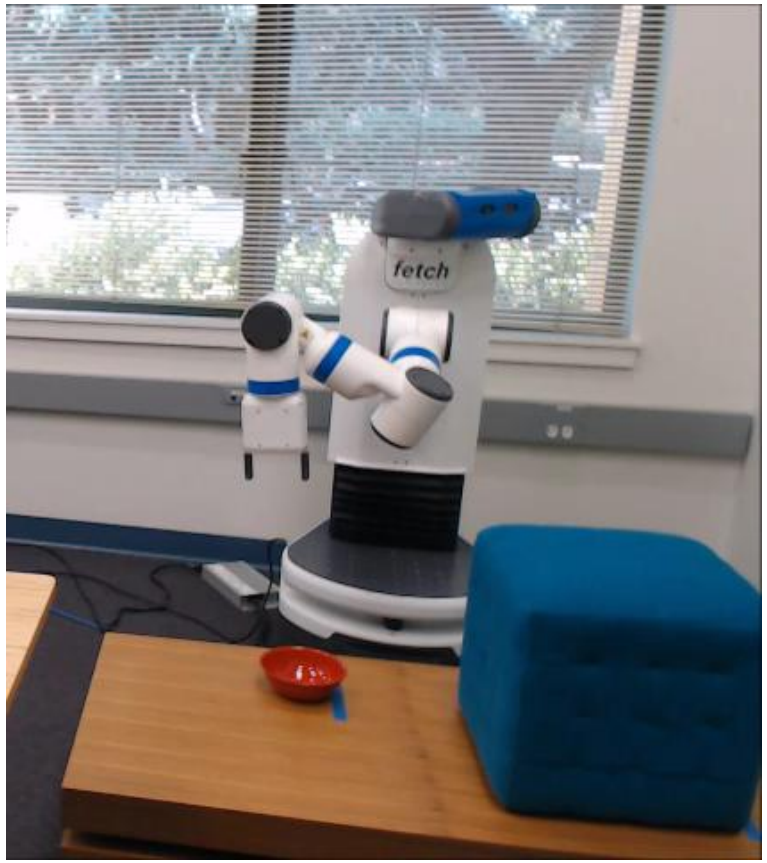
Training language models to follow instructions with human feedback
Ouyang et al., NeurIPS 2022

Today…

- Learning from human feedback
  - Pairwise comparisons
  - Reinforcement learning from human feedback (RLHF)
  - Comparative language feedback

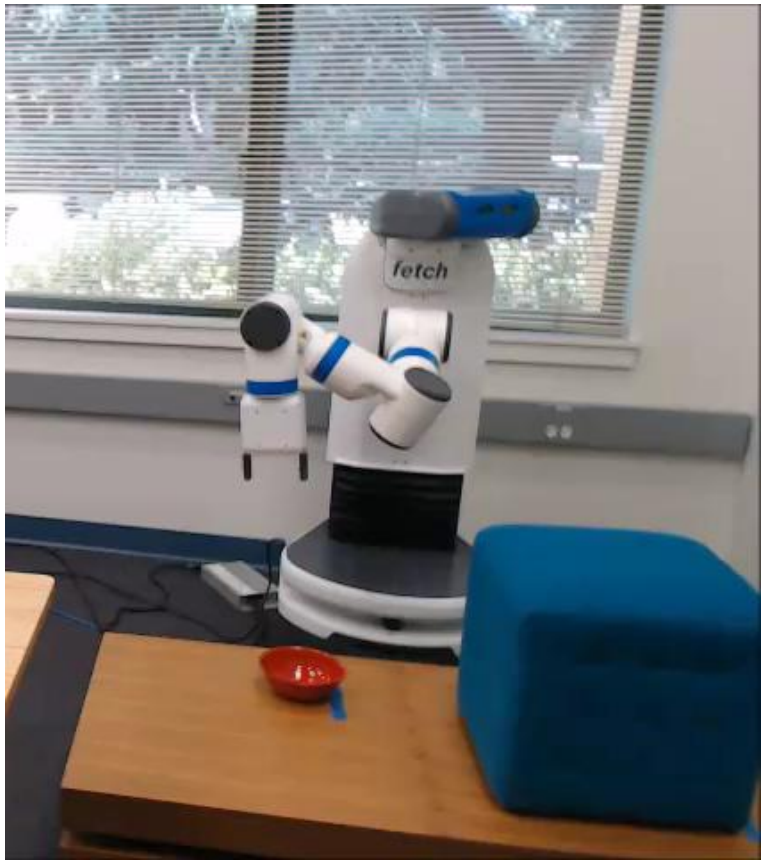
Comparisons take too long

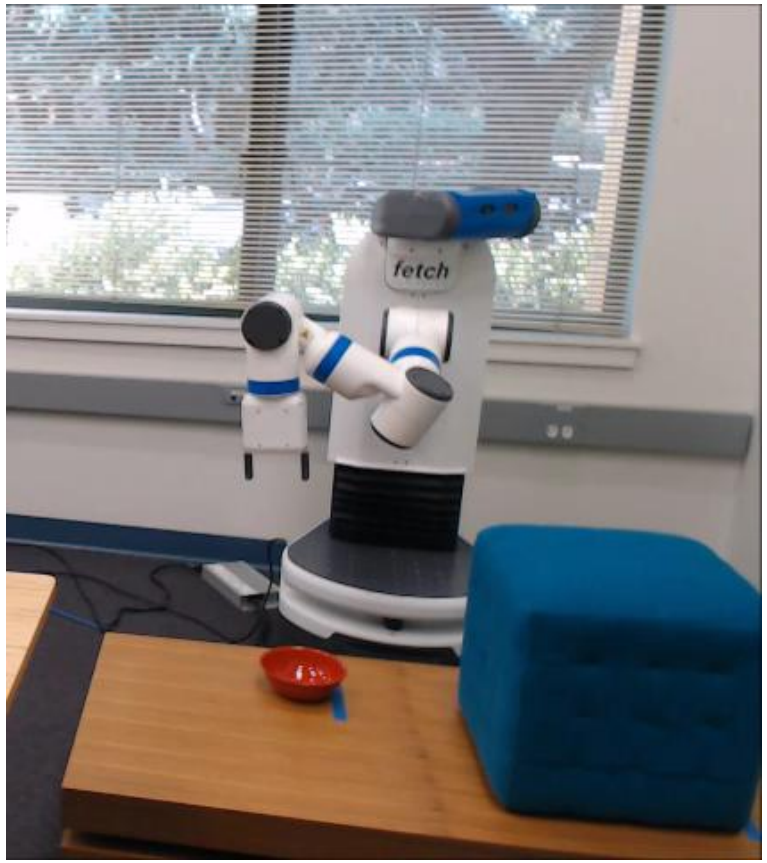




Bıyık, Huynh, Kochenderfer, Sadigh. IJRR'23.

