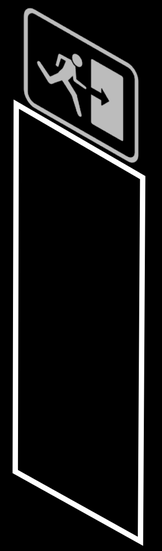
An Optimization-Centric Theory of Mind for Human-Robot Interaction

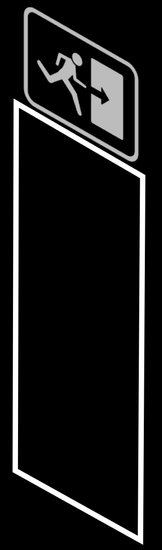
Anca Dragan

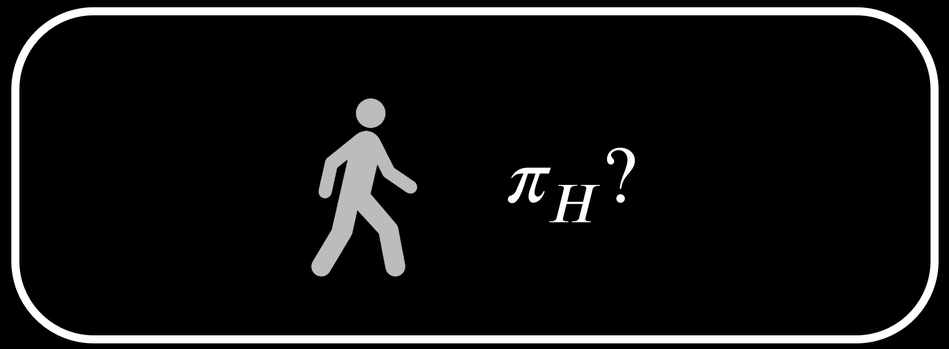

Human
Being
100%
ORGANIC

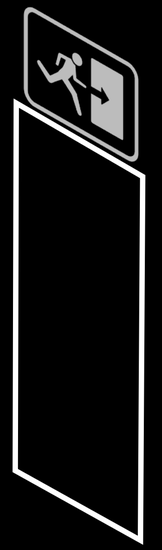
$$\max_{\pi_R} \mathbb{E}[U(\pi_R)]$$

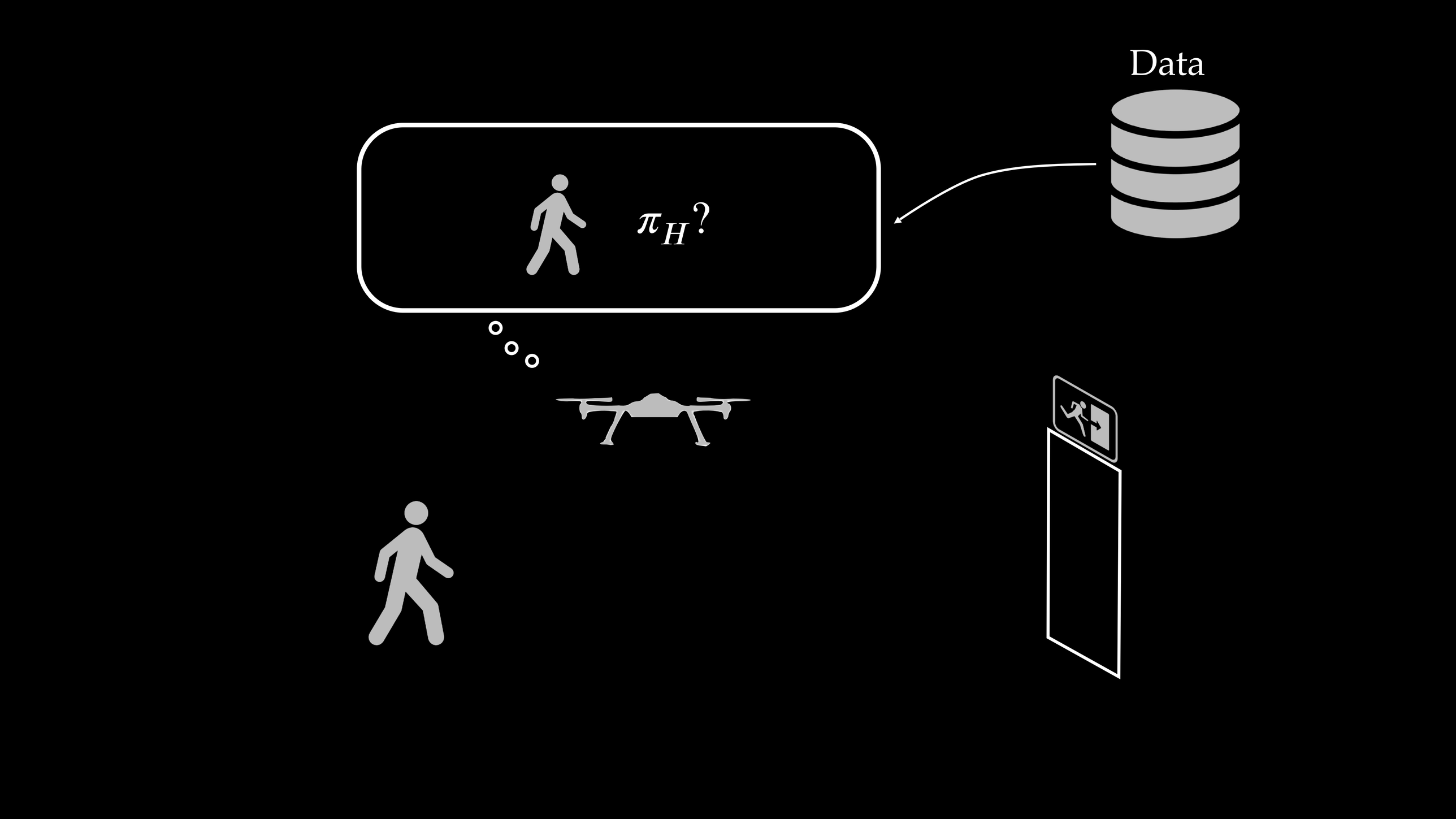


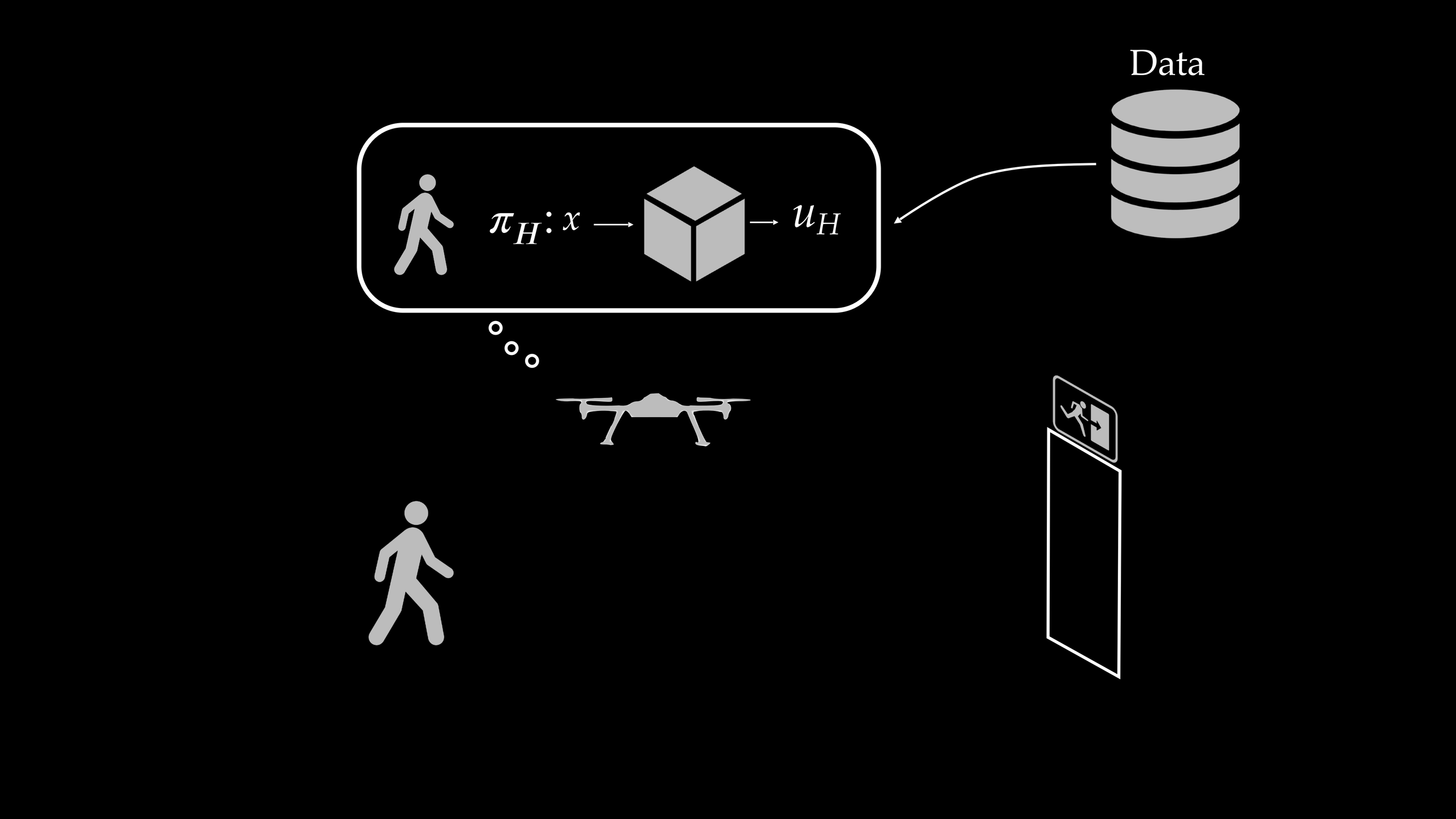
$$\max_{\pi_R} \mathbb{E}[U(\pi_R, \pi_H)]$$



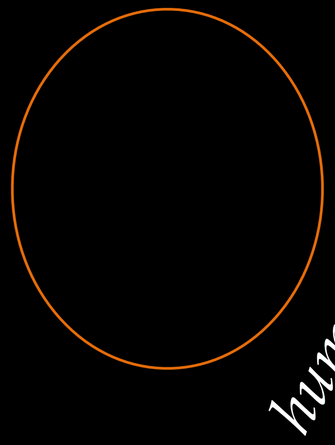

π_H?



