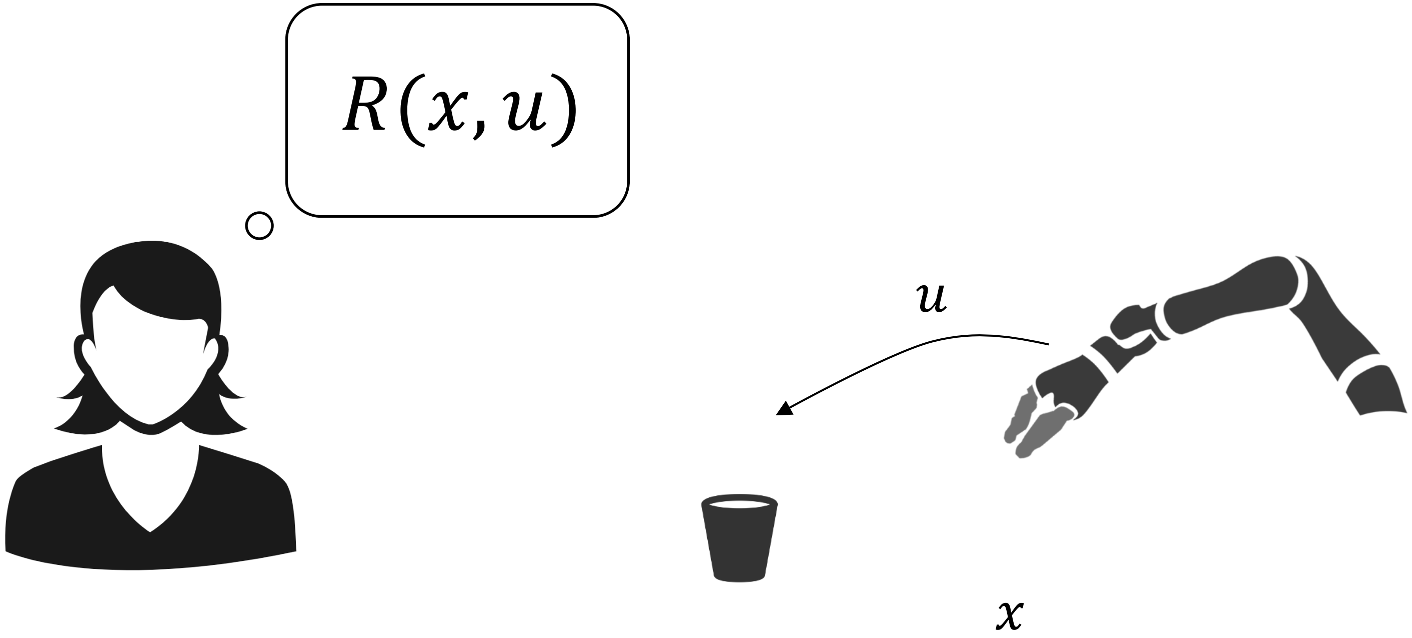
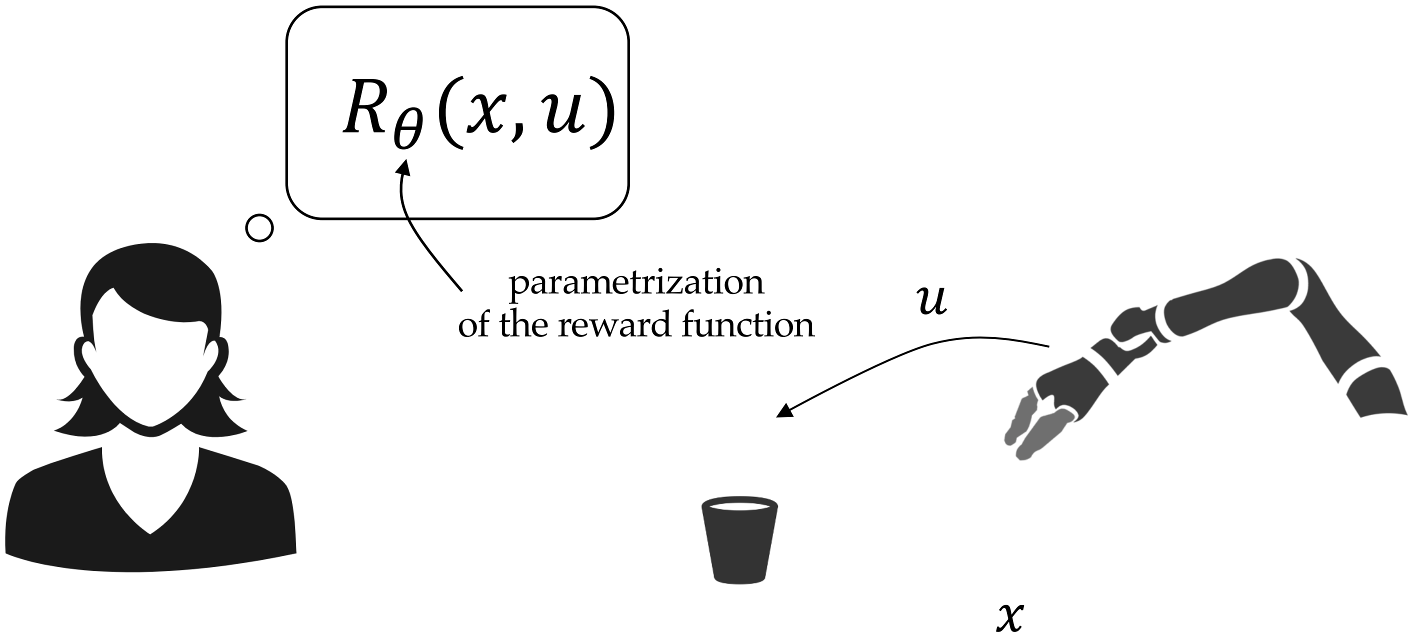
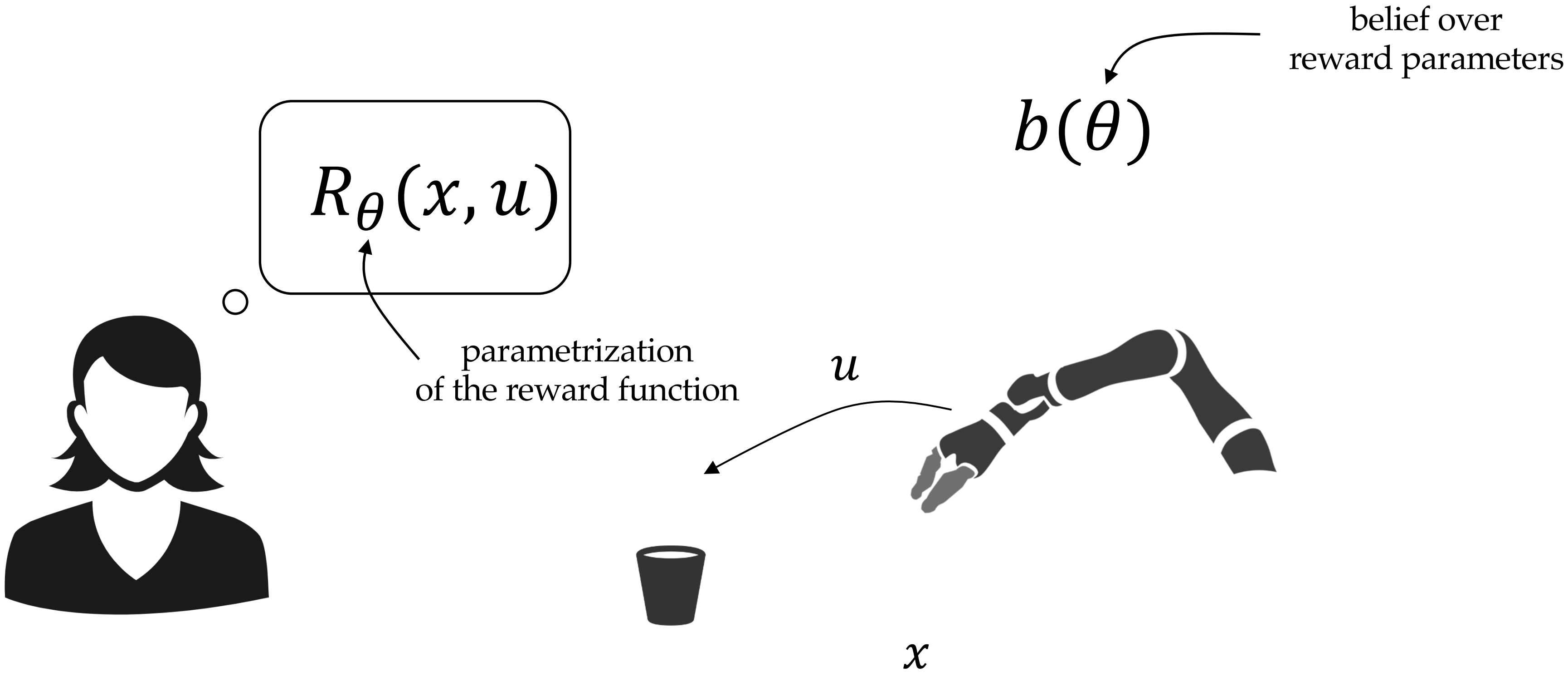
Learning Objectives and Preferences:

How? Actively

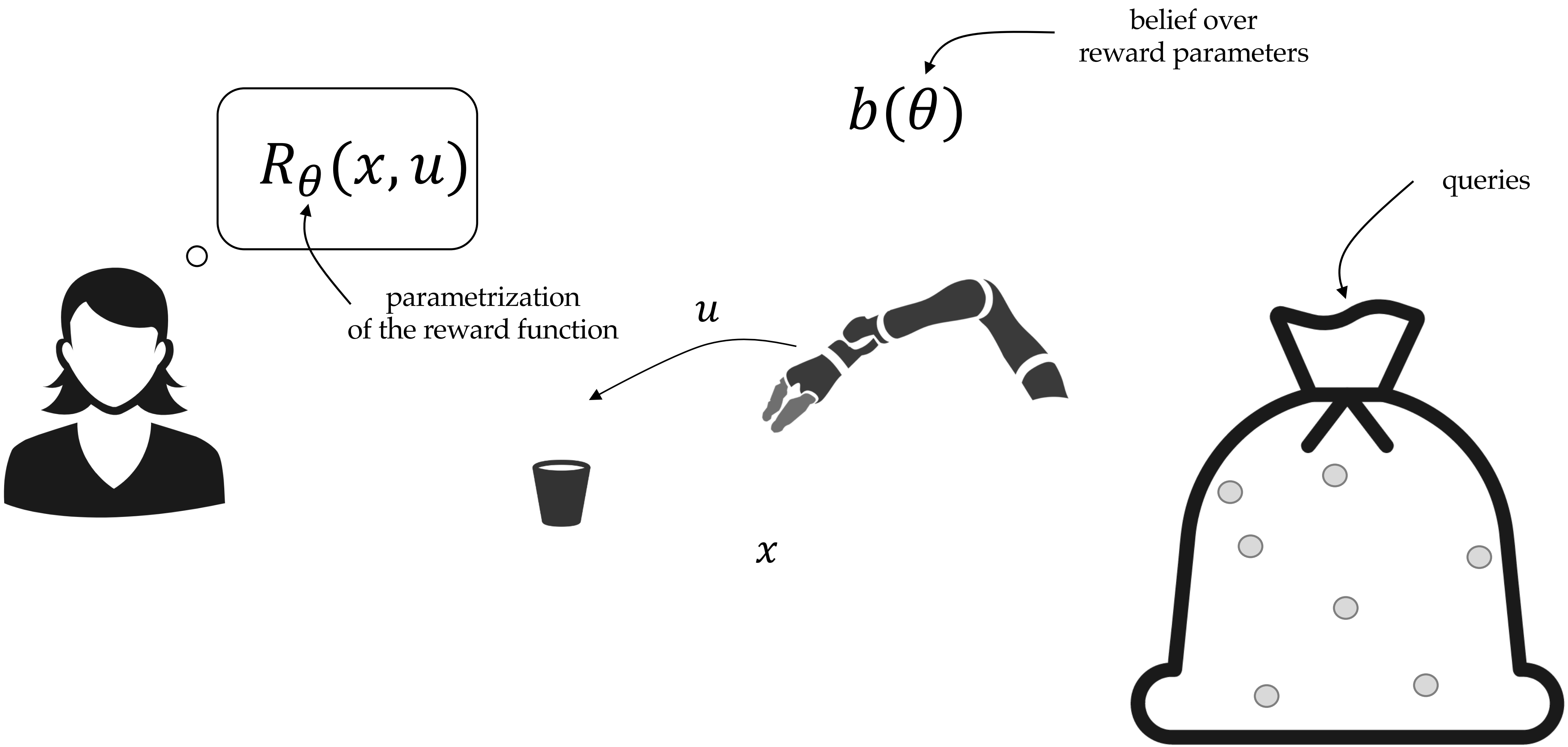

$R(x, u)$
u
x



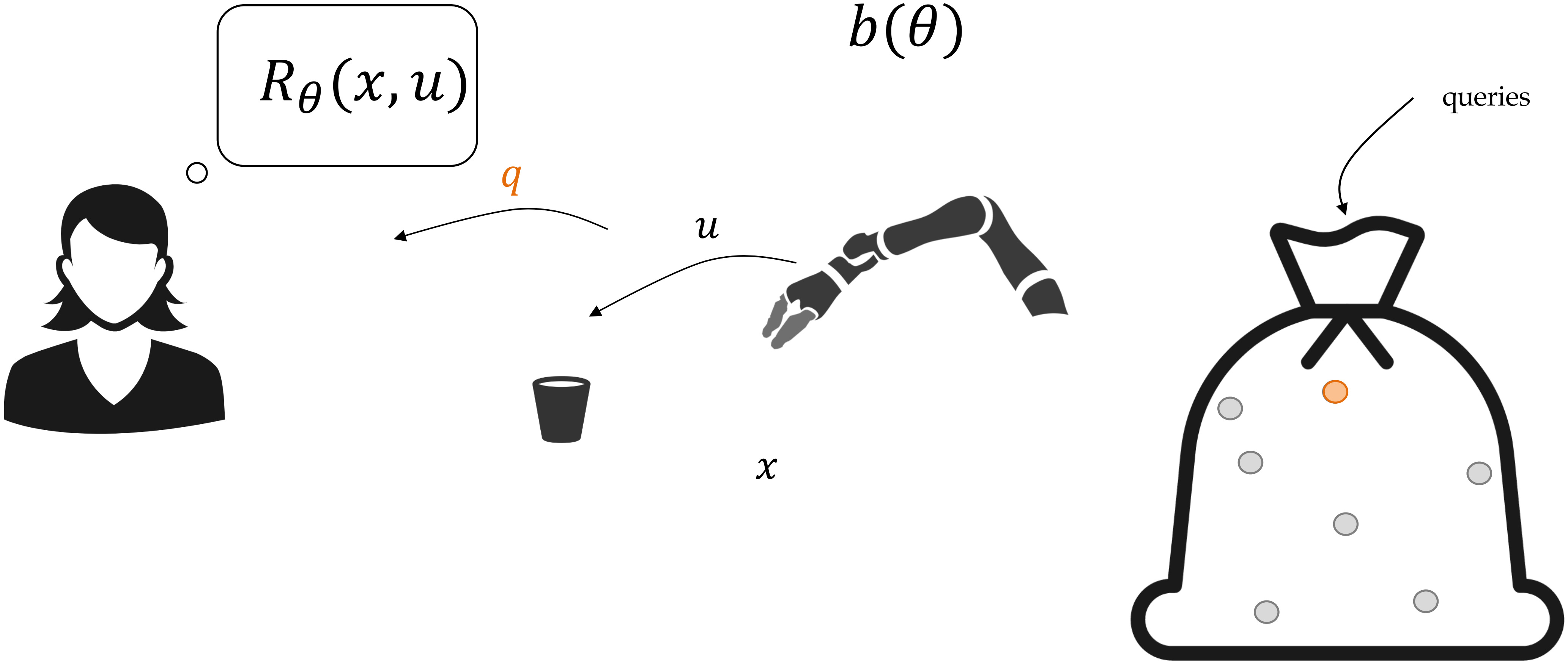
$R_\theta(x, u)$
parametrization
of the reward function
u
x



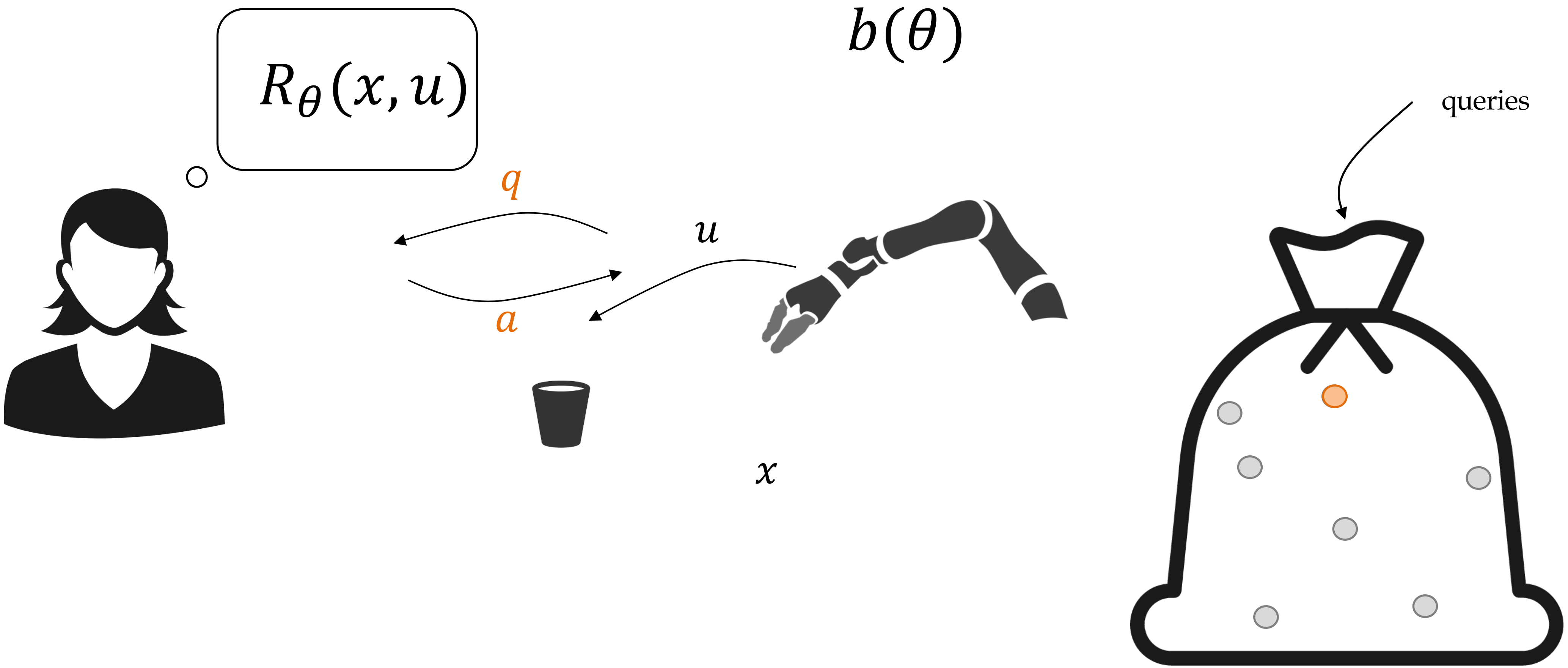
belief over
reward parameters
$b(\theta)$
$R_\theta(x, u)$
parametrization
of the reward function
u
x



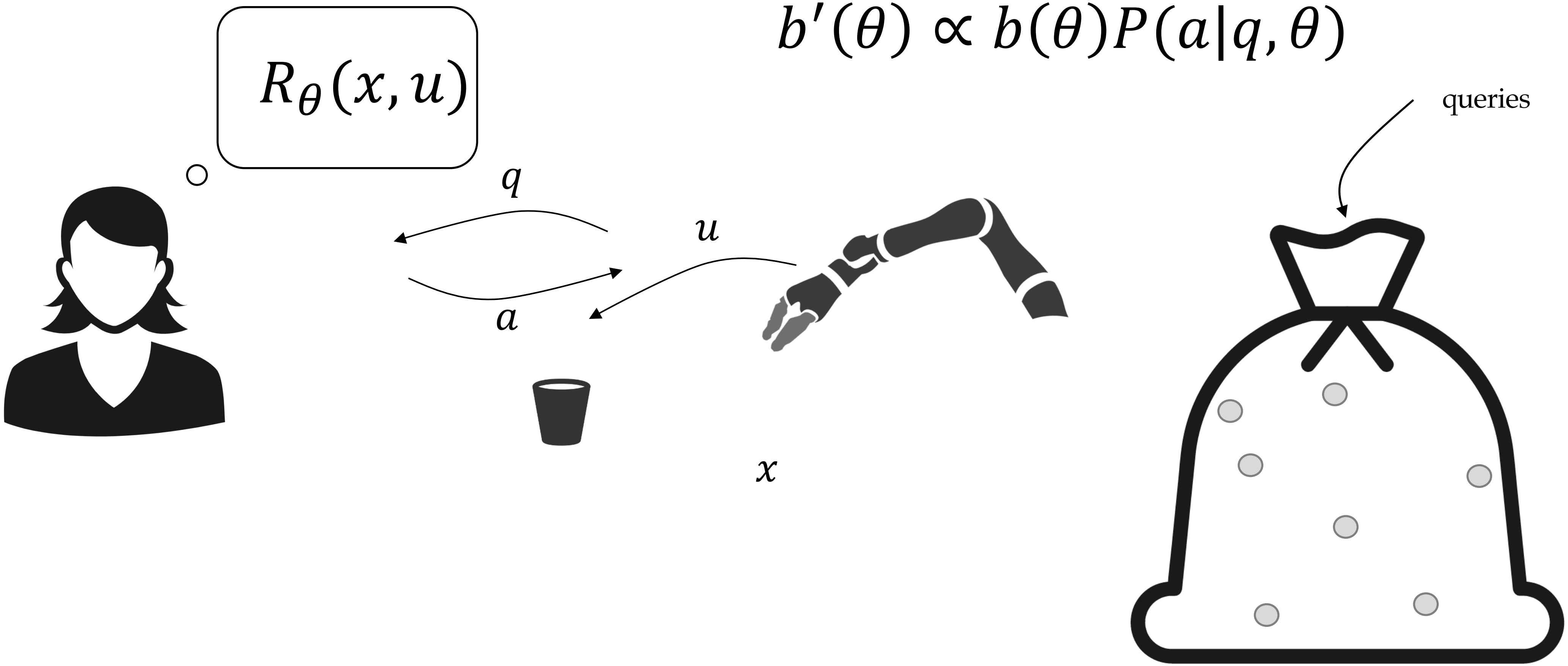
$R_\theta(x, u)$
parametrization of the reward function
belief over reward parameters
$b(\theta)$
queries
u
x



$R_\theta(x, u)$
q
u
$b(\theta)$
queries
x



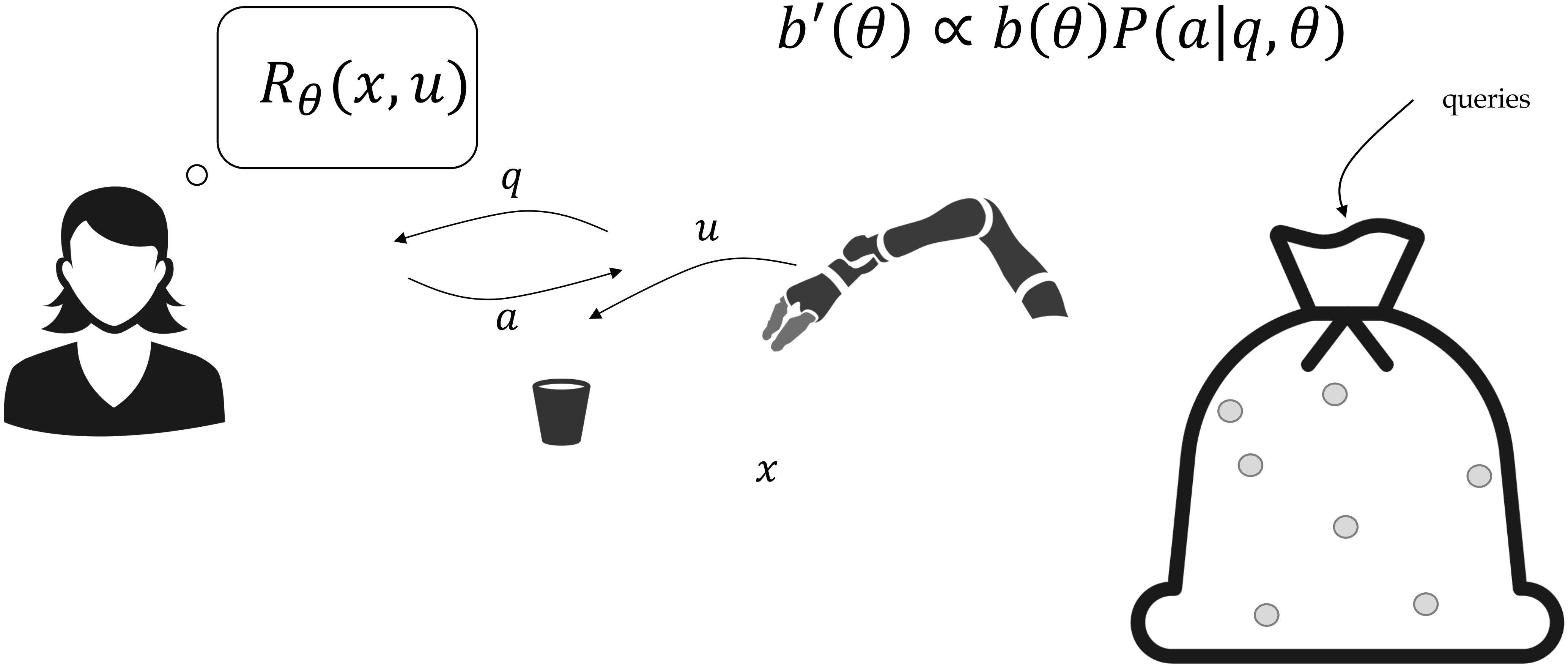
$R_\theta(x, u)$
$b(\theta)$
queries
q
a
u
x