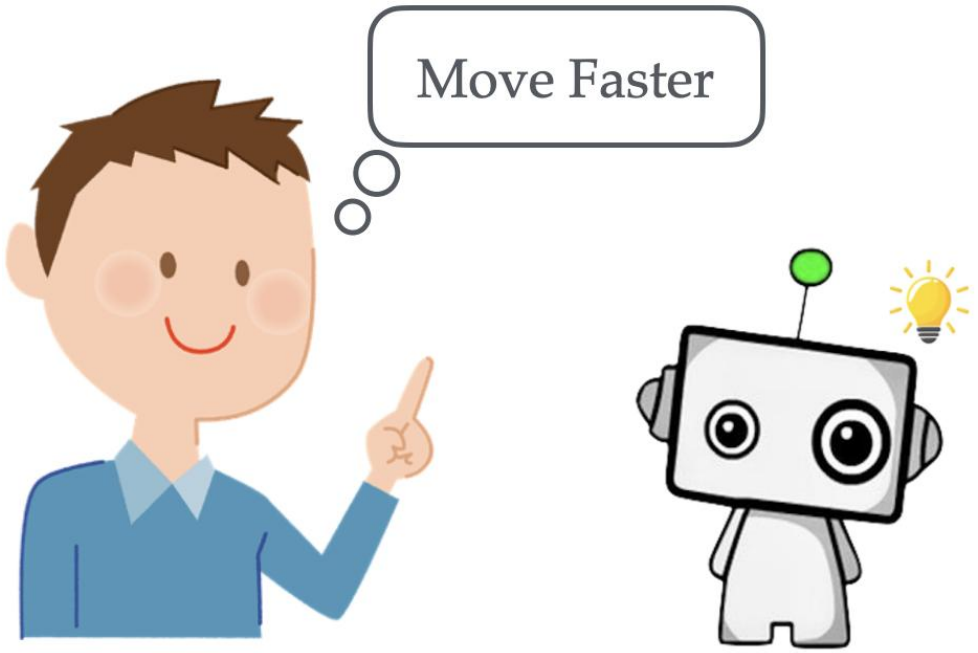
Why do you prefer that?



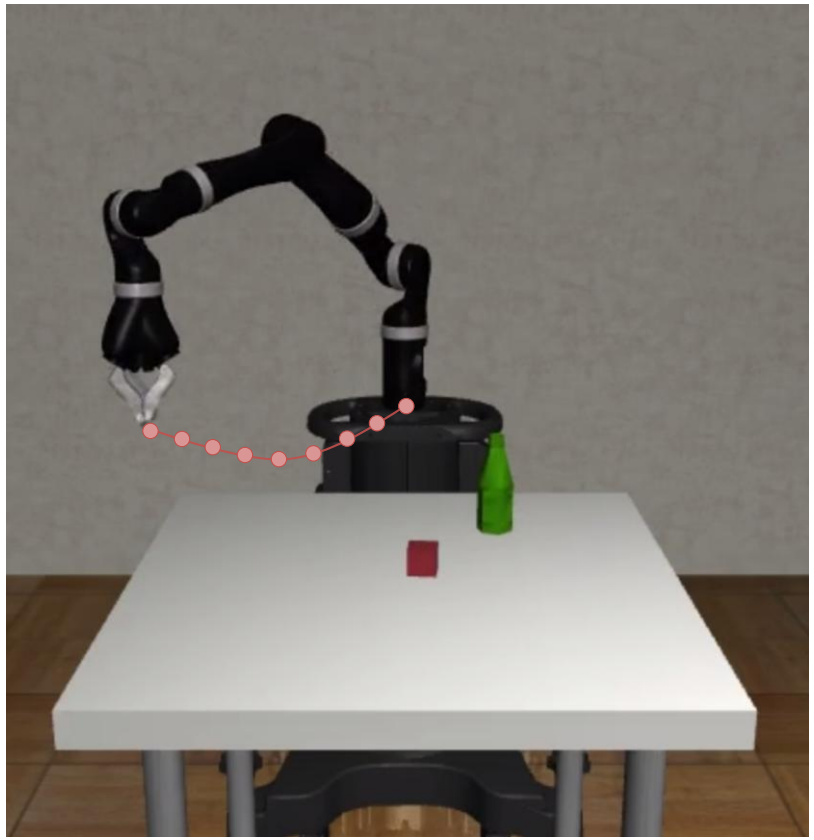


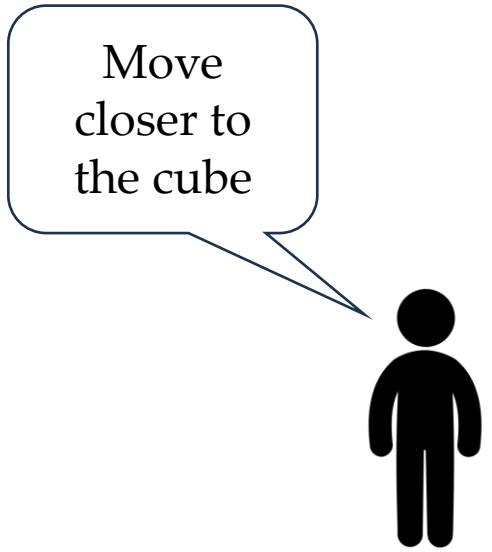
Bıyık, Huynh, Kochenderfer, Sadigh. IJRR'23.