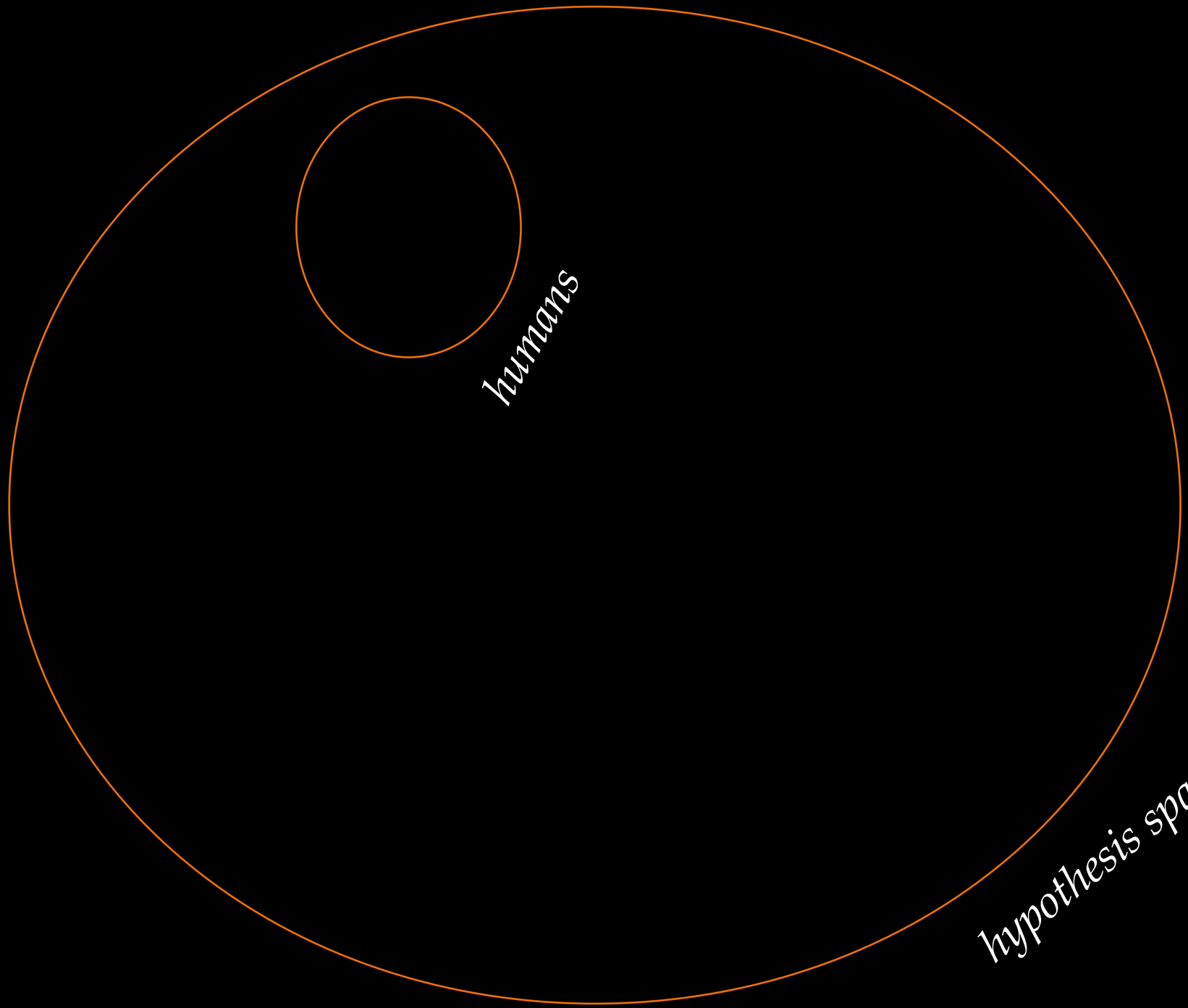

Data
π_H?

