# Optimizing Robot Action for & around People

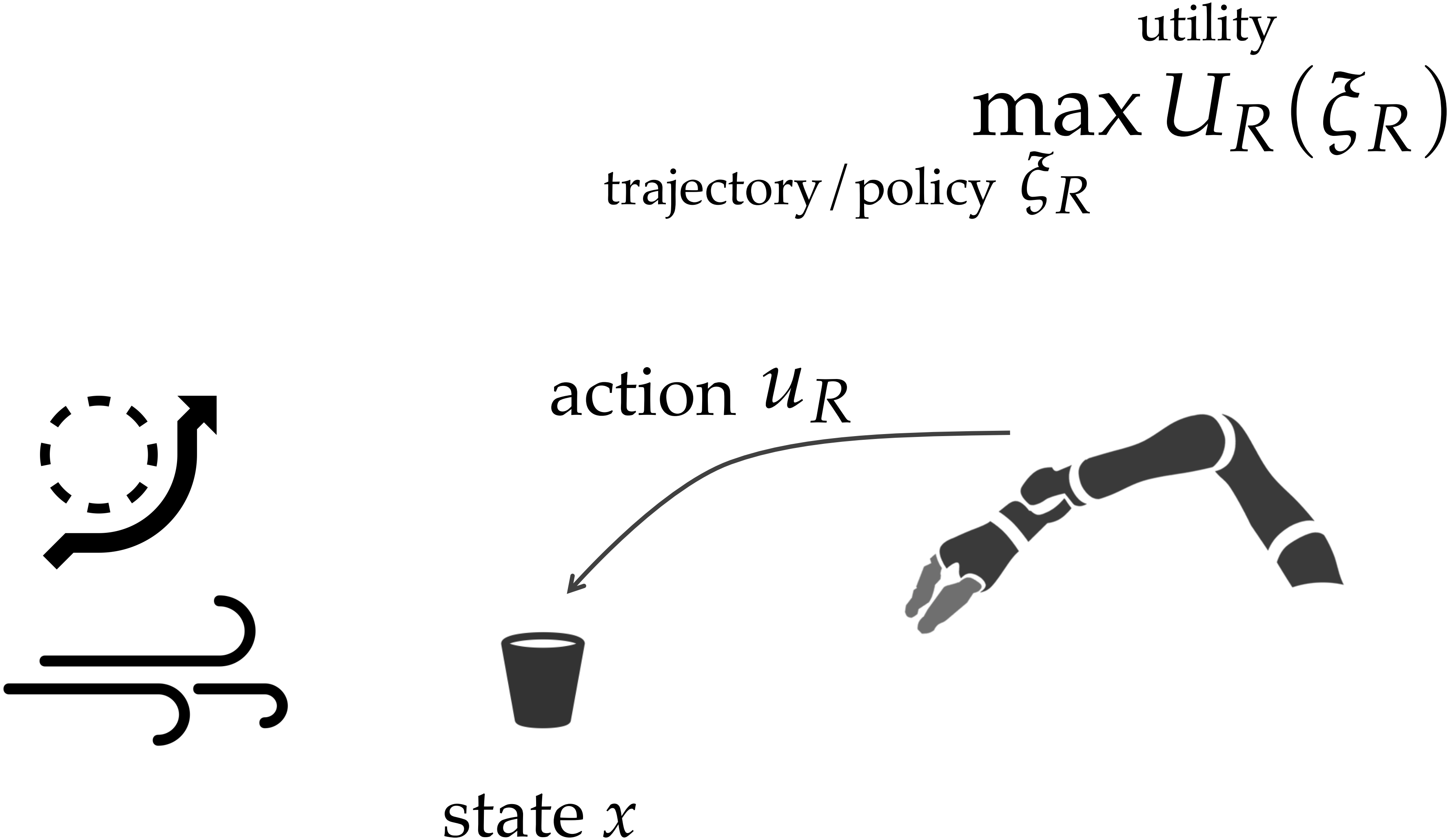
Anca Dragan

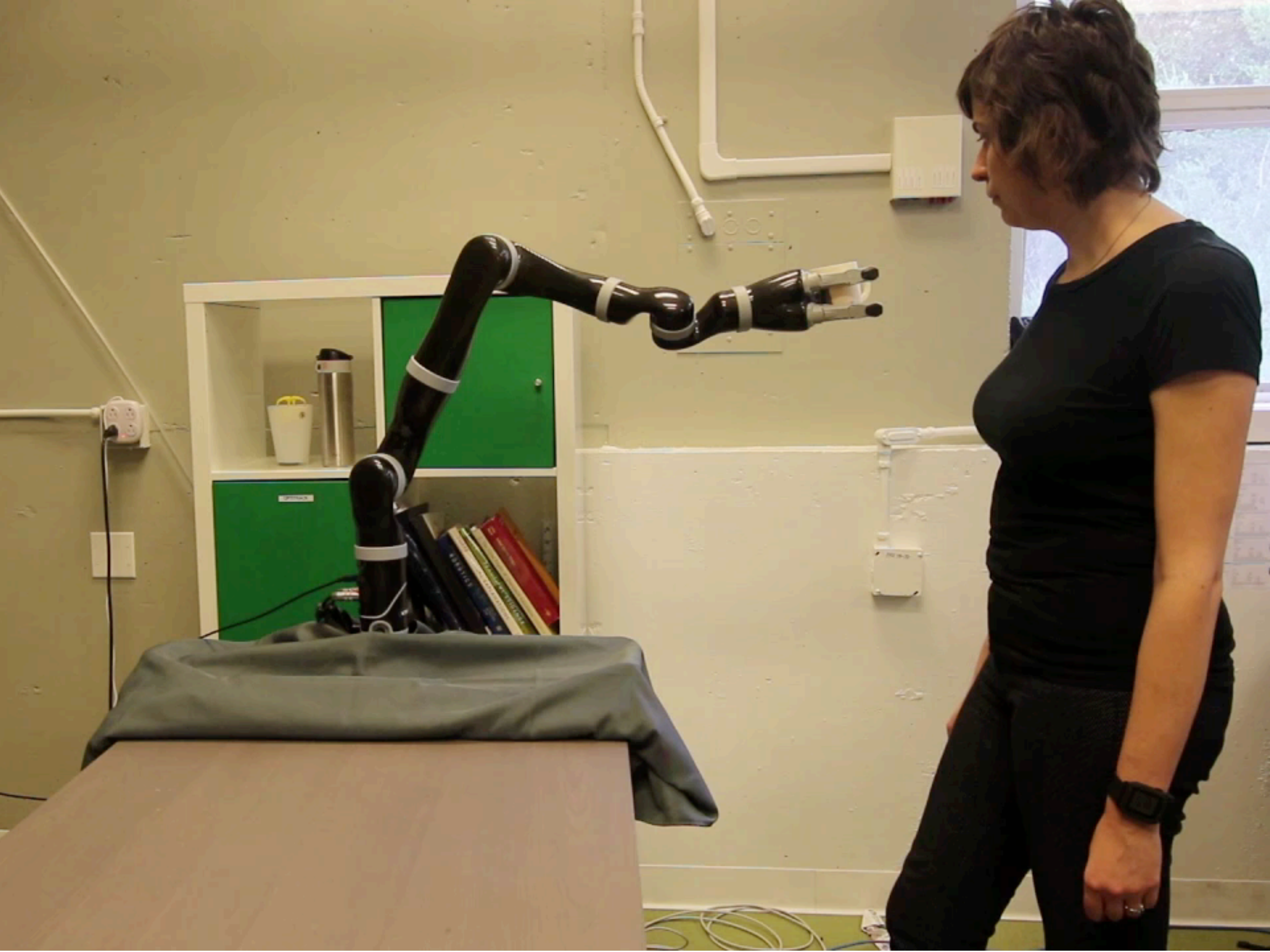
InterACT Laboratory

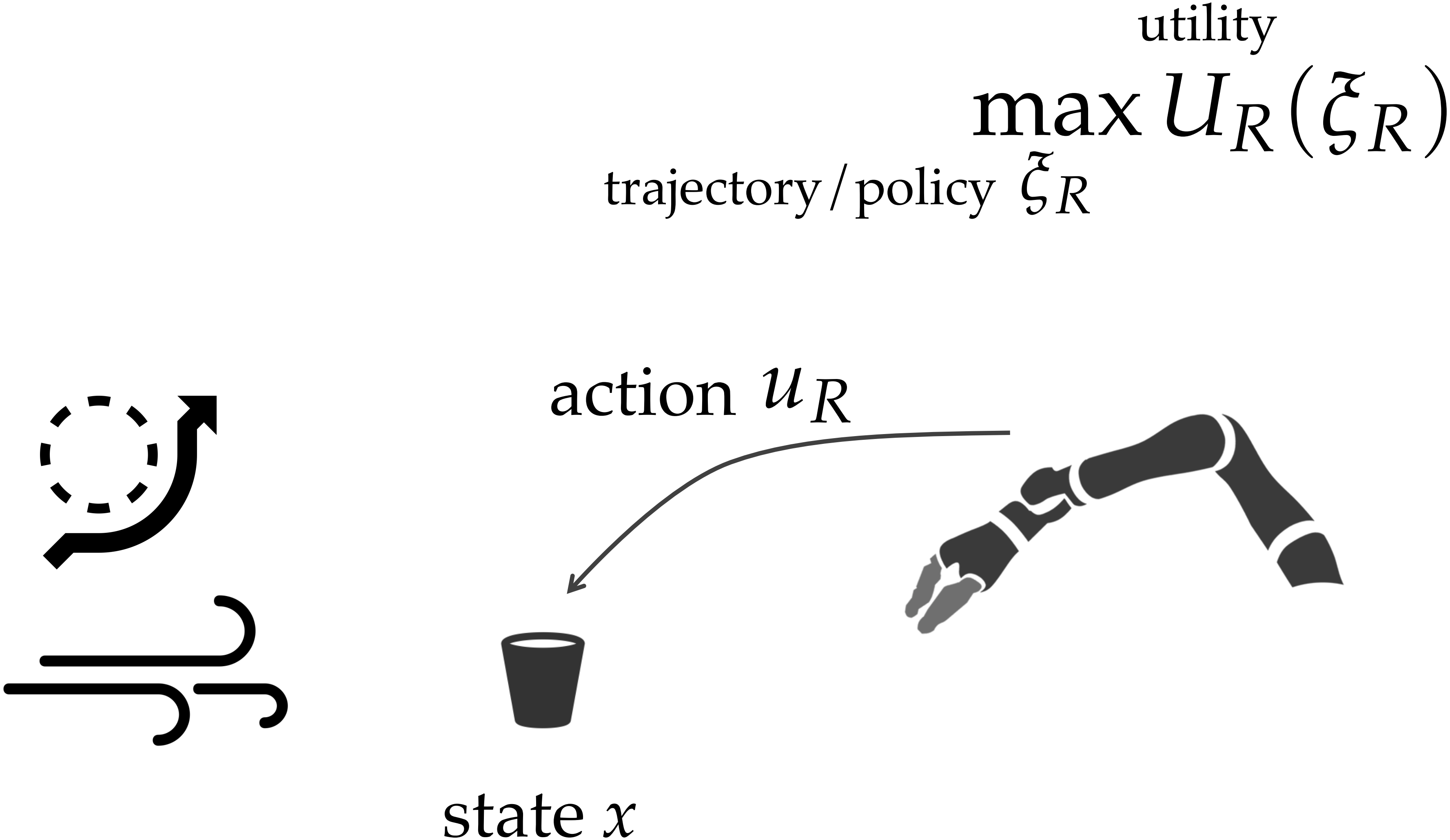
BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

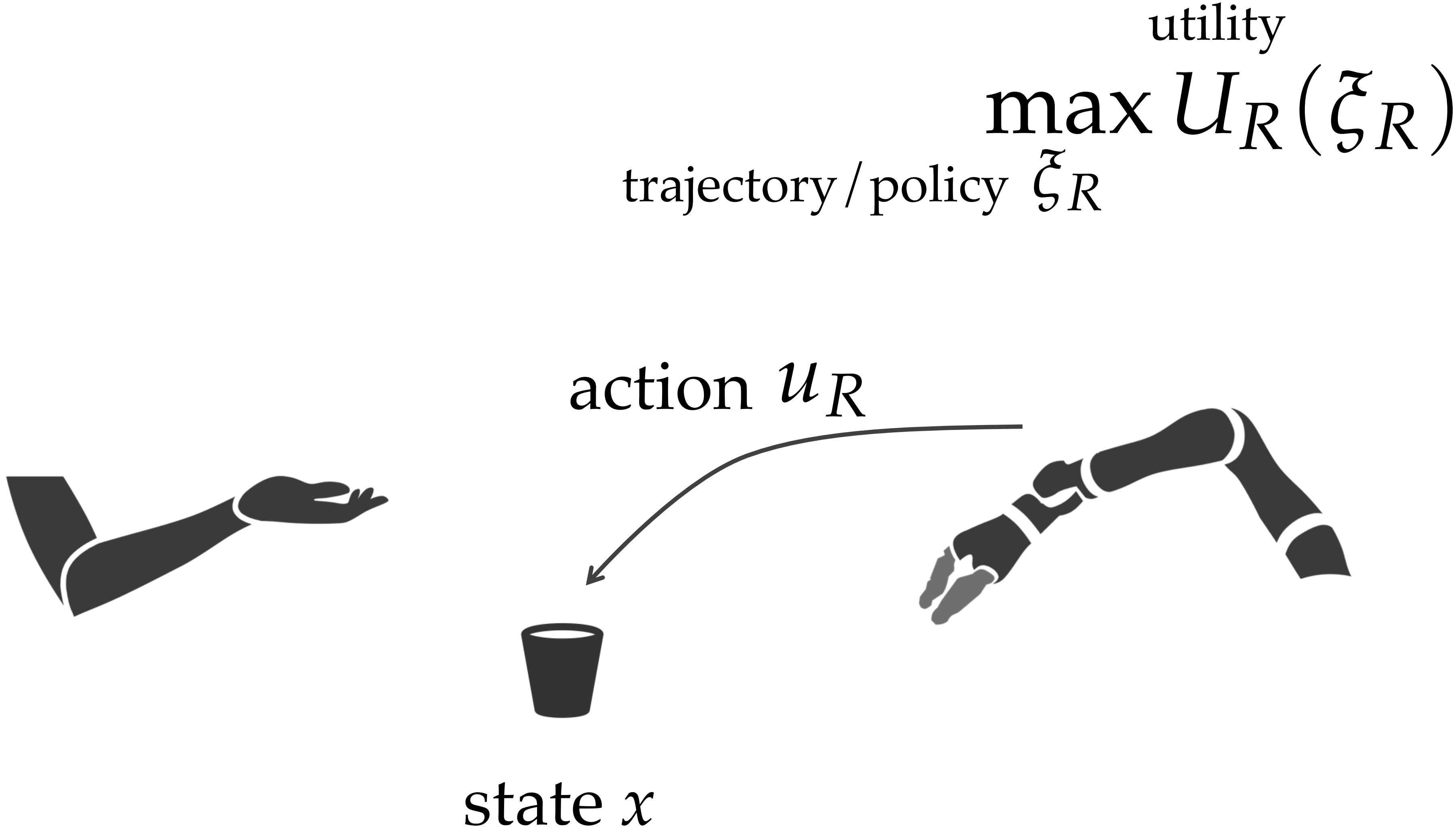

utility
$\max U_R(\xi_R)$
trajectory / policy $\xi_R$
action $u_R$
state $x$


Boston Dynamics


utility
$\max U_R(\xi_R)$
trajectory / policy $\xi_R$
action $u_R$
state $x$

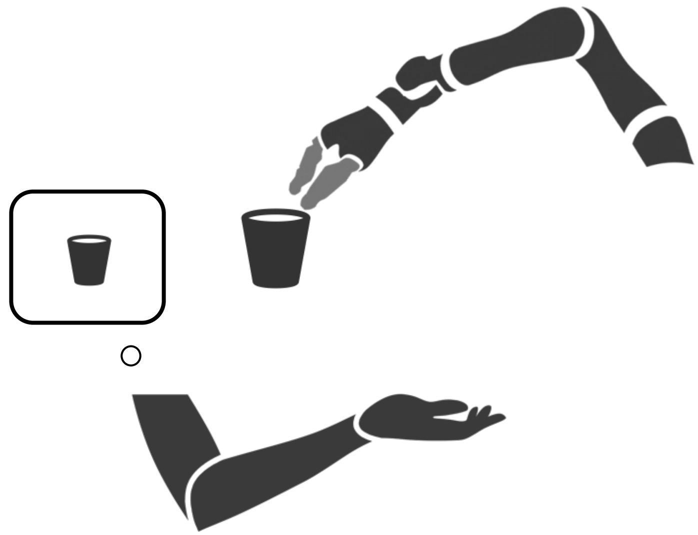



utility
$\max U_R(\xi_R)$
trajectory / policy $\xi_R$
action $u_R$
state $x$

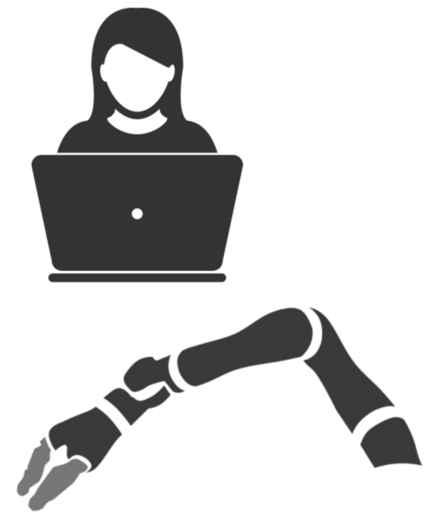

utility
$\max U_R(\xi_R)$
trajectory / policy $\xi_R$
action $u_R$
state $x$

# 3 types of people in a robot's life

person in its environment

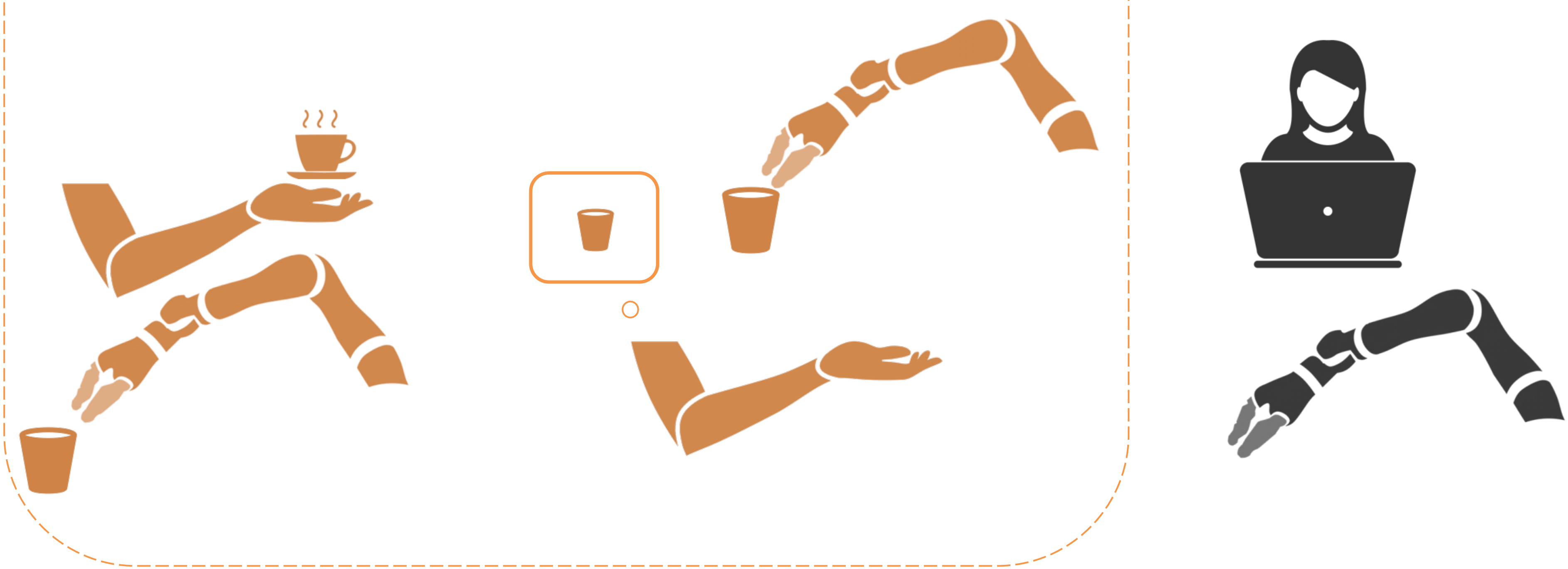


its end-user



its designer

[RSS16,IROS16,AURO17,
ISER16,HRI17a,WAFR16,
HRI16,ACL17,RSS17a, HRI18]



Optimize utility in coordination with people.

[NIPS16,ICRA16,CDC16,
HRI17,ICRA17,IJCAI17a,IJCAI17b,
RSS17b,CoRL17, ISRR17, NIPS17, HRI18a, HRI18b]

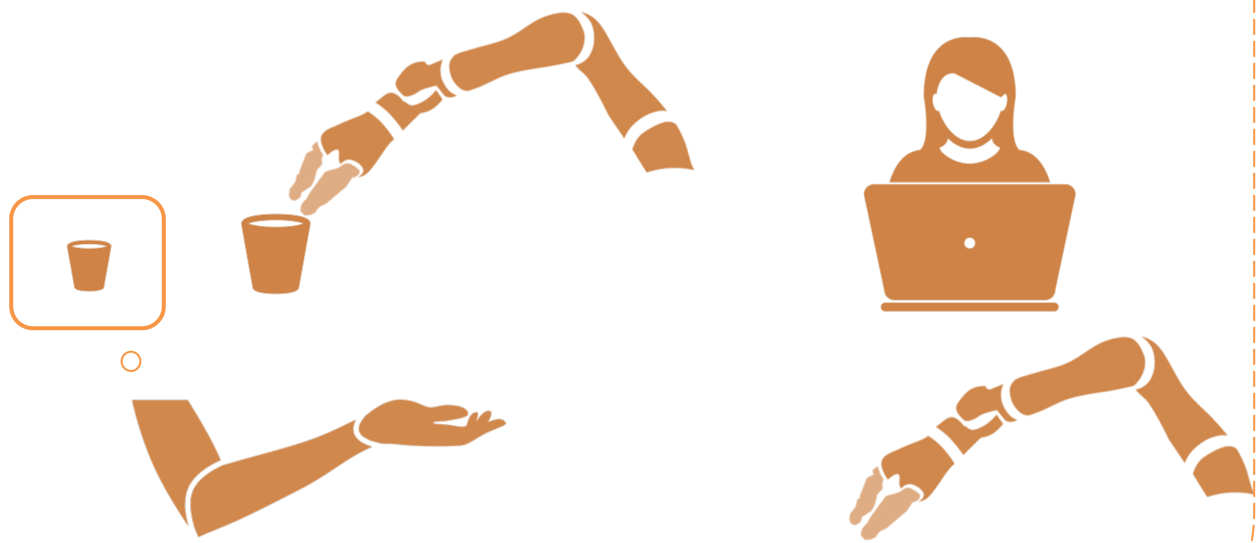
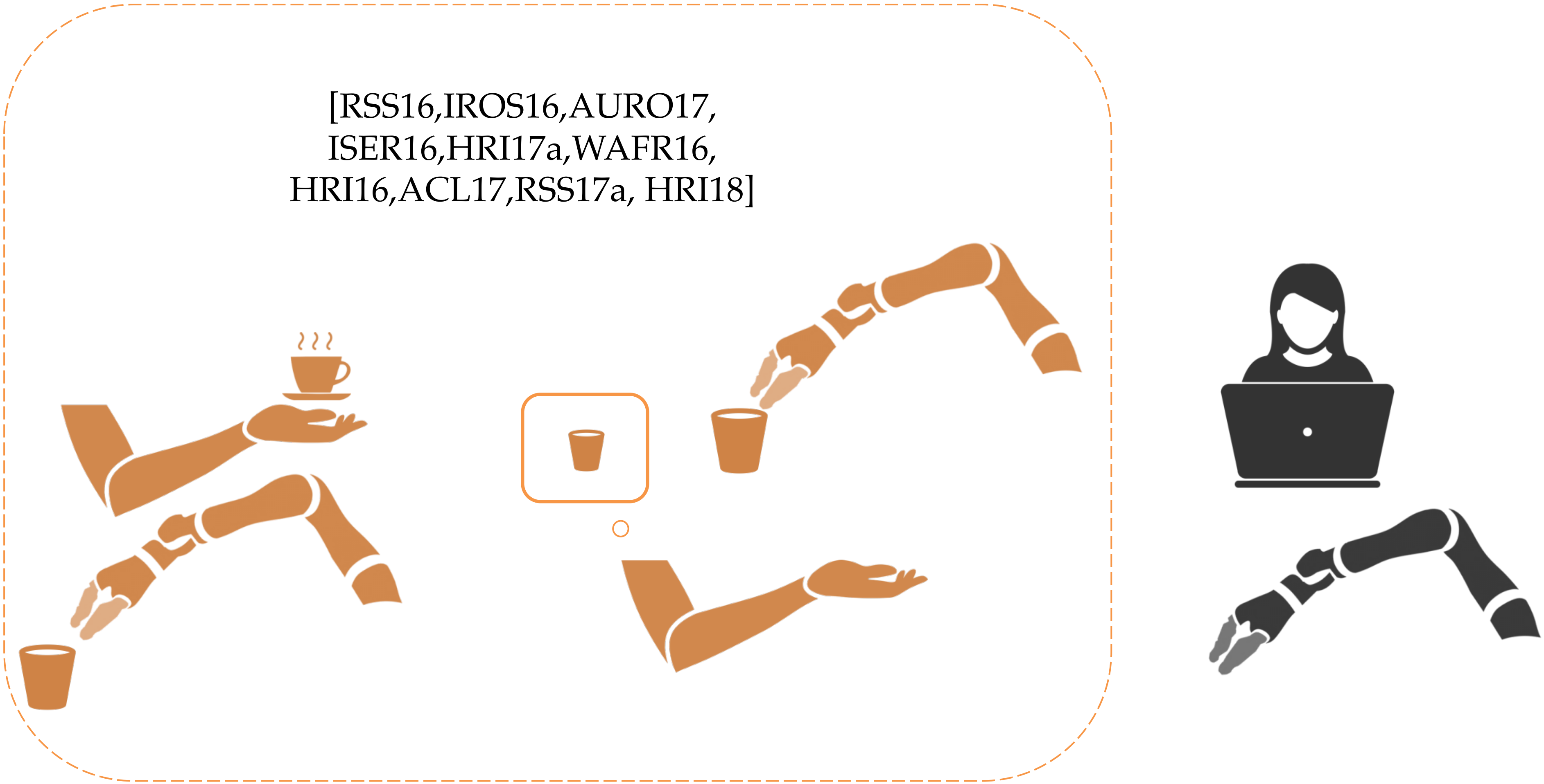
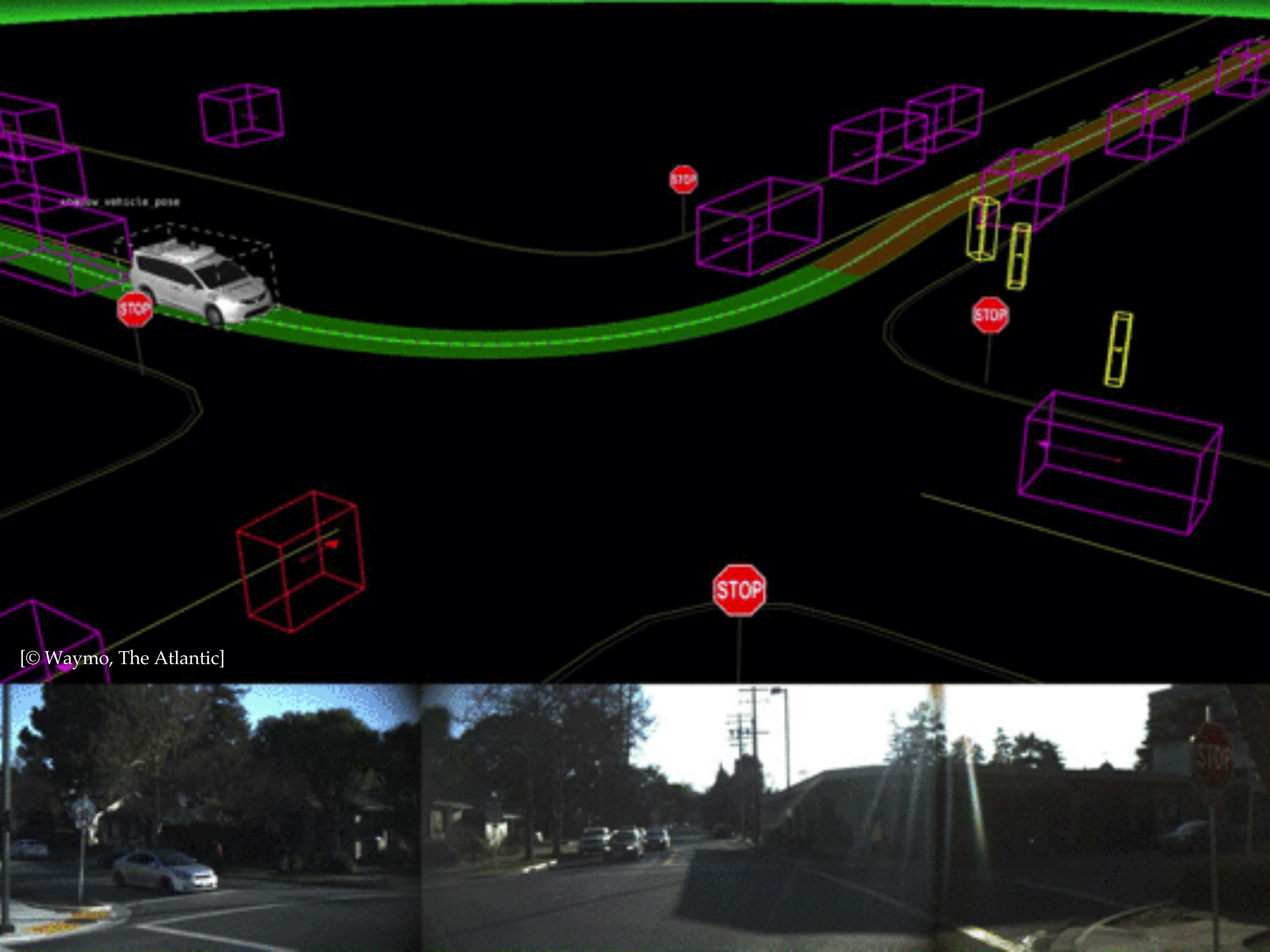


Figure out what utility to optimize.




Optimize utility in coordination with people.

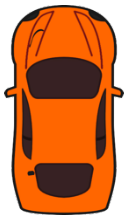


[© Waymo, The Atlantic]

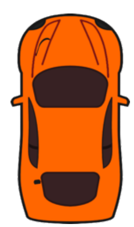
# Maximize robot utility..

$$\xi_R^* = \arg\max_{\xi_R} U_R(\xi_R)$$

robot plan



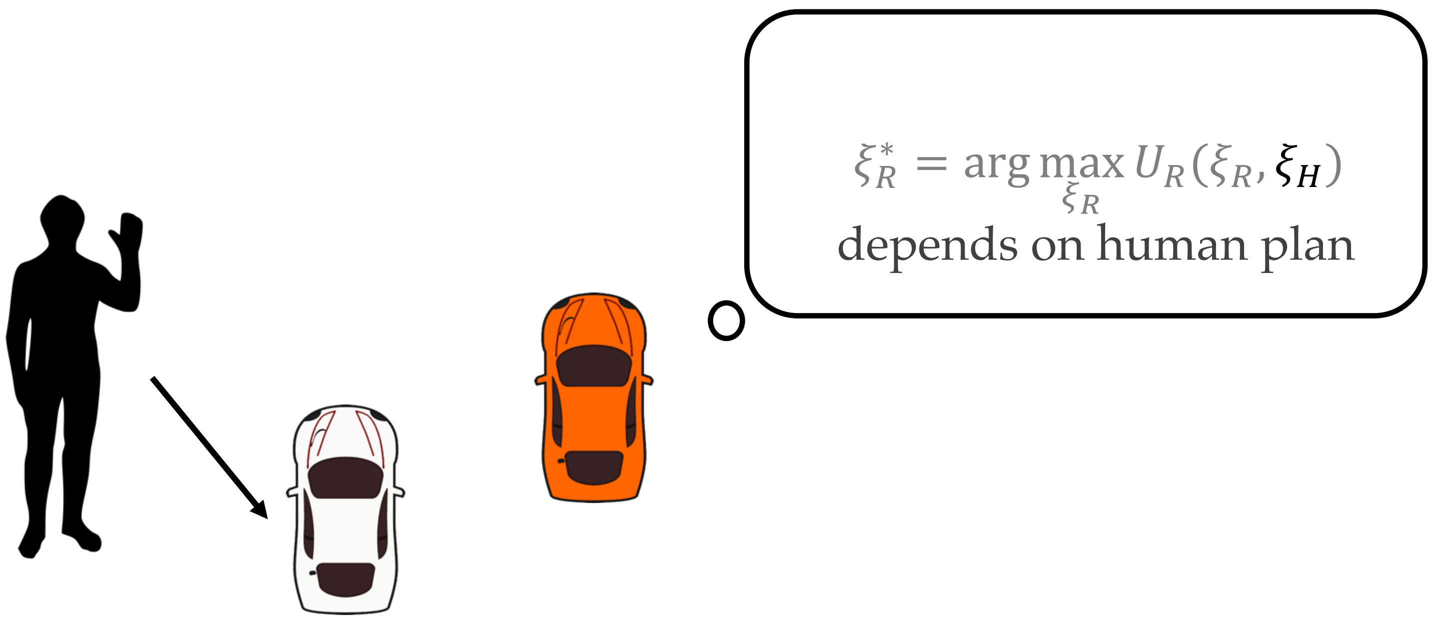
# Maximize robot utility..



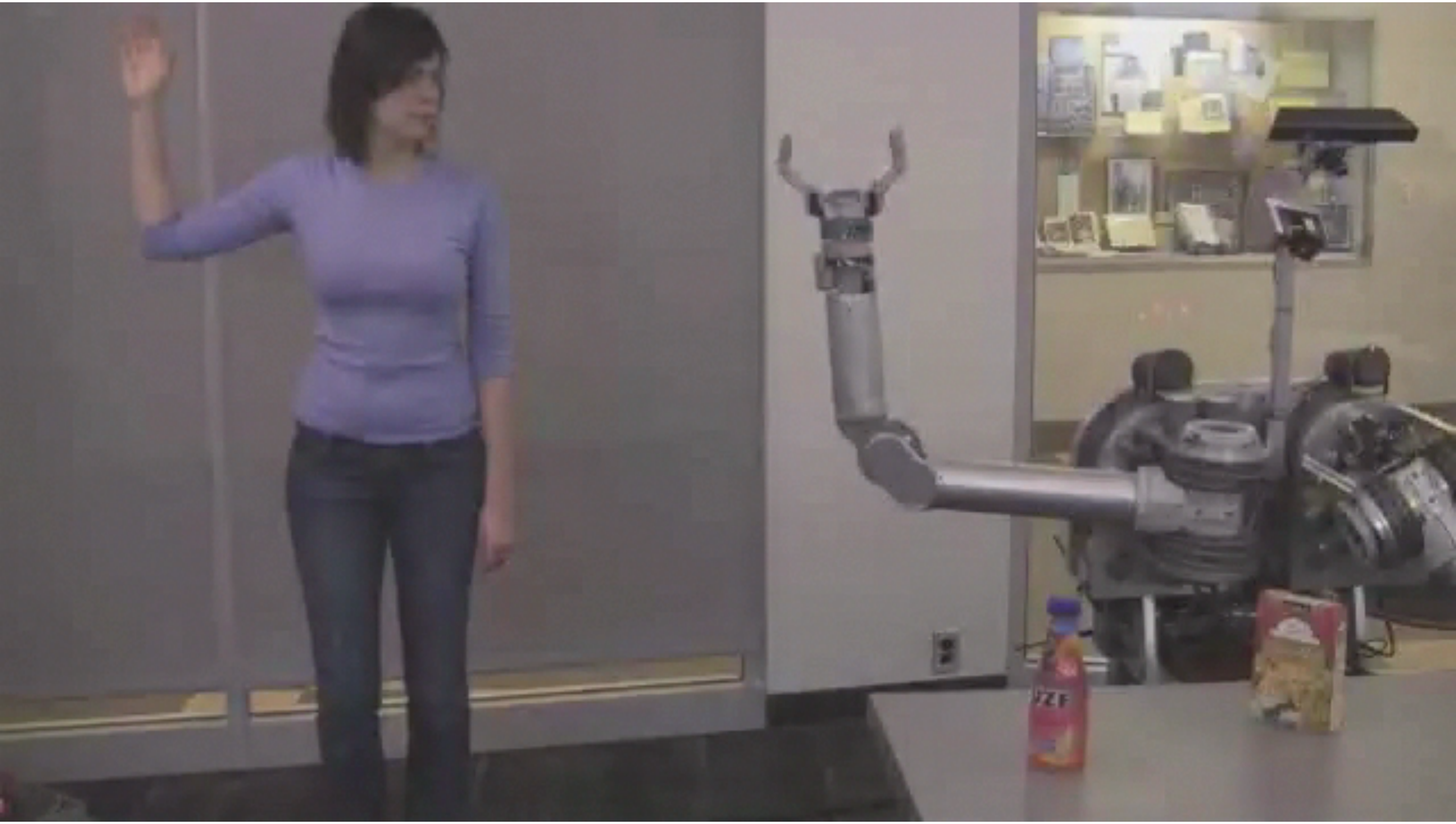
maximizes robot utility

$$\xi_R^* = \arg\max_{\xi_R} U_R(\xi_R)$$

# When the human is also acting.



# Predict H action, optimize R action in response



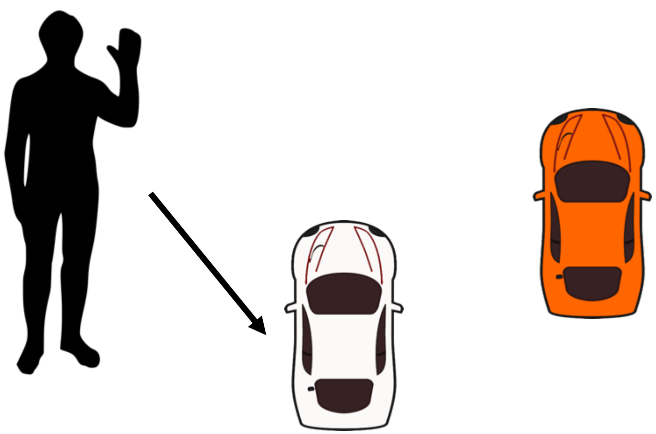
*Formalizing Assistive Teleoperation [RSS'12]*