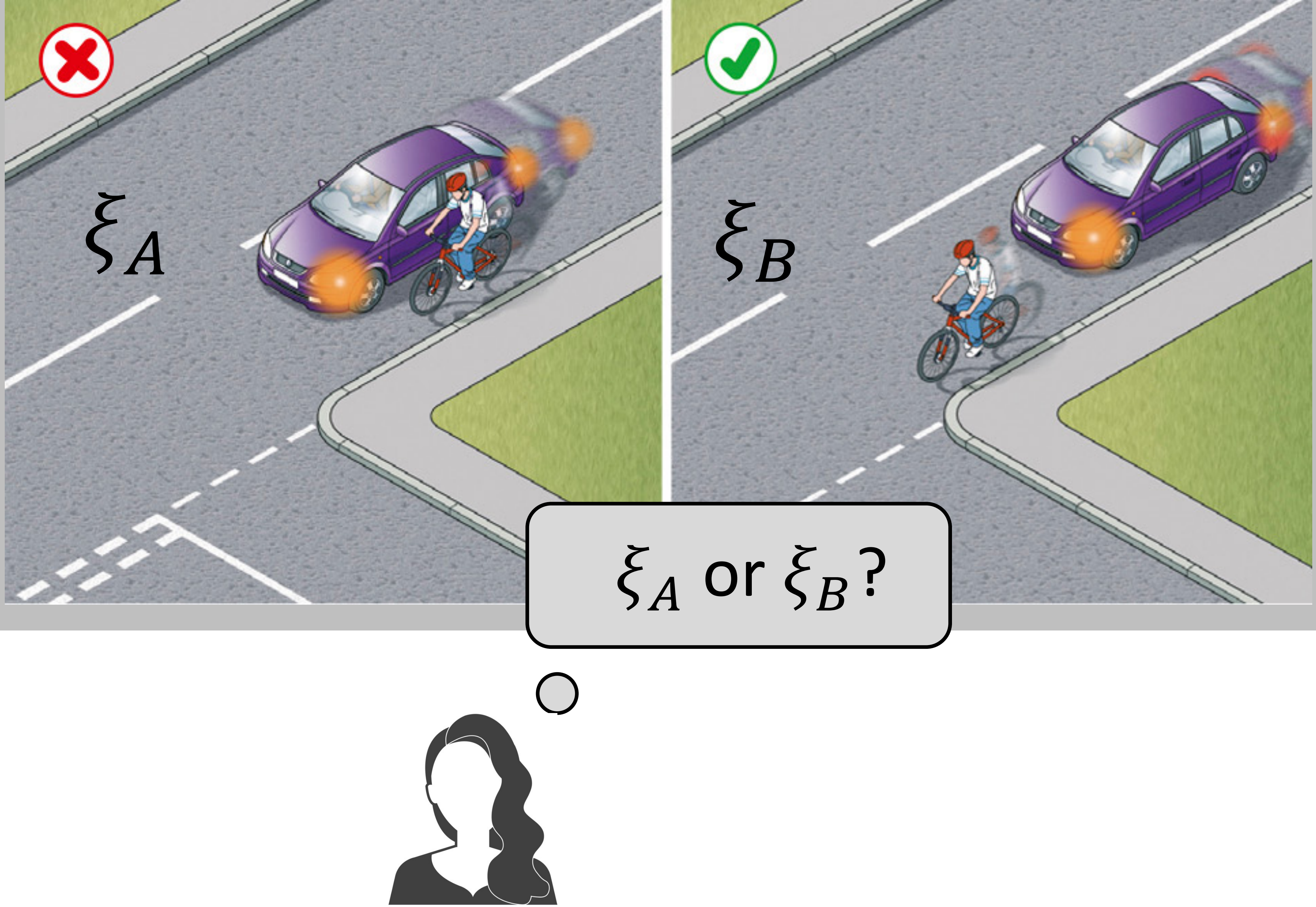


$b'(\theta) \propto b(\theta)P(a|q,\theta)$
$R_\theta(x,u)$
queries
q
u
a
x

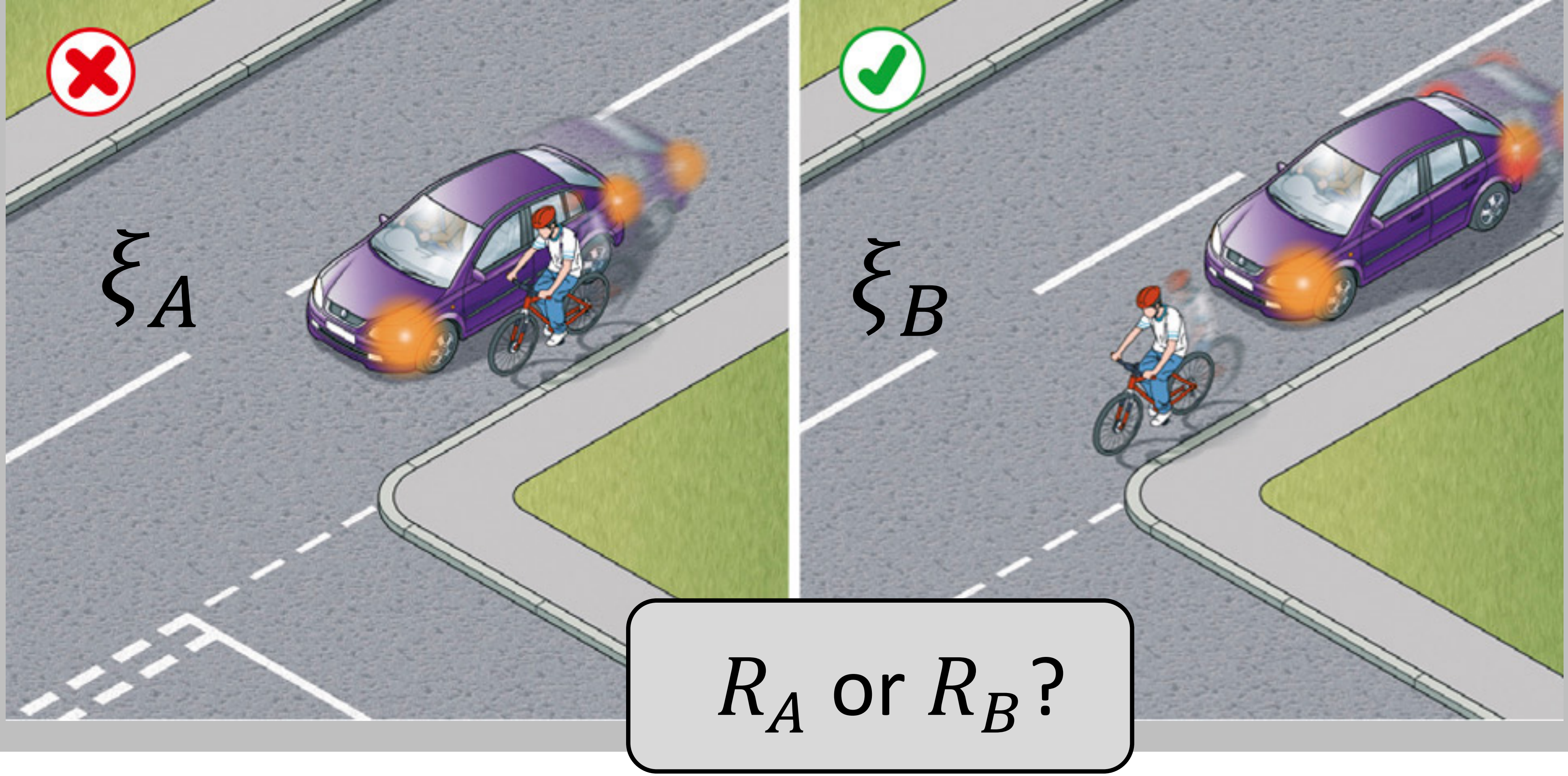
Where do queries come from?

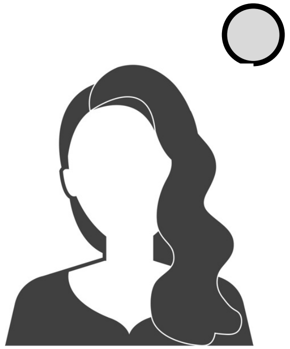
$$b'(\theta) \propto b(\theta)P(a|q,\theta)$$



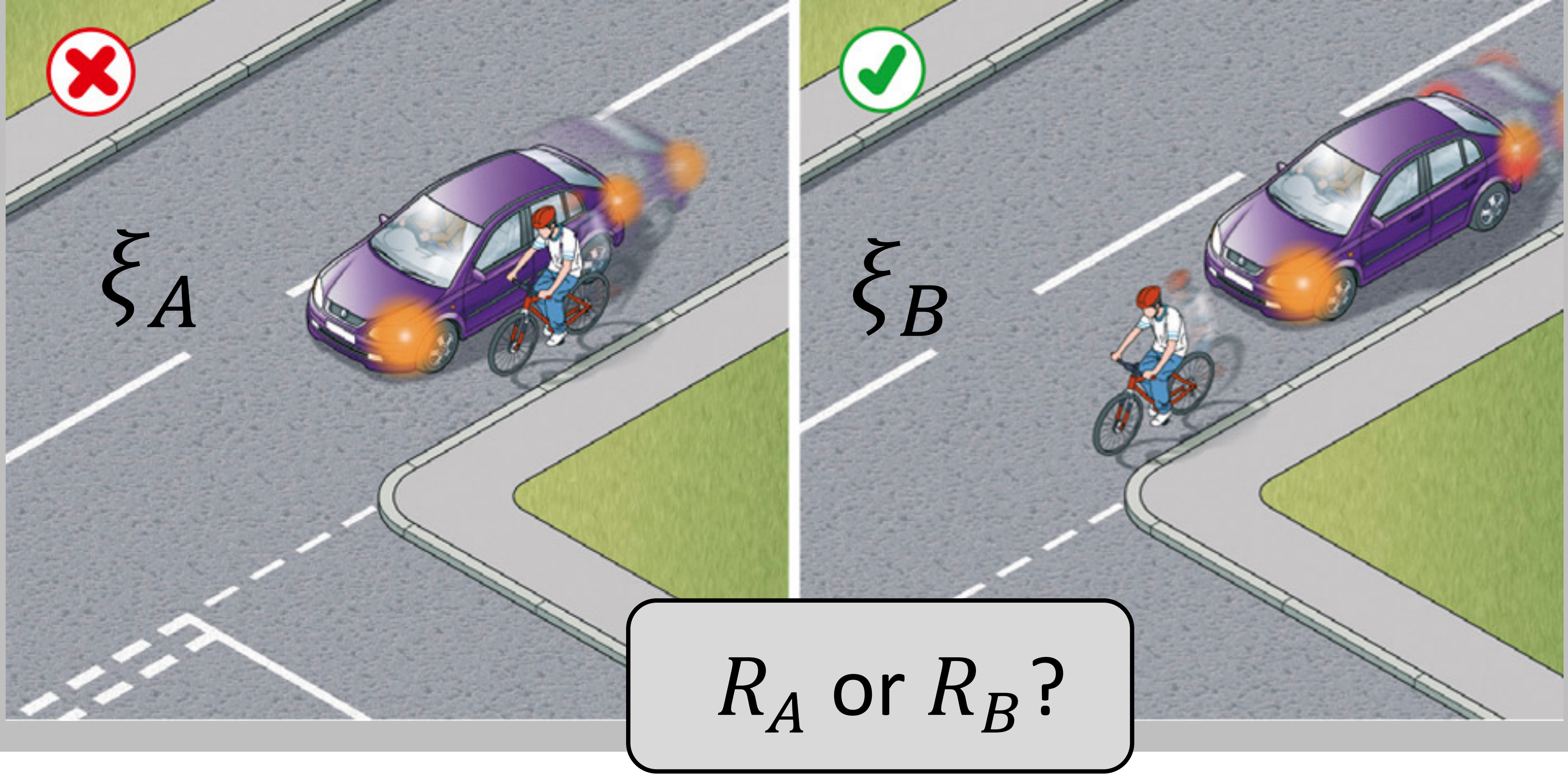


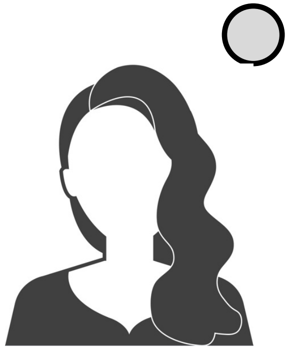
ξ_A
ξ_B
ξ_A or ξ_B?





$$R = \boldsymbol{\theta} \cdot \phi(\xi)$$

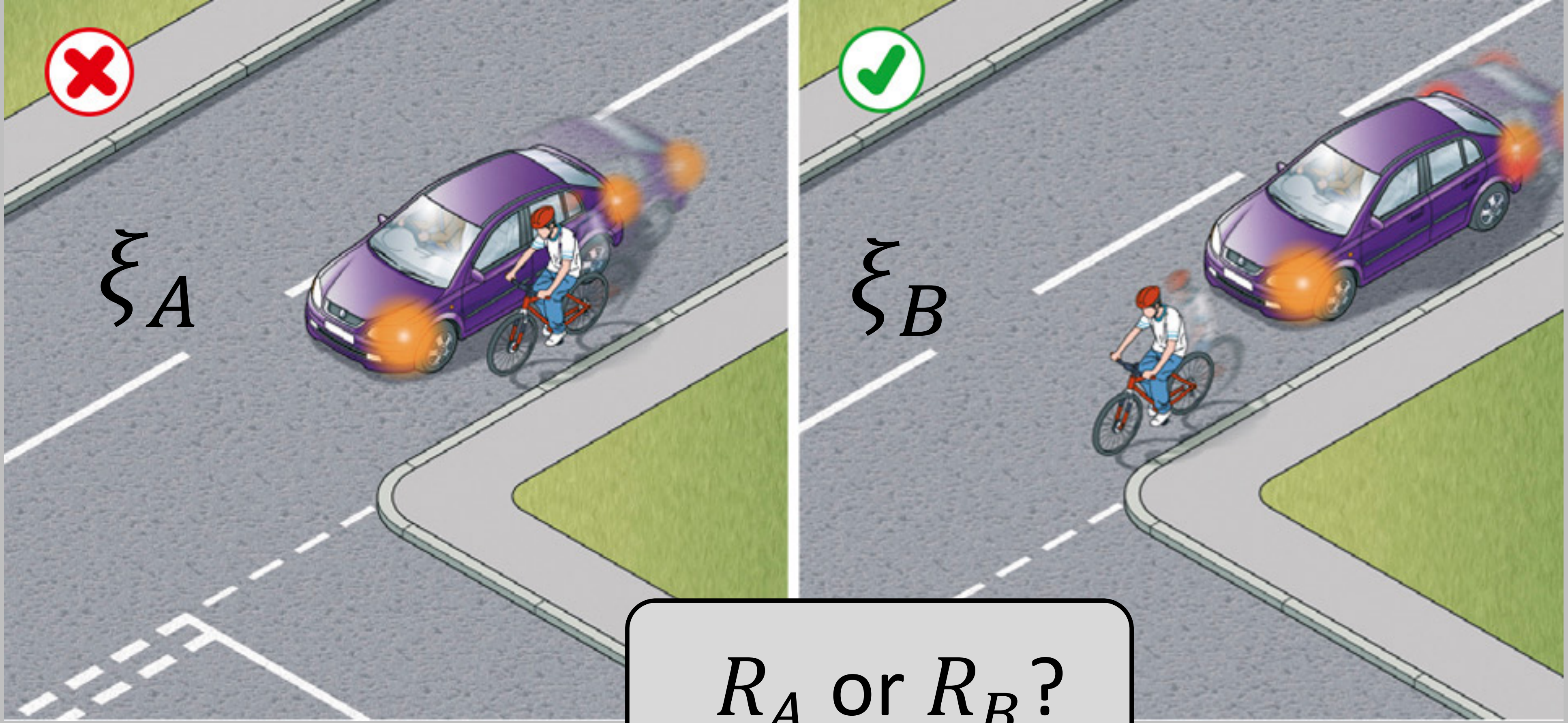




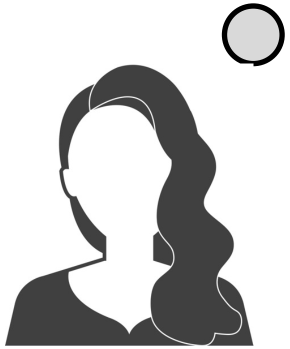
$$R = \boldsymbol{\theta} \cdot \phi$$

$$\boldsymbol{\theta} \cdot \phi_A > \boldsymbol{\theta} \cdot \phi_B$$

or $$\boldsymbol{\theta} \cdot \phi_A < \boldsymbol{\theta} \cdot \phi_B$$



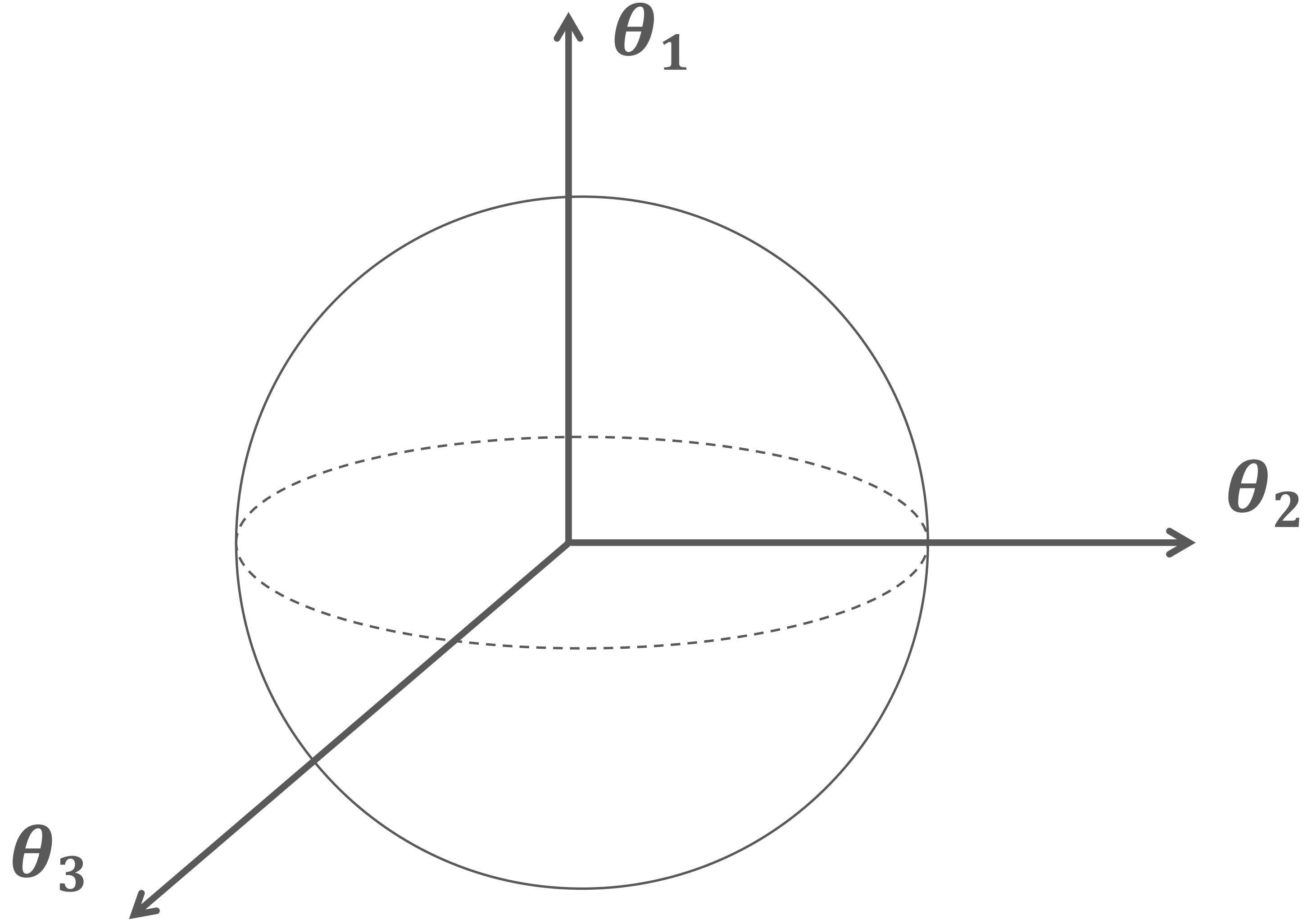
R_A or R_B?



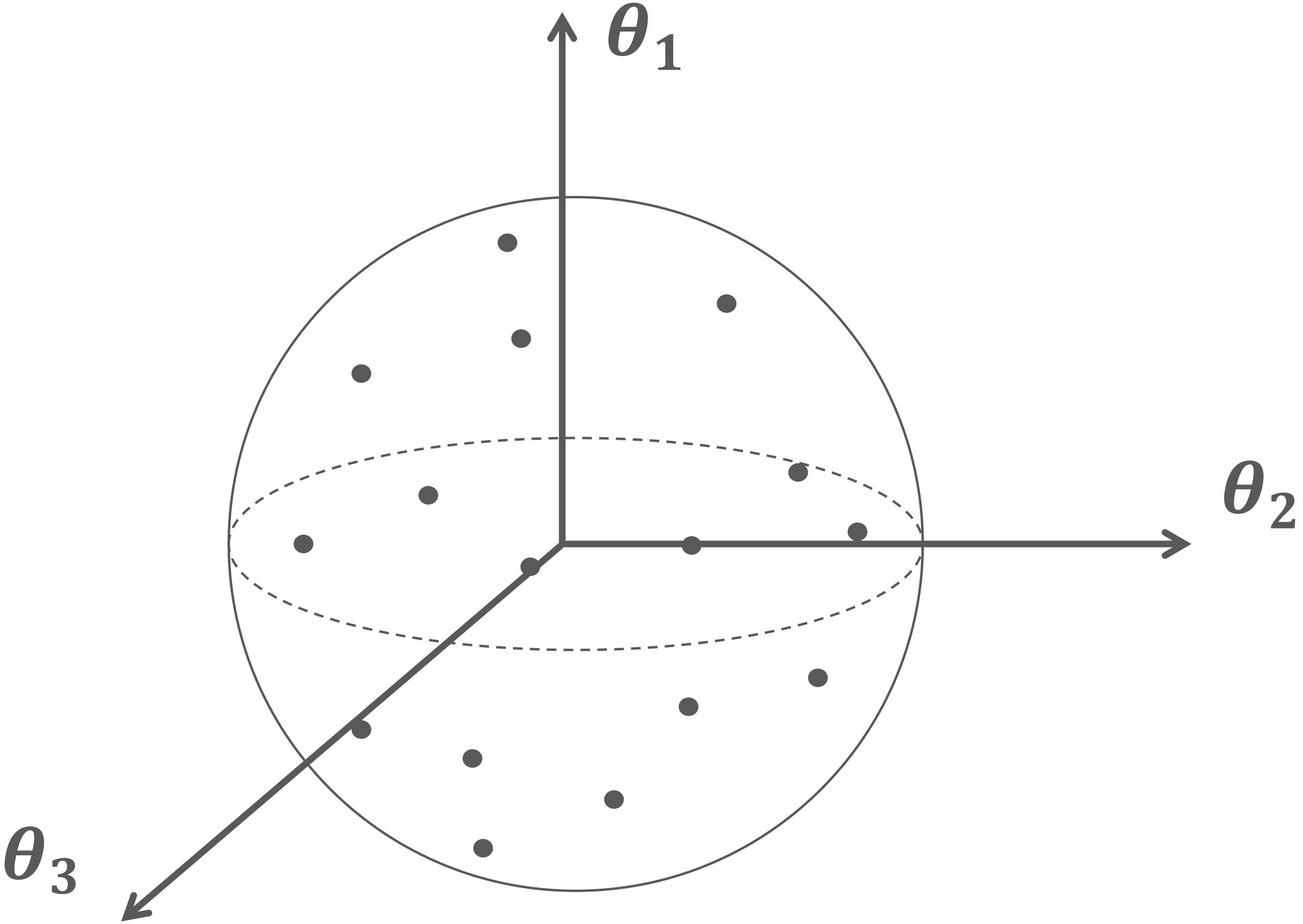
$$R = \boldsymbol{\theta} \cdot \phi$$

$$\psi = (\phi_A - \phi_B)$$

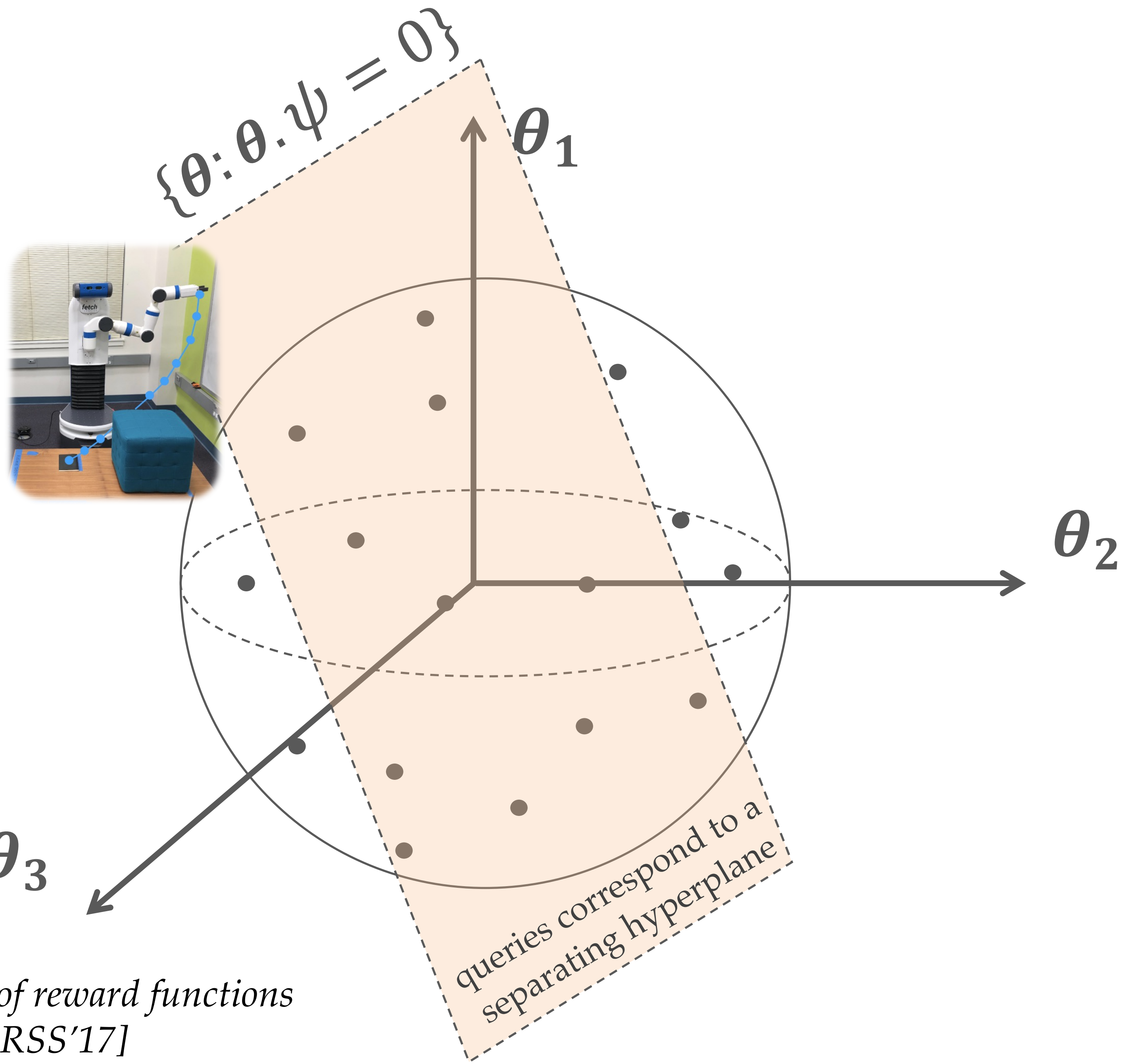
$$\boldsymbol{\theta} \cdot \psi > 0 \text{ or } \boldsymbol{\theta} \cdot \psi < 0$$



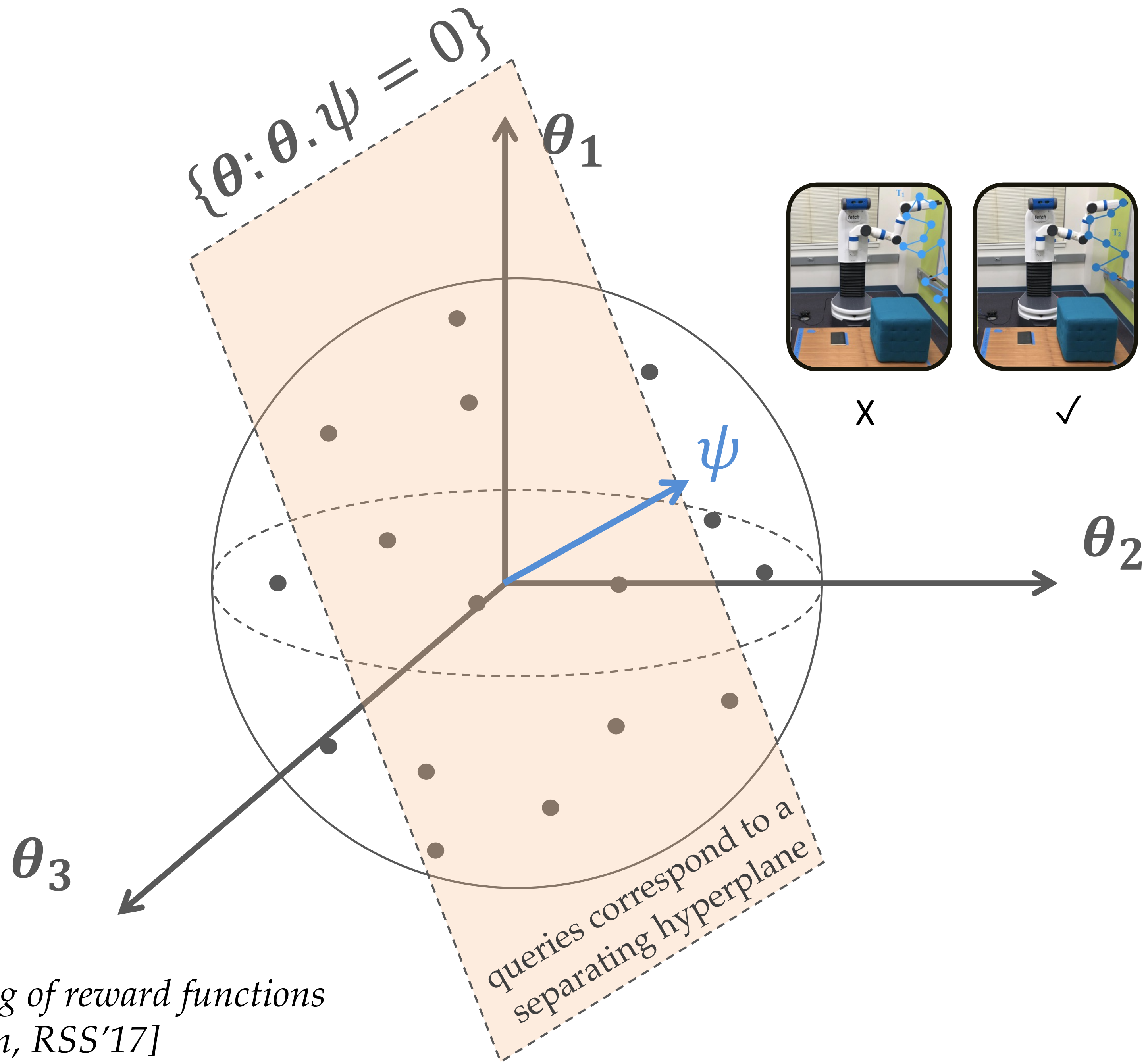
Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]



Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]

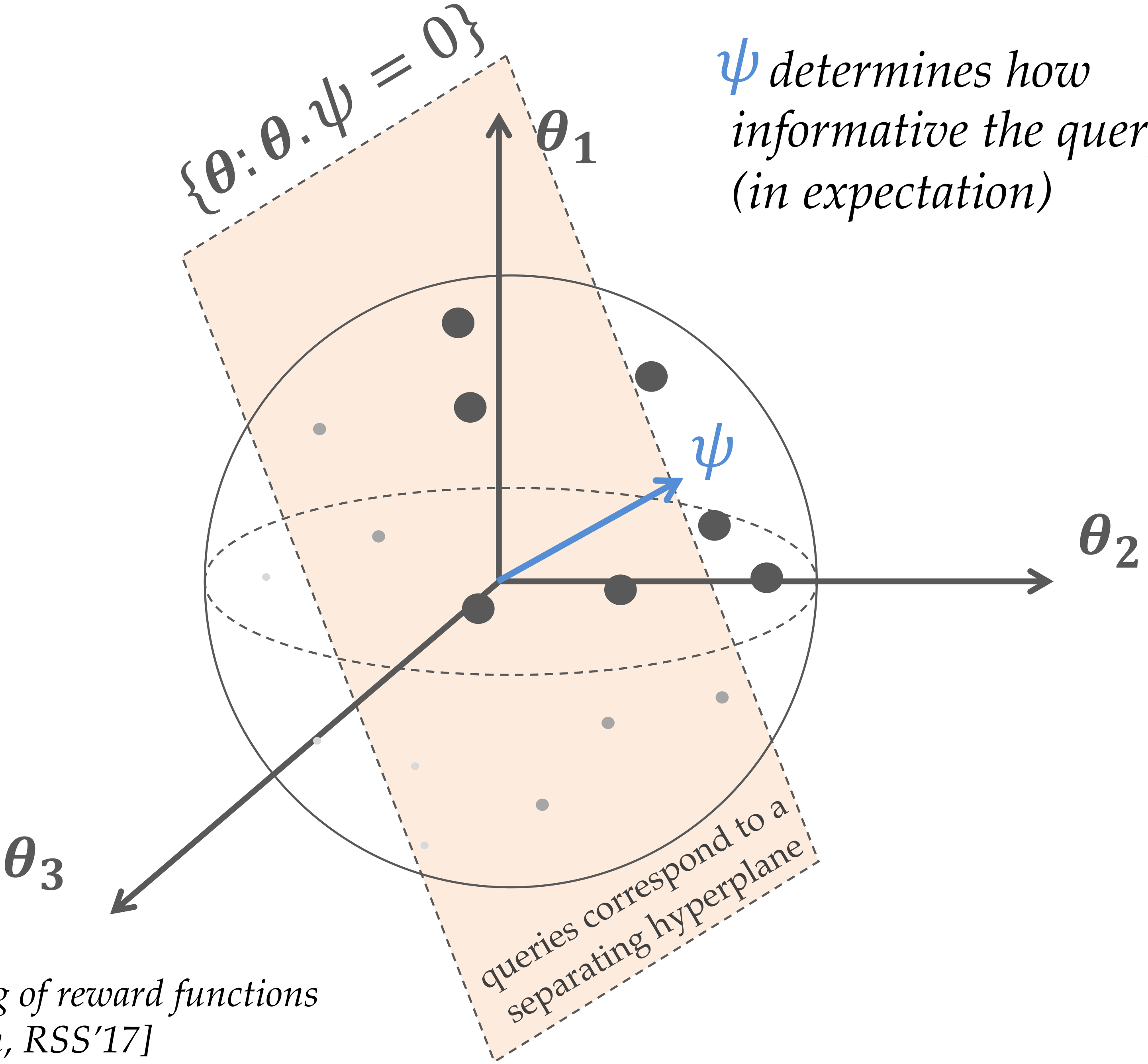


Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]

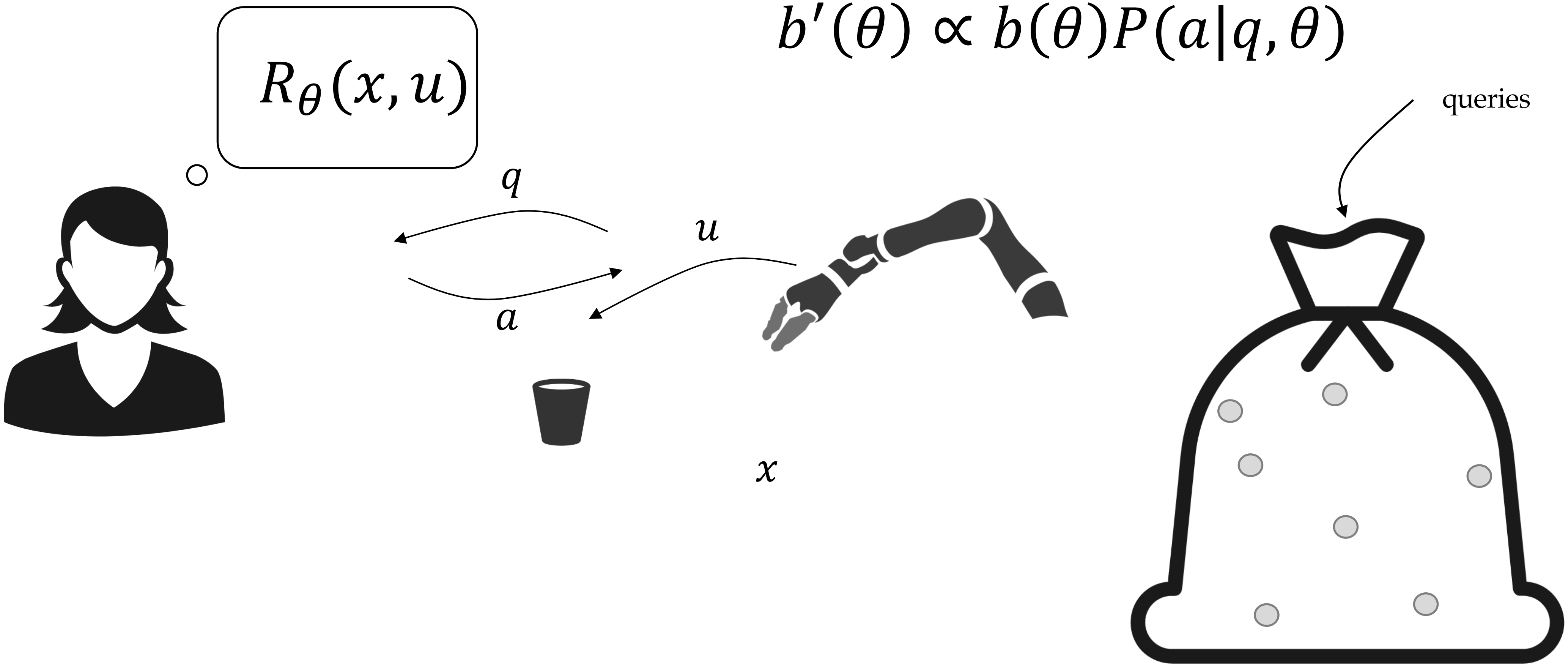


Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]

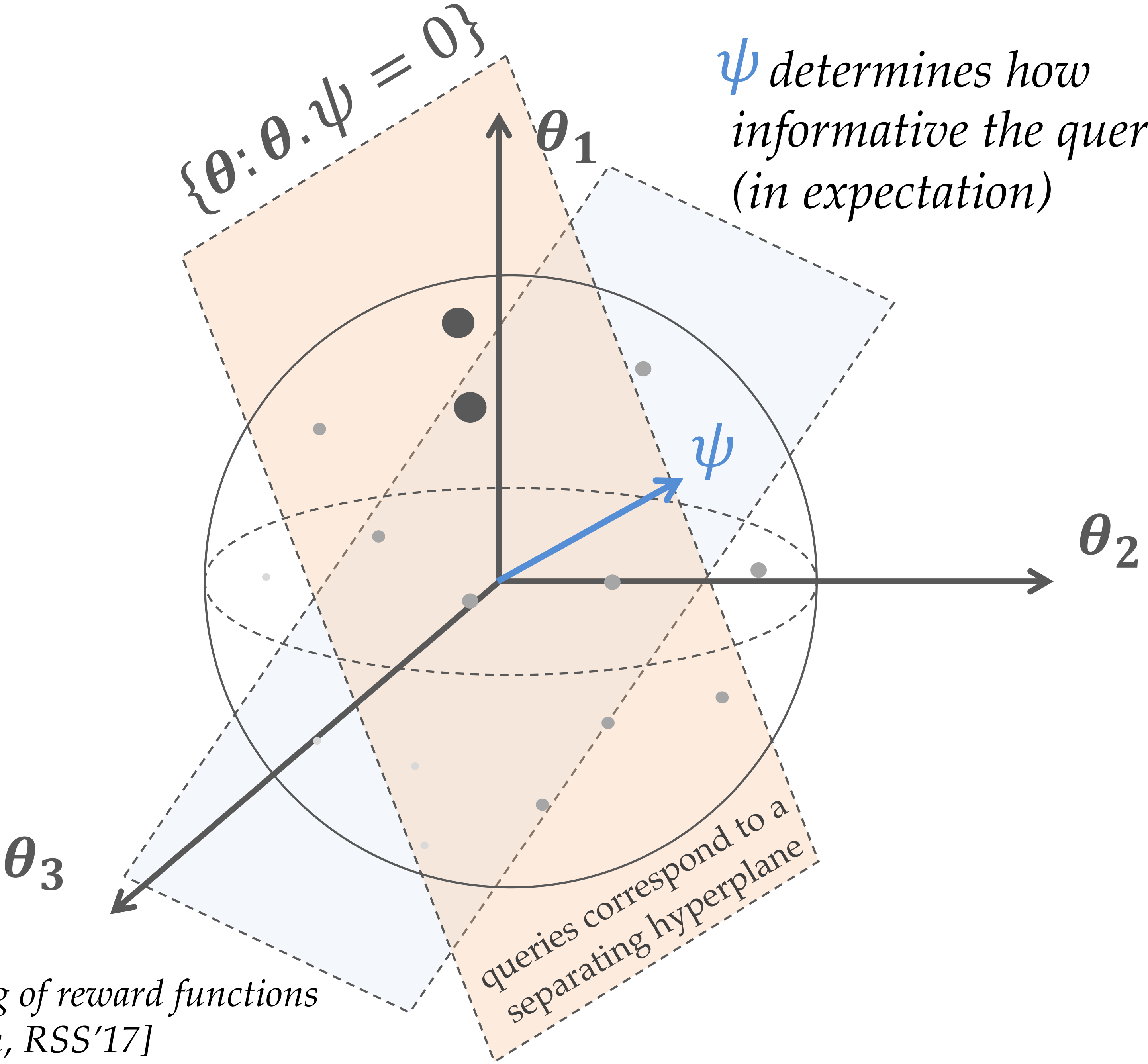
ψ *determines how informative the query is (in expectation)*



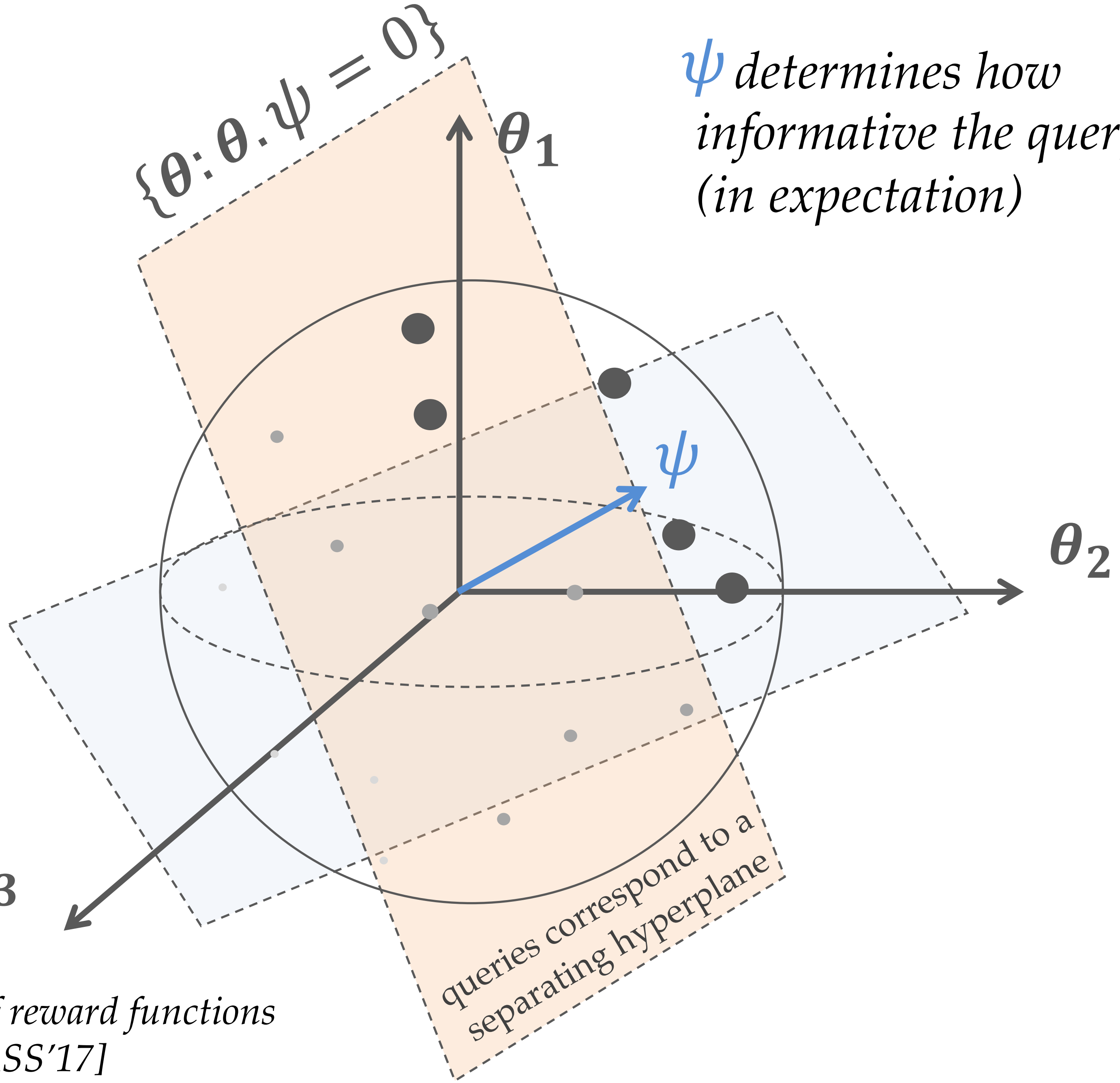
Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]


$R_\theta(x, u)$
q
a
u
x
$b'(\theta) \propto b(\theta)P(a|q, \theta)$
queries

ψ determines how informative the query is (in expectation)

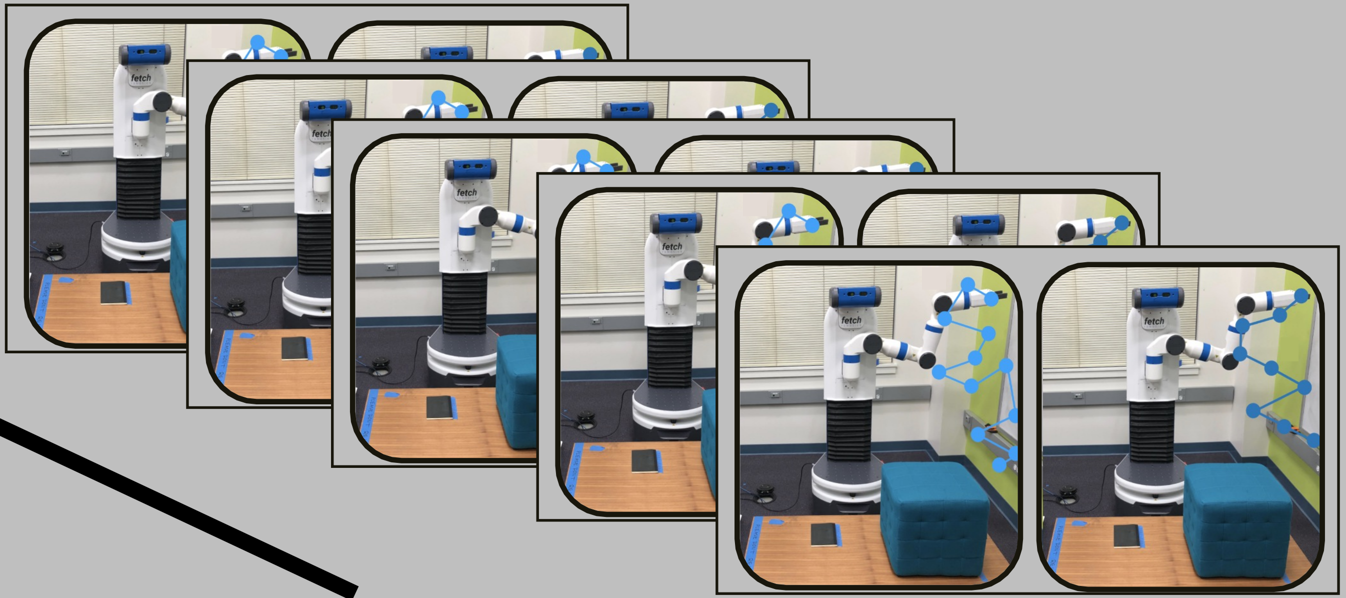


Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]

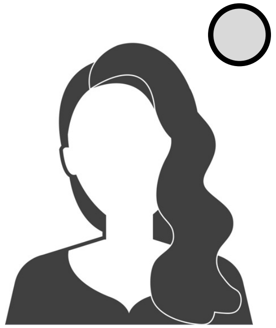


ψ determines how informative the query is (in expectation)

Active preference based learning of reward functions [Sadigh, Seshia, Sastry, Dragan, RSS'17]

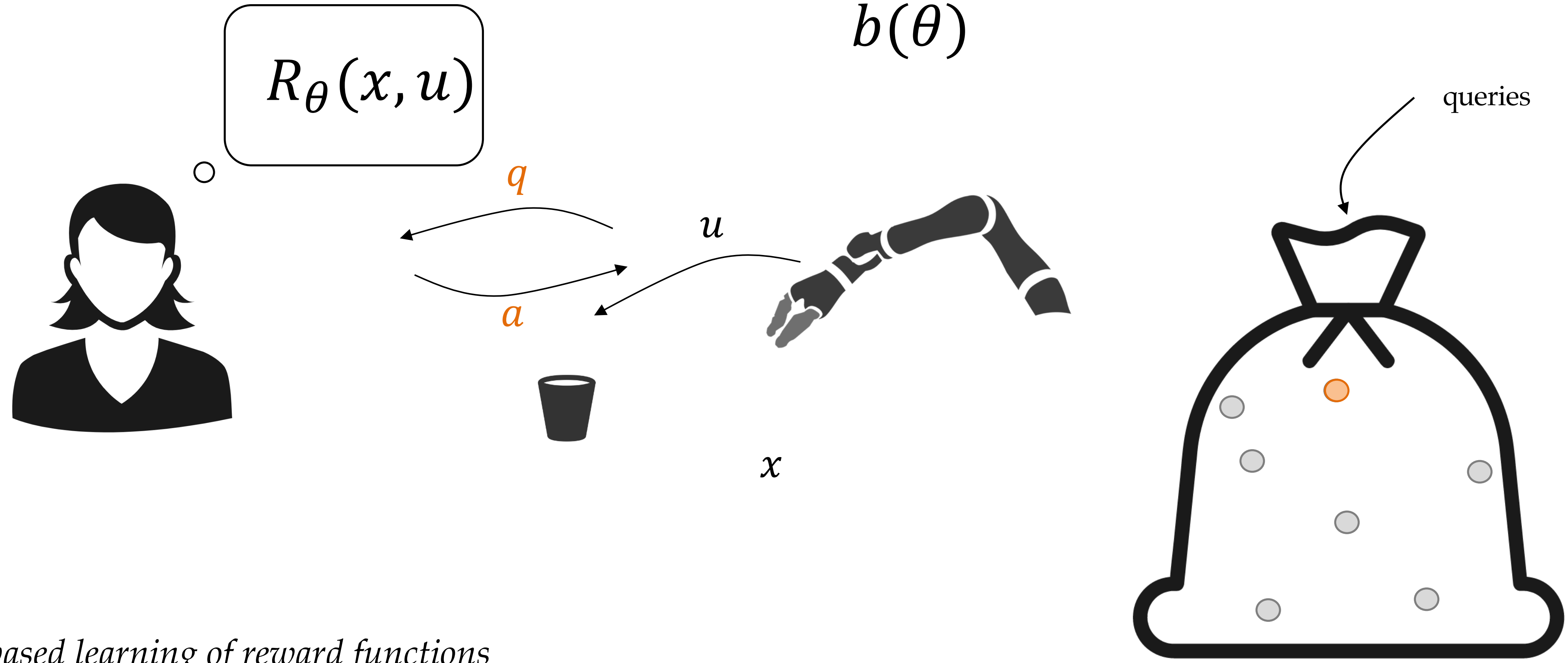


Most informative,
diverse sequence of queries



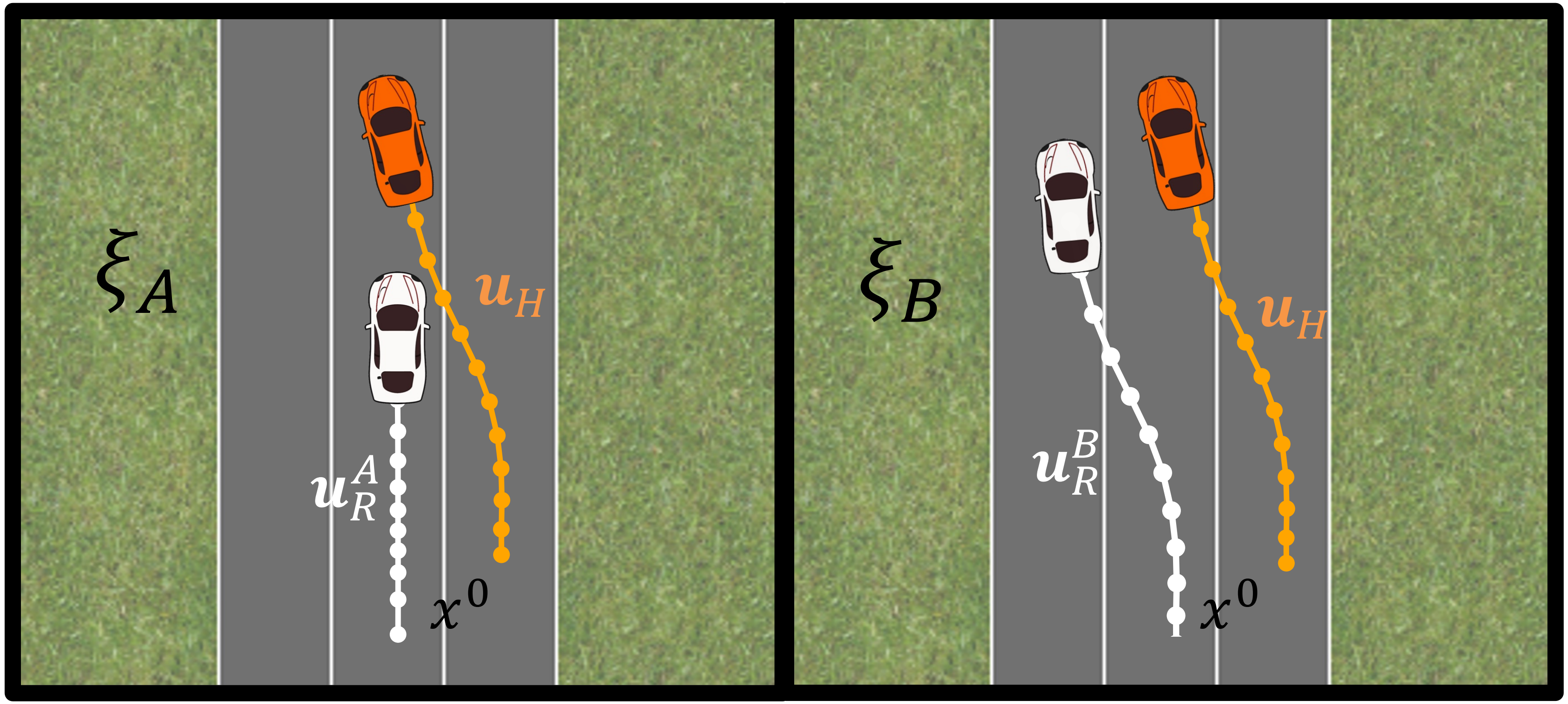
ξ_A or ξ_B?

Queries should be actively selected.

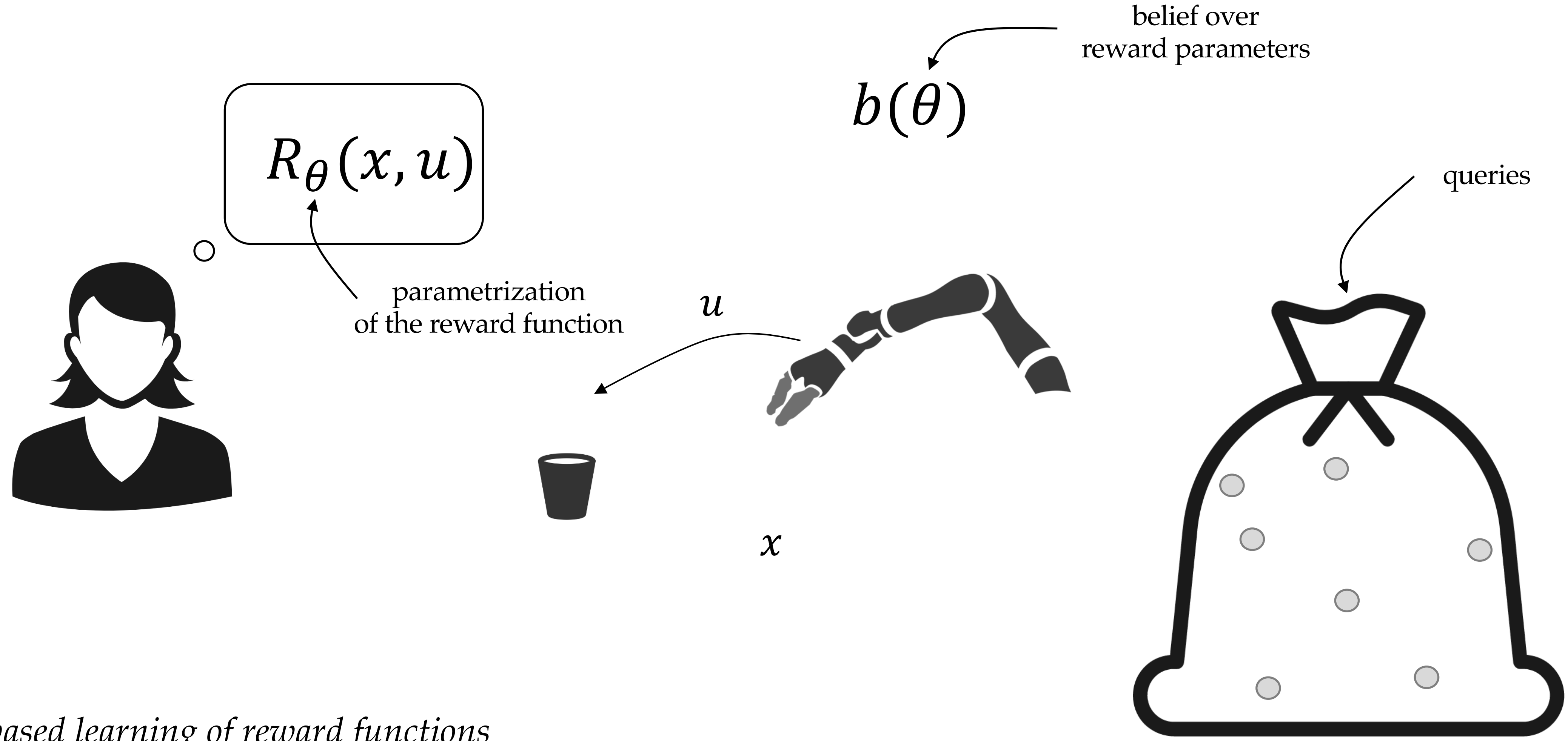


Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]

Challenge:
Queries lie in a continuous and high-dimensional space.