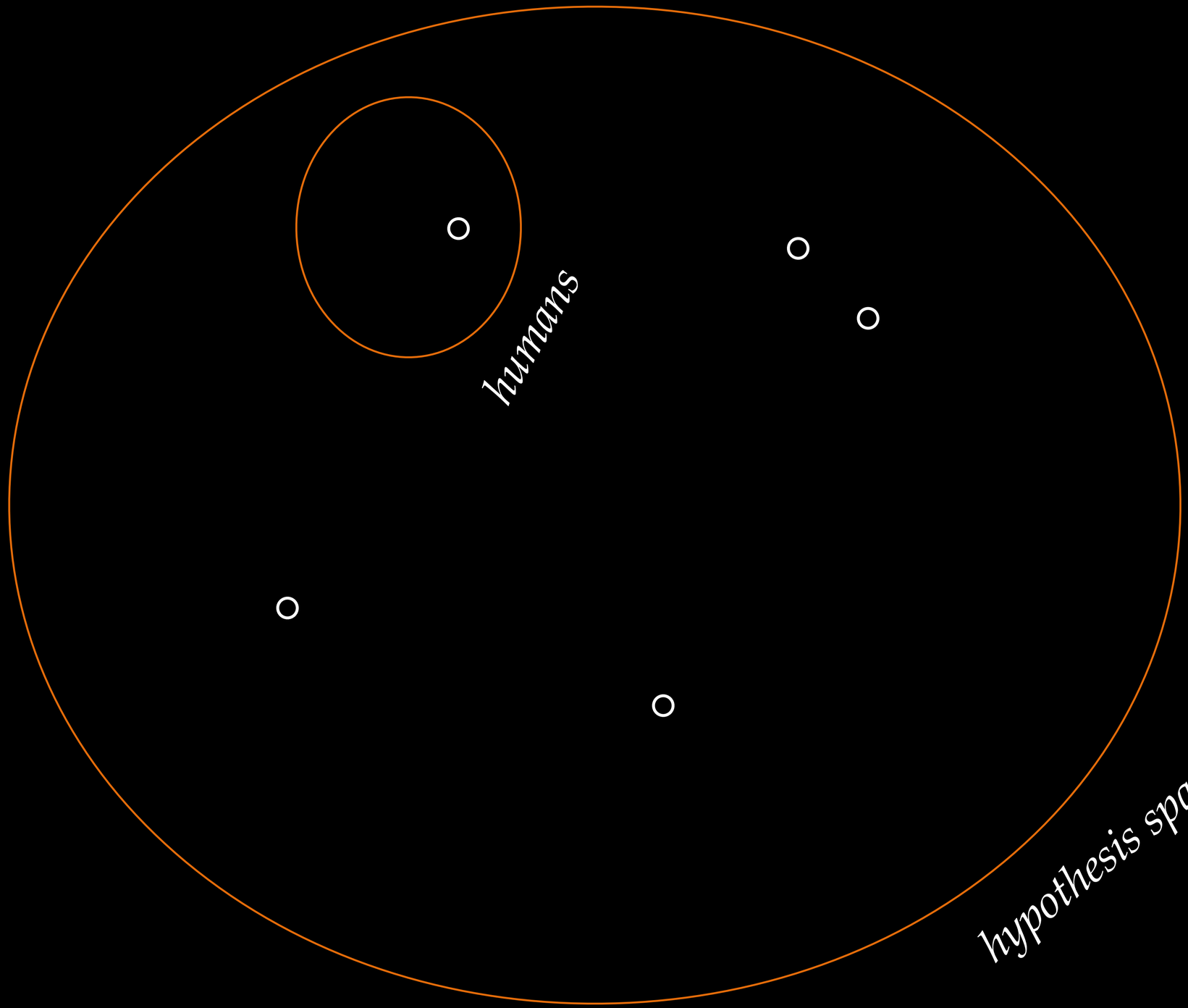

Data
$\pi_H : x \rightarrow$
$\rightarrow u_H$

all policies

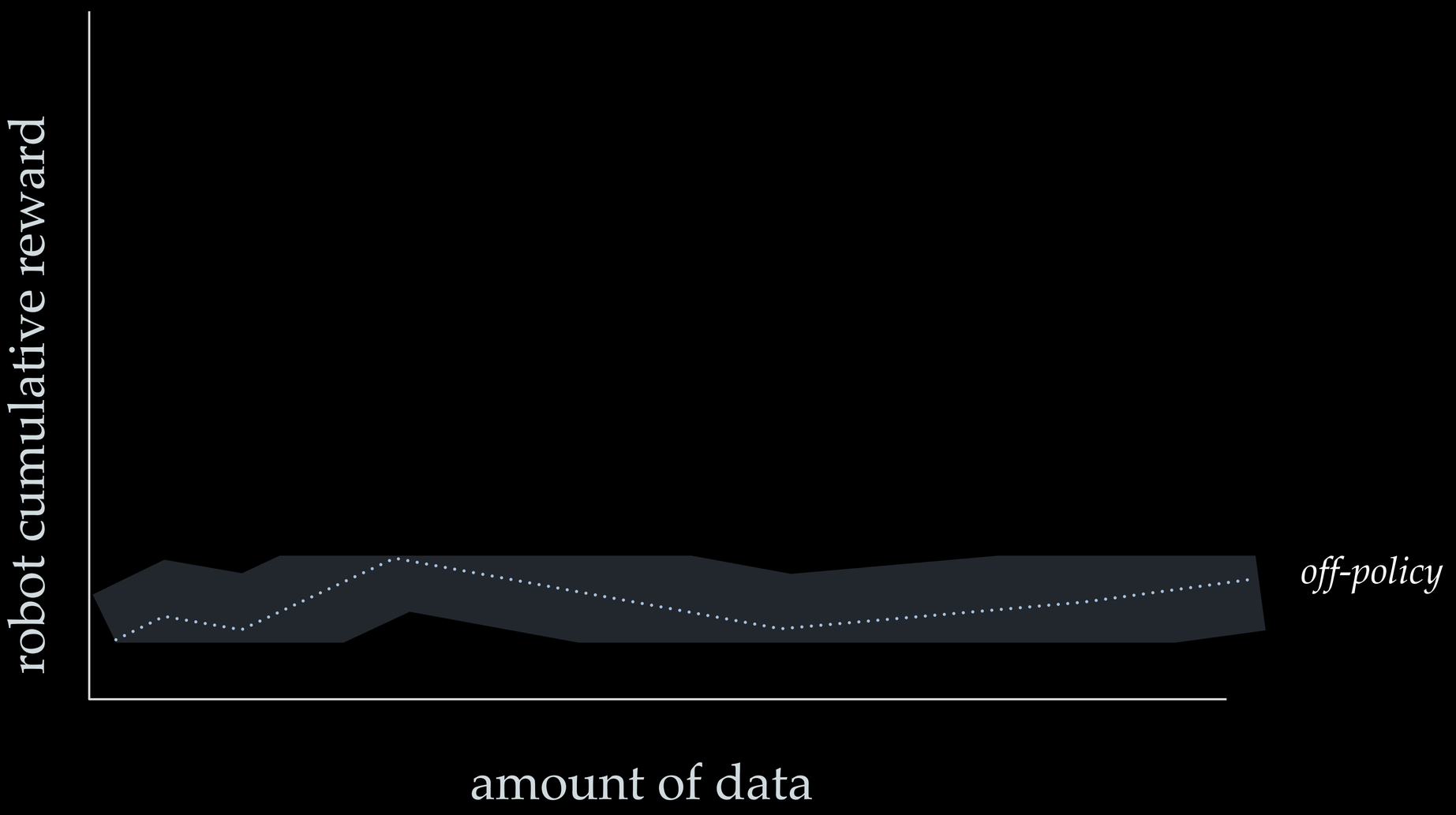


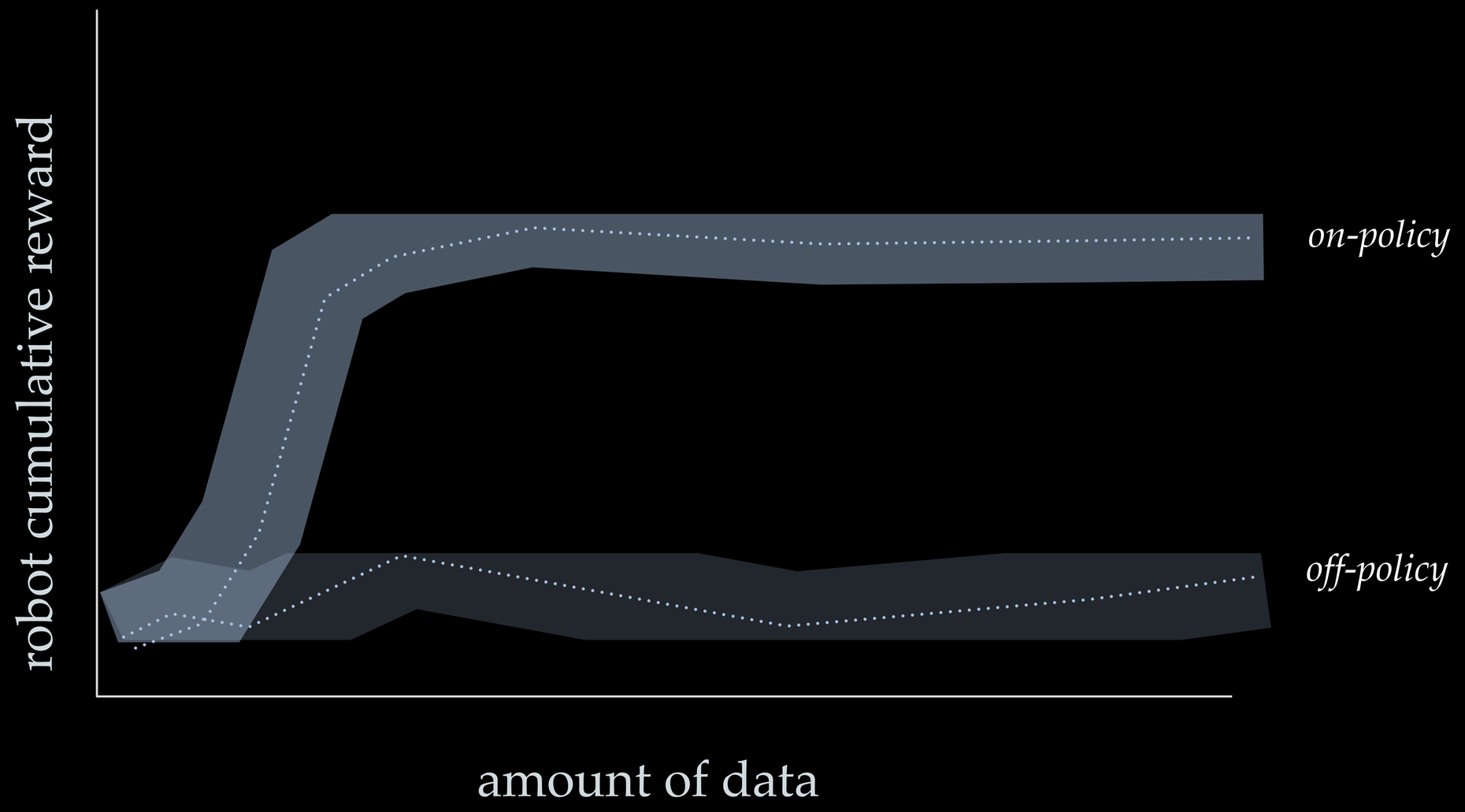
all policies

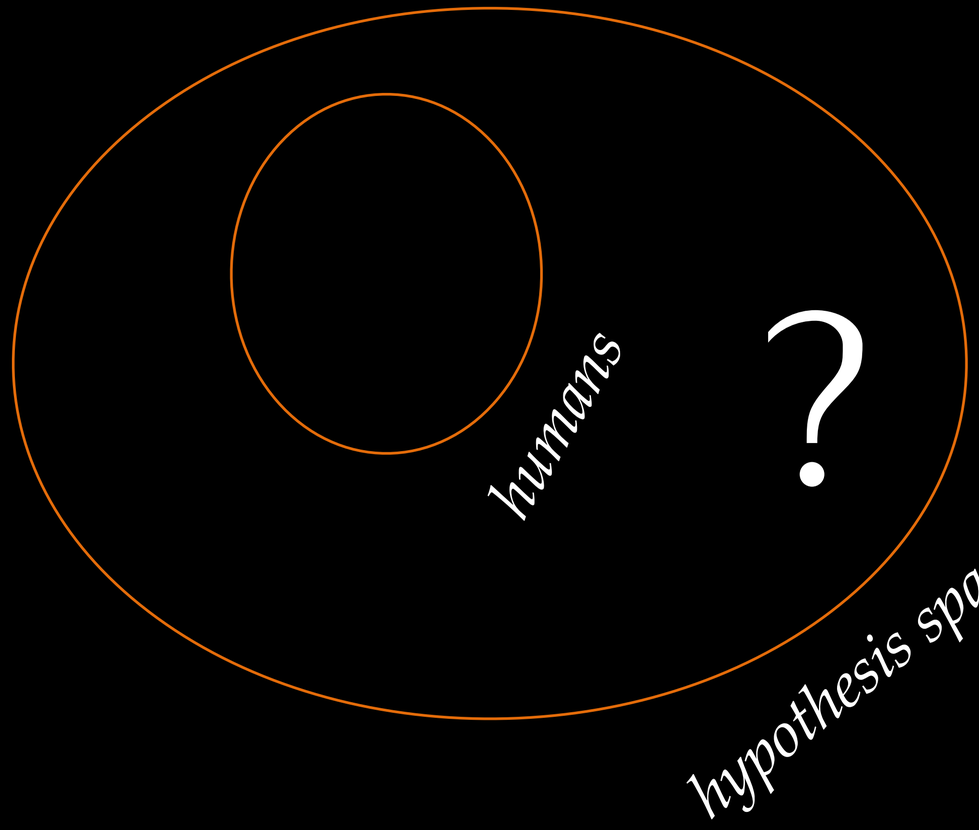

humans
hypothesis space
all policies


humans
hypothesis space
all policies


robot cumulative reward
amount of data


robot cumulative reward
amount of data
off-policy

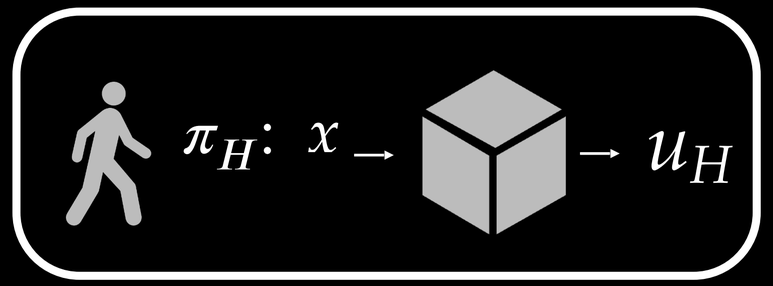

robot cumulative reward
amount of data
on-policy
off-policy

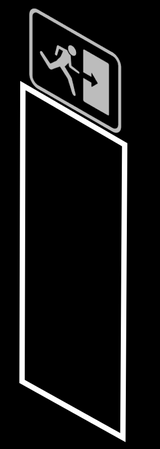


all policies

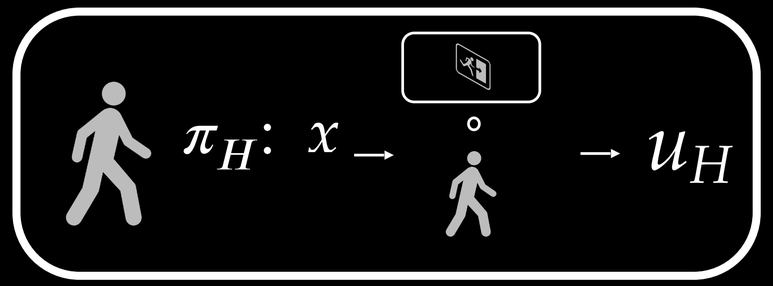
What is the right inductive bias for HRI?

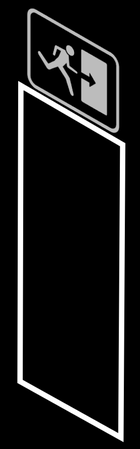
Humans as black-box policies





Humans as intent-driven agents

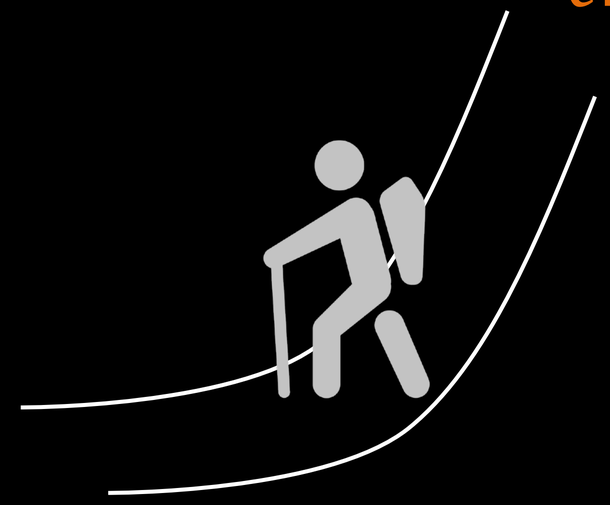




Humans have intent

Humans have intent

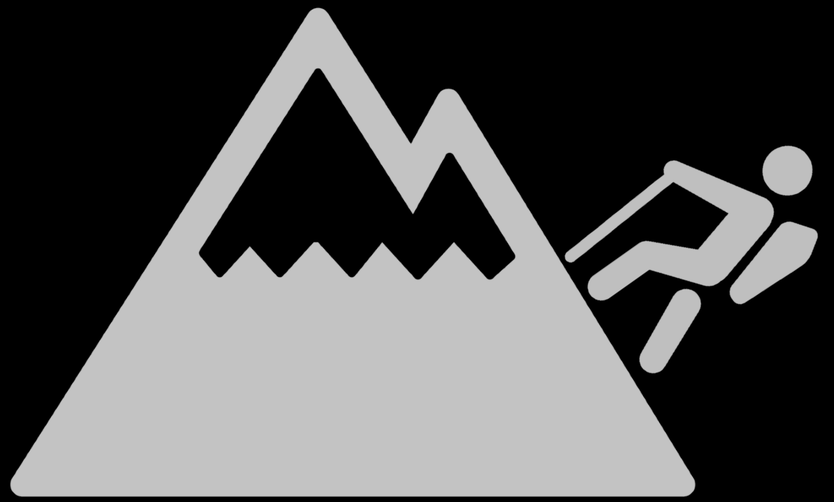
intent



inductive bias

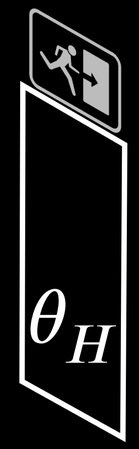
Humans have intent

inductive bias



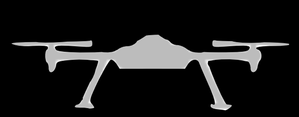
intent via utility

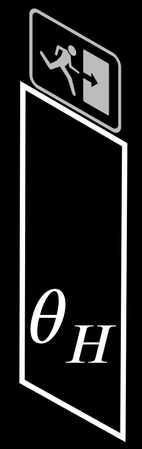
$$U_H(x, u_H; \theta_H)$$



intent via utility

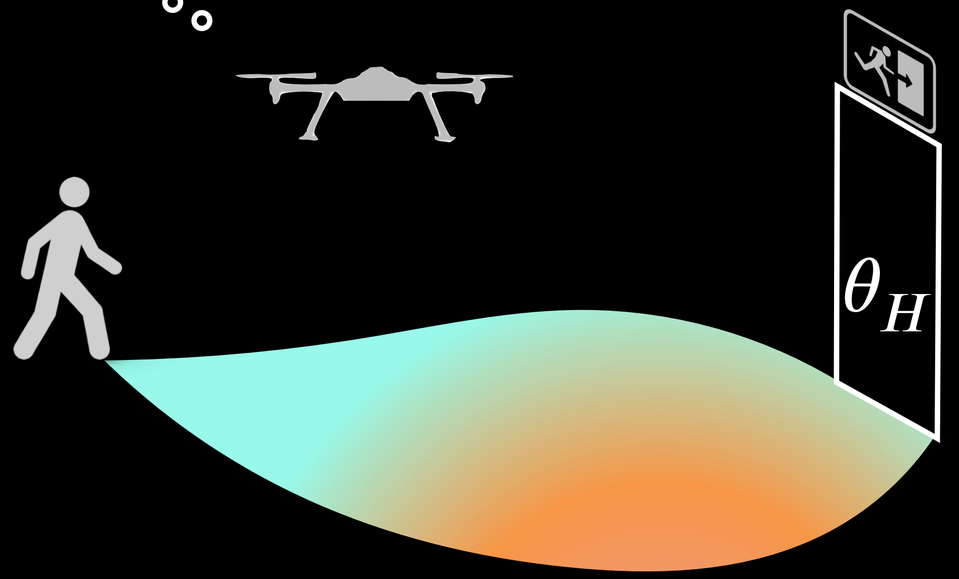
$$\max_{P} \; H(P)$$
$$s.t. \; \mathbb{E}[Q_{\theta_H}] = Q^*_{\theta_H} - \epsilon$$