*Formalizing Assistive Teleoperation [RSS'12]*

# HRI as predict-then-react

$$\xi_H^* = \arg\max_{\xi_H} U_H(\xi_H)$$

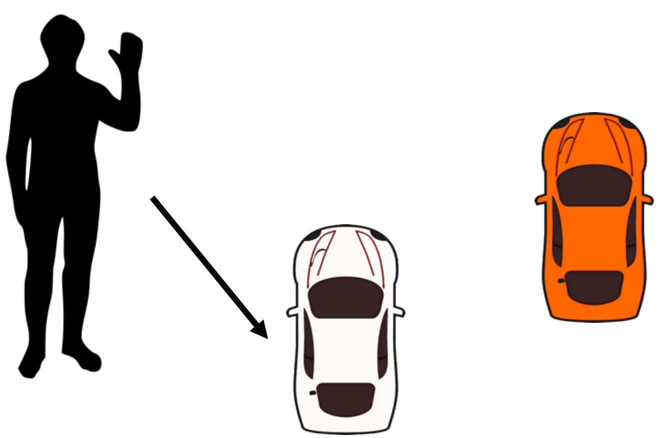
predicted plan



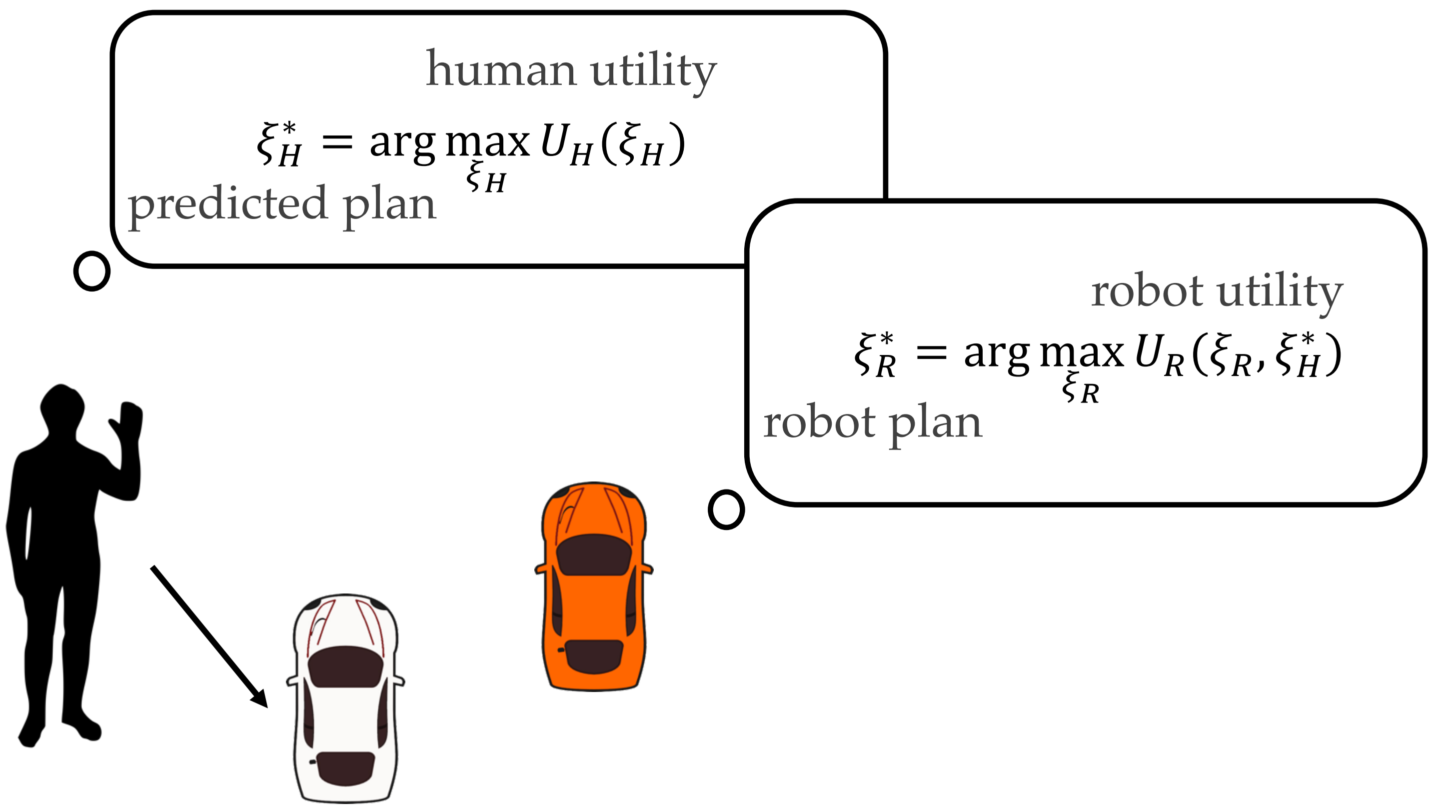
# HRI as predict-then-react

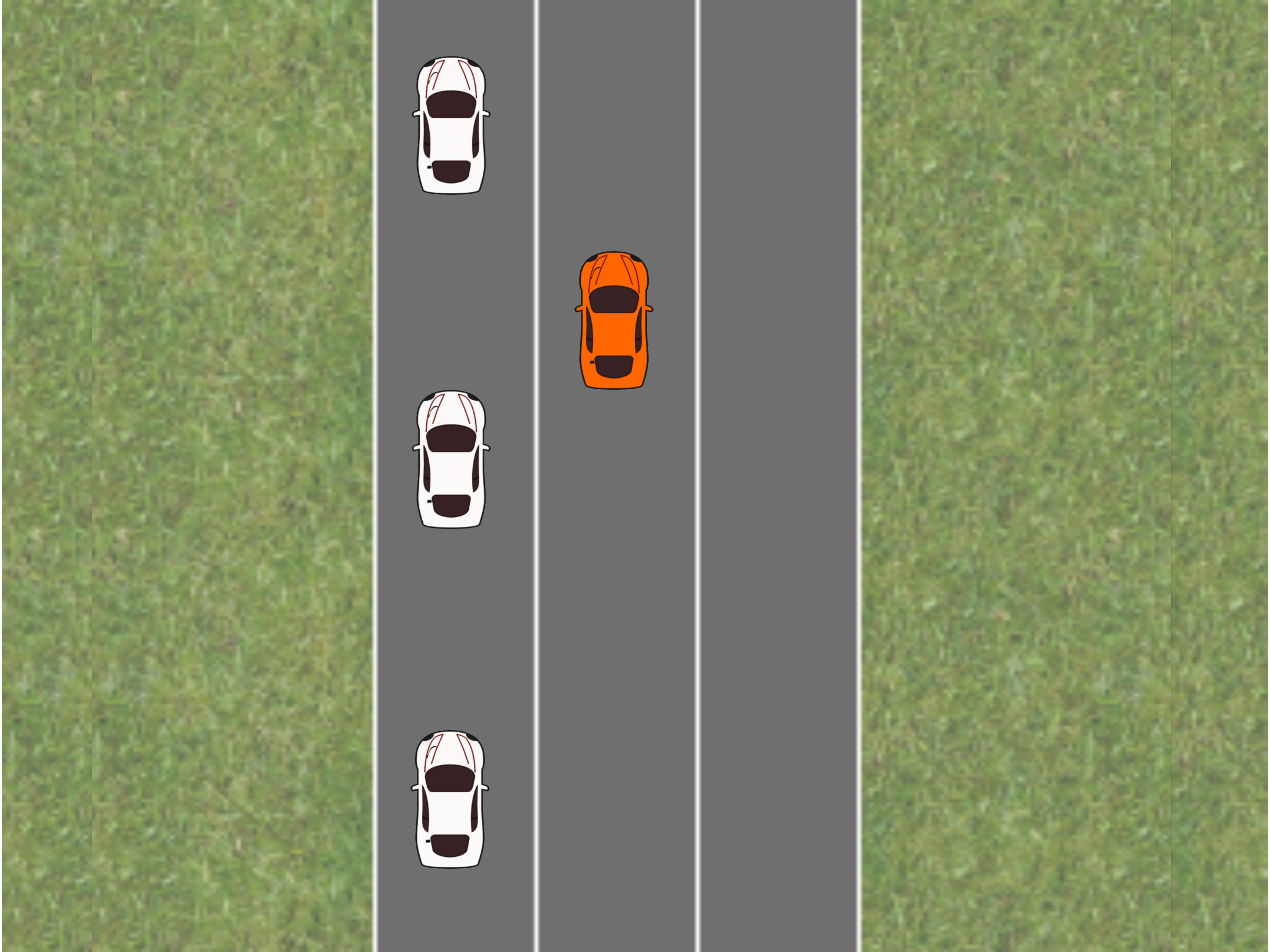
maximizes human utility

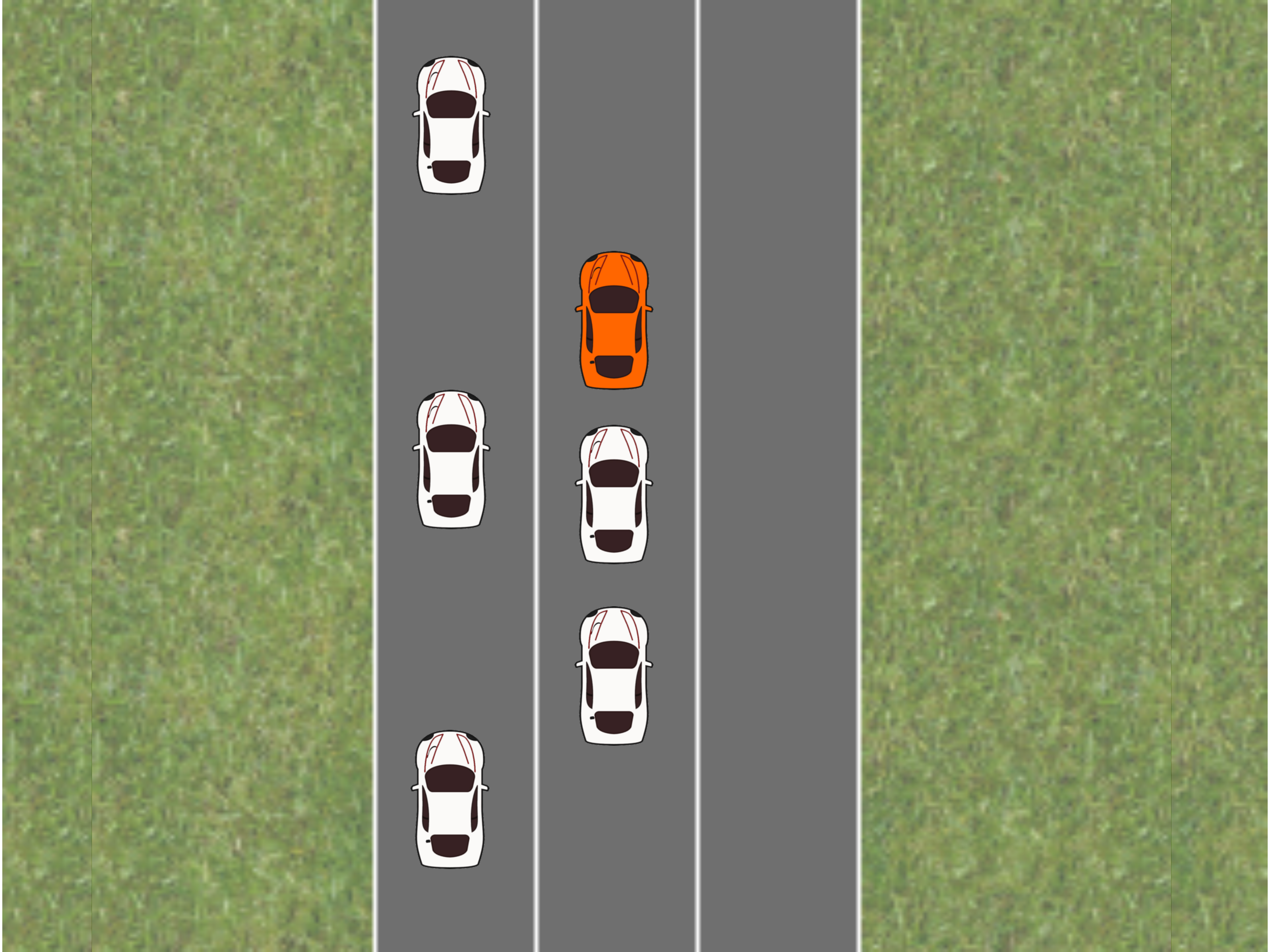
$$\xi_H^* = \arg\max_{\xi_H} U_H(\xi_H)$$

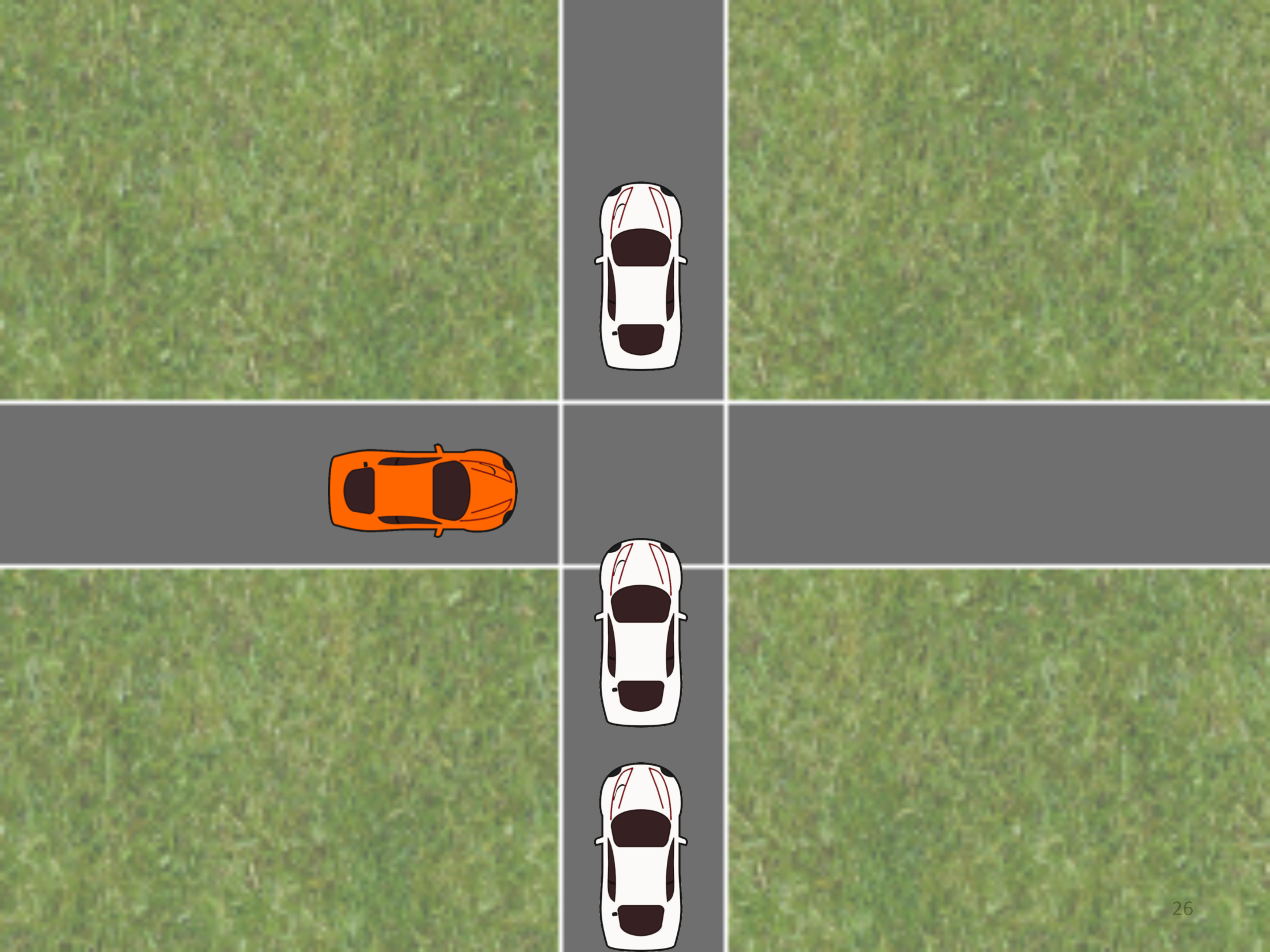


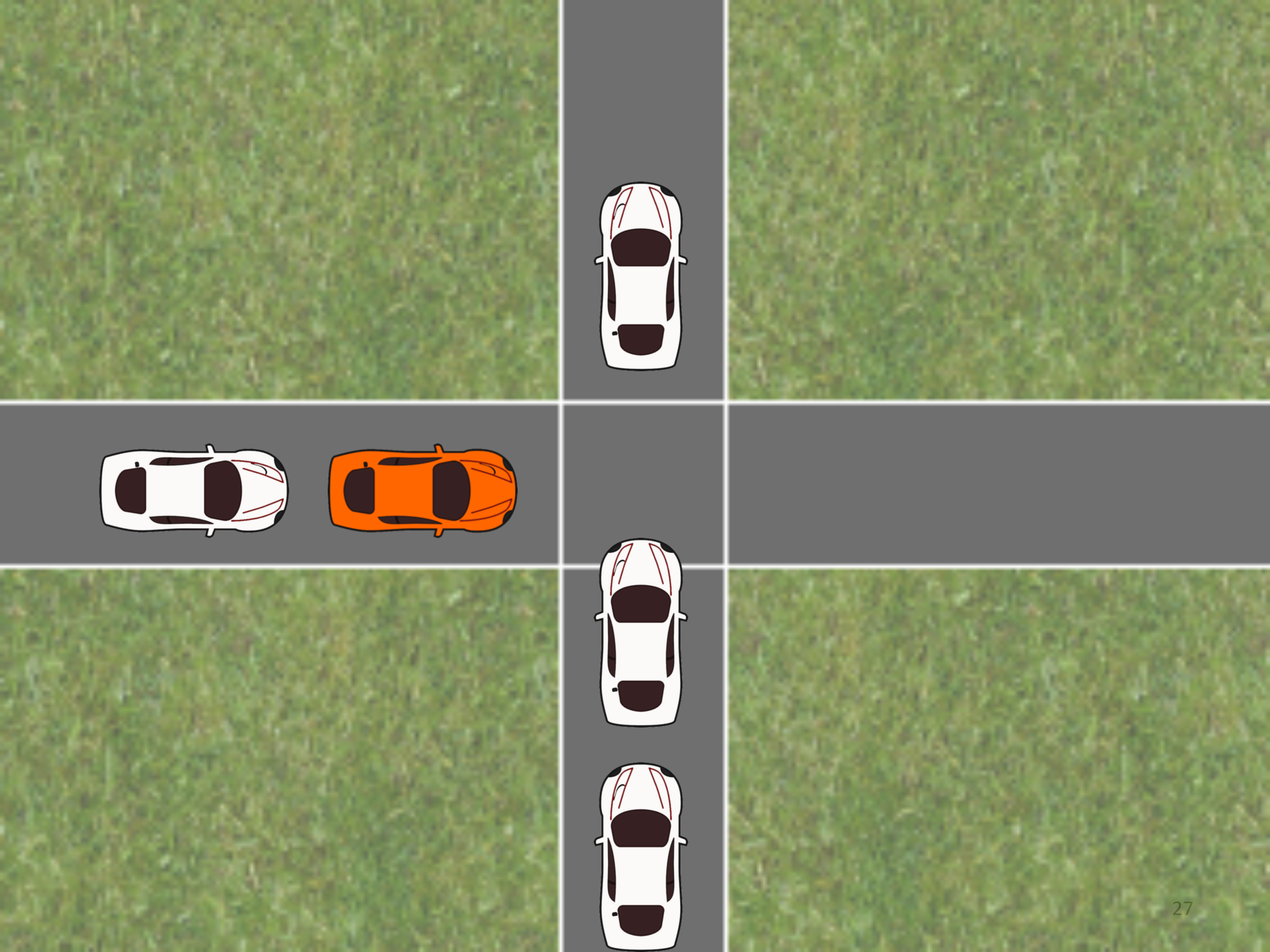
# HRI as predict-then-react



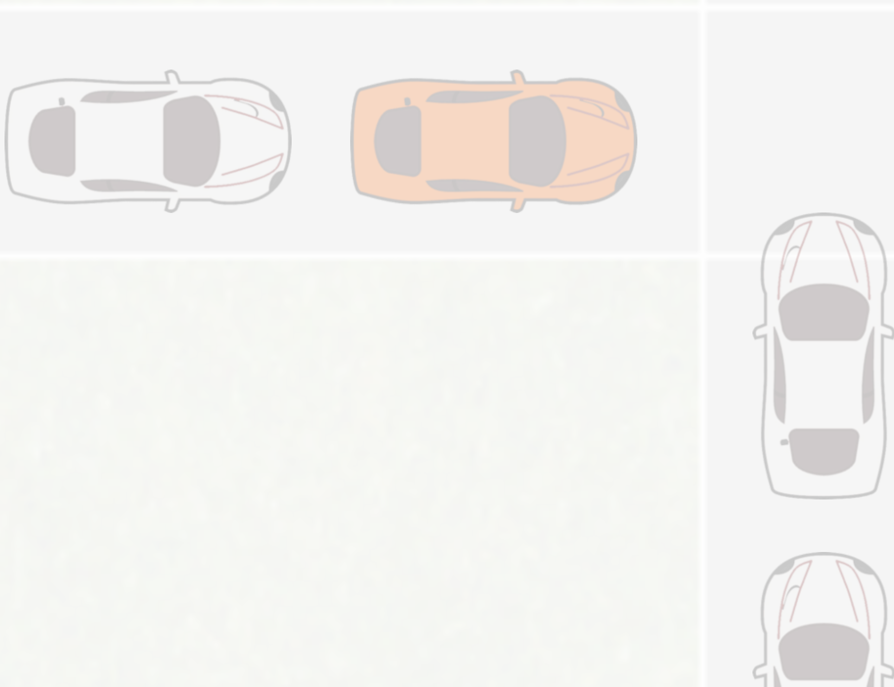




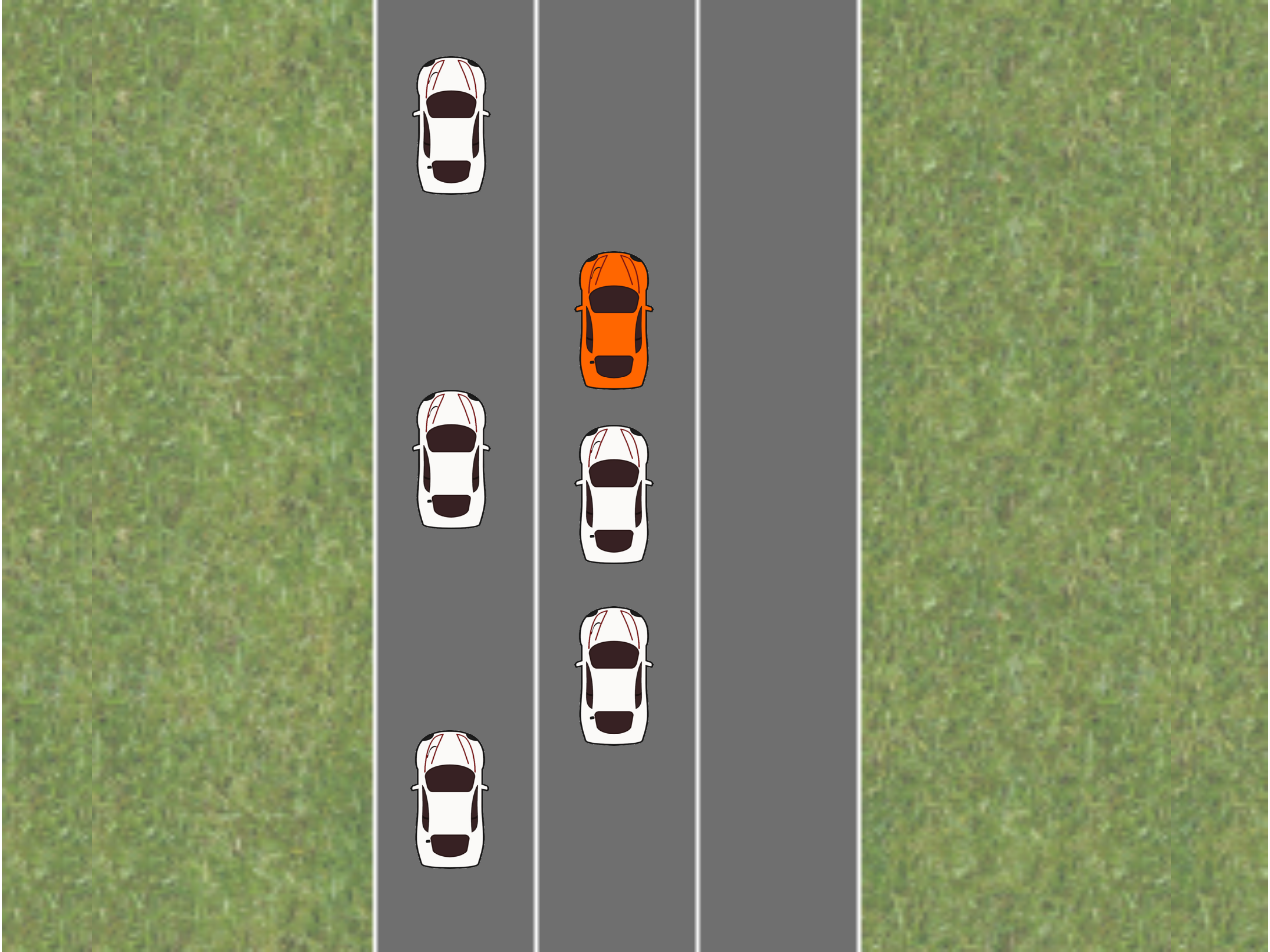




*One Google car [..] couldn't get through a four-way stop because its sensors kept waiting for other (**human**) drivers [..]. The human **drivers kept inching forward** looking for the advantage – paralyzing Google's robot.*



"Google's Driverless Cars Run Into Problem: Cars With Drivers"[Richtel&Dougherty]

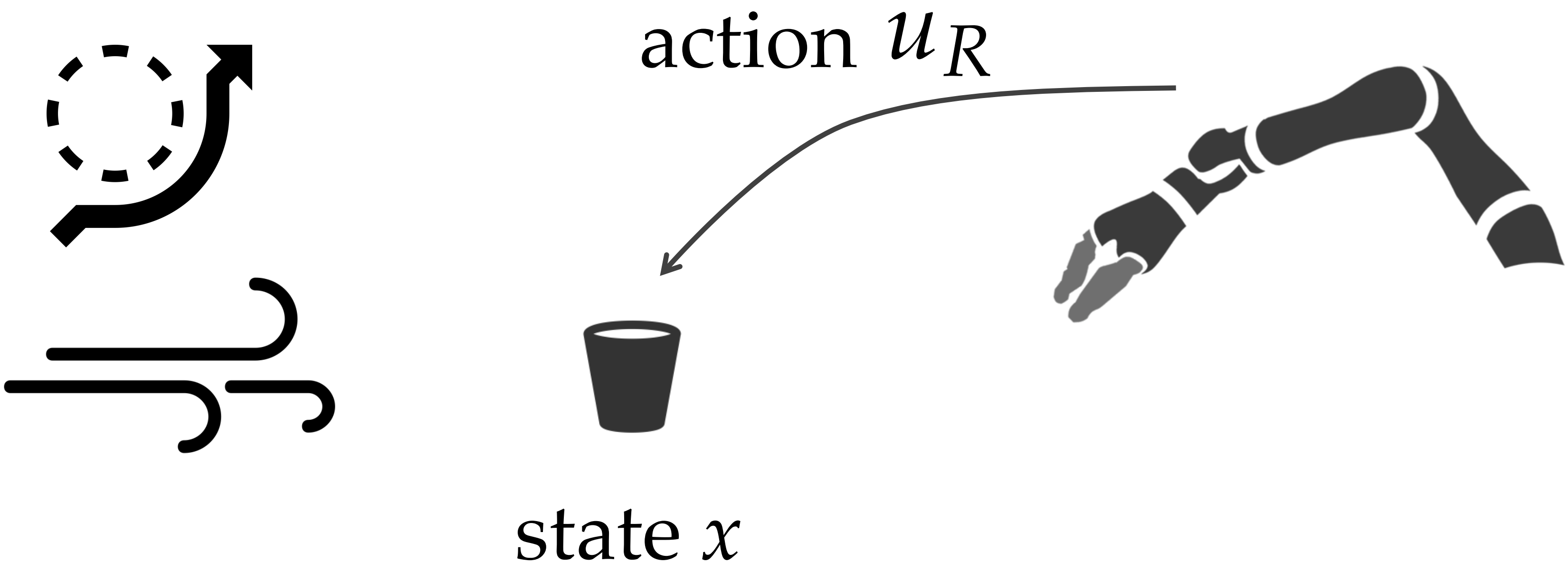



THE POWER YOU CAN TRUST
FORTUNER
CIVIC

Robot actions
affect human actions.

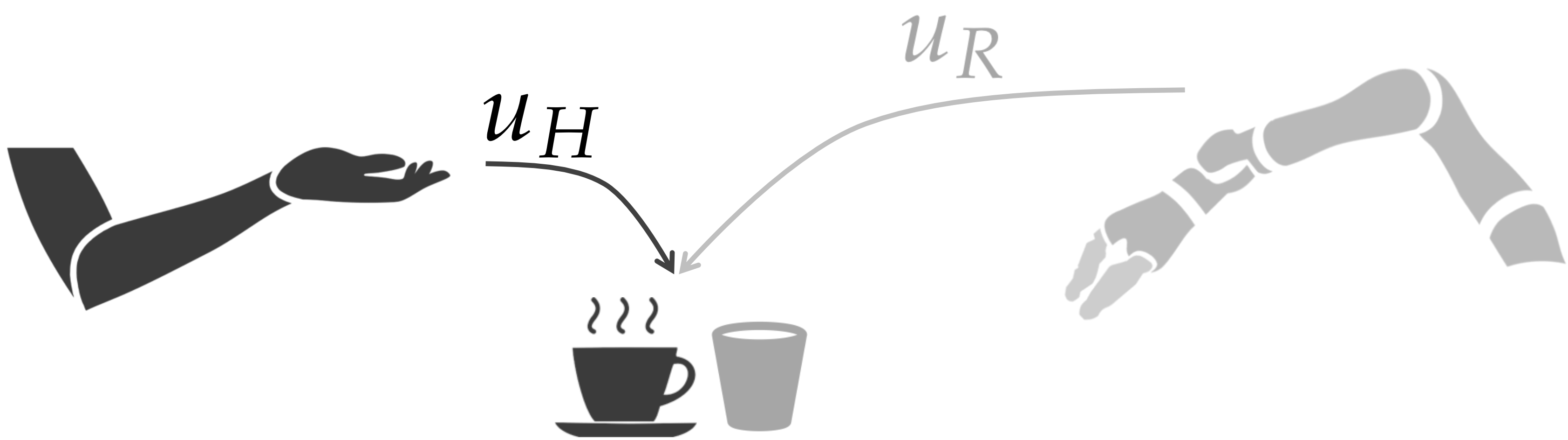
Leveraging this effect can make seemingly impossible plans possible.

$$\max_{\xi_R} U_R(\xi_R)$$

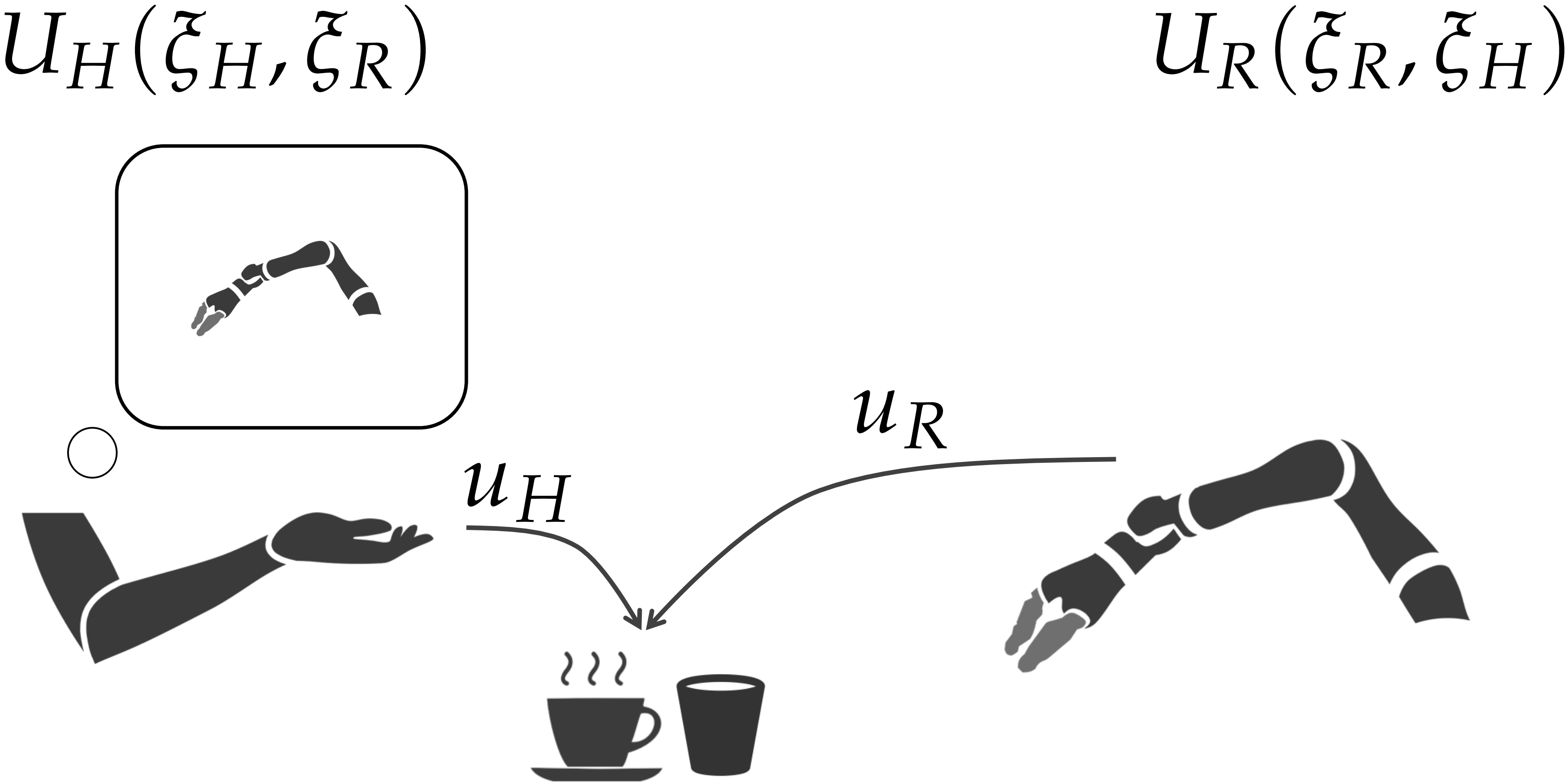



People are not obstacles or disturbances.

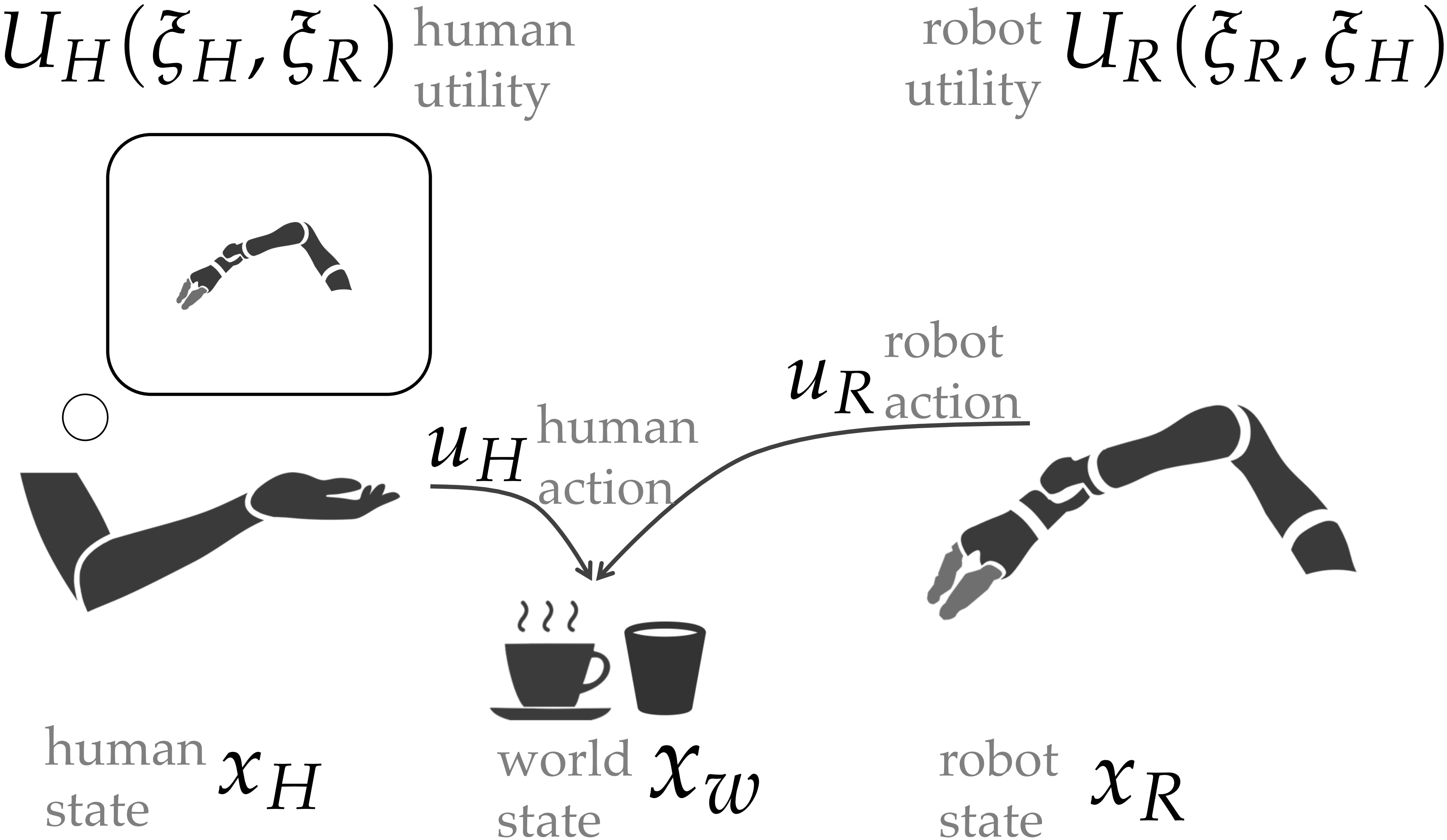
$$\max_{\xi_H} U_H(\xi_H)$$



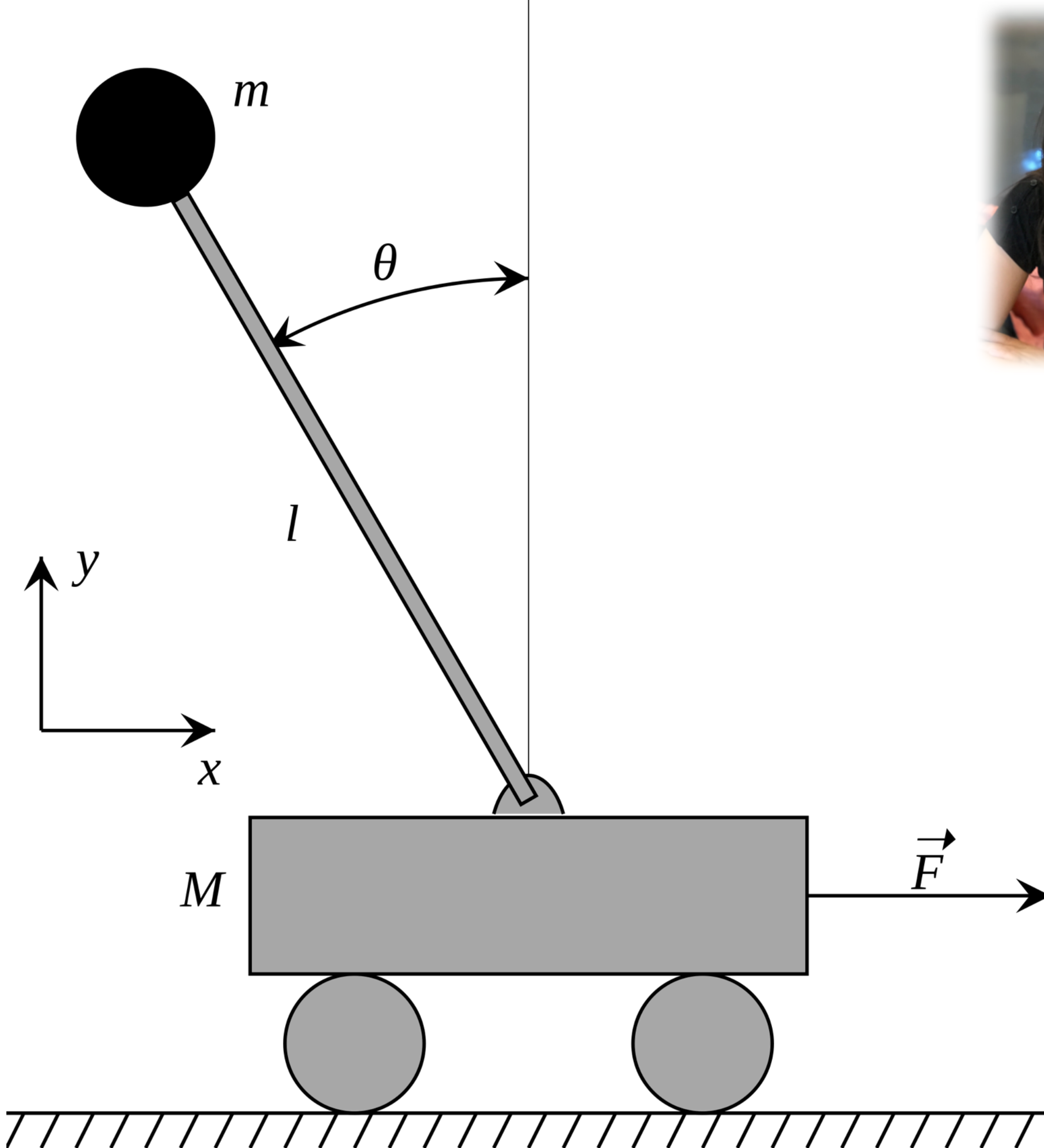

People do not act in isolation.



