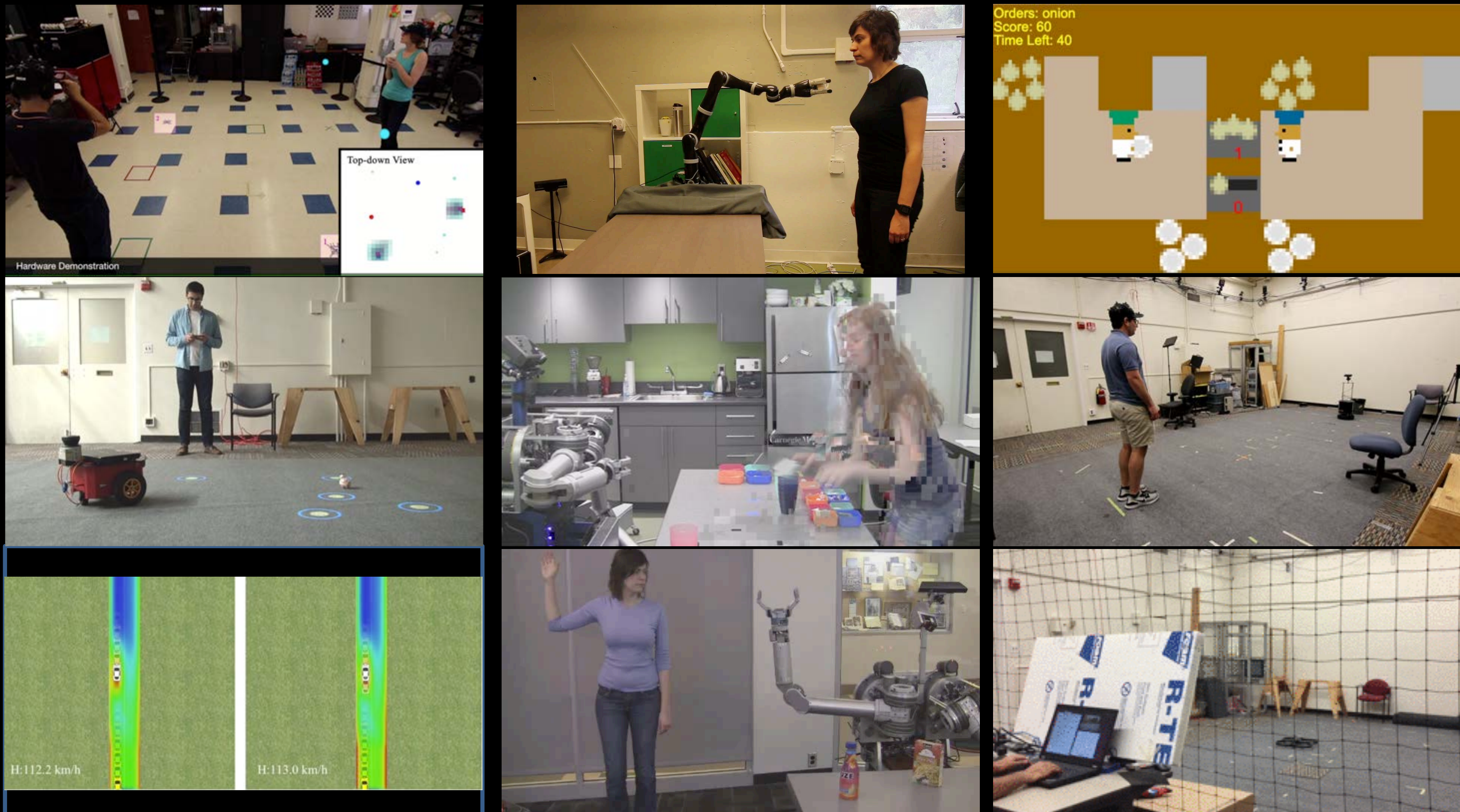
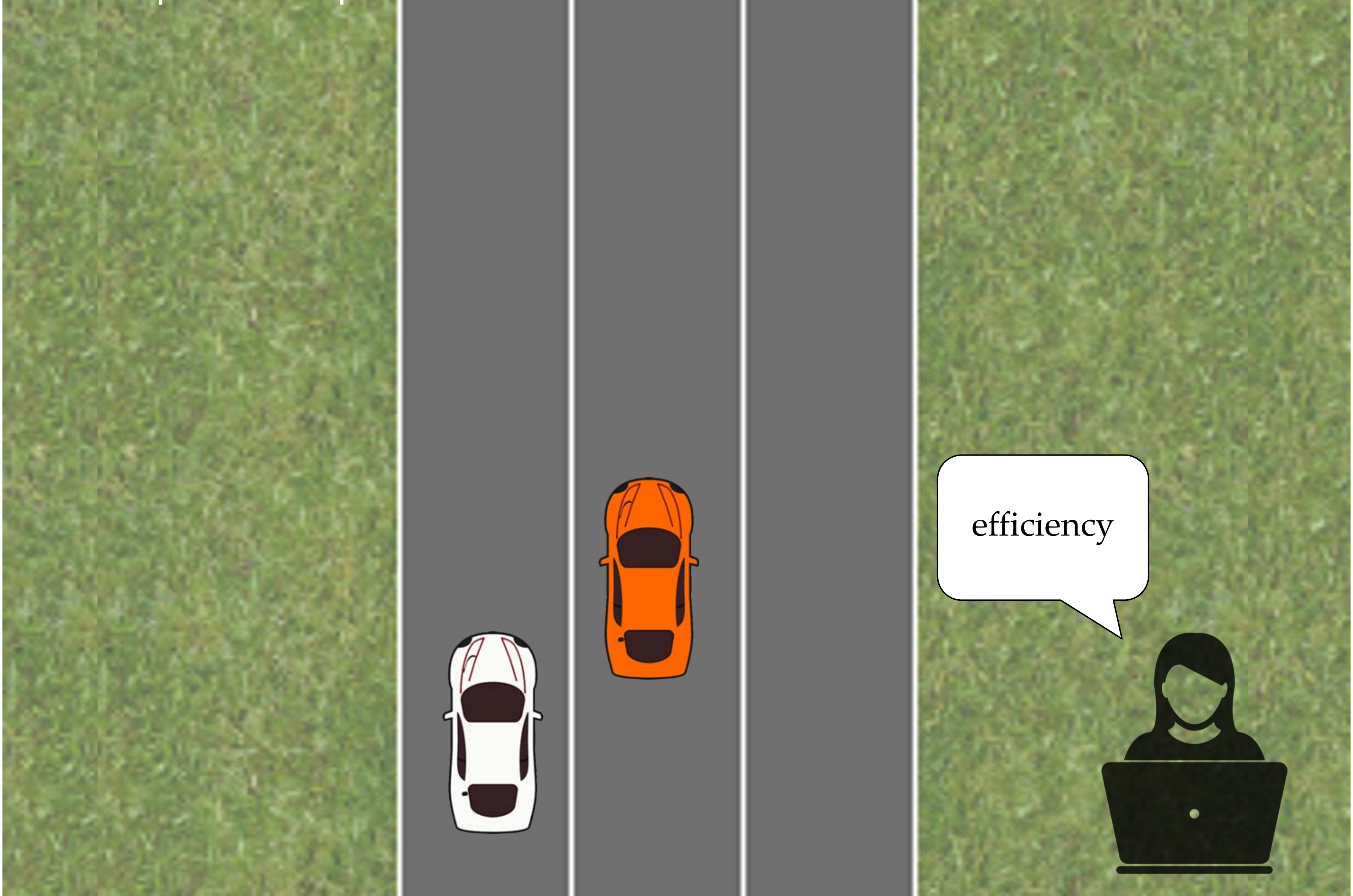
Optimizing Intended Reward Functions

Anca Dragan

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

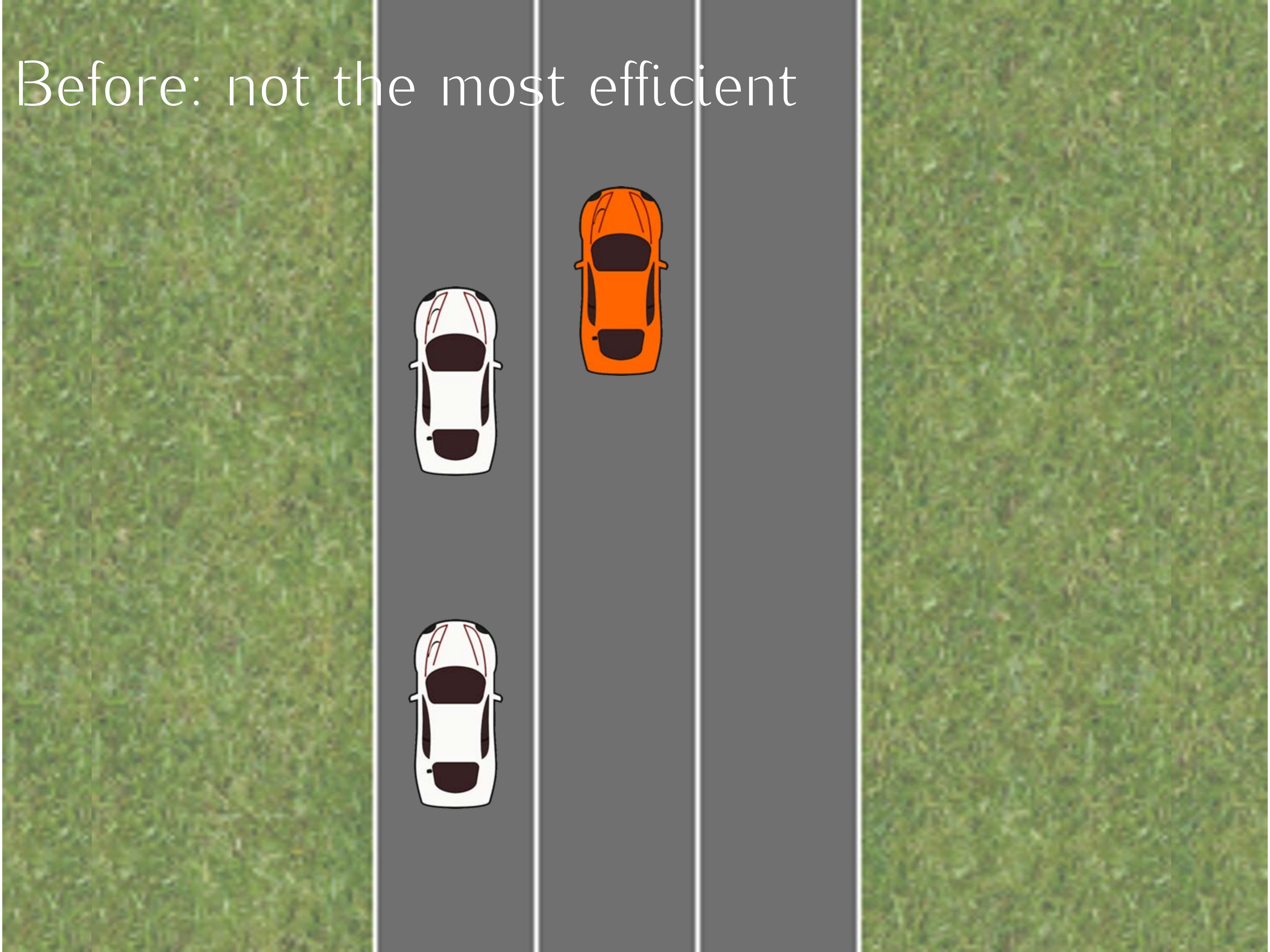

Hardware Demonstration
Top-down View
Orders: onion
Score: 60
Time Left: 40
H:112.2 km/h
H:113.0 km/h

Example: Capture influence on human action



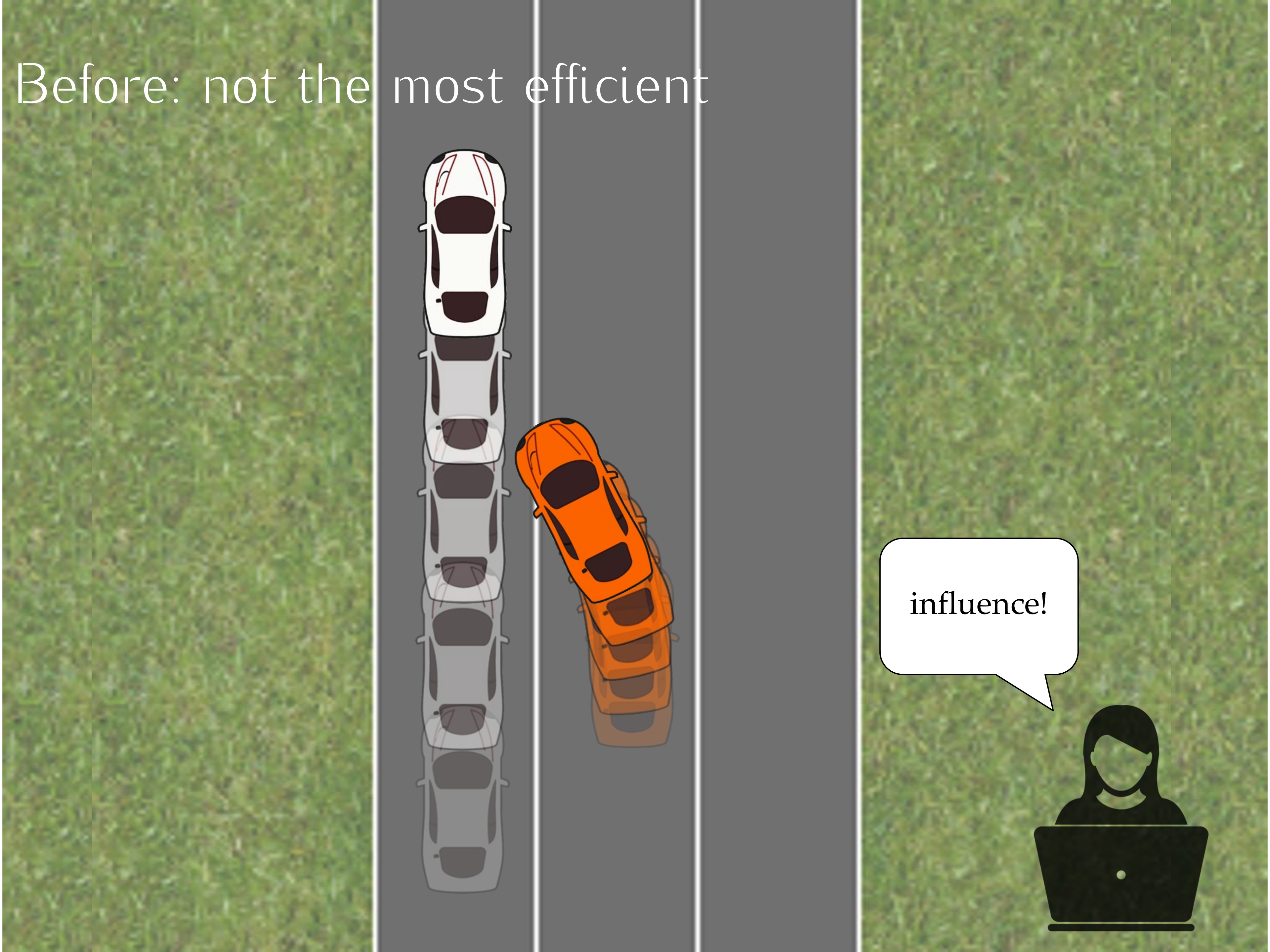

Before: not the most efficient

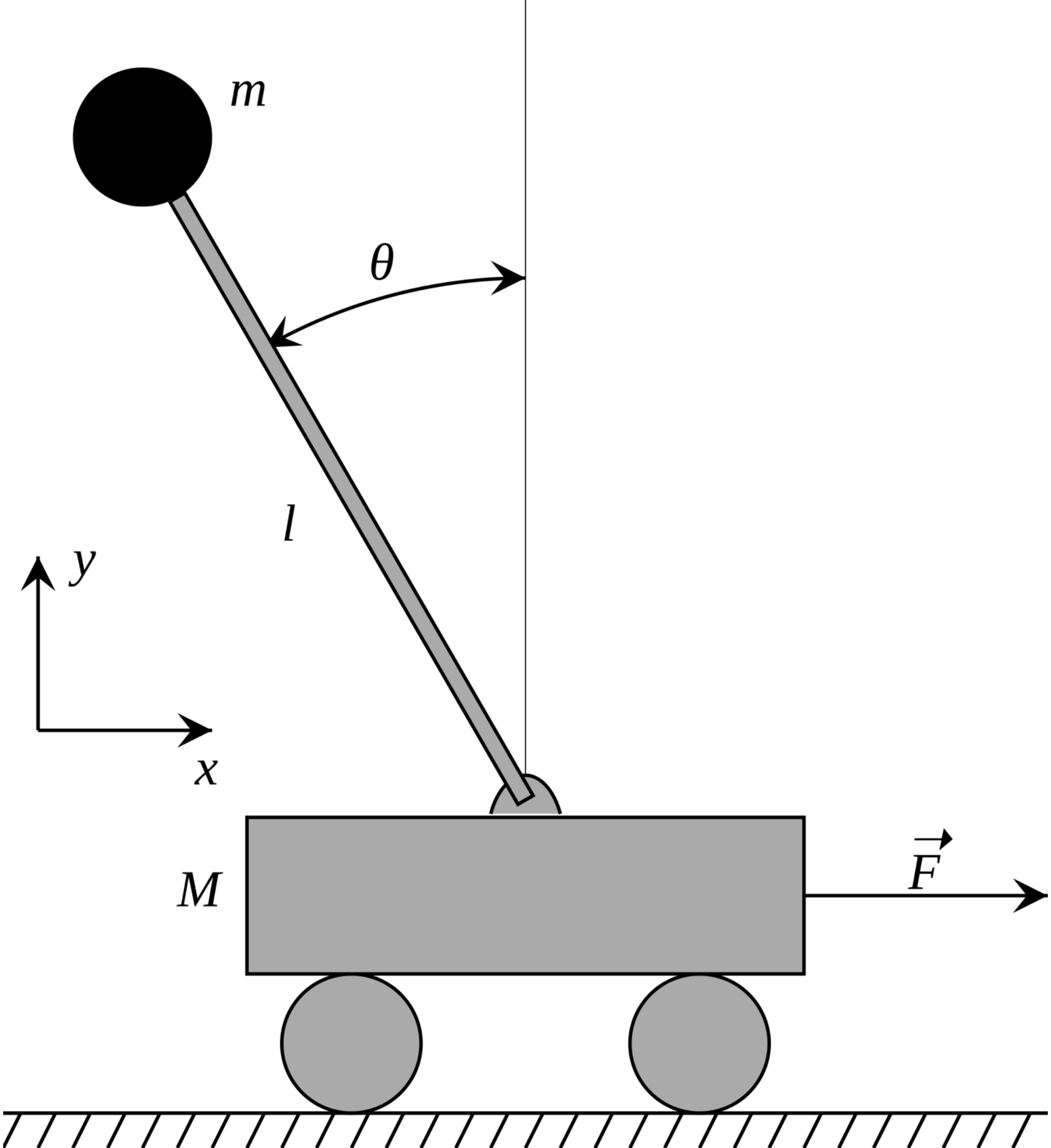
Before: not the most efficient




THE POWER YOU CAN TRUST
FORTUNER
CIVIC

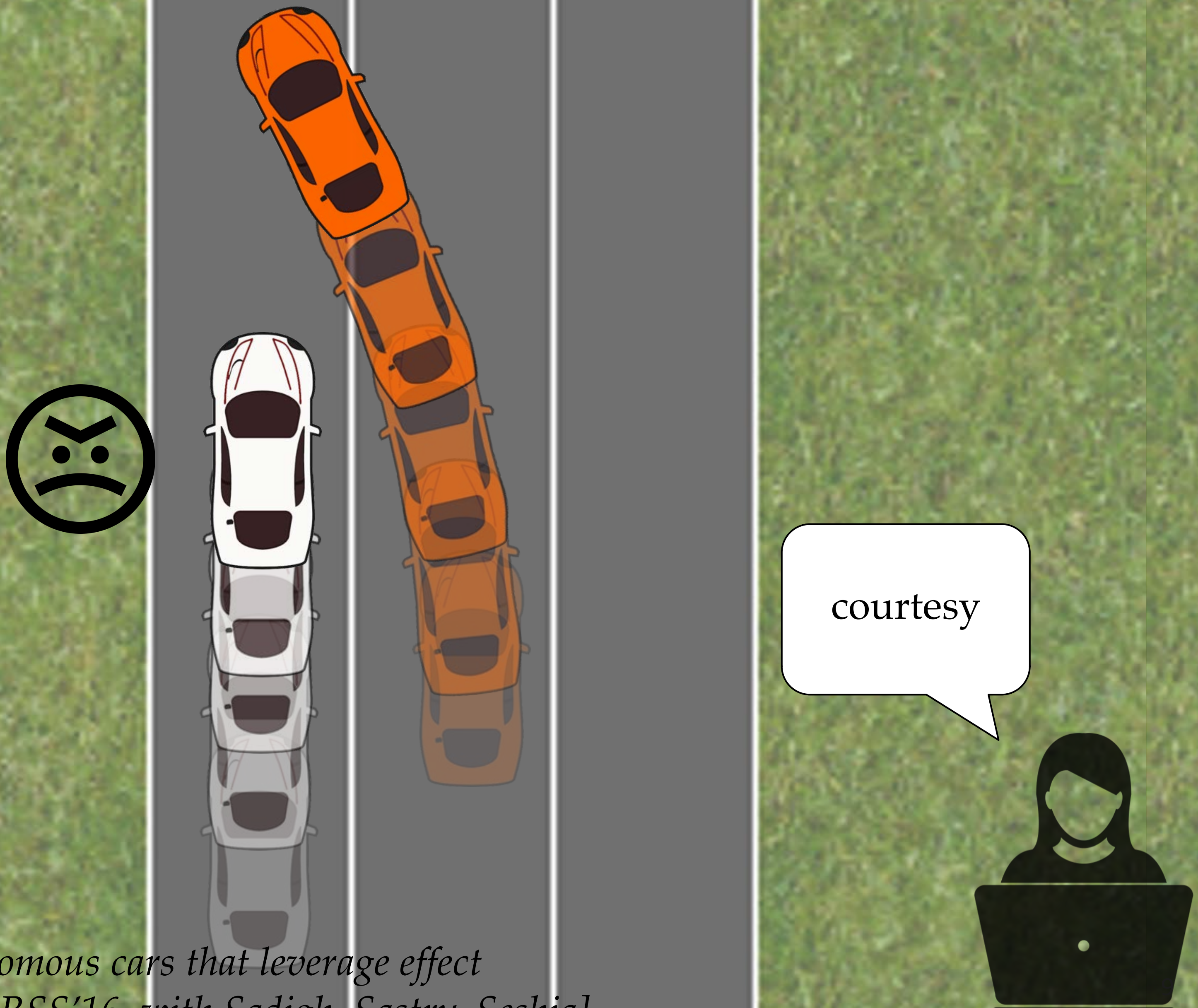
Before: not the most efficient





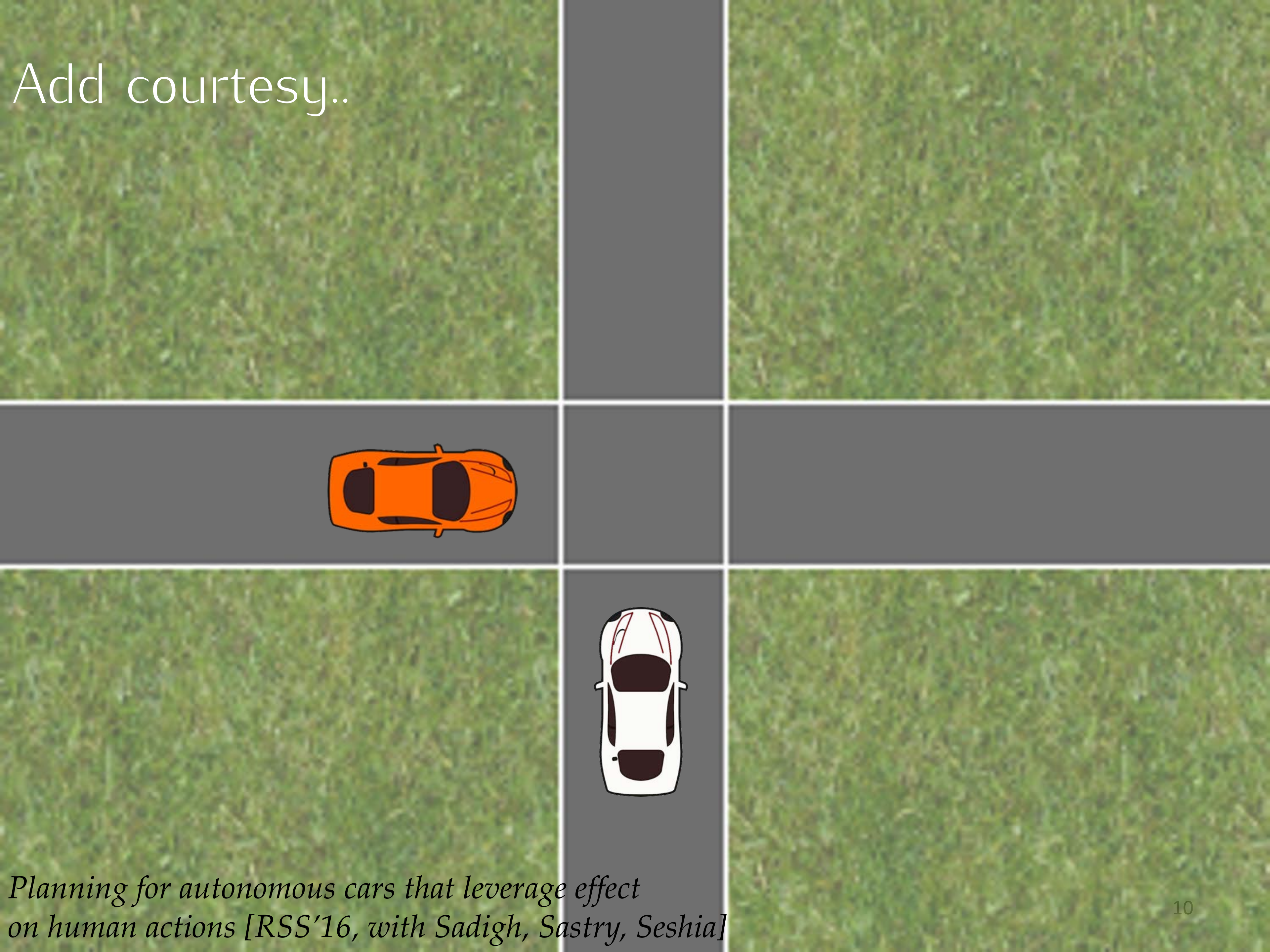
m
θ
l
y
x
M
F

Now: sometimes "too" efficient



Planning for autonomous cars that leverage effect on human actions [RSS'16, with Sadigh, Sastry, Seshia]

Add courtesy..