Humans as noisy-rational agents

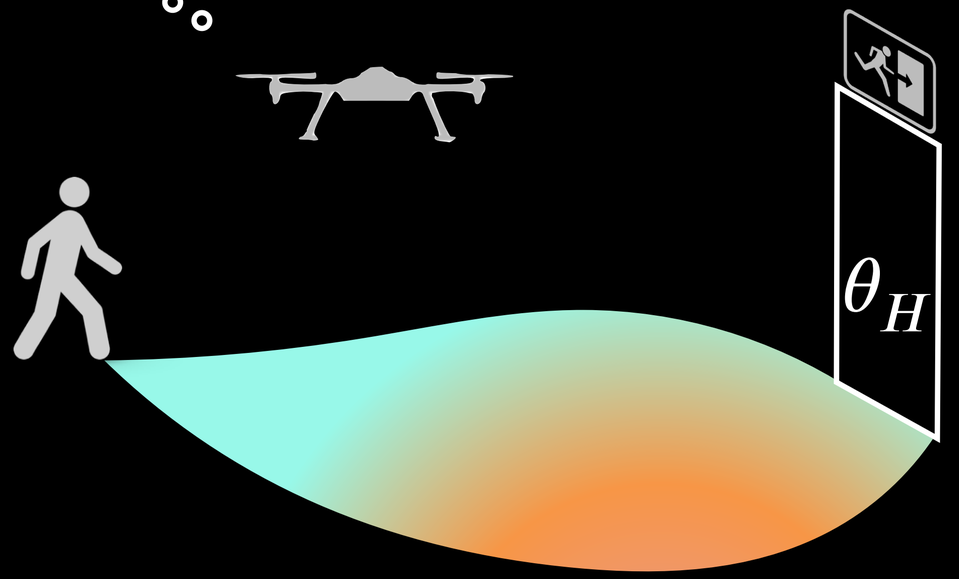
$$P(u_H \mid x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$$



Humans as noisy-rational agents

don't learn all this

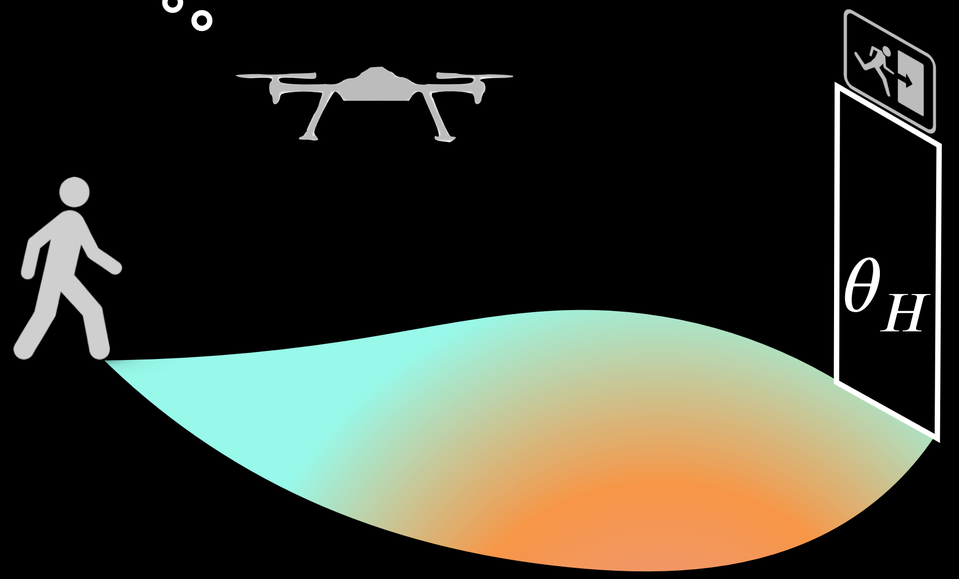
$$P(u_H \mid x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$$



Humans as noisy-rational agents with unknown utilities

learn this

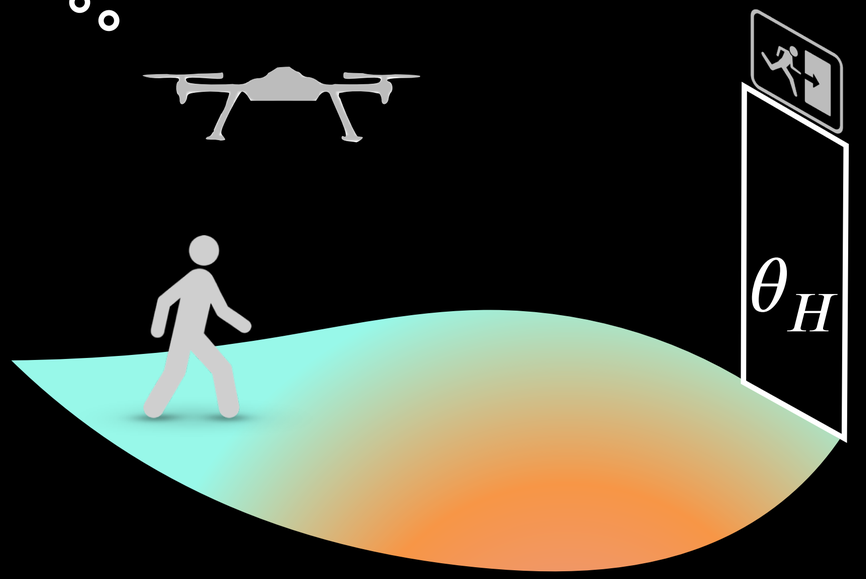
$$P(u_H \mid x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$$



Humans as noisy-rational agents with unknown utilities

learn this

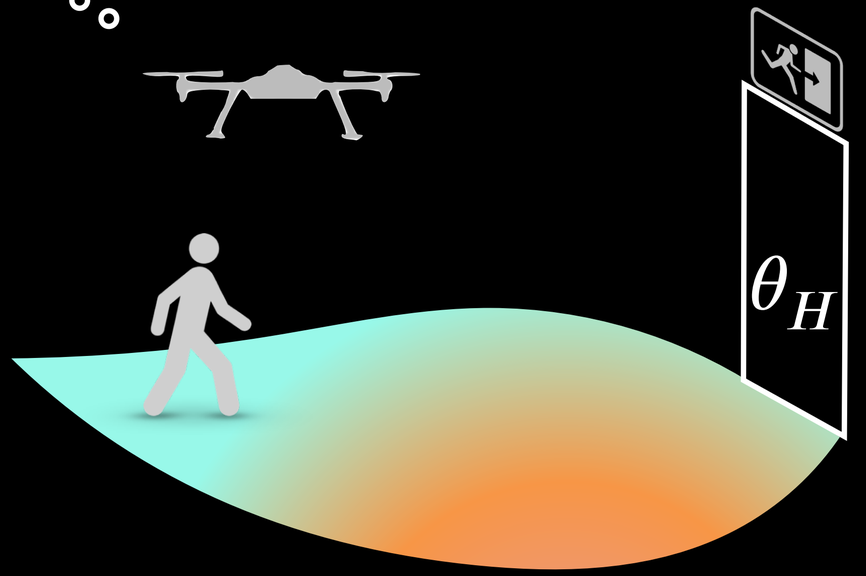
$$P(u_H \mid x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$$



Humans as noisy-rational agents with unknown utilities

$$P(u_H \mid x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$$

$$b'(\theta_H) \propto b(\theta_H) P(u_H \mid x, \theta_H)$$



Human's goal

Quadcopter

Quadcopter's goal

$t+2$

$t+1$

Human

Noisy-rationality as a unifying way to interpret behavior

action (demonstration)

$$P(u_H | x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$$

Noisy-rationality as a unifying way to interpret behavior

action (demonstration) $u_H > u \forall u$ $P(u_H | x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$

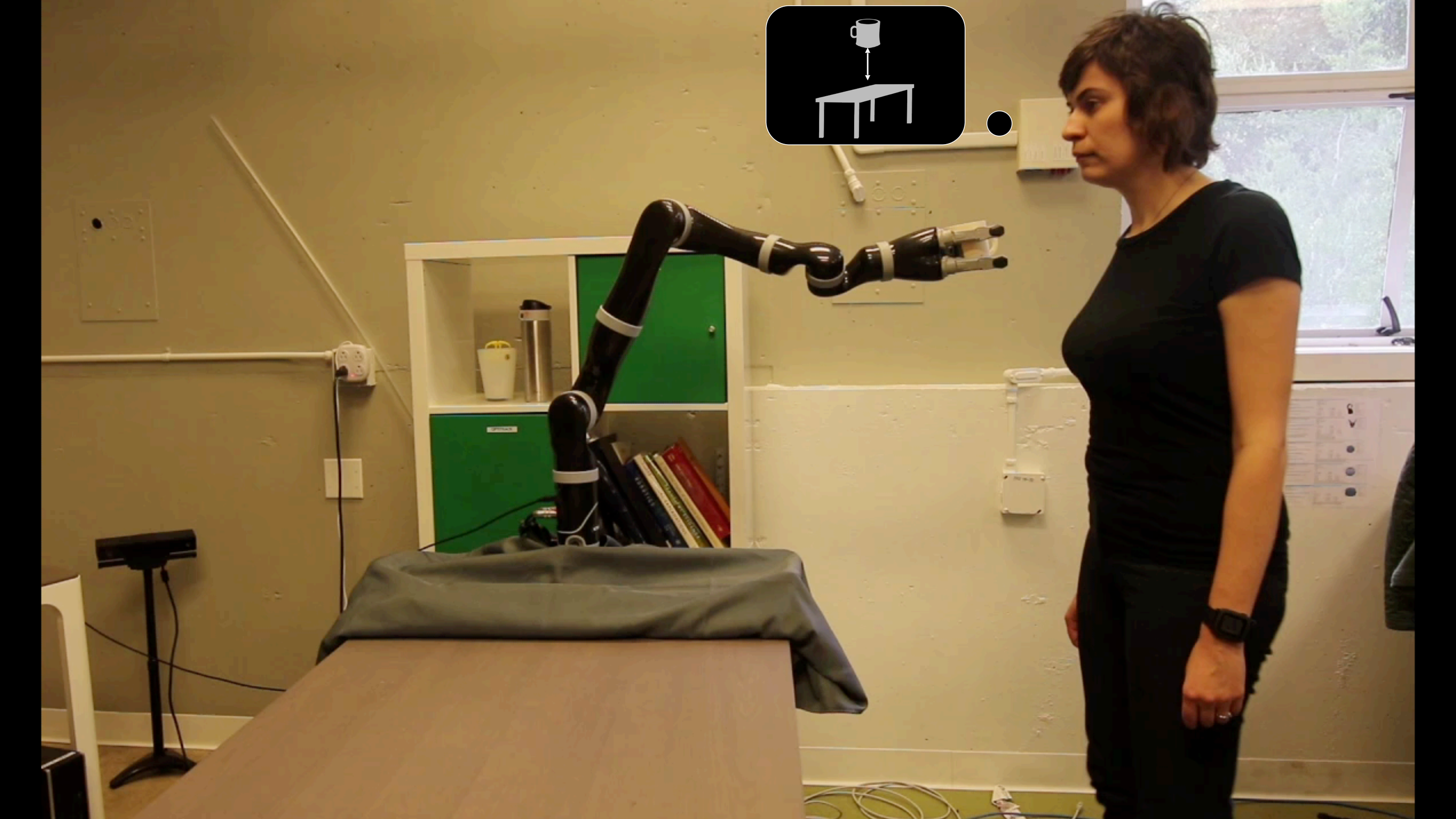
Noisy-rationality as a unifying way to interpret behavior