Comparative Language Feedback



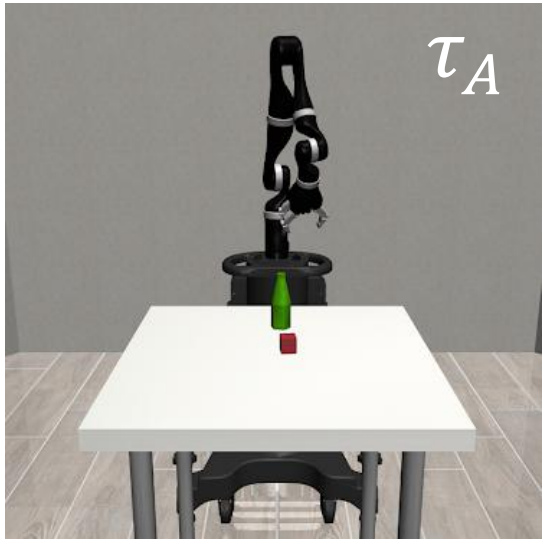
Comparative Language Feedback

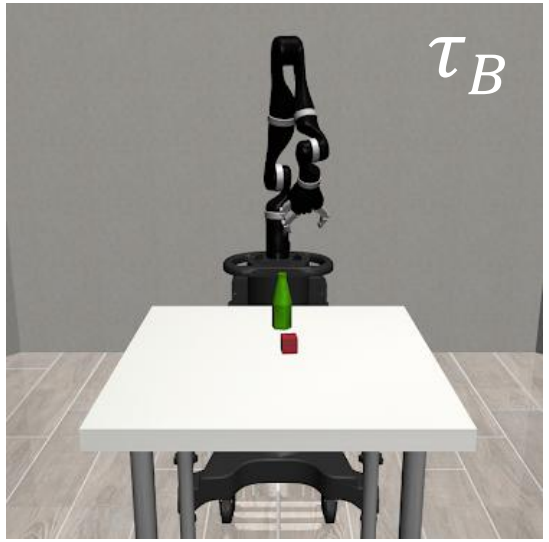




Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

Shared Latent Space



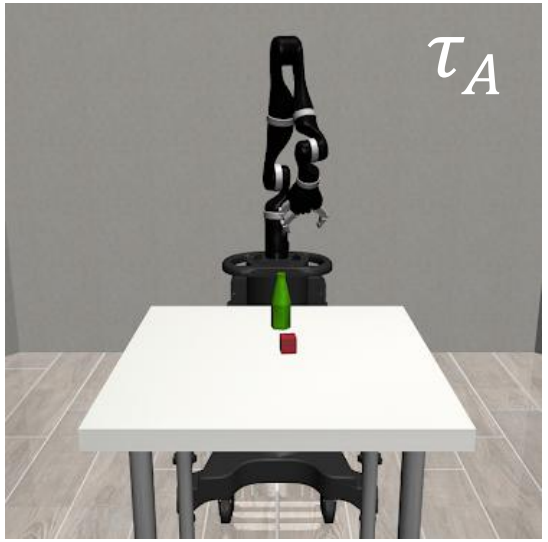


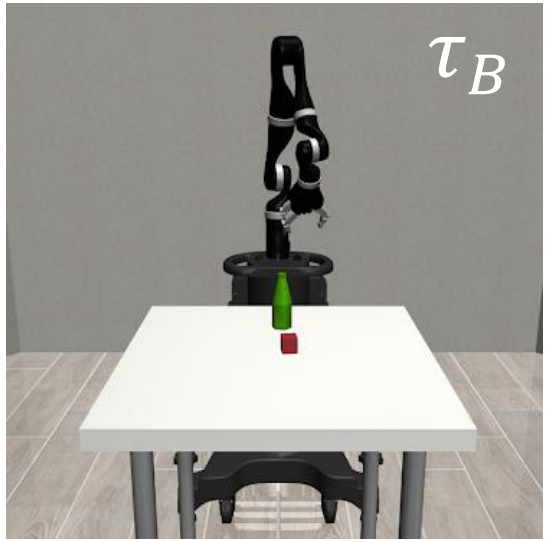
τ^B moves faster than τ^A

Just describes a difference between two trajectories!

Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

Shared Latent Space





τ^B moves faster than τ^A

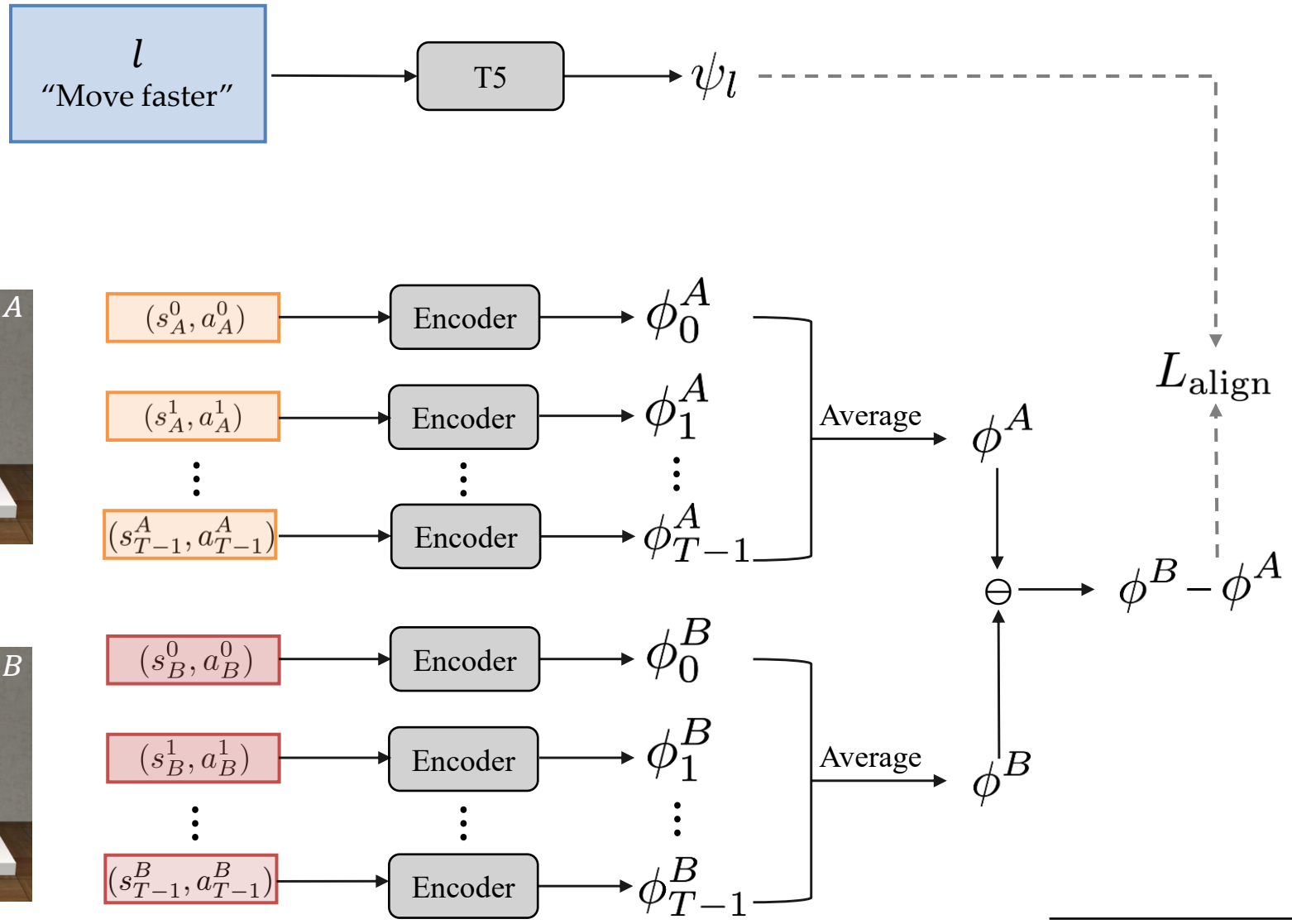
τ^B moves more quickly than τ^A

τ^A moves slower than τ^B

Just describes a difference between two trajectories!

Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

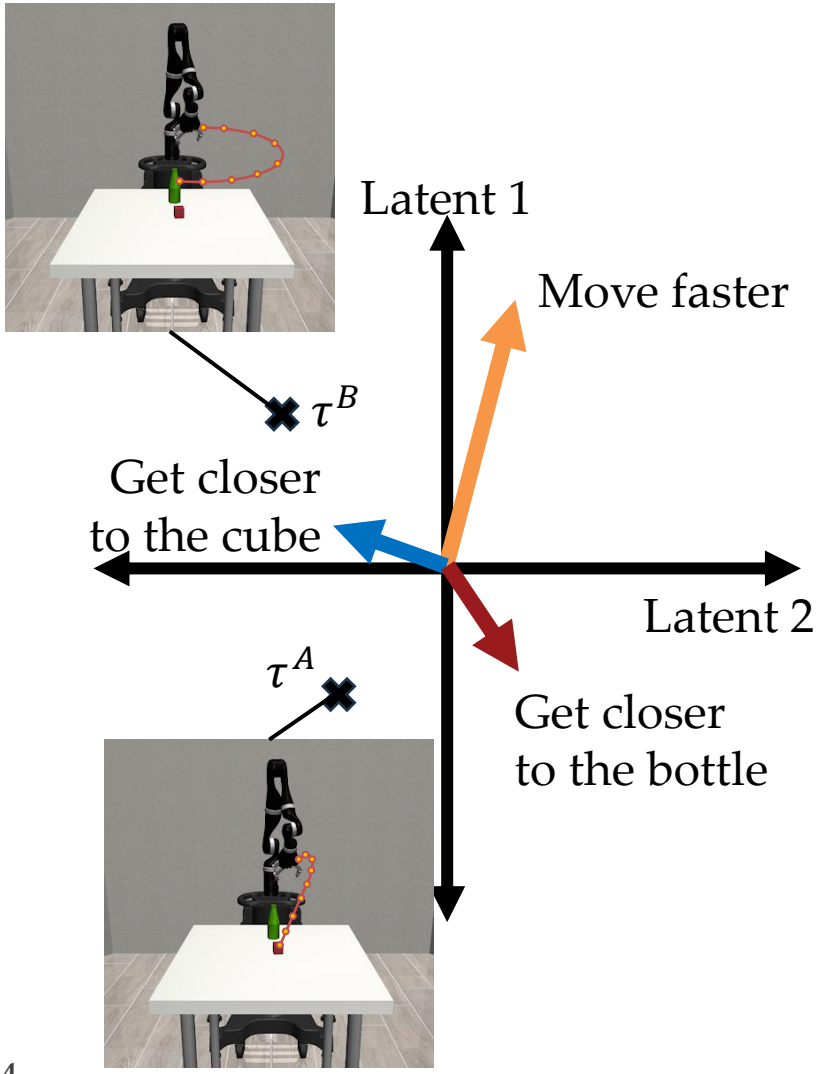
Language feedback



	Accuracy
Robosuite	84.9%
Meta-World	82.9%

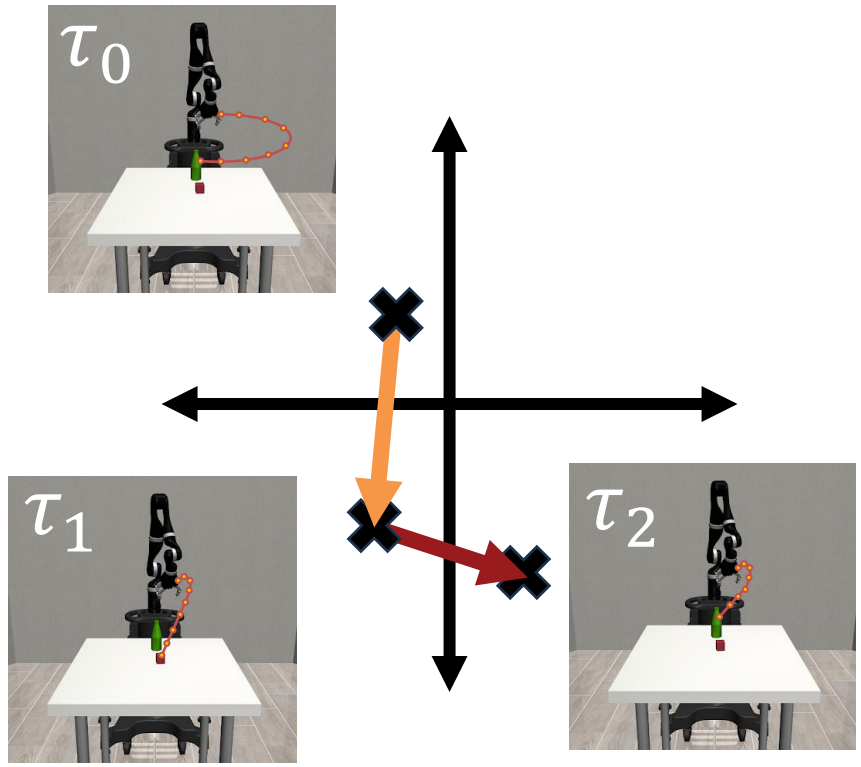
Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

Shared Latent Space



Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

Using the Latent Space



: τ_0?