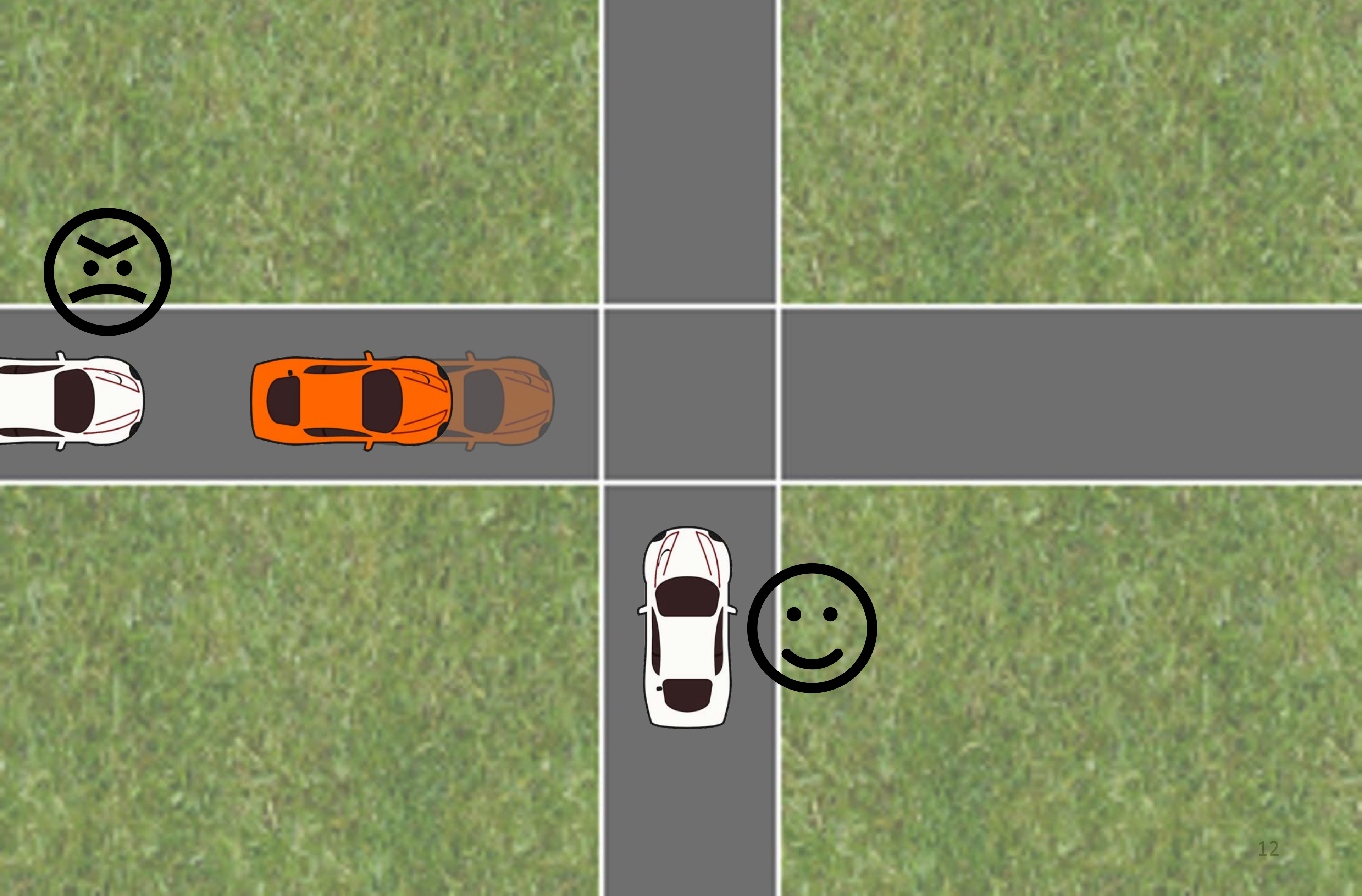


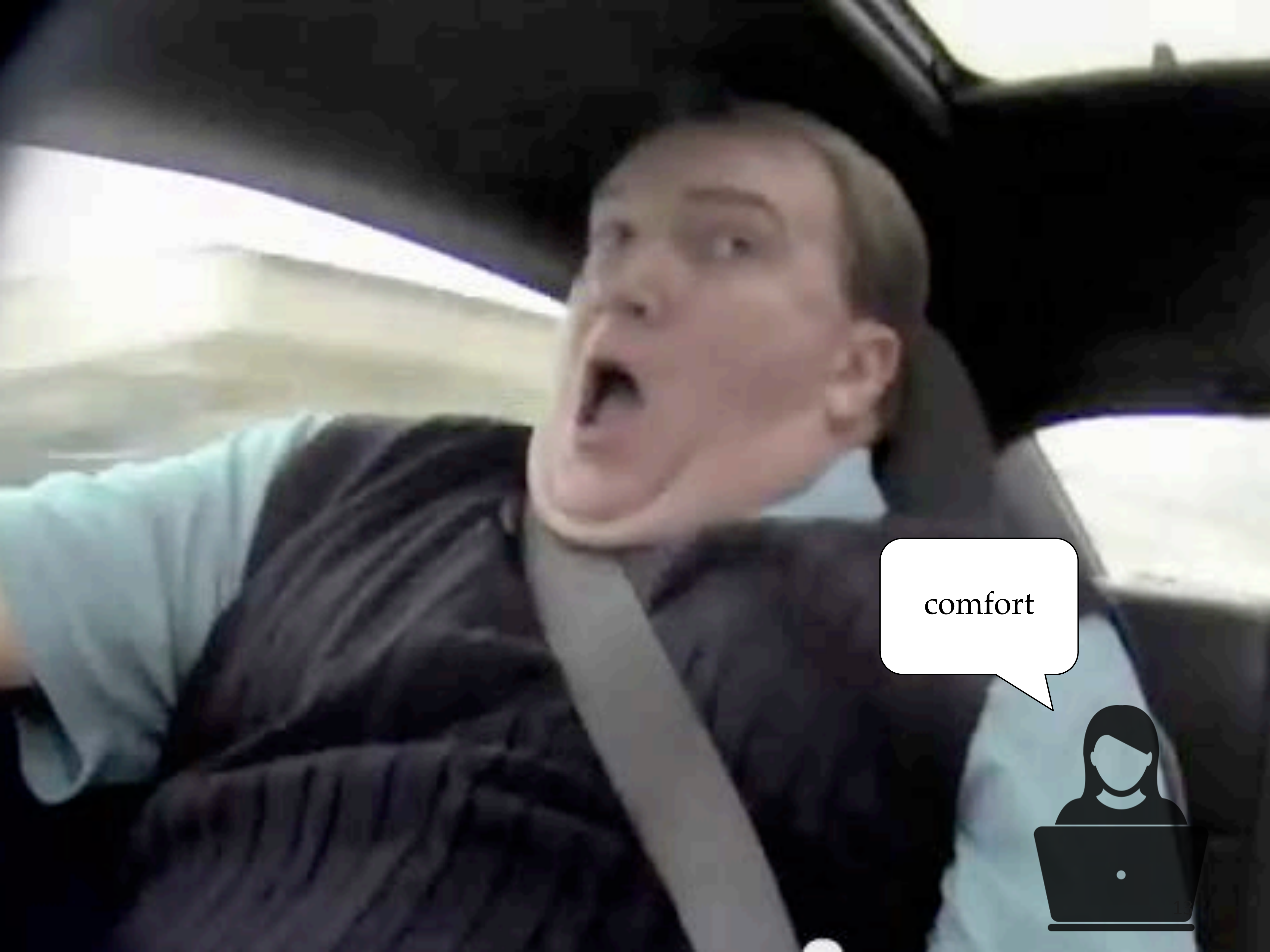
Planning for autonomous cars that leverage effect on human actions [RSS'16, with Sadigh, Sastry, Seshia]

But now, car inches backwards to get you to go!

Planning for autonomous cars that leverage effect on human actions [RSS'16, with Sadigh, Sastry, Seshia]

But now, car inches backwards to get you to go!




comfort

optimization, search, constraint satisfaction, satisficing, RL…

task specification $\longrightarrow$ behavior

cost, reward, goal, loss, constraints,…

?????

optimization, search, constraint satisfaction, satisficing, RL…

task specification ⟶ behavior

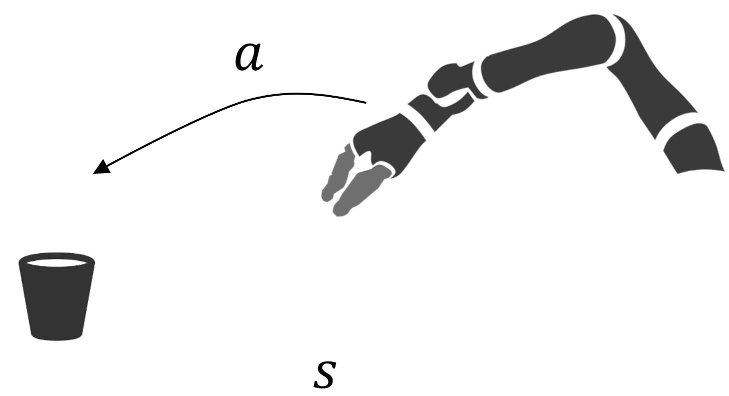
cost, reward, goal, loss, constraints,…

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

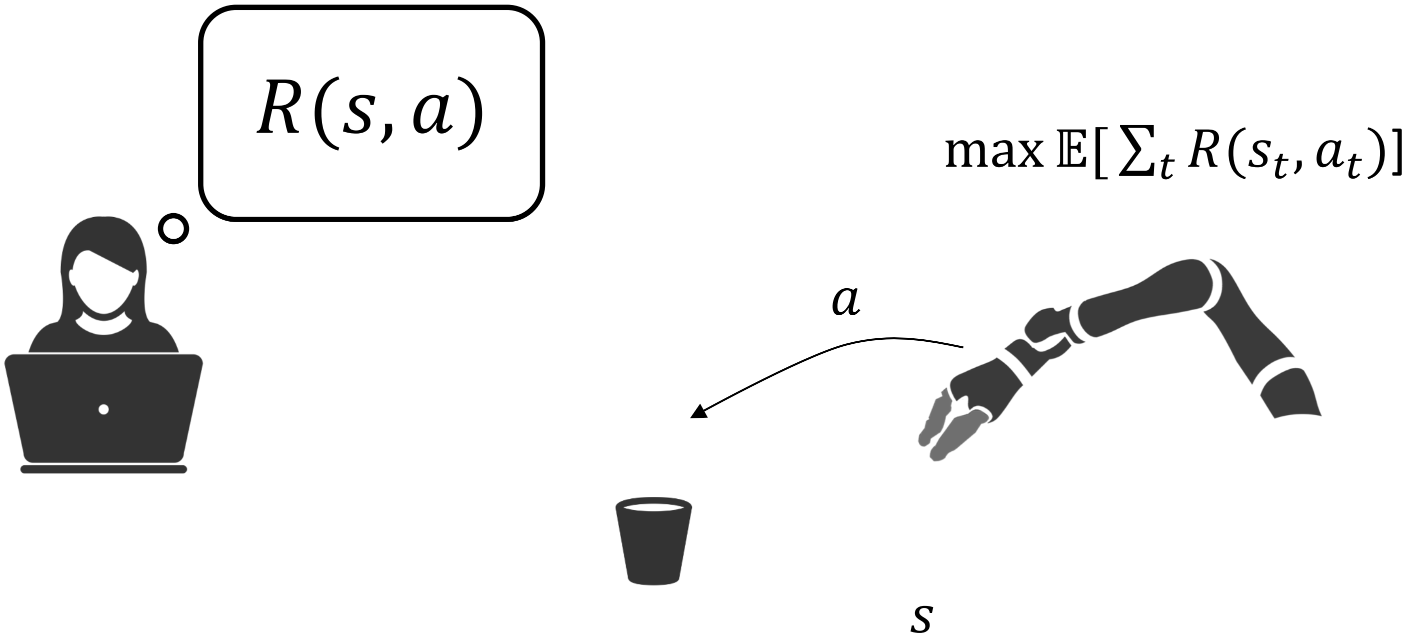
What we pretend AI is:

$$R(s, a)$$

$$\max \mathbb{E}[\sum_t R(s_t, a_t)]$$



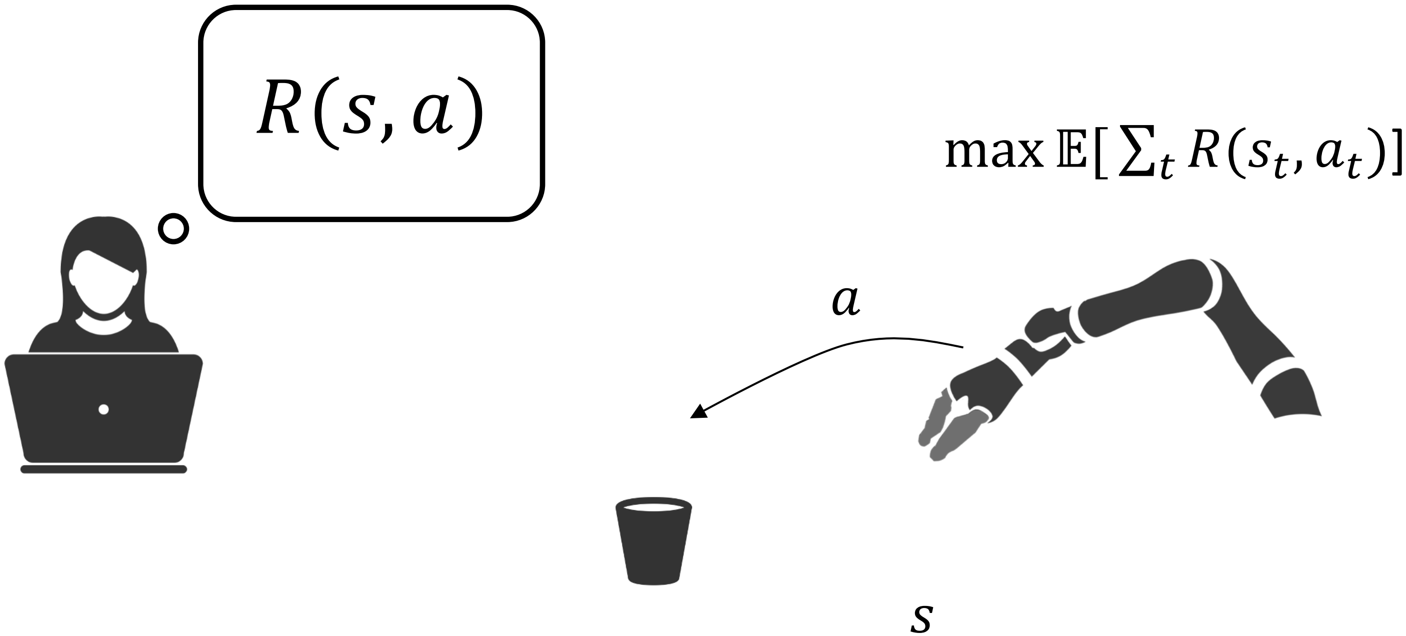
What AI actually is:



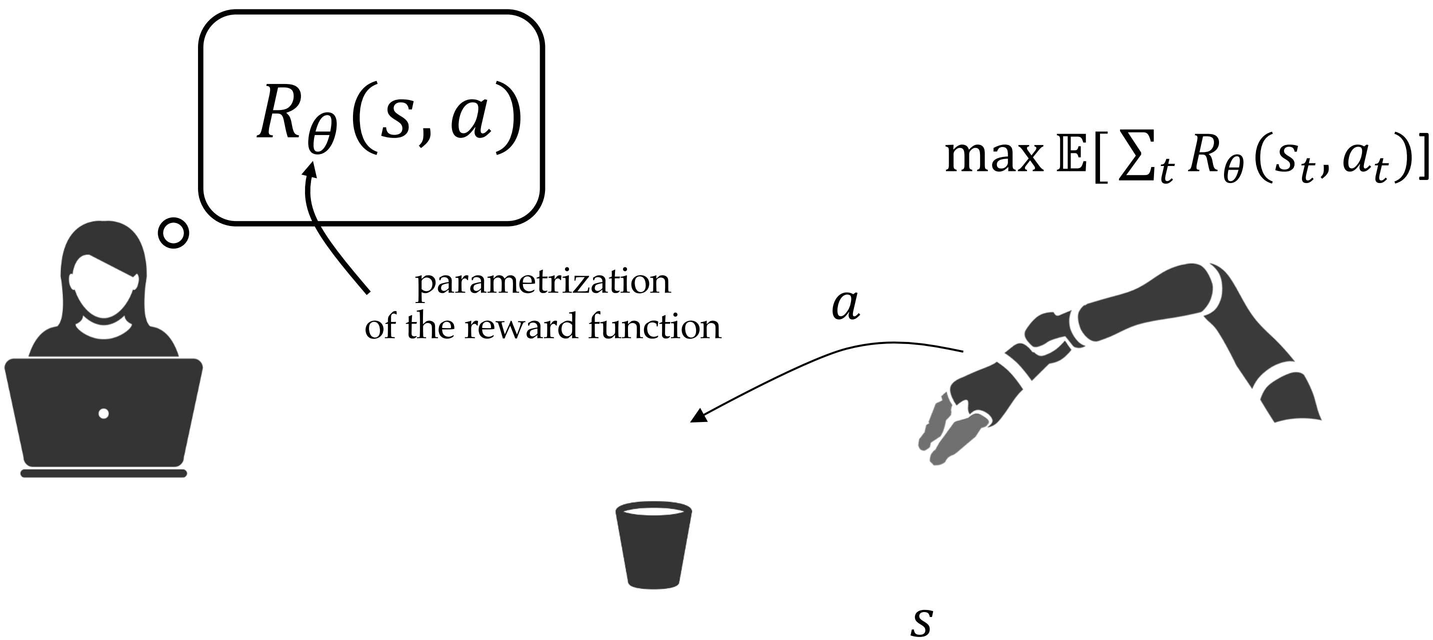
$$\text{AI} \neq \text{optimize specified reward}$$

$$\text{AI} = \text{optimize } \underline{\text{intended}} \text{ reward}$$

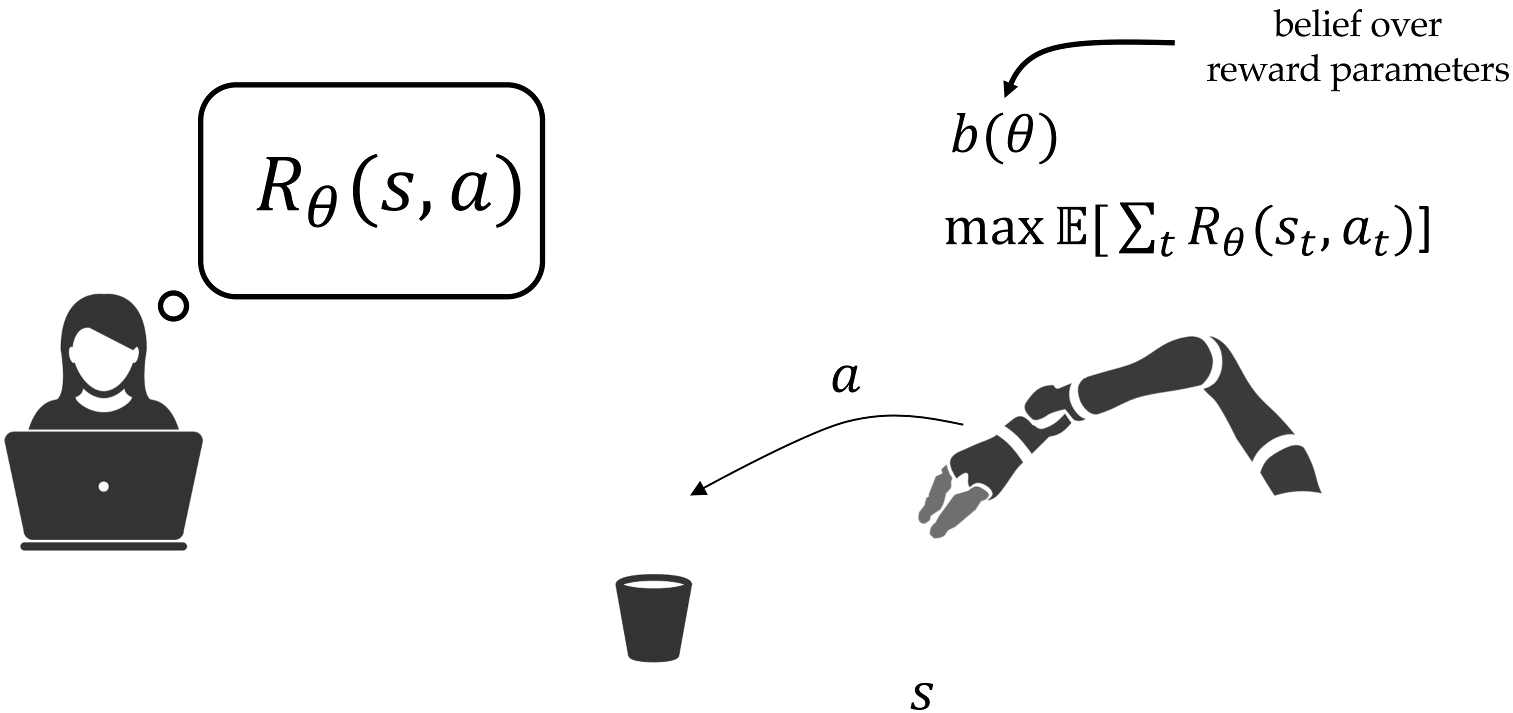
Optimize intended reward



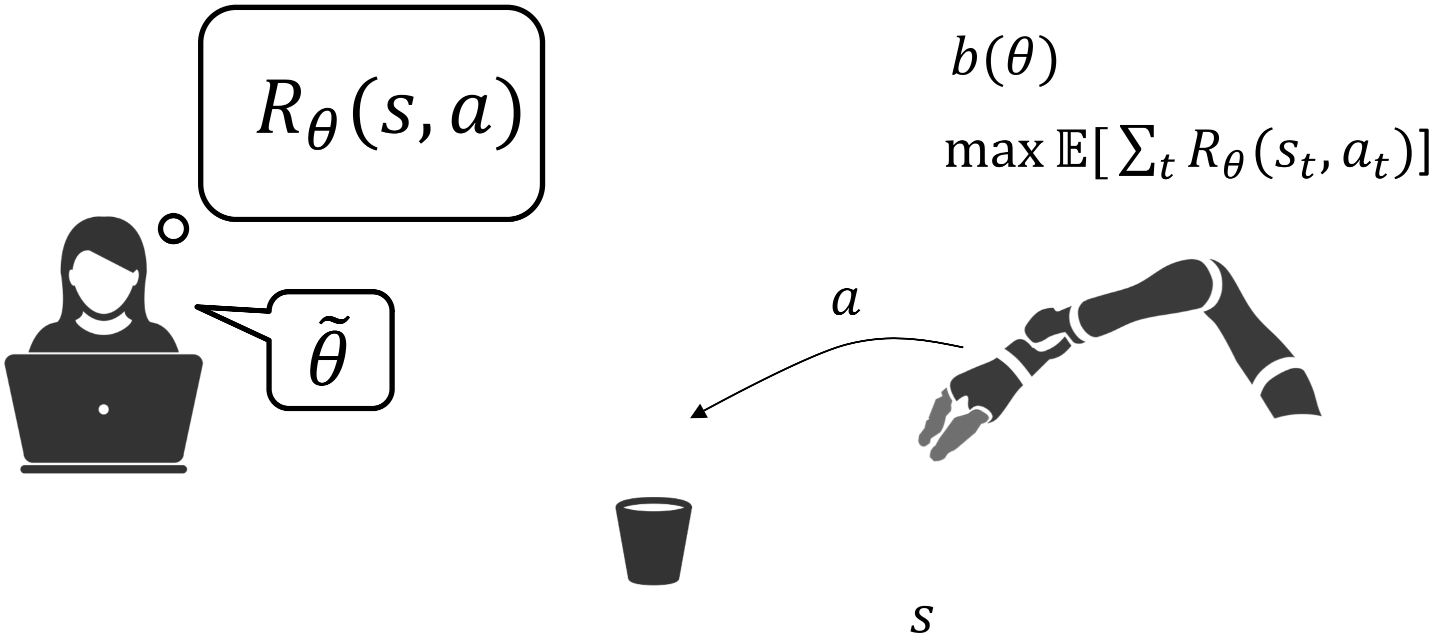
Optimize intended reward



Optimize intended reward



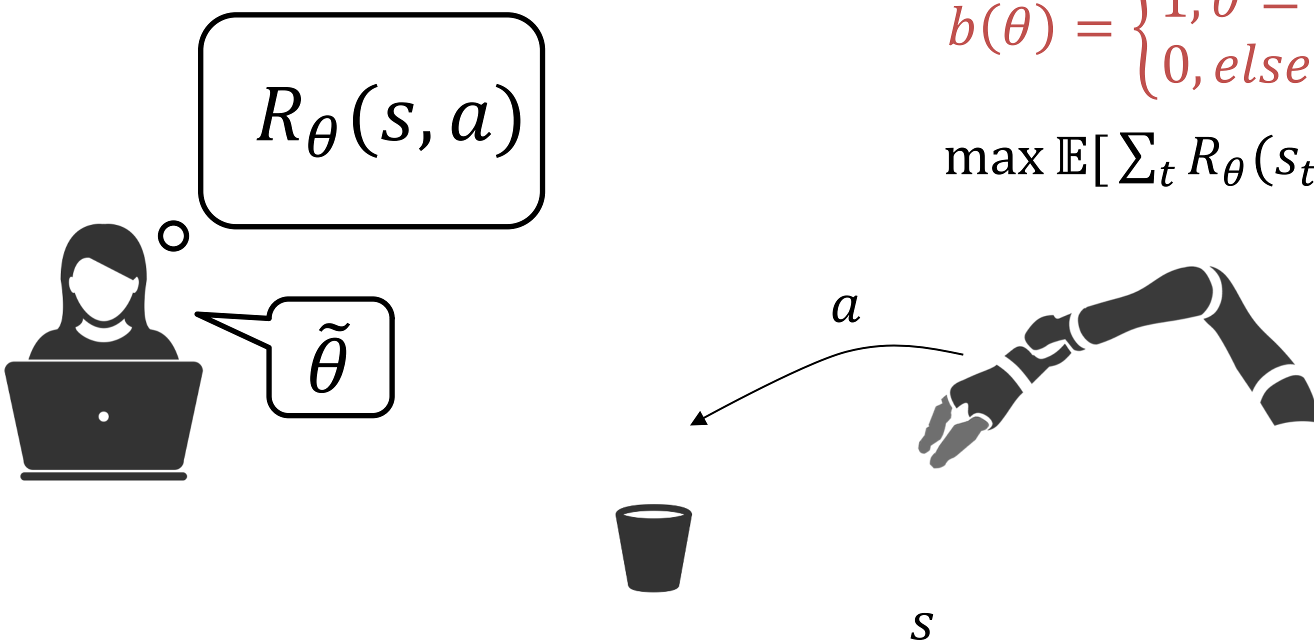
Why treat specified rewards as definition?