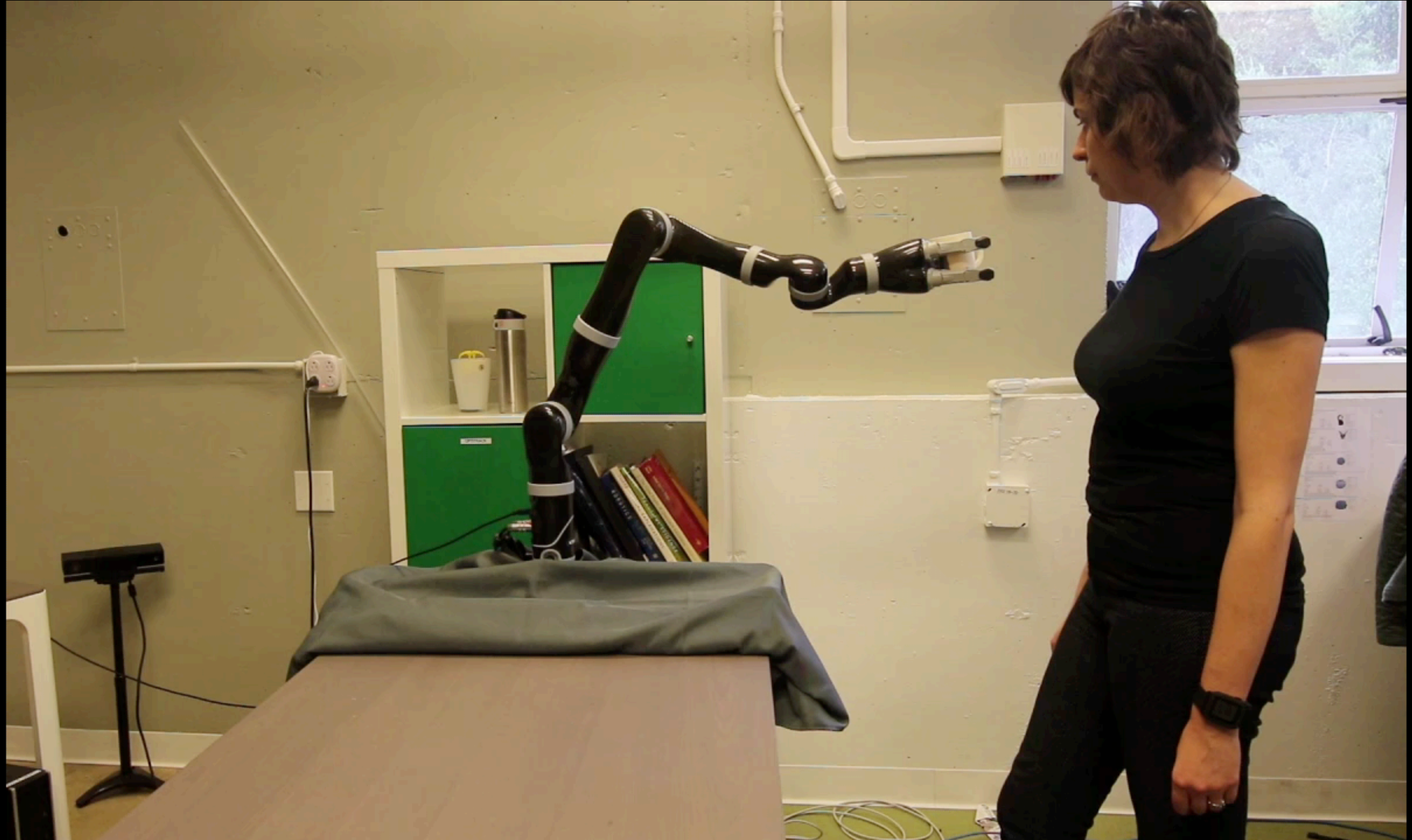
action (demonstration) $u_H > u \forall u$

$$P(u_H | x, \theta_H) = \frac{e^{\beta Q(x, u_H; \theta_H)}}{\int e^{\beta Q(x, u; \theta_H)} du}$$

Noisy-rationality as a unifying way to interpret behavior

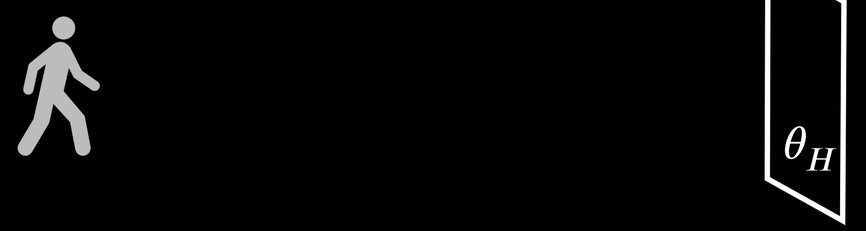
action (demonstration)	$u_H > u \forall u$	$P(u_H \mid x, \theta_H) = \frac{e^{\beta Q(x,u_H;\theta_H)}}{\int e^{\beta Q(x,u;\theta_H)} du}$
comparison	$u_A > u_B$	$P(u_A \mid x, u_A, u_B, \theta_H) = \frac{e^{\beta Q(x,u_A;\theta_H)}}{e^{\beta Q(x,u_A;\theta_H)} + e^{\beta Q(x,u_B;\theta_H)}}$
correction	$u_H + u_R > u_R$	$P(u_H \mid x, u_R, \theta_H) = \frac{e^{\beta Q(x,u_H+u_R;\theta_H)}}{e^{\beta Q(x,u_R;\theta_H)} + e^{\beta Q(x,u_R+u_H;\theta_H)}}$
stop	$u_0 > u_R$	$P(u_0 \mid x, u_R, \theta_H) = \frac{e^{\beta Q(x,u_0;\theta_H)}}{e^{\beta Q(x,u_0;\theta_H)} + e^{\beta Q(x,u_R;\theta_H)}}$
proxy reward	$\tilde{\theta} > \theta \forall \theta$	$P(\tilde{\theta} \mid \theta_H, x) = \frac{e^{\beta Q(x,u^*_{\tilde{\theta}};\theta_H)}}{\int e^{\beta Q(x,u^*_{\bar{\theta}};\theta_H)} d\bar{\theta}}$
current state	$x_H > x_T \forall x_T$	$P(x_H \mid \theta) = \int P(u_{0..T} \mid x_0, \theta) dx_0 du_{0..T} \; s.t. \; x_T = x_H$