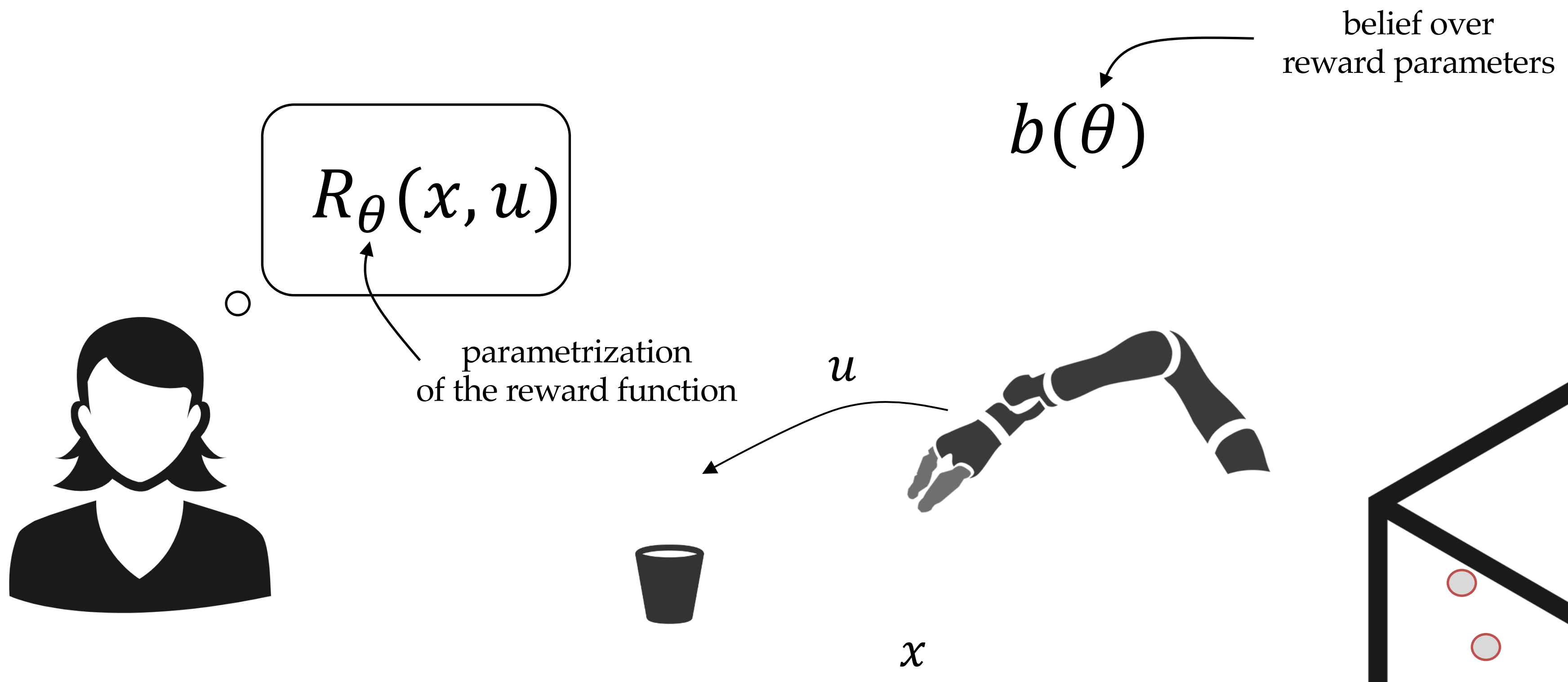


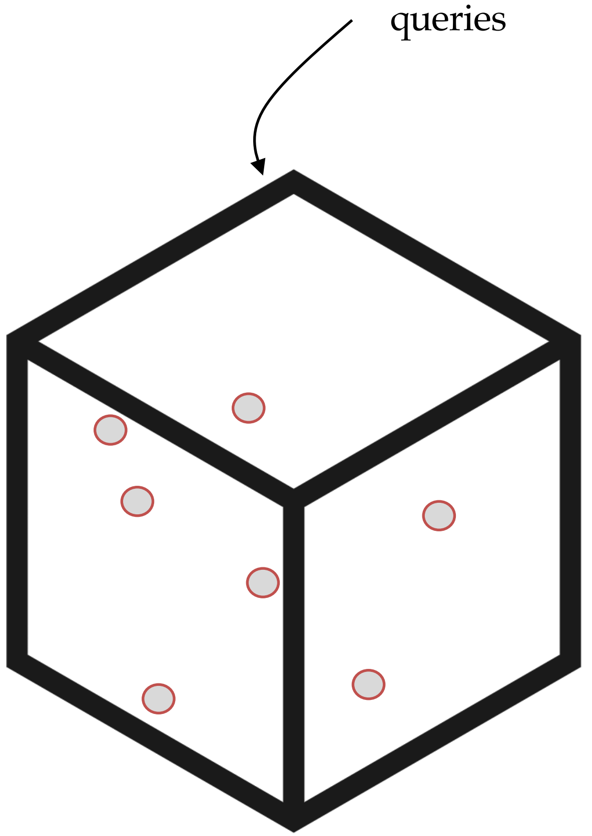
Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]



Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]

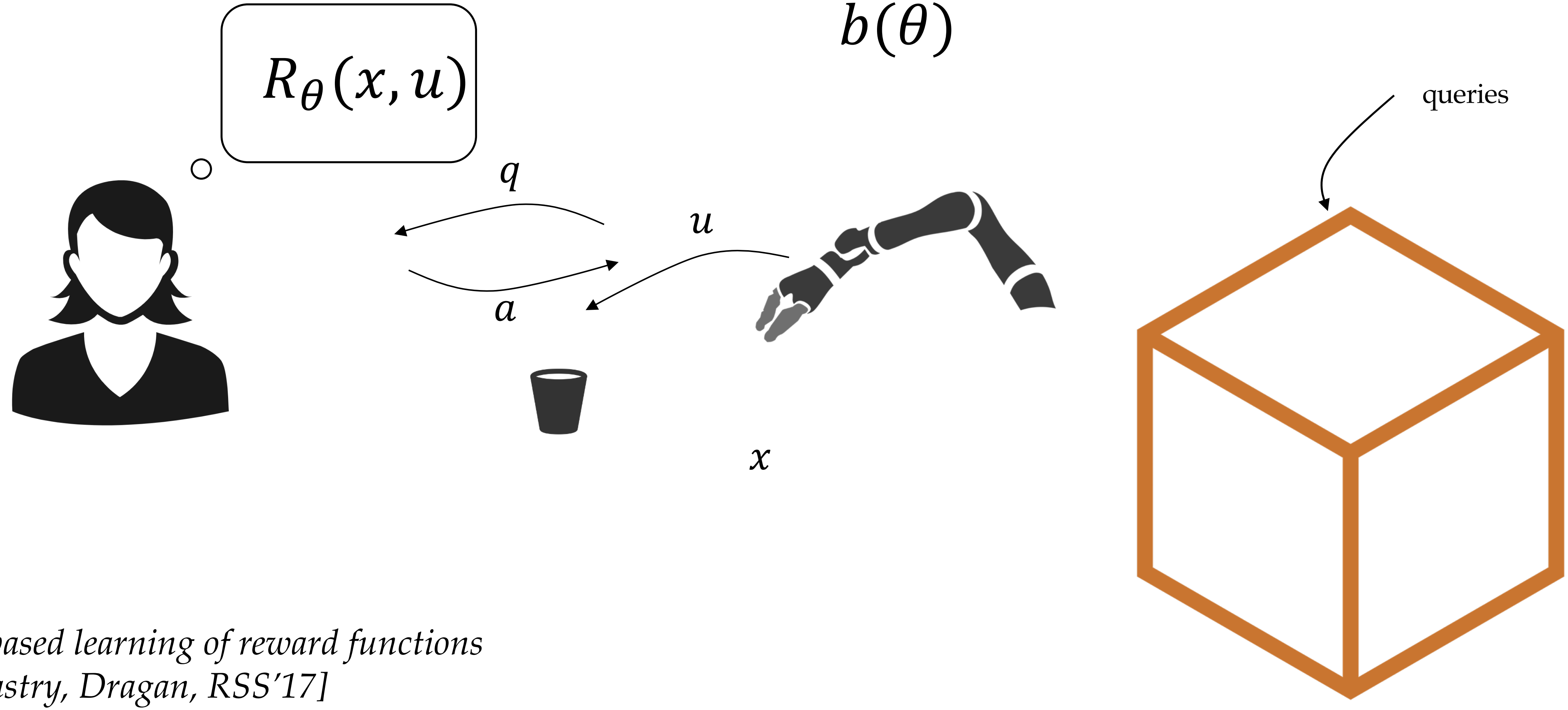
Real robots don't get handed a neat query set





Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]

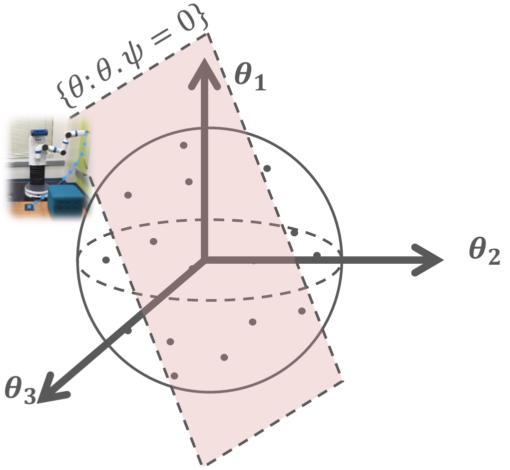
They have to synthesize their queries from scratch.



Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]

Queries should be actively synthesized.

Actively *synthesizing* queries



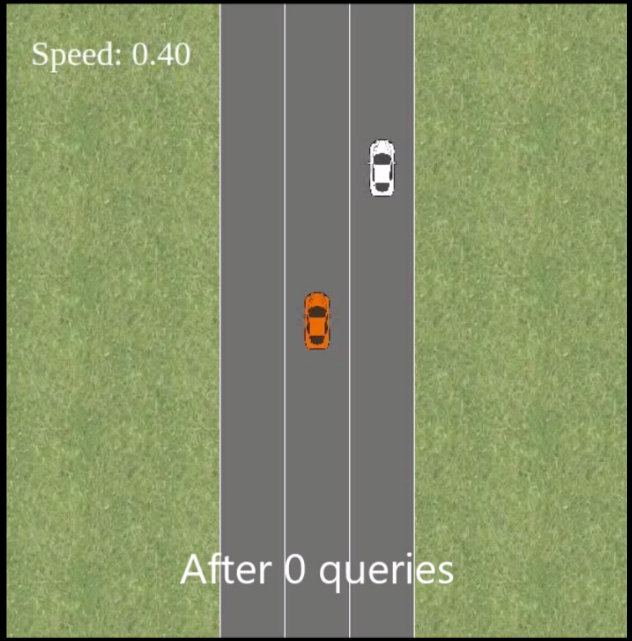
$$\max_{\psi} \quad \overbrace{\min\{\mathbb{E}[1 - f_{\psi}(\theta)], \mathbb{E}[1 - f_{-\psi}(\theta)]\}}^{\textit{minimum volume removed}}$$

Subject to $\psi \in \mathbb{F}$

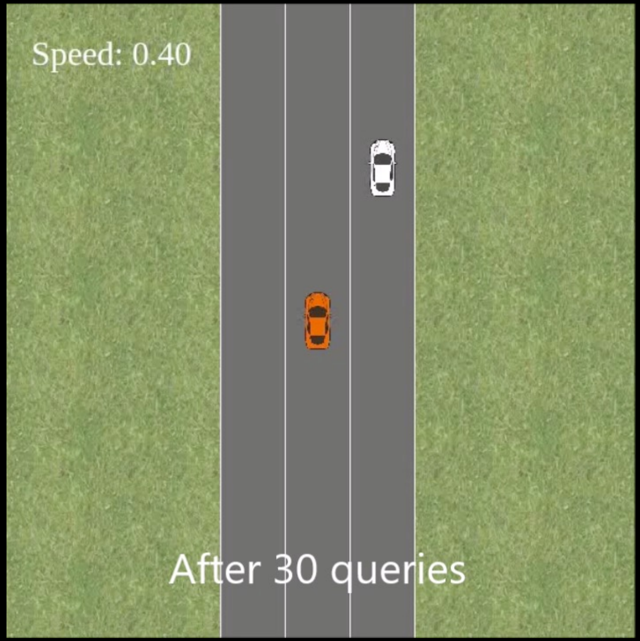
$$\mathbb{F} = \{\psi : \psi = \Phi(\xi_A) - \Phi(\xi_B), \xi_A, \xi_B \in \Xi\}$$

Human update function $\quad f_{\psi}(\boldsymbol{\theta}) = \min(1, \exp(I_t \boldsymbol{\theta}^{\top} \psi))$

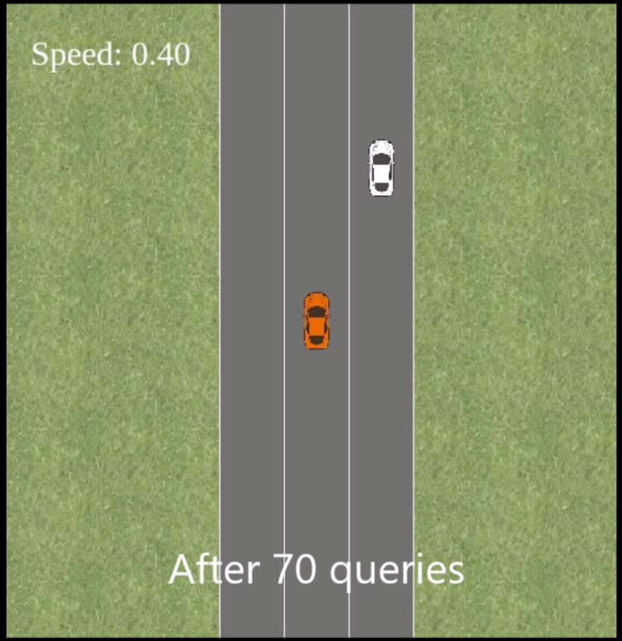
Active preference based learning of reward functions
[Sadigh, Seshia, Sastry, Dragan, RSS'17]



No prior preference



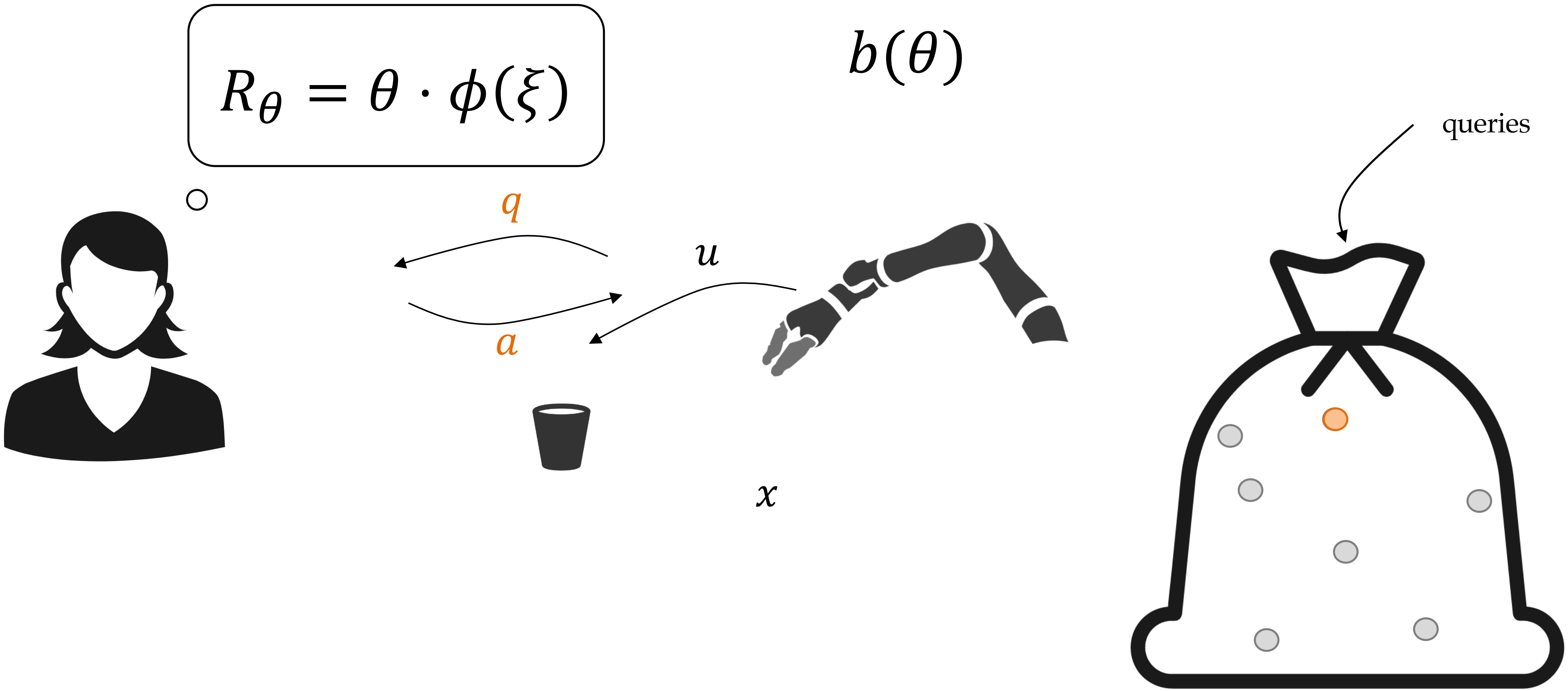
Learns *heading* preferences

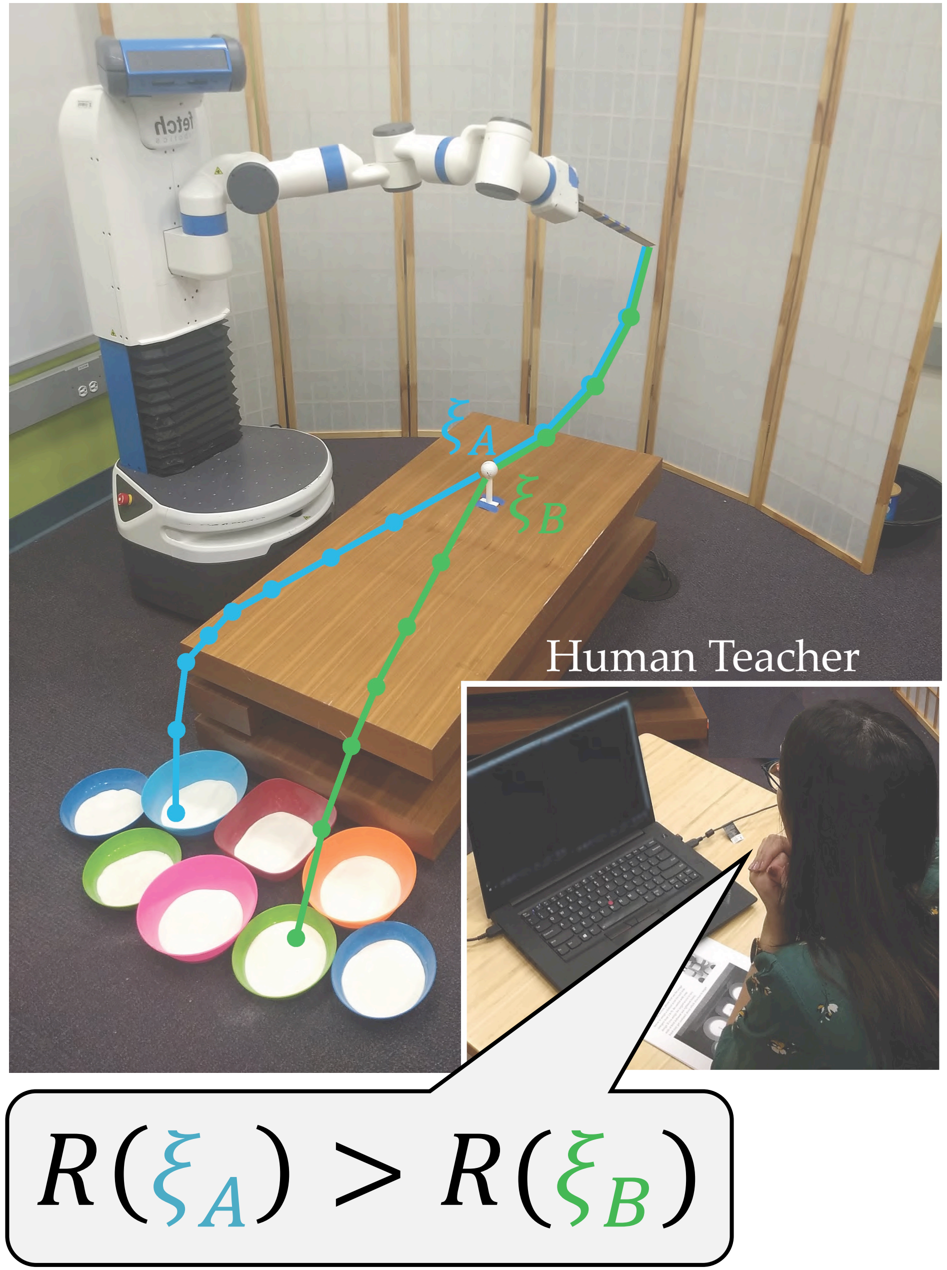


Learns *collision avoidance* preferences

Batch active preference based learning of reward functions
[Biyik, Sadigh, CoRL'18]

Queries should be actively synthesized.





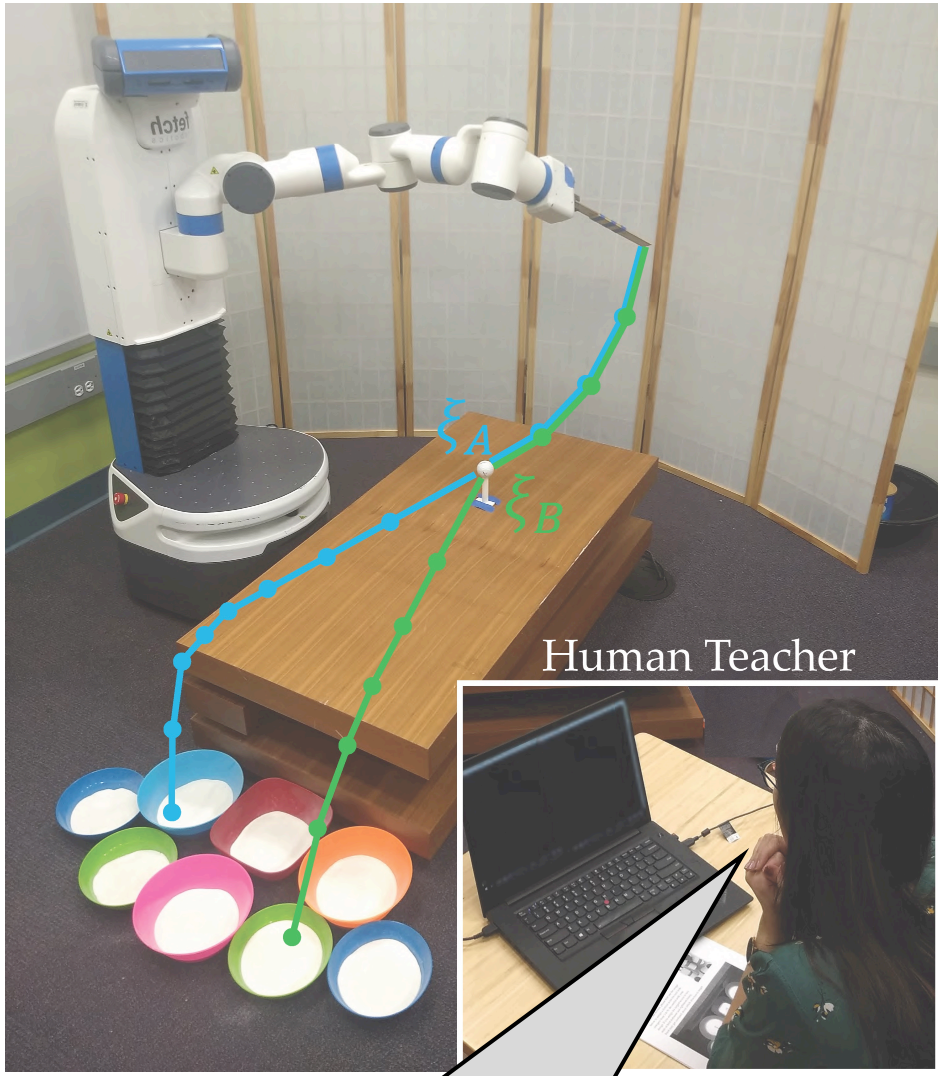
ξ_A
ξ_B
Human Teacher
$R(\xi_A) > R(\xi_B)$

$$R(\xi_A) = \theta \cdot \phi(\xi_A)$$

Designing features is hard.

$e^{-c_4 d_4}$ where d_4 is the final horizontal distance between the object and the center of the closest basket, and $c_4 = 3$.