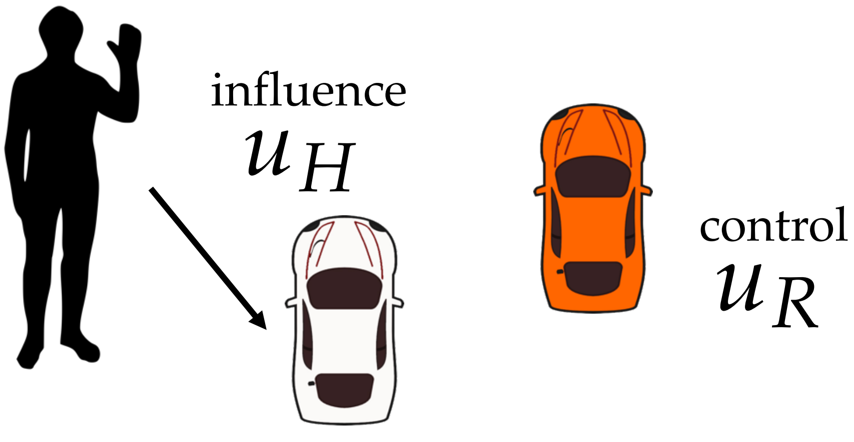

Actual interaction is game-theoretic.




Actual interaction is game-theoretic.




Dorsa Sadigh

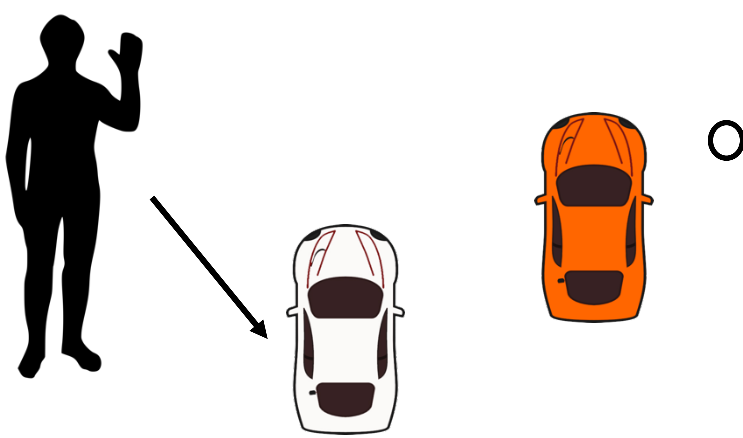


influence
$u_H$
control
$u_R$

# HRI as predict-then-react

$$\xi_H^* = \arg\max_{\xi_H} U_H(\xi_H)$$

$$\xi_R^* = \arg\max_{\xi_R} U_R(\xi_R, \xi_H^*)$$

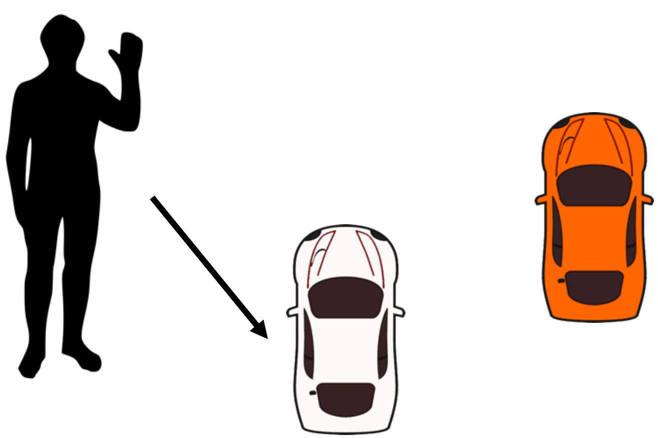


*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*

# HRI as an underactuated system

$$\xi_H^*(\xi_R) = \arg\max_{\xi_H} U_H(\xi_H, \xi_R)$$

$$\xi_R^* = \arg\max_{\xi_R} U_R(\xi_R, \xi_H^*)$$

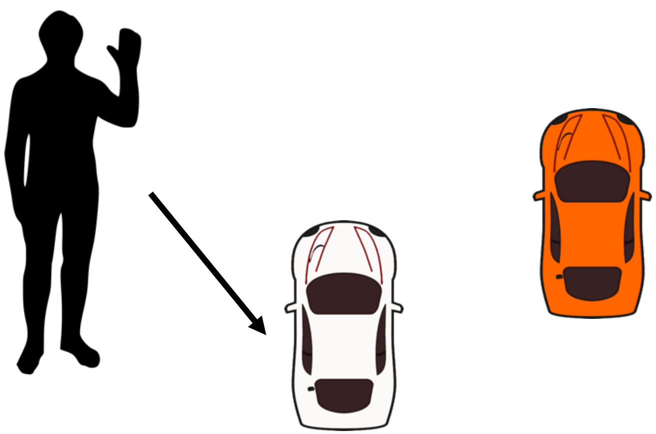


*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*

# HRI as an underactuated system

$$\xi_H^*(\xi_R) = \arg\max_{\xi_H} U_H(\xi_H, \xi_R)$$

$$\xi_R^* = \arg\max_{\xi_R} U_R(\xi_R, \xi_H^*(\xi_R))$$



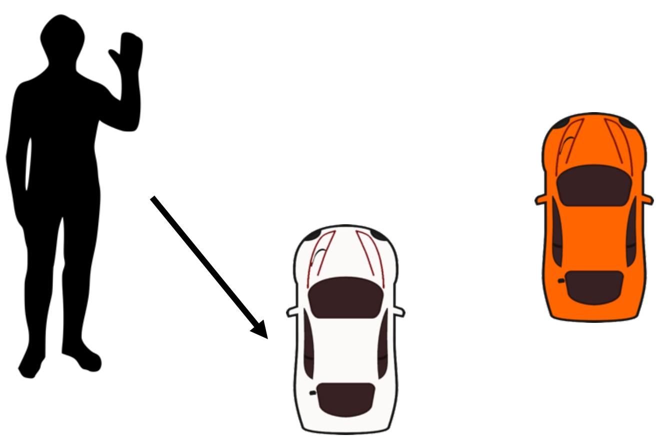
*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*

# HRI as an underactuated system

$$\xi_H^*(\xi_R) = \arg\max U_H(\xi_H, \xi_R)$$

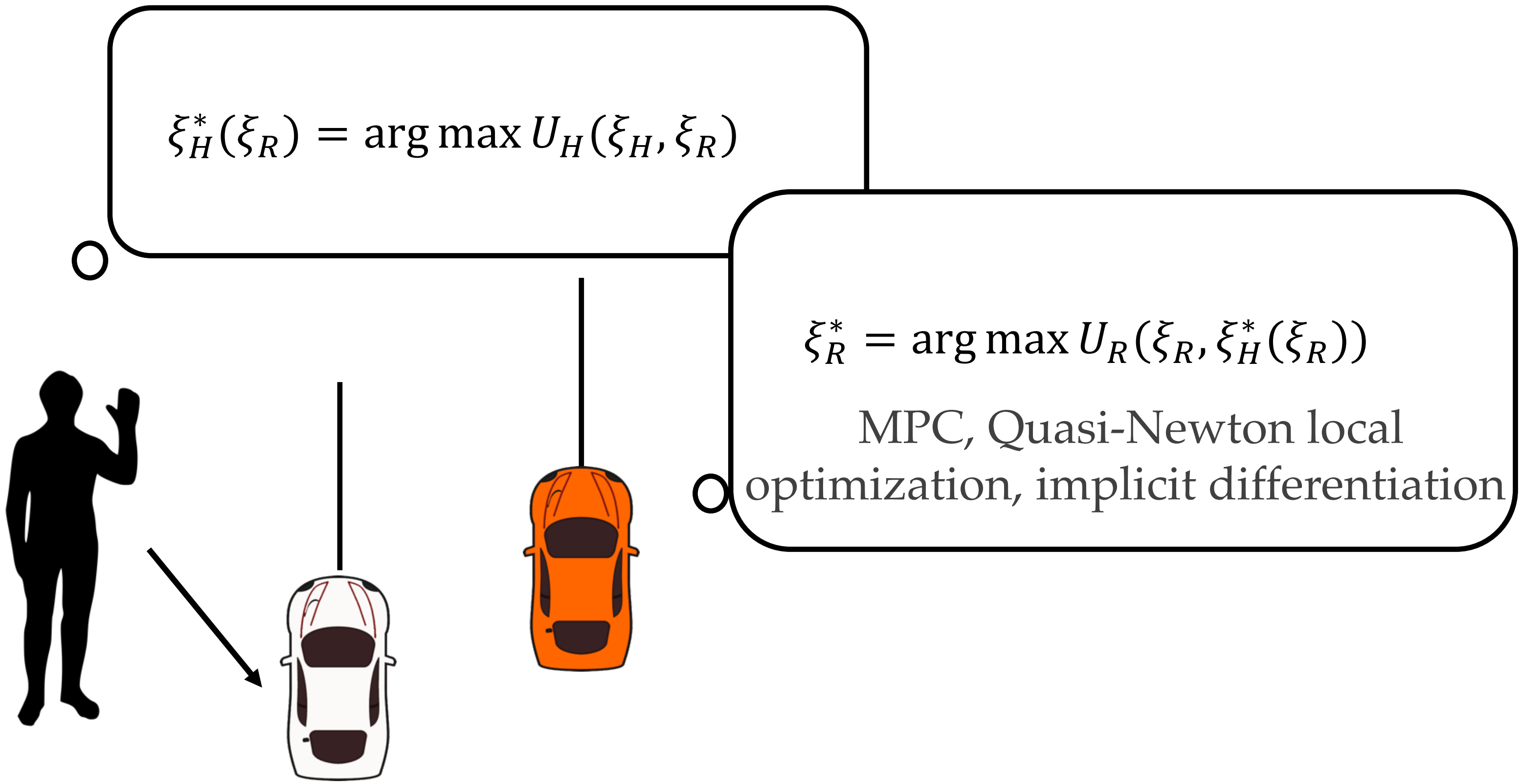
$$\xi_R^* = \arg\max U_R(\xi_R, \xi_H^*(\xi_R))$$

MPC, Quasi-Newton local optimization, implicit differentiation



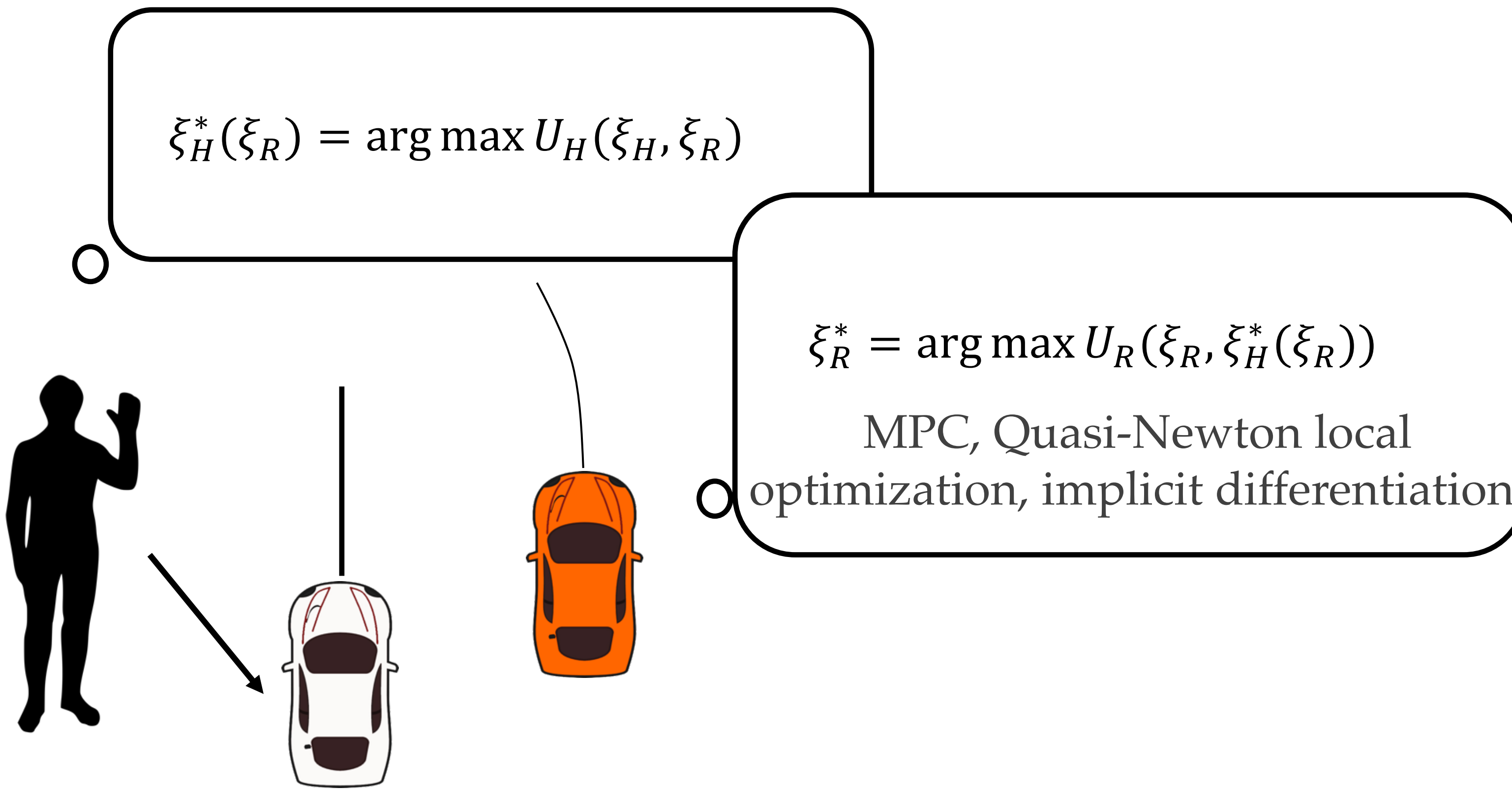
*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*

# HRI as an underactuated system



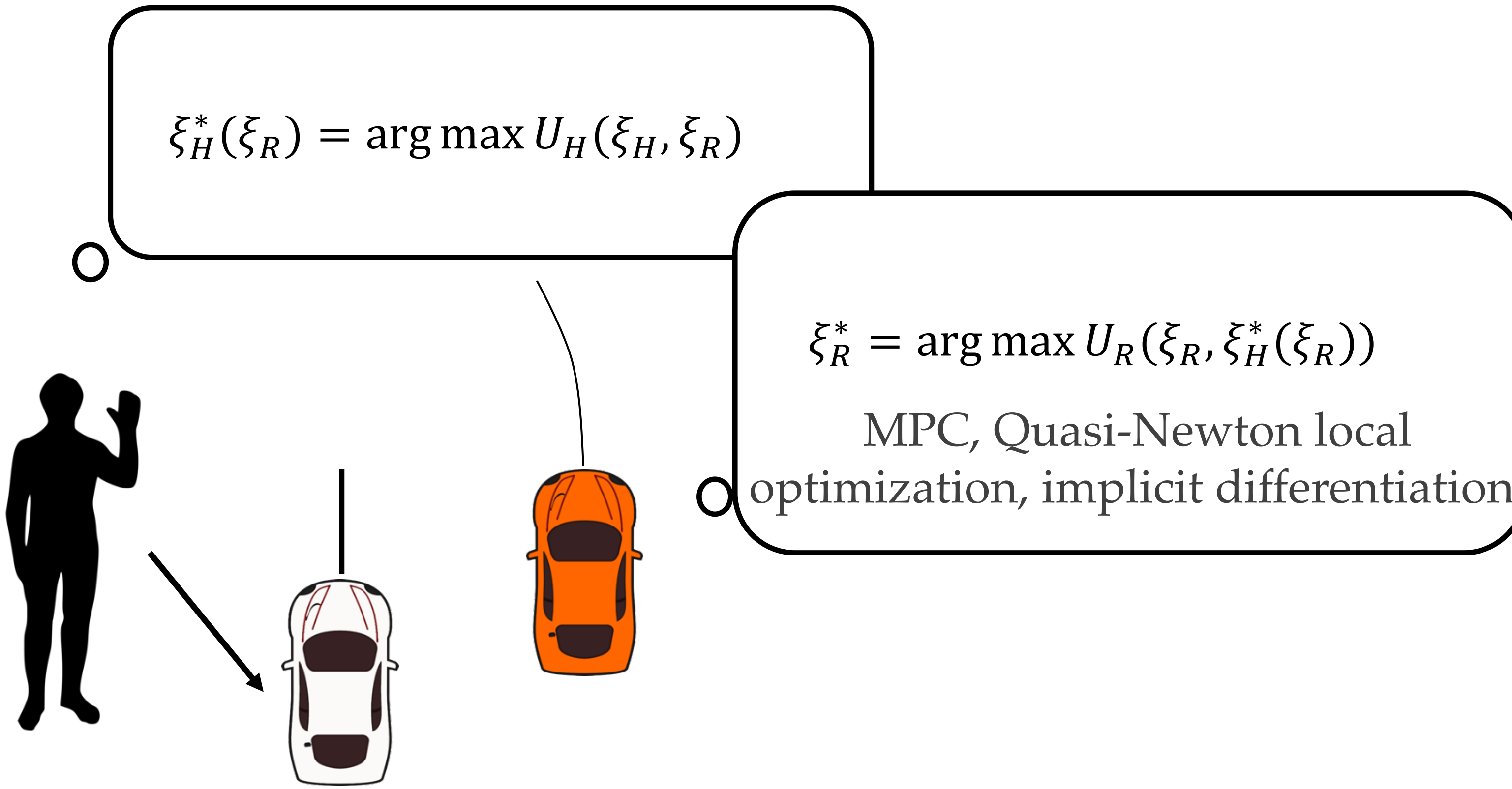

*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*

# HRI as an underactuated system




*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*

# HRI as an underactuated system



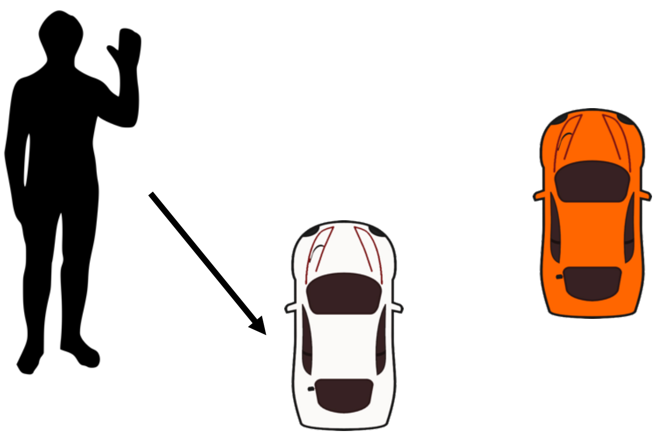

*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*

# HRI as an underactuated system

hidden

$$\xi_H^*(\xi_R;\theta) = \arg\max U_H(\xi_H,\xi_R;\theta)$$

$$\xi_R^* = \arg\max U_R(\xi_R,\xi_H^*(\xi_R;\theta))$$



*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*

# HRI as an underactuated system

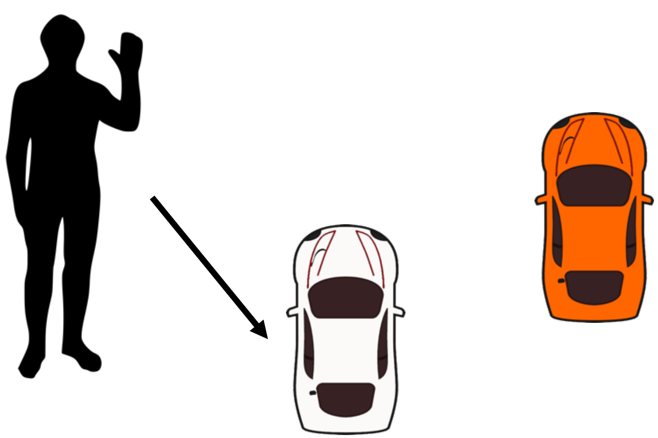
hidden

$$\xi_H^*(\xi_R;\theta) = \arg\max U_H(\xi_H,\xi_R;\theta)$$

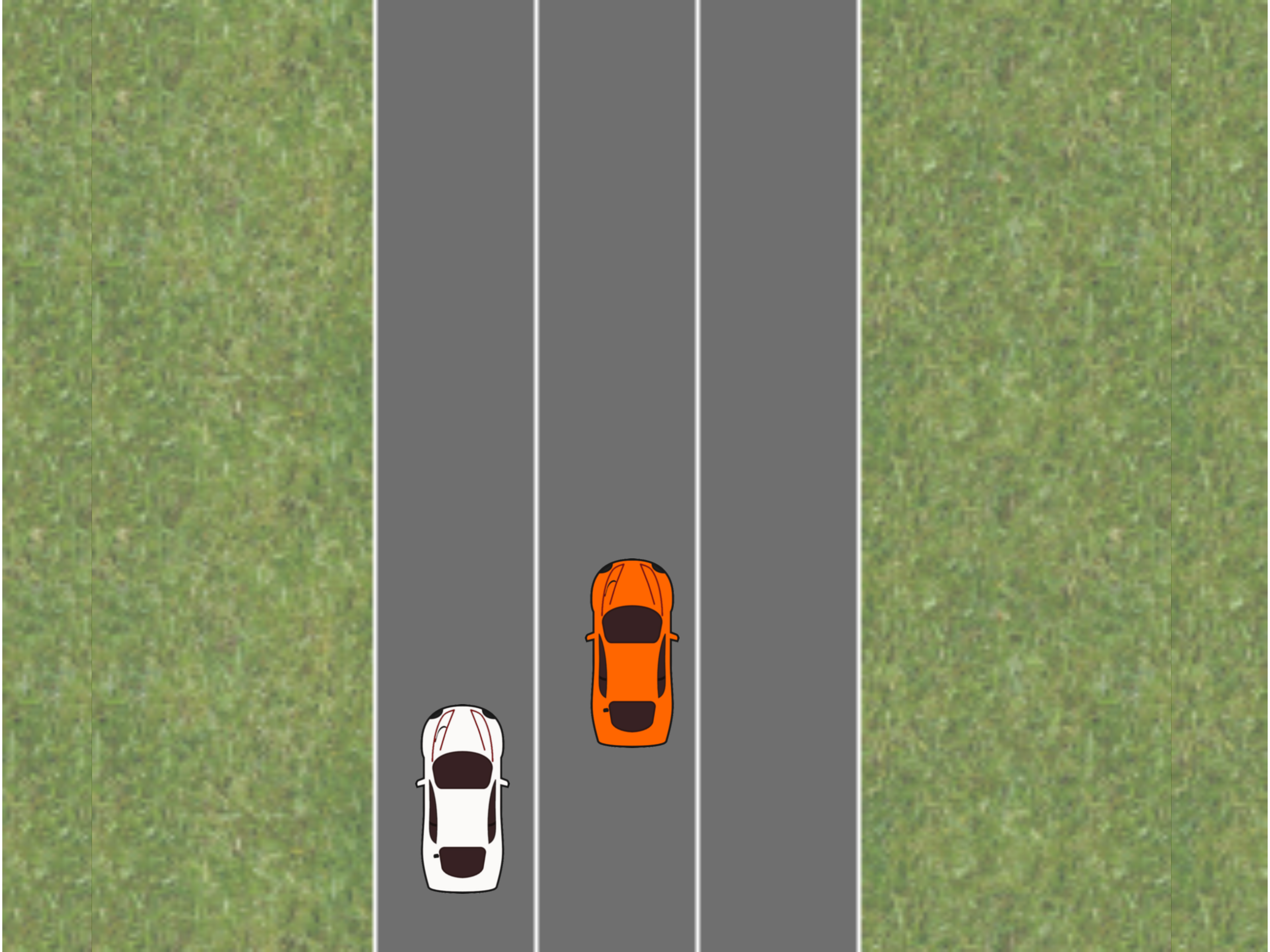
$$\theta^* = \arg\max P(\xi_H^D|\xi_R,\theta)$$ offline IRL

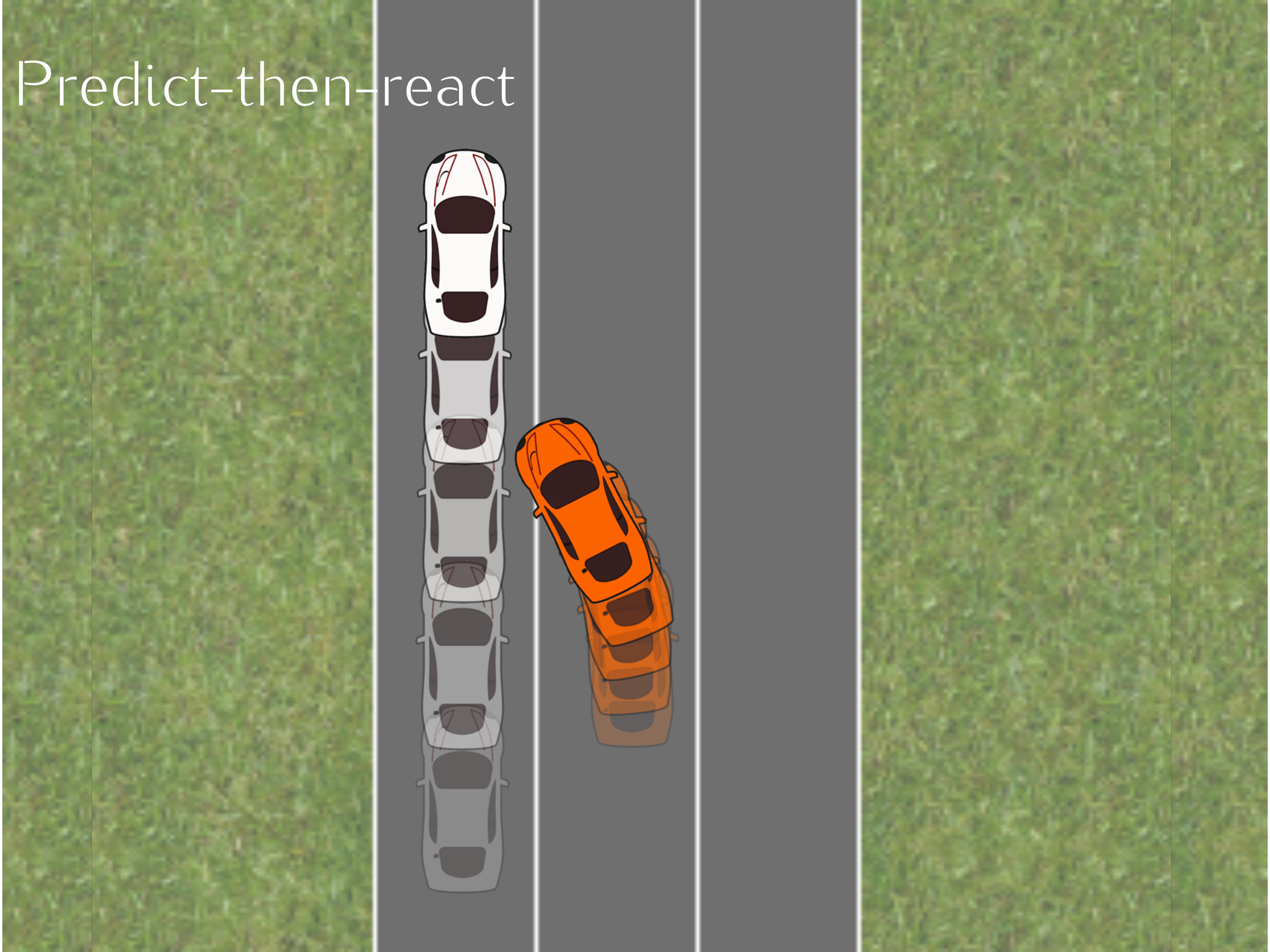
$$P(\xi_H^D|\xi_R,\theta) \propto e^{U_H(\xi_H^D,\xi_R;\theta)}$$

$$\xi_R^* = \arg\max U_R(\xi_R,\xi_H^*(\xi_R;\theta))$$



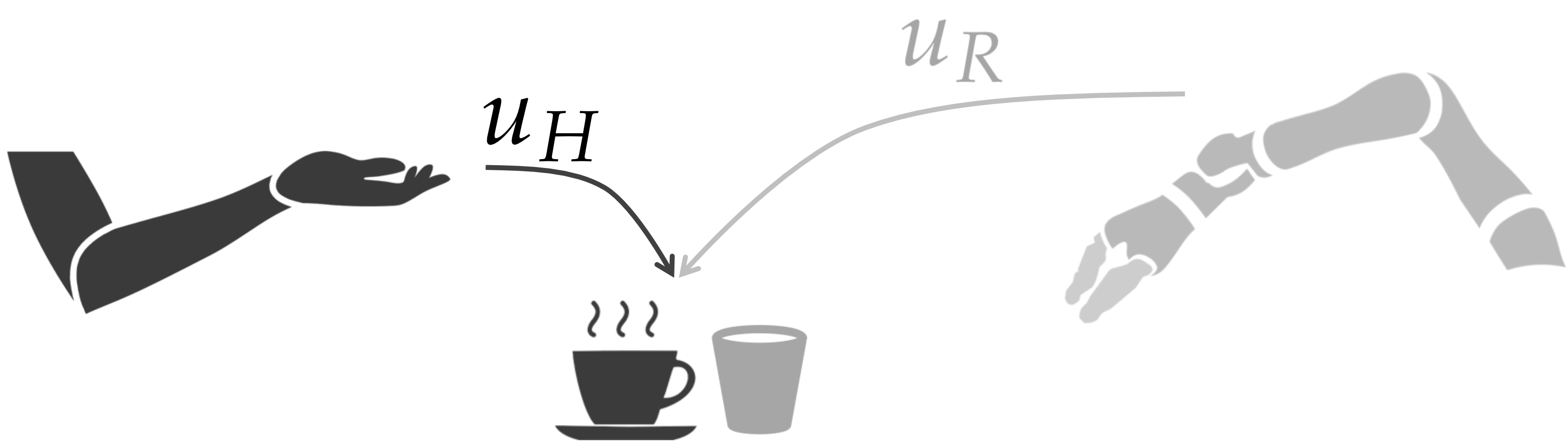
*Planning for Autonomous Cars that Leverage Effects on Human Actions [RSS'16]*




Predict-then-react


Underactuated System, $U_H$ Learned Offline

$$\max_{\xi_H} U_H(\xi_H)$$



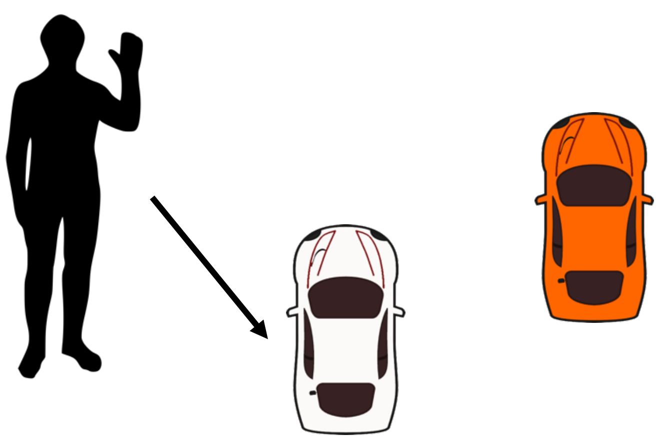

People do not act in isolation.


SPEED
SOCIETY
2015/02/06 23:09:54

# Adapting to the individual driver

$$\xi_H^*(\xi_R;\theta) = \arg\max U_H(\xi_H, \xi_R; \theta)$$

$$\xi_R^* = \arg\max U_R(\xi_R, \xi_H^*(\xi_R; \hat{\theta}))$$



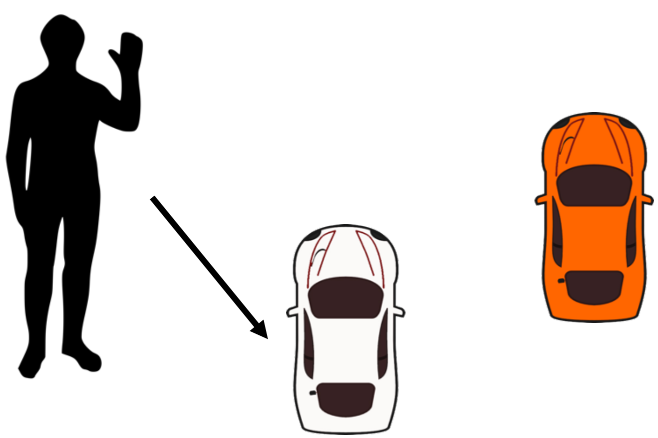
*Information Gathering Actions over Human Internal State [IROS'16]*

# Adapting to the individual driver

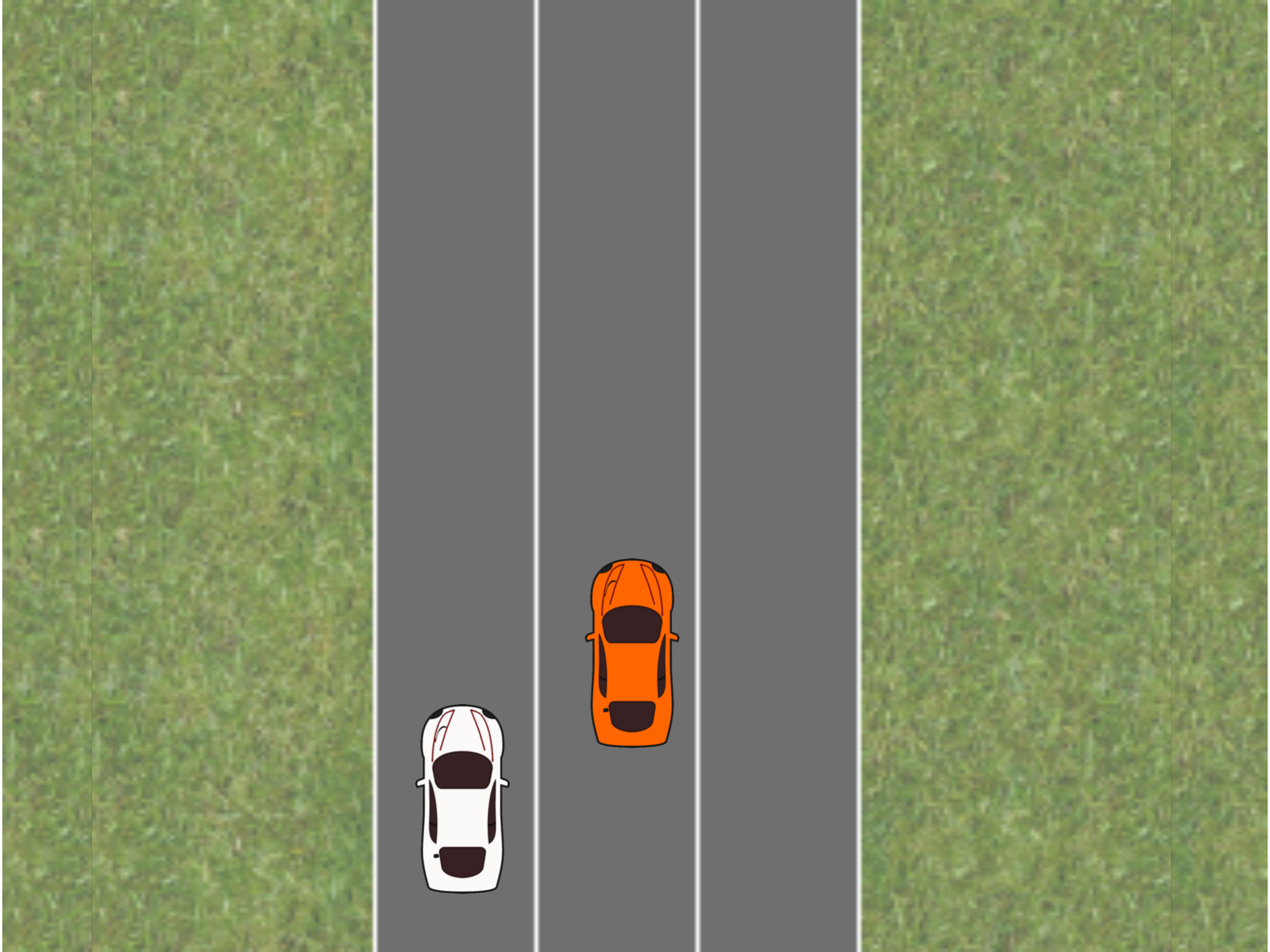
$$\xi_H^*(\xi_R; \theta) = \arg\max U_H(\xi_H, \xi_R; \theta)$$

$$\hat{\theta} = \arg\max b(\theta)$$
$$\xi_R^* = \arg\max U_R(\xi_R, \xi_H^*(\xi_R; \hat{\theta}))$$

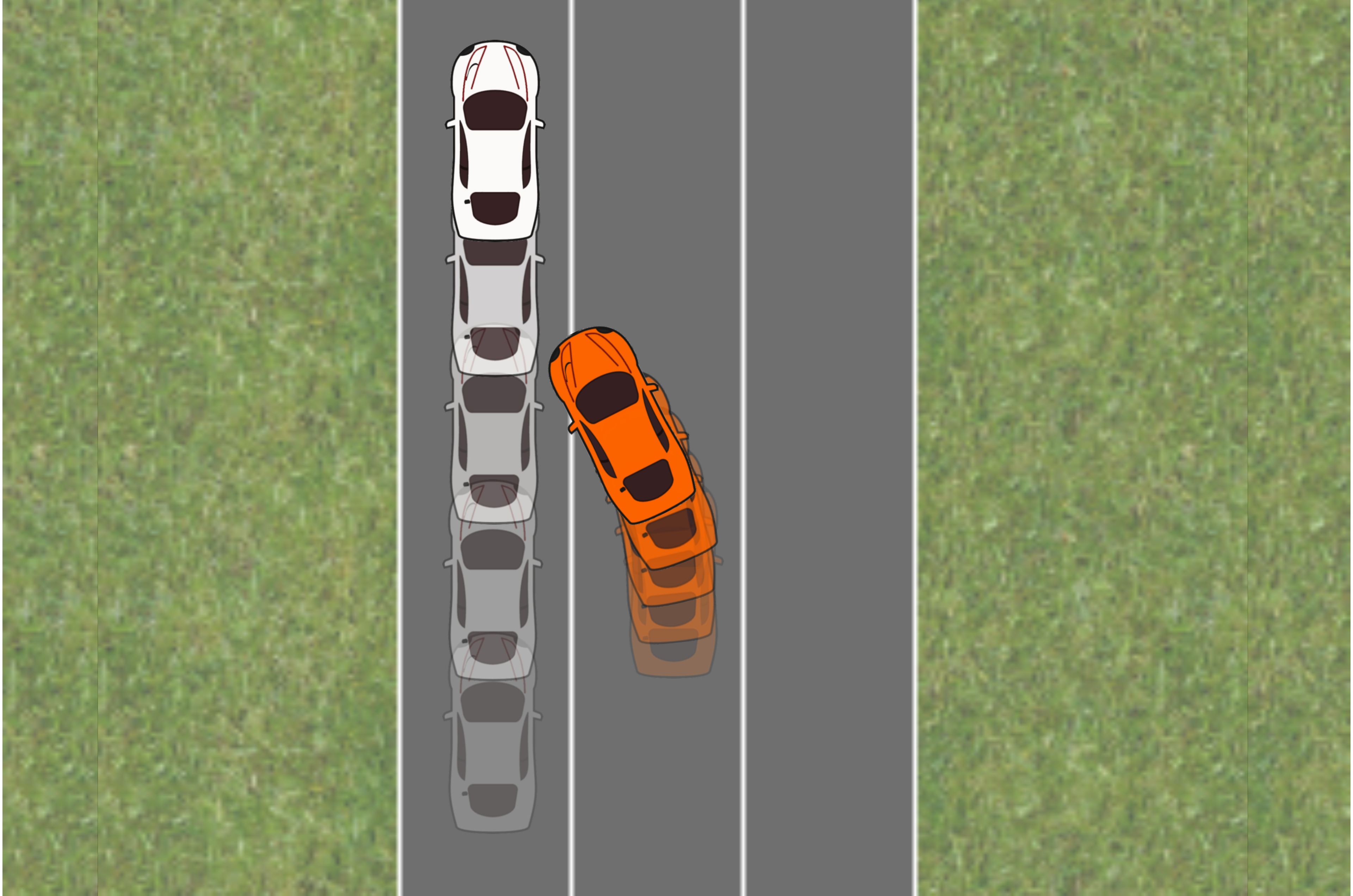
$$b'(\theta) \propto P(u_H | x, u_R, \theta) b(\theta)$$