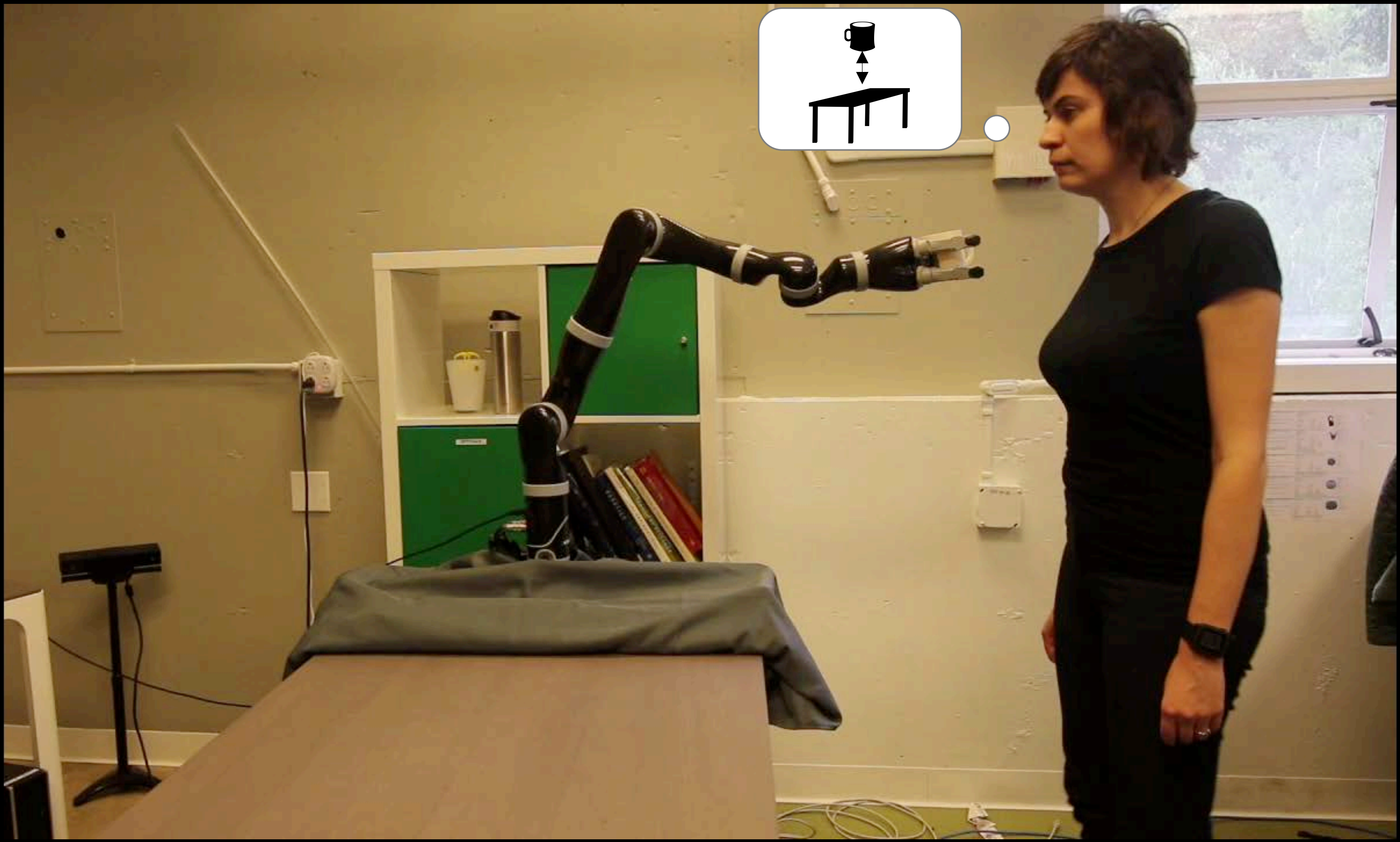


Why treat specified rewards as definition?

$$b(\theta) = \begin{cases} 1, \theta = \tilde{\theta} \\ 0, else. \end{cases}$$

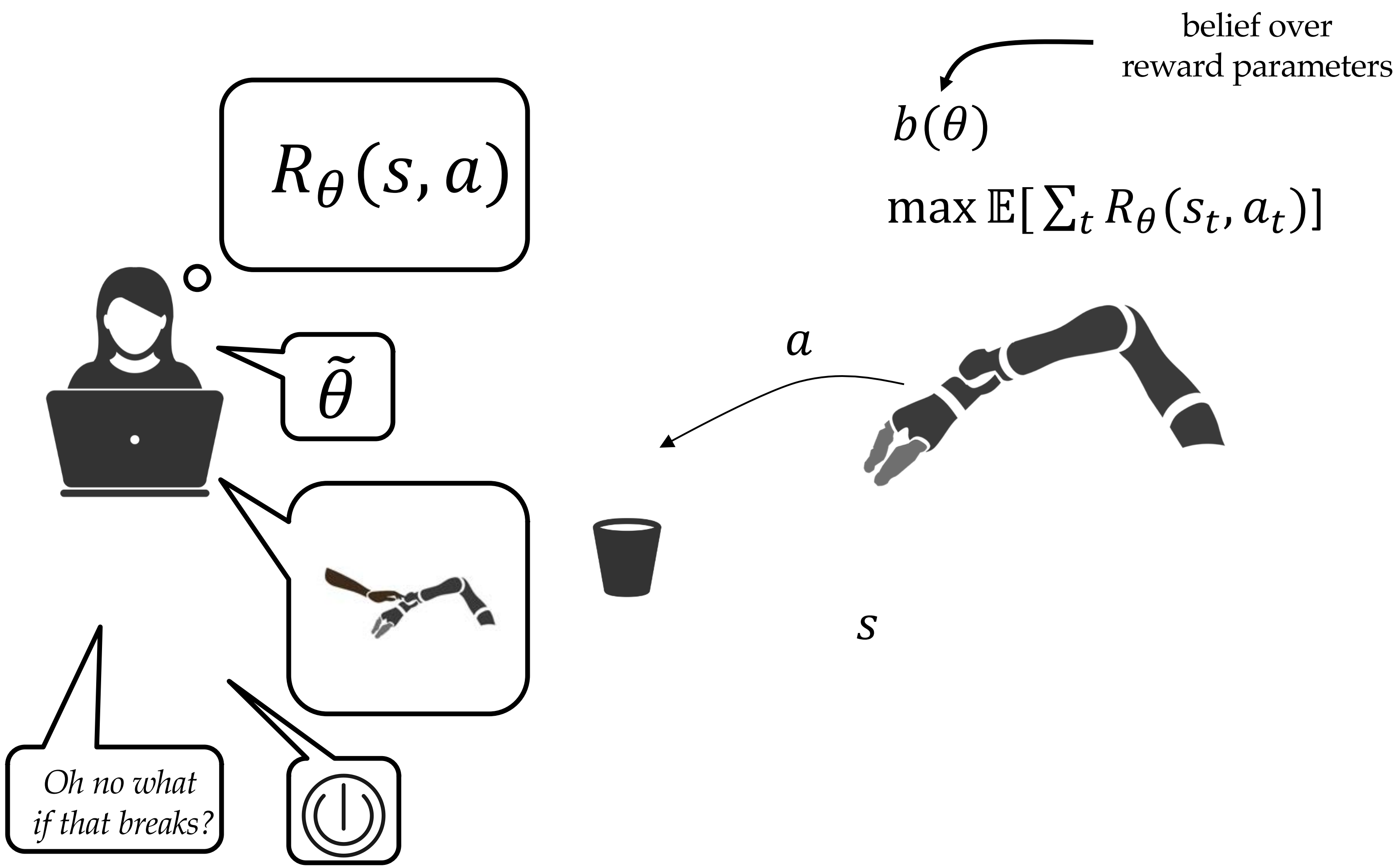
$$\max \mathbb{E}[\sum_t R_\theta(s_t, a_t)]$$

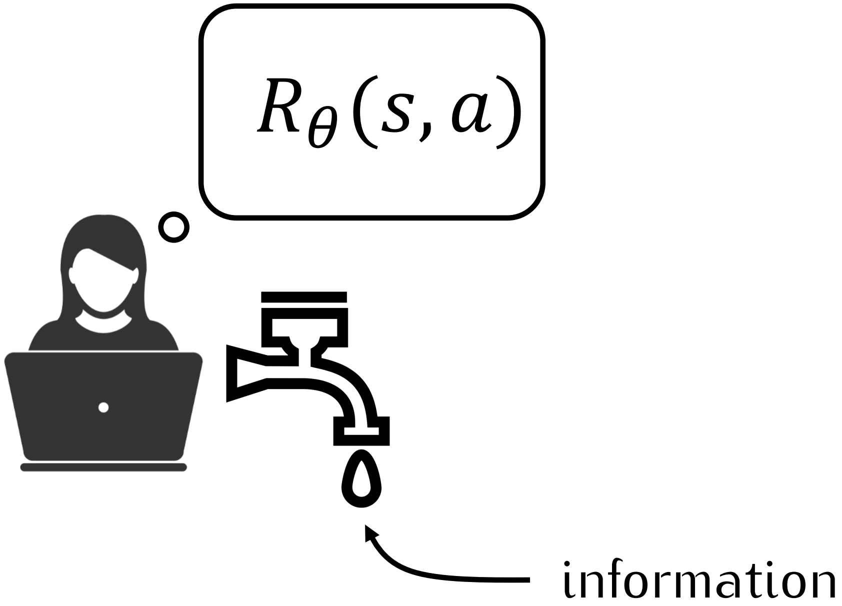




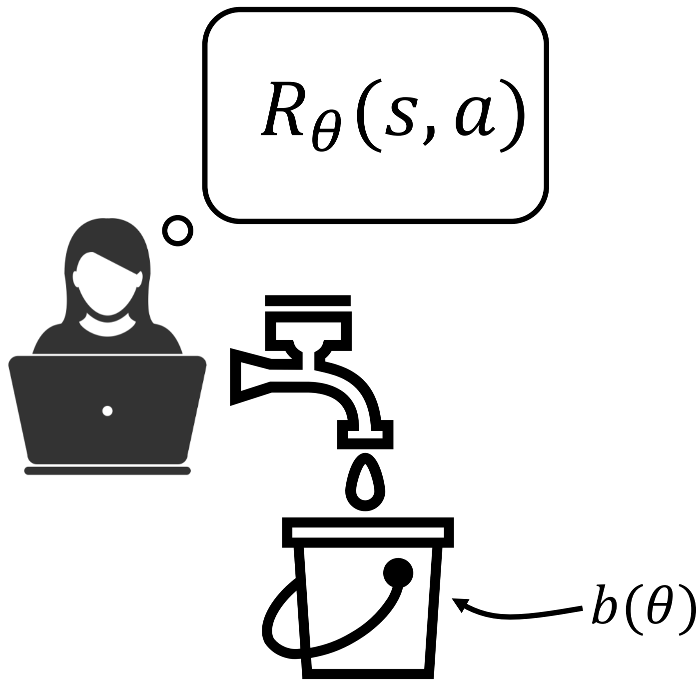
Agents overlearn from specified rewards,
but underlearn from other sources.

Optimize intended reward



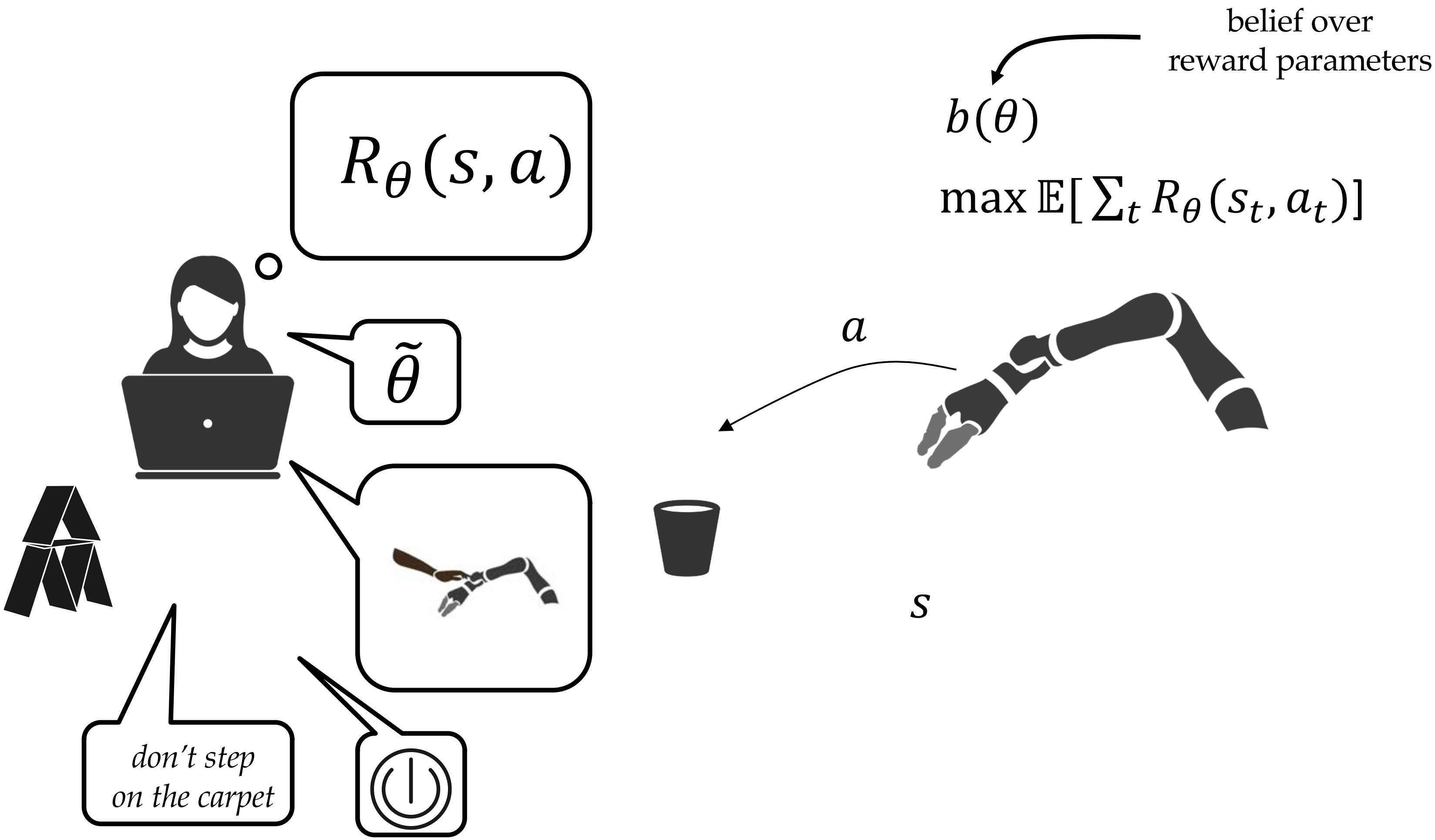


Humans leak information about the reward.

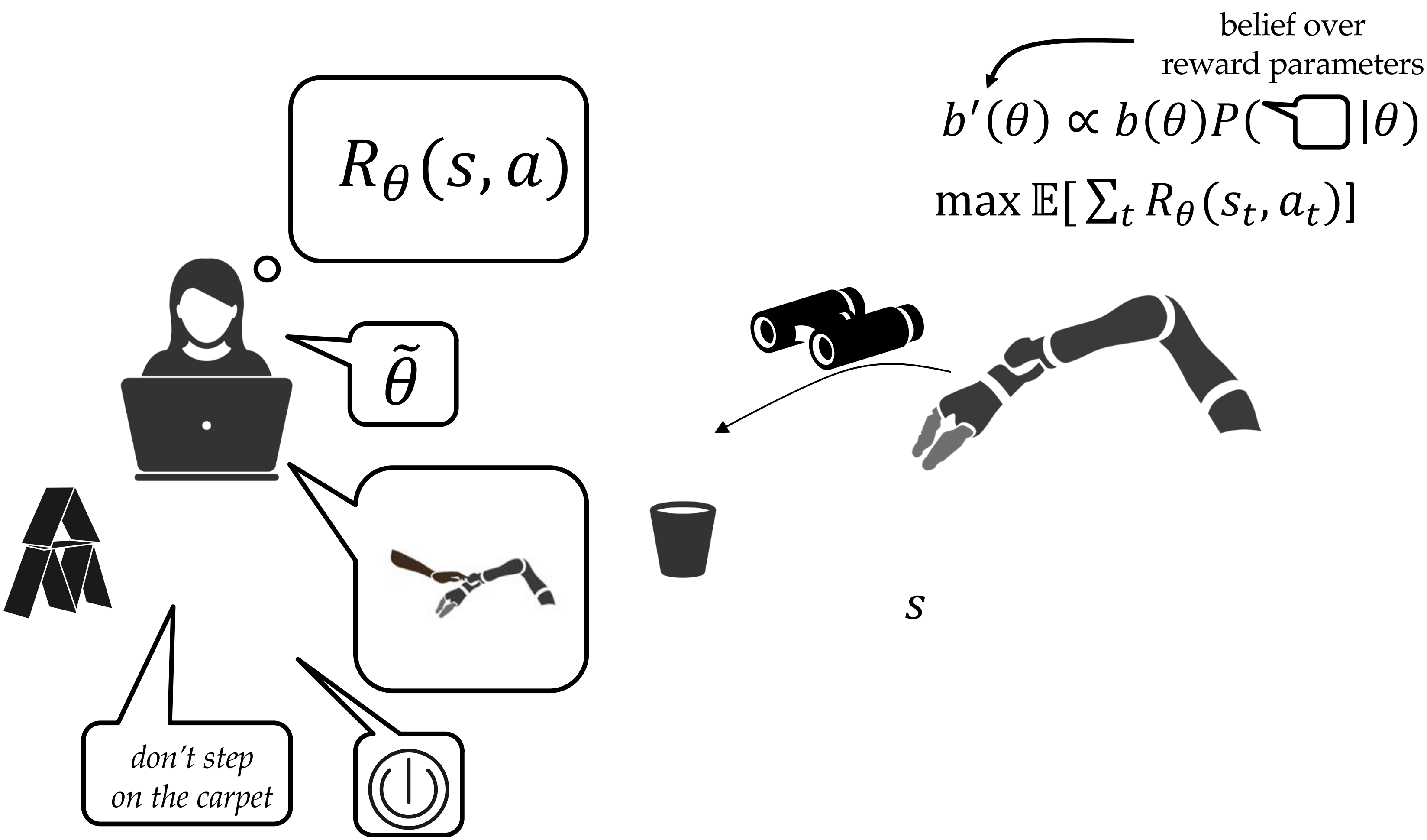


How should the robot extract
it into an updated belief?

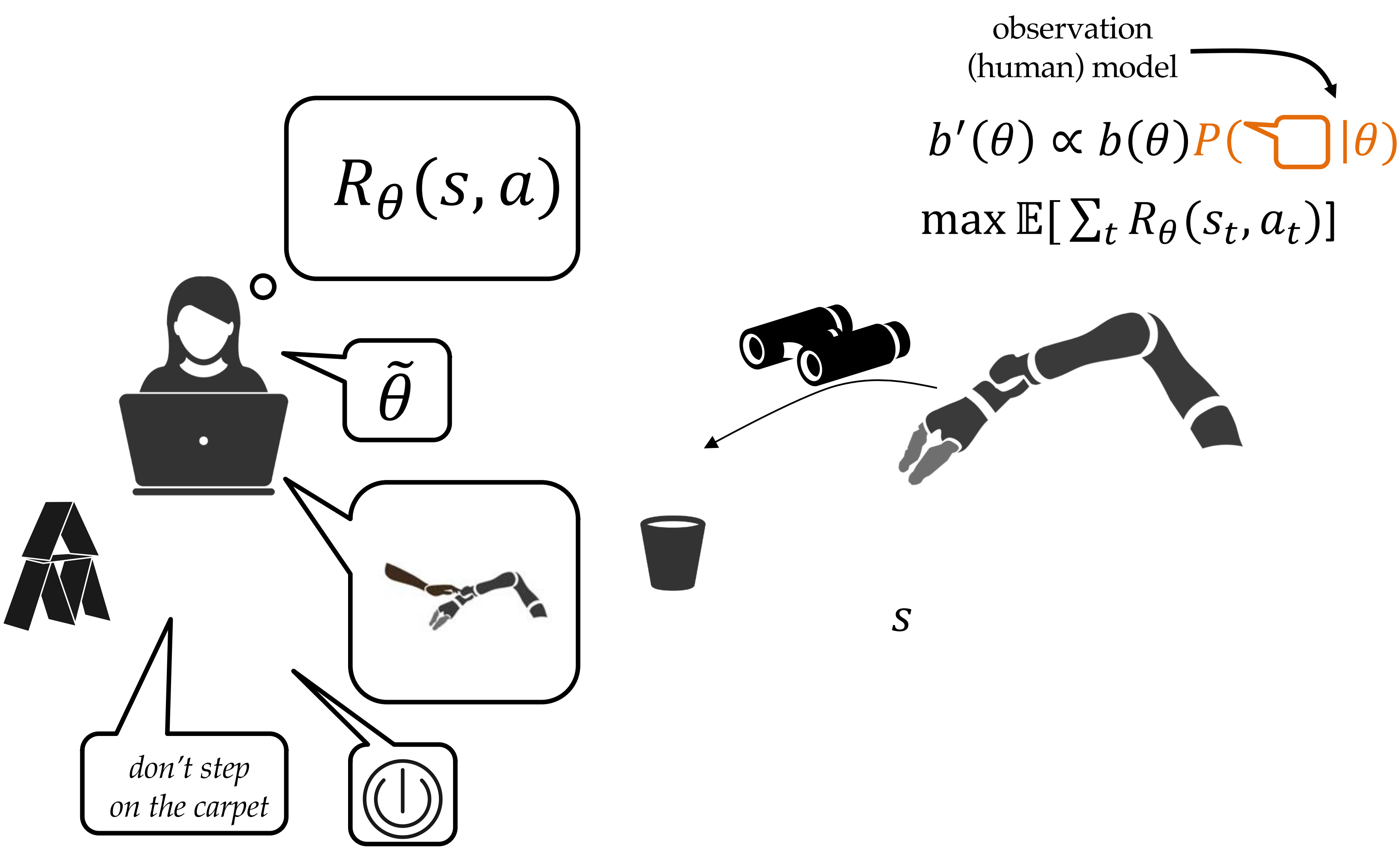
Human feedback, from specifying a reward to turning the robot off, is evidence about the intended reward.