The average of $e^{-c_2 d_2^2 - c_3 d_3^2}$ over the trajectory, where d_2 and d_3 are the horizontal and vertical distances between the ego car and the other car, respectively; and $c_2 = 7$, $c_3 = 3$

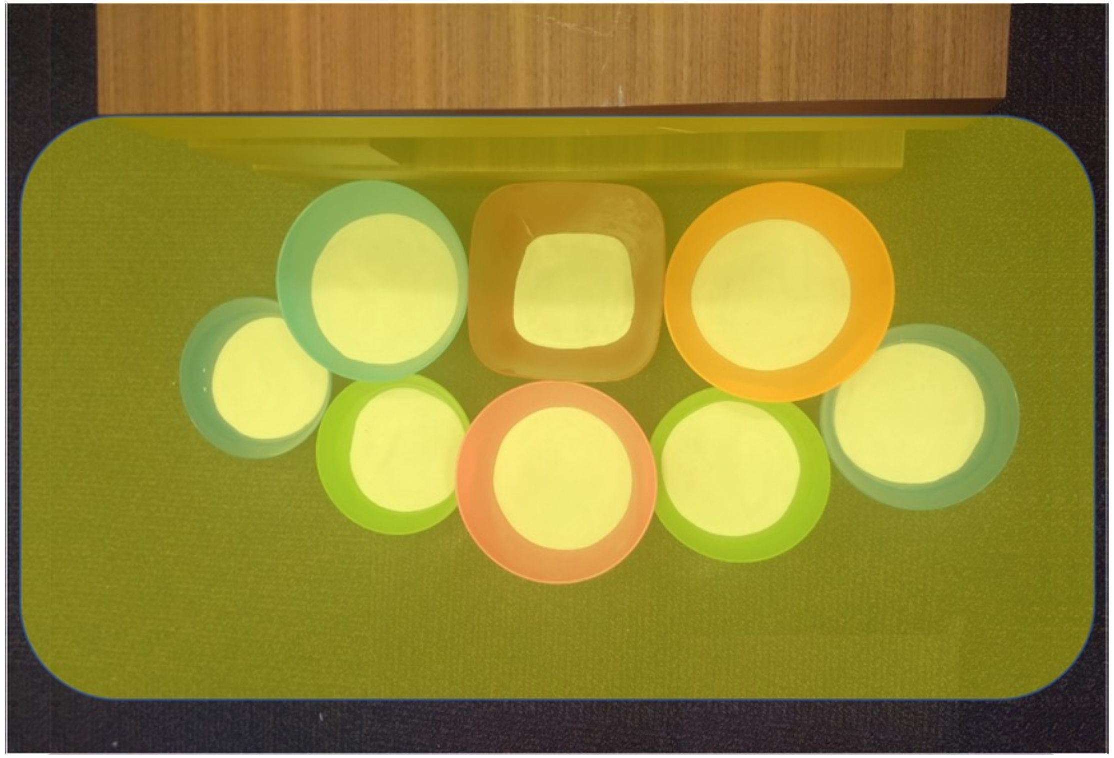


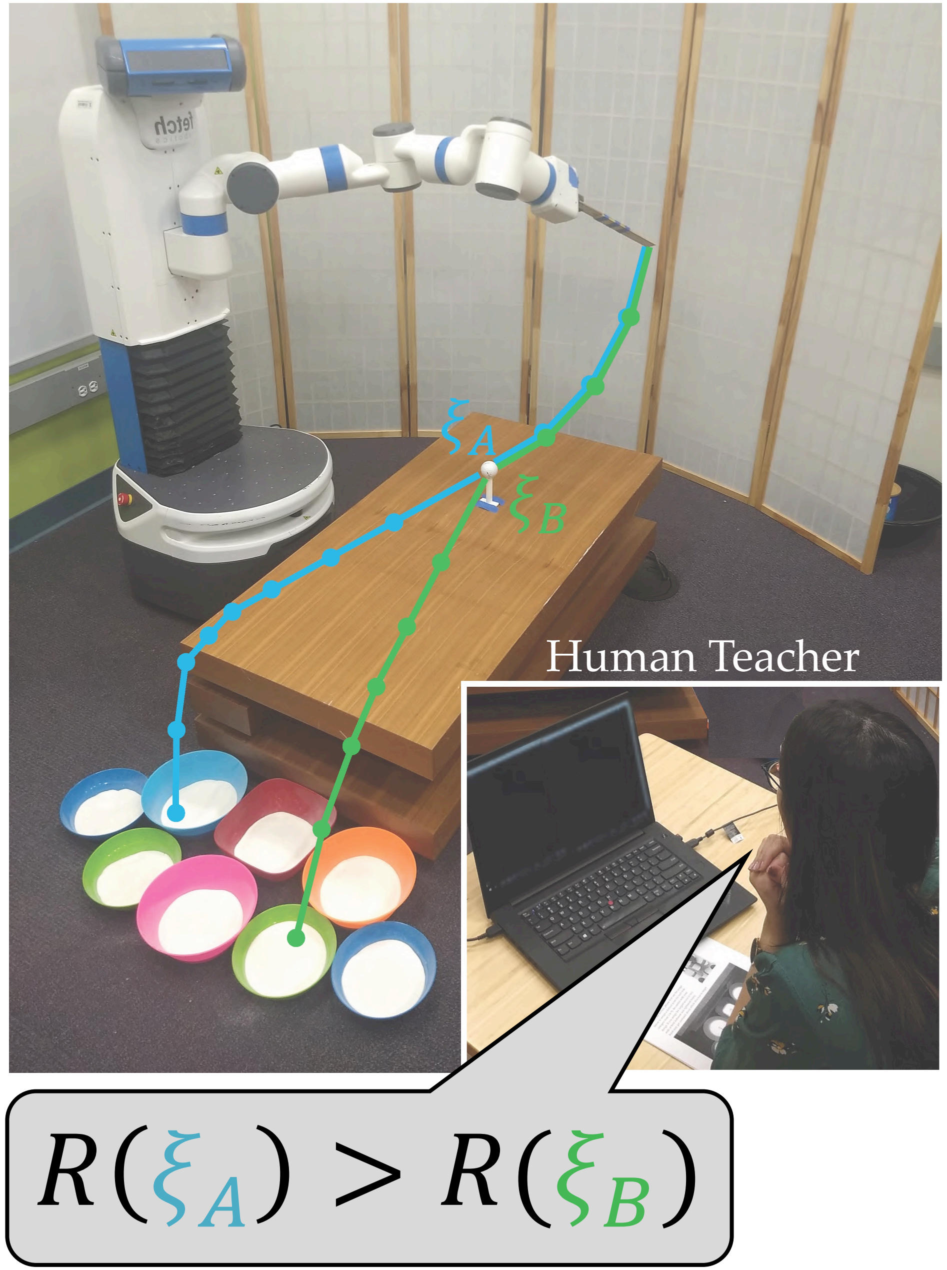
$$R(\xi_A) > R(\xi_B)$$

[Sadigh'17]
[Basu'18]
[Biyik'18,'19]
[Katz'19]
[Palan'19]
[Wilde'20]

Trajectory Features: Shot Speed, Shot Angle

$$R(\xi_A) = \theta(\phi(\xi_A))$$



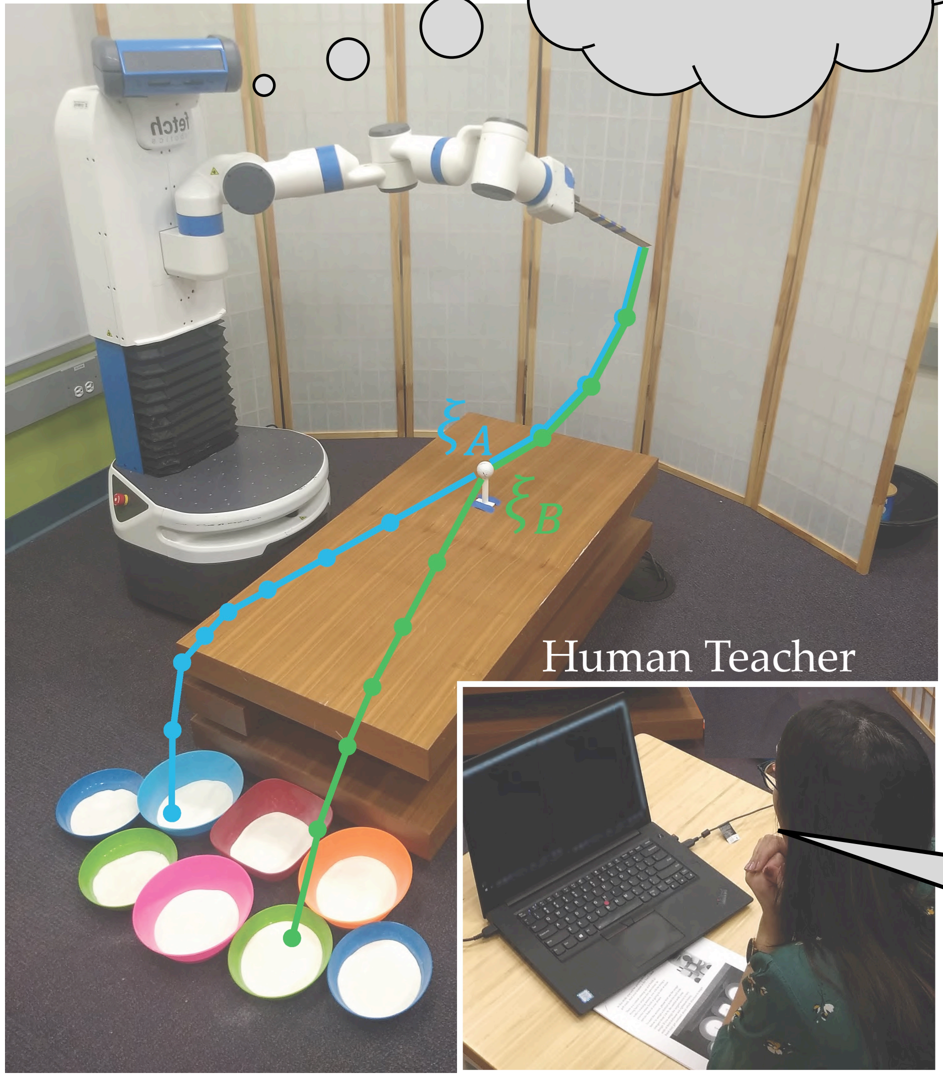


1. What if our reward function is nonlinear?

Approximate as a GP

Maximize Info Gain

$$(\xi_A, \xi_B)$$



$$P(\xi_A \mid \xi_A, \xi_B) = \Phi\left(\frac{R(\xi_A) - R(\xi_B)}{\sqrt{2}\sigma}\right)$$

$$P(f \mid \quad) \propto P(f)P(\quad \mid f)$$

I prefer ξ_A over ξ_B

Active Preference-based Gaussian Process Regression for Reward Learning
[*Biyik, Huynh, Kochenderfer, Sadigh. RSS'20*]

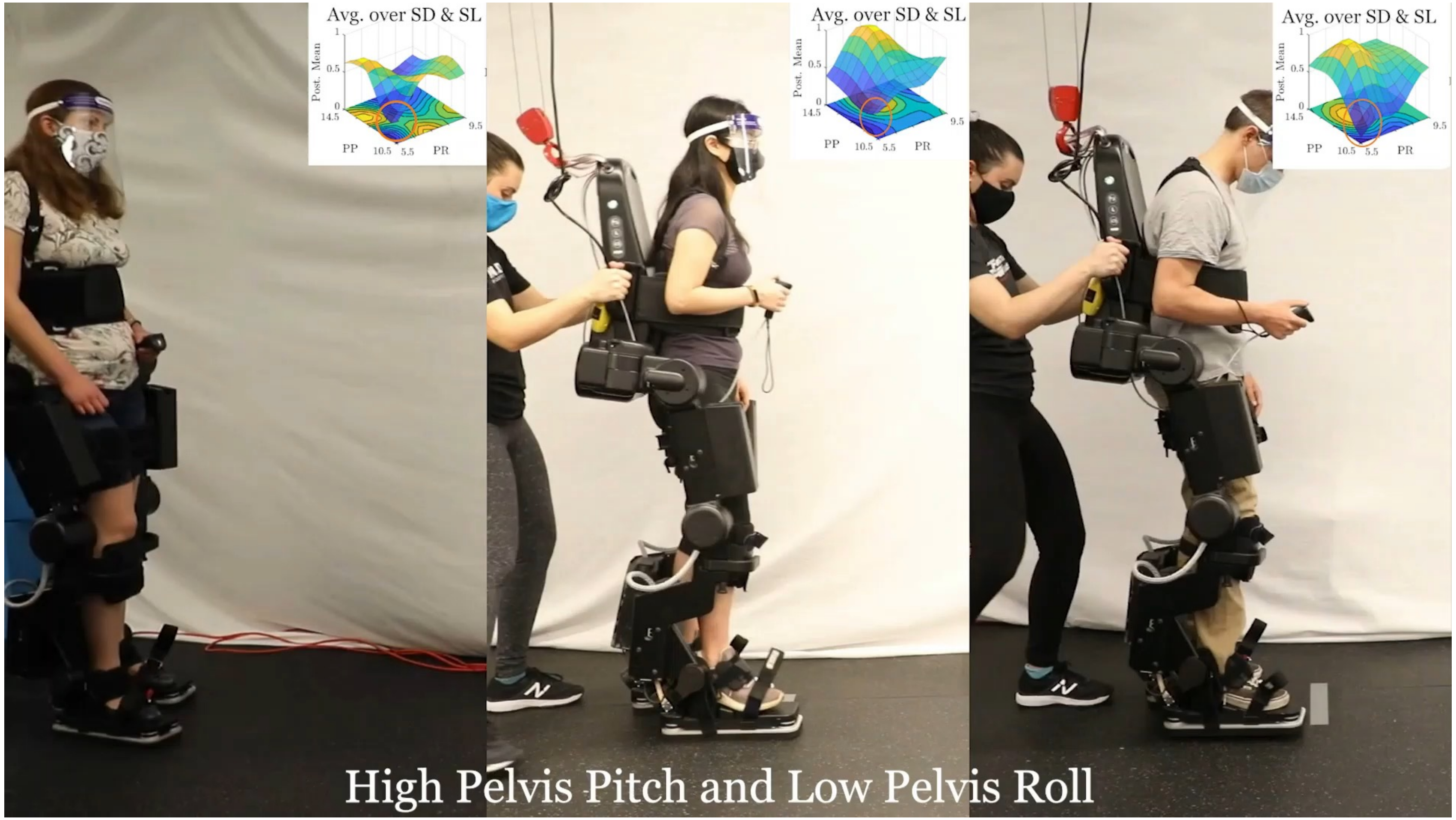
Learned Policies

Linear Reward

GP Reward

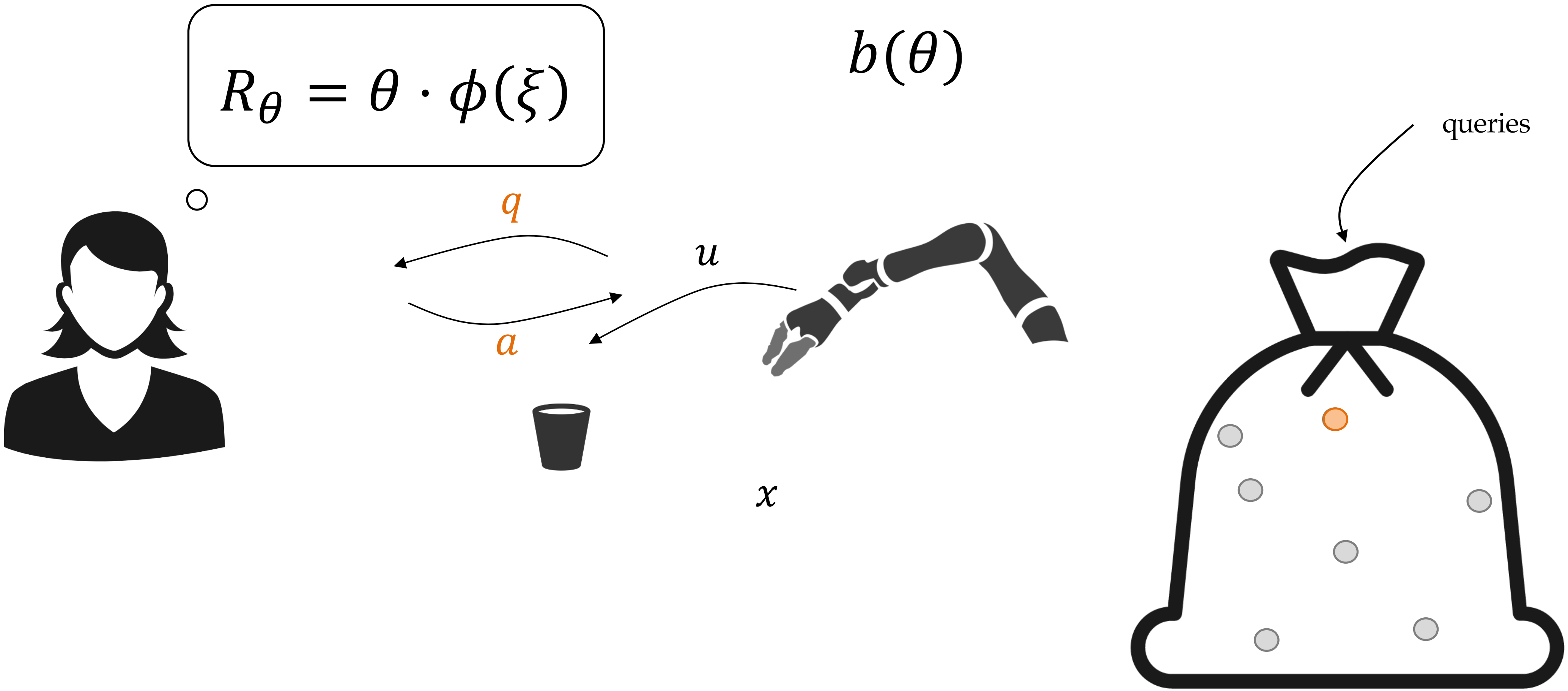
Active Preference-based Gaussian Process Regression for Reward Learning
[*Biyik, Huynh, Kochenderfer, Sadigh. RSS'20*]

Nonlinear Rewards for Exoskeletons

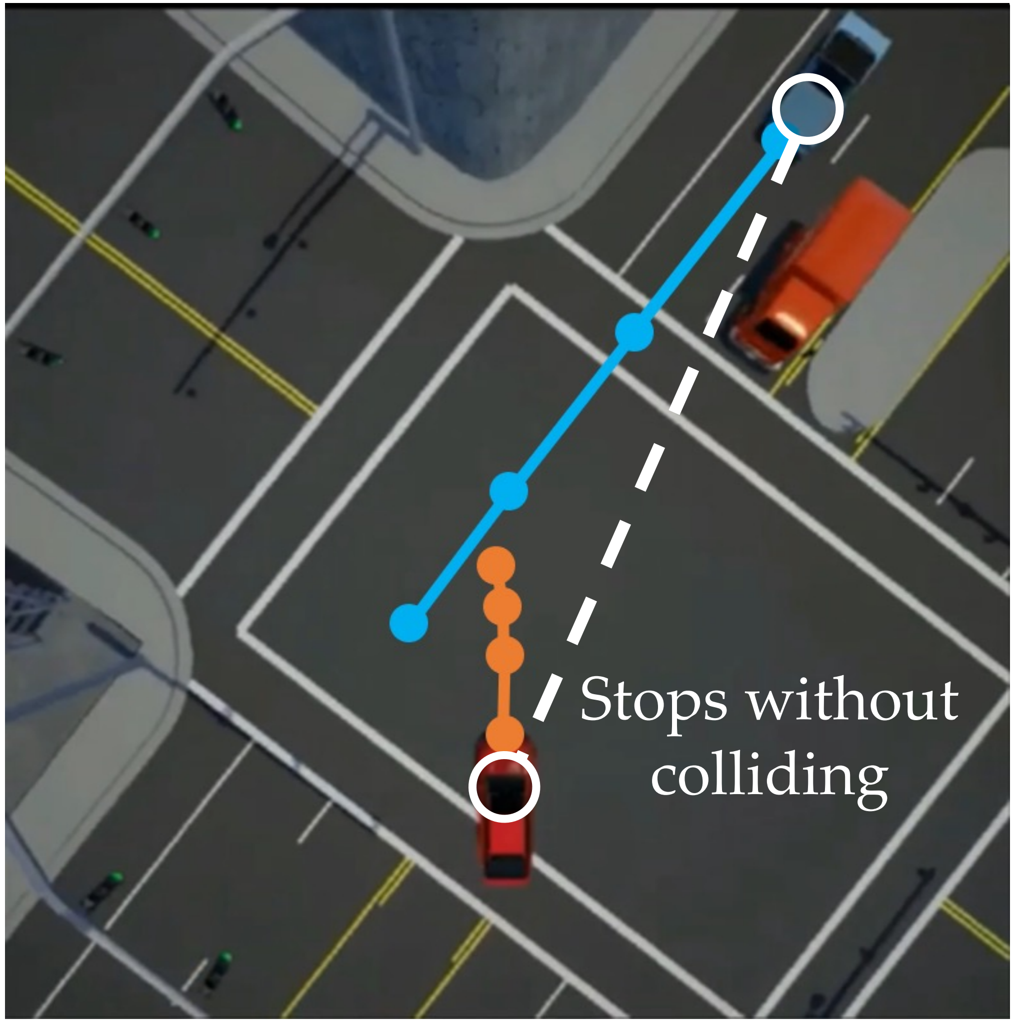


ROIAL: Region of Interest Active Learning for Characterizing Exoskeleton Gait Preference Landscapes
[*Li, Tucker, Biyik, Novoseller, Burdick, Sui, Sadigh, Yue, Ames, ICRA'21*]

Queries should be actively synthesized.



The red car will make an unprotected left turn…

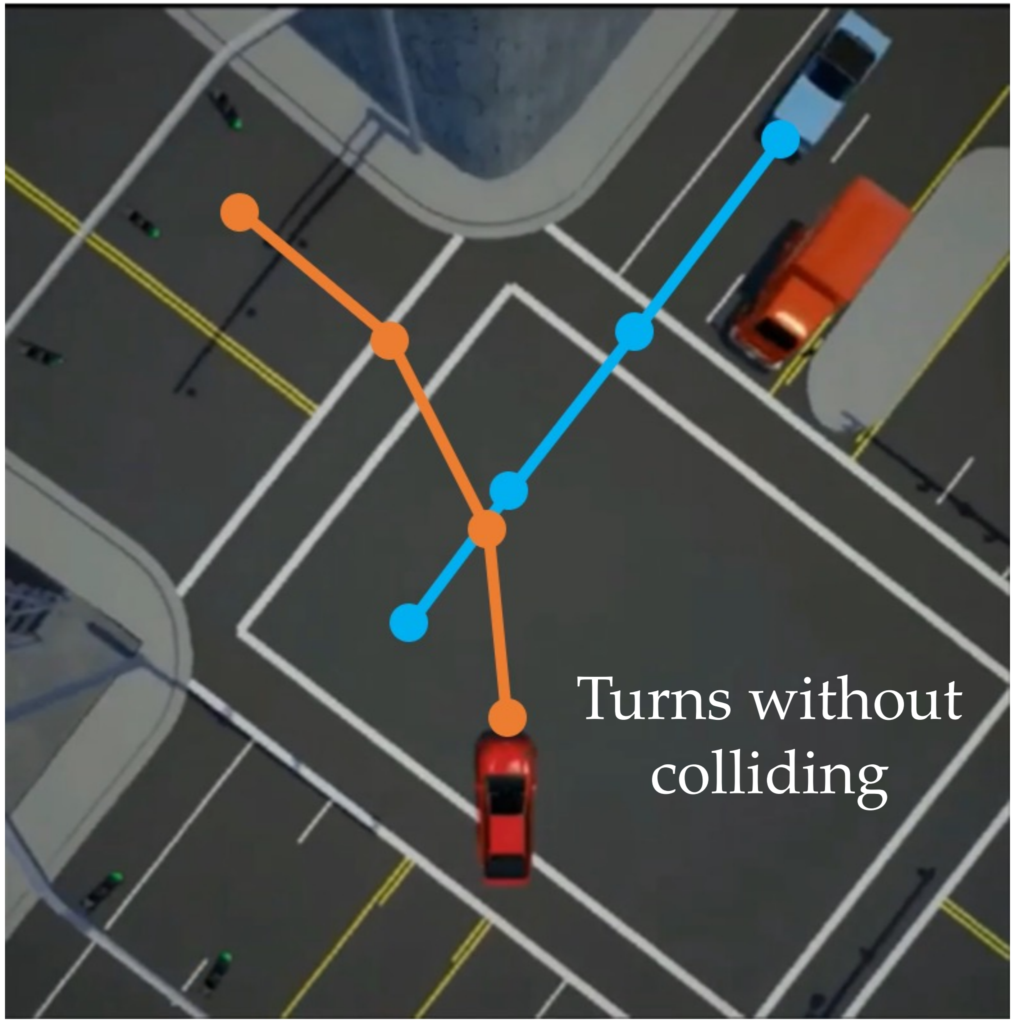


…but realizes the blue car is coming.

Learning from

A timid driver

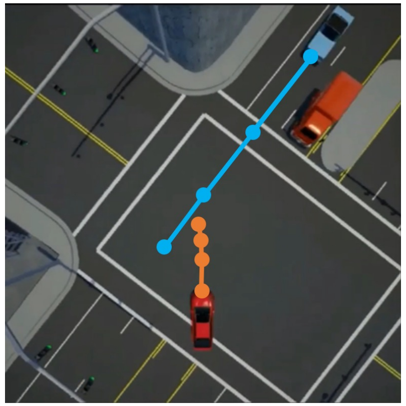
The red car will make an unprotected left turn…



…but realizes the blue car is coming.

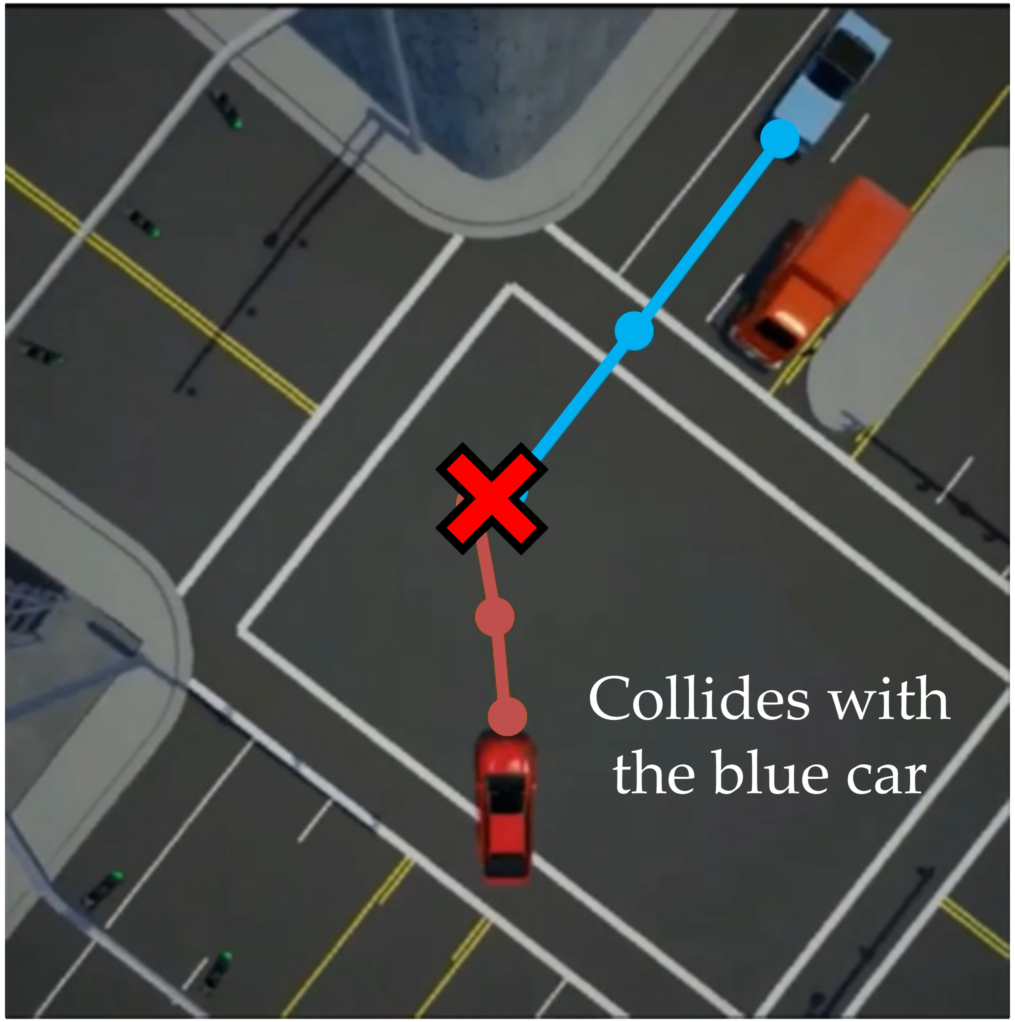
Learning from

A timid driver



An aggressive driver

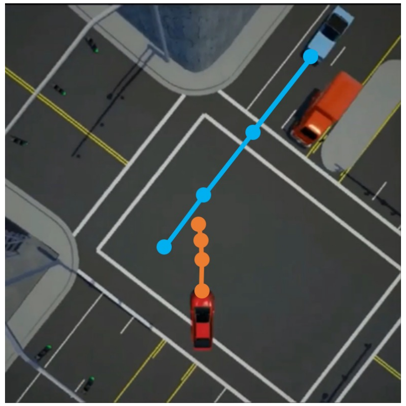
The red car will make an unprotected left turn…



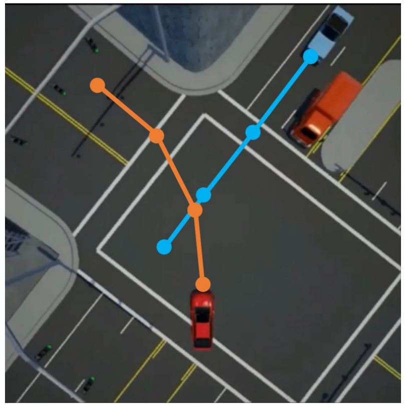
…but realizes the blue car is coming.