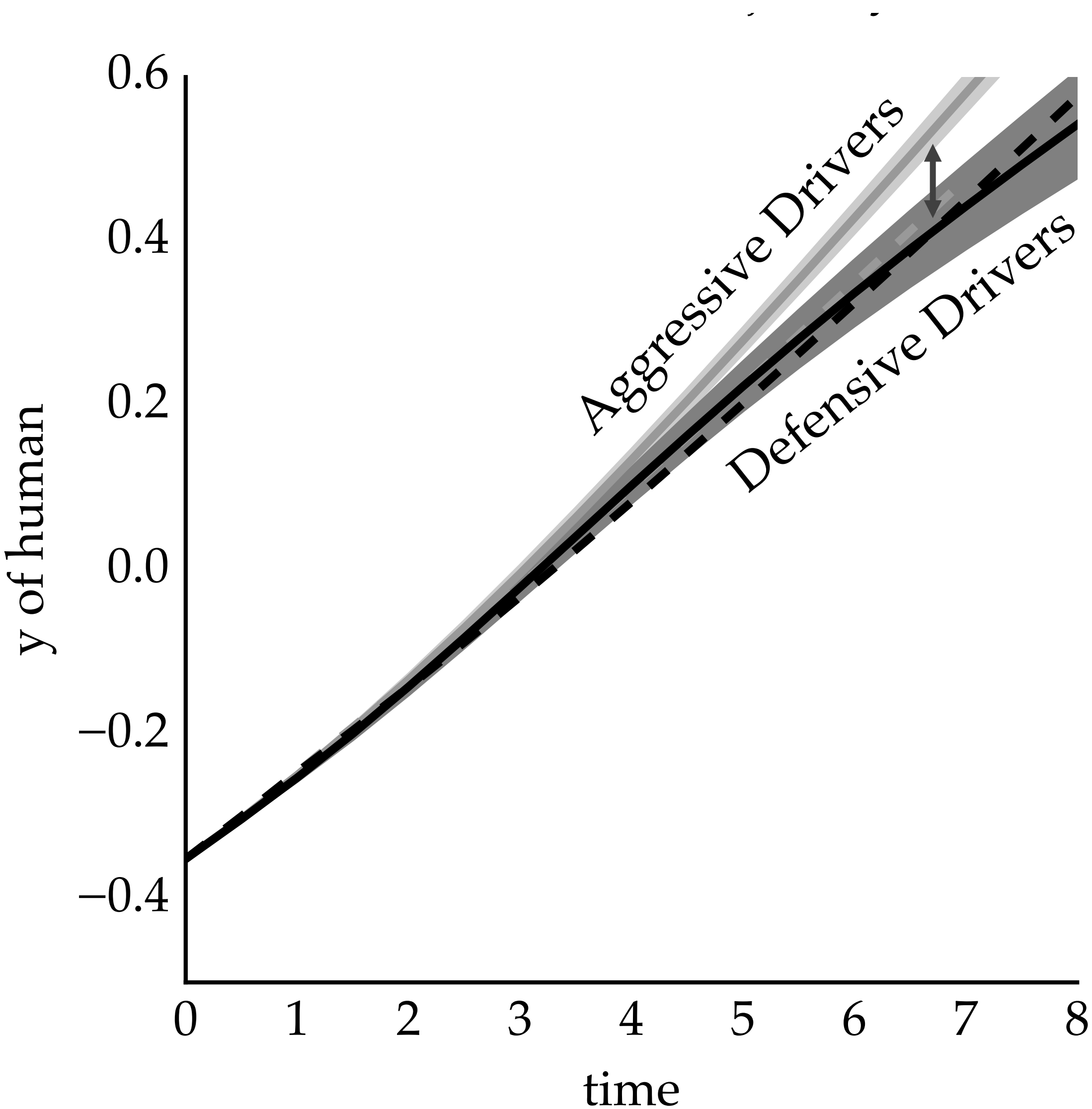
*Information Gathering Actions over Human Internal State [IROS'16]*



# Back to ultra-defensive

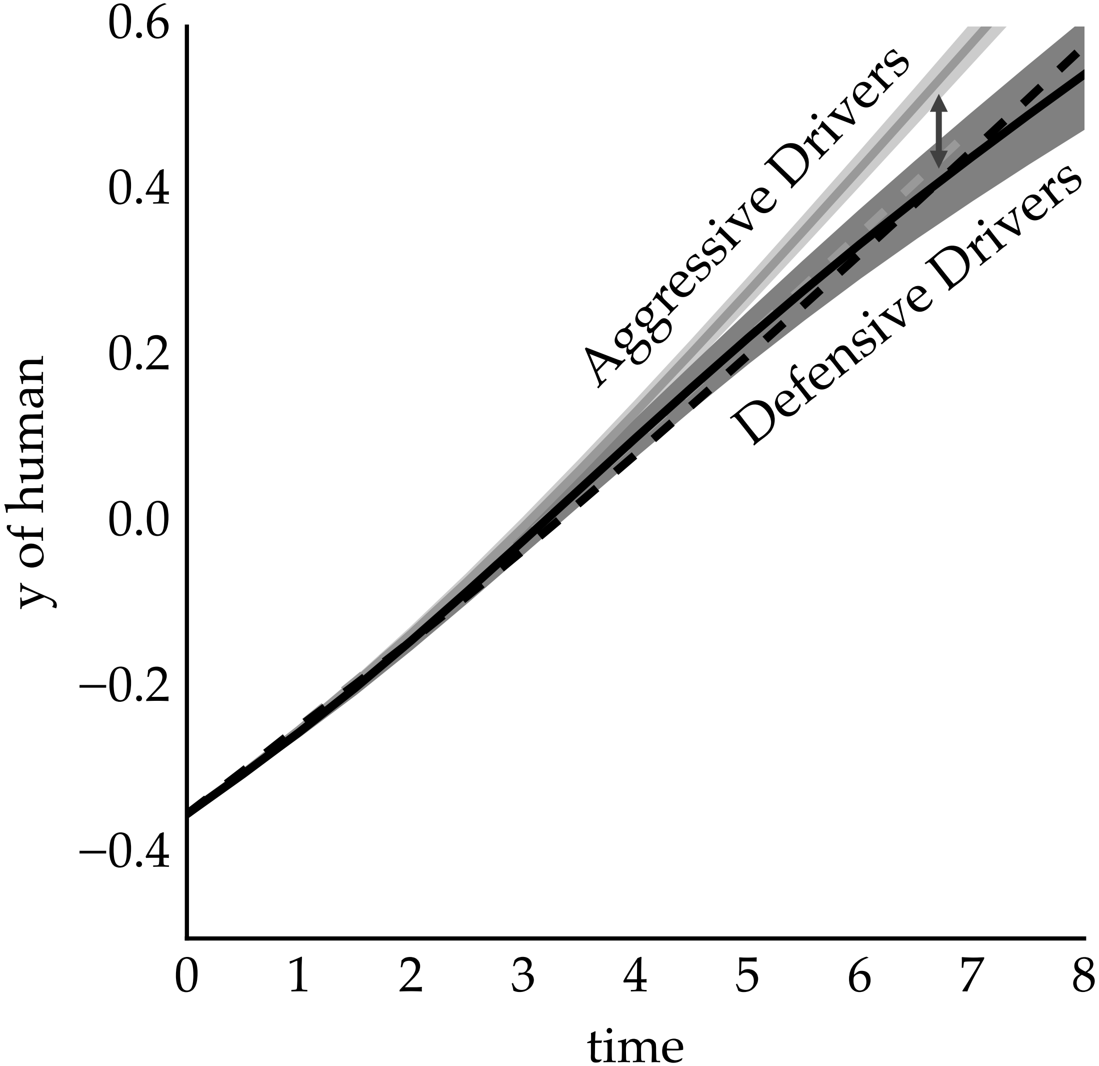


# All Users Drive in Almost the Same Way

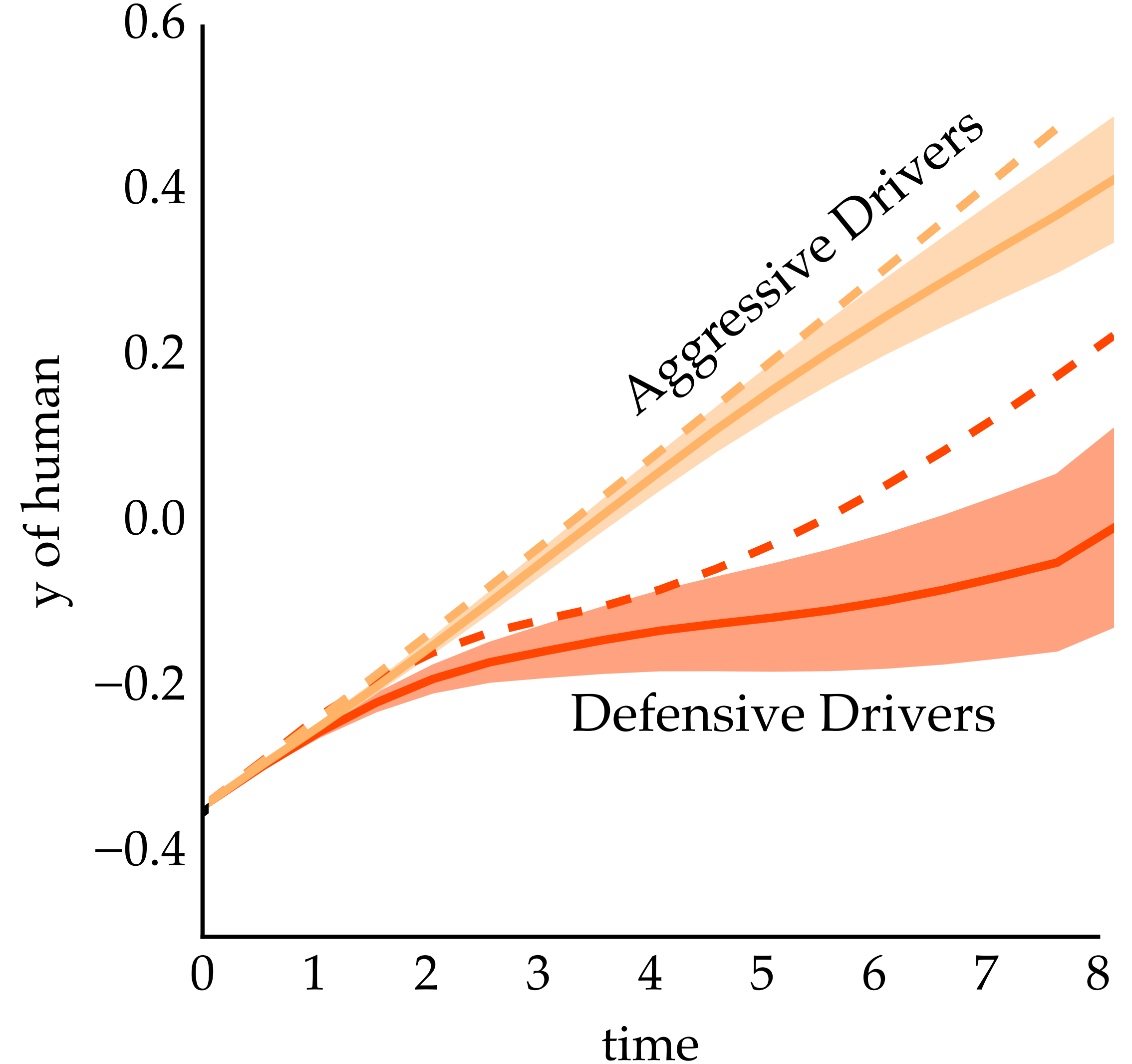



gifsboom
.net

# All Users Drive in Almost the Same Way



# Idea: Leverage the robot's actions!

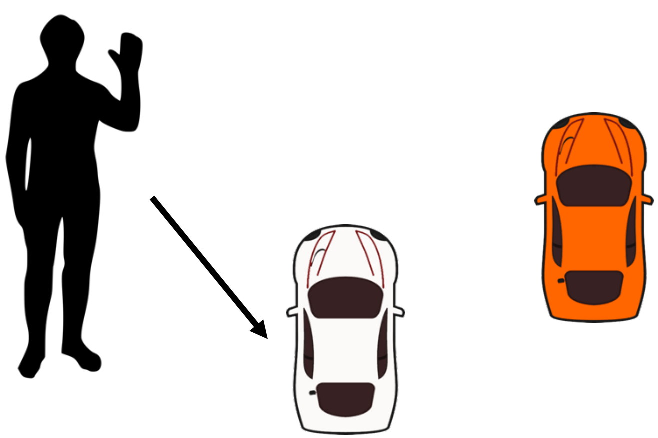


# Adapting to the individual driver

$$\xi_H^*(\xi_R;\theta) = \arg\max U_H(\xi_H, \xi_R; \theta)$$

$$\hat{\theta} = \arg\max b(\theta)$$
$$\xi_R^* = \arg\max U_R(\xi_R, \xi_H^*(\xi_R; \hat{\theta}))$$

$$b'(\theta) \propto P(u_H | x, u_R, \theta) b(\theta)$$



*Information Gathering Actions over Human Internal State [IROS'16]*

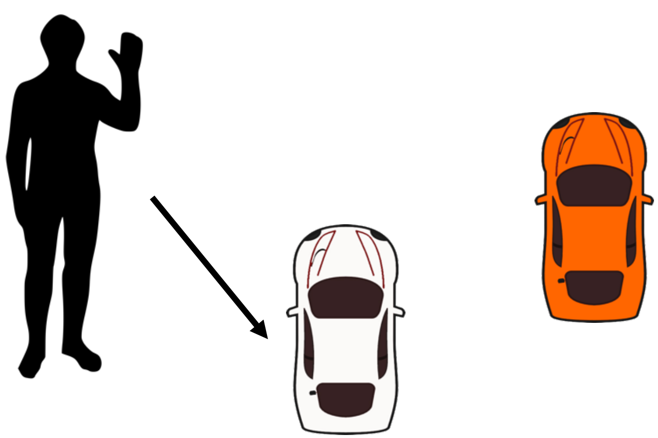
# Actively estimating driver style

$$\xi_H^*(\xi_R;\theta) = \arg\max U_H(\xi_H,\xi_R;\theta)$$

$$\hat{\theta} = \arg\max b(\theta)$$

$$\xi_R^* = \arg\max U_R(\xi_R,\xi_H^*(\xi_R;\hat{\theta})) + \lambda\mathbb{E}[H(b) - H(b')]$$

$$b'(\theta) \propto P(u_H|x,u_R,\theta)b(\theta)$$

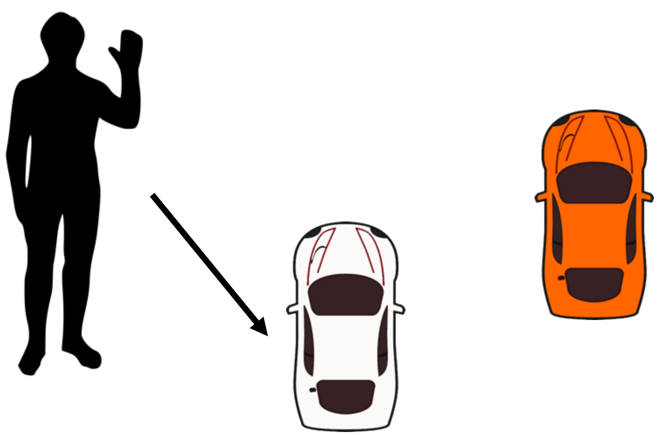


*Information Gathering Actions over Human Internal State [IROS'16]*

# HRI as an underactuated system

$$\xi_H^*(\xi_R;\theta) = \arg\max U_H(\xi_H, \xi_R; \theta)$$

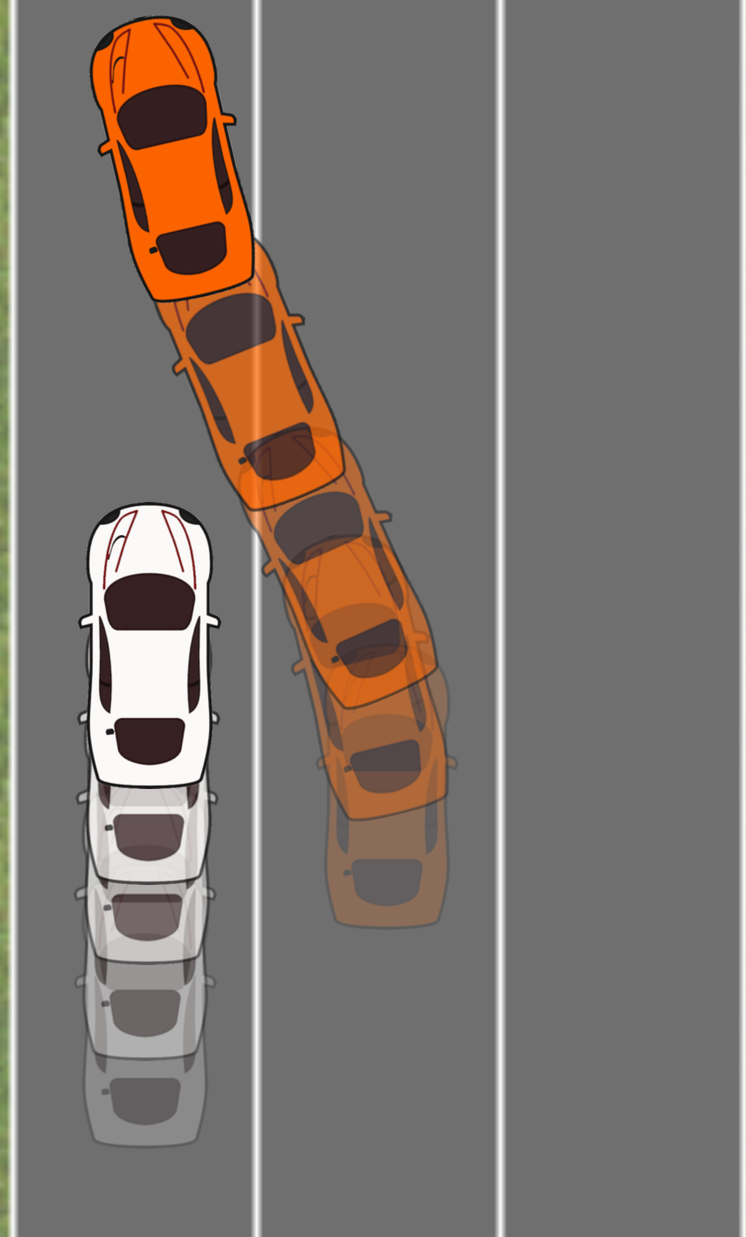
$$\xi_R^* = \arg\max U_R(\xi_R, \xi_H^*(\xi_R;\theta))$$



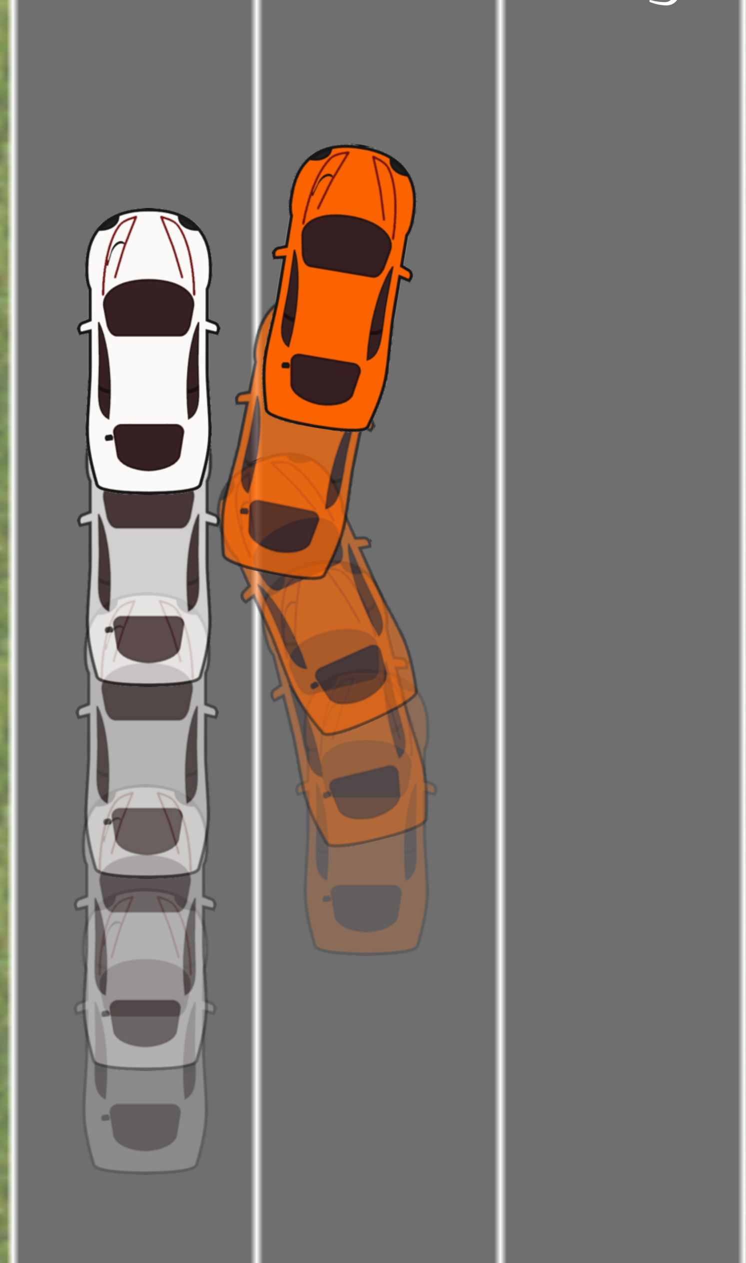
*Information Gathering Actions over Human Internal State [IROS'16]*