: Move slower

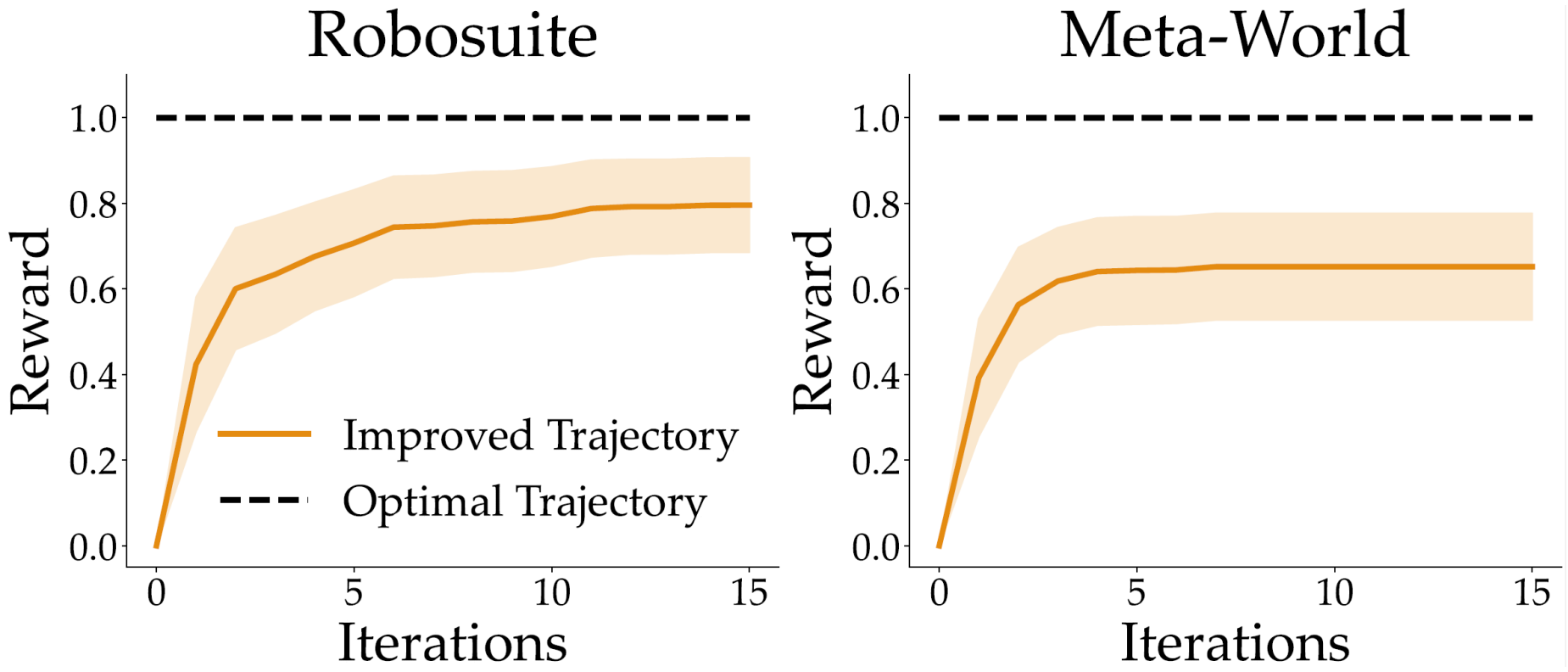
: τ_1?

: Move closer to the bottle

: τ_2?

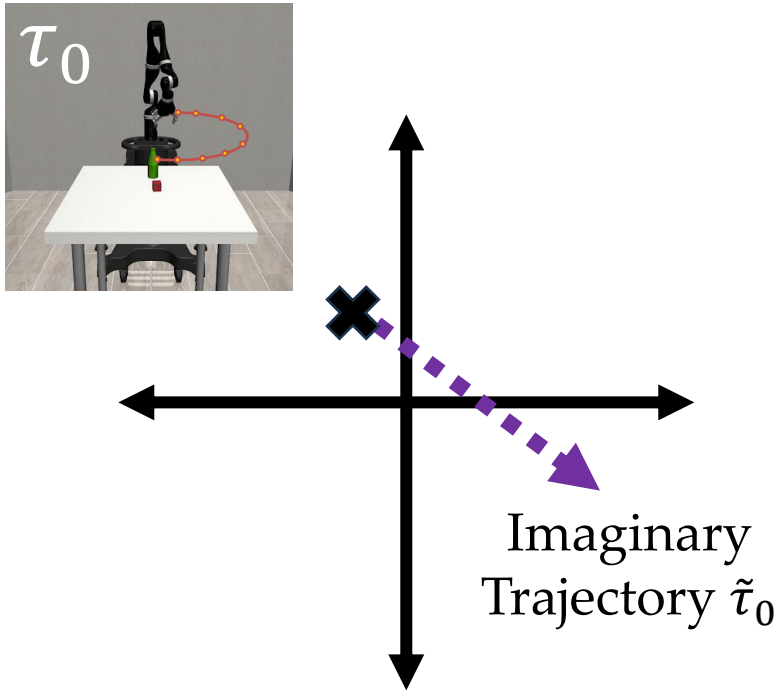
Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