Learning from

A timid driver

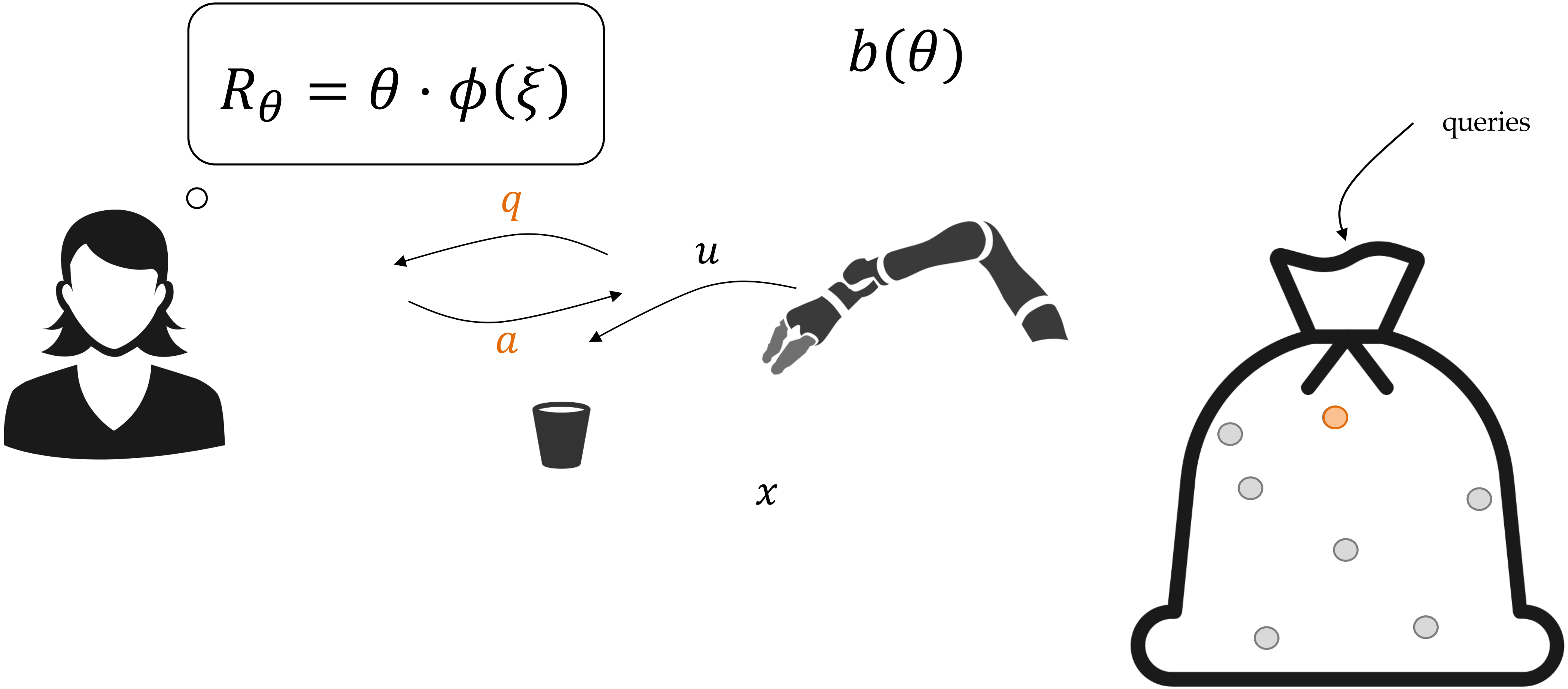


An aggressive driver

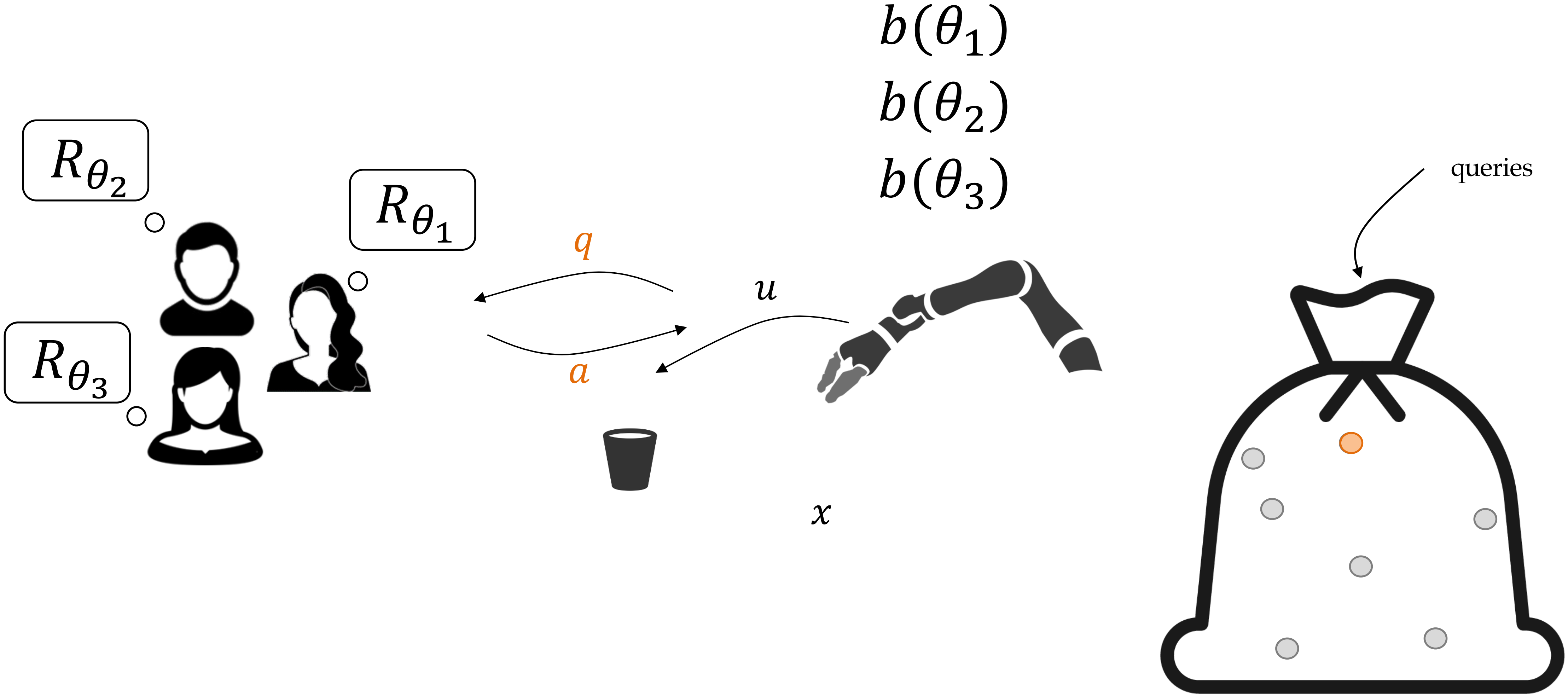


Both types of drivers

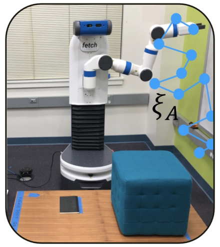
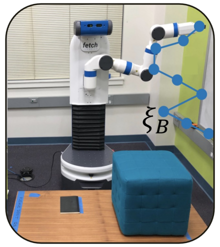
Queries should be actively synthesized.

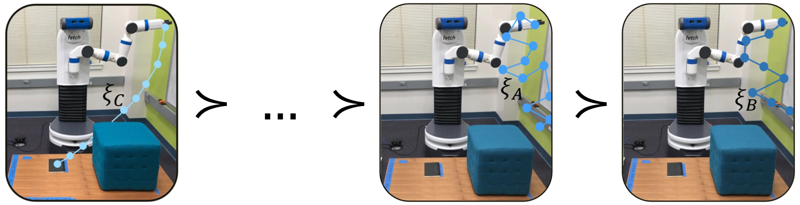


Queries should be actively synthesized.

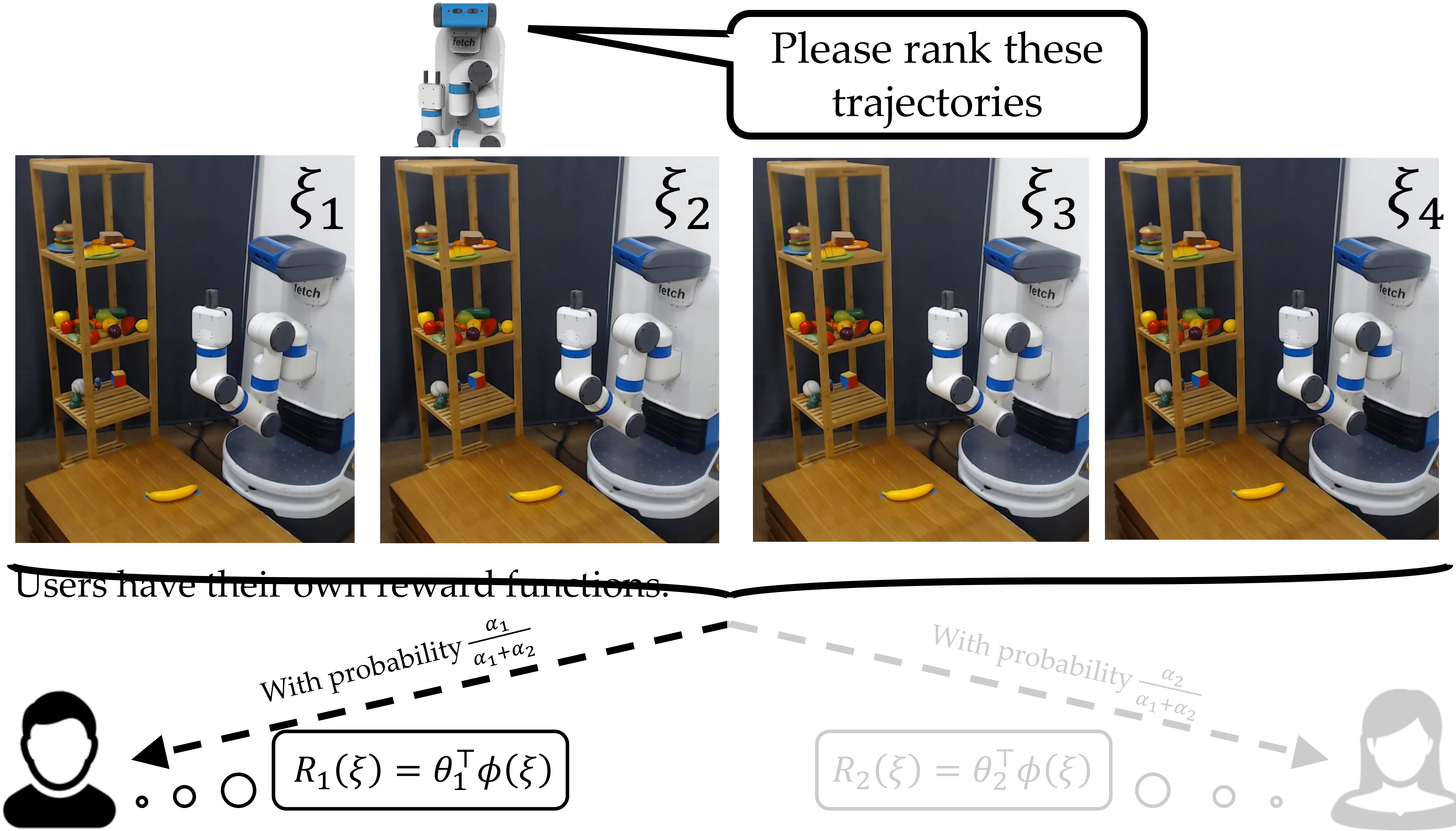


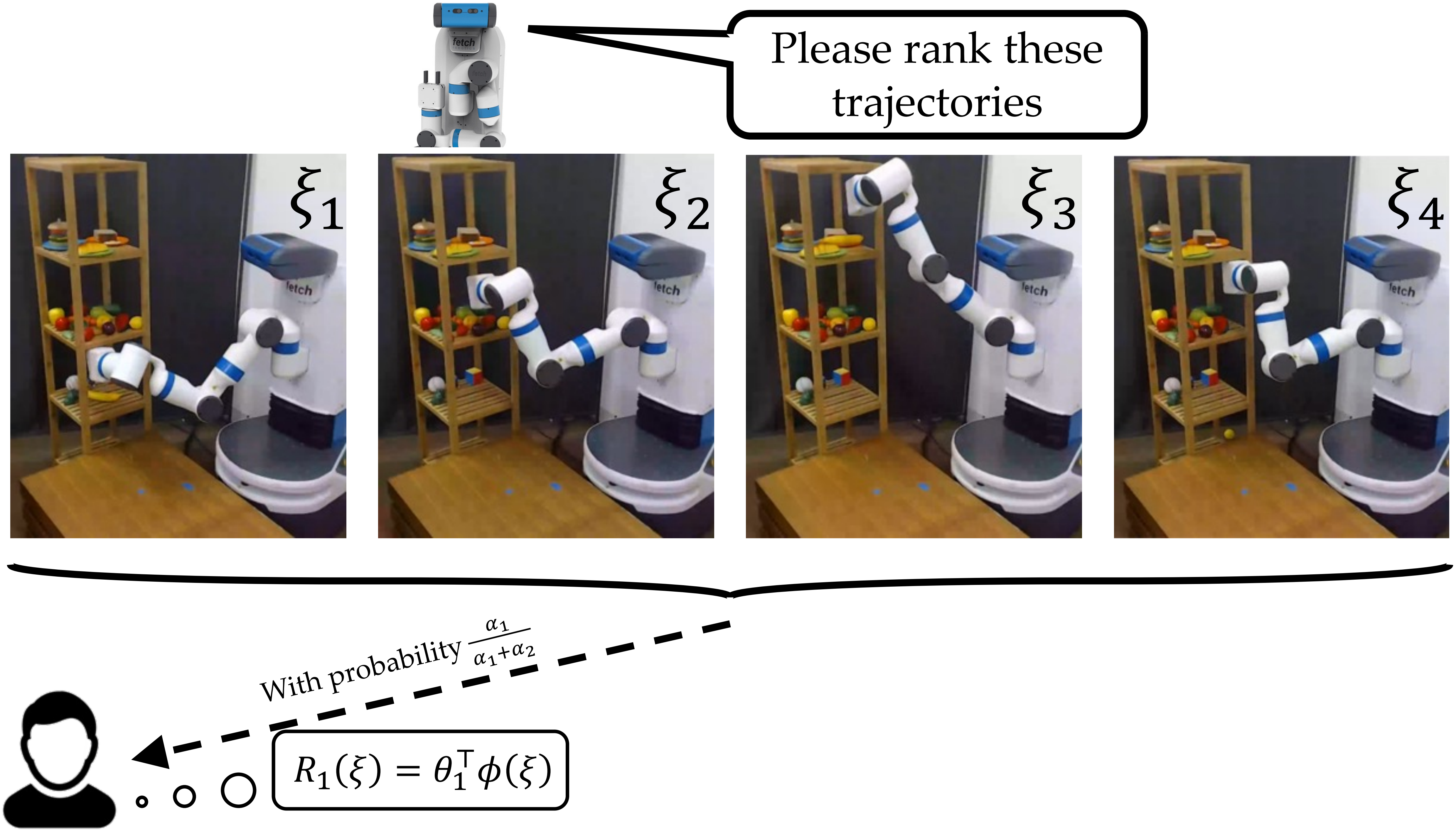
1. What if our reward function is nonlinear?
2. What if our reward function is multimodal?

	Unimodal	Multimodal
Learning from Comparisons  $\succ$ 	Gervasio et al., 1999 Akrour et al., 2012 Sadigh et al., 2017 Christiano et al., 2017 Bıyık et al., 2020 Tucker et al., 2020 Wilde et al., 2021	Impossible ? (Zhao et al., 2016)

	Unimodal	Multimodal
Learning from Rankings 	Gervasio et al., 1999 Akrour et al., 2012 Sadigh et al., 2017 Christiano et al., 2017 Bıyık et al., 2020 Tucker et al., 2020 Wilde et al., 2021	This work

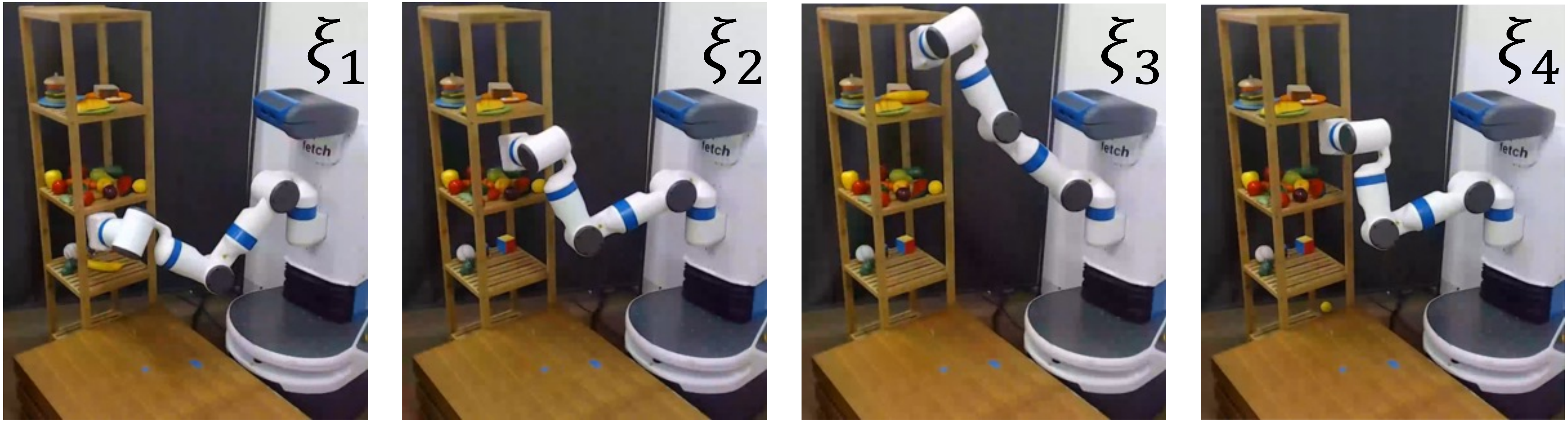
Learning Multimodal Rewards from Rankings
[Myers, Bıyık, Anari, Sadigh, CoRL'21]


Please rank these trajectories
ξ_1
ξ_2
ξ_3
ξ_4
Users have their own reward functions.
With probability $\frac{\alpha_1}{\alpha_1+\alpha_2}$
With probability $\frac{\alpha_2}{\alpha_1+\alpha_2}$
$R_1(\xi) = \theta_1^\top \phi(\xi)$
$R_2(\xi) = \theta_2^\top \phi(\xi)$



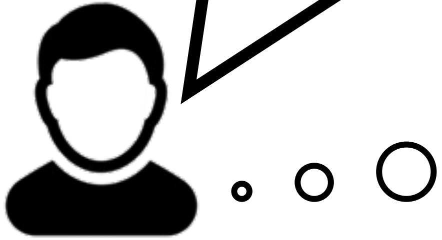
Please rank these trajectories
ξ_1
ξ_2
ξ_3
ξ_4
With probability $\frac{\alpha_1}{\alpha_1+\alpha_2}$
$R_1(\xi) = \theta_1^\top \phi(\xi)$

Please rank these trajectories

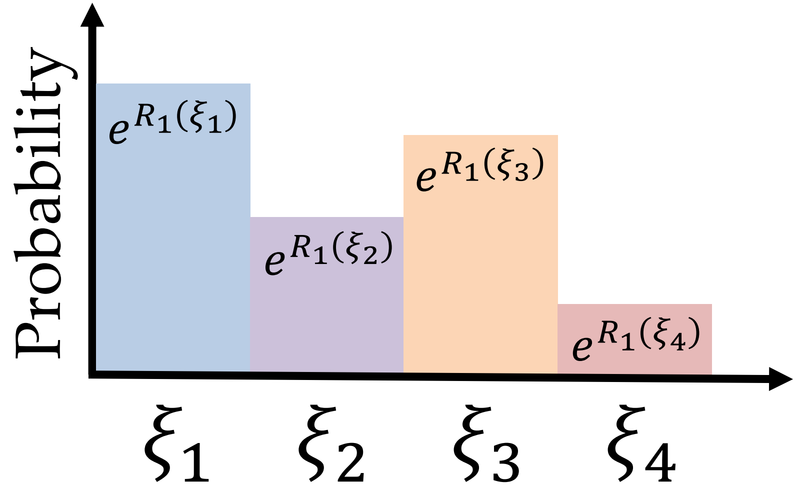


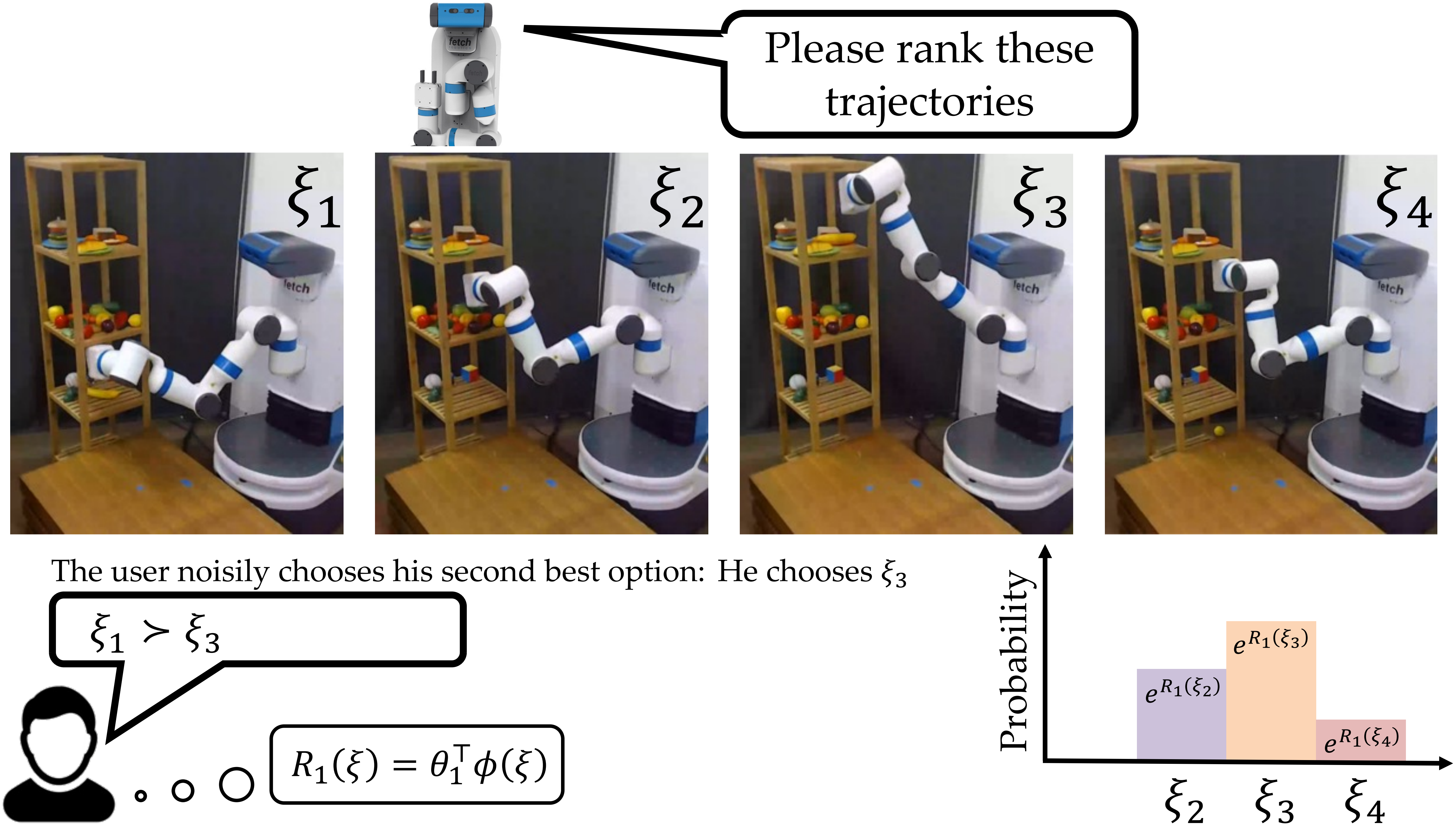
The user noisily chooses his best option: He chooses ξ_1