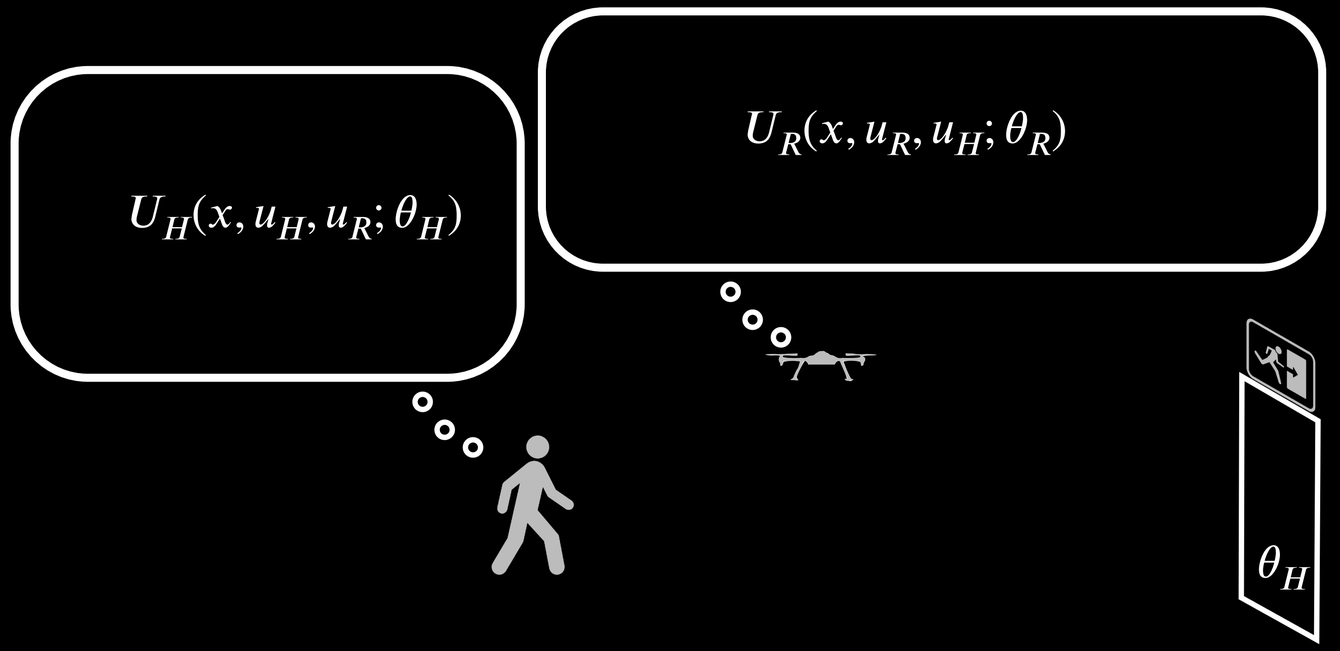


missing the future robot actions

$$P(u_H \mid x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$$



Noisy-rationality generalizes to dynamic games



Strategic level: dynamic programming

$$V_R(x, T+1) = 0 \quad V_H(x, T+1) = 0$$

$t = T \ldots 0$

$\forall x$

$\forall u_R$

$\forall u_H$

$$Q_H(x, u_H, u_R) = U_H(x, u_H, u_R)] + V(x', t+1)$$

$$P(u_H \mid x, u_R) \propto e^{Q_H(x, u_H, u_R)}$$

$$Q_R(x, u_R) = \mathbb{E}_{u_H \mid u_R} U_R(x, u_R, u_H) + V_R(x, t+1)$$

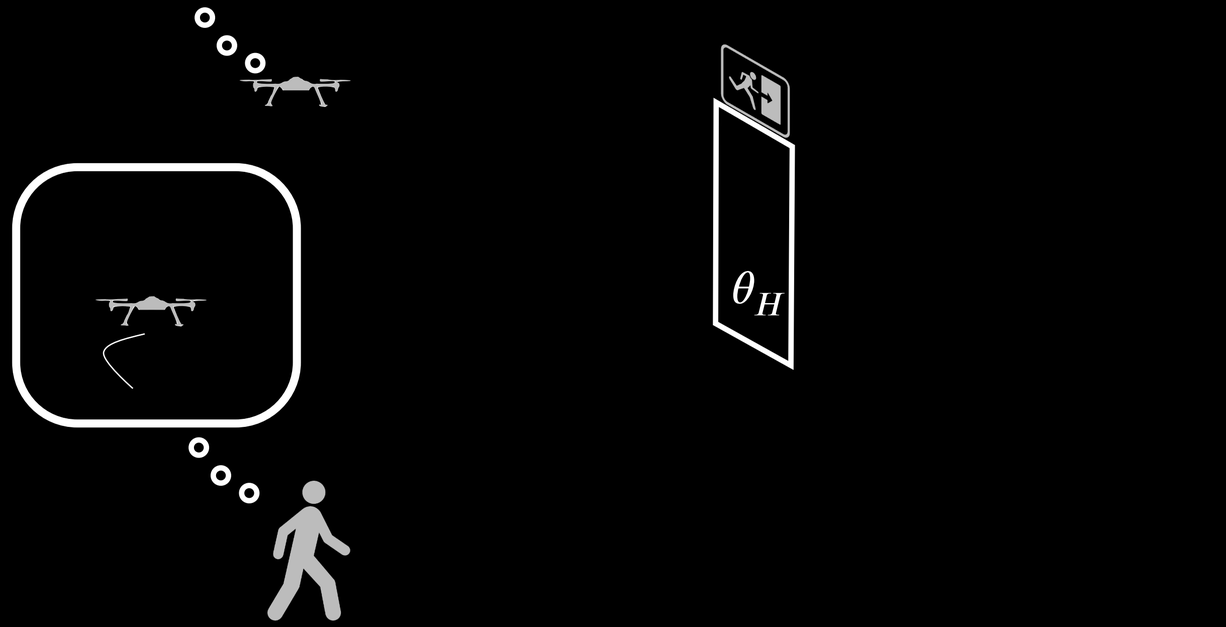
$$u_R^* = \arg\max Q_R(x, u_R) \quad V_R(x, t) = Q_R(x, u_R^*)$$

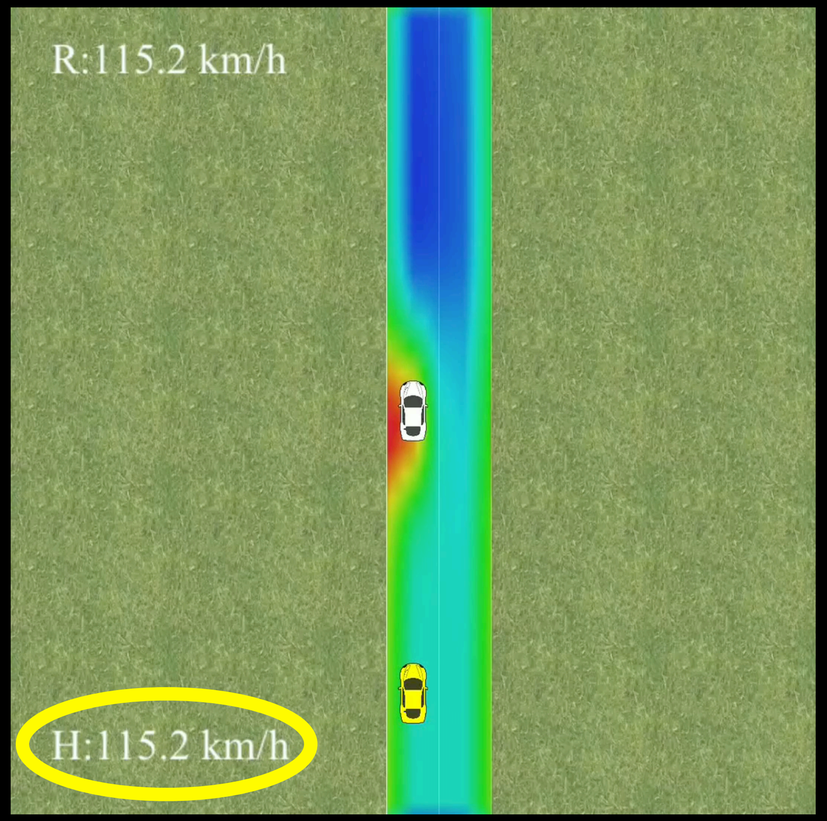
$$V_H(x, t) = \mathbb{E}_{u_H \mid u_R^*} Q_H(x, u_H, u_R^*)$$

Tactical level: use strategic value

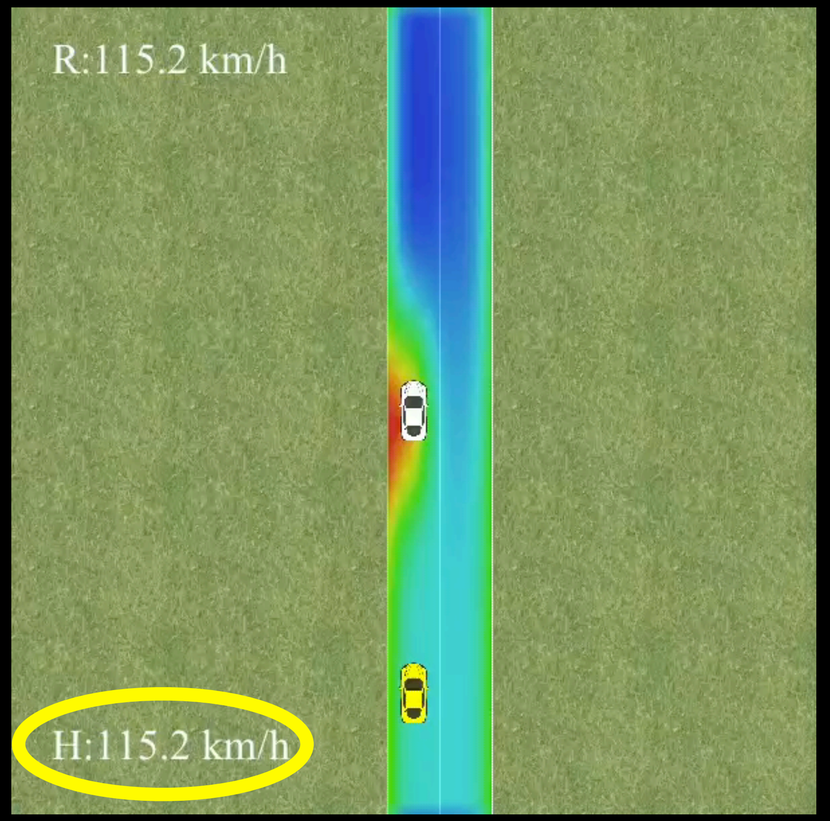
$$u_H^*(u_R) = \underset{u}{\operatorname{argmax}} U_H(x_0, u, u_R; \theta_H) \quad + V_H(g(x^T))$$

$$u_R^* = \underset{u}{\operatorname{argmax}} U_R(x_0, u_R, u_H^*(u_R); \theta_R) \quad + V_R(g(x^T))$$