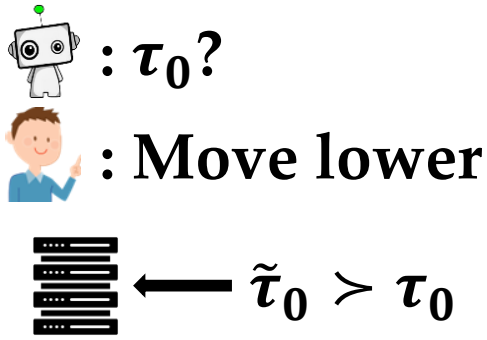
Using the Latent Space



Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

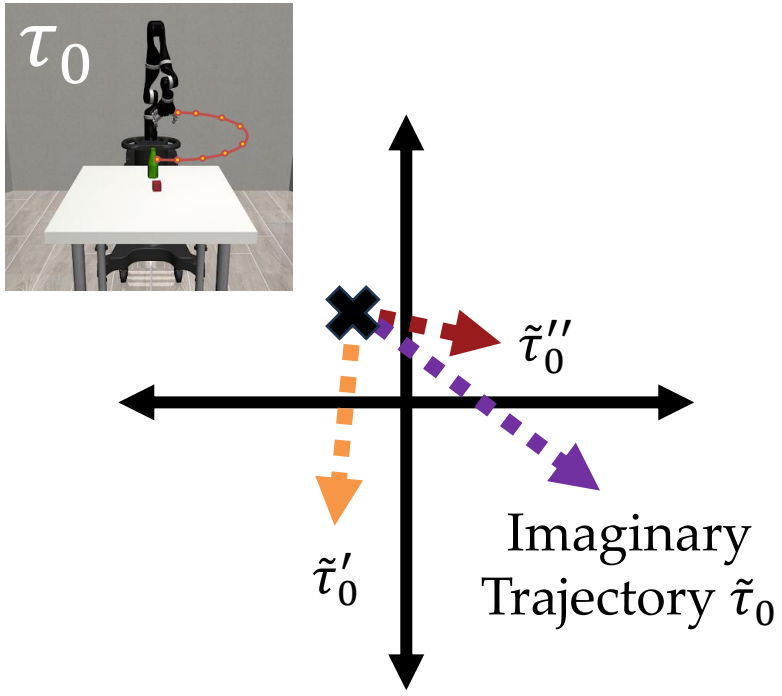
Using the Latent Space

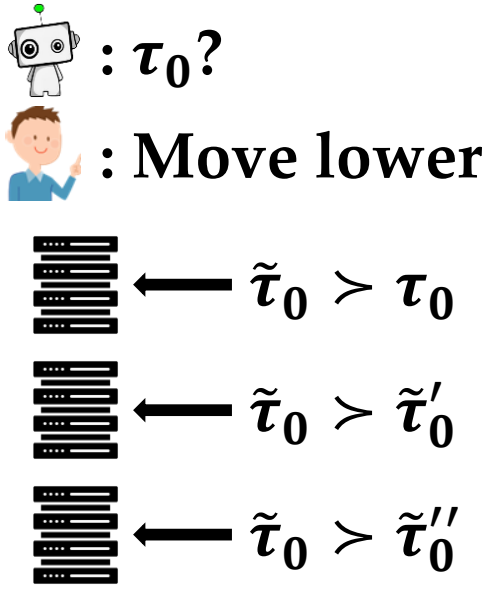




Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

Using the Latent Space



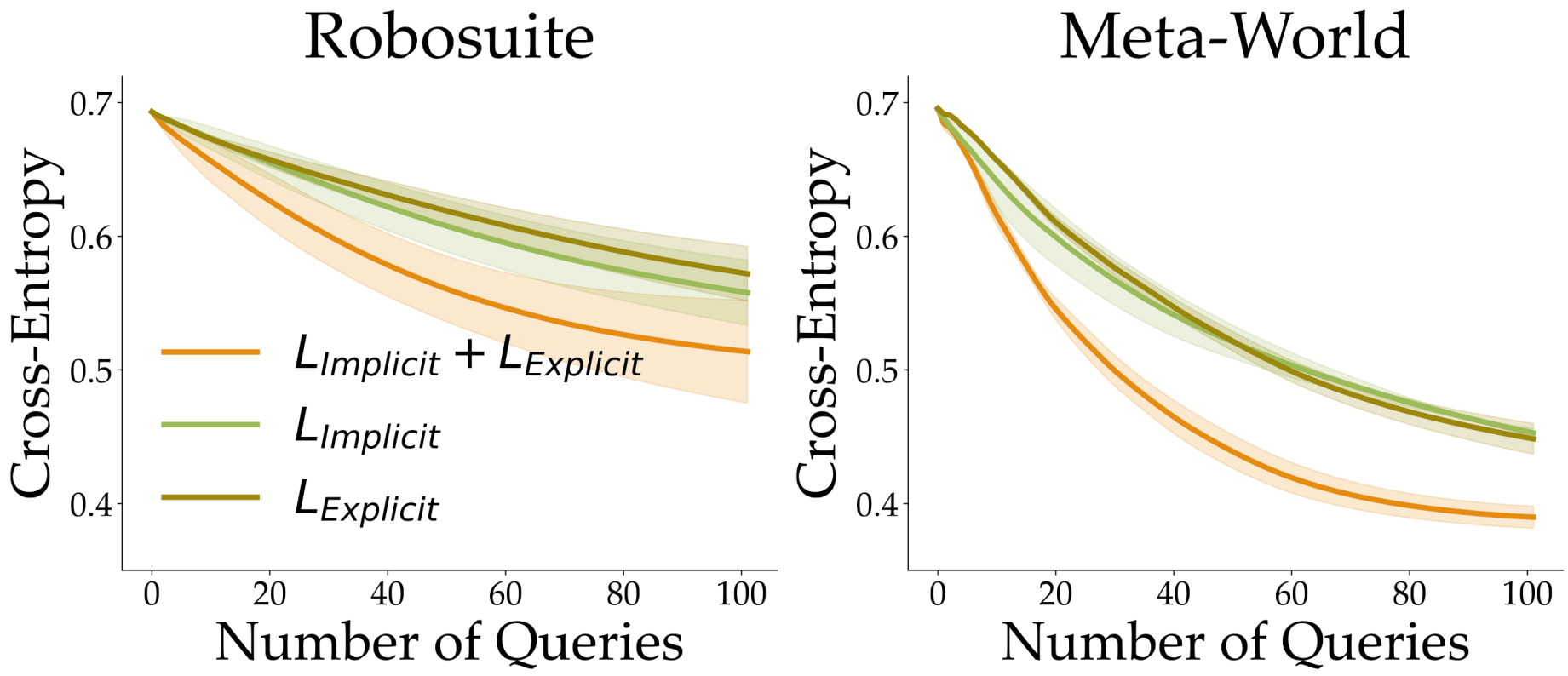


The user intentionally chose to say "move lower"

Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

Using the Latent Space

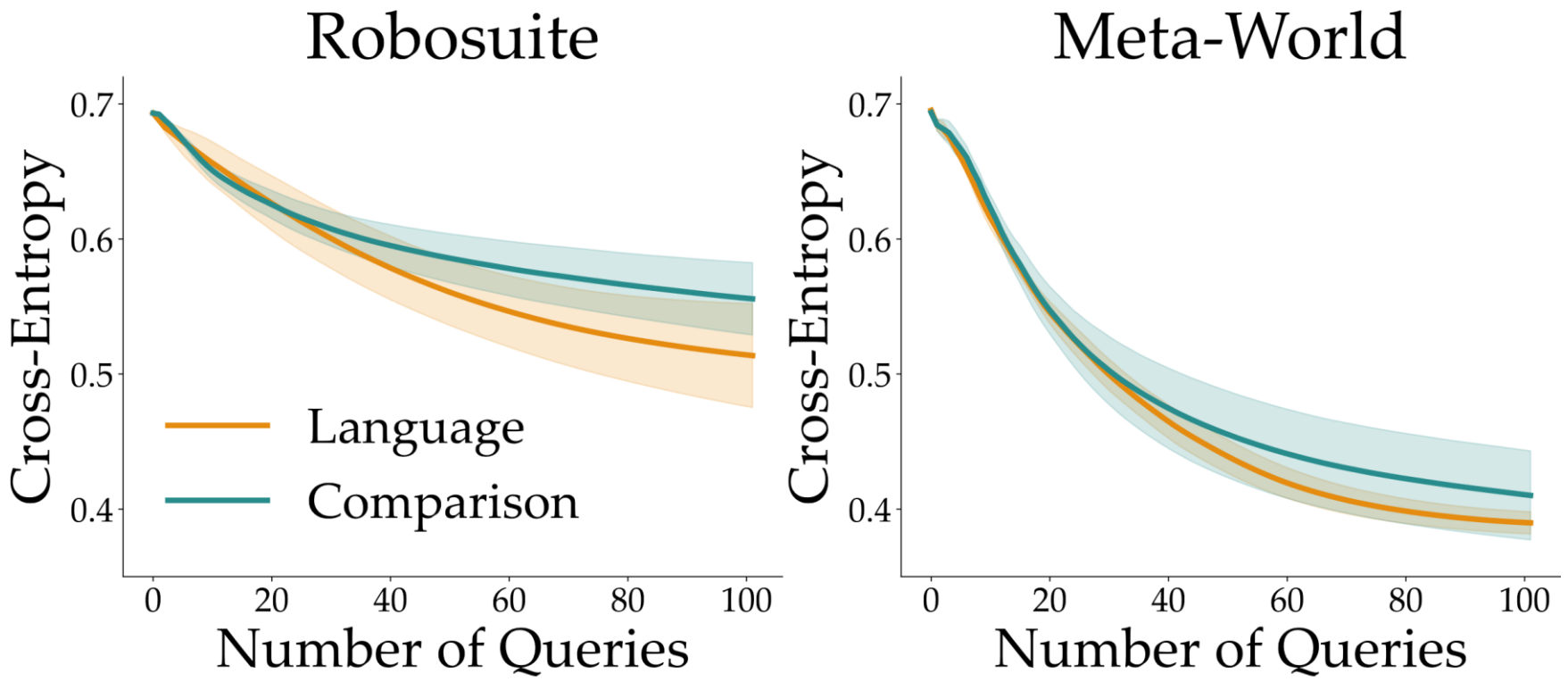
Lower is better



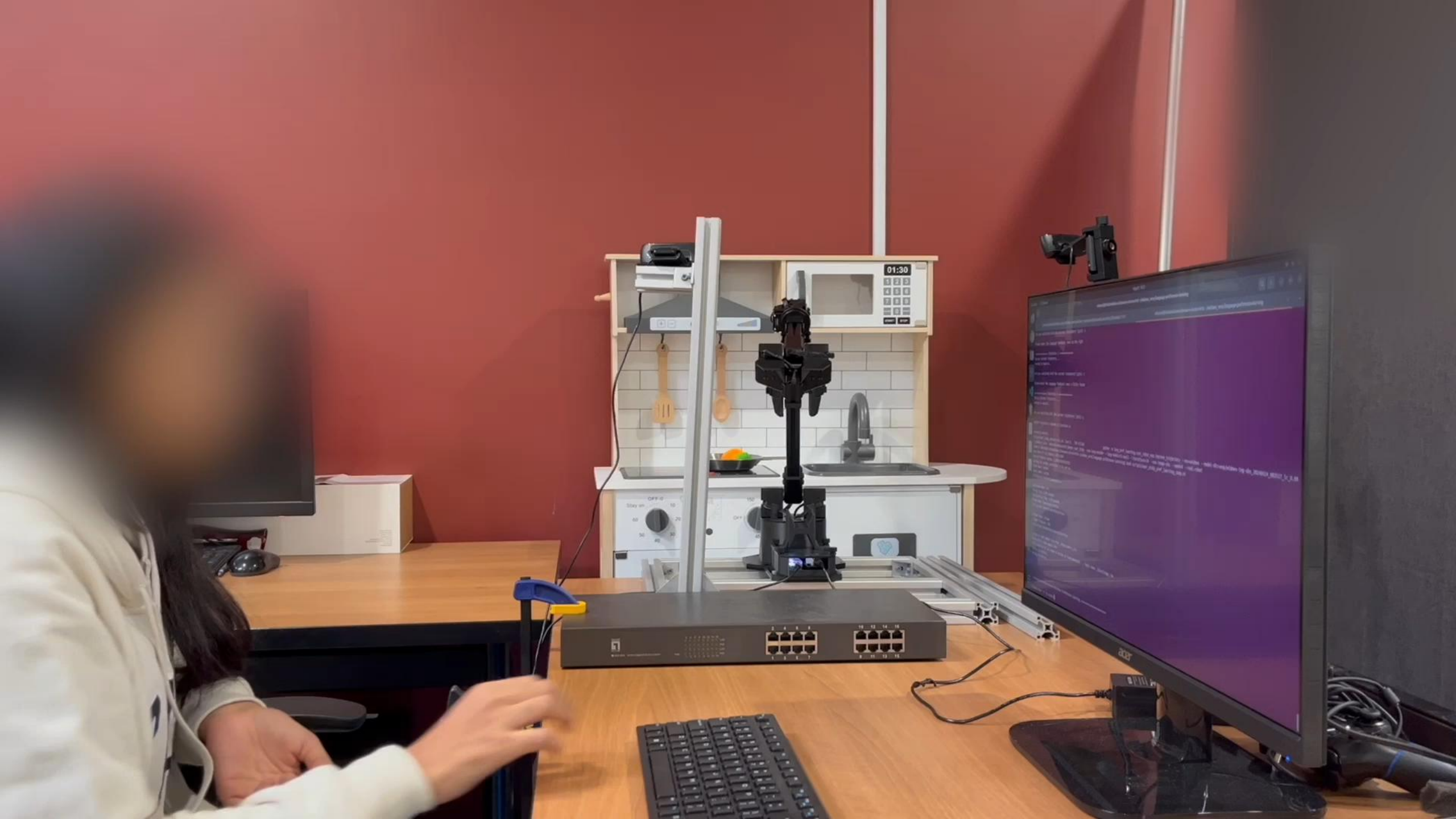
Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