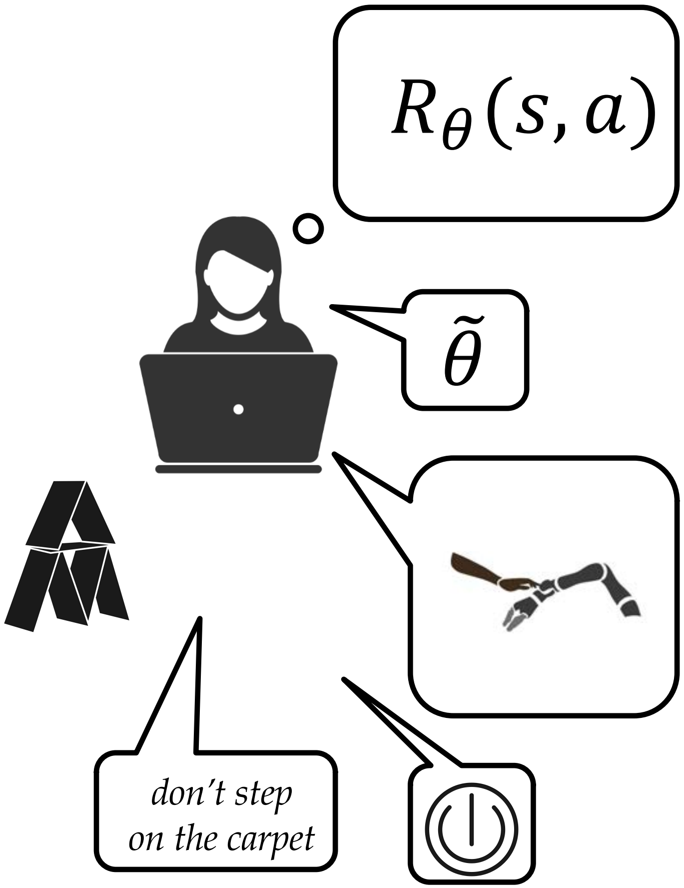
Human feedback, from specifying a reward to turning the robot off, is evidence about the intended reward.



Human feedback, from specifying a reward to turning the robot off, is evidence about the intended reward.



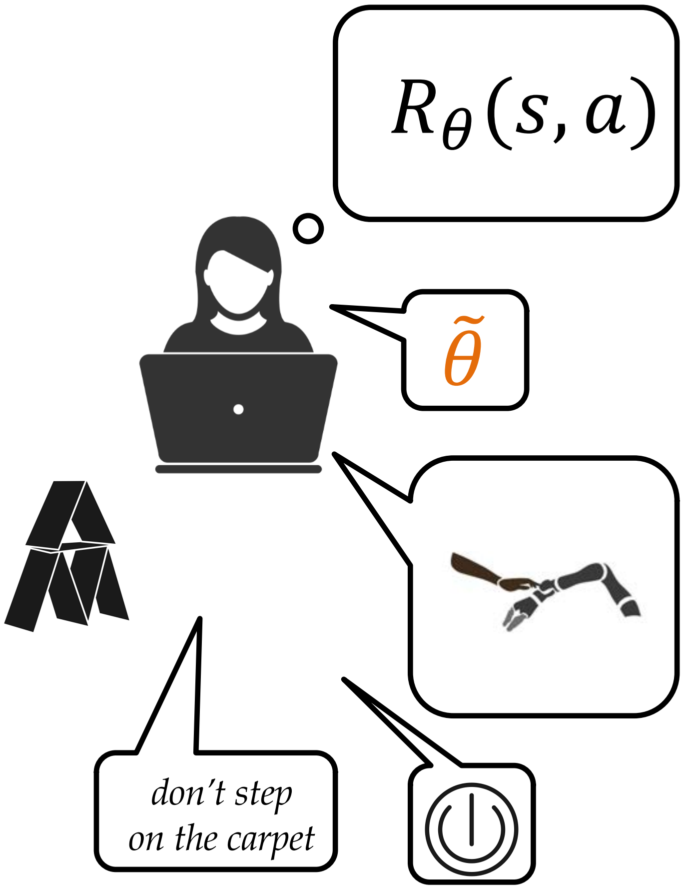
What is a human model that can be used to make sense of all these types of human feedback?



observation
(human) model

$$b'(\theta) \propto b(\theta) P(\quad | \theta)$$

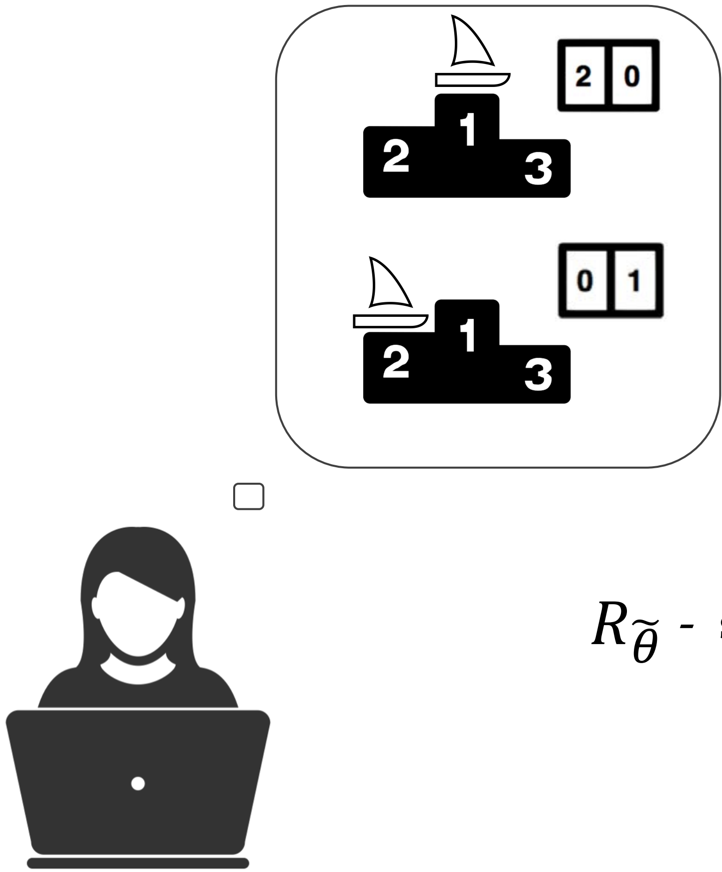
How can we model reward design/specification as a noisy and suboptimal process?



observation (human) model

$$b'(\theta) \propto b(\theta) P(\tilde{\theta} \mid \theta)$$

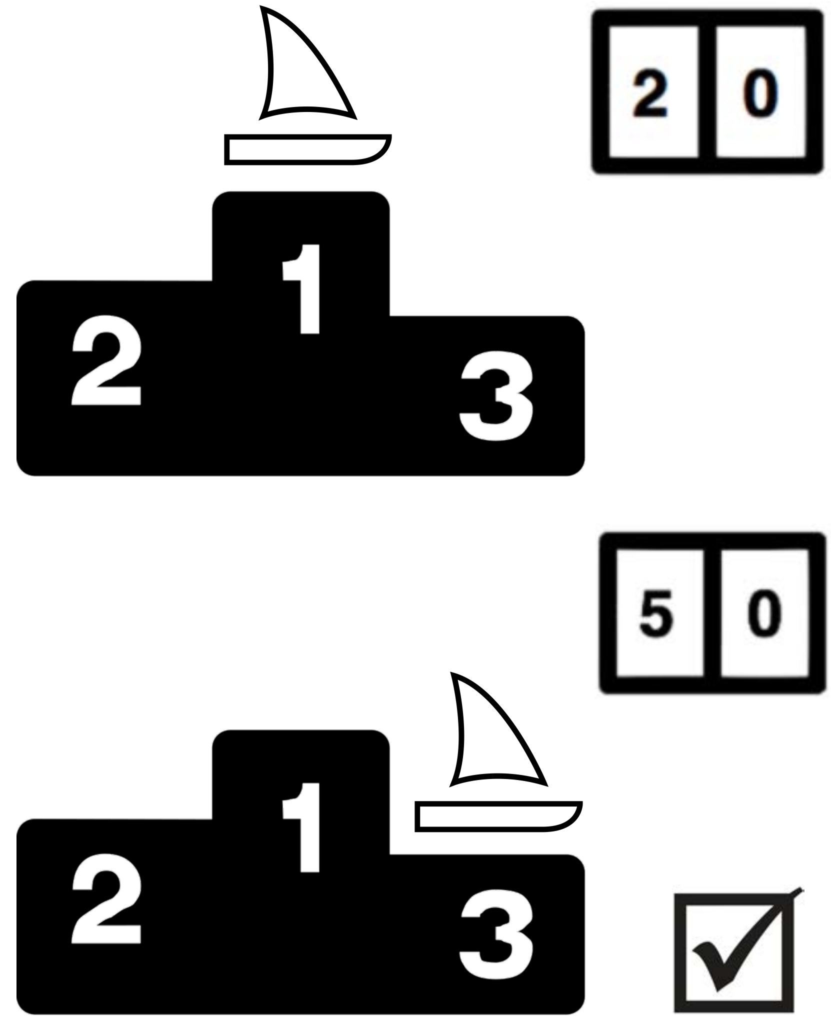
Development



$R_{\widetilde{\theta}}$ - score

score and winning were
correlated at development time…

Deployment



… but no longer
correlated at deployment time

We <u>only</u> know this about the true reward:

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

What you specify is contextualized by the state you specify it in. Robots should interpret it as such.

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

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

Inverse Reward Design
[NIPS'17 with Menell, Milli, Abbeel, Russell]

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

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

Inverse Reward Design
[NIPS'17 with Menell, Milli, Abbeel, Russell]

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

(approximately)
optimal trajectories

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

Inverse Reward Design
[NIPS'17 with Menell, Milli, Abbeel, Russell]

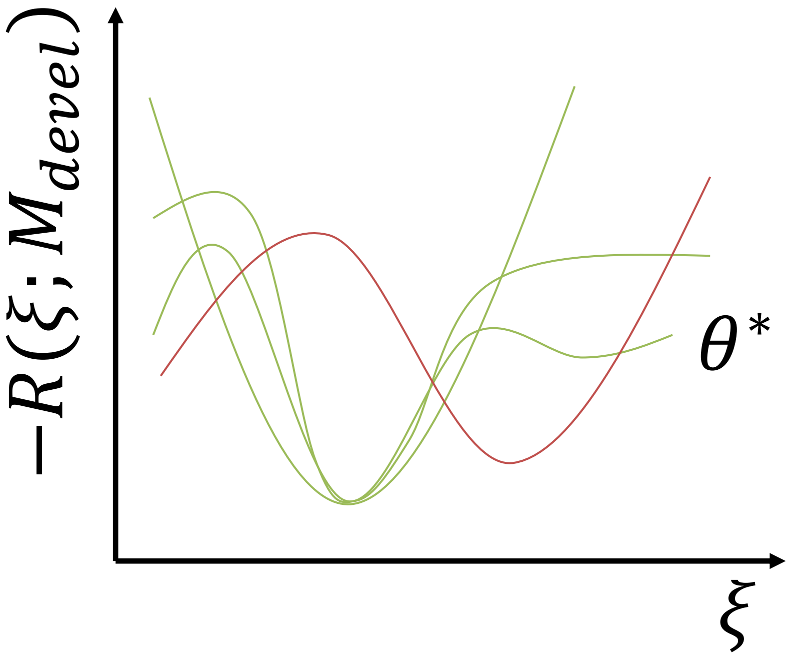
The behavior incentivized by the specified reward in development has high true reward

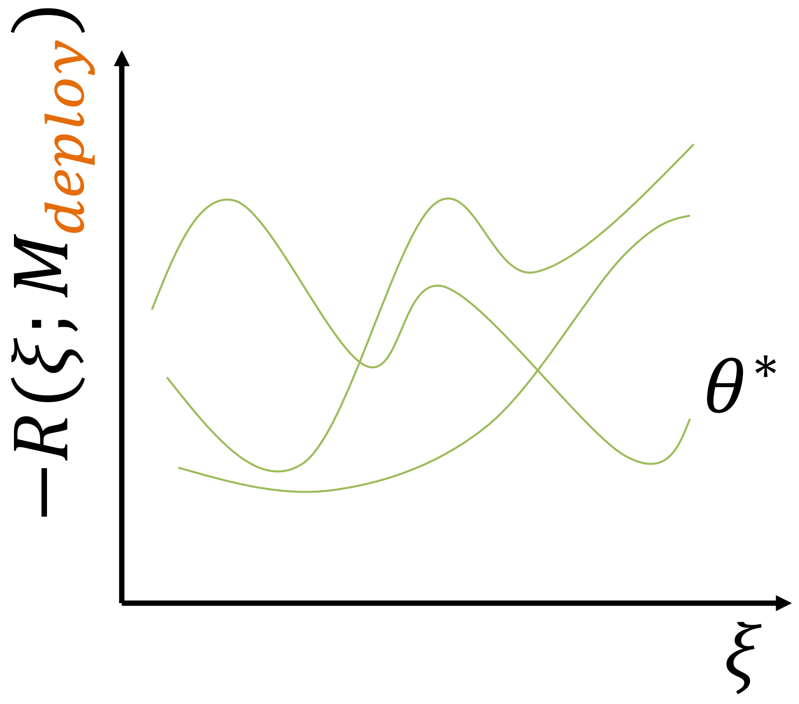
$$P\left(\tilde{\theta} \middle| \theta^*, M_{devel}\right) \propto e^{\beta \mathbb{E}\left[R_{\theta^*}(\xi; M_{devel}) \,\middle|\, \xi \sim P(\xi | \tilde{\theta}, M_{devel})\right]}$$

Inverse Reward Design
[NIPS'17 with Menell, Milli, Abbeel, Russell]

The behavior incentivized by the specified cost in development has low true cost

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

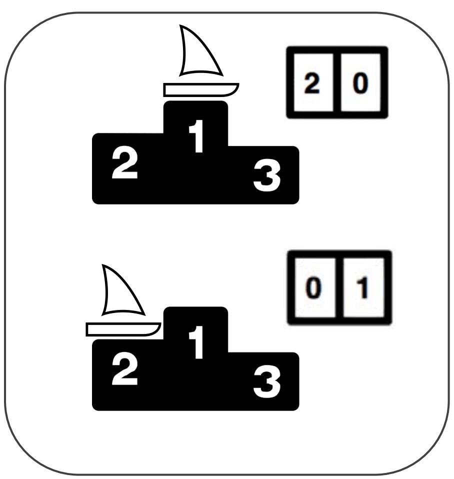




The behavior incentivized by the specified cost in development has low true cost

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

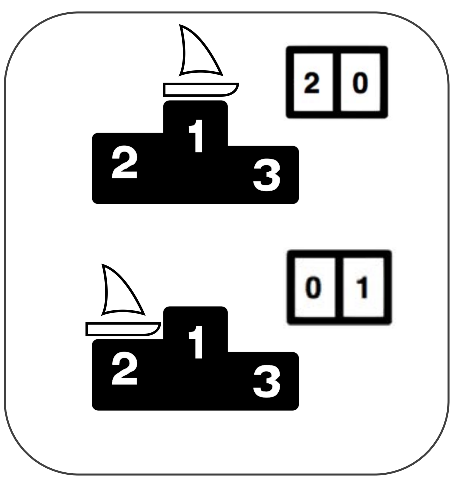
M_{devel}



The behavior incentivized by the specified cost in development has low true cost

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

M_{devel}



θ_1 maximizing winning

θ_2 maximizing score

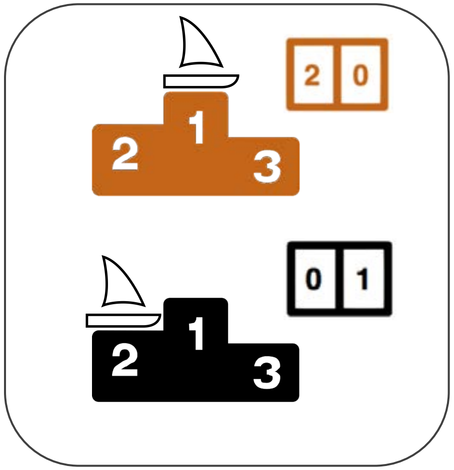
θ_3 minimizing winning

θ_4 minimizing score

The behavior incentivized by the specified cost in development has low true cost

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

M_{devel}



θ_1 maximizing winning

θ_2 maximizing score

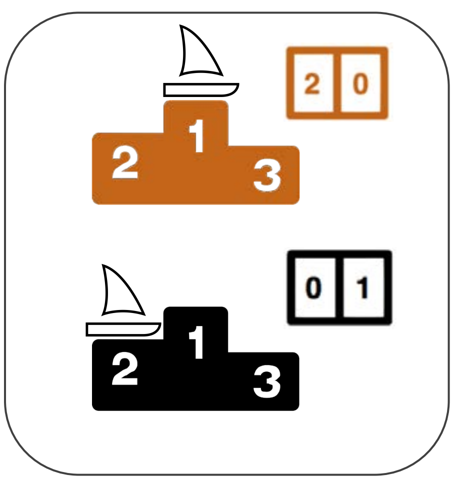
θ_3 minimizing winning

θ_4 minimizing score

The behavior incentivized by the specified cost in development has low true cost

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

M_{devel}



θ_1 maximizing winning

θ_2 maximizing score

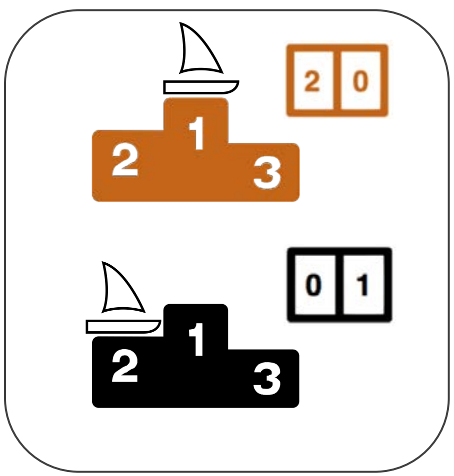
θ_3 minimizing winning

θ_4 minimizing score

The behavior incentivized by the specified cost in development has low true cost

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

M_{devel}

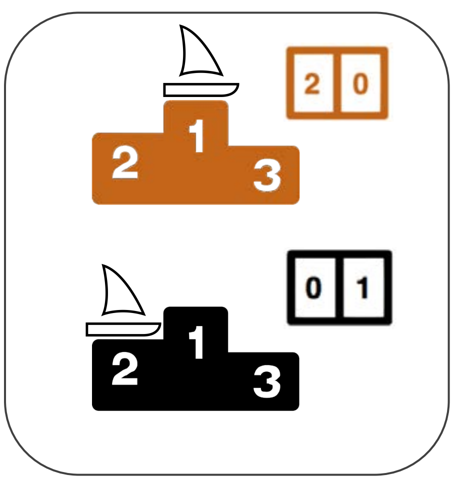


θ_1	θ_2	θ_3	θ_4
maximizing winning	maximizing score	minimizing winning	minimizing score
	✓	✗	✗

The behavior incentivized by the specified cost in development has low true cost

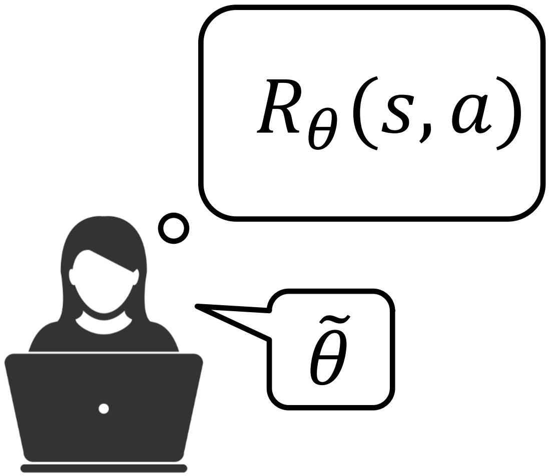
$$P(\tilde{\theta} \mid \theta^*, M_{devel}) \propto e^{\beta \mathbb{E}[R_{\theta^*}(\xi; M_{devel}) \mid \xi \sim P(\xi \mid \tilde{\theta}, M_{devel})]}$$

M_{devel}



θ_1	θ_2	θ_3	θ_4
maximizing winning	maximizing score	minimizing winning	minimizing score
✓	✓	✗	✗

Specified rewards as evidence about the reward



risk-averse planning

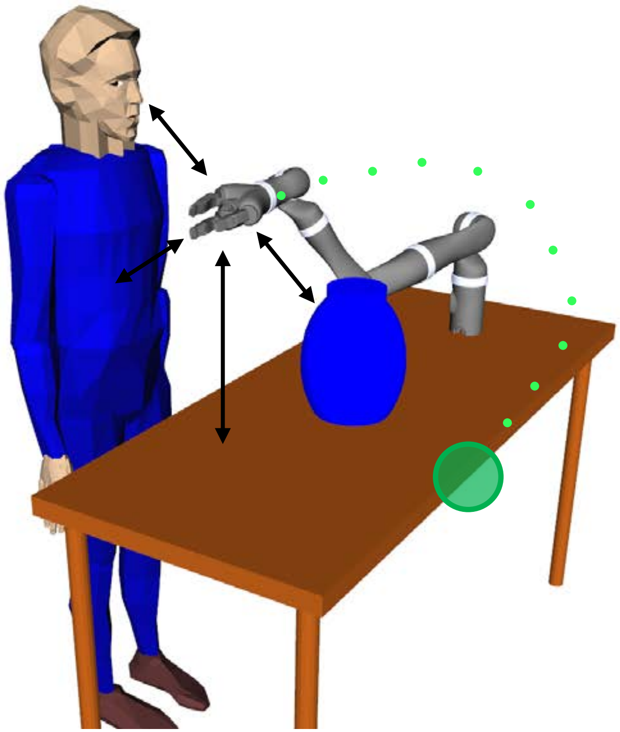
$$\max_{\xi} \min_{\theta \in \{\theta_i \sim b'(\theta)\}} R_\theta(\xi;\, M_{test})$$

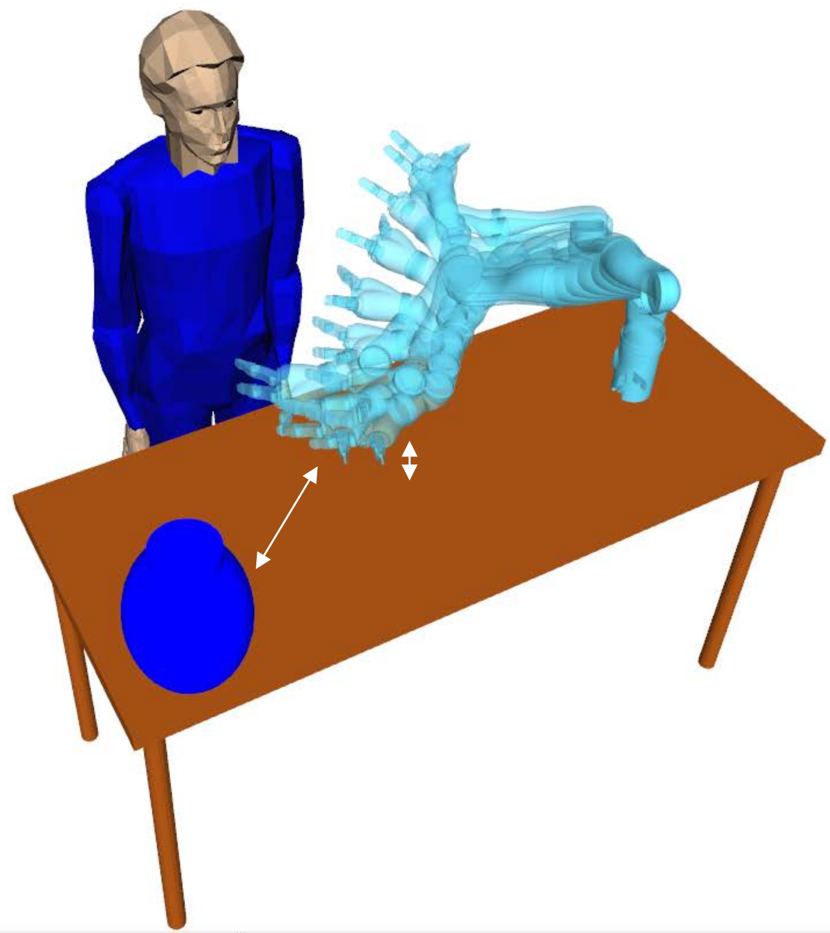
$$b'(\theta) \propto b(\theta) P(\tilde{\theta} | \theta)$$