# Estimating Human Driver Style Online

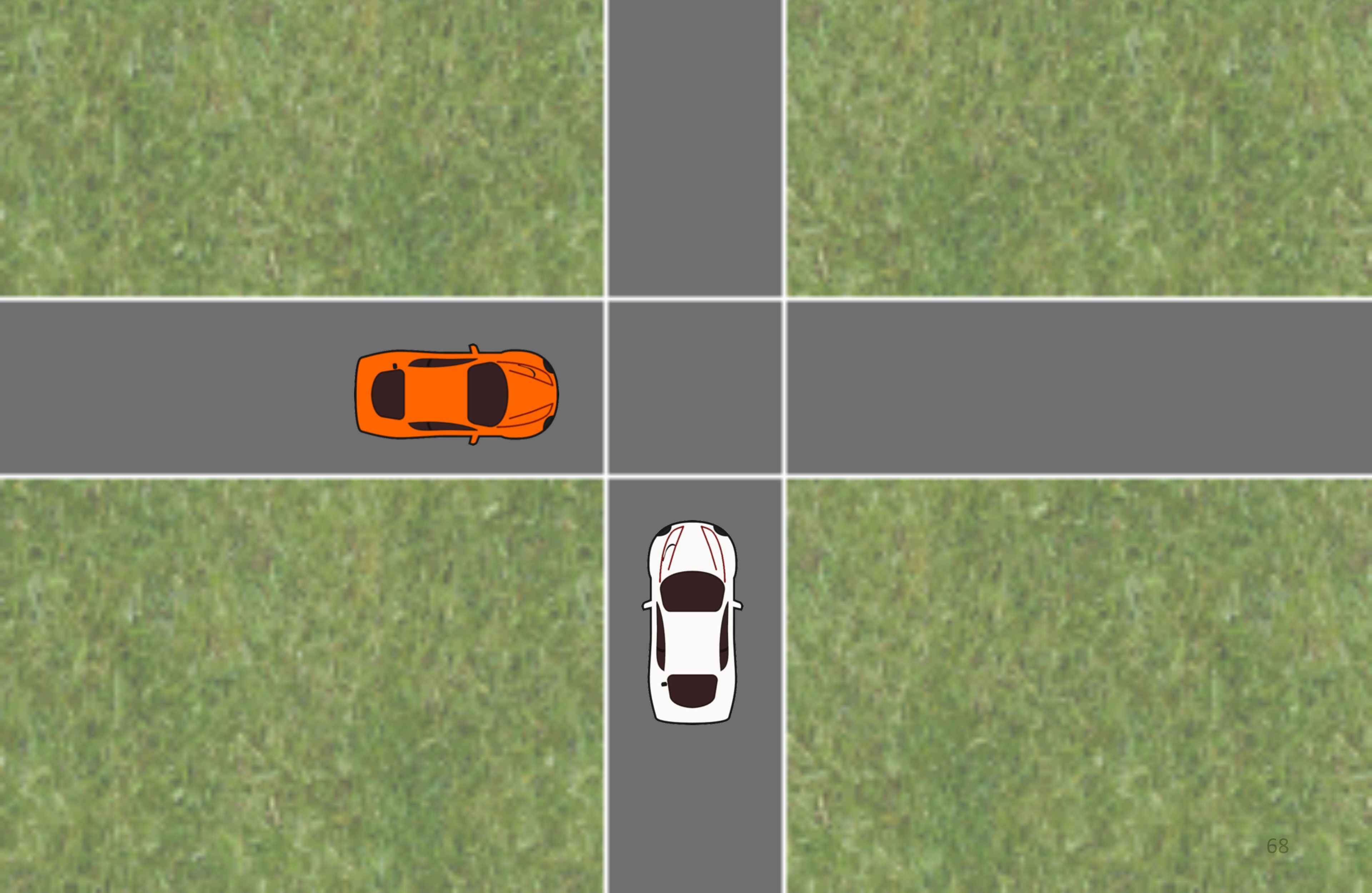
# Estimating Human Driver Style Online



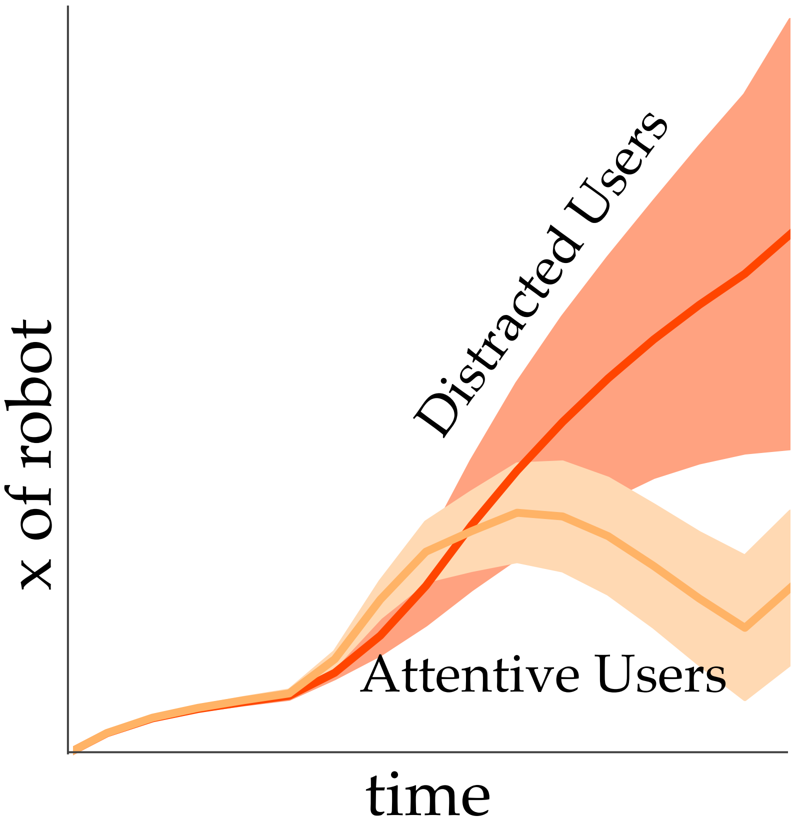
# Estimating Human Driver Style Online

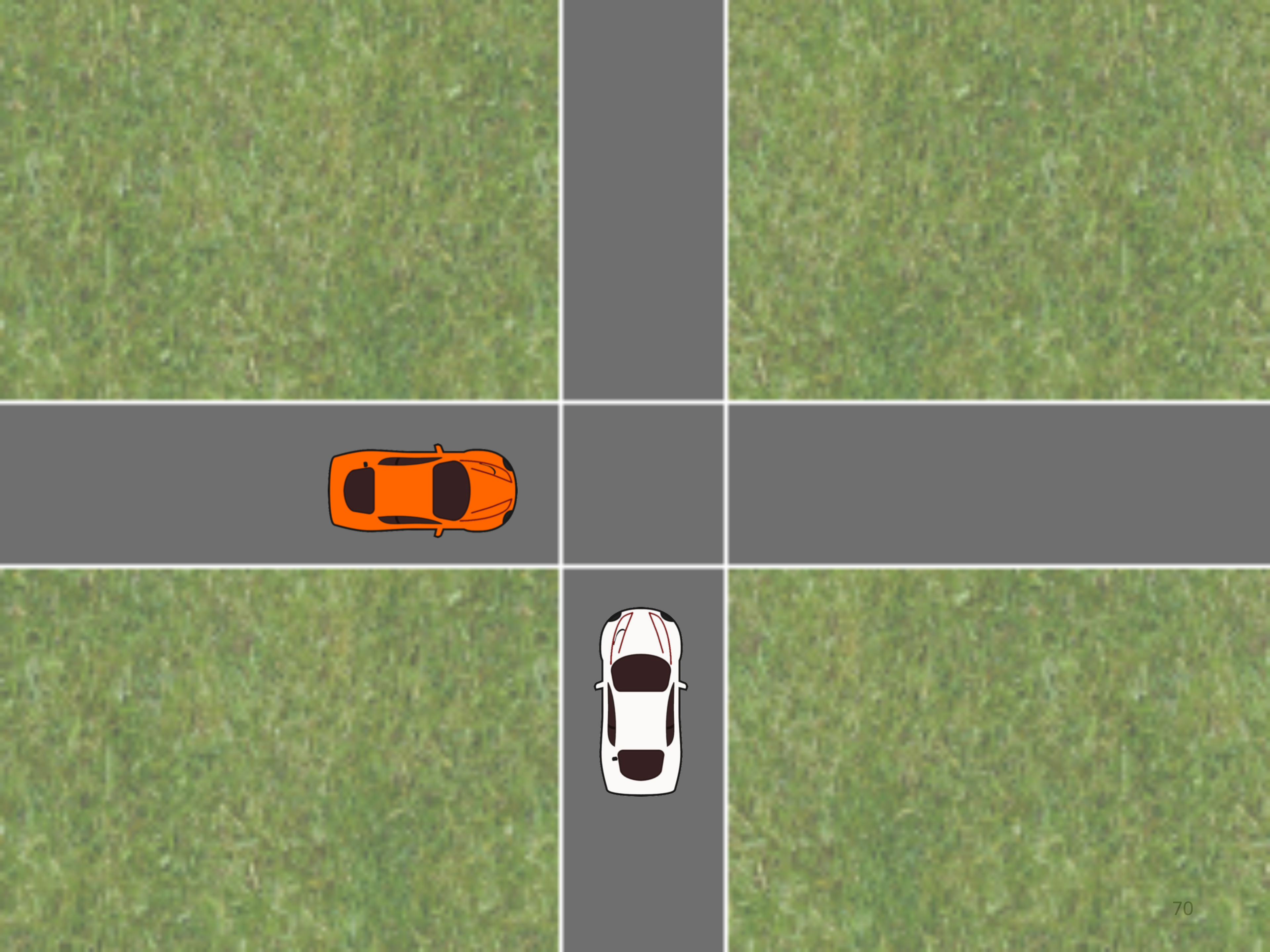


# Coordination at 4-Way Stops

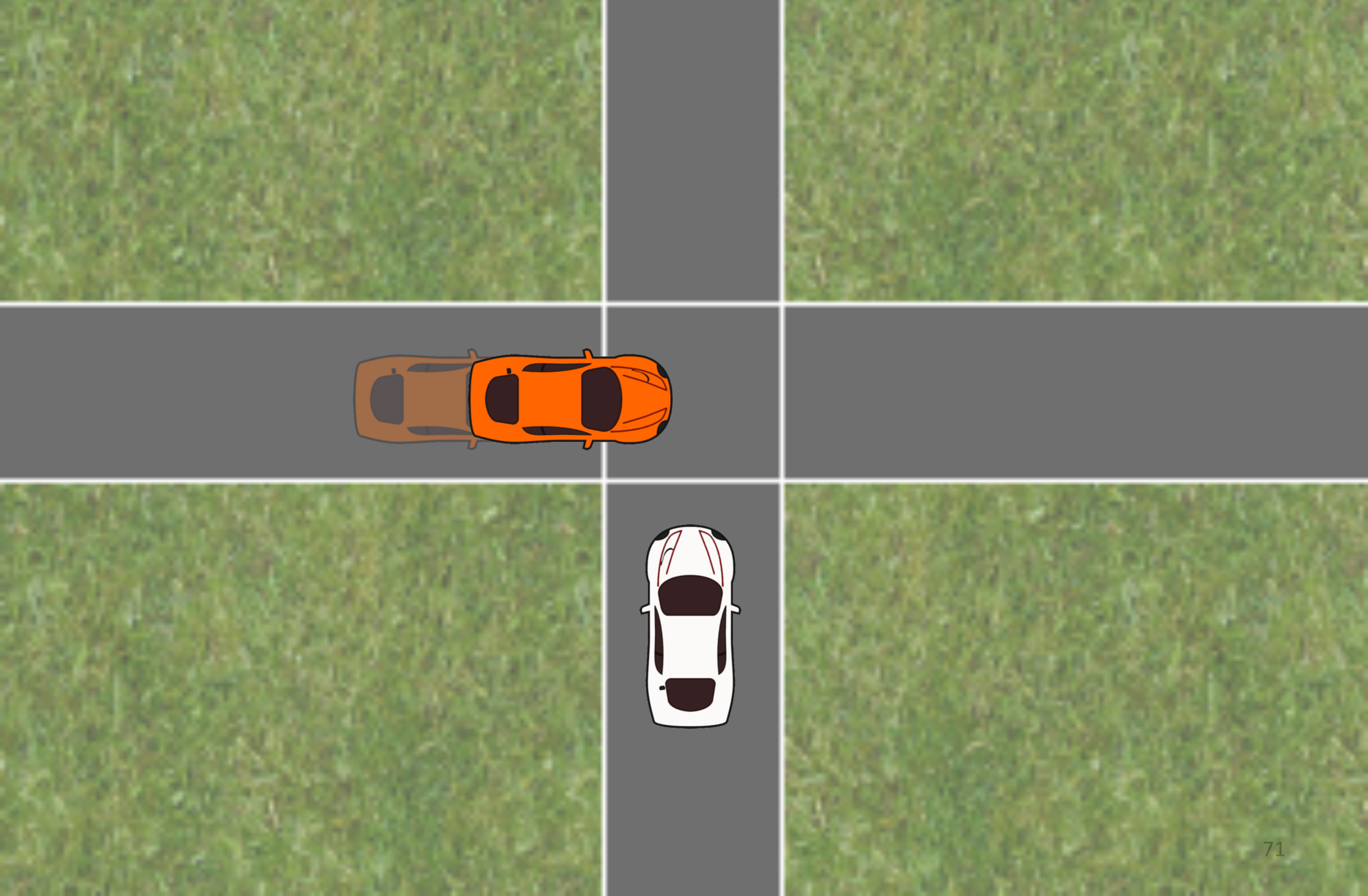


# Robot Trajectories

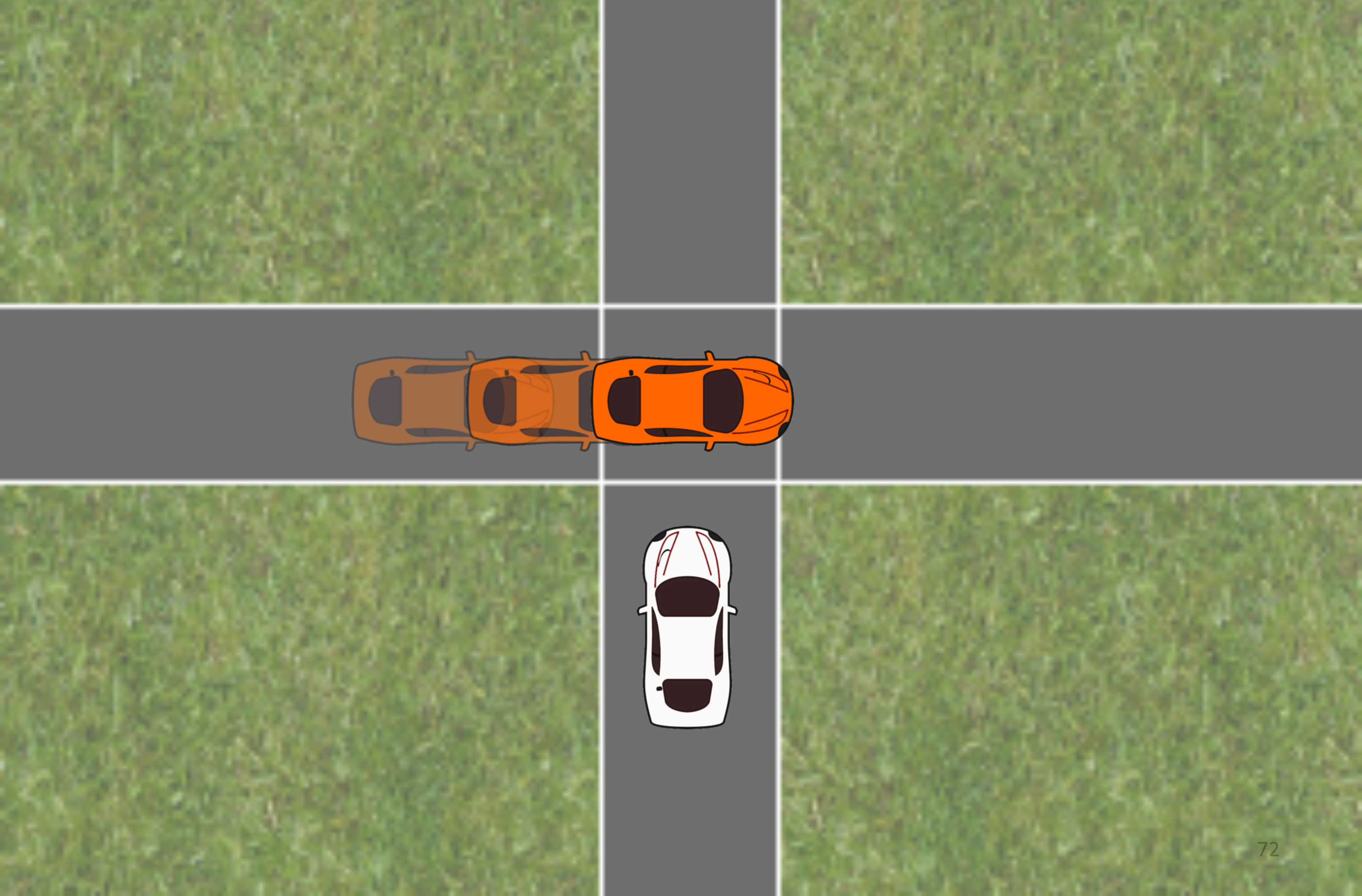




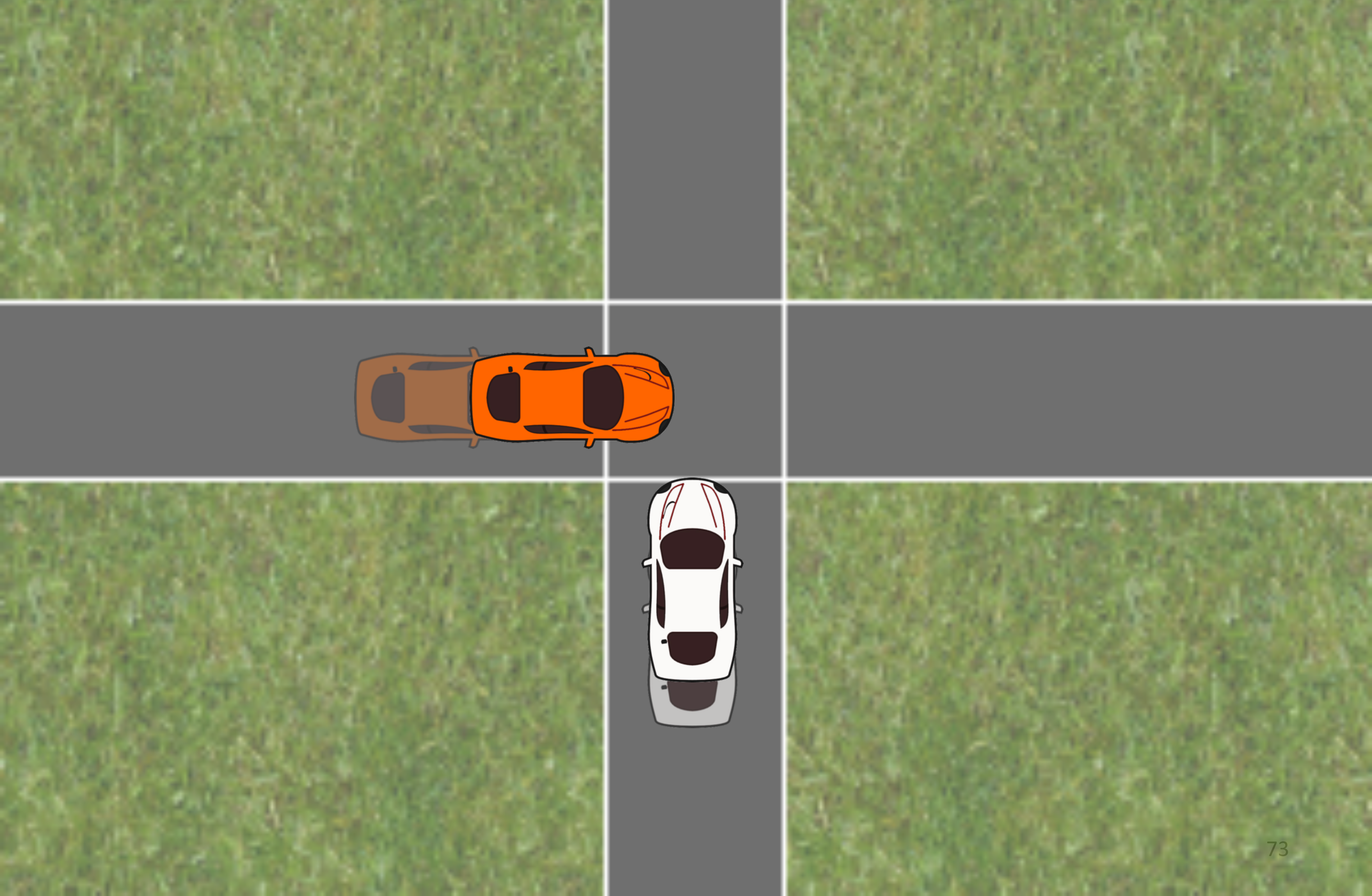
# Inch Forward



# Attentive Users: Continue

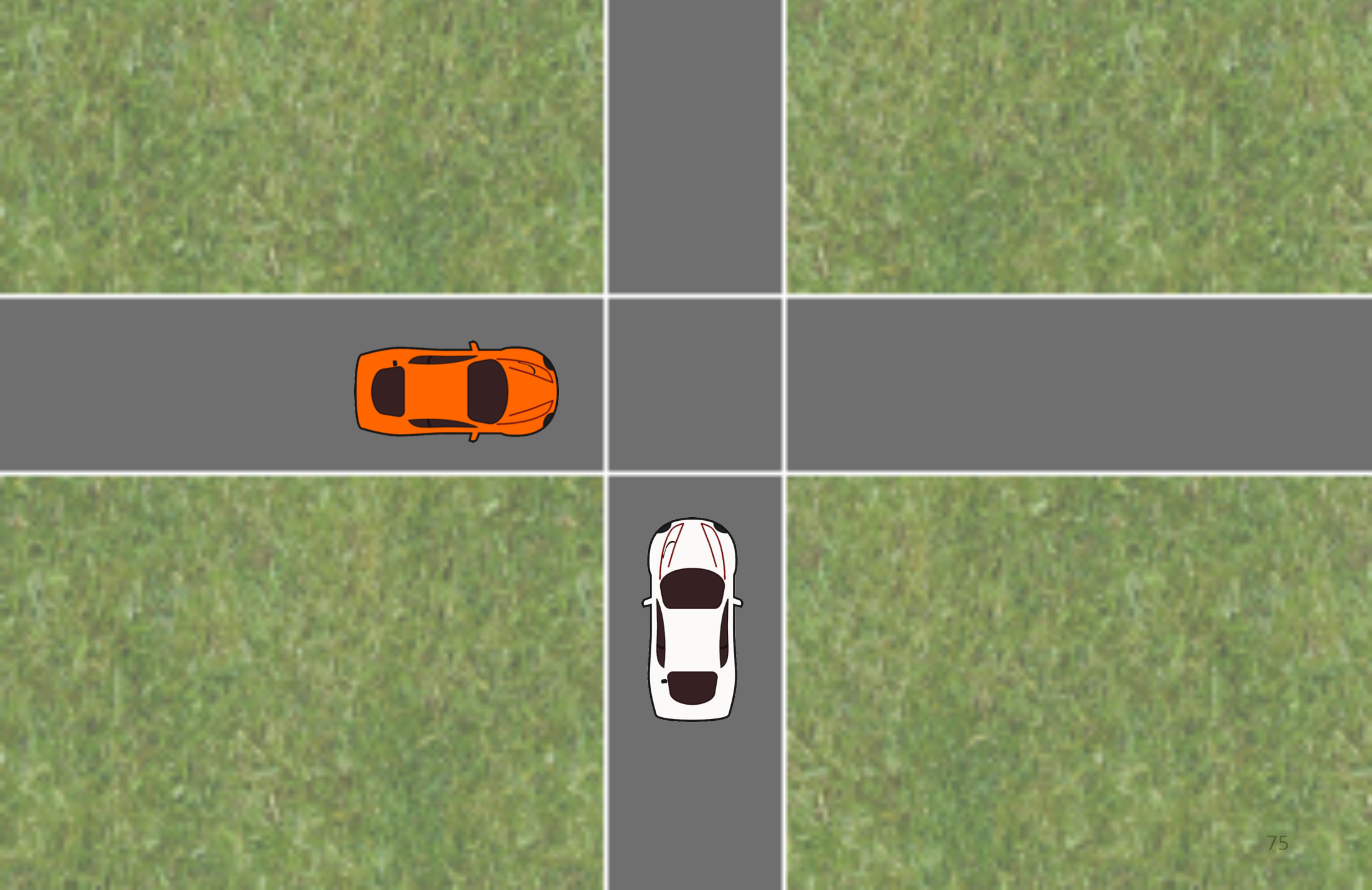


# Inch Forward

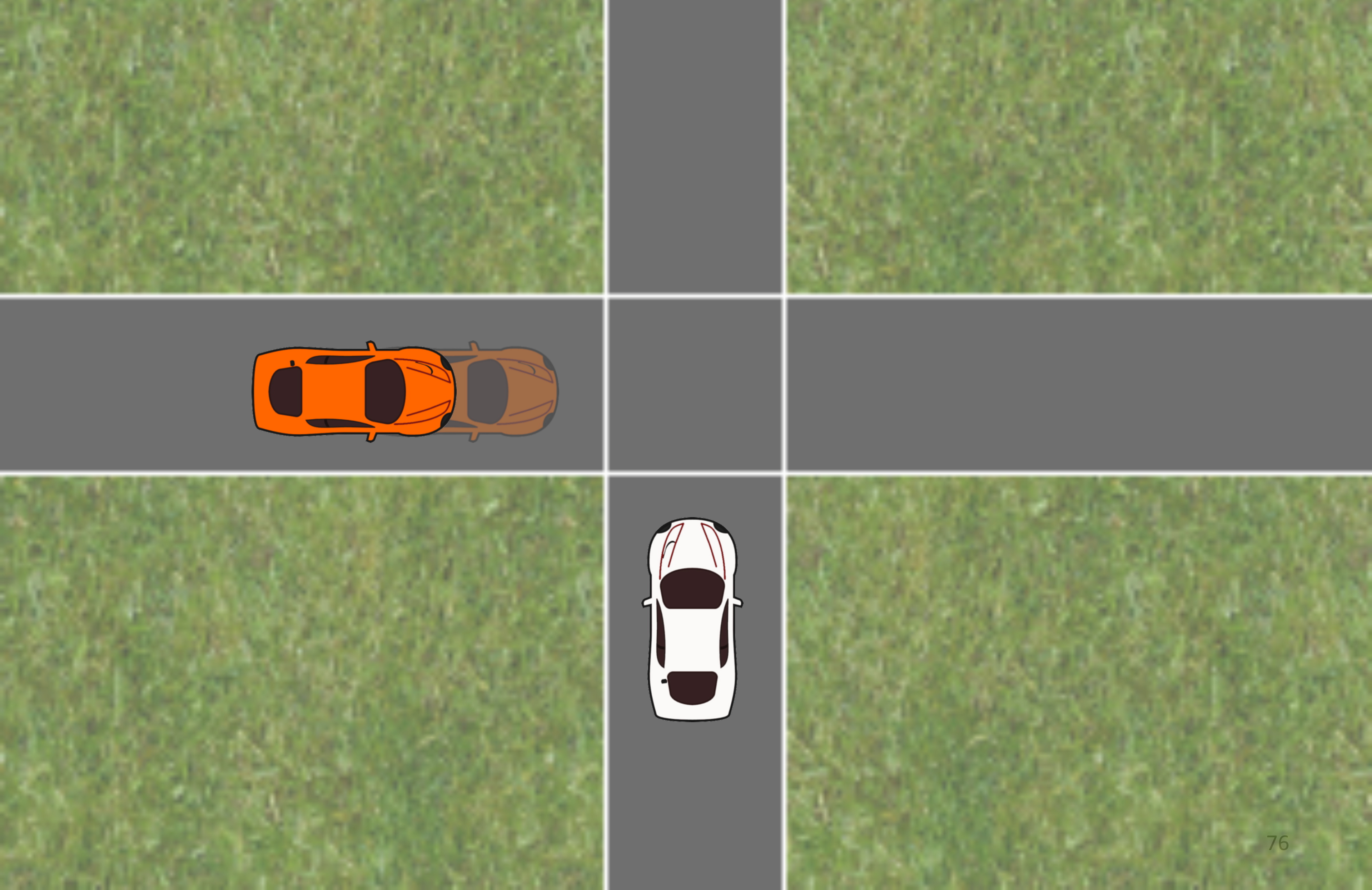


# Distracted Users: Go Back

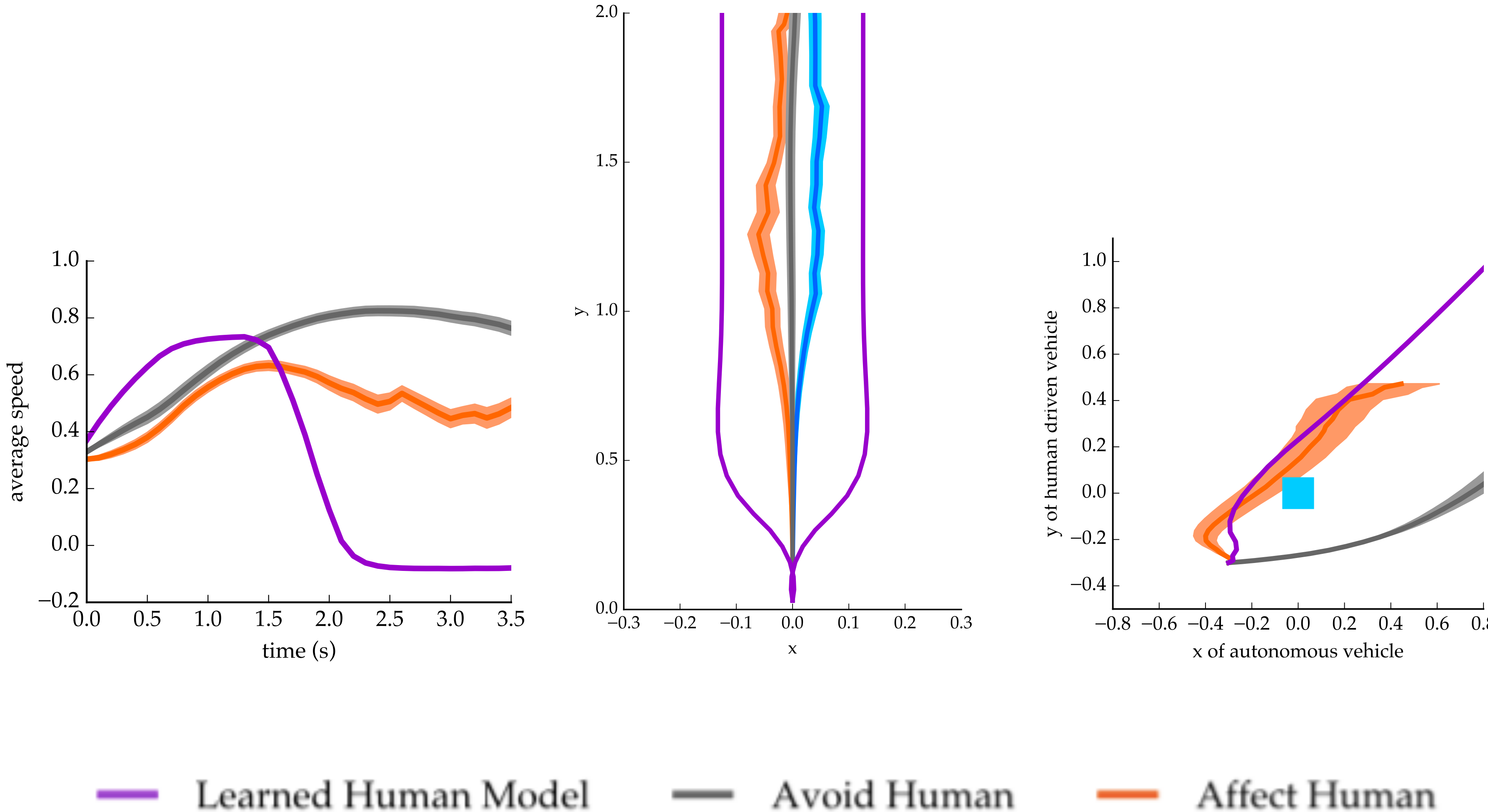
# $U_R$: Human Should Go First

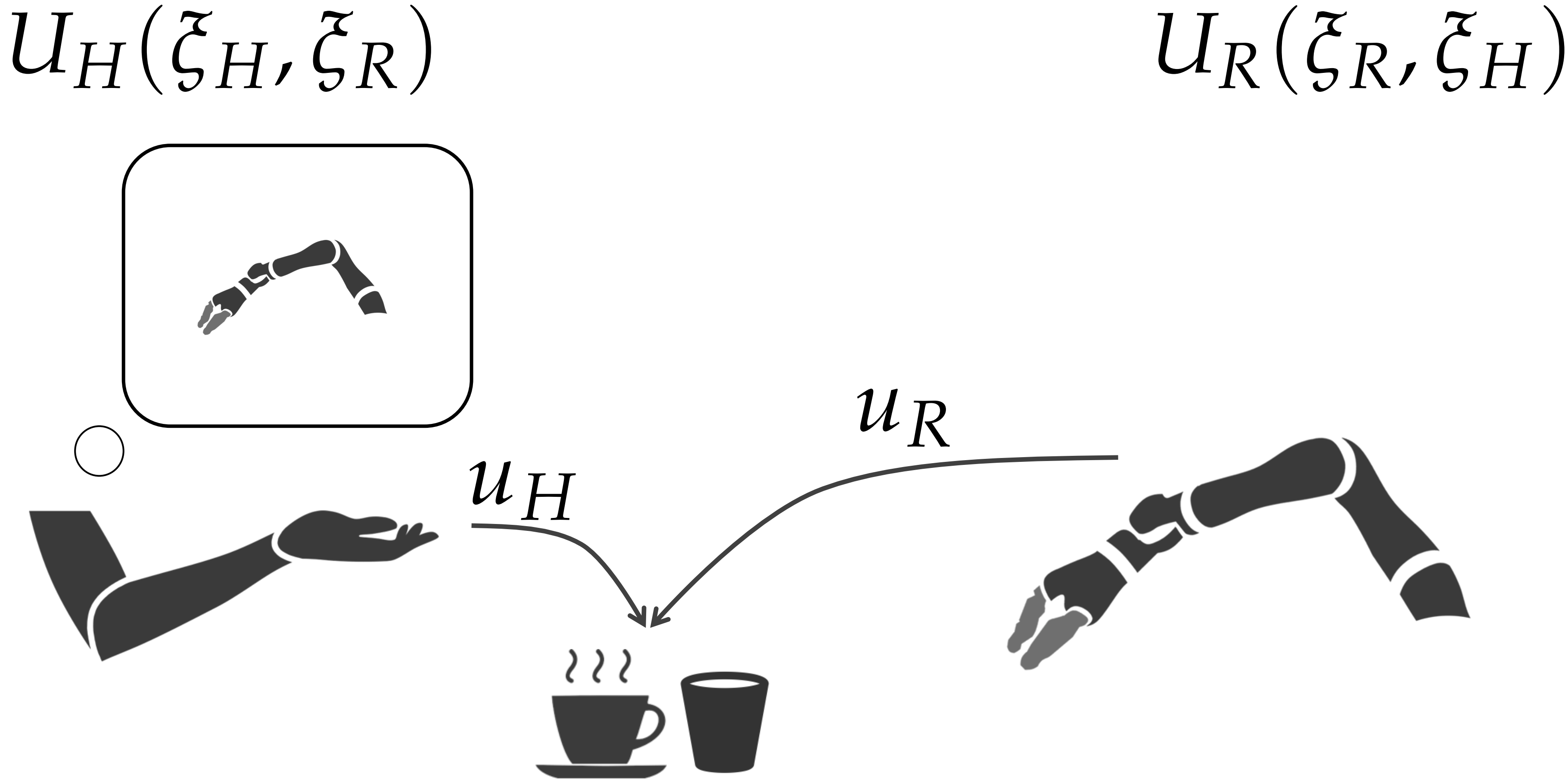


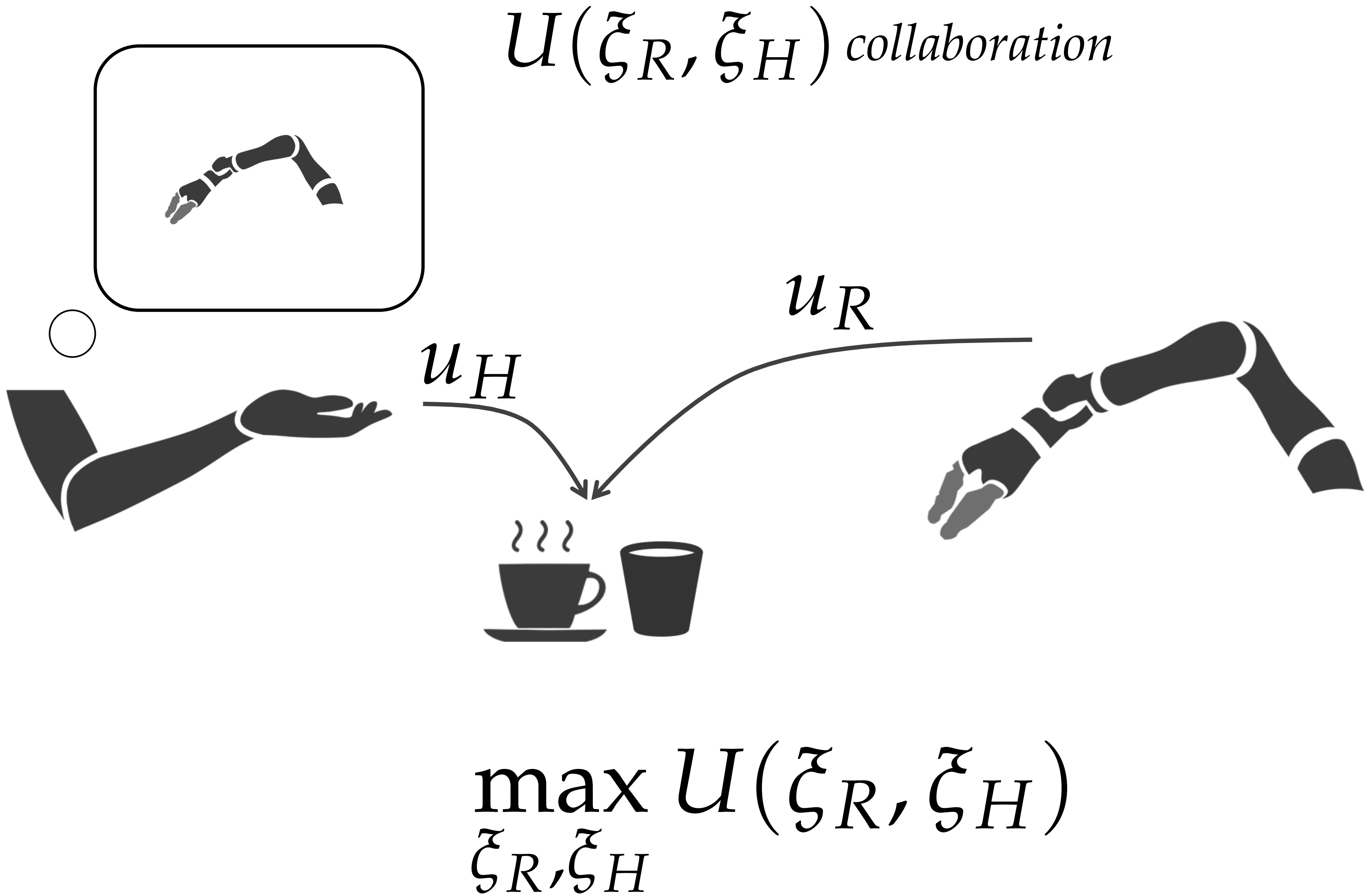
# $U_R$: Human Should Go First



# Communication-like strategies emerged from optimizing in a system that accounts for human reactions.



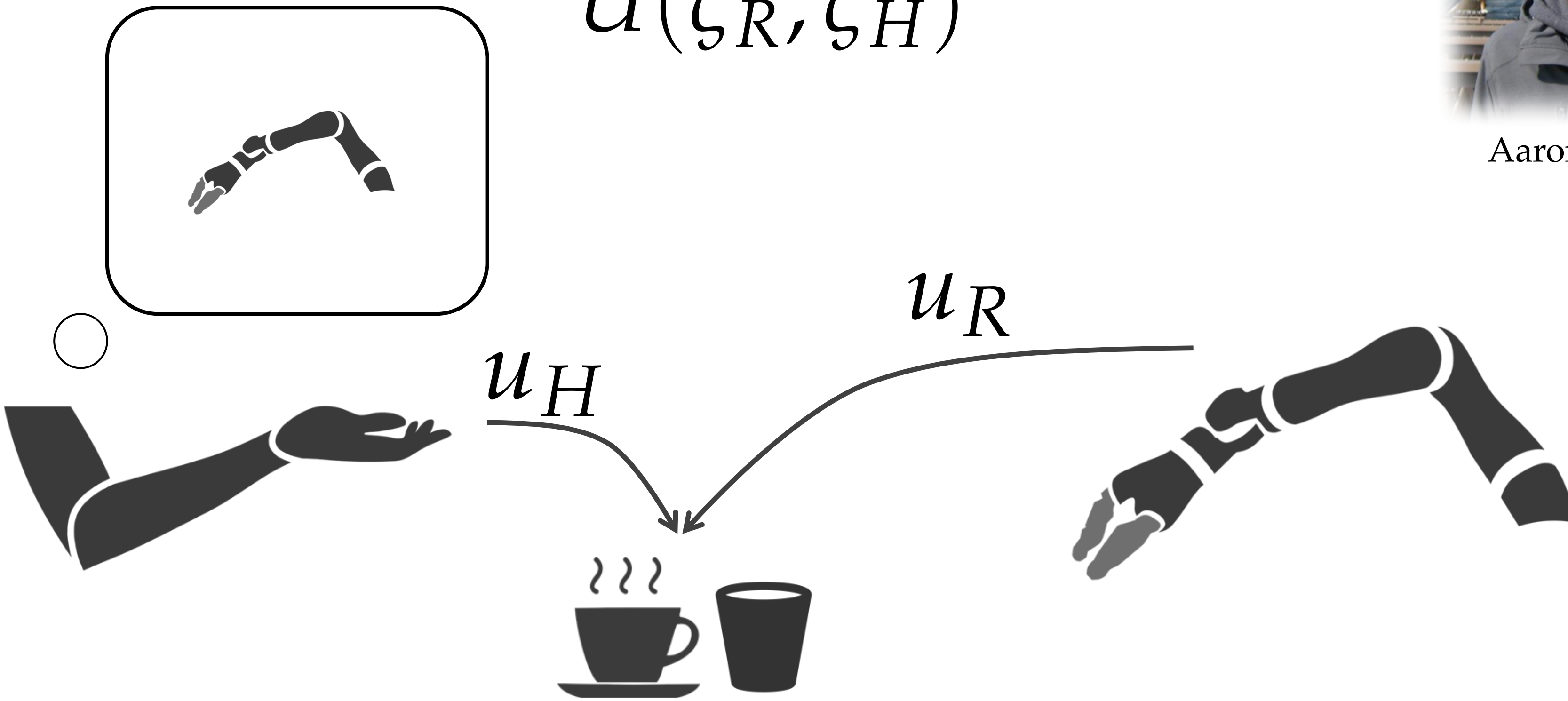

$U_H(\xi_H, \xi_R)$
$U_R(\xi_R, \xi_H)$
$u_R$
$u_H$



$U(\xi_R, \xi_H)$ *collaboration*
$u_R$
$u_H$
$\max_{\xi_R, \xi_H} U(\xi_R, \xi_H)$

Aaron Bestick

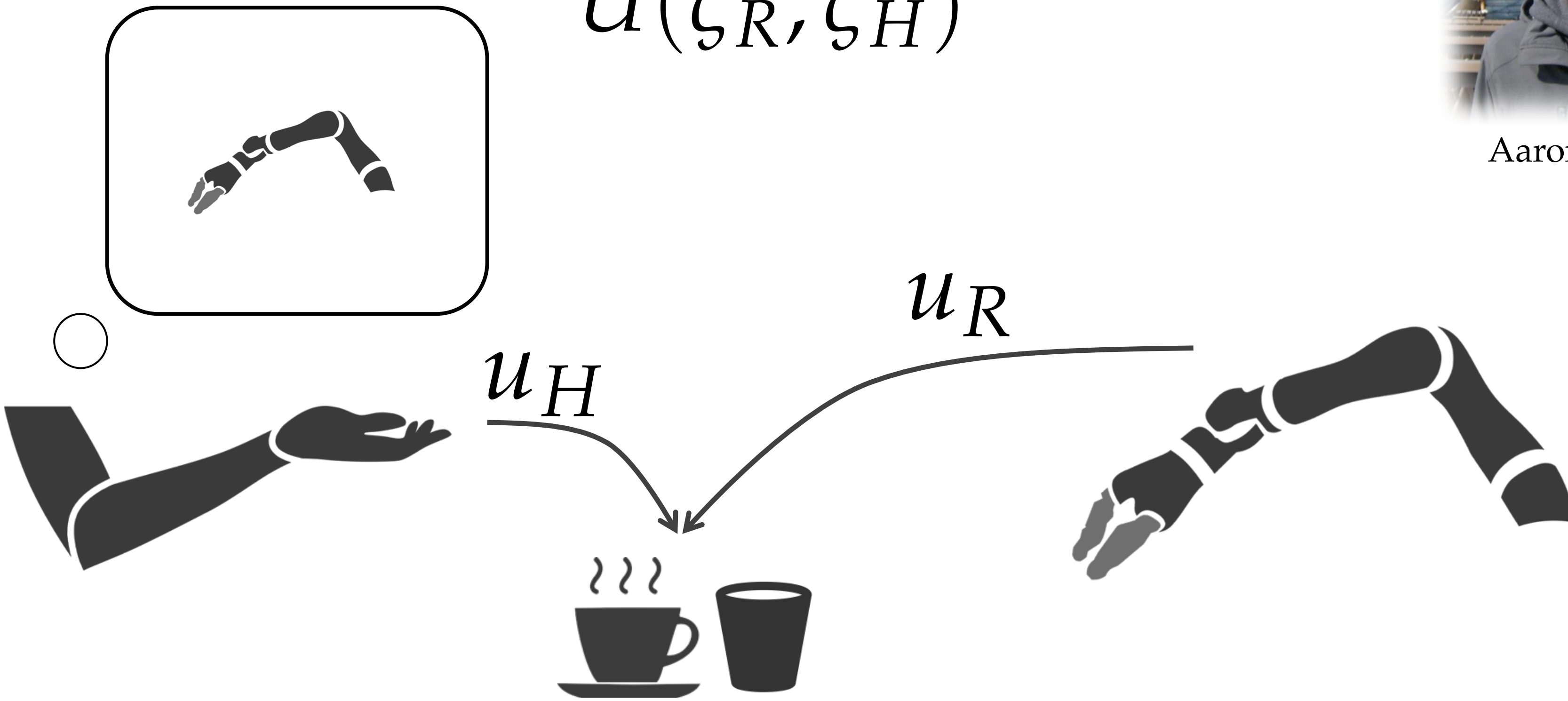
$$U(\xi_R, \xi_H)$$




*myopic human optimization*

$$\max_{u_H} U(u_H, u_R)$$

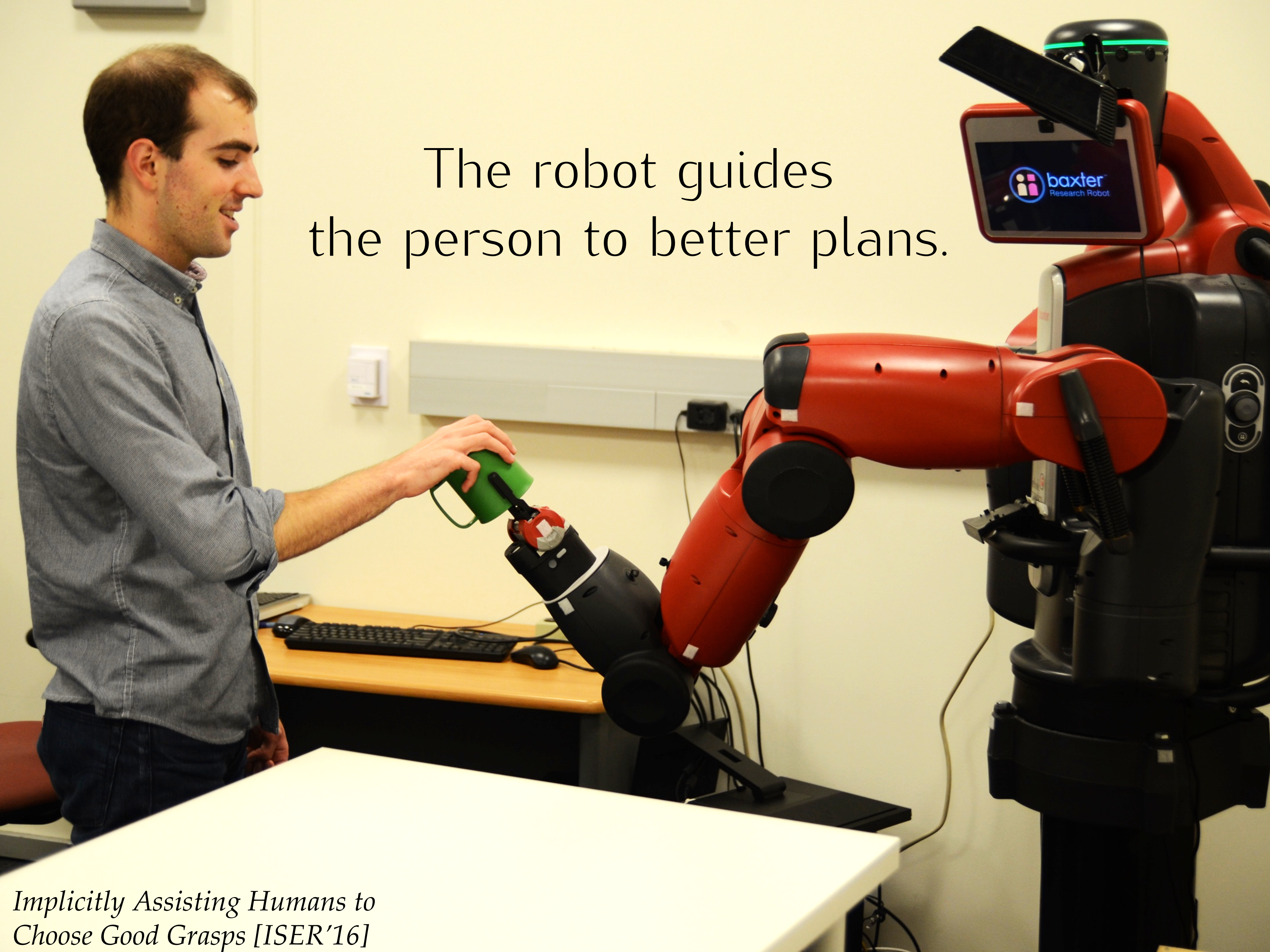
$$U(\xi_R, \xi_H)$$

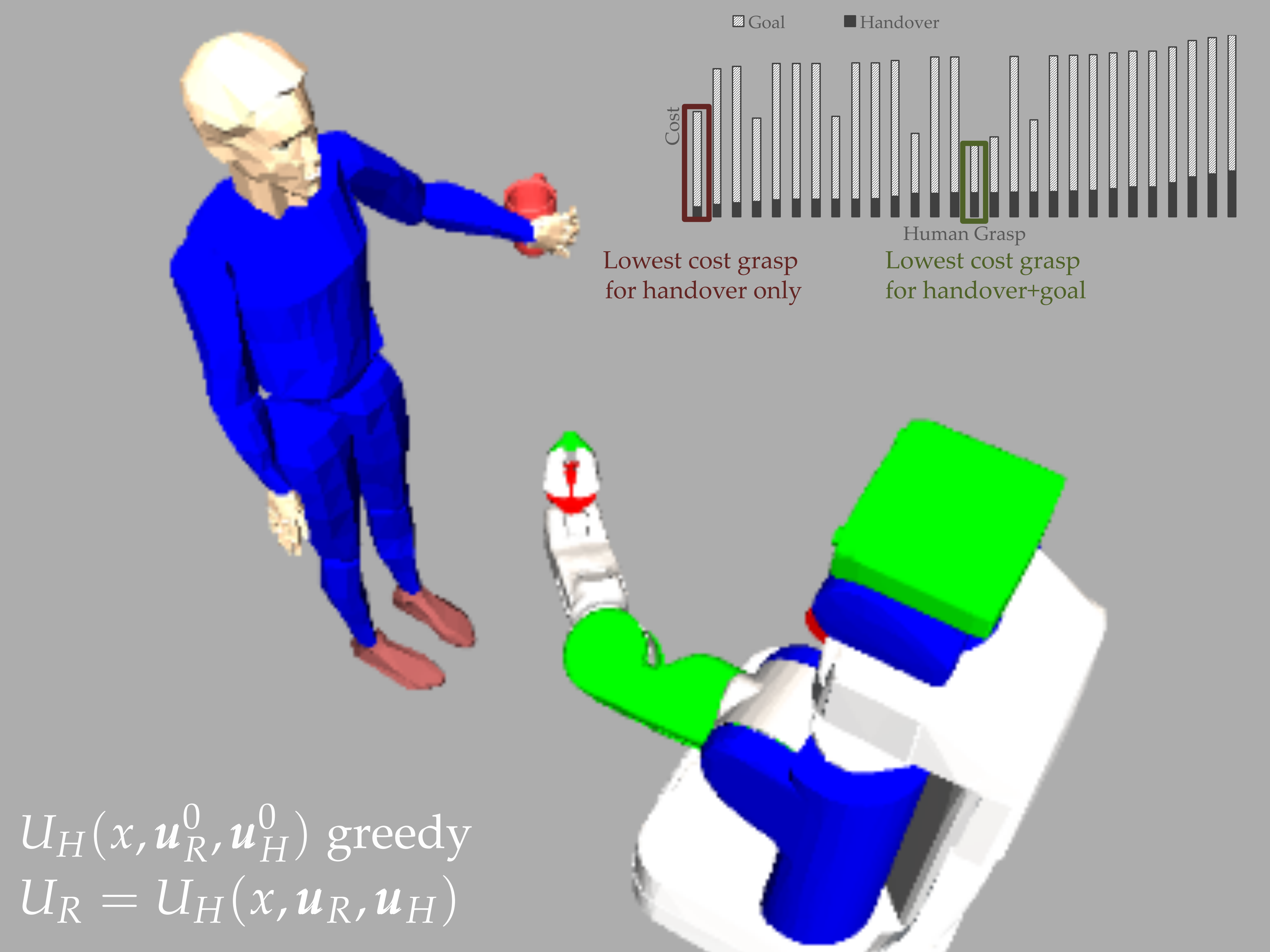



Aaron Bestick

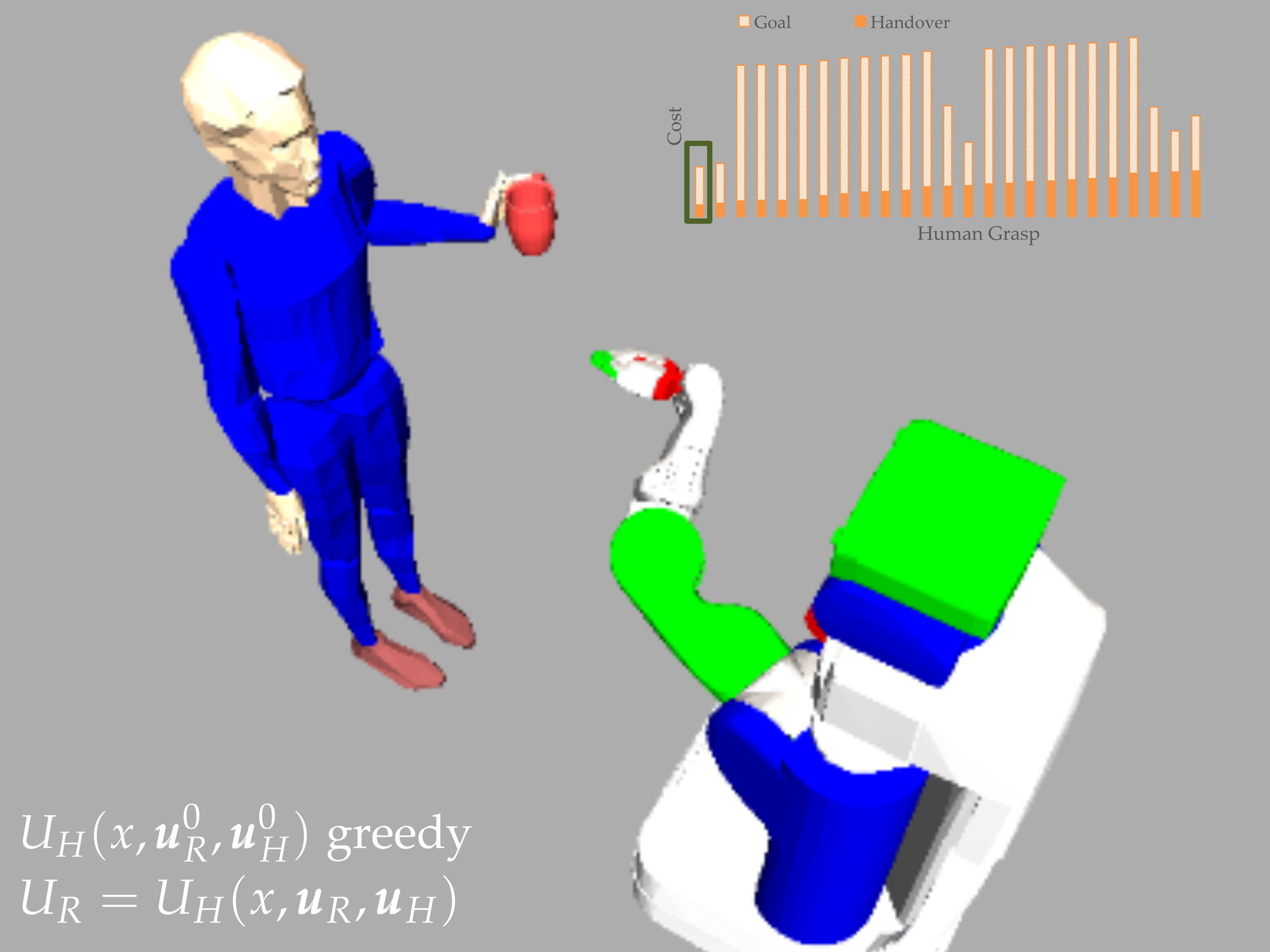
*myopic human optimization*

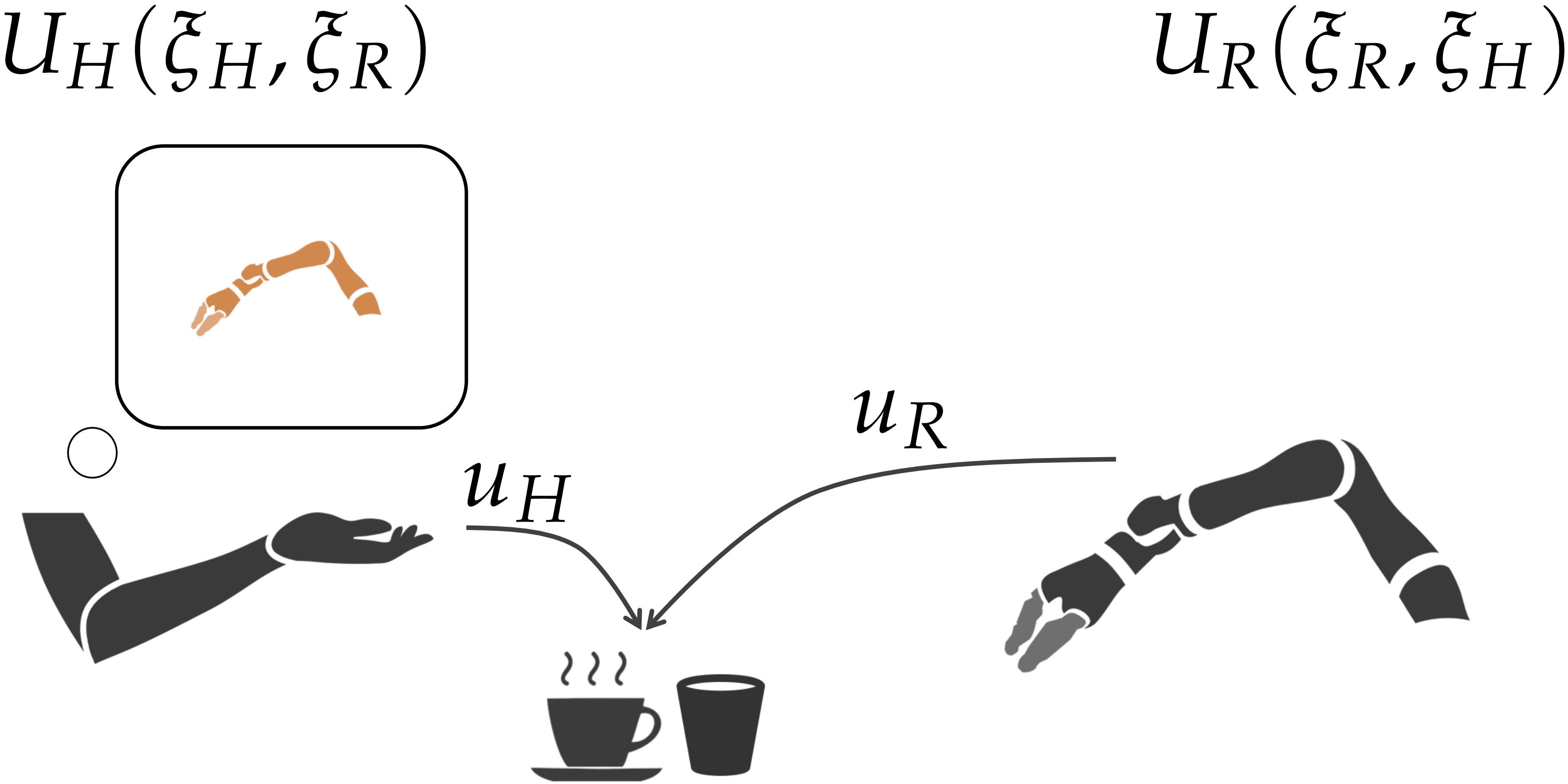
$$\max_{u_H} U(u_H, u_R) \qquad \max_{\xi_R} U(\xi_R, \xi_H(\xi_R))$$