Using the Latent Space

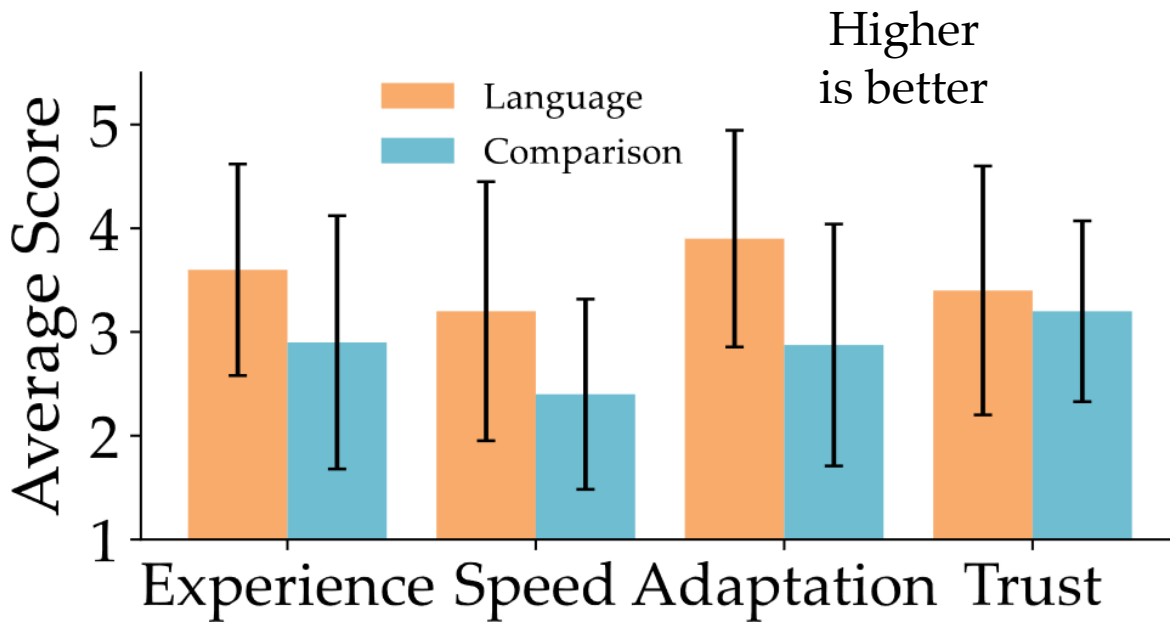
Lower is better

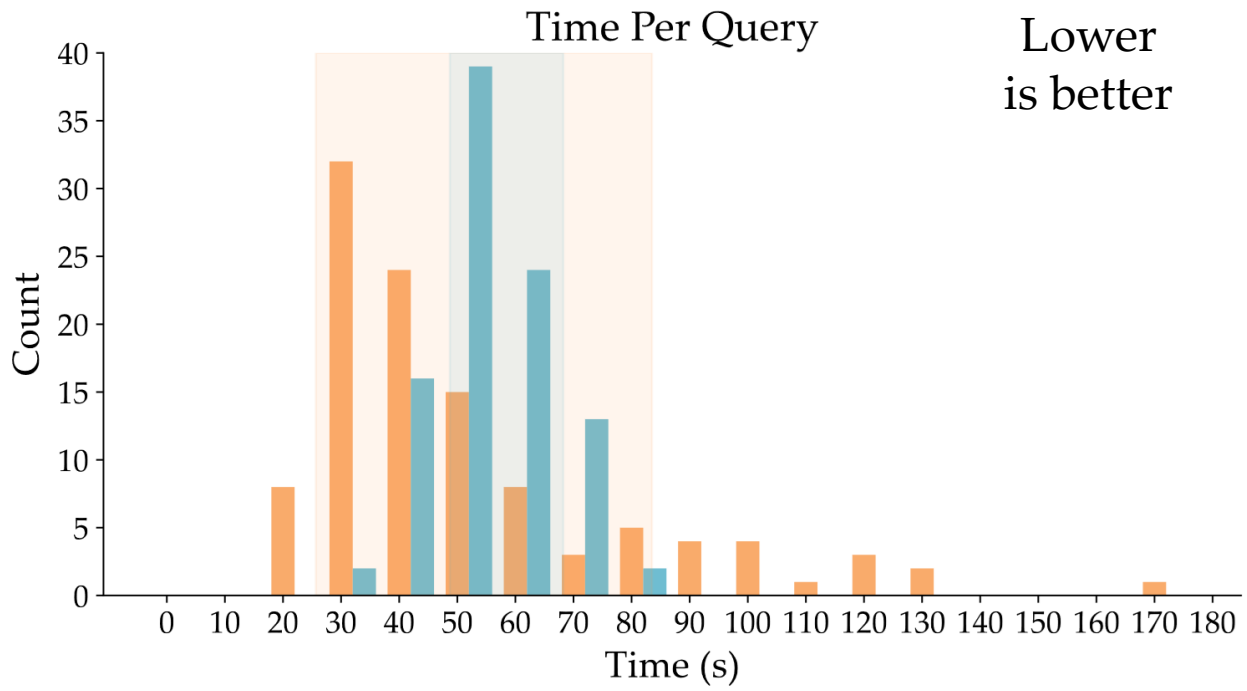


Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.


01:30
acer

User Studies





Yang, Jun, Tien, Russell, Dragan, Bıyık. CoRL'24.

Key Takeaway:

Comparative language feedback brings the benefits of pairwise comparisons and is preferred by humans.

Comparative language for learning

The first law of thermodynamics, also known as the law of energy conservation, states that energy cannot be created or destroyed in an isolated system. Instead, the total energy of a system and its surroundings remains constant. It can be transformed from one form to another or transferred between the system and its surroundings. Mathematically, it is expressed as $\Delta U = Q - W$, where ΔU is the change in internal energy of the system, Q is the heat added to the system, and W is the work done by the system. This principle underscores that all energy changes in a system must account for energy conservation.

Make it less technical

The first law of thermodynamics says that energy in a system can't be created or destroyed, only changed from one form to another or moved around. For example, when you heat water on a stove, the heat energy from the stove is transferred to the water, raising its temperature. The total amount of energy stays the same; it just changes its form or location.

Today…

- Learning from human feedback
  - Pairwise comparisons
  - Reinforcement learning from human feedback (RLHF)
  - Comparative language feedback