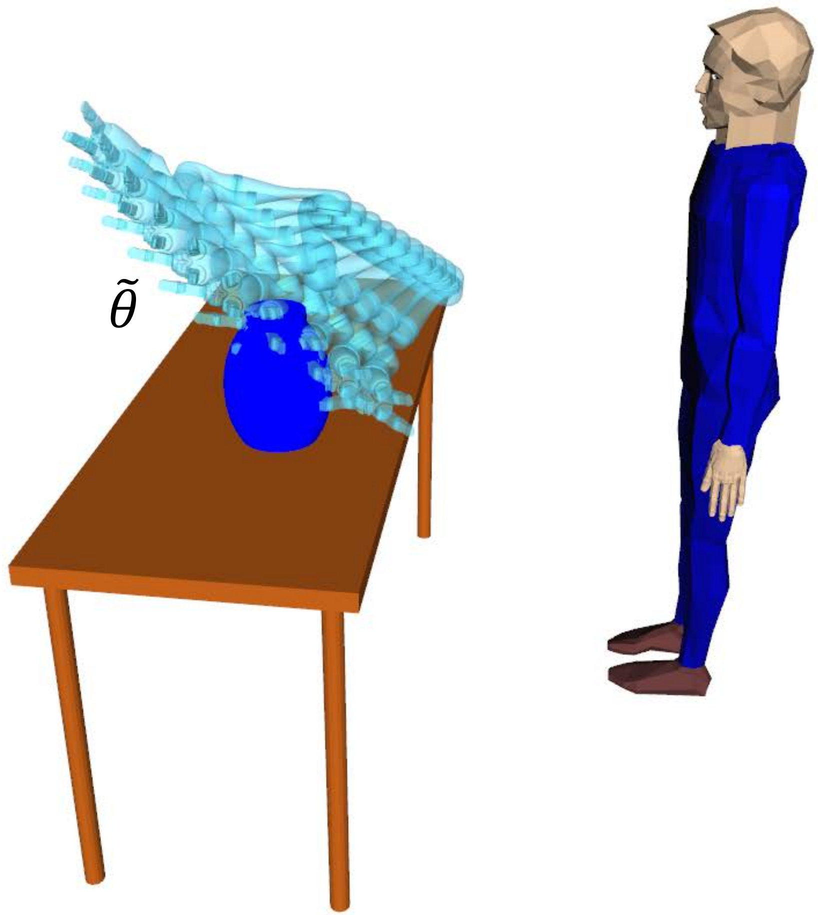
plan in expectation

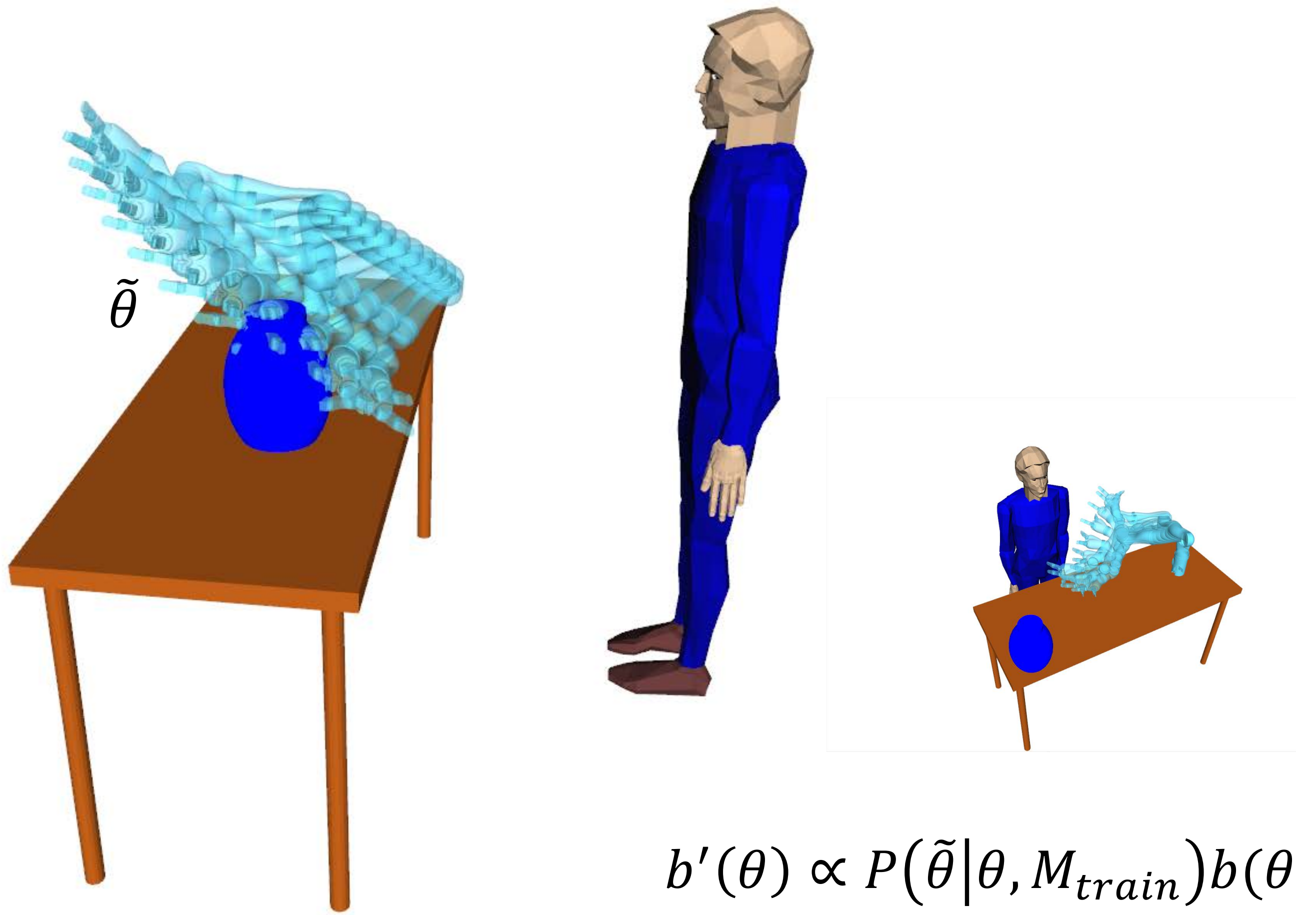
$$\max_{\xi} \mathbb{E}[R_\theta(\xi; M_{test}) | \theta \sim b'(\theta)]$$


KINOVA

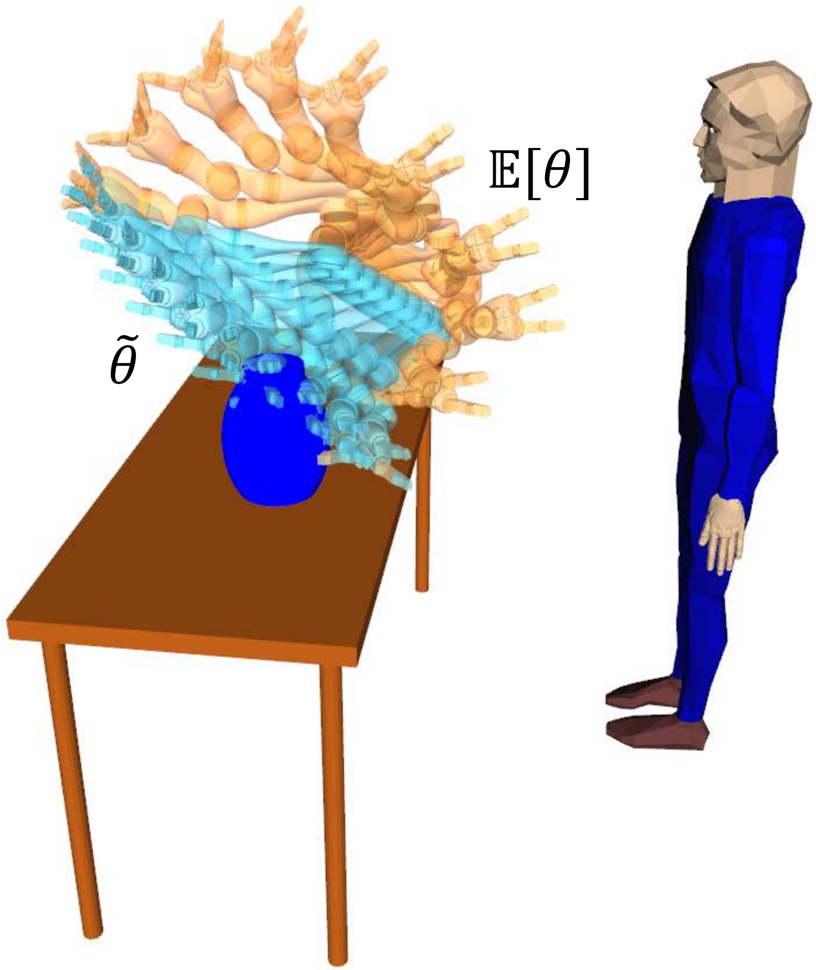




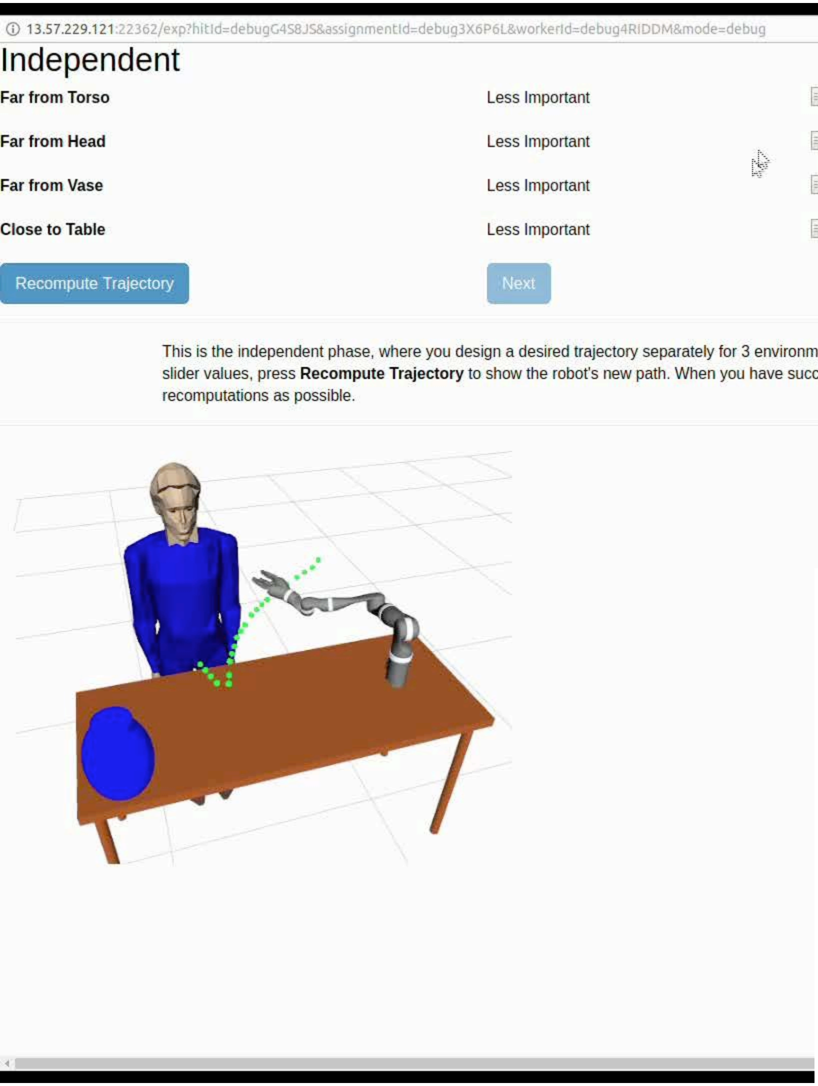

$\tilde{\theta}$

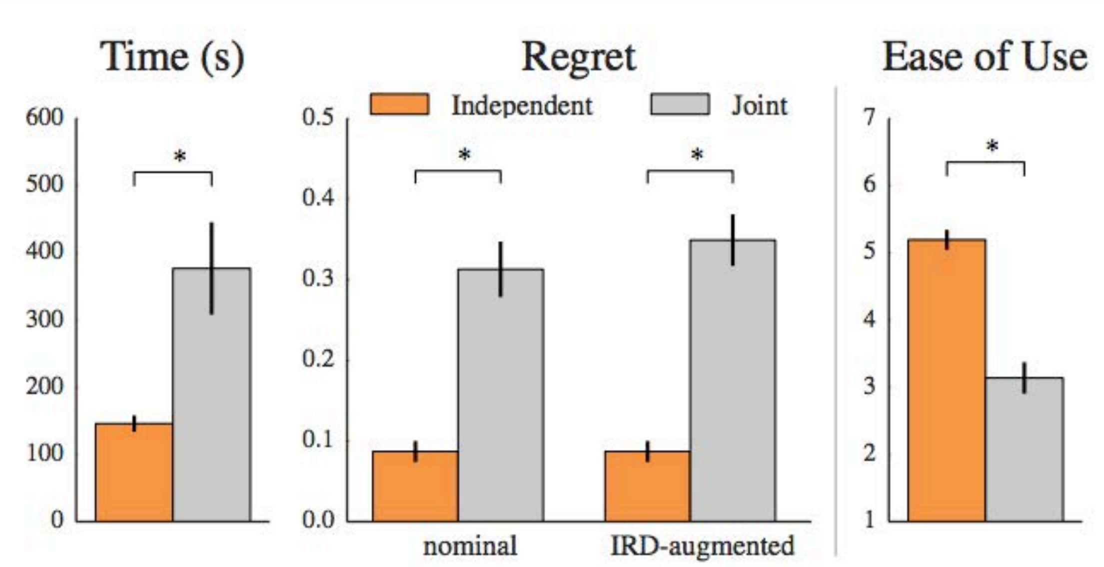


$$b'(\theta) \propto P\left(\tilde{\theta} \middle| \theta, M_{train}\right) b(\theta)$$


$\mathbb{E}[\theta]$
$\tilde{\theta}$

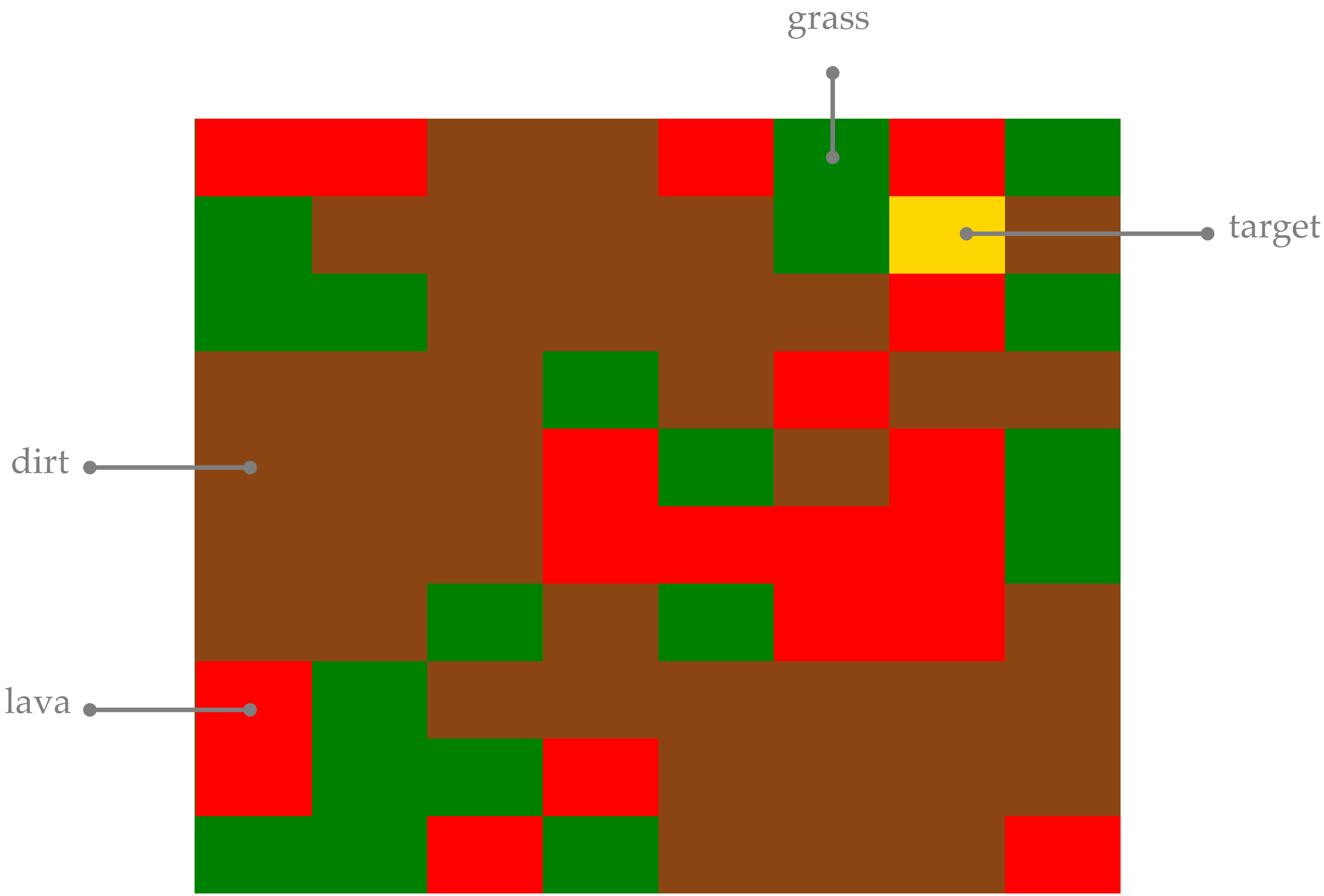
Easier, faster, lower regret



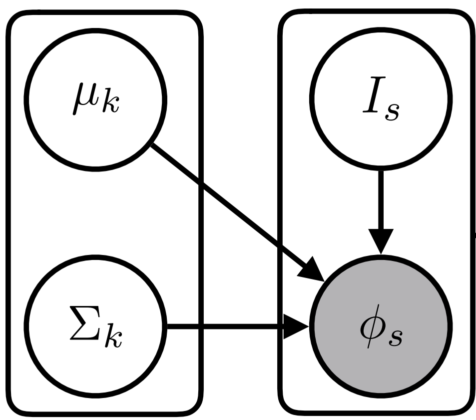


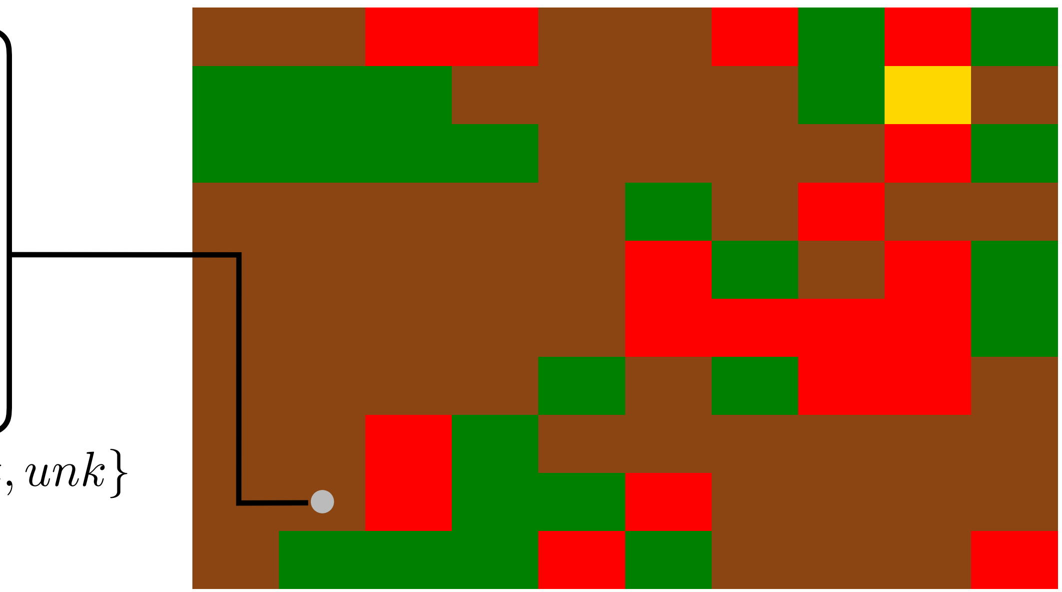
Limitations / ongoing work ..

What if we don't know the important features?!



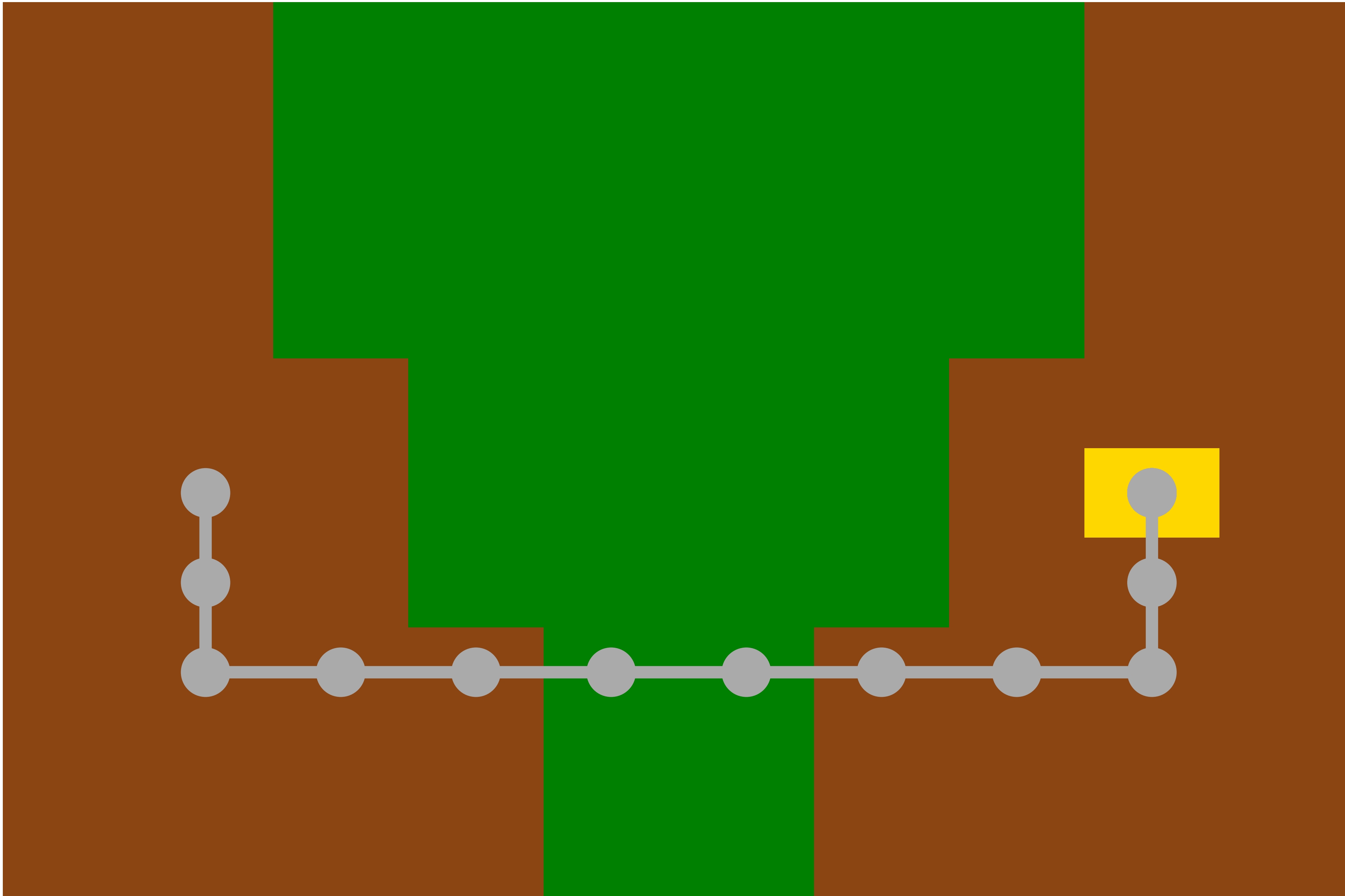
Inference from raw observations, no direct indicators...