The robot guides
the person to better plans.
baxter
Research Robot
Implicitly Assisting Humans to
Choose Good Grasps [ISER'16]

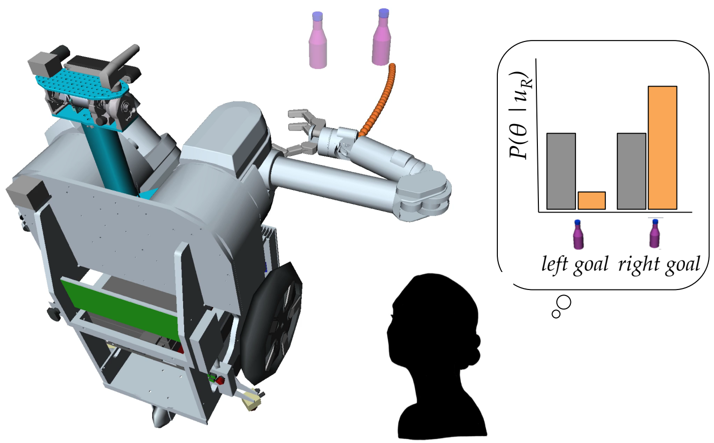



$$U_H(x, \boldsymbol{u}_R^0, \boldsymbol{u}_H^0) \text{ greedy}$$
$$U_R = U_H(x, \boldsymbol{u}_R, \boldsymbol{u}_H)$$

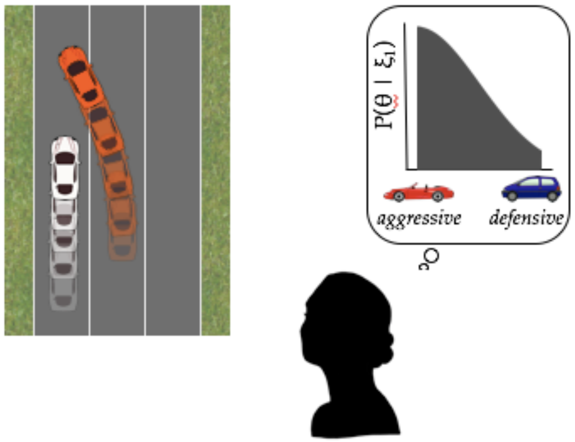

Goal
Handover
Cost
Human Grasp
$U_H(x, \boldsymbol{u}_R^0, \boldsymbol{u}_H^0)$ greedy
$U_R = U_H(x, \boldsymbol{u}_R, \boldsymbol{u}_H)$


$U_H(\xi_H, \xi_R)$
$U_R(\xi_R, \xi_H)$
$u_H$
$u_R$

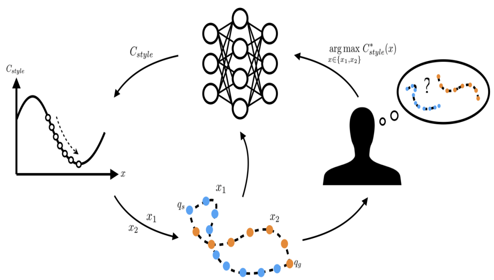
# Expressive Robots



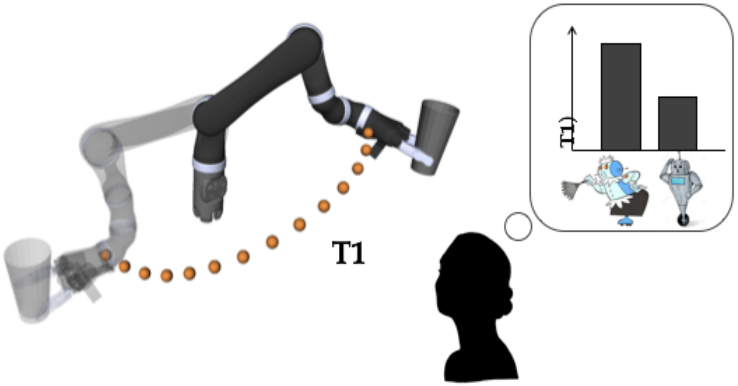

Goals [RSS'13]
best paper finalist




Utility [RSS'17]

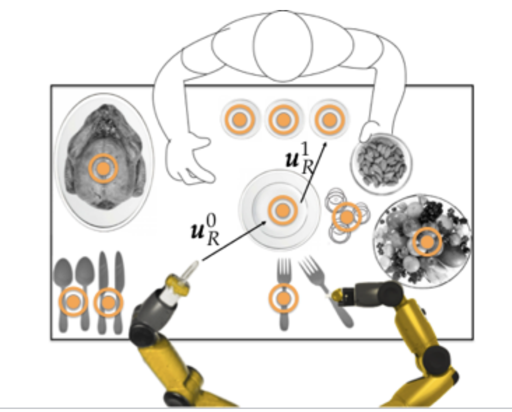


Style [in review]

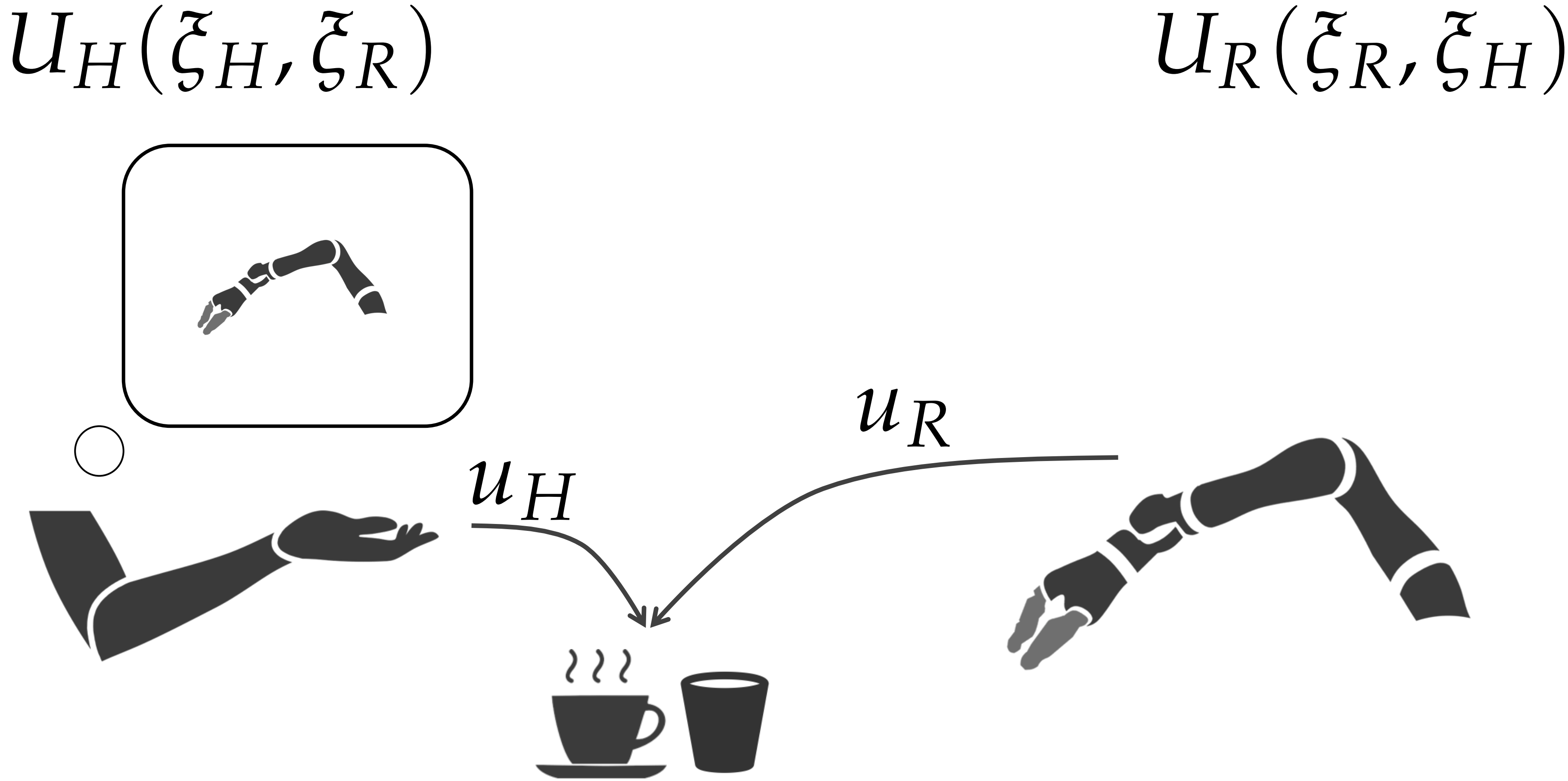



Timing [HRI'17]

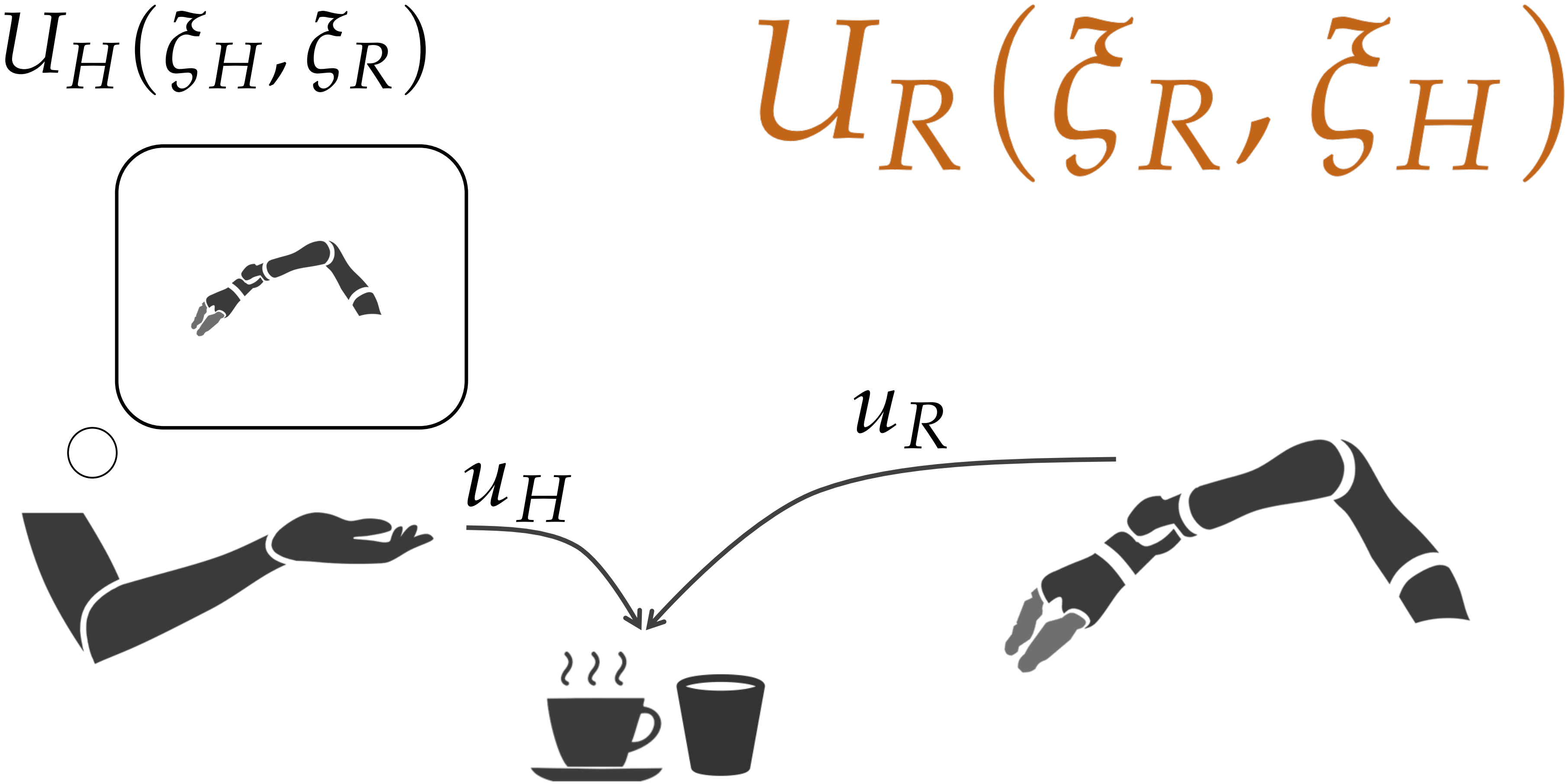
Incapability [HRI'18]
best paper finalist

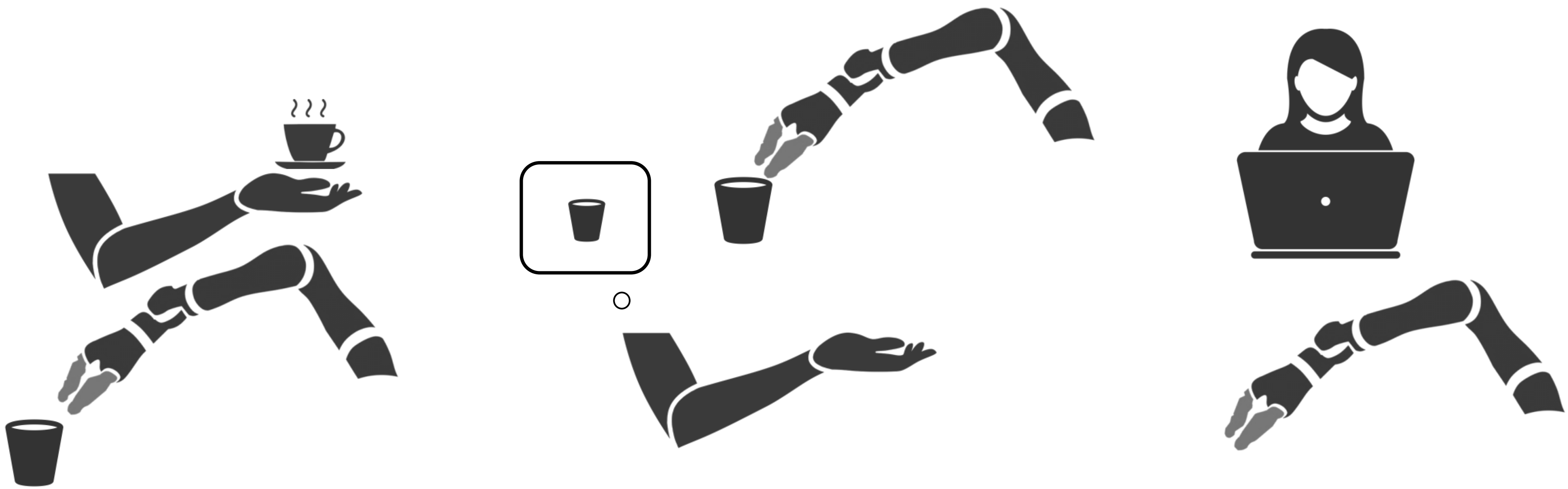


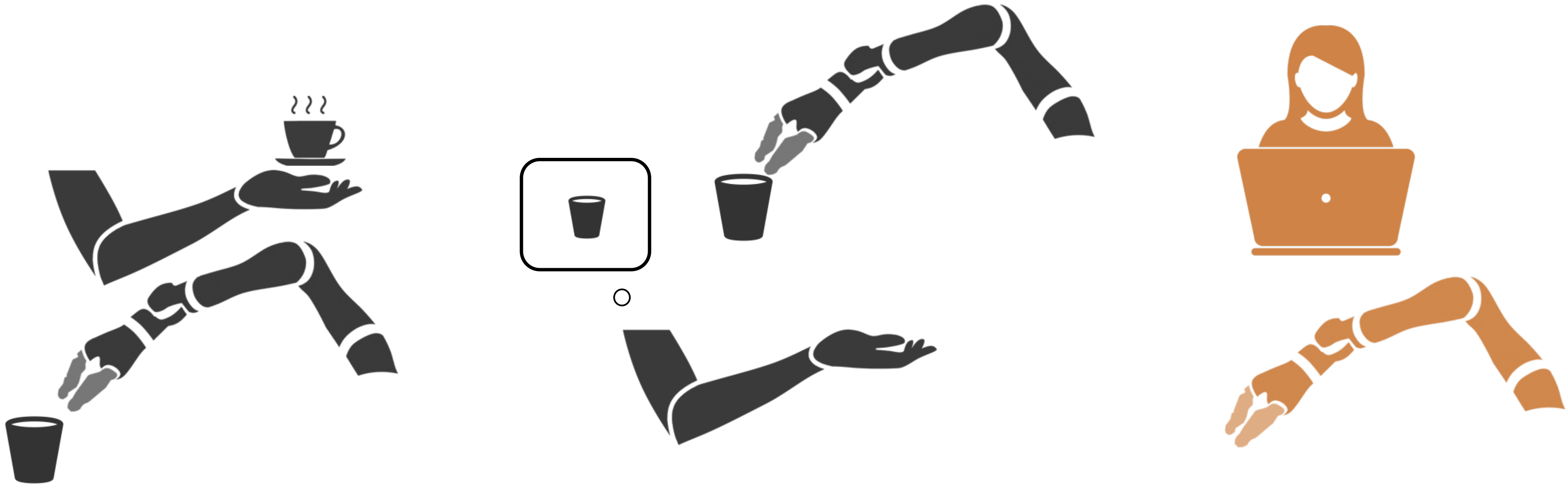

Task Plans[WAFR'16]




Coordination requires reasoning about effects on human actions and beliefs.


$u_H(\xi_H, \xi_R)$
$u_R(\xi_R, \xi_H)$
$u_R$
$u_H$





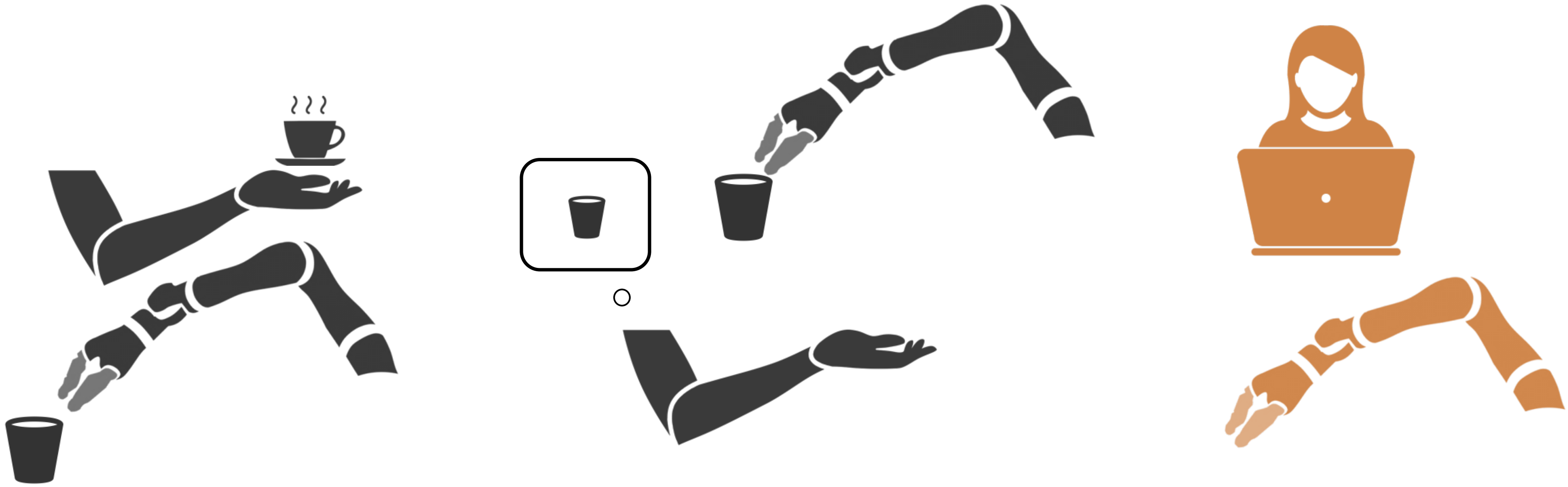
# Faulty Reward Functions in the Wild

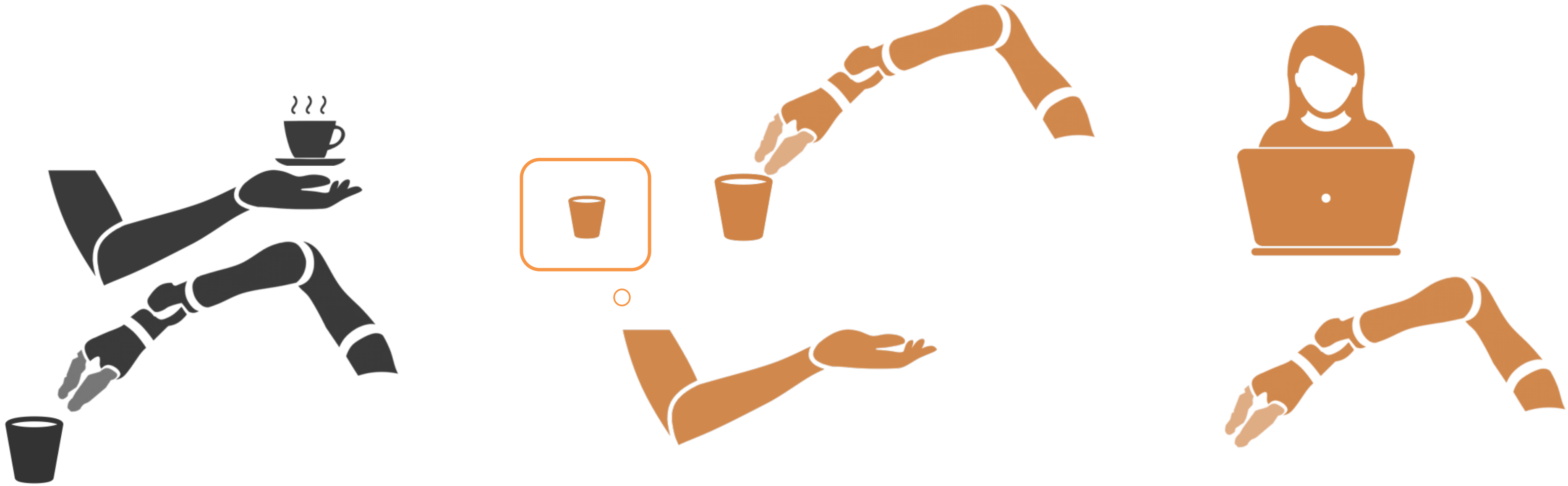
JACK CLARK & DARIO AMODEI

DECEMBER 21, 2016

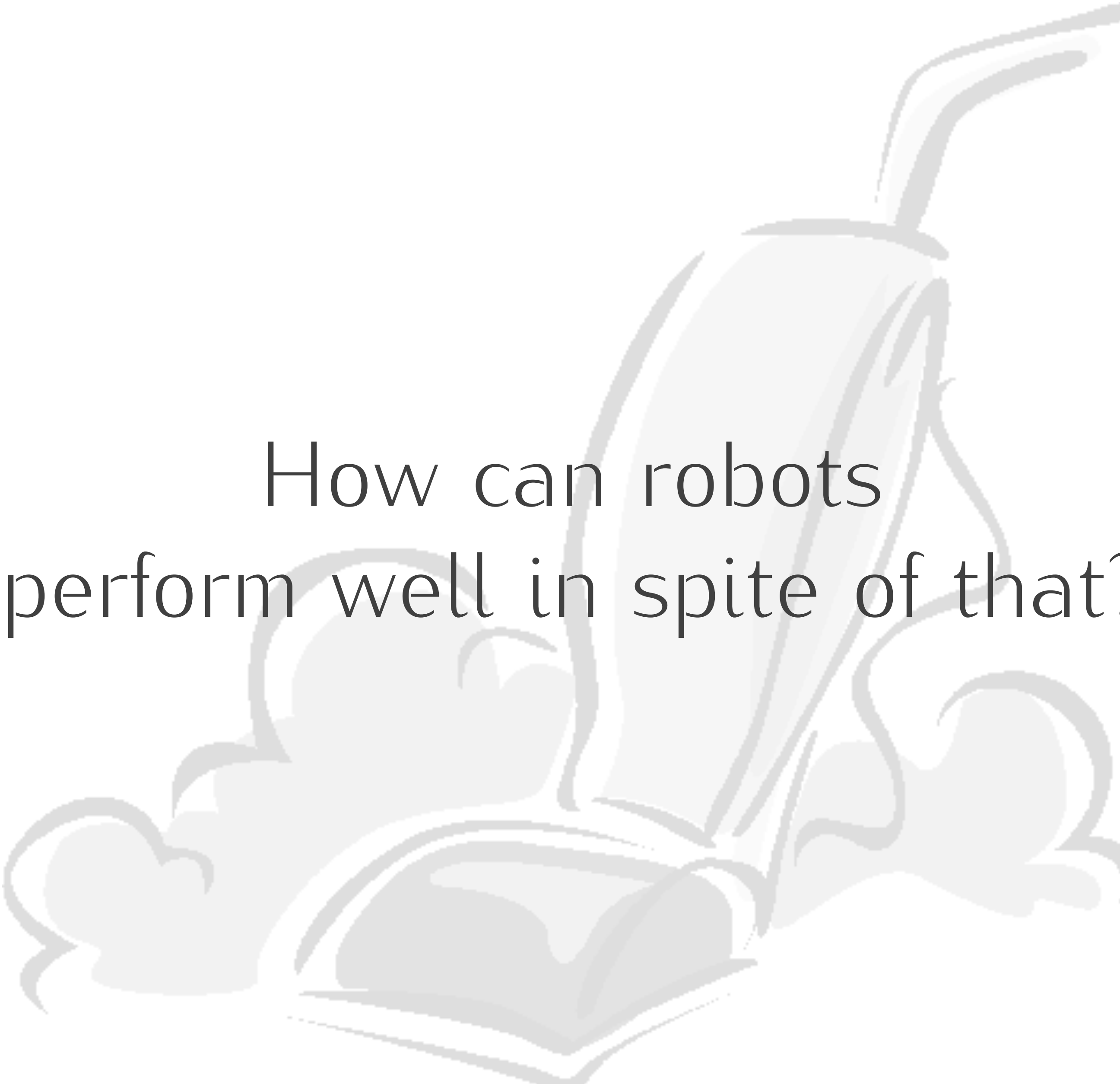
**Reinforcement learning algorithms can break in surprising, counterintuitive ways.** In this post we'll explore one failure mode, which is where you misspecify your reward function.


START
START
SCORE
0
LAPS
--/3
TIME
0:01
TURBO
MORE GAMES





We are bad at specifying utility functions for robots.



How can robots
perform well in spite of that?

[NIPS16,ICRA16,CDC16,
HRI17,ICRA17,IJCAI17a,IJCAI17b,
RSS17b,CoRL17, ISRR17, NIPS17, HRI18a, HRI18b]

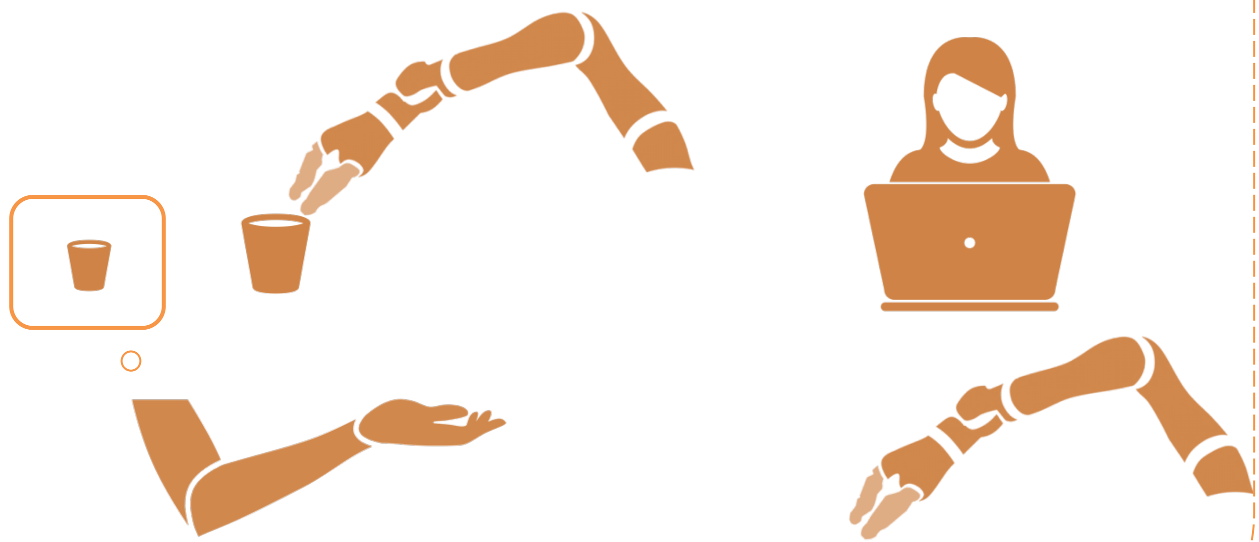
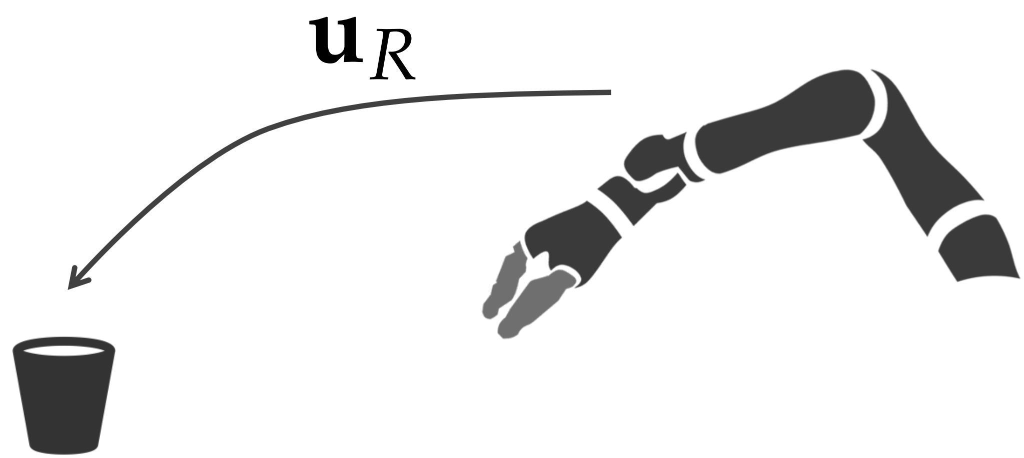
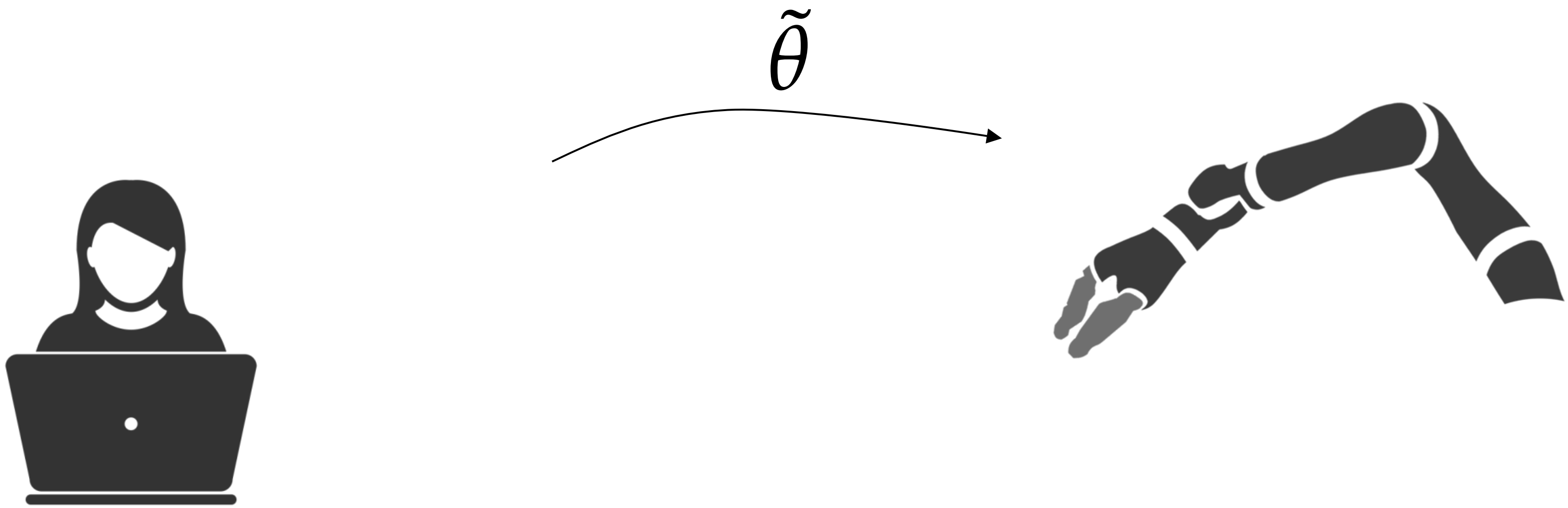


Figure out what utility to optimize.