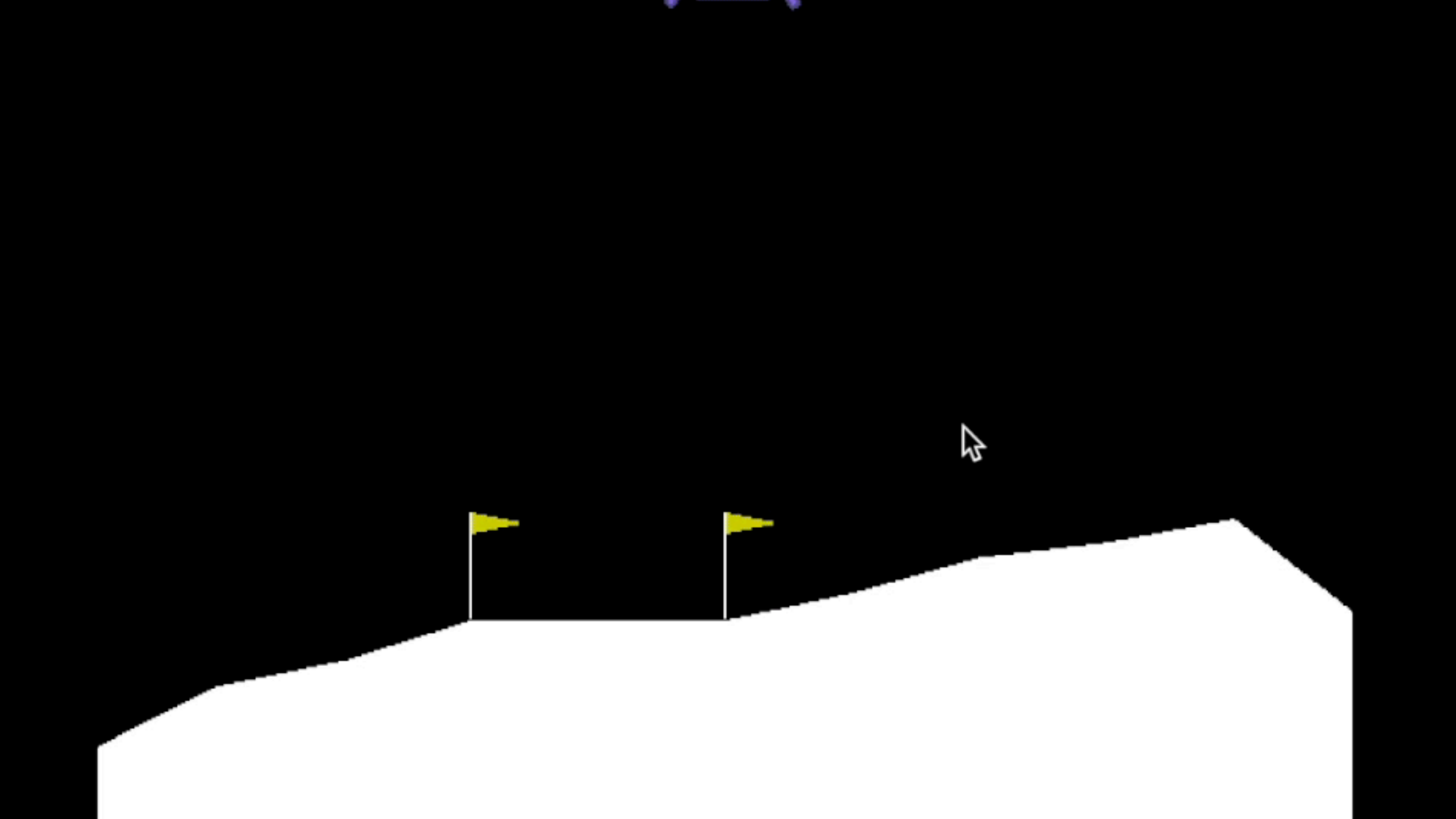


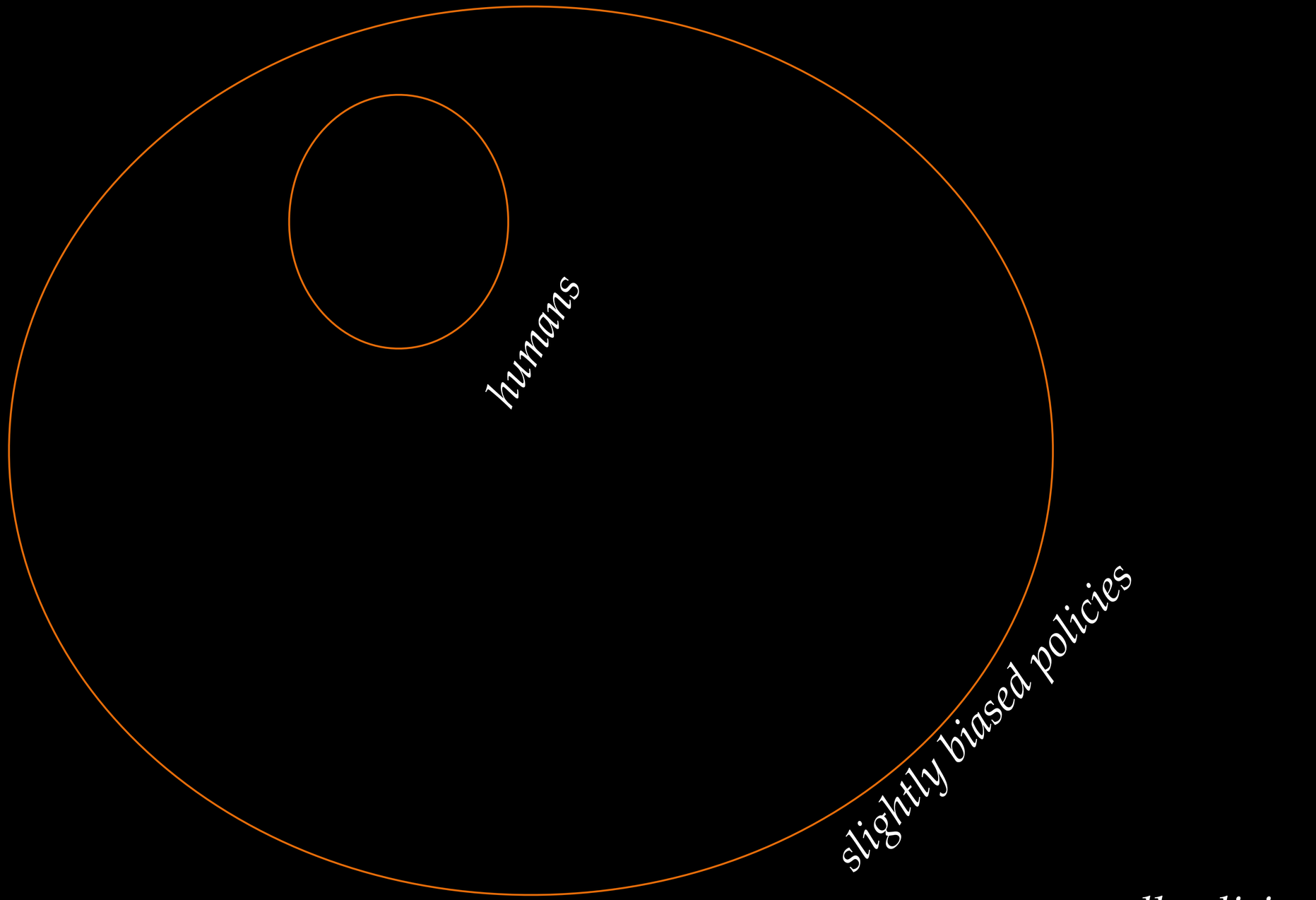


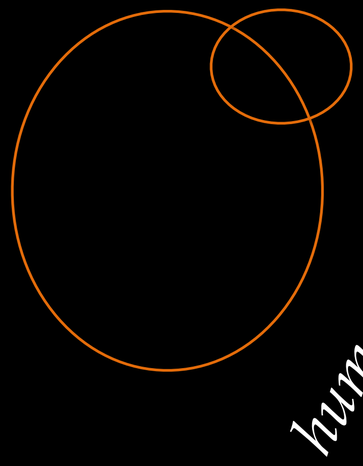
Lead vehicle decelerates



Lead vehicle accelerates



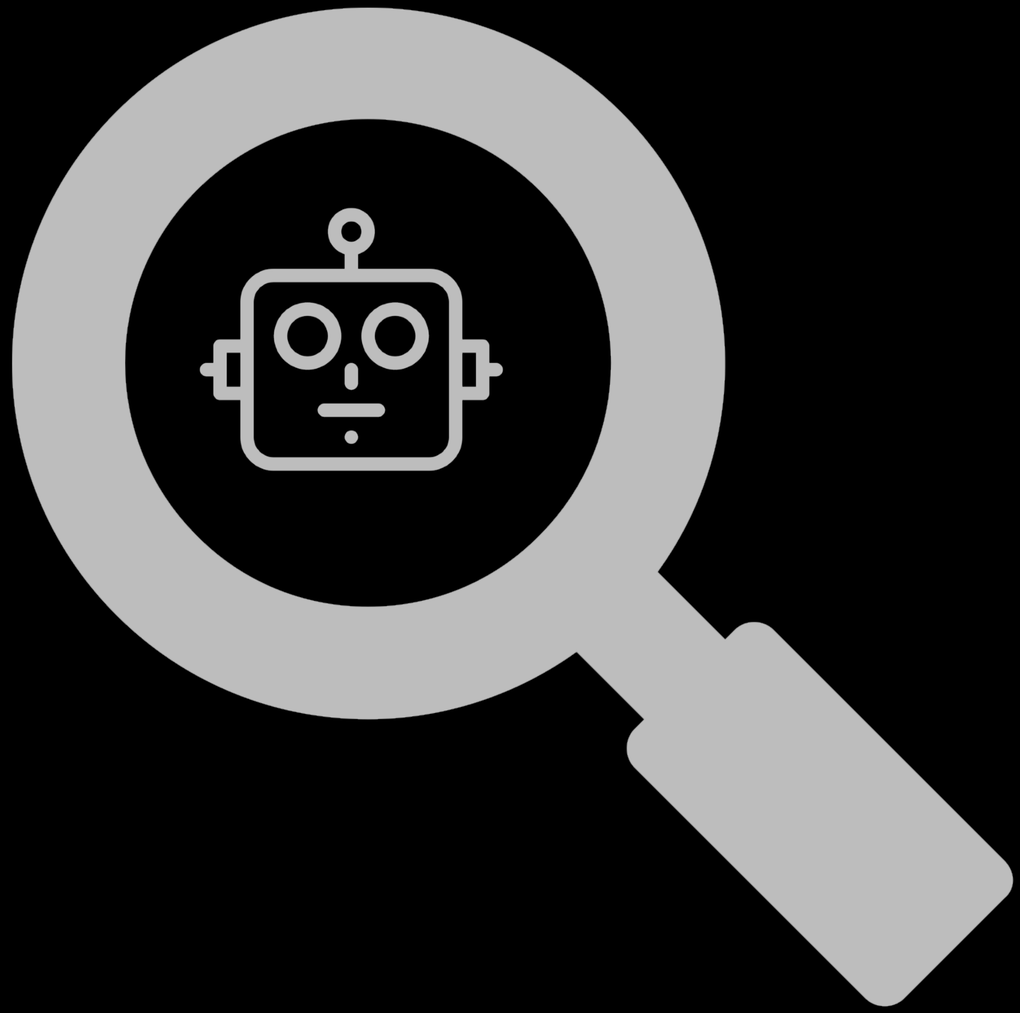

humans
slightly biased policies
all policies



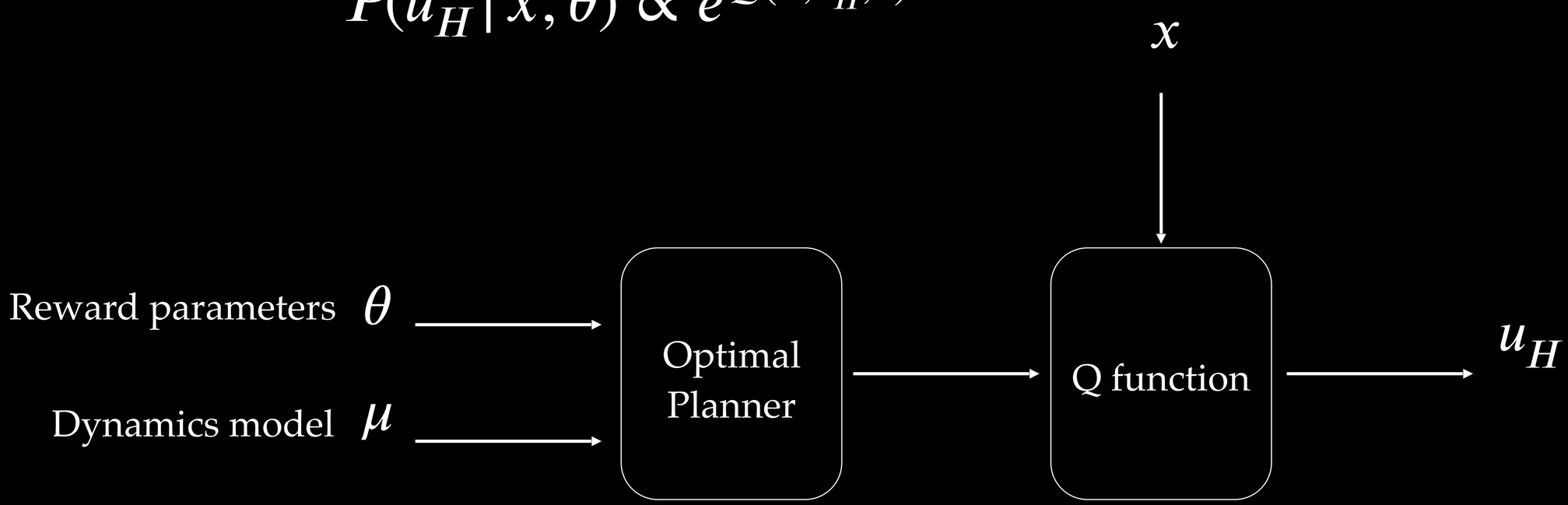