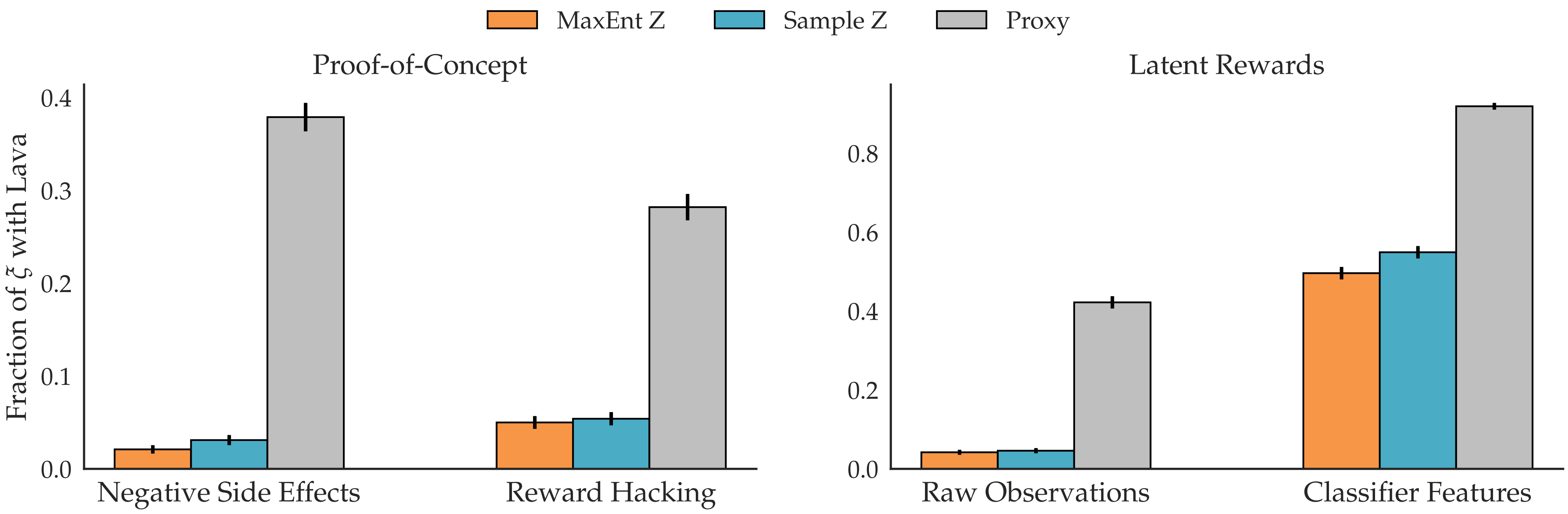


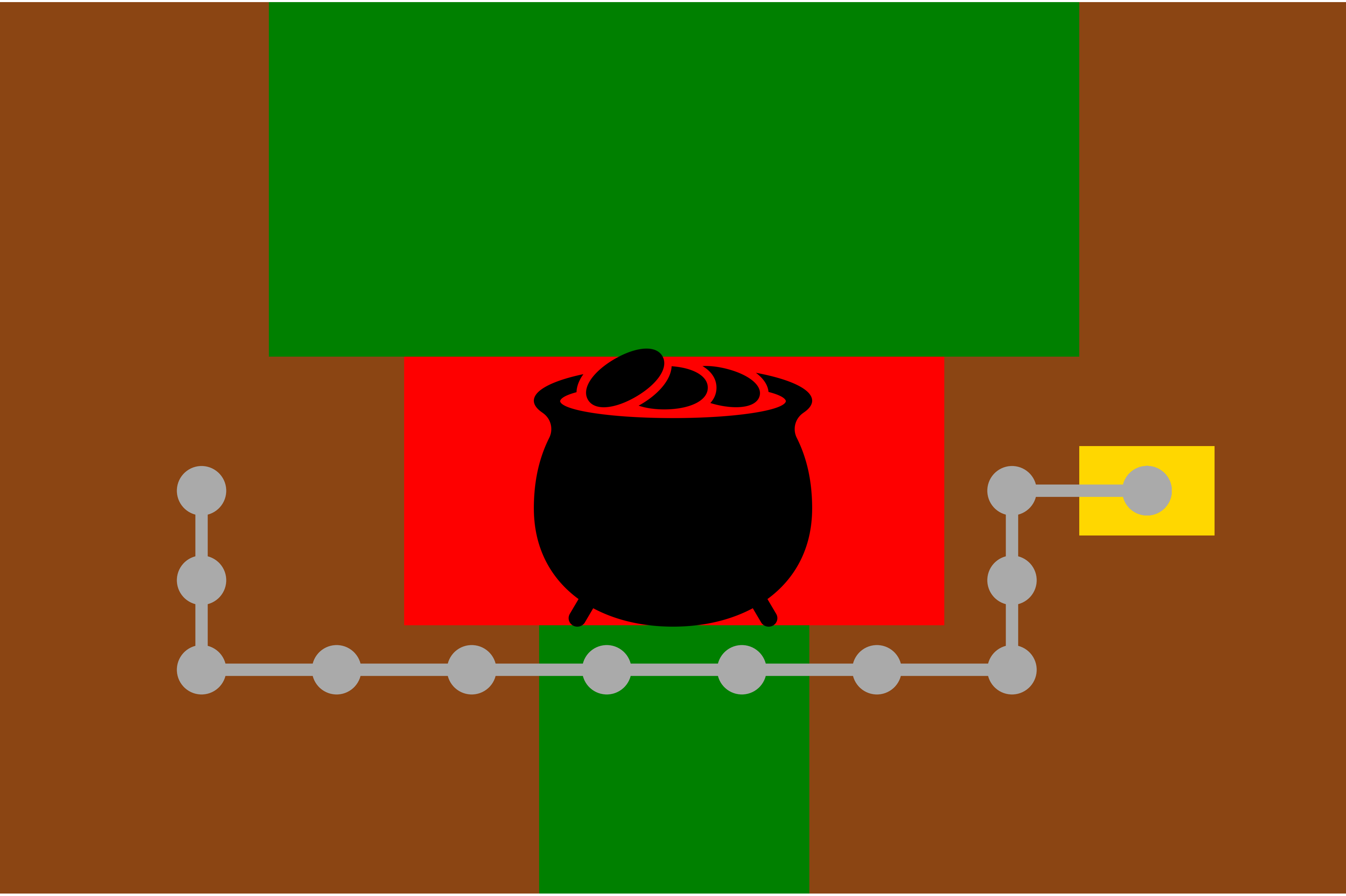
$I_s \in \{grass, dirt, target, unk\}$

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

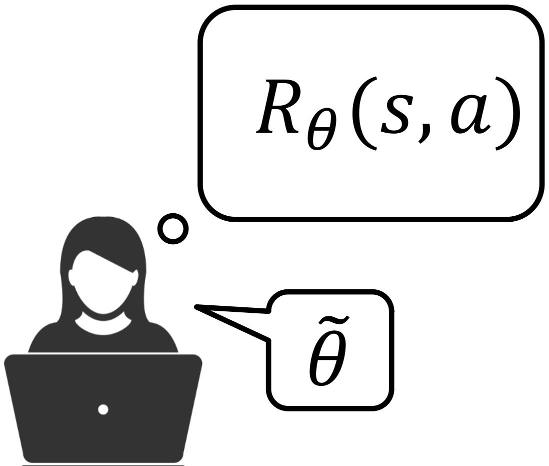


The agent can avoid unintended consequences, even when the features that matter are latent!

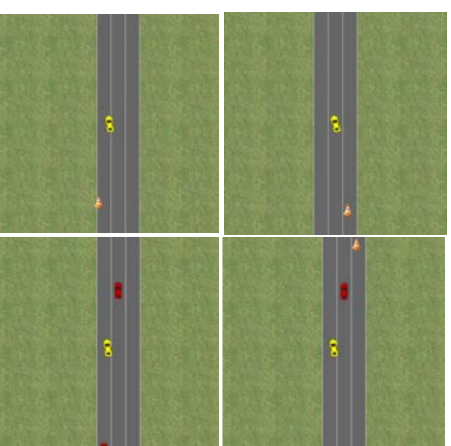




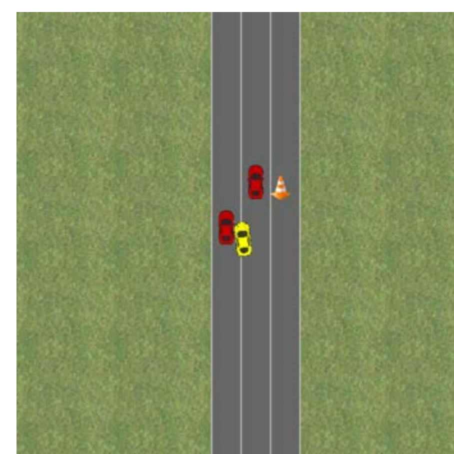
Leverage the posterior to identify edge cases



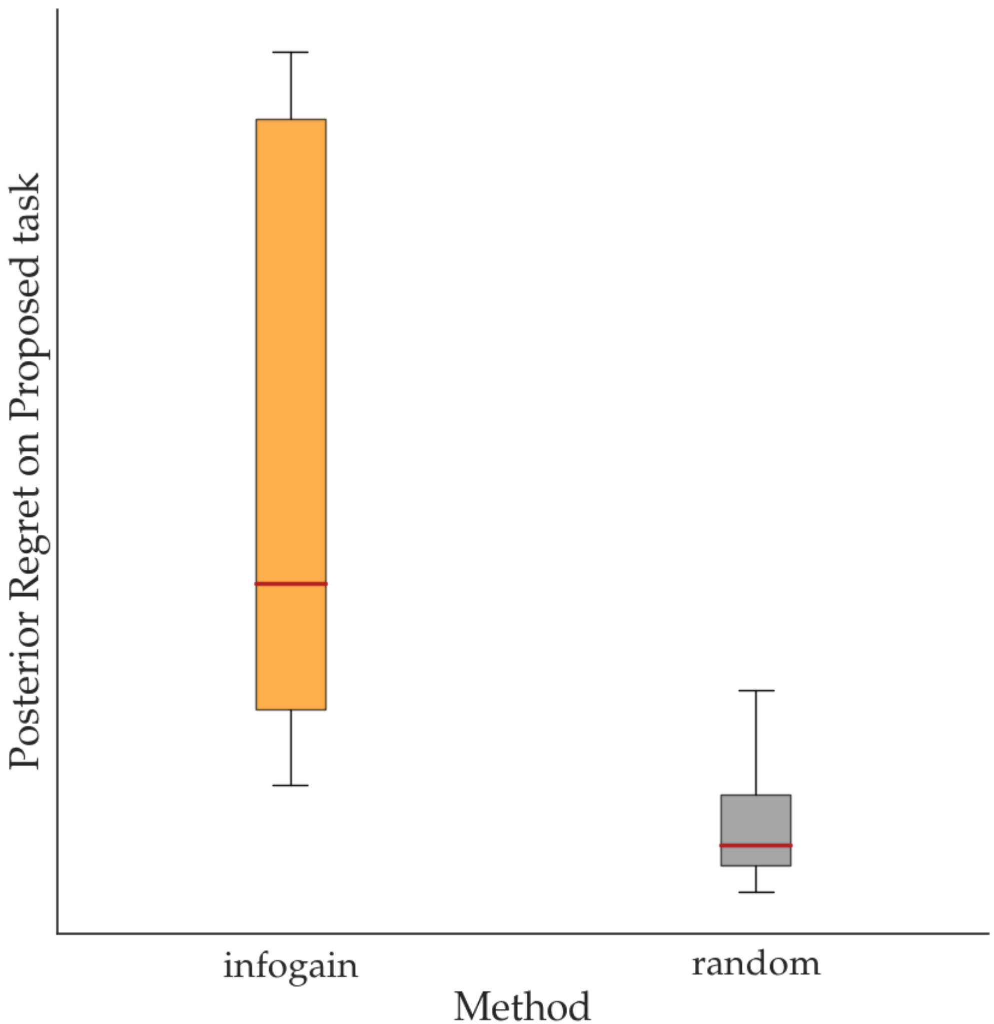
$$b'(\theta) \propto b(\theta) P(\tilde{\theta} | \theta)$$

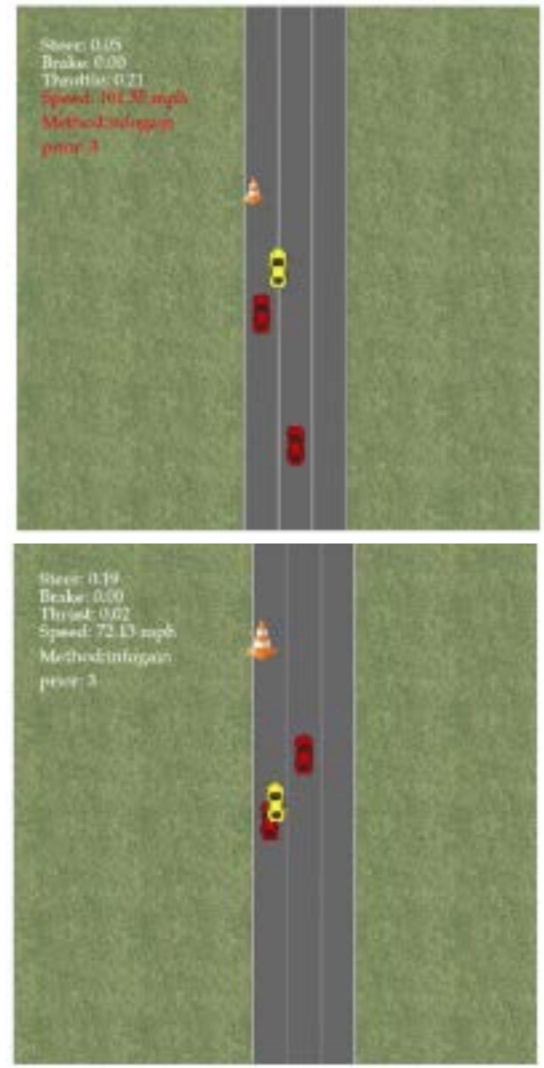


$$M = \arg\max_{m} \mathbb{E}_b [H(b) - H(b'|m)]$$

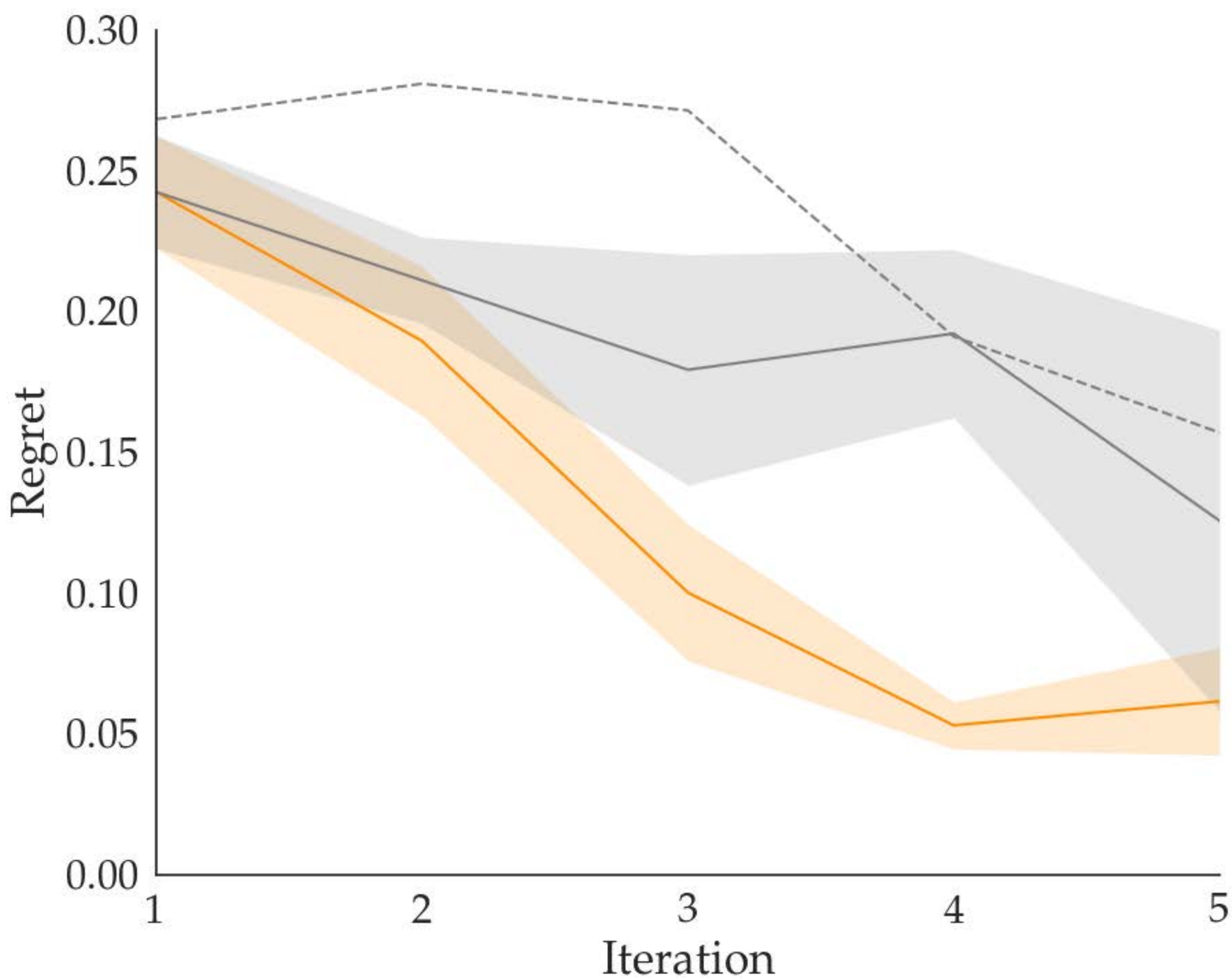


This finds edge-case environments that break the current reward function.



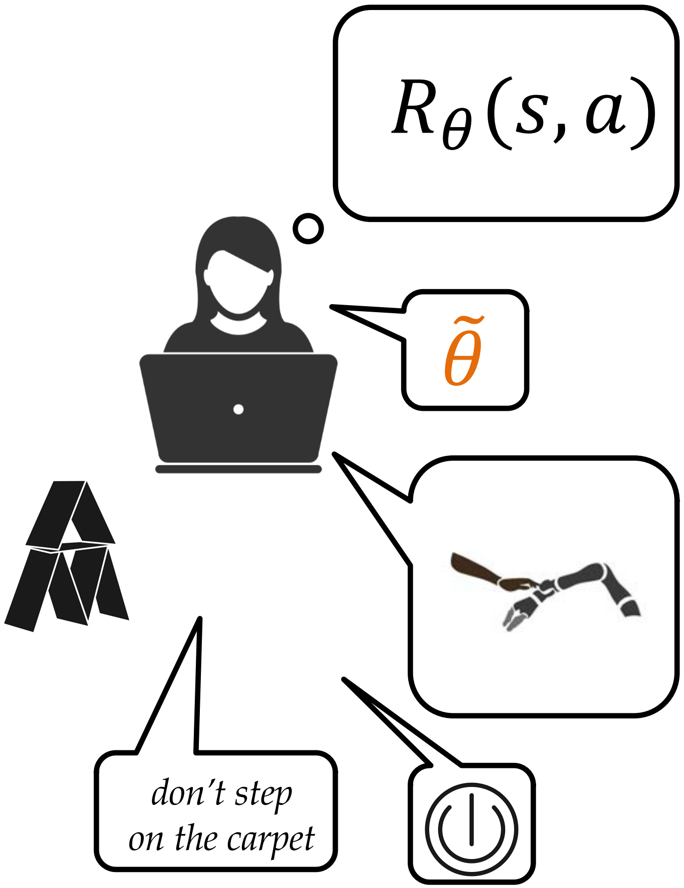


By exposing the designer to these edge cases, regret on held-out environments goes down quickly.



$\theta^T \phi(\xi)$

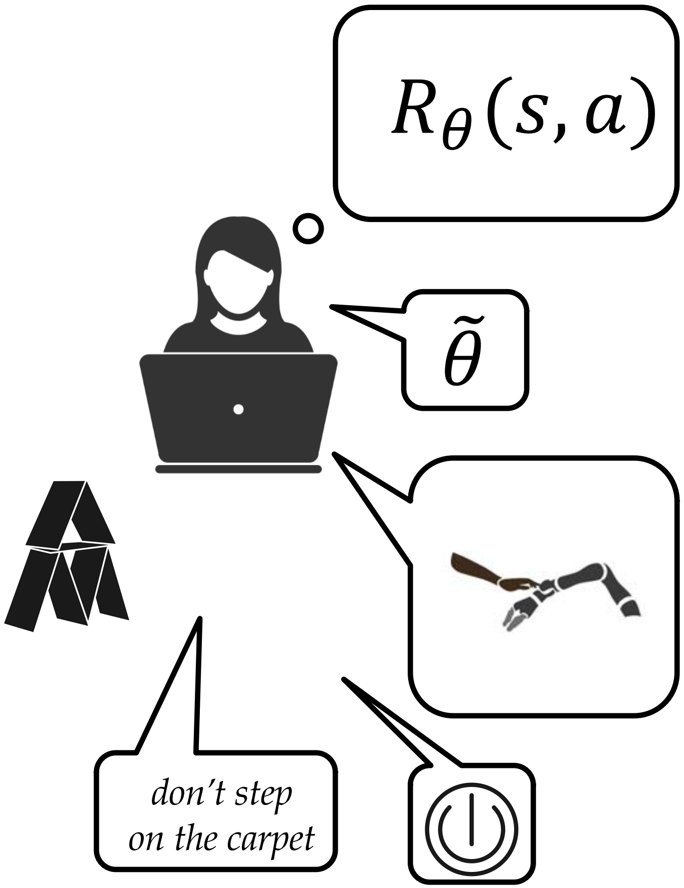
Specified rewards are evidence about the reward.



observation
(human) model

$$b'(\theta) \propto b(\theta) P(\tilde{\theta} \mid \theta)$$

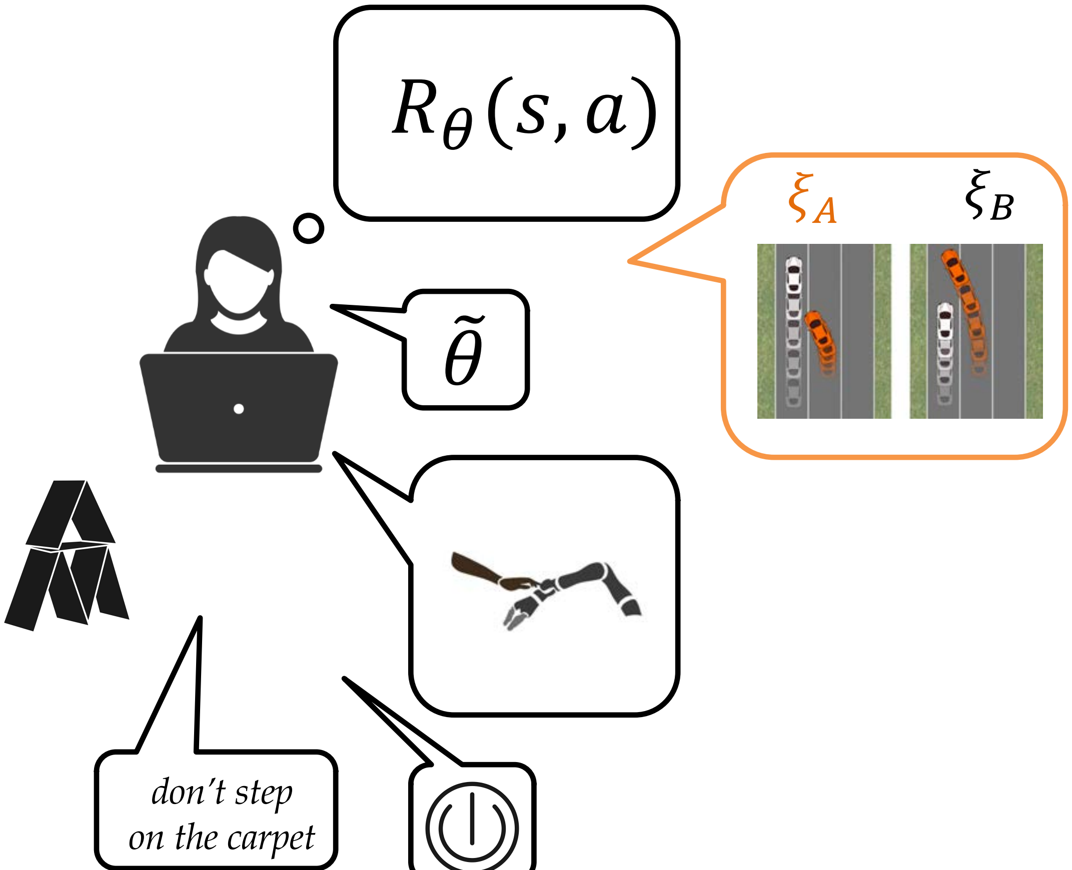
What is a human model that can be used to make sense of all these types of human feedback?



observation (human) model

$$b'(\theta) \propto b(\theta) P(\square | \theta)$$

We know what to do for comparisons.

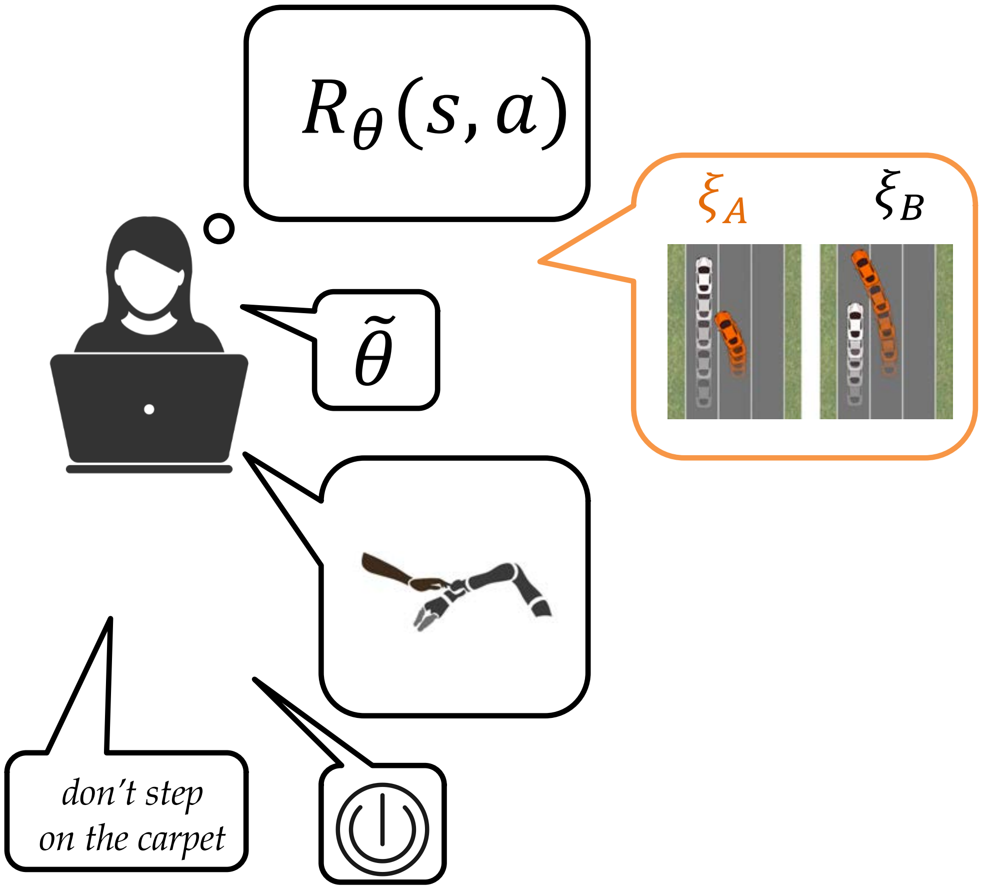


observation
(human) model

$$b'(\theta) \propto b(\theta) P(\square | \theta)$$

Active Preference-Based Learning of Reward Functions
[RSS'17 with Sadigh, Seshia, Sastry]

We know what to do for comparisons: model feedback as a reward-rational choice.