ξ_1

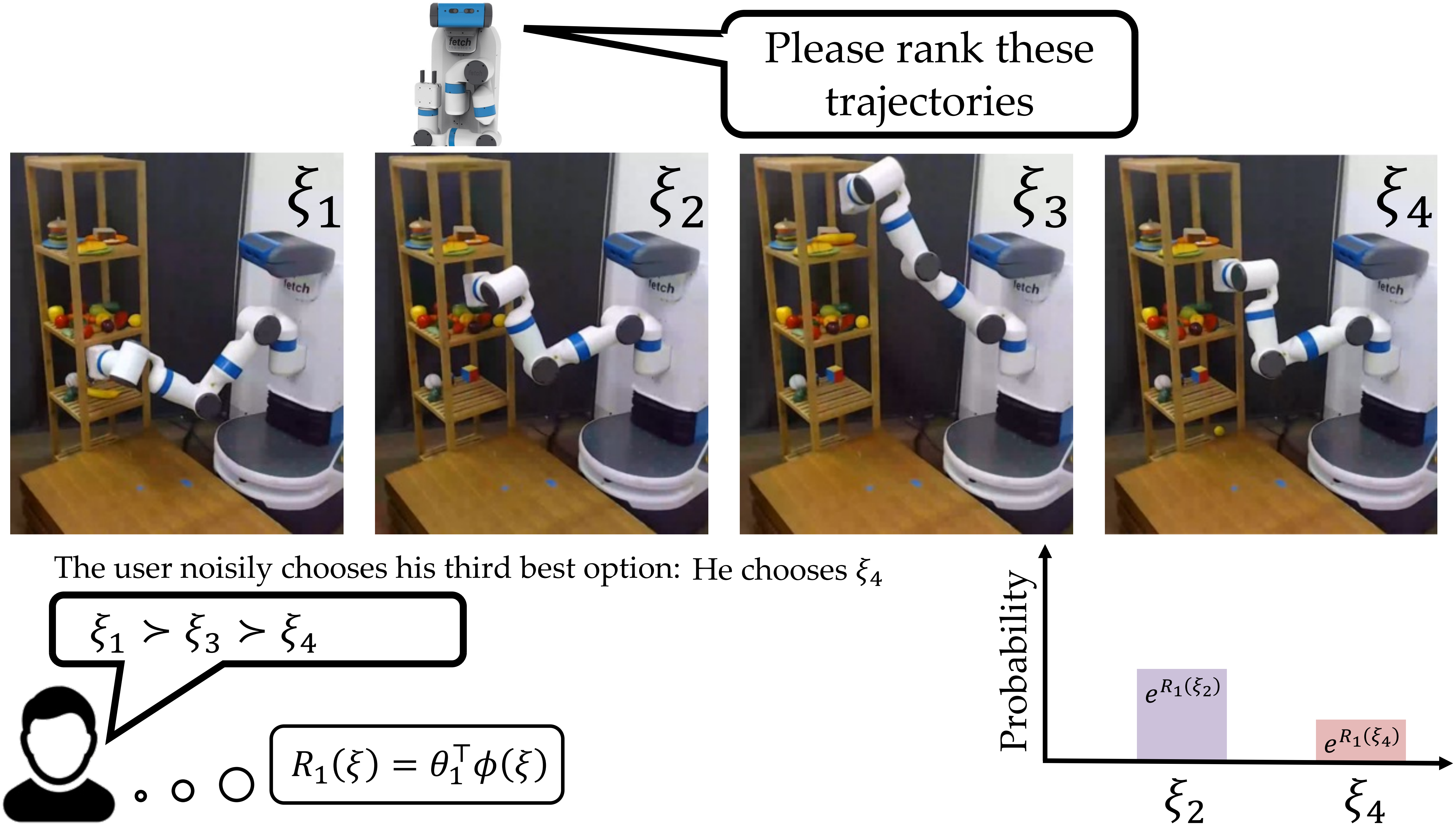


$R_1(\xi) = \theta_1^\top \phi(\xi)$

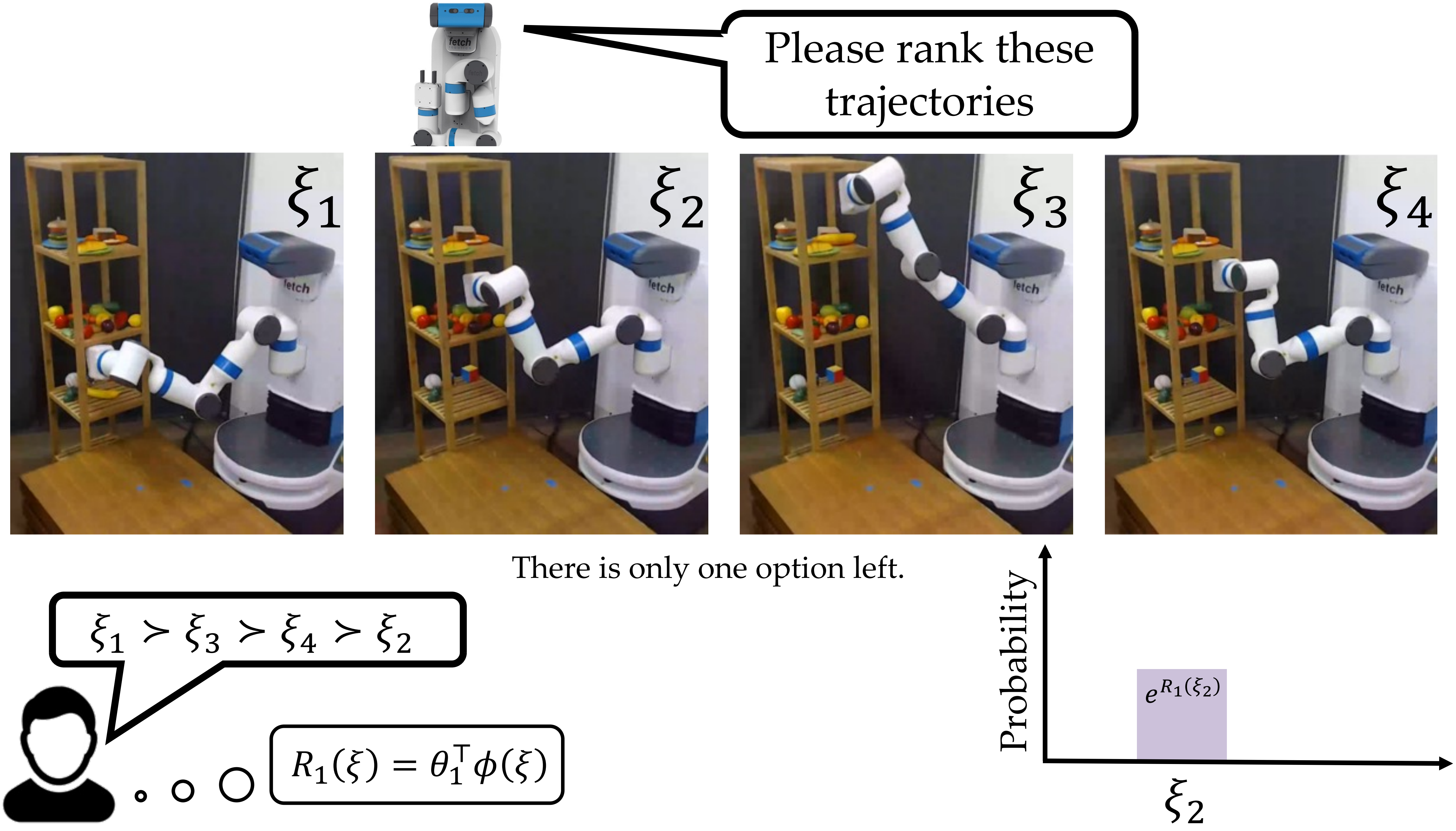




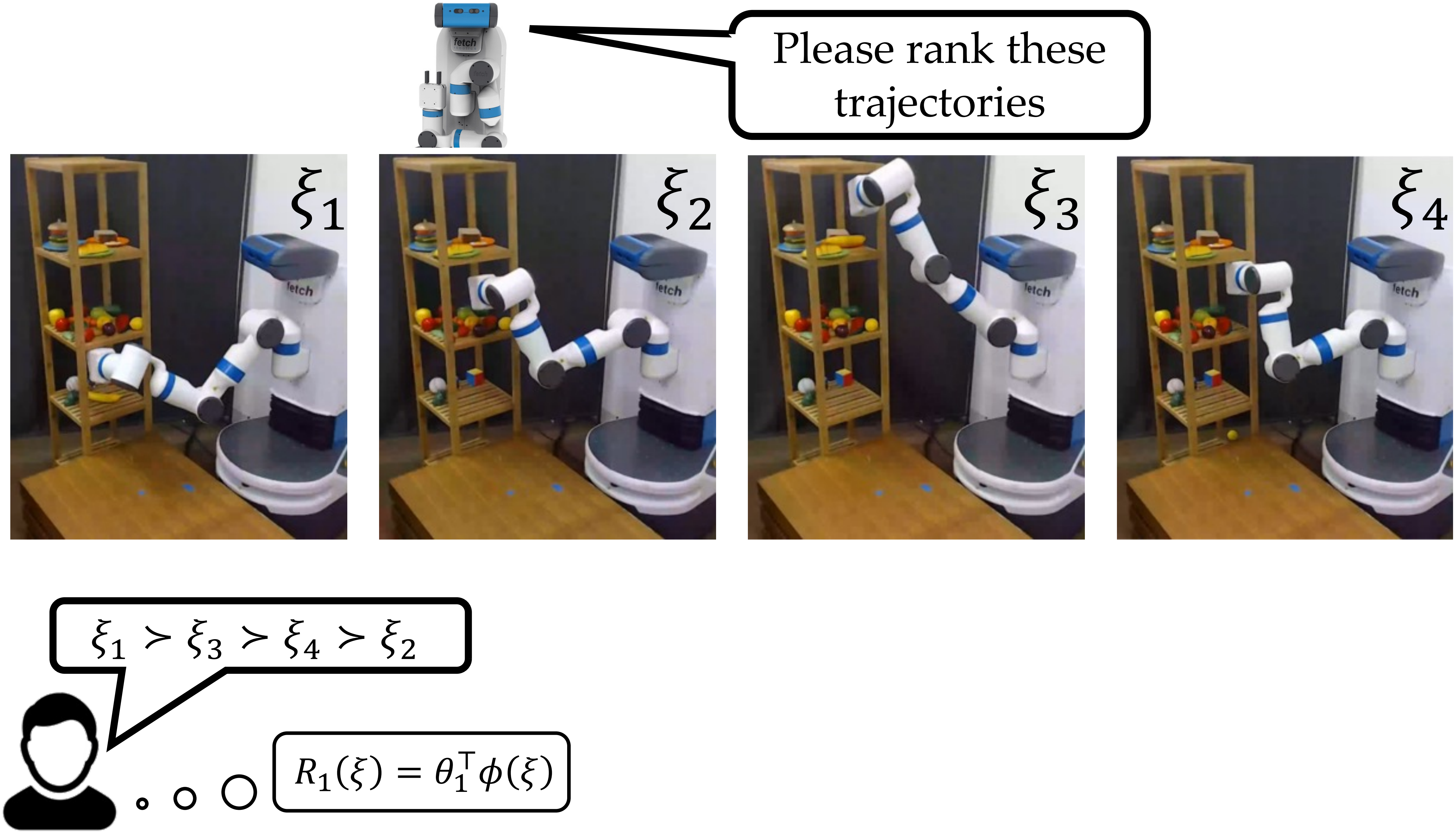
Please rank these trajectories
ξ_1
ξ_2
ξ_3
ξ_4
The user noisily chooses his second best option: He chooses ξ_3
$\xi_1 \succ \xi_3$
$R_1(\xi) = \theta_1^\top \phi(\xi)$
Probability
$e^{R_1(\xi_2)}$
$e^{R_1(\xi_3)}$
$e^{R_1(\xi_4)}$
ξ_2 ξ_3 ξ_4



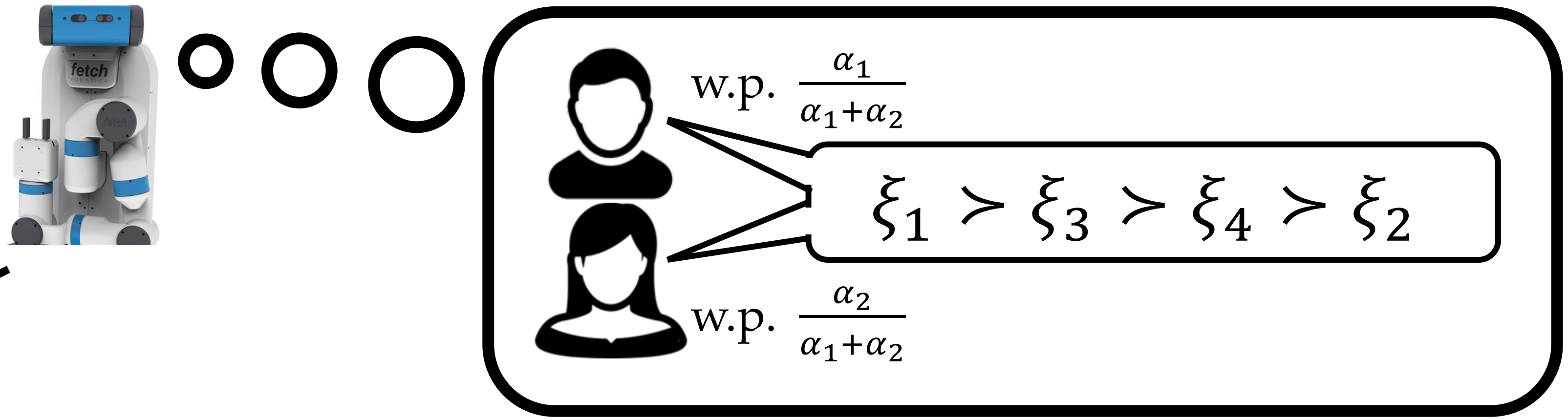
Please rank these trajectories
ξ_1
ξ_2
ξ_3
ξ_4
The user noisily chooses his third best option: He chooses ξ_4
$\xi_1 \succ \xi_3 \succ \xi_4$
$R_1(\xi) = \theta_1^\top \phi(\xi)$
Probability
$e^{R_1(\xi_2)}$
$e^{R_1(\xi_4)}$
ξ_2
ξ_4



Please rank these trajectories
ξ_1
ξ_2
ξ_3
ξ_4
There is only one option left.
$\xi_1 \succ \xi_3 \succ \xi_4 \succ \xi_2$
$R_1(\xi) = \theta_1^\top \phi(\xi)$
Probability
$e^{R_1(\xi_2)}$
ξ_2



Please rank these trajectories
ξ_1
ξ_2
ξ_3
ξ_4
$\xi_1 \succ \xi_3 \succ \xi_4 \succ \xi_2$
$R_1(\xi) = \theta_1^\top \phi(\xi)$

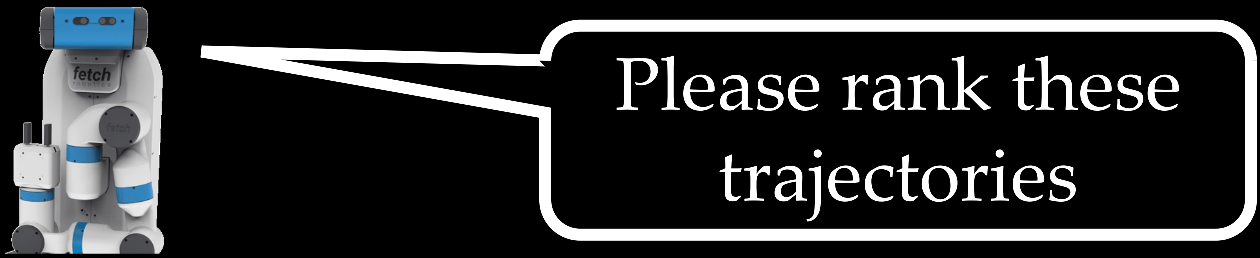


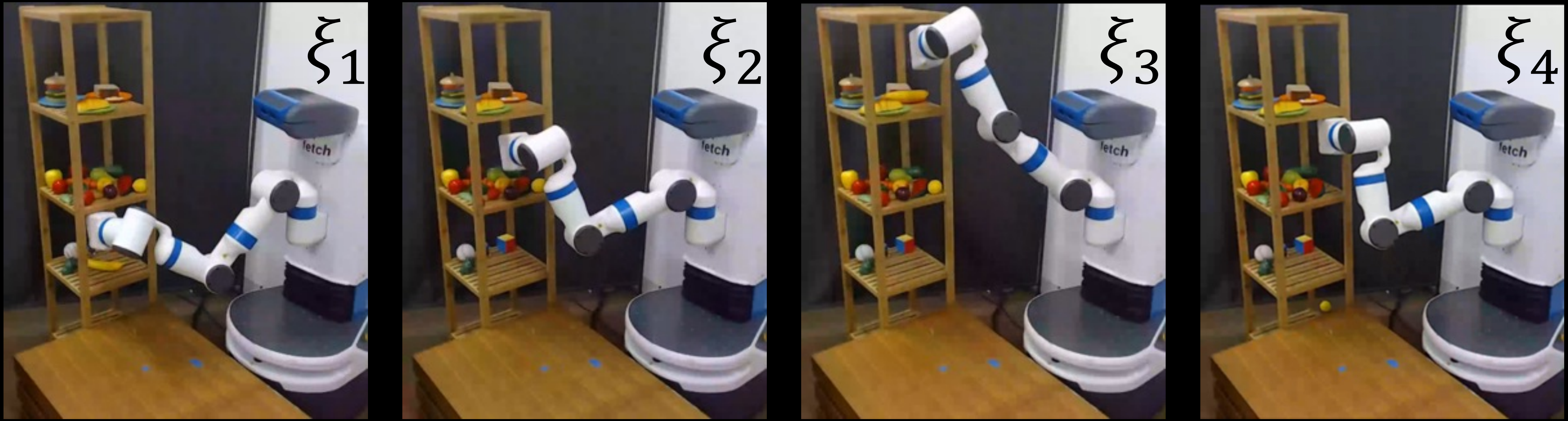
Starts with a prior $P(\alpha, \theta_1, \theta_2)$

Posterior: $P(\alpha, \theta_1, \theta_2 \mid \xi_1 \succ \xi_3 \succ \xi_4 \succ \xi_2) \propto P(\alpha, \theta_1, \theta_2) \underbrace{P(\xi_1 \succ \xi_3 \succ \xi_4 \succ \xi_2 \mid \alpha, \theta_1, \theta_2)}_{\textit{Probability of observing the ranking}}$

$$P(\xi_1 \succ \xi_3 \succ \xi_4 \succ \xi_2 \mid \alpha, \theta_1, \theta_2) =$$
$$P(\;\mid \alpha) P(\xi_1 \succ \xi_3 \succ \xi_4 \succ \xi_2 \mid \;, \theta_1) + P(\;\mid \alpha) P(\xi_1 \succ \xi_3 \succ \xi_4 \succ \xi_2 \mid \;, \theta_2)$$

Learning Multimodal Rewards from Rankings
[Myers, Bıyık, Anari, Sadigh, CoRL'21]





How does the robot choose which trajectories to show to the users?

We actively query the users by maximizing information gain.

Maximize Information Gain

Unimodal: $\max_{\text{query}} I(\theta\,;\ \text{response} \mid \text{query})$

Multimodal: $\max_{\text{query}} I\big(\alpha, (\theta_i)_{i=1}^{M}\,;\ \text{response} \mid \text{query}\big)$

$$\min_{\text{query}} \mathbb{E}_{\text{response},\alpha,(\theta_i)_{i=1}^{M}\mid\text{query}} \log \frac{\mathbb{E}_{\alpha',(\theta'_i)_{i=1}^{M}} P\big(\text{response} \mid \text{query}, \alpha', (\theta'_i)_{i=1}^{M}\big)}{P\big(\text{response} \mid \text{query}, \alpha, (\theta_i)_{i=1}^{M}\big)}$$

Bıyık et al. CoRL'19

Maximize Information Gain

Unimodal: $\max_{\text{query}} I(\theta\, ;\, \text{response} \mid \text{query})$

Multimodal: $\max_{\text{query}} I\Big(\alpha, (\theta_i)_{i=1}^{M}\, ;\, \text{response} \mid \text{query}\Big)$

$$\min_{\text{query}} \mathbb{E}_{\text{response},\alpha,(\theta_i)_{i=1}^{M} \mid \text{query}} \log \frac{\mathbb{E}_{\alpha',(\theta_i')_{i=1}^{M}}\, P\big(\text{response} \mid \text{query}, \alpha', (\theta_i')_{i=1}^{M}\big)}{P\big(\text{response} \mid \text{query}, \alpha, (\theta_i)_{i=1}^{M}\big)}$$

Maximize Information Gain

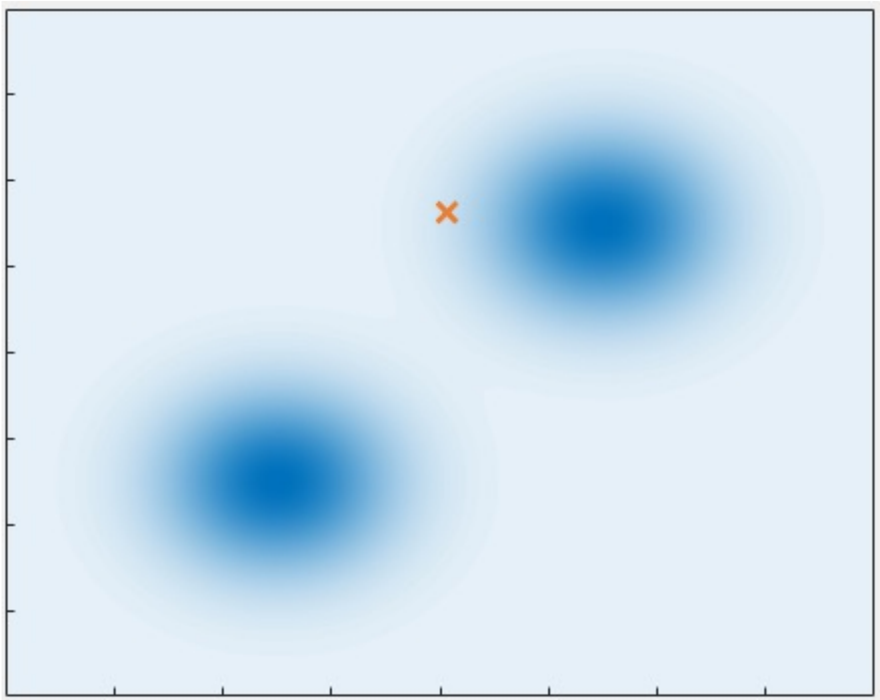
Unimodal: $\max_{\text{query}} I(\theta\,;\ \text{response} \mid \text{query})$

Multimodal: $\max_{\text{query}} I\left(\alpha, (\theta_i)_{i=1}^{M}\,;\ \text{response} \mid \text{query}\right)$

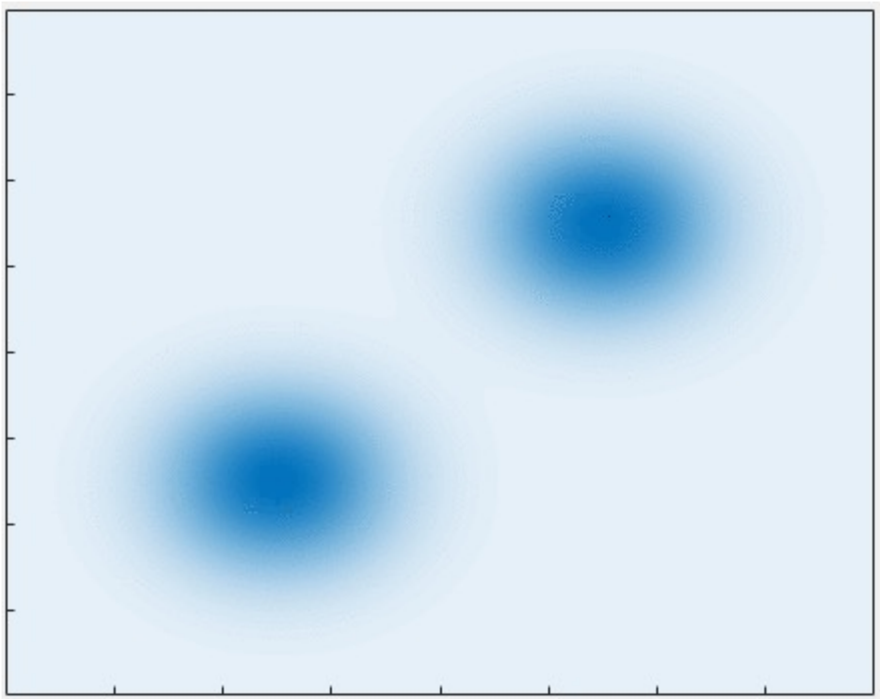
$$\min_{\text{query}} \mathbb{E}_{\text{response},\alpha,(\theta_i)_{i=1}^{M}\mid\text{query}} \log \frac{\mathbb{E}_{\alpha',(\theta_i')_{i=1}^{M}} P\left(\text{response} \mid \text{query}, \alpha', (\theta_i')_{i=1}^{M}\right)}{P\left(\text{response} \mid \text{query}, \alpha, (\theta_i)_{i=1}^{M}\right)}$$

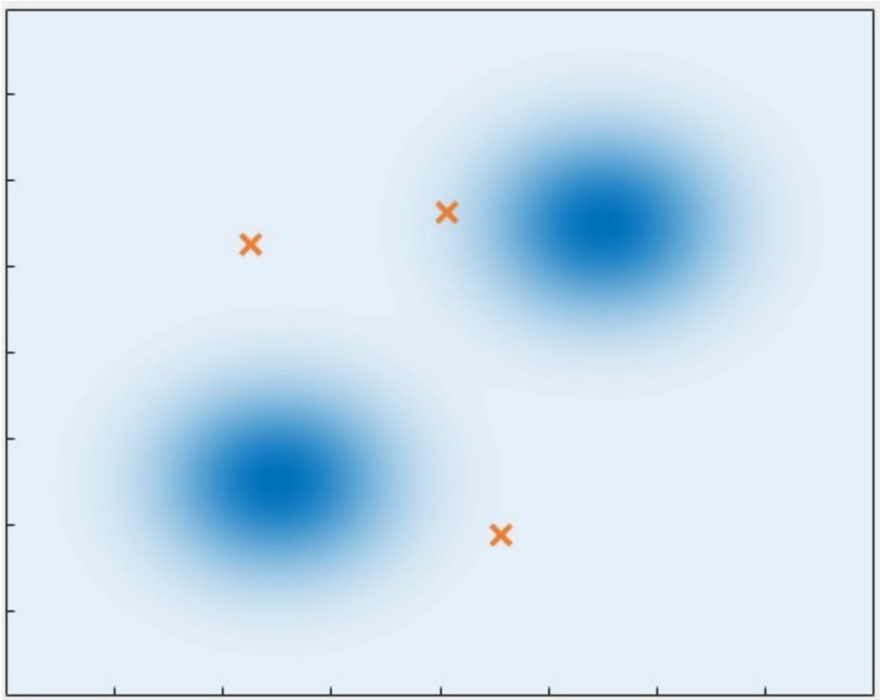
Sample from: $P\left(\alpha, (\theta_i)_{i=1}^{M}\right)$ Multimodal!

Sampling from a Multimodal Distribution

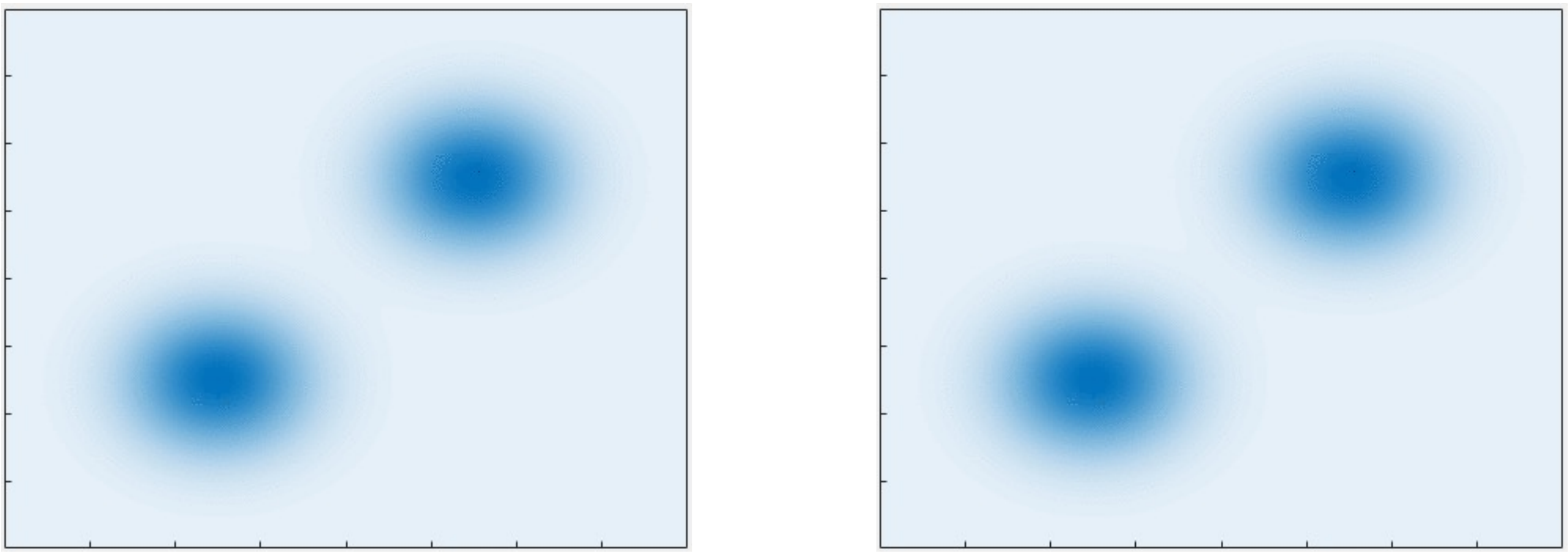


Sampling from a Multimodal Distribution





Sampling from a Multimodal Distribution



Maximize Information Gain

Unimodal: $\max_{\text{query}} I(\theta\,;\,\text{response} \mid \text{query})$

Multimodal: $\max_{\text{query}} I\left(\alpha, (\theta_i)_{i=1}^{M}\,;\,\text{response} \mid \text{query}\right)$

$$\min_{\text{query}} \mathbb{E}_{\text{response},\alpha,(\theta_i)_{i=1}^{M}\mid\text{query}} \log \frac{\mathbb{E}_{\alpha',(\theta_i')_{i=1}^{M}} P\left(\text{response} \mid \text{query}, \alpha', (\theta_i')_{i=1}^{M}\right)}{P\left(\text{response} \mid \text{query}, \alpha, (\theta_i)_{i=1}^{M}\right)}$$

Maximize Information Gain

Unimodal: $$\max_{\text{query}} I(\theta \,;\, \text{response} \mid \text{query})$$

Multimodal: $$\max_{\text{query}} I\left(\alpha, (\theta_i)_{i=1}^{M} \,;\, \text{response} \mid \text{query}\right)$$

$$\min_{\text{query}} \mathbb{E}_{\text{response},\alpha,(\theta_i)_{i=1}^{M}\mid\text{query}} \log \frac{\mathbb{E}_{\alpha',(\theta_i')_{i=1}^{M}} P\left(\text{response} \mid \text{query}, \alpha', (\theta_i')_{i=1}^{M}\right)}{P\left(\text{response} \mid \text{query}, \alpha, (\theta_i)_{i=1}^{M}\right)}$$

Active Method Better Learns

Middle Shelf

After 10 random queries:

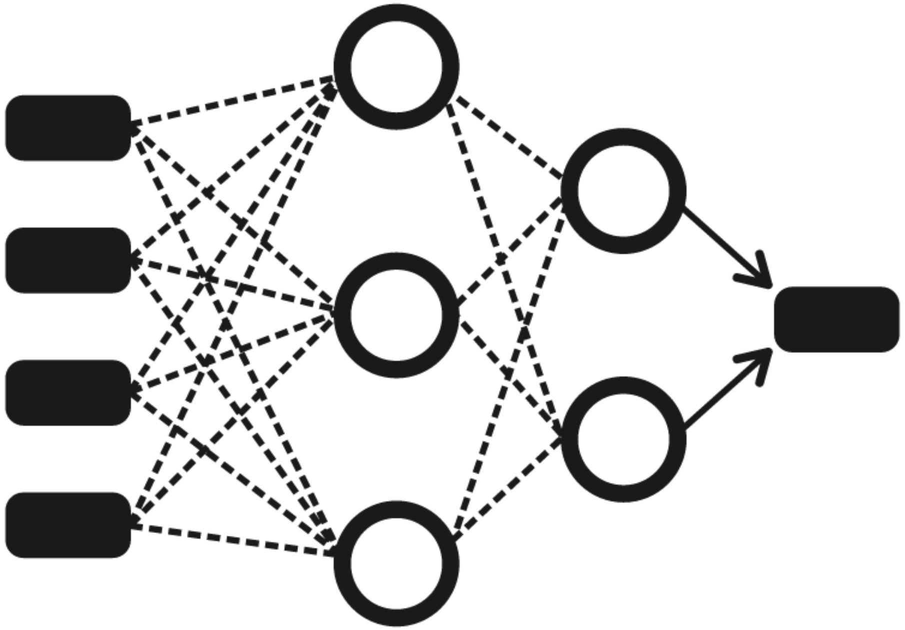
Random querying did not properly learn not dropping items yet

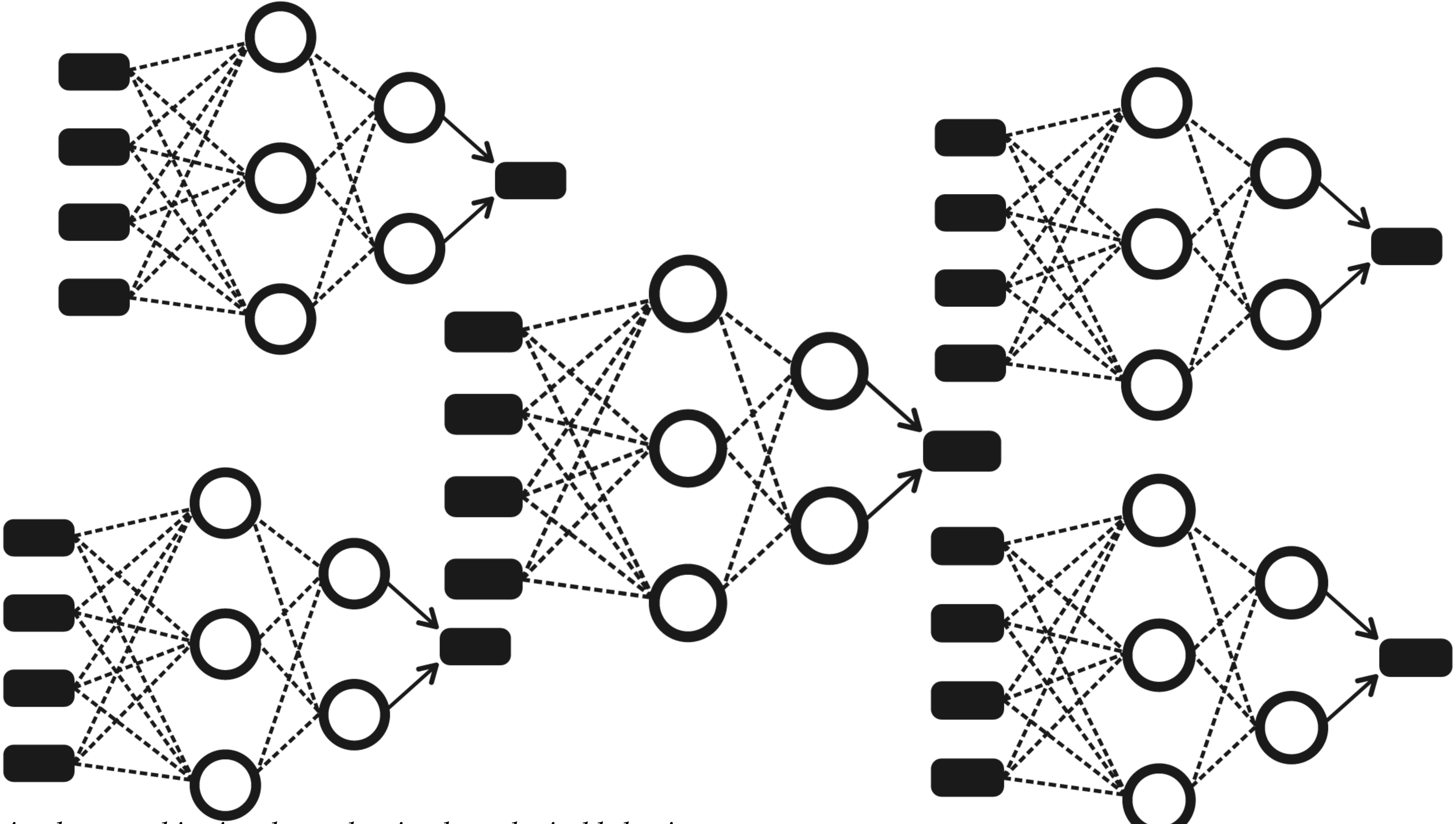
After 10 information-optimal queries:

Learning Multimodal Rewards from Rankings
[Myers, Bıyık, Anari, Sadigh, CoRL'21]

1. What if our reward function is nonlinear?
2. What if our reward function is multimodal?

1. What if our reward function is nonlinear?
2. What if our reward function is multimodal?
3. What if we have a neural reward function?