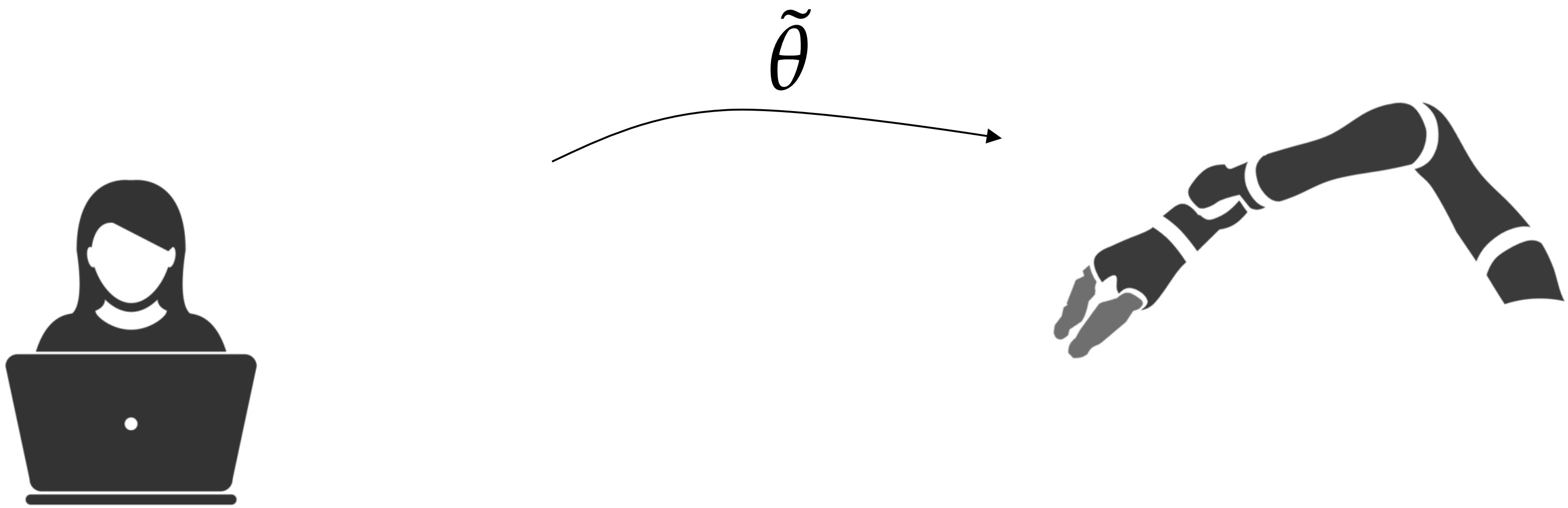
$$U_R(x_0, \mathbf{u}_R; \theta)$$

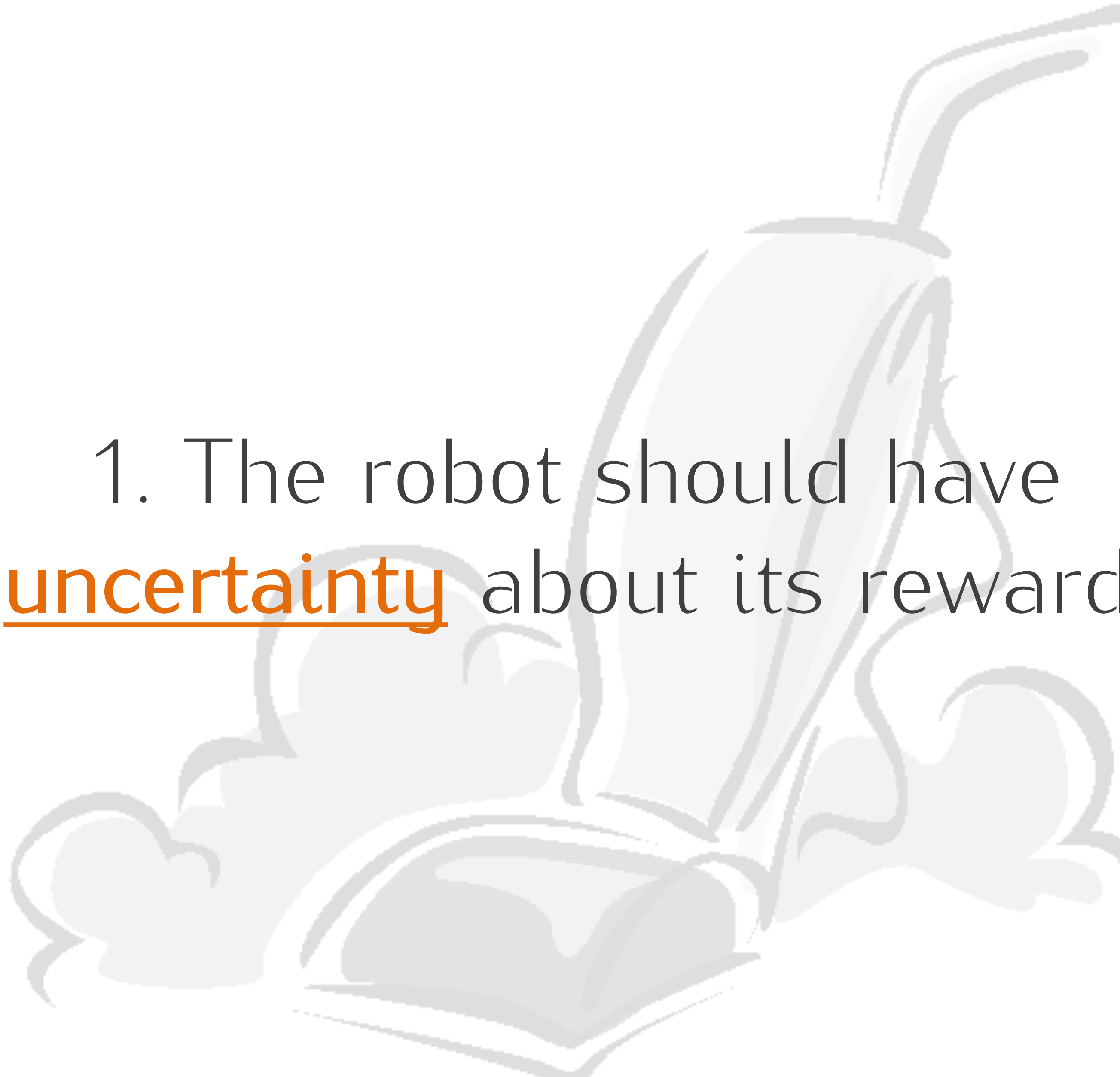


$$U_R(x_0, \mathbf{u}_R; \theta)$$

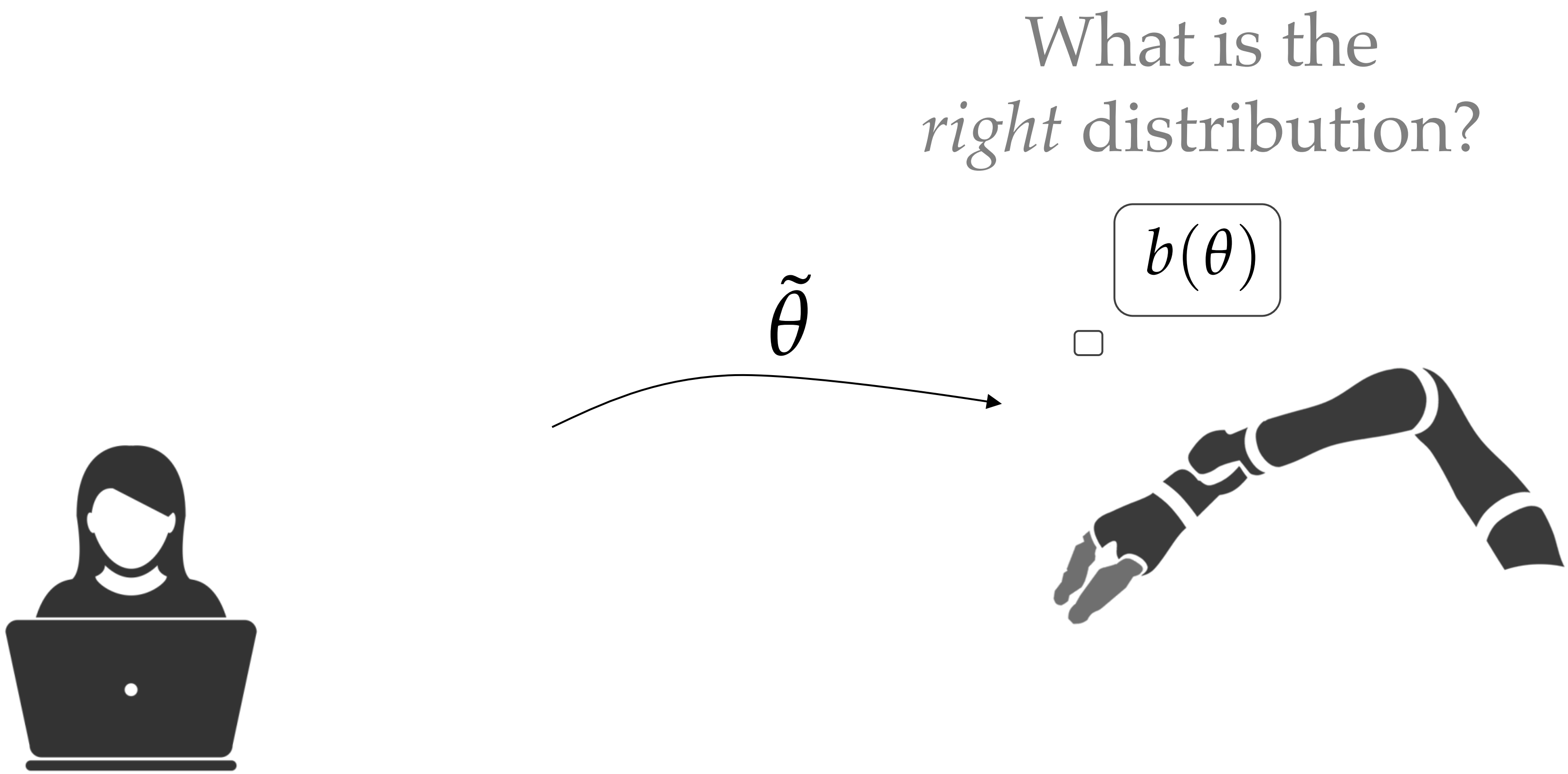


$$U_R(x_0, \mathbf{u}_R; \tilde{\theta})$$

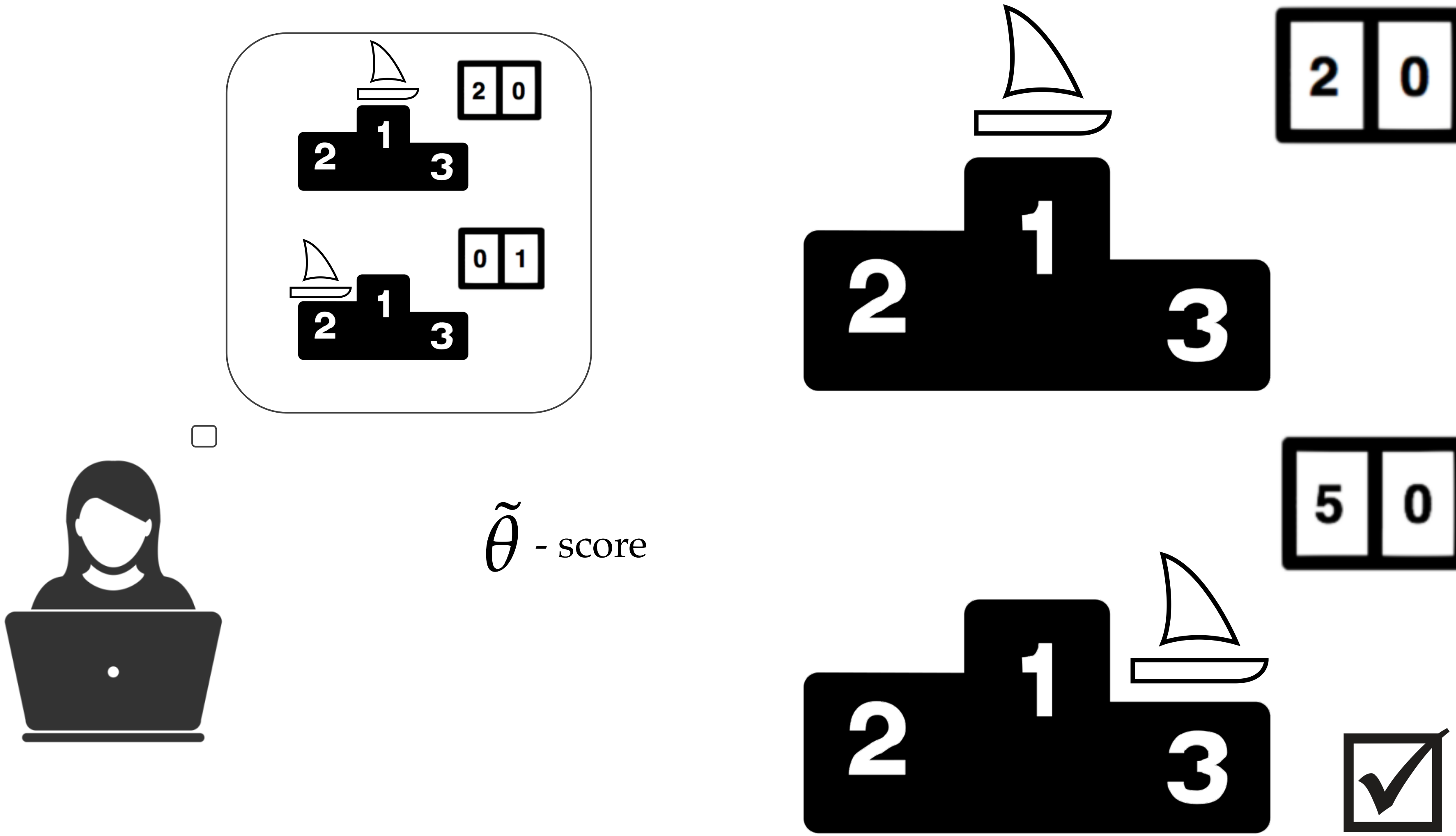




1. The robot should have <u>uncertainty</u> about its reward.

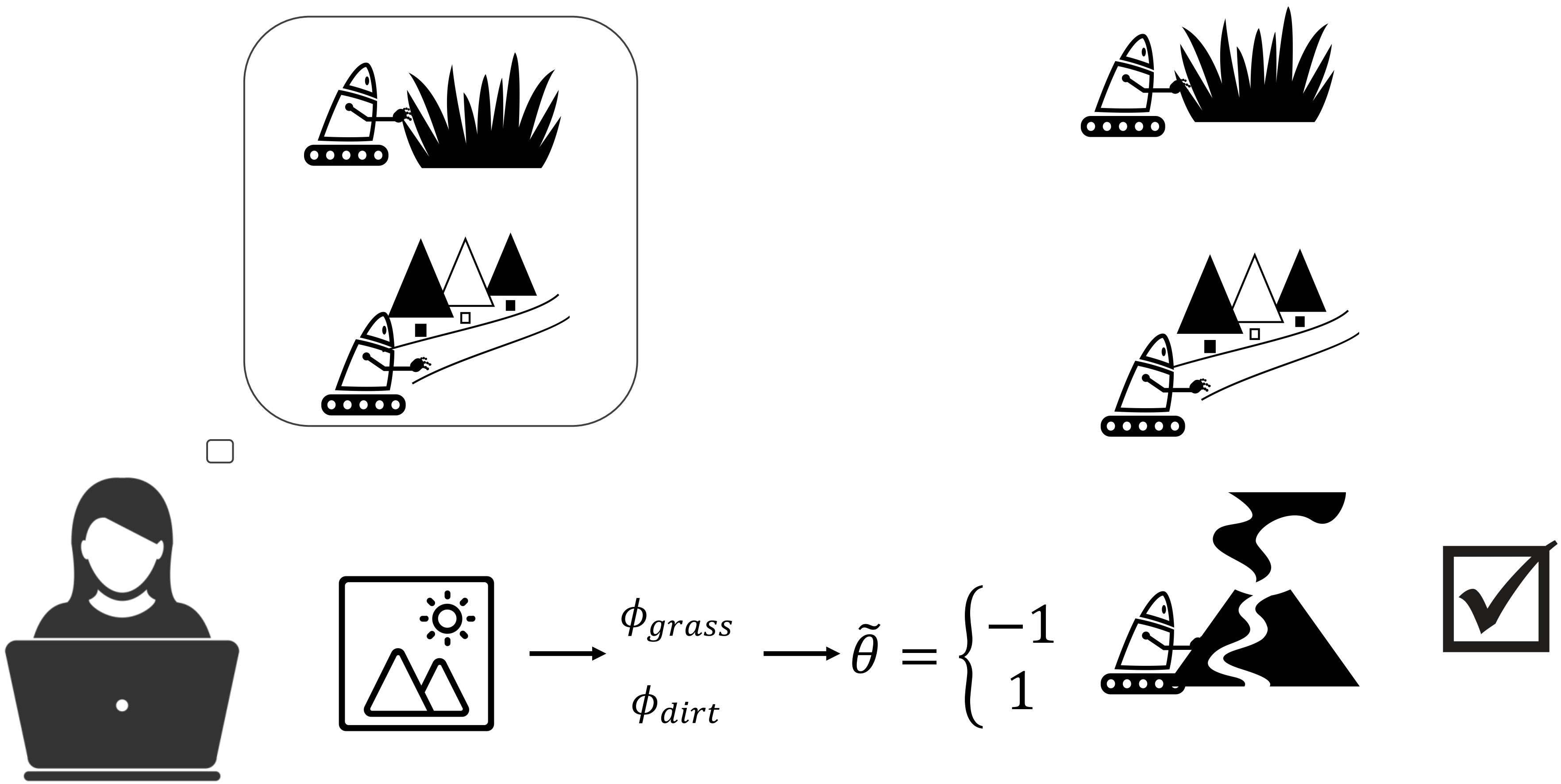


What is the *right* distribution?
$b(\theta)$
$\tilde{\theta}$




score and winning were *correlated* at training time…

… but no longer correlated at test time




lava was not
present at training time

... but appeared
at test time

Dylan Hadfield-Menell

Smitha Milli

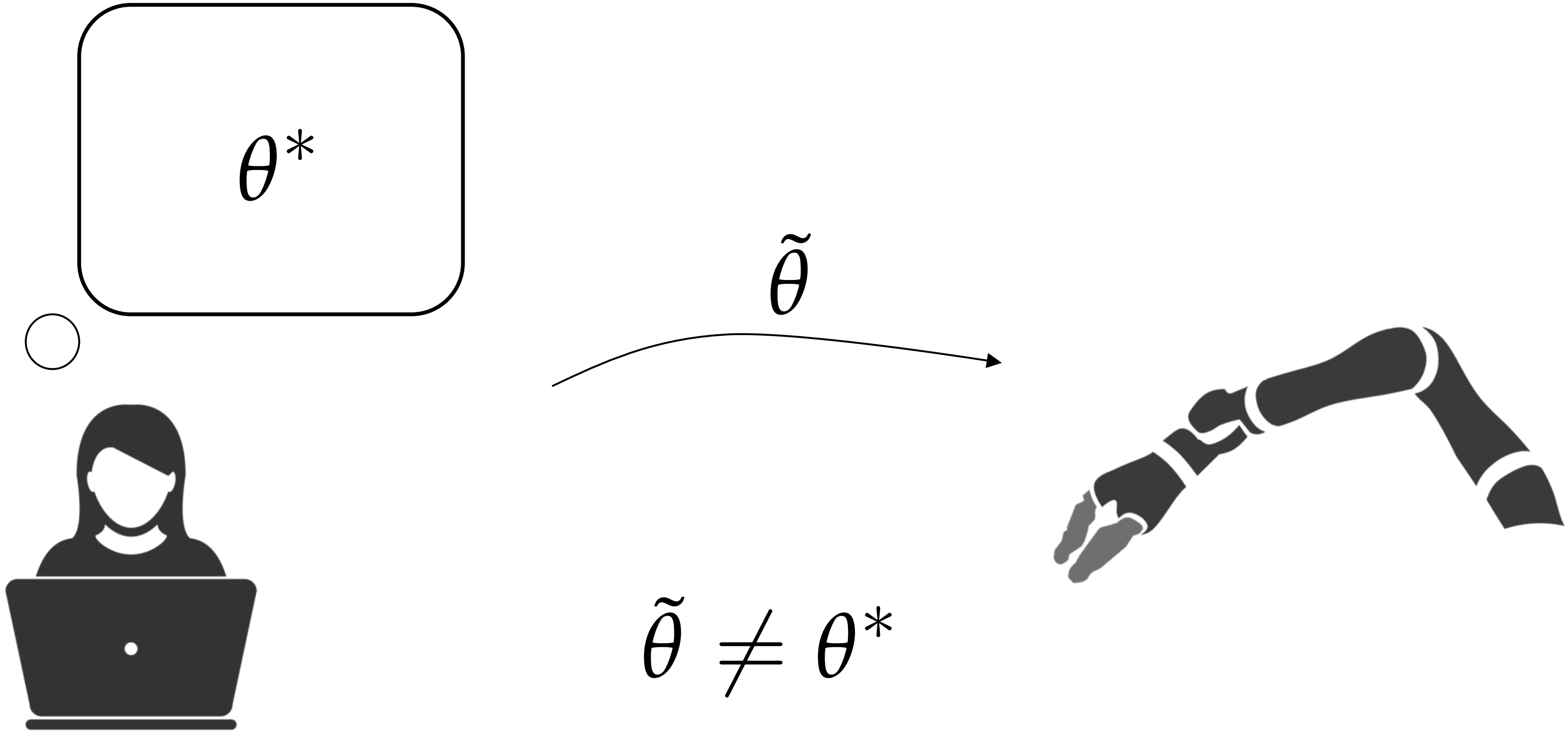
2. All we know about the <u>true</u> reward is that the <u>specified</u> reward works well in the <u>training</u> envs.

Dylan Hadfield-Menell

Smitha Milli

2. The behavior incentivized by the specified reward in training has high <u>true</u> reward.

# Reward Design



# Inverse Reward Design

$\theta^*$

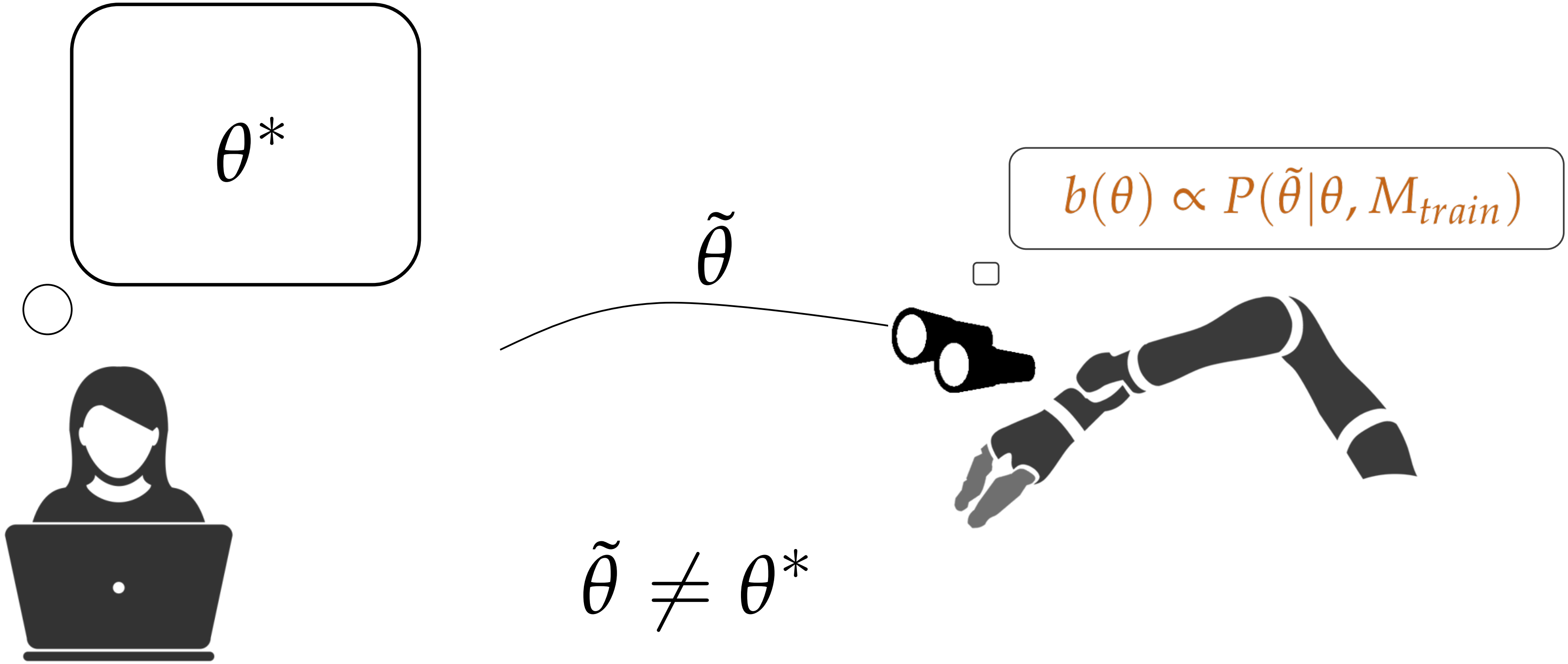
$\tilde{\theta}$

$b(\theta) \propto P(\tilde{\theta}|\theta, M_{train})$

$\tilde{\theta} \neq \theta^*$

*Inverse Reward Design*
*[NIPS'17,oral]*

# Inverse Reward Design




*Inverse Reward Design*
*[NIPS'17,oral]*

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta}\middle|\theta^*, M_{train}\right) \propto \mathrm{e}^{\beta\mathbb{E}[R(\xi;\theta^*,M_{train})\,|\,\xi\sim P(\xi|\tilde{\theta},M_{train})]}$$

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta}\middle|\theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*,M_{train}) \mid \xi \sim P(\xi|\tilde{\theta},M_{train})]}$$

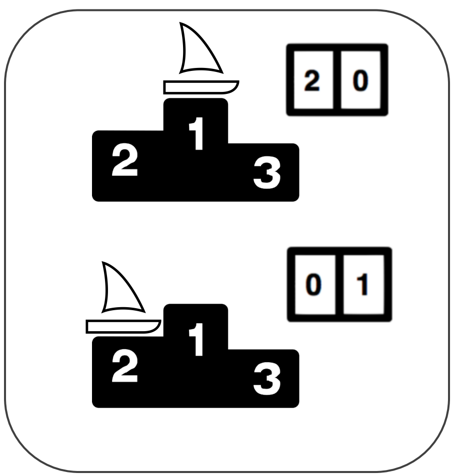
The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta} \middle| \theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*,M_{train}) \mid \xi \sim P(\xi|\tilde{\theta},M_{train})]}$$

# The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta}\middle|\theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*, M_{train}) \mid \xi \sim P(\xi|\tilde{\theta}, M_{train})]}$$

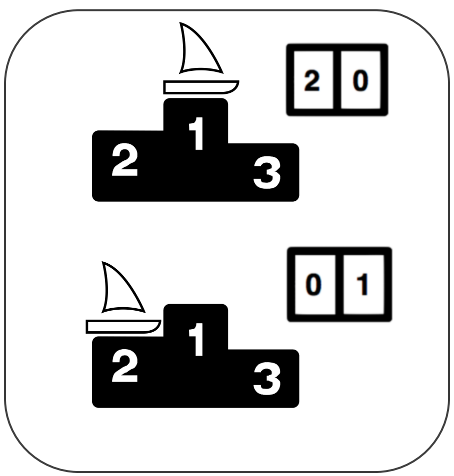
$M_{train}$



The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta}\middle|\theta^*, M_{train}\right) \propto e^{\beta\mathbb{E}[R(\xi;\theta^*,M_{train})\,|\,\xi\sim P(\xi|\widetilde{\theta},M_{train})]}$$

$M_{train}$




$\theta_1$ maximizing winning

$\theta_2$ maximizing score

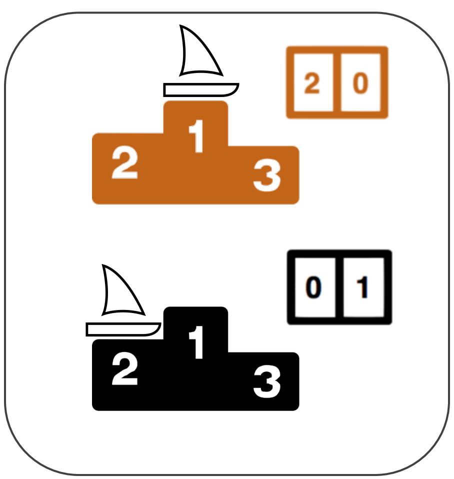
$\theta_3$ minimizing winning

$\theta_4$ minimizing score

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta} \middle| \theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*,M_{train}) \mid \xi \sim P(\xi|\tilde{\theta},M_{train})]}$$

$M_{train}$




$\theta_1$ maximizing winning

$\theta_2$ maximizing score

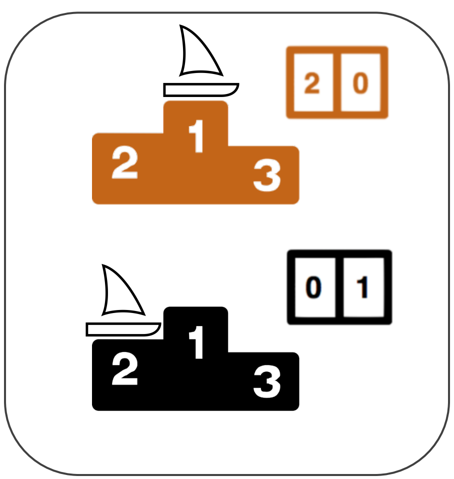
$\theta_3$ minimizing winning

$\theta_4$ minimizing score

The behavior incentivized by the specified reward in training has high true reward

$$P(\tilde{\theta} | \theta^*, M_{train}) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*, M_{train}) \,|\, \xi \sim P(\xi | \tilde{\theta}, M_{train})]}$$

$M_{train}$




$\theta_1$ maximizing winning

$\theta_2$ maximizing score

$\theta_3$ minimizing winning

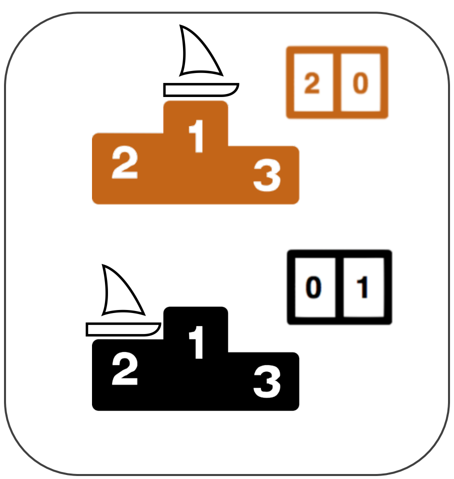
$\theta_4$ minimizing score

?

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta} \middle| \theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*, M_{train}) \,|\, \xi \sim P(\xi | \tilde{\theta}, M_{train})]}$$

$M_{train}$




$\theta_1$ maximizing winning

$\theta_2$ maximizing score

✓

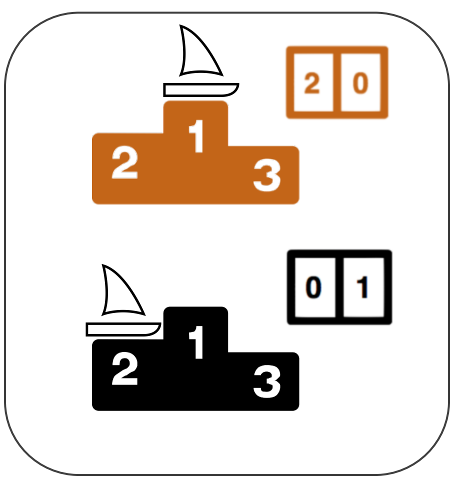
$\theta_3$ minimizing winning

$\theta_4$ minimizing score

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta}\middle|\theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*,M_{train}) \mid \xi\sim P(\xi|\tilde{\theta},M_{train})]}$$

$M_{train}$

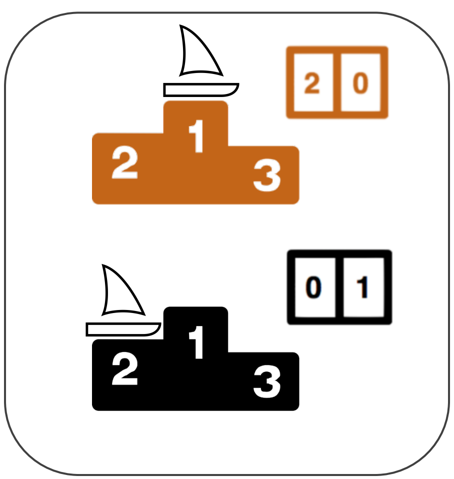



| $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|---|---|---|---|
| maximizing winning | maximizing score | minimizing winning | minimizing score |
| | ✓ | | ? |

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta} \middle| \theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*, M_{train}) \mid \xi \sim P(\xi | \tilde{\theta}, M_{train})]}$$

$M_{train}$

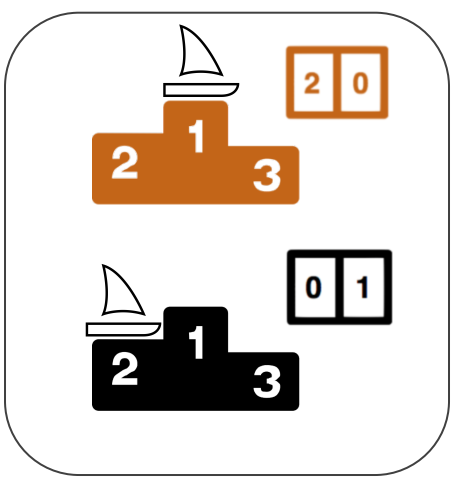



| $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|---|---|---|---|
| maximizing winning | maximizing score | minimizing winning | minimizing score |
| | ✓ | | ✗ |

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta} \middle| \theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*, M_{train}) \mid \xi \sim P(\xi|\tilde{\theta}, M_{train})]}$$

$M_{train}$

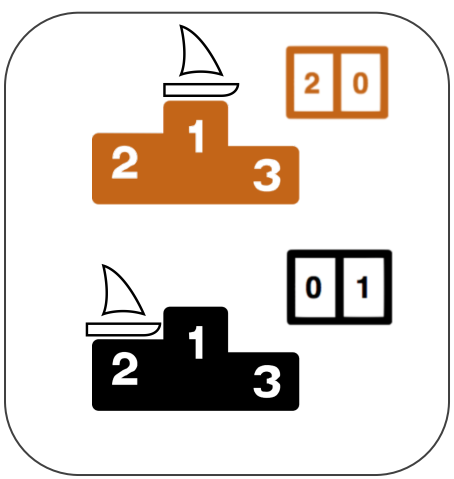


| $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|---|---|---|---|
| maximizing winning | maximizing score | minimizing winning | minimizing score |
| | ✓ | ✗ | ✗ |

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta}\middle|\theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*, M_{train}) \mid \xi \sim P(\xi|\tilde{\theta}, M_{train})]}$$

$M_{train}$

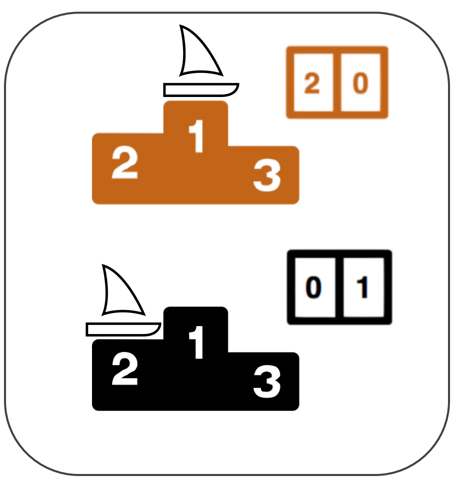



| $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|---|---|---|---|
| maximizing winning | maximizing score | minimizing winning | minimizing score |
| ? | ✓ | ✗ | ✗ |

The behavior incentivized by the specified reward in training has high true reward

$$P\left(\tilde{\theta} \middle| \theta^*, M_{train}\right) \propto e^{\beta \mathbb{E}[R(\xi;\theta^*, M_{train}) \,|\, \xi \sim P(\xi | \tilde{\theta}, M_{train})]}$$

$M_{train}$



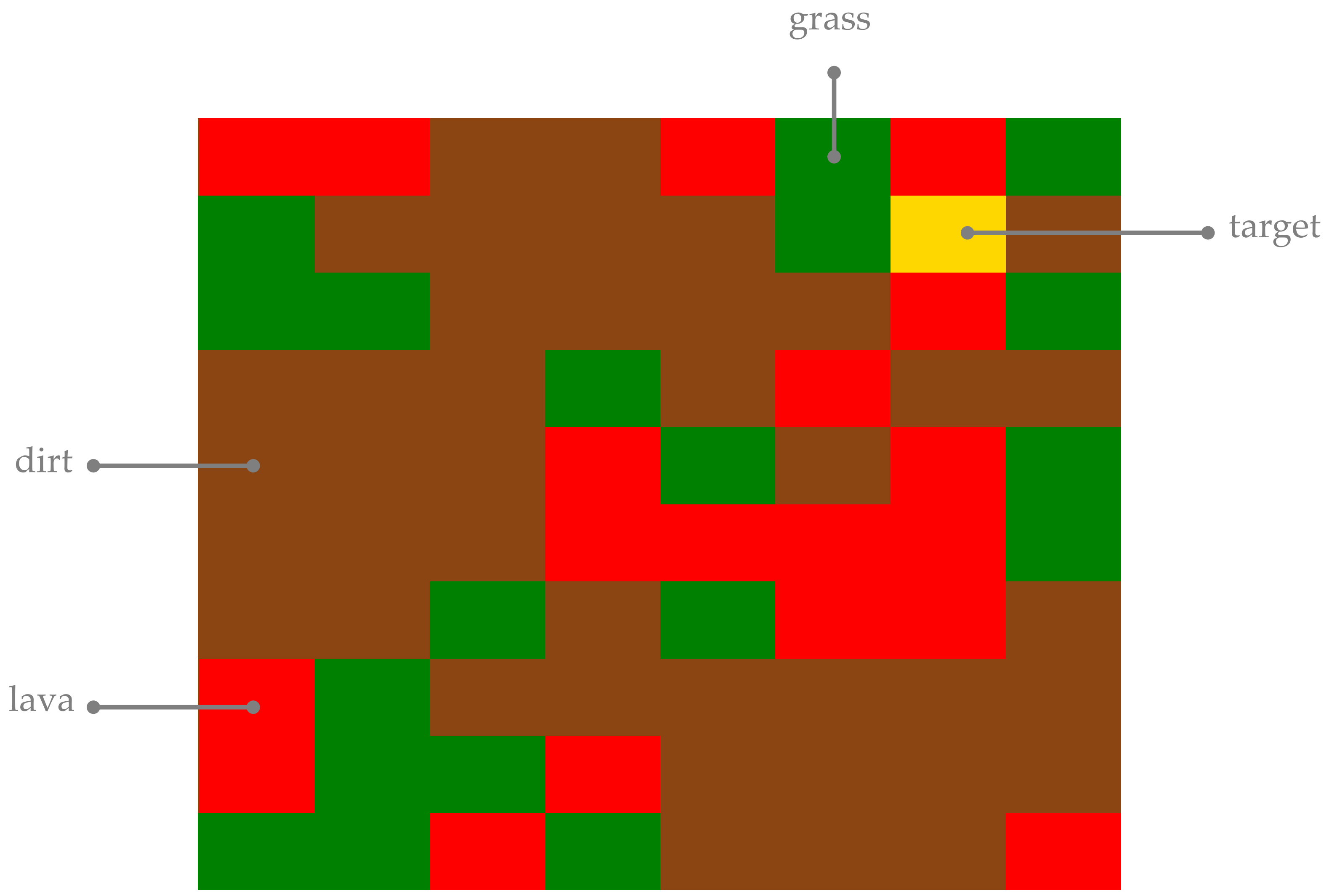

| $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|---|---|---|---|
| maximizing winning | maximizing score | minimizing winning | minimizing score |
| ✓ | ✓ | ✗ | ✗ |

What is the
*right* distribution?

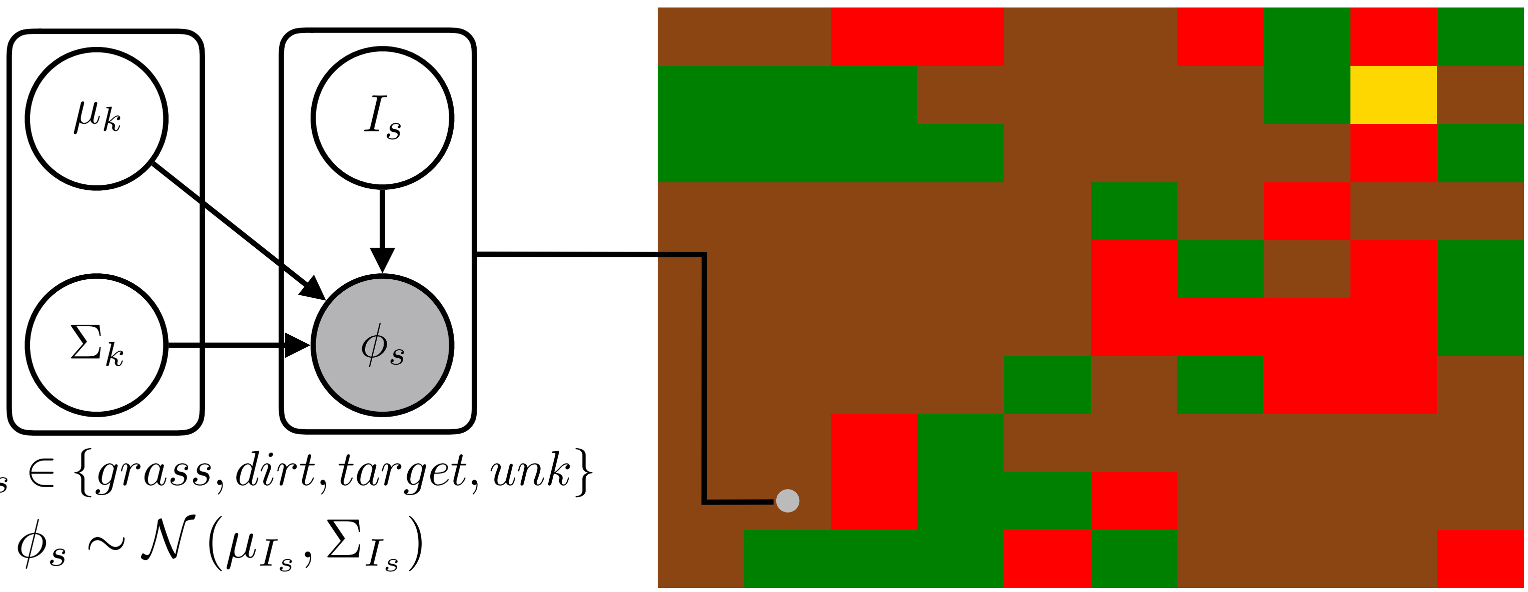
$b(\theta)$

$\tilde{\theta}$

# “La-Va-Land”



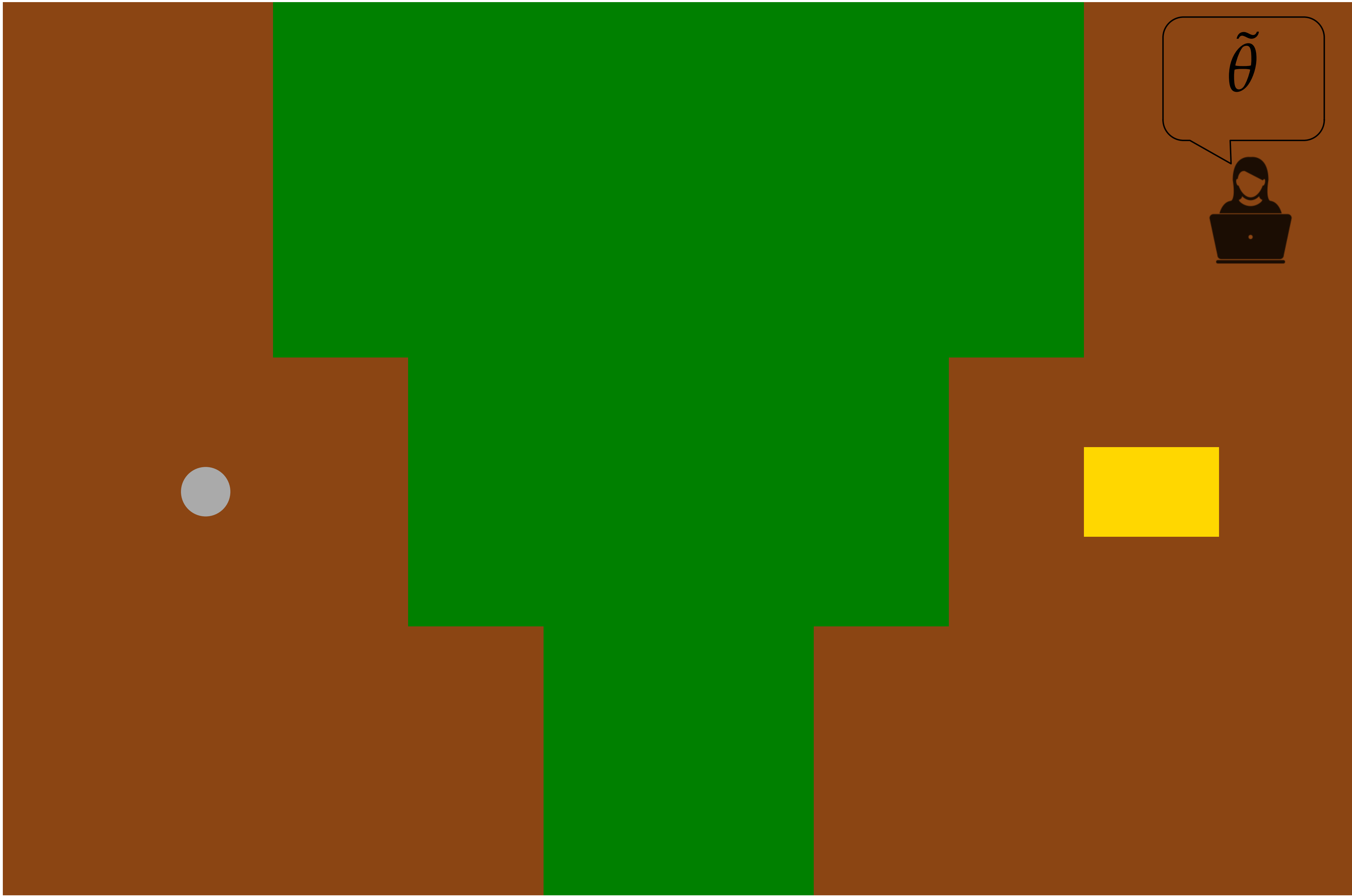
# Raw observations, no direct indicators...