observation (human) model

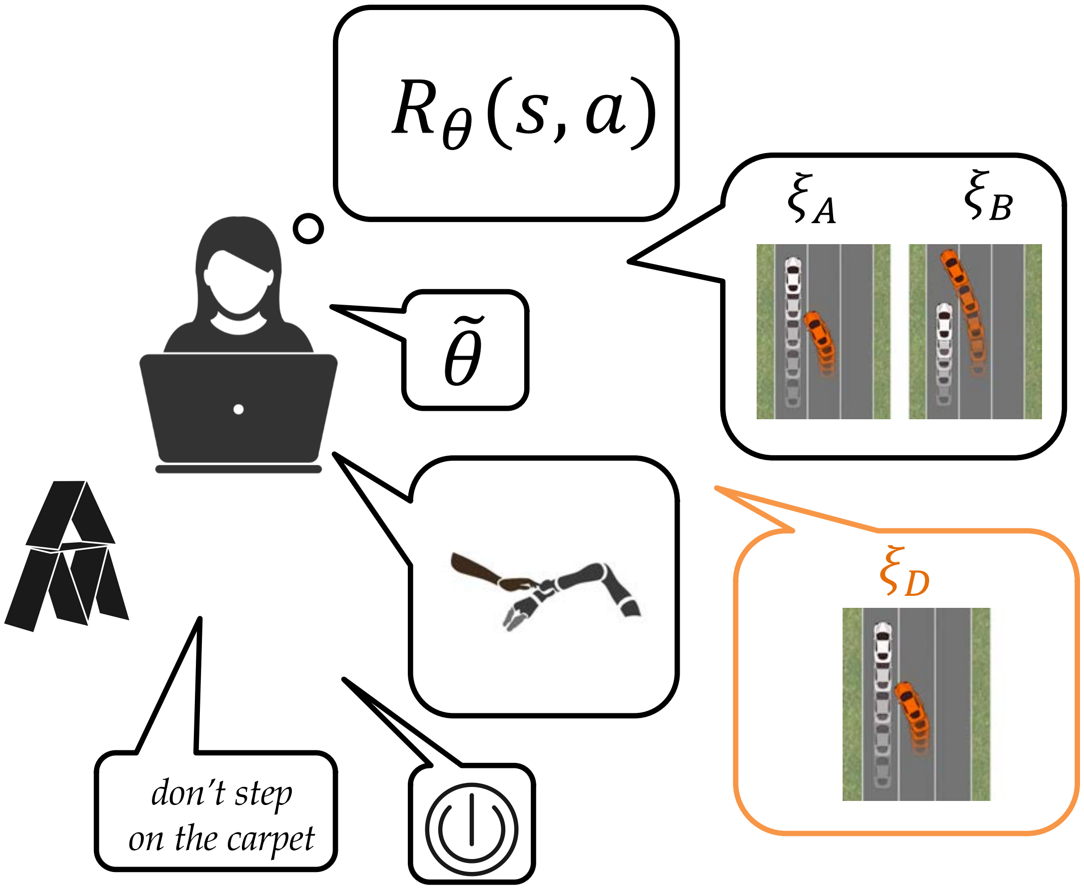
$$b'(\theta) \propto b(\theta) P(\square | \theta)$$

choices: $\{\xi_A, \xi_B\}$

choose based on reward: $R_\theta(\xi_A)\ vs\ R_\theta(\xi_B)$

$$P(\xi_A|\theta) = \frac{e^{R_\theta(\xi_A)}}{e^{R_\theta(\xi_A)} + e^{R_\theta(\xi_B)}}$$

We know what to do for demonstrations: model the demo as a reward-rational implicit choice.



observation (human) model

$$b'(\theta) \propto b(\theta) P(\square \mid \theta)$$

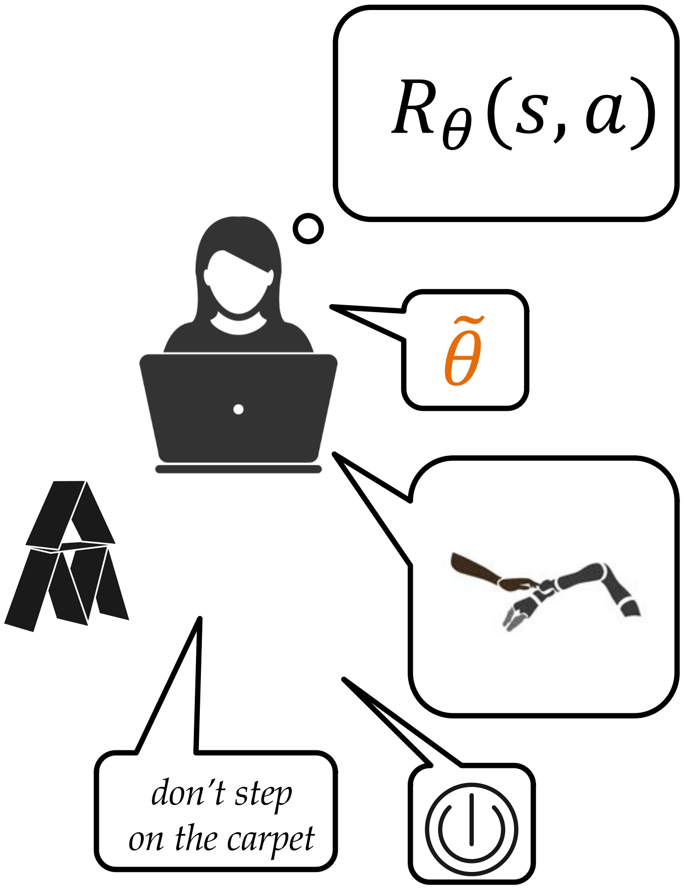
choices: $\{\xi_i\}$

choose based on reward: $R_\theta(\xi_D)\ vs\ R_\theta(\xi) \forall \xi$

$$P(\xi_D \mid \theta) = \frac{e^{R_\theta(\xi_D)}}{\sum_\xi e^{R_\theta(\xi)}}$$

[Ramachandran et al., Bayesian Inverse Reinforcement Learning]

We know what to do for specified rewards

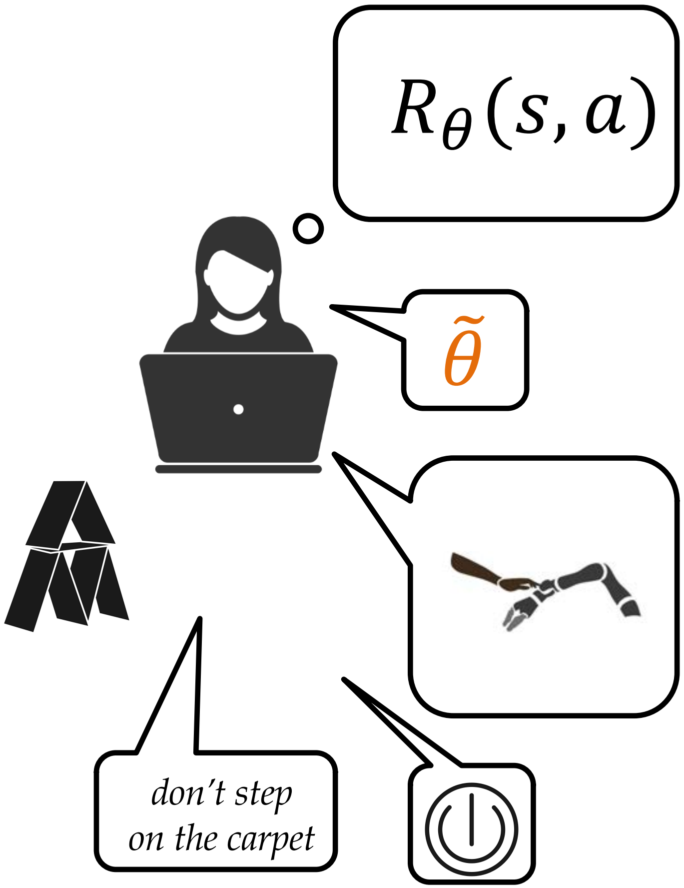


observation
(human) model

$$b'(\theta) \propto b(\theta) P(\quad|\theta)$$

$$P(\tilde{\theta}|\theta) = \frac{e^{\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]}}{\sum_{\bar{\theta}} e^{\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\bar{\theta}, M_{devel})]}}$$

We know what to do for specified rewards: model them as a reward-rational implicit choice.



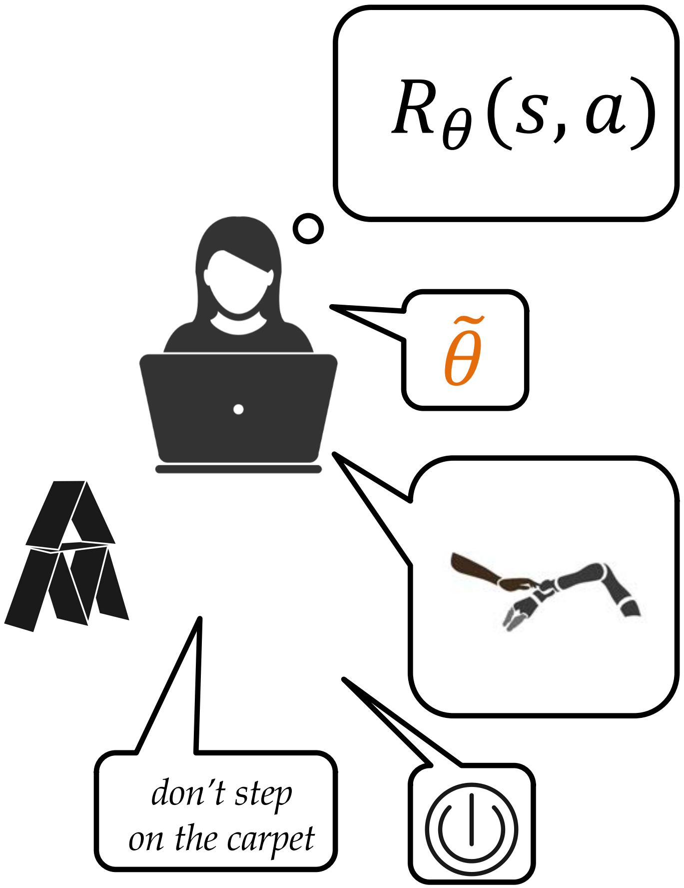
observation (human) model

$$b'(\theta) \propto b(\theta) P(\quad|\theta)$$

choices: $\{\theta_i\}$

$$P(\tilde{\theta}|\theta) = \frac{e^{\mathbb{E}[R_\theta(\xi)|\xi\sim P(\xi|\tilde{\theta},M_{devel})]}}{\sum_{\bar{\theta}} e^{\mathbb{E}[R_\theta(\xi)|\xi\sim P(\xi|\bar{\theta},M_{devel})]}}$$

We know what to do for specified rewards: model them as a reward-rational implicit choice.



observation (human) model

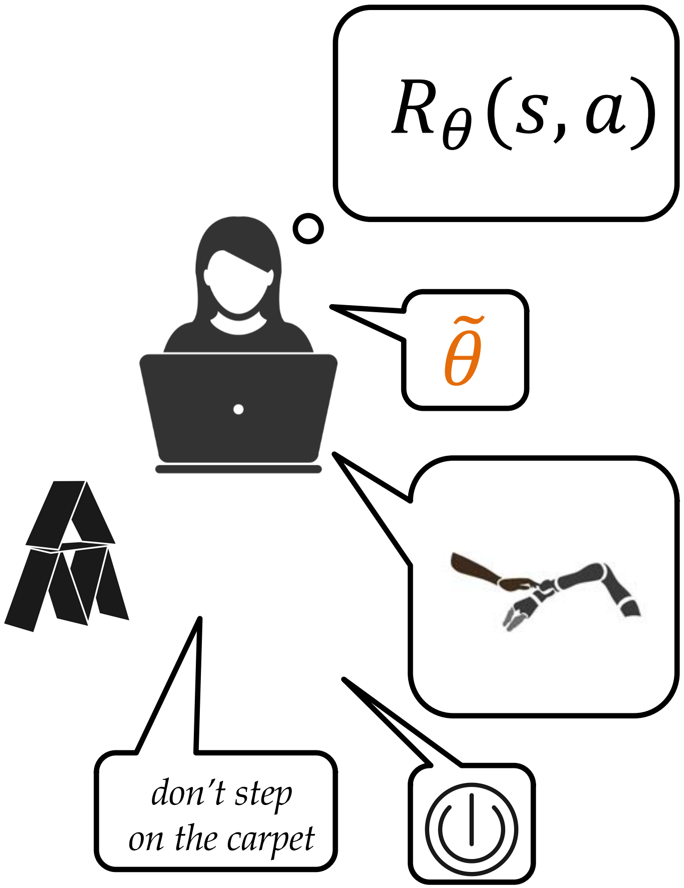
$$b'(\theta) \propto b(\theta)P(\square|\theta)$$

choices: $\{\theta_i\}$

choose based on reward: $R_\theta(\tilde{\theta})$?!

$$P(\tilde{\theta}|\theta) = \frac{e^{\mathbb{E}[R_\theta(\xi)|\xi\sim P(\xi|\tilde{\theta},M_{devel})]}}{\sum_{\bar{\theta}} e^{\mathbb{E}[R_\theta(\xi)|\xi\sim P(\xi|\bar{\theta},M_{devel})]}}$$

We know what to do for specified rewards: model them as a reward-rational implicit choice.



observation (human) model

$$b'(\theta) \propto b(\theta) P(\quad|\theta)$$

choices: $\{\theta_i\}$

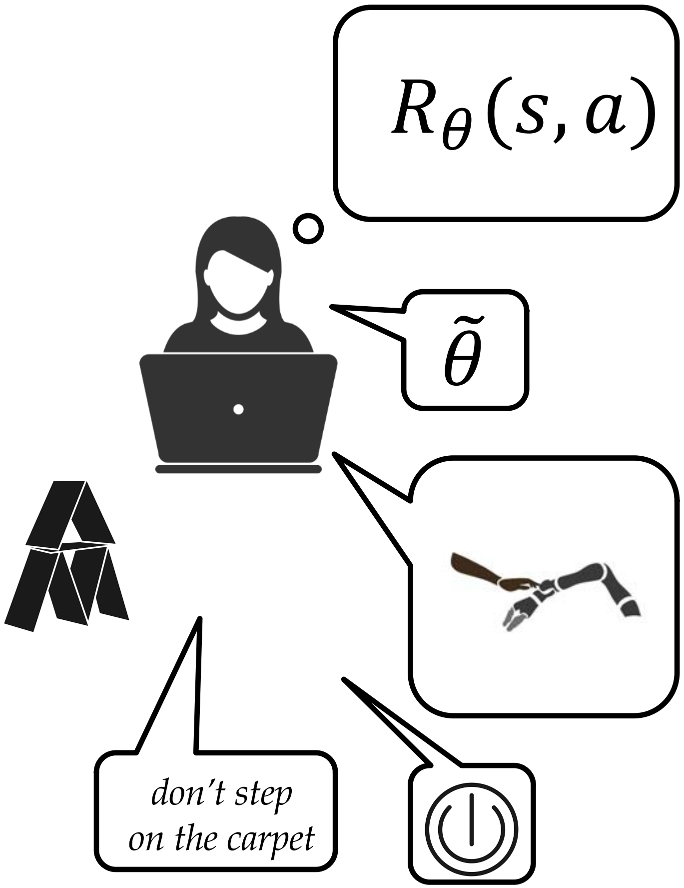
choose based on reward: $\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]$

vs

$\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\bar{\theta}, M_{devel})] \forall \bar{\theta}$

$$P(\tilde{\theta}|\theta) = \frac{e^{\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\tilde{\theta}, M_{devel})]}}{\sum_{\bar{\theta}} e^{\mathbb{E}[R_\theta(\xi)|\xi \sim P(\xi|\bar{\theta}, M_{devel})]}}$$

Reward-rational (implicit) choices



observation (human) model

$$b'(\theta) \propto b(\theta) P(\quad|\theta)$$

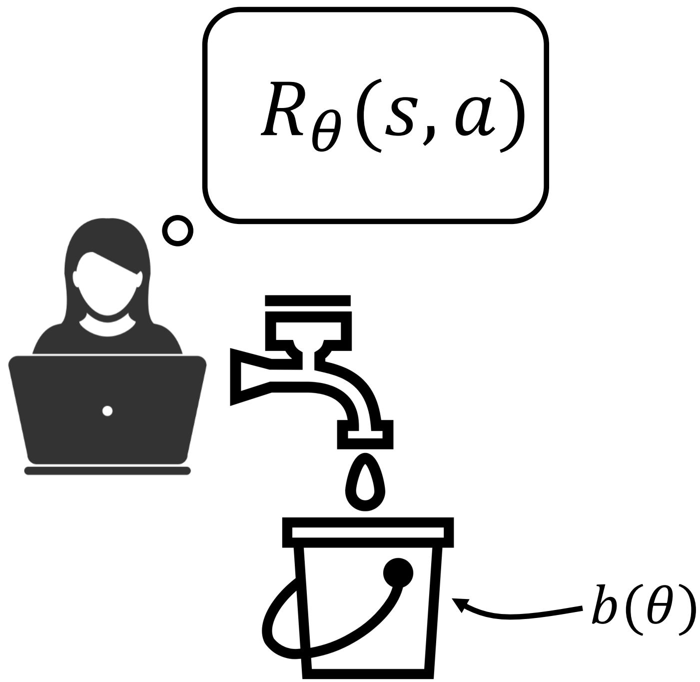
choices: $\{\, c \}$

choose based on reward: $\mathbb{E}[R_\theta(\xi)|\xi \sim \psi(c^*)]$

vs

$\mathbb{E}[R_\theta(\xi)|\xi \sim \psi(c)] \forall c$

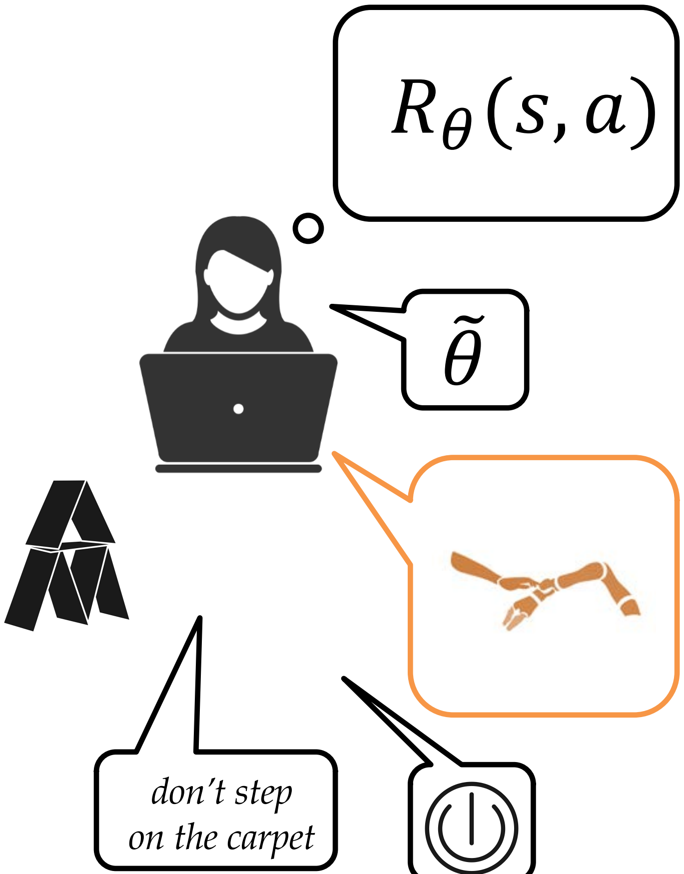
$$P(c^*|\theta) = \frac{e^{\mathbb{E}[R_\theta(\xi)|\xi \sim \psi(c^*)]}}{\sum e^{\mathbb{E}[R_\theta(\xi)|\xi \sim \psi(c)]}}$$



How should the robot extract
the leaked information into an updated belief?

Key idea: Interpret any type of human feedback as a reward-rational implicit choice.

Human feedback as a reward-rational implicit choice.



observation (human) model

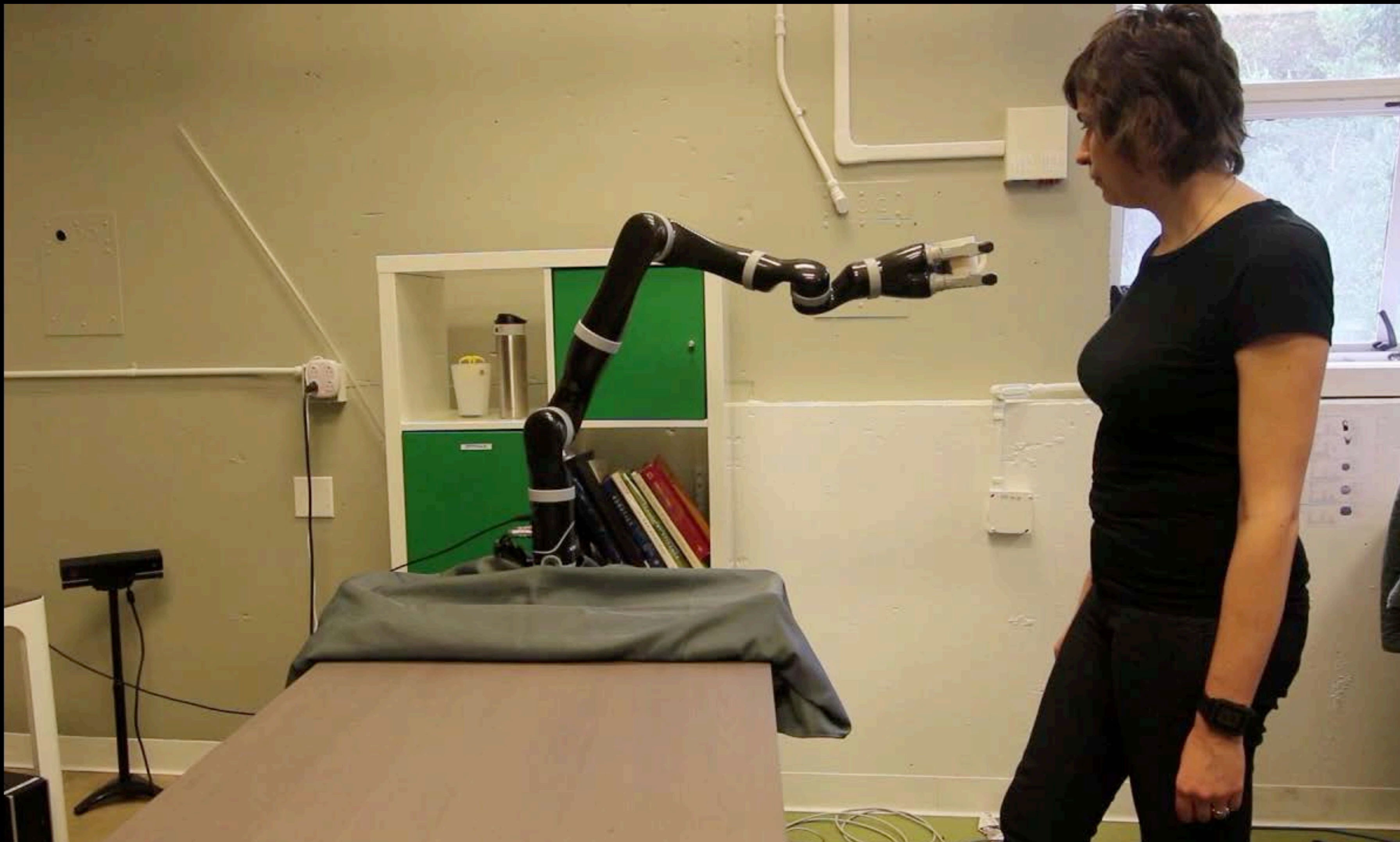
$$b'(\theta) \propto b(\theta) P(\quad |\theta)$$

choices: $\{\tau\}$ (external torques)

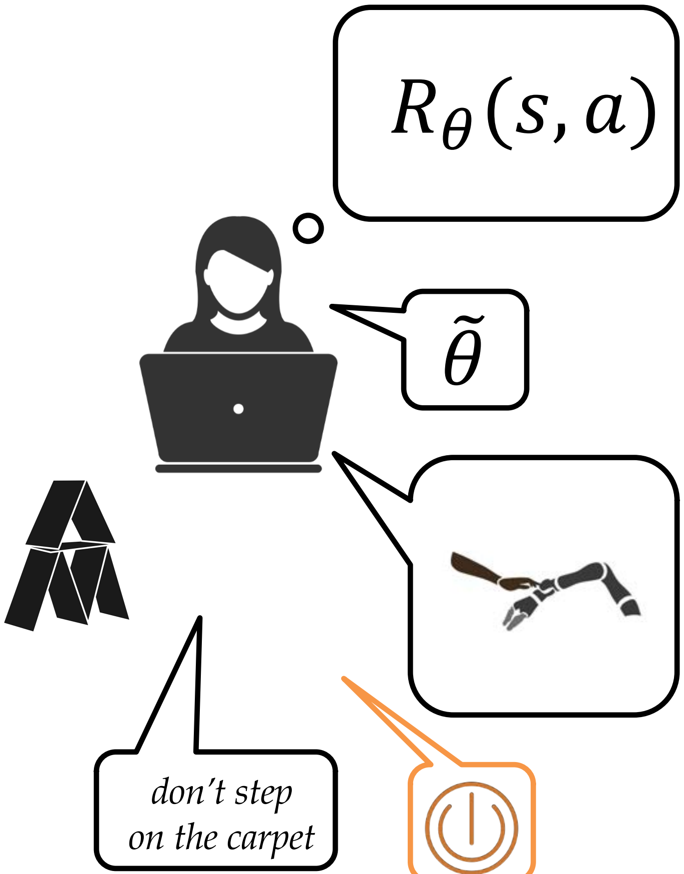
choose based on reward: $R_\theta(\xi(\xi_{original}, \tau))$

(deformed trajectories)

Learning Robot Objectives from Physical Human Interaction
[CoRL'18 with Bajcsy, Losey, O'Malley]



Human feedback as a reward-rational implicit choice.