Learning human objectives by evaluating hypothetical behavior
[Reddy, Dragan, Levine, Legg, Leike, ICML'20]

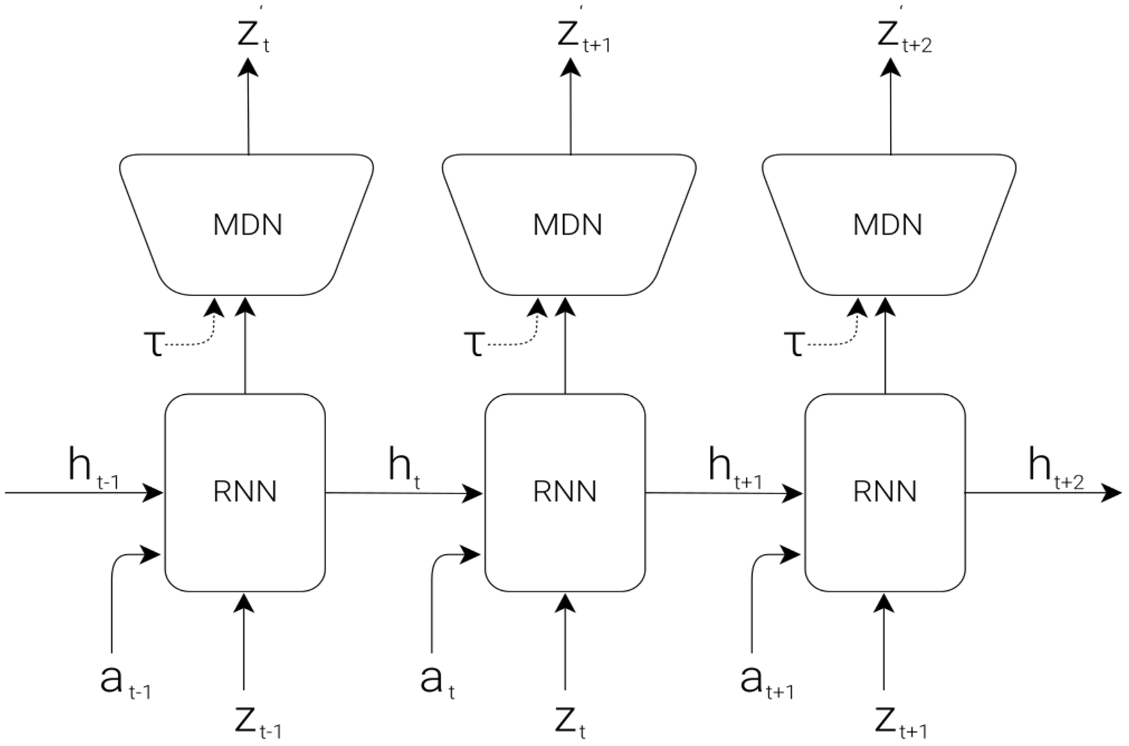
Actively synthesizing queries

$$\max_{\xi} \quad \text{EnsableDisagreement}(\mathrm{R}_\theta(\xi))$$

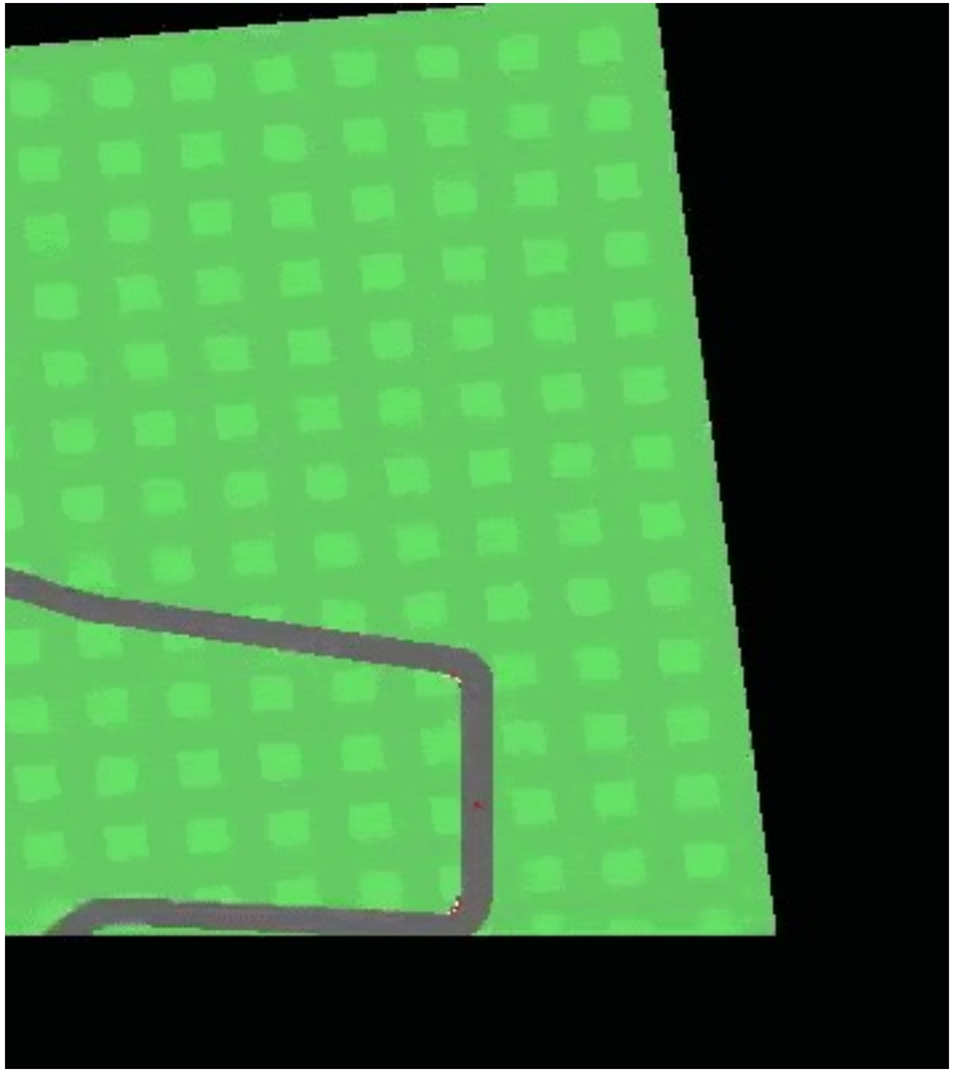
$$\text{Subject to} \quad p_\phi(\xi) > \tau$$

learned dynamics model

Learning human objectives by evaluating hypothetical behavior
[Reddy, Dragan, Levine, Legg, Leike, ICML'20]

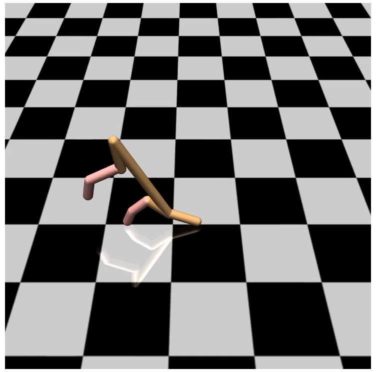
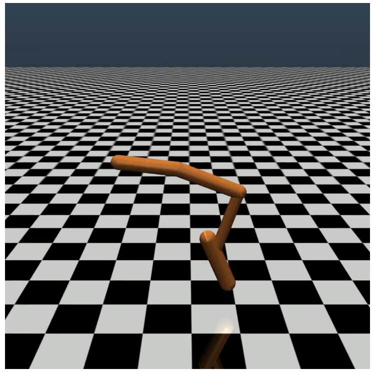


VAE + RNN + Mixture Density Network

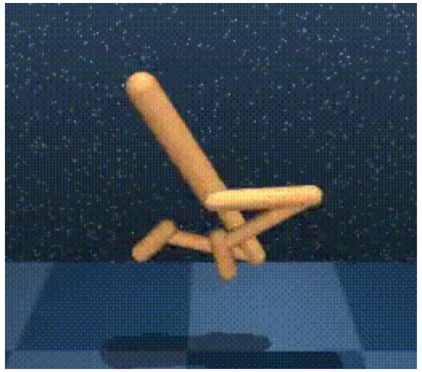


Learning human objectives by evaluating hypothetical behavior
[Reddy, Dragan, Levine, Legg, Leike, ICML'20]

Active Learning of Neural Rewards

Christiano et al. NeurIPS‘17

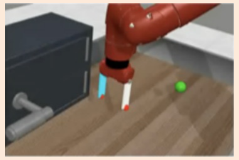


Lee et al. ICML‘21

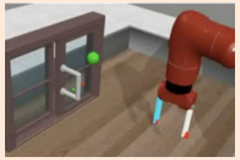
- Generating trajectories takes too much time.
- Human data are expensive.

Pre-training

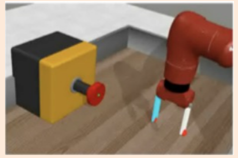
Prior Tasks



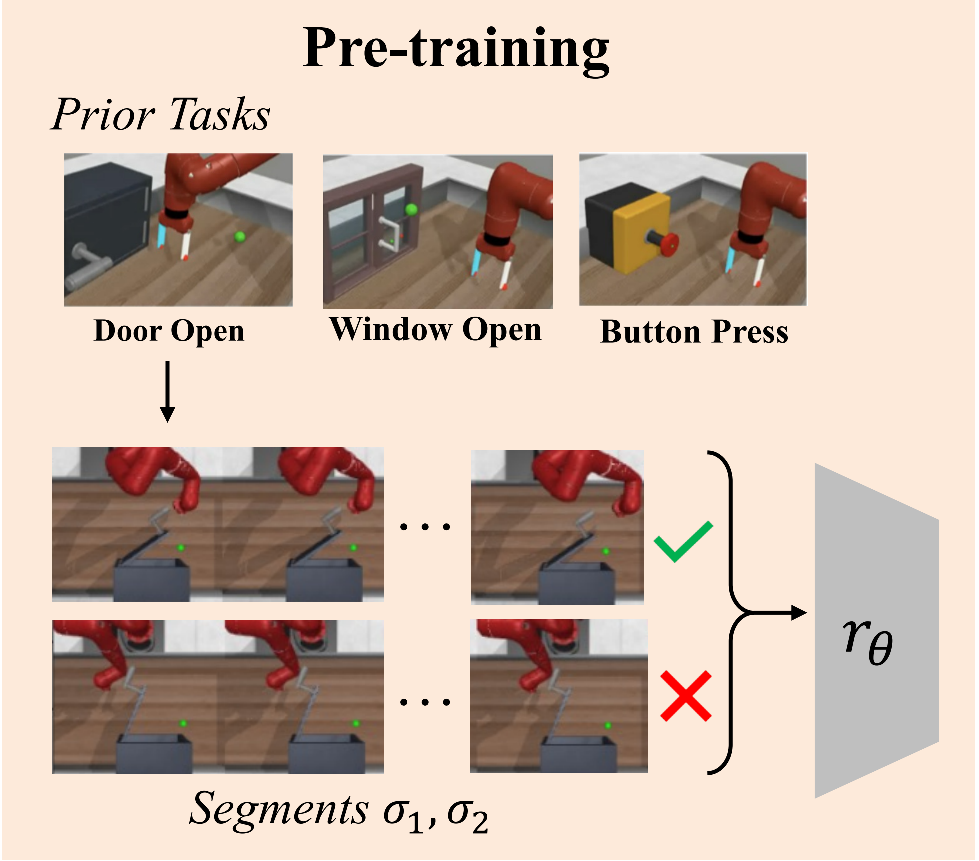
Door Open



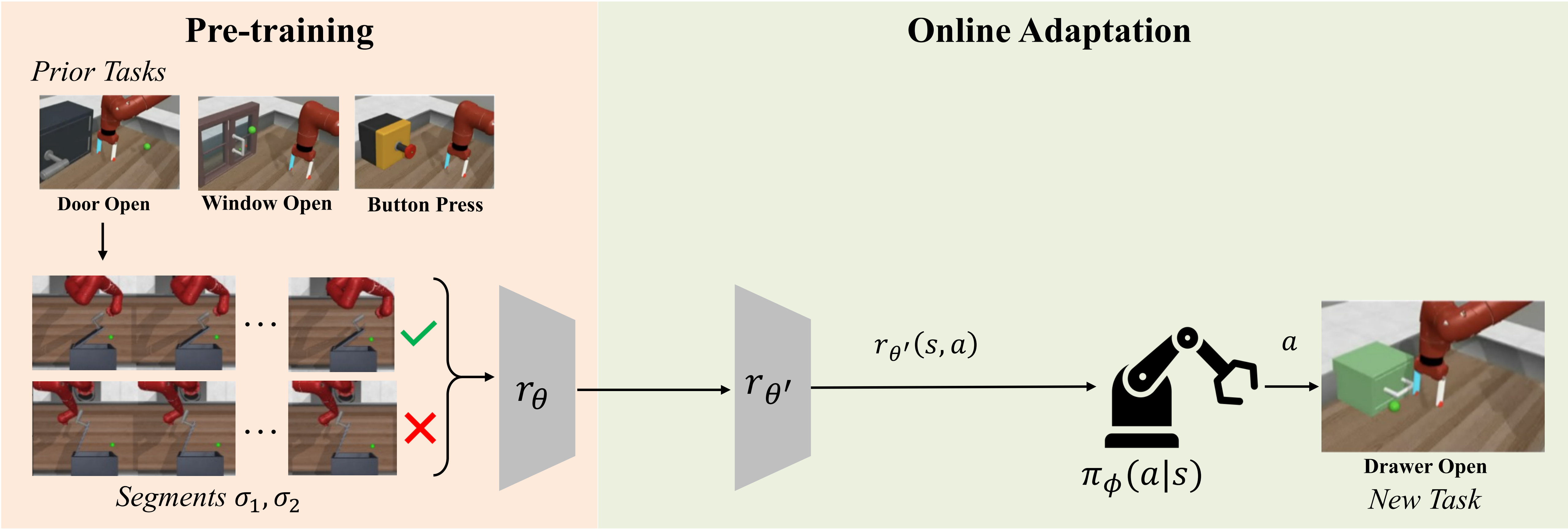
Window Open



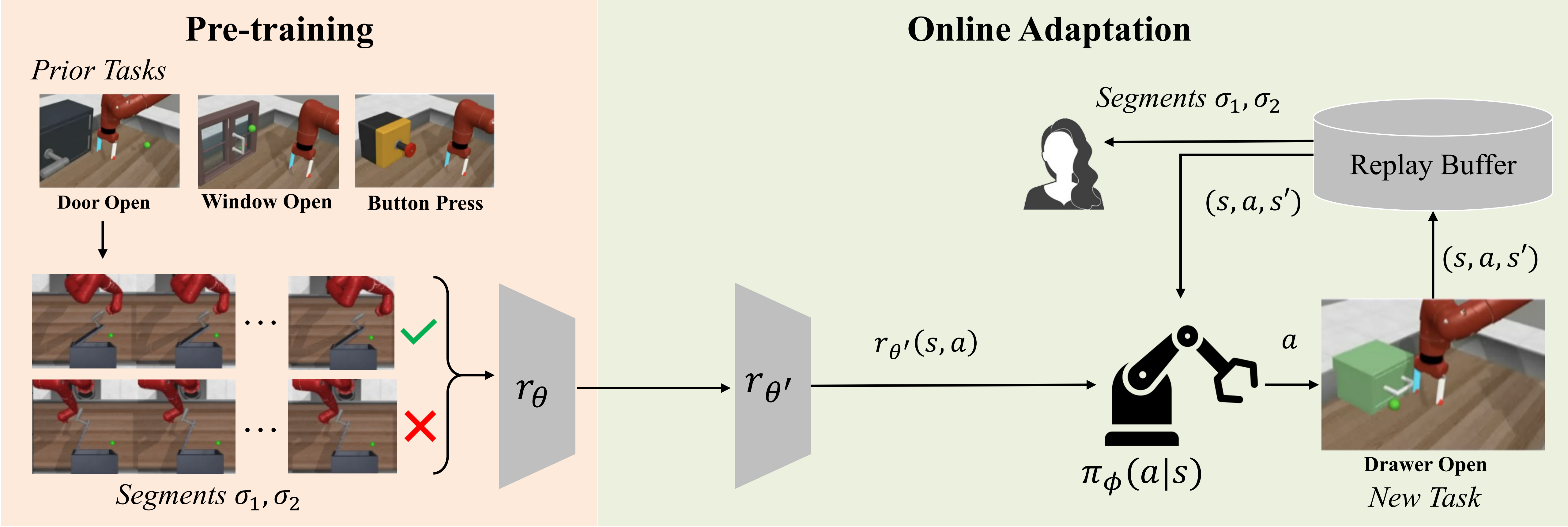
Button Press

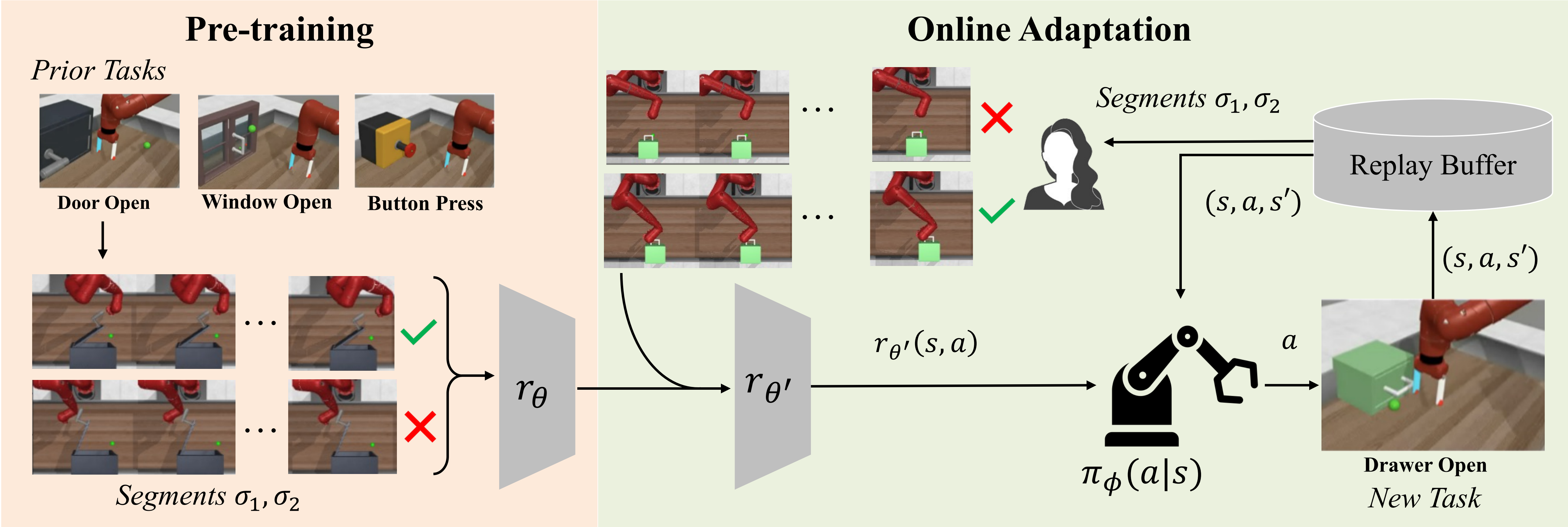

Pre-training
Prior Tasks
Door Open
Window Open
Button Press
r_θ
Segments σ_1, σ_2


Pre-training
Online Adaptation
Prior Tasks
Door Open
Window Open
Button Press
r_θ
Segments σ_1, σ_2



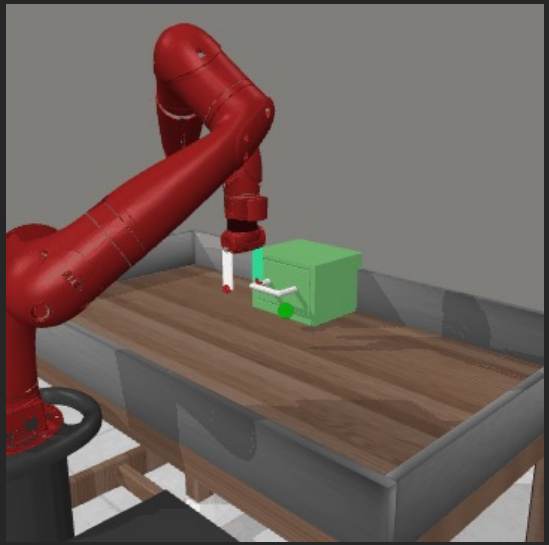
Pre-training
Online Adaptation
Prior Tasks
Door Open
Window Open
Button Press
Segments σ_1, σ_2
r_θ
$r_{\theta'}$
$r_{\theta'}(s,a)$
a
$\pi_\phi(a|s)$
Drawer Open
New Task


Pre-training
Online Adaptation
Prior Tasks
Door Open
Window Open
Button Press
Segments σ_1, σ_2
r_θ
$r_{\theta'}$
$r_{\theta'}(s,a)$
Segments σ_1, σ_2
Replay Buffer
(s,a,s')
(s,a,s')
a
$\pi_\phi(a|s)$
Drawer Open
New Task

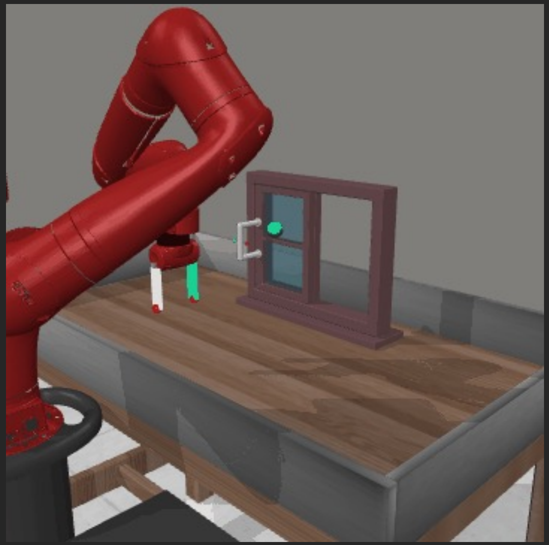


Pre-training
Prior Tasks
Door Open
Window Open
Button Press
Segments σ_1, σ_2
r_θ
Online Adaptation
Segments σ_1, σ_2
Replay Buffer
(s, a, s')
(s, a, s')
$r_{\theta'}$
$r_{\theta'}(s, a)$
a
$\pi_\phi(a|s)$
Drawer Open
New Task

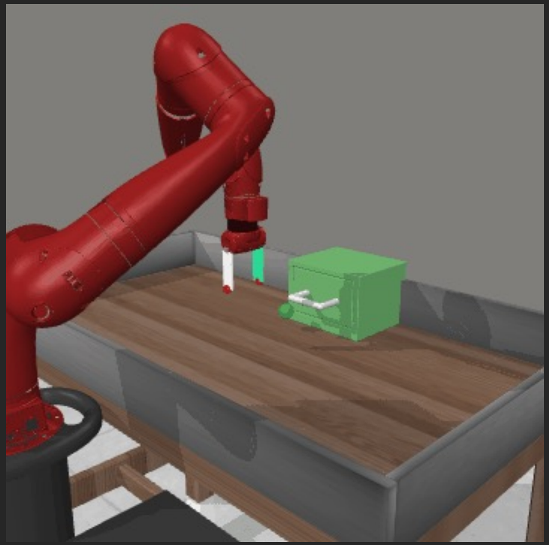
OURS



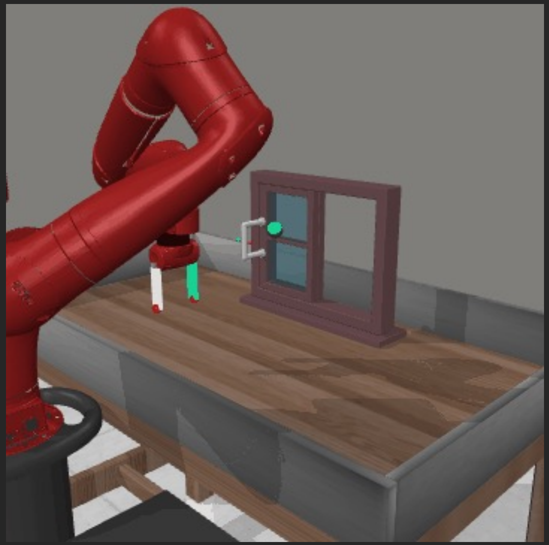
OURS



PEBBBLE



PEBBBLE



1. What if our reward function is nonlinear?
2. What if our reward function is multimodal?
3. What if we have a neural reward function?