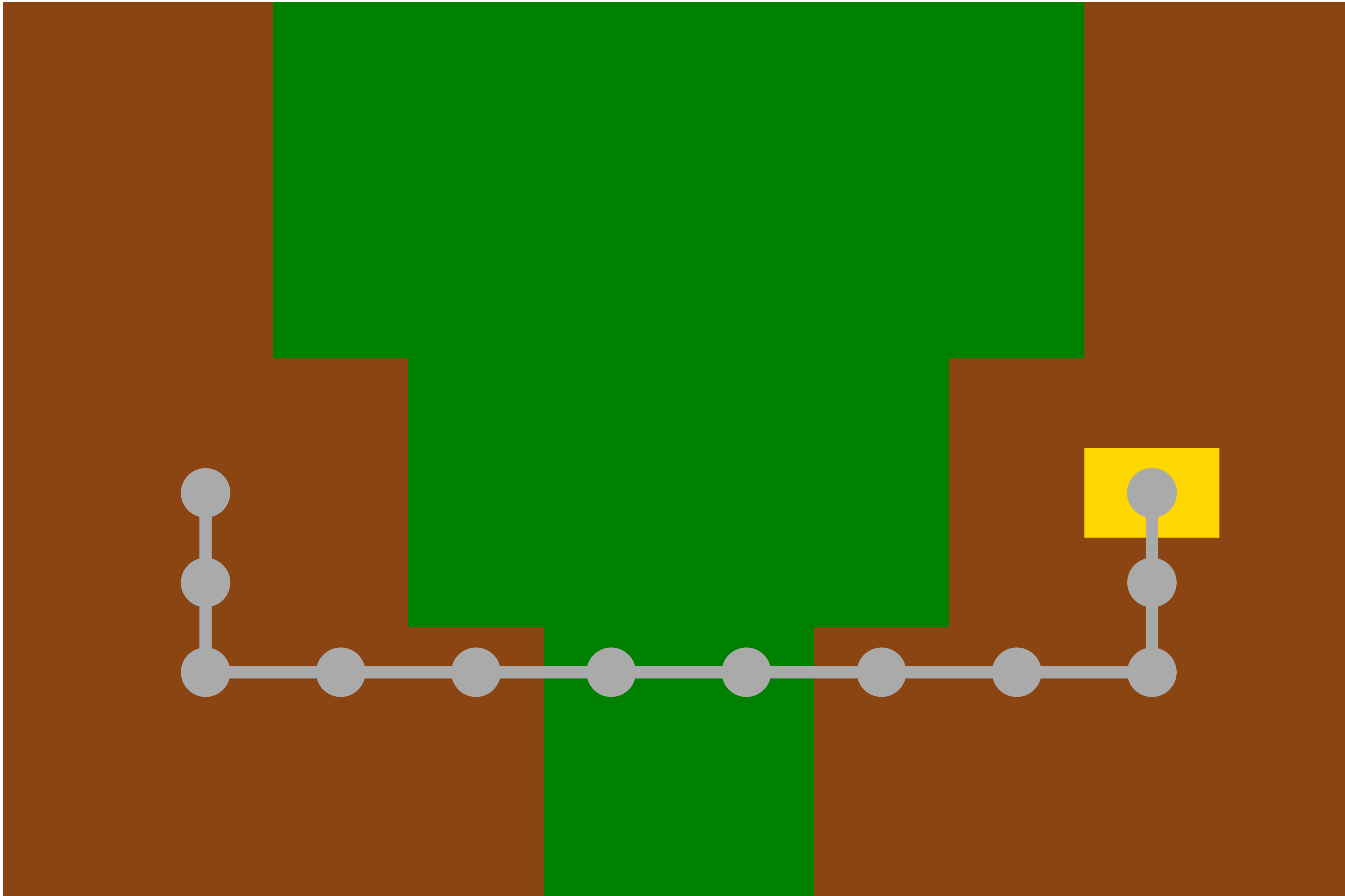

$$I_s \in \{grass, dirt, target, unk\}$$

$$\phi_s \sim \mathcal{N}\left(\mu_{I_s}, \Sigma_{I_s}\right)$$

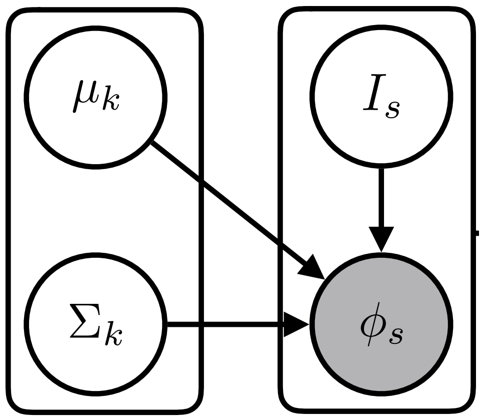
$M_{train}$



$M_{train} \quad \tilde{\theta}$

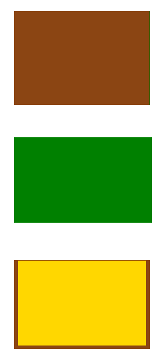


# Designer has proxy based on indicators (forgets lava)

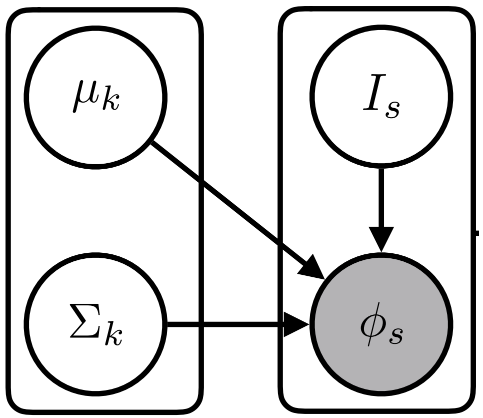



$$I_s \in \{grass, dirt, target, unk\}$$
$$\phi_s \sim \mathcal{N}\left(\mu_{I_s}, \Sigma_{I_s}\right)$$

$$\tilde{R}(I_s)$$

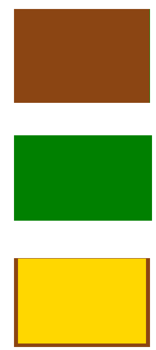


Designer has proxy based on indicators (forgets lava), and builds classifiers from raw obs to indicators



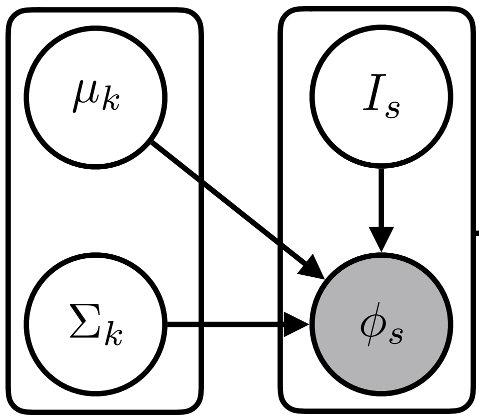

$$I_s \in \{grass, dirt, target, unk\}$$
$$\phi_s \sim \mathcal{N}\left(\mu_{I_s}, \Sigma_{I_s}\right)$$

$$\tilde{R}(I_s)$$



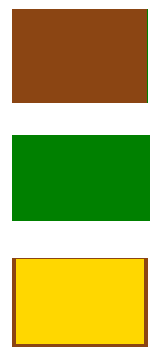
$$\phi_s \rightarrow dirt$$
$$\phi_s \rightarrow grass$$
$$\phi_s \rightarrow target$$

Designer has proxy based on indicators (forgets lava), and regresses proxy based on observations.



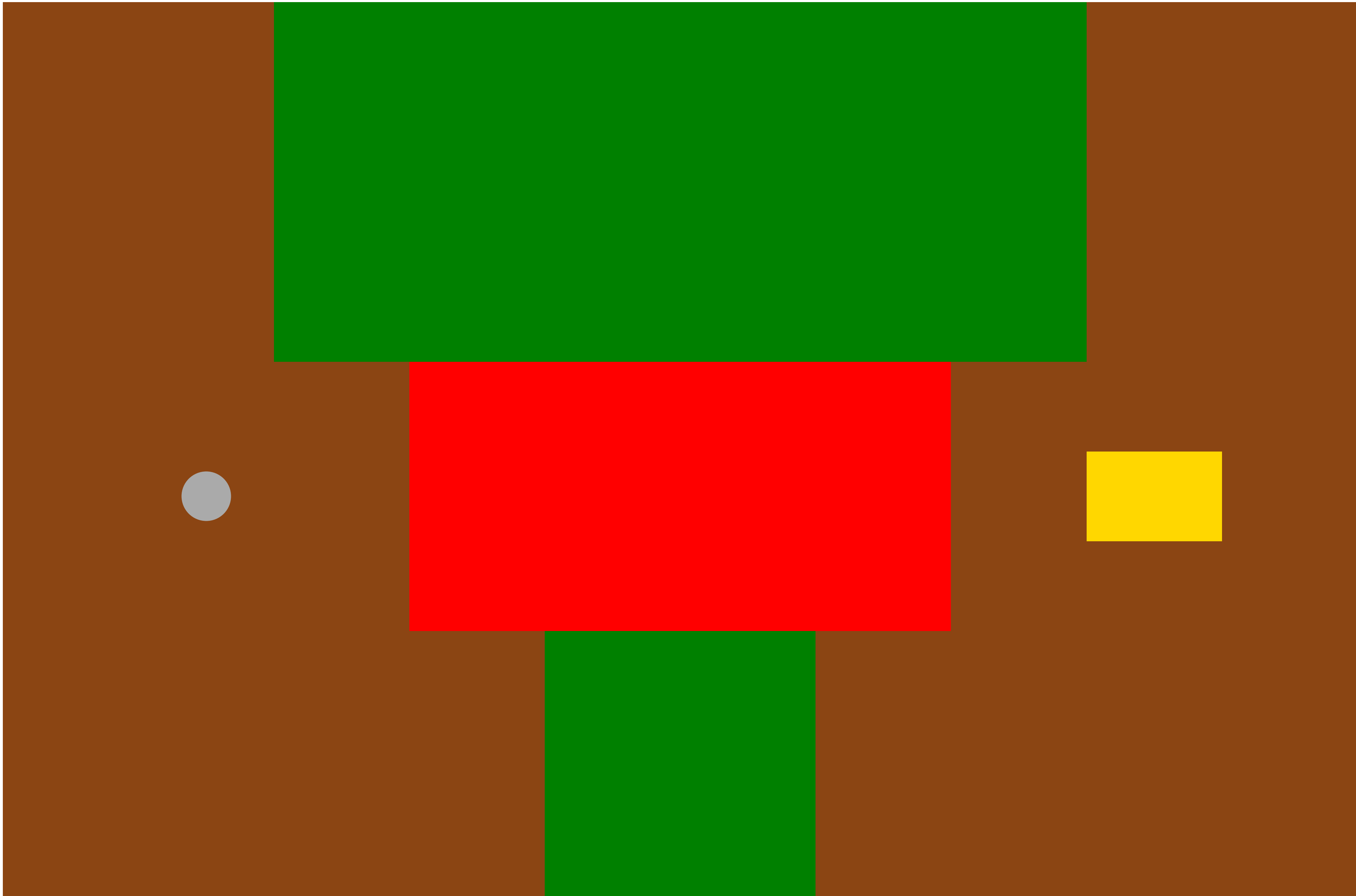

$$I_s \in \{grass, dirt, target, unk\}$$
$$\phi_s \sim \mathcal{N}\left(\mu_{I_s}, \Sigma_{I_s}\right)$$

$$\tilde{R}(I_s)$$

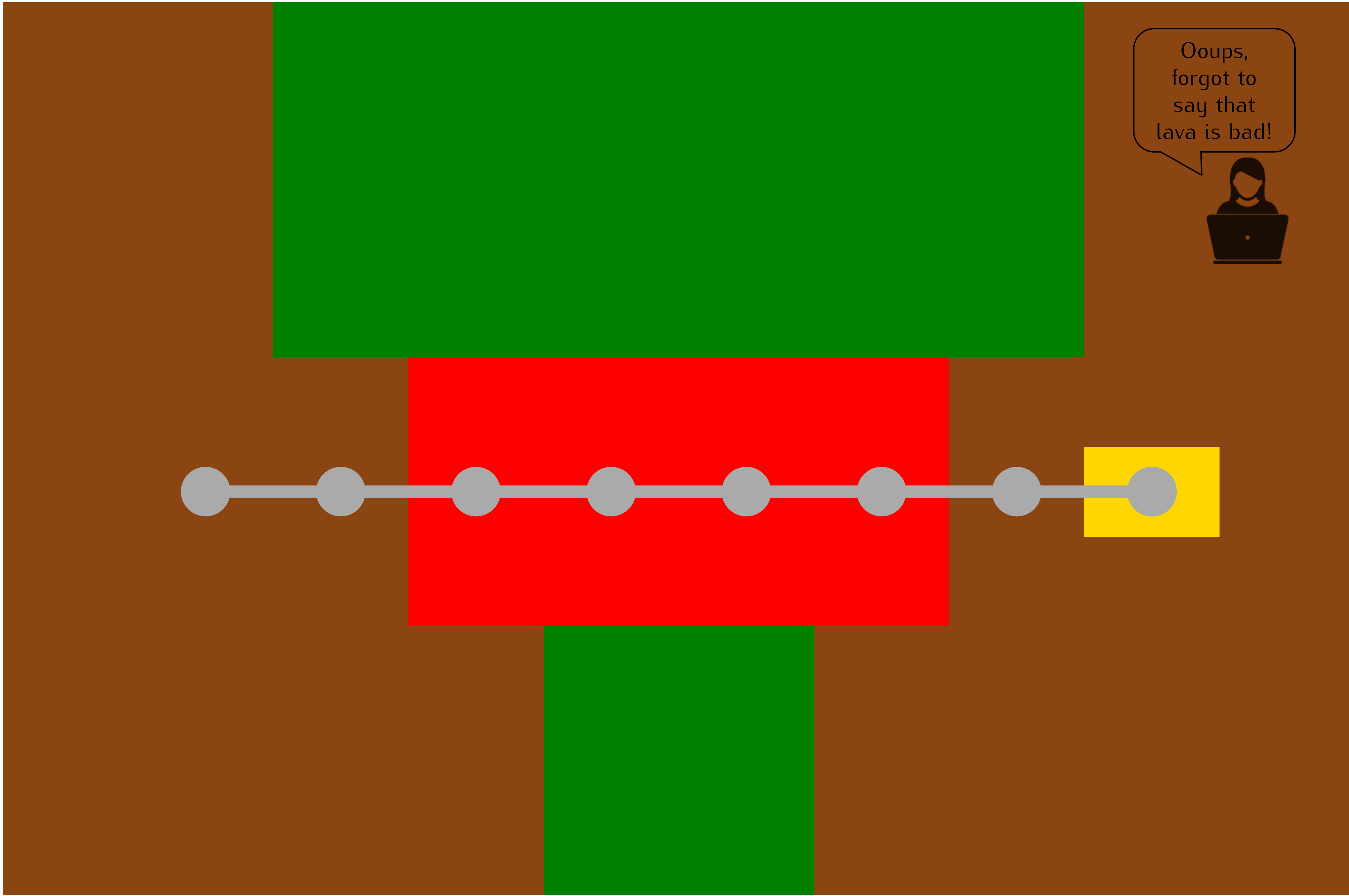


$$\tilde{\theta}^T \phi_s = \tilde{R}(I_s)$$

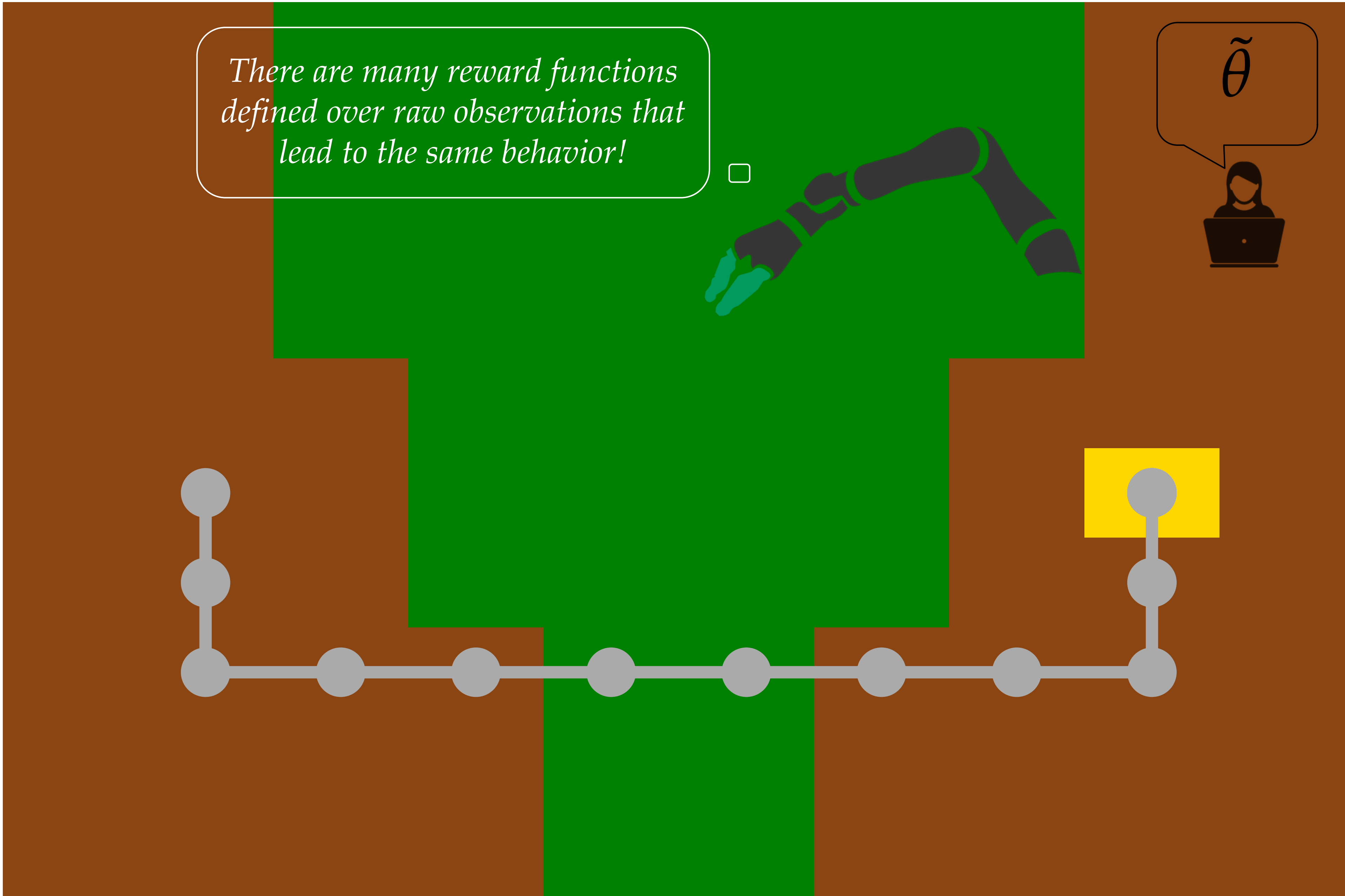
$M_{test}$



$M_{test}$ $\tilde{\theta}$

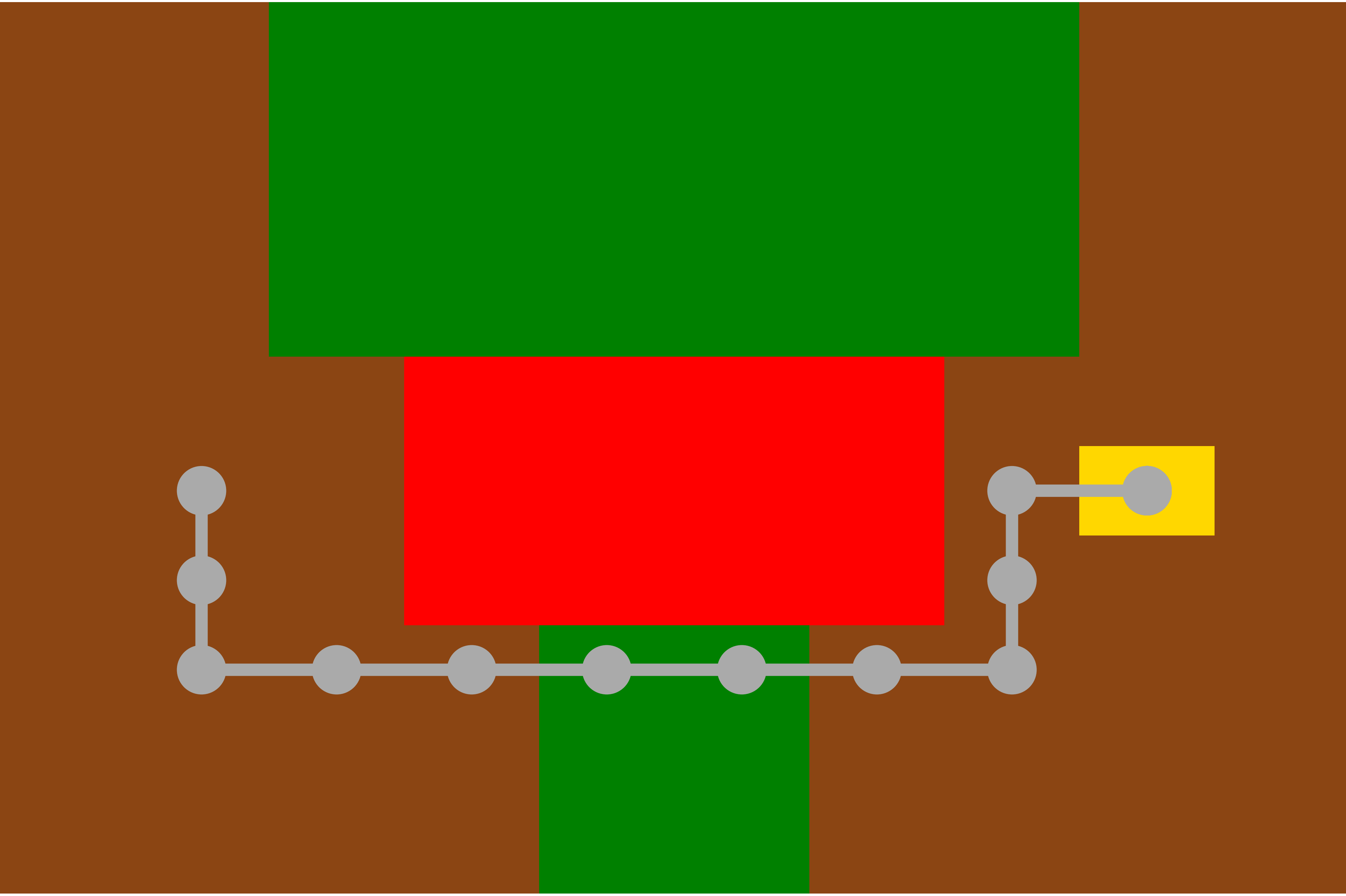


$M_{train}$

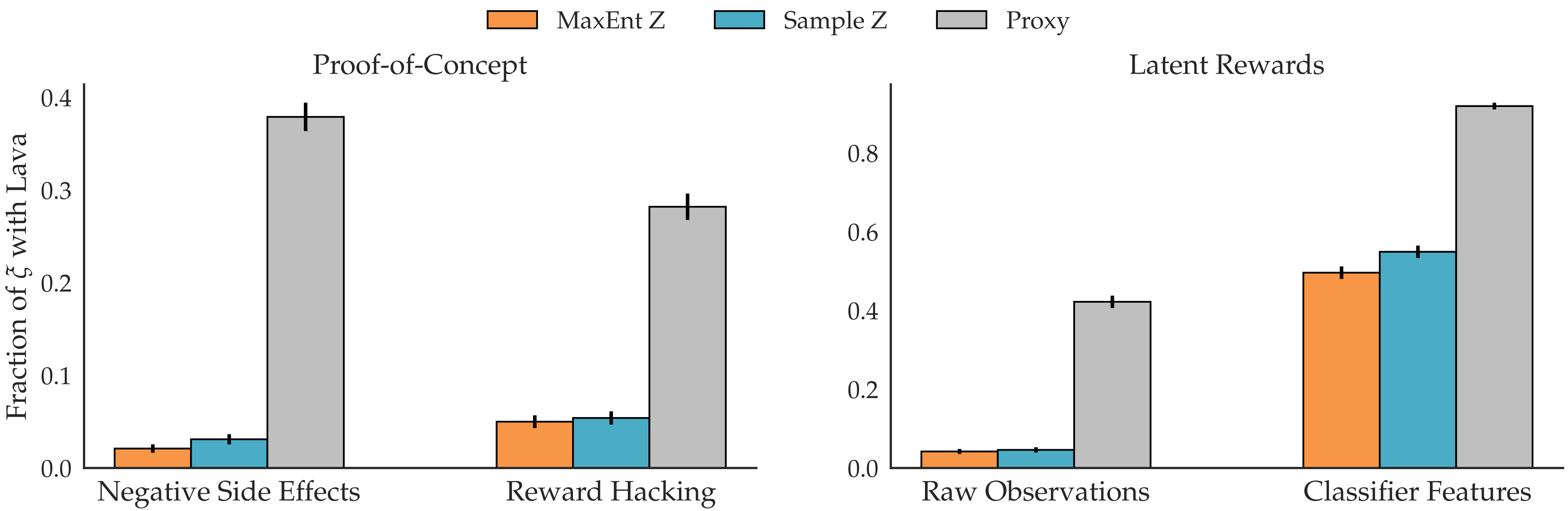


There are many reward functions defined over raw observations that lead to the same behavior!
$\tilde{\theta}$

$M_{test}$



# The agent can avoid unintended consequences, even when the features that matter are latent!

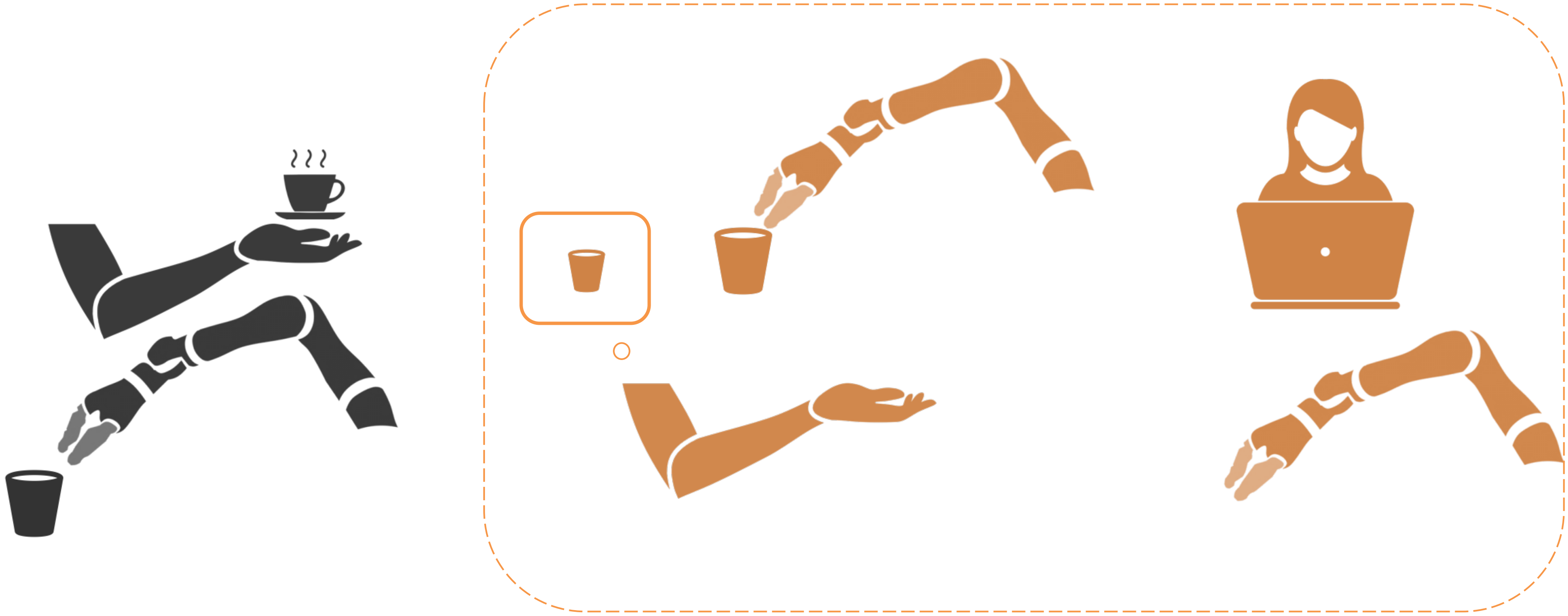


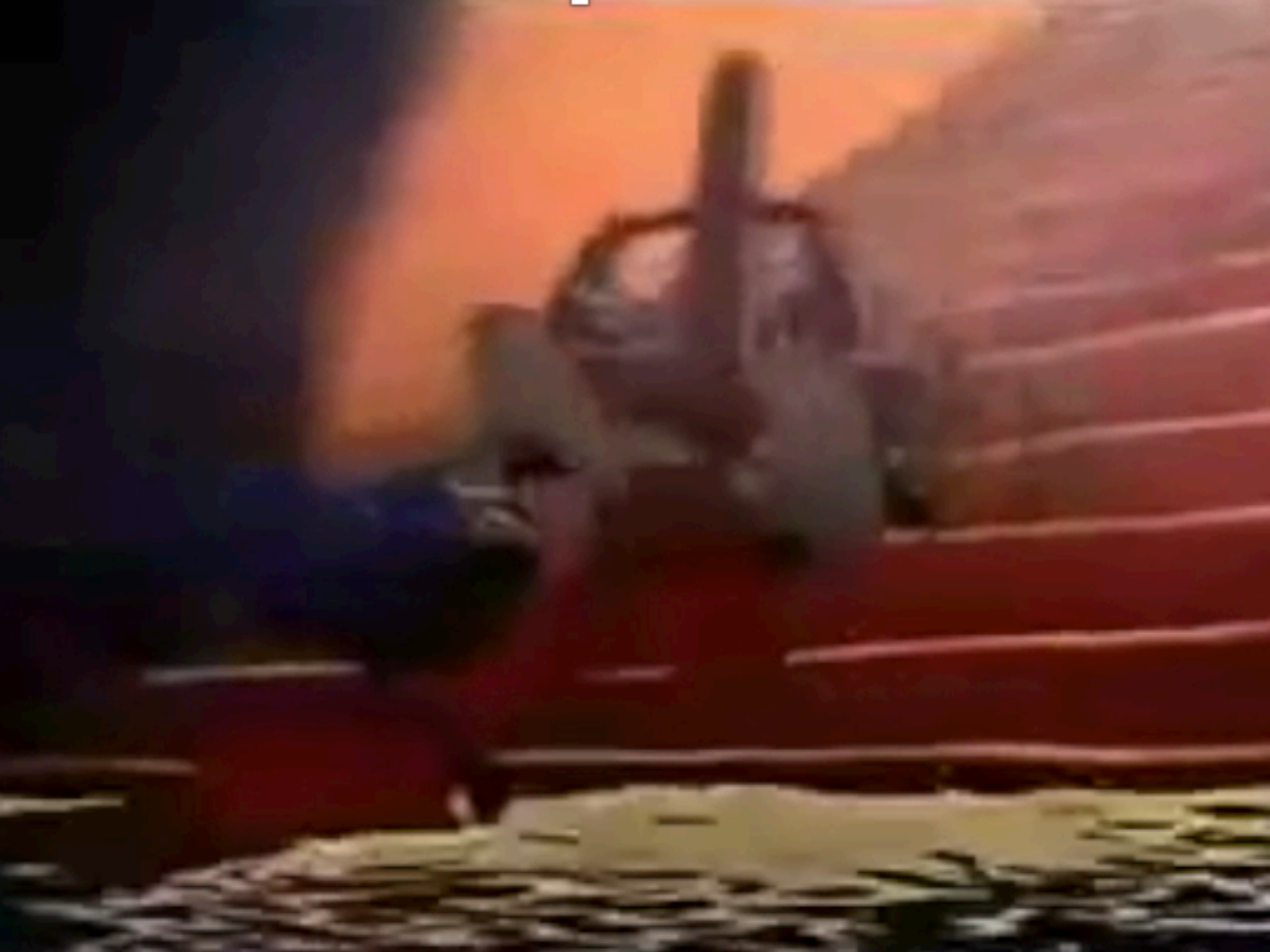
# Simplifying motion planning cost tuning

# Simplifying Reward Design through Divide-and-Conquer

Robotics: Science and Systems, 2018

Specified rewards are
observations about the true desired reward.



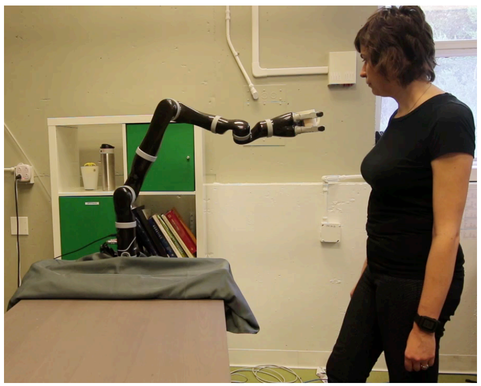


Human **guidance**

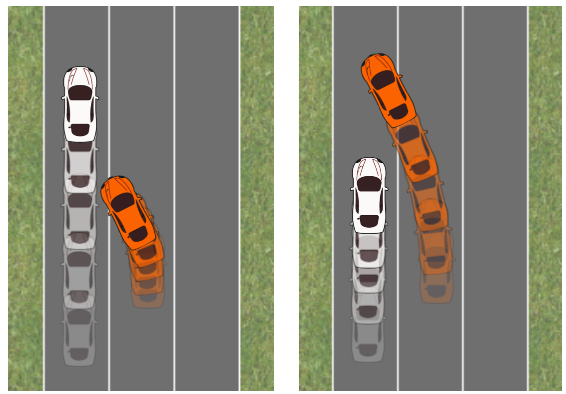
is observation about the true reward.

# Learning from rich guidance modalities

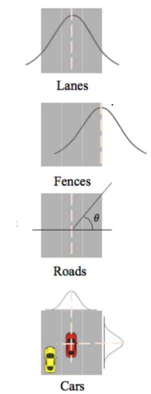
$$b'(\theta) \propto \prod P(u_H | x, \theta) b(\theta)$$



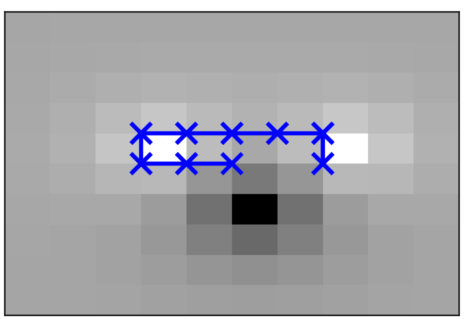
Corrections [CoRL'17]



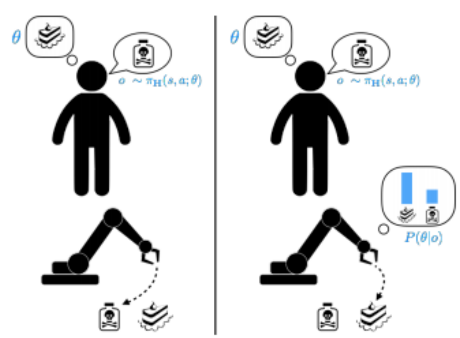
Comparisons [RSS'17]



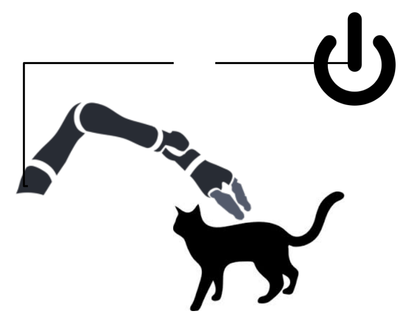

Feature queries [in review]



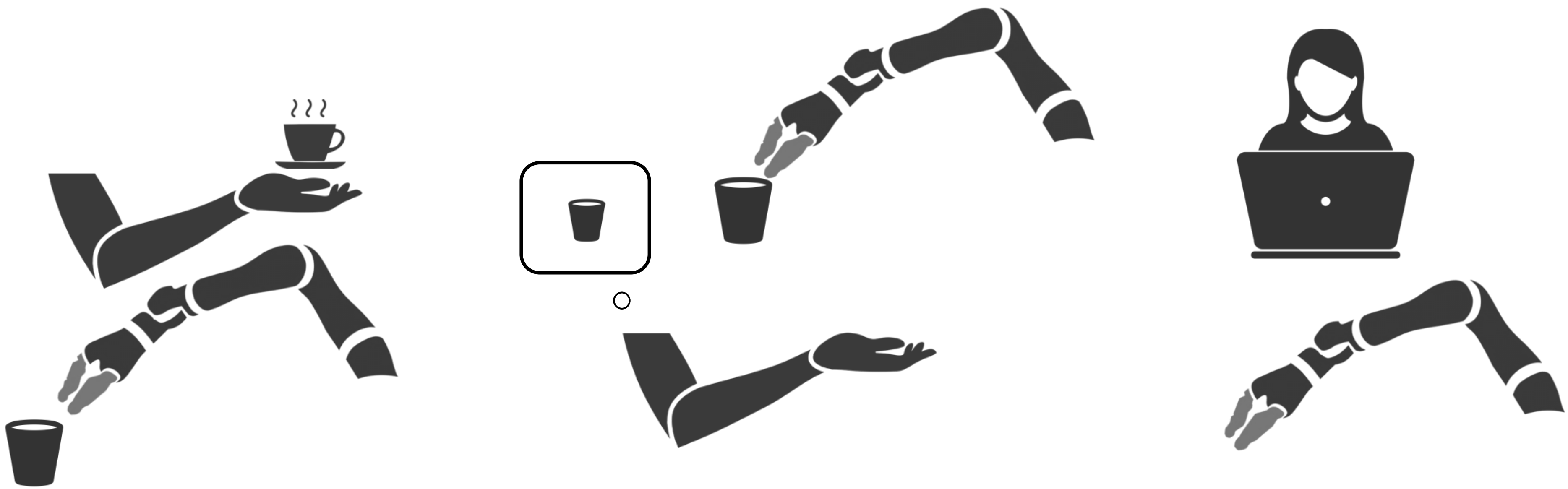
Human teaching [NIPS'16]




Orders [IJCAI'17a]



ShutDown command [IJCAI'17b]



InterACT

Laboratory

*Generating Plans that Predict Themselves [WAFR'16]*