The off-switch game
[IJCAI'17 with Menell, Milli, Abbeel, Russell]

observation (human) model

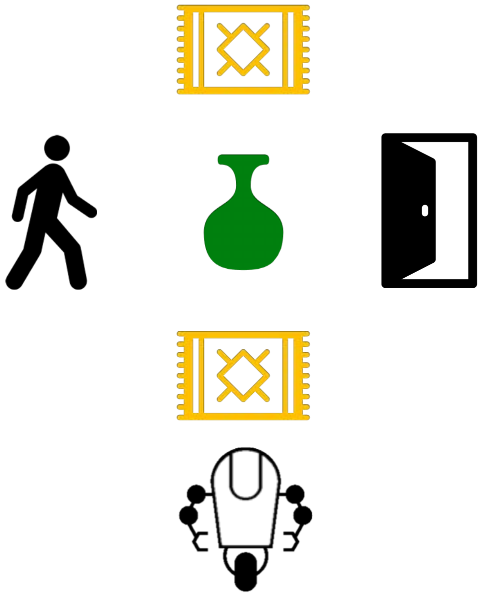
$$b'(\theta) \propto b(\theta) P(\quad | \theta)$$

choices: $\{press\ button, do\ nothing\}$

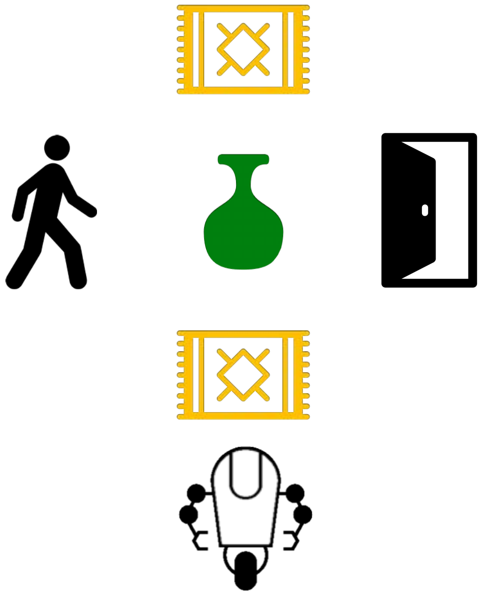
choose based on reward:

$$R_\theta(\xi_{stopped})$$
$$vs$$
$$R_\theta(\xi_{planned})$$

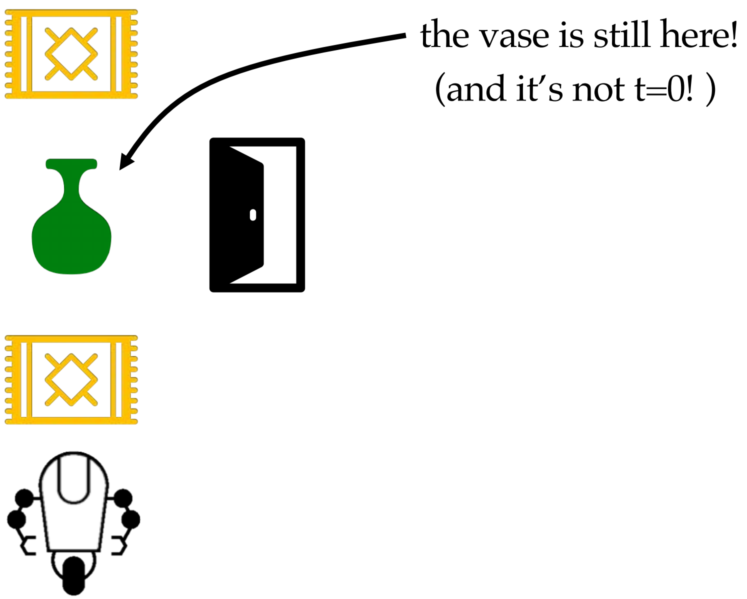
So far, we've talked about sources of information that look at human behavior:

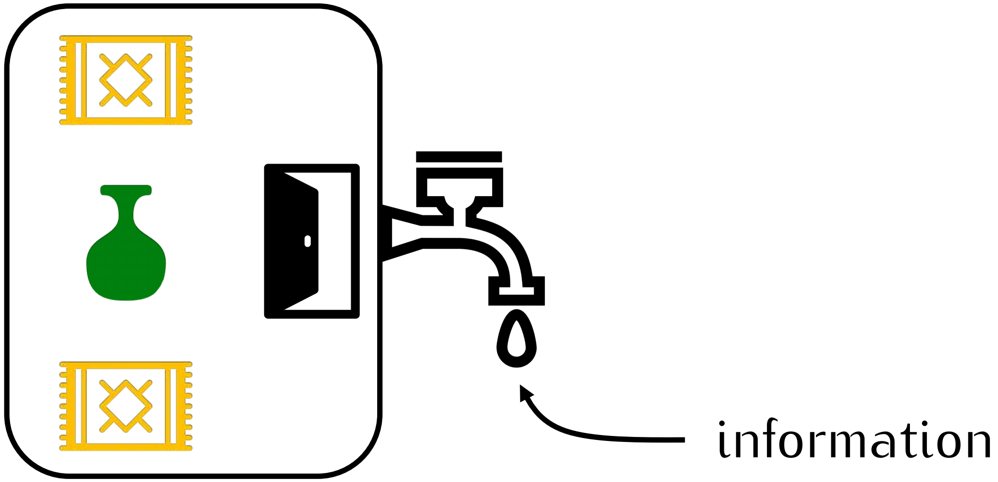


To know that you shouldn't break the vase, you need to see some behavior, e.g.:



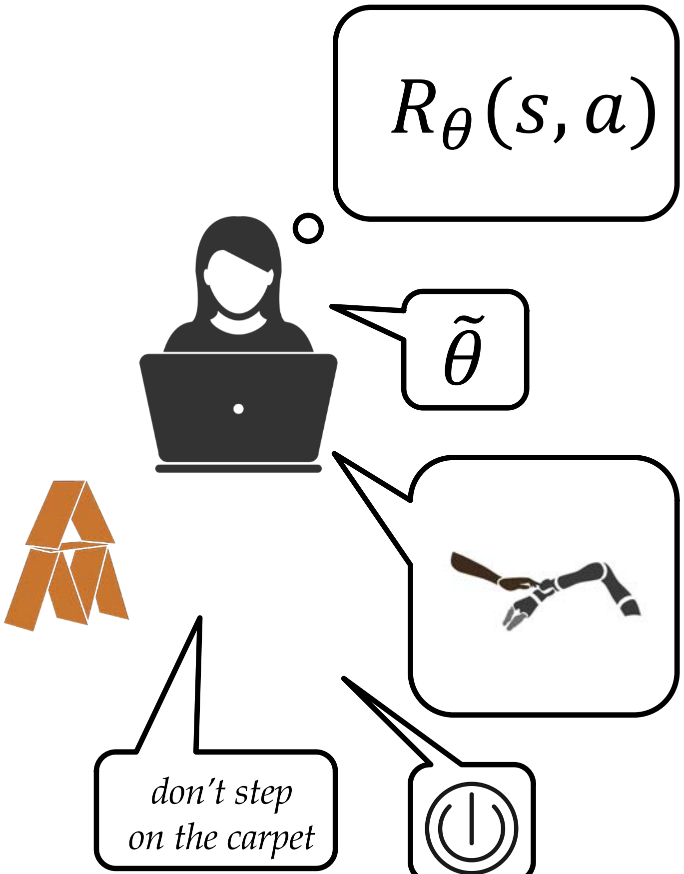
What if we don't see any behavior?





When the agent is deployed in an environment that the human has been acting in, the state of the environment has information about the human's intended reward.

The state of the environment as a reward-rational implicit choice.



observation (human) model

$$b'(\theta) \propto b(\theta) P(\square | \theta)$$

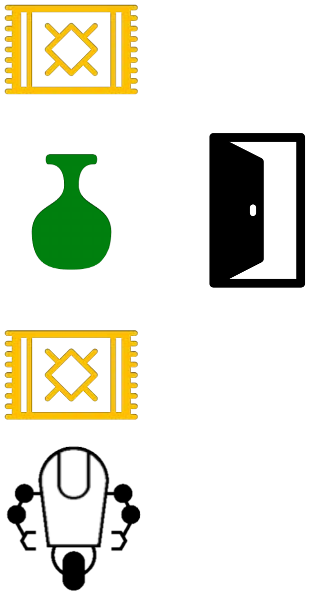
choices: $\{ s_0 \}$ (states)

choose based on reward: $\mathbb{E}[R_\theta(\xi_{-T:0}) | \xi(0) = s_0]$

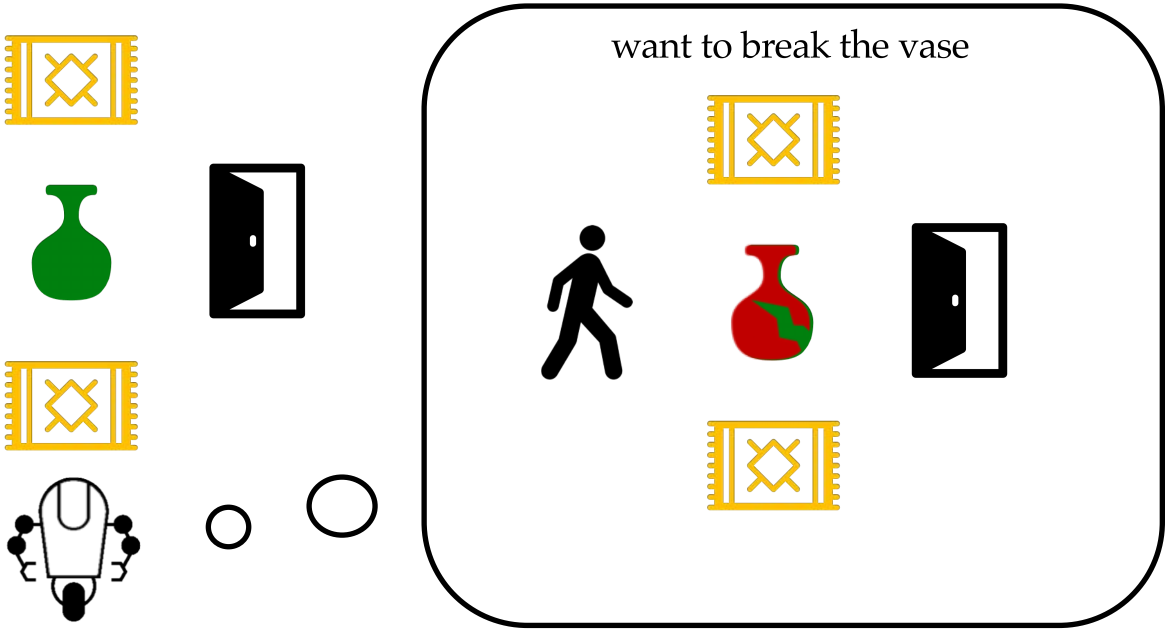
(trajectories that end at the observed state)

Preferences implicit in the state of the world
[ICLR'19 with Shah, Krashenninikov, Alexander, Abbeel]

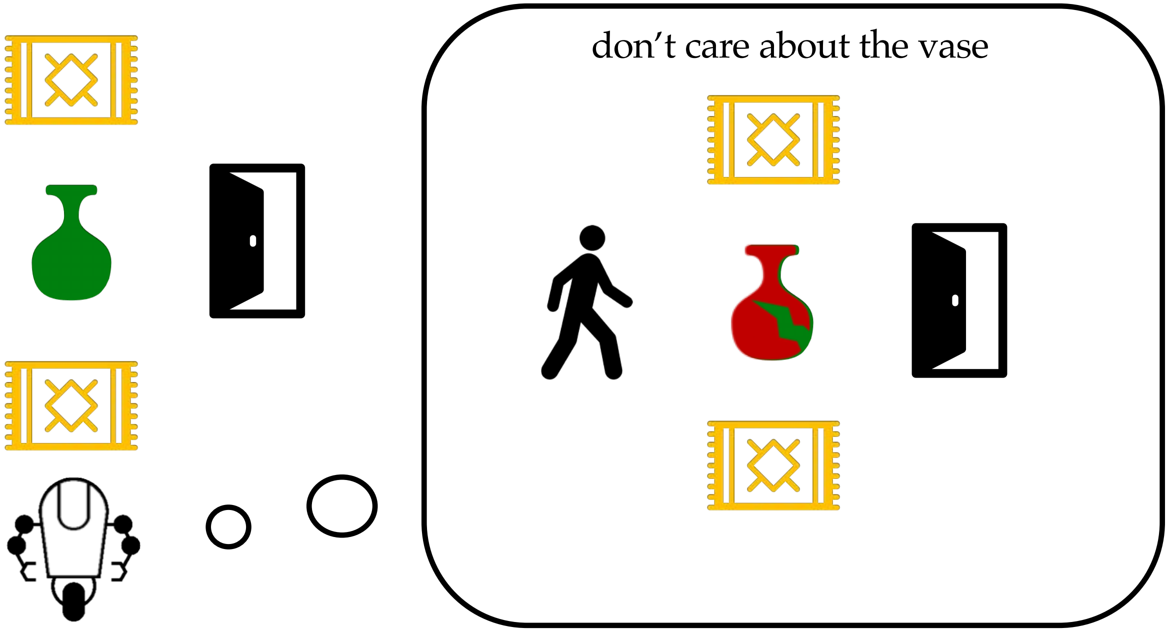
What reward function is the state consistent with?



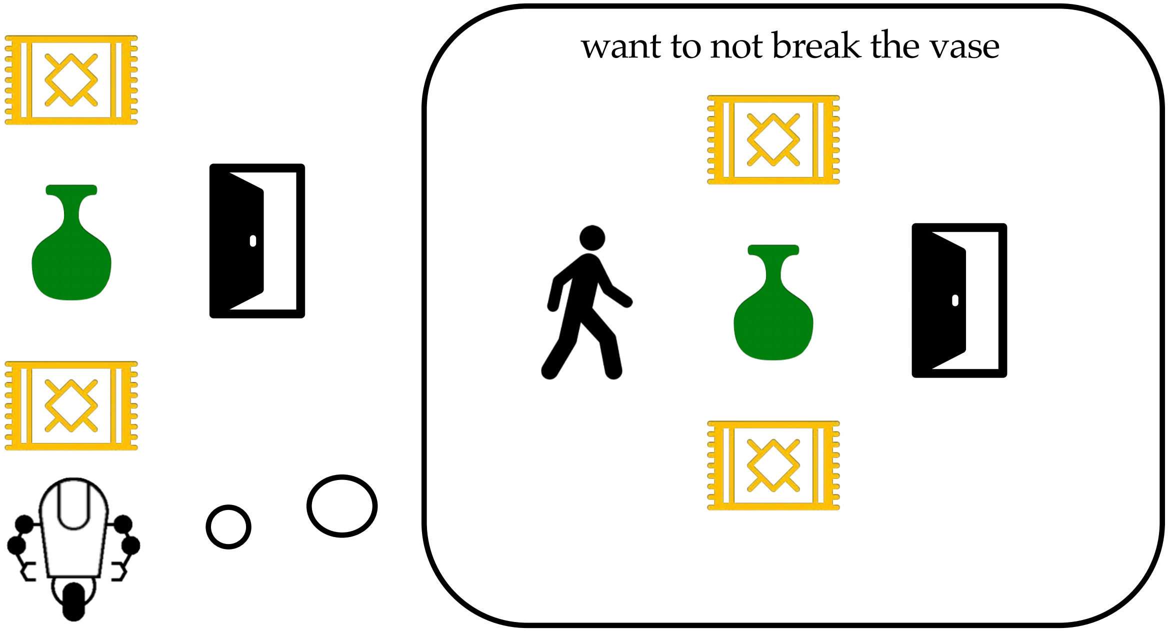
What reward function is the state consistent with?



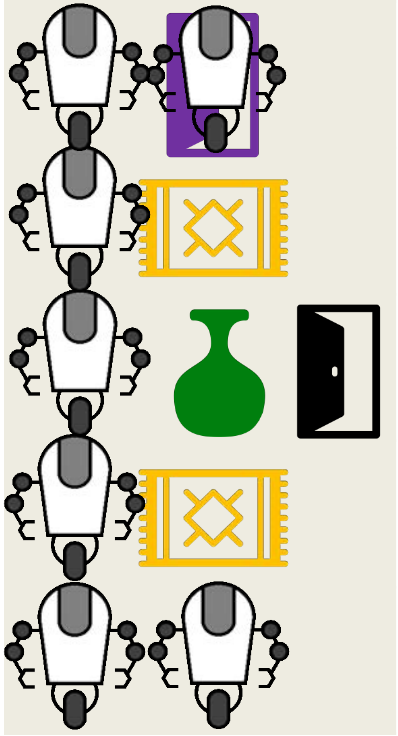
What reward function is the state consistent with?



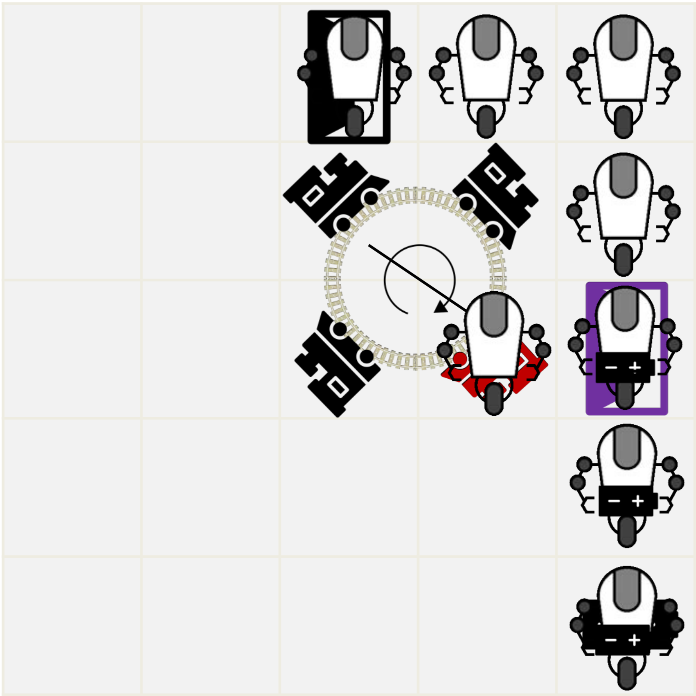
What reward function is the state consistent with?



Side effects: Room with vase



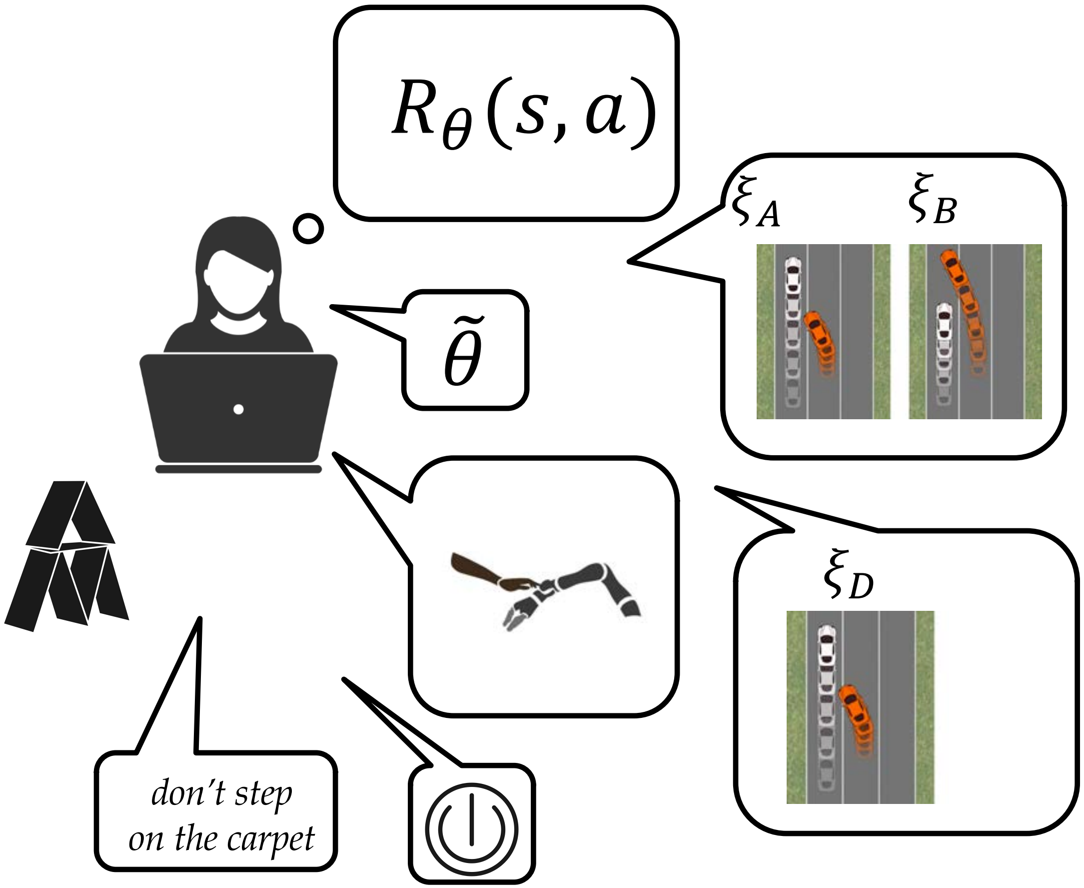
Desirable side effects: Batteries



AI $\neq$ optimize specified reward

AI $=$ optimize intended reward

Human feedback as reward-rational implicit choice



observation (human) model

$$b'(\theta) \propto b(\theta) P(\quad | \theta)$$

choices: $\{\, c \}$

choose based on reward: $\mathbb{E}[R_\theta(\xi) | \xi \sim \psi(c^*)]$

vs

$\mathbb{E}[R_\theta(\xi) | \xi \sim \psi(c)] \forall c$

$$P(c^* | \theta) = \frac{e^{\mathbb{E}[R_\theta(\xi) | \xi \sim \psi(c^*)]}}{\sum e^{\mathbb{E}[R_\theta(\xi) | \xi \sim \psi(c)]}}$$

Agents overlearn from specified rewards,
but leave other information on the table.

We can read the right amount of information into each source by interpreting them as reward-rational implicit choices.

Agents overlearn from specified rewards,
but leave other information on the table.

We can read the right amount of information into each source by interpreting them as reward-rational implicit choices.

???

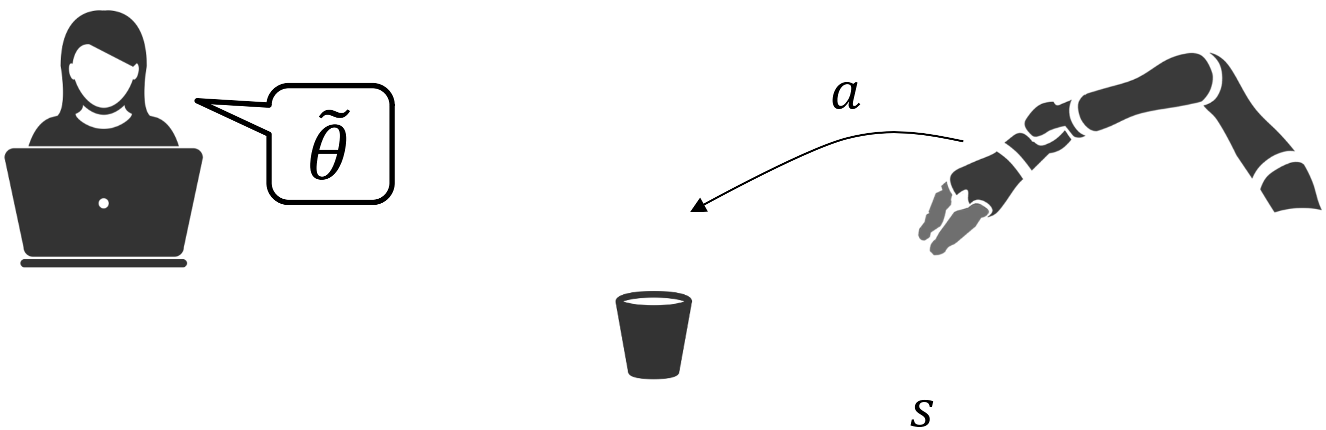
?????

optimization, search, constraint satisfaction, satisficing, RL…

task specification ⟶ behavior

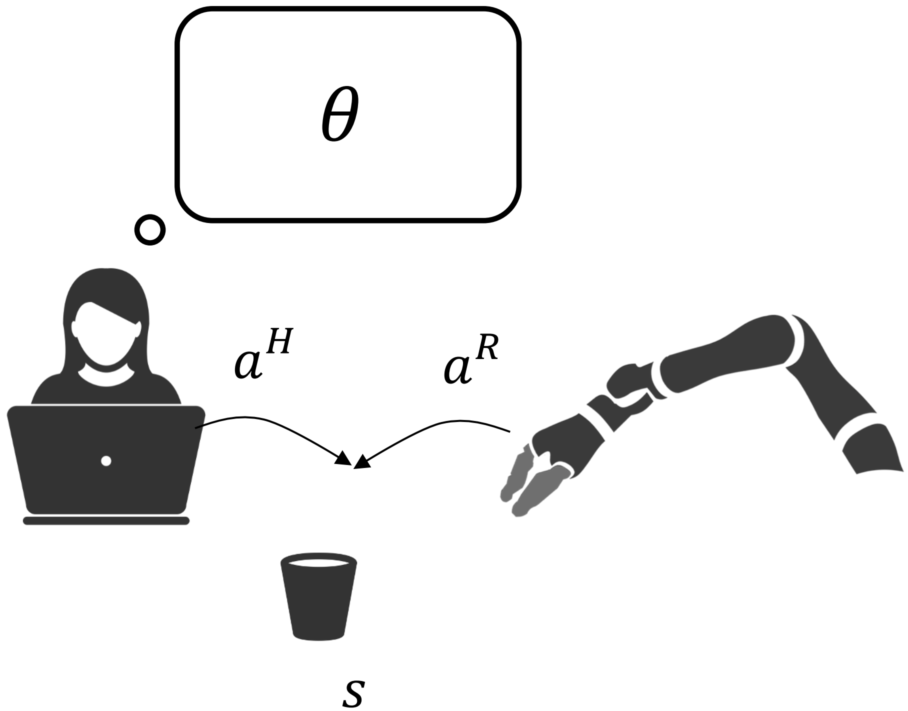
cost, reward, loss, constraints,…

$$\max \mathbb{E}[\textstyle\sum_t R_{\tilde{\theta}}(s_t, a_t)]$$



Assistance Games

$$\max \mathbb{E}[\sum_t R_\theta(s_t, a_t^R, a_t^H)]$$



Thanks!