all policies

Humans have intent

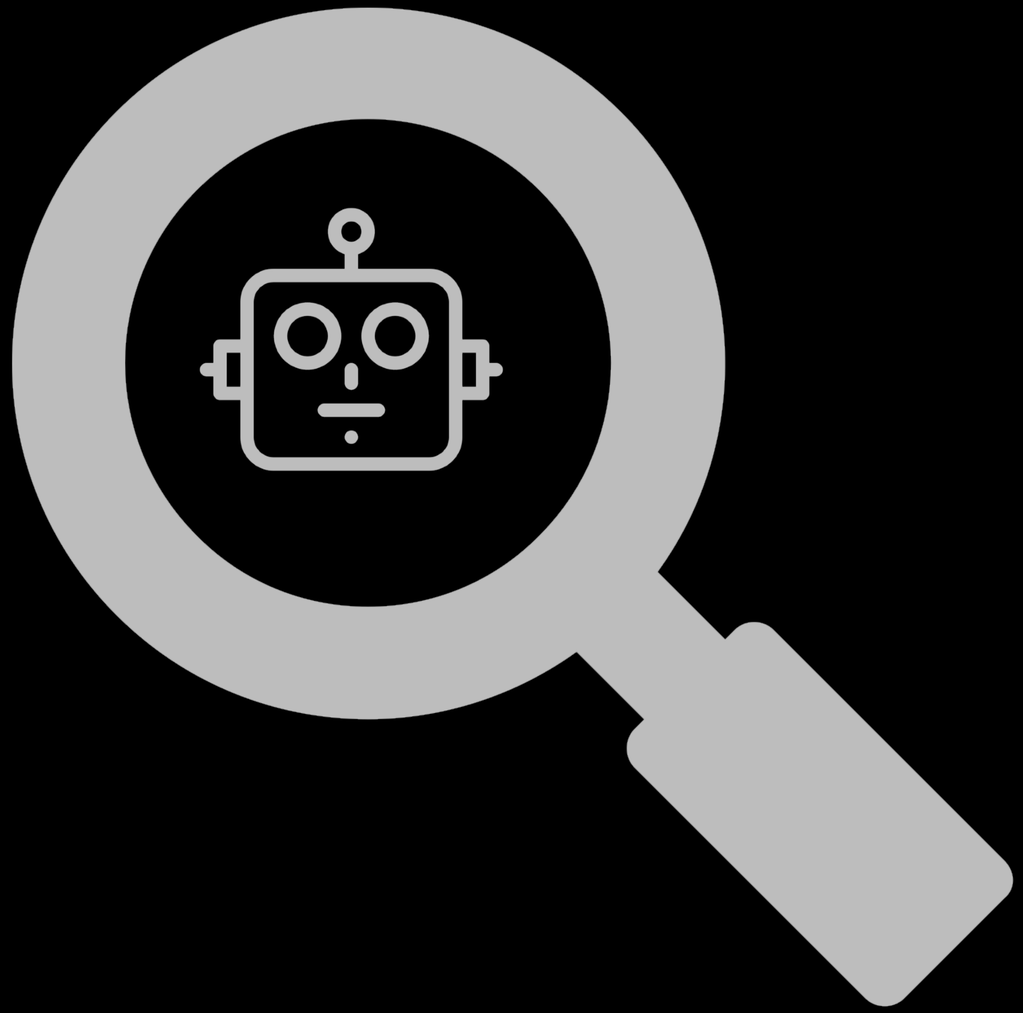
inductive bias



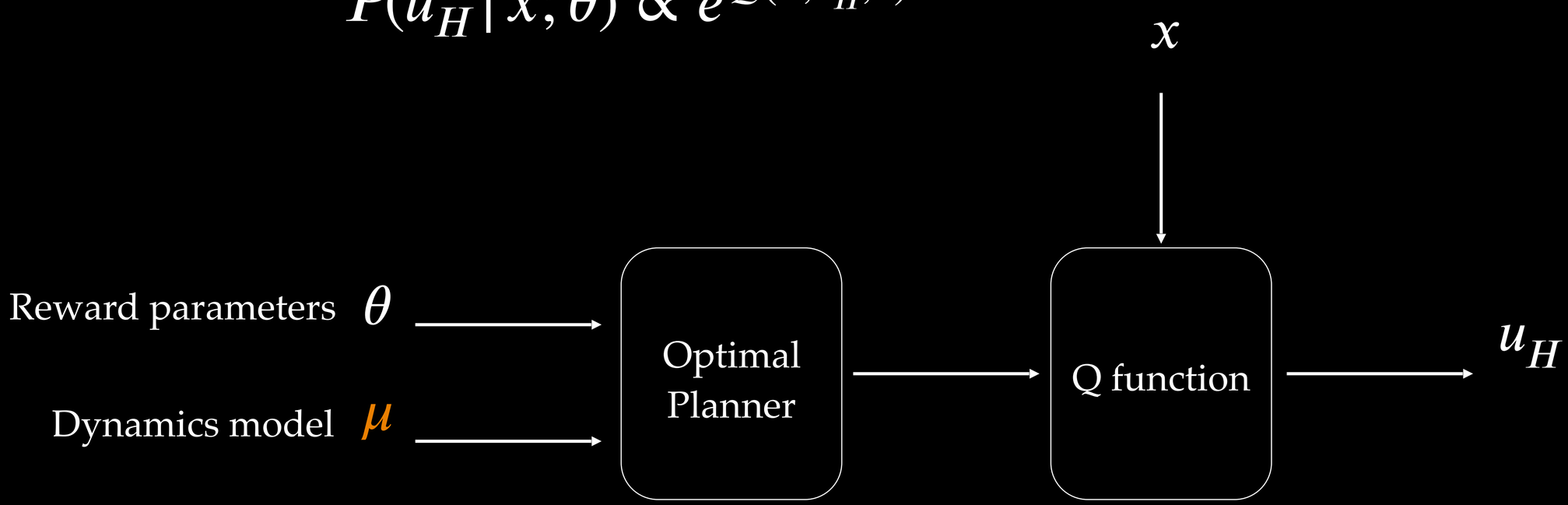
$$P(u_H \mid x; \theta) \propto e^{Q(x, u_H; \theta)}$$

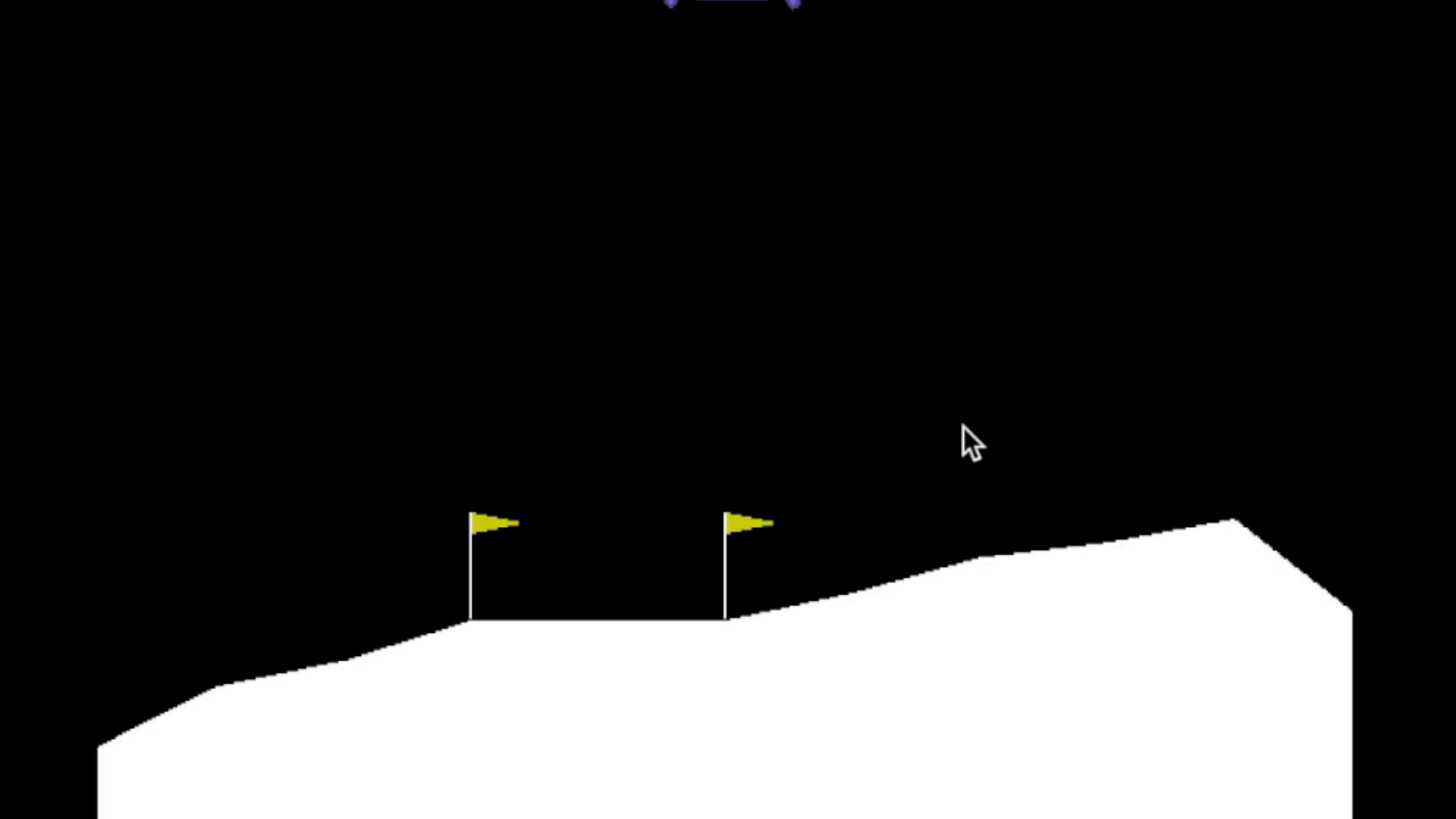


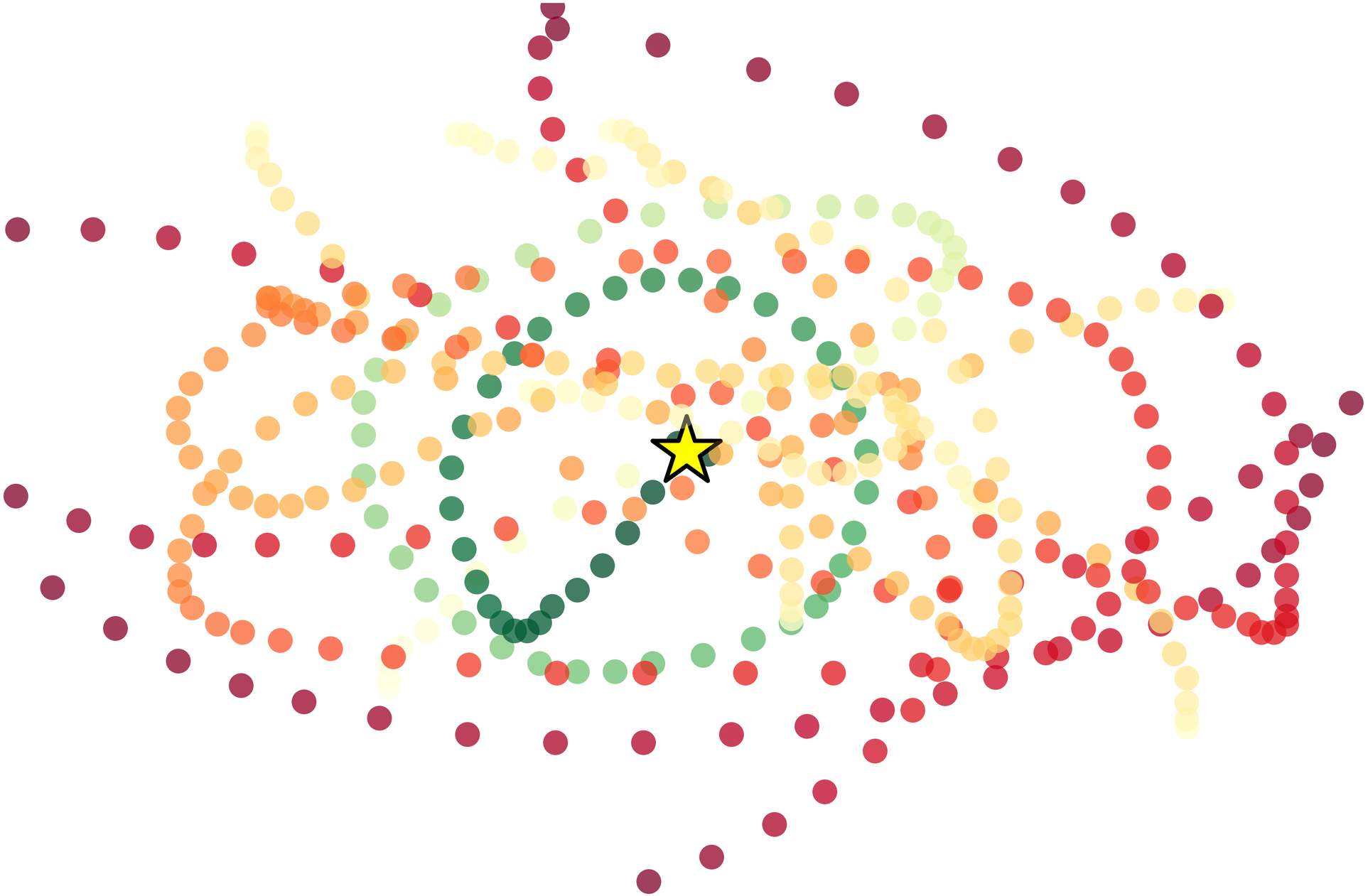
When are robots not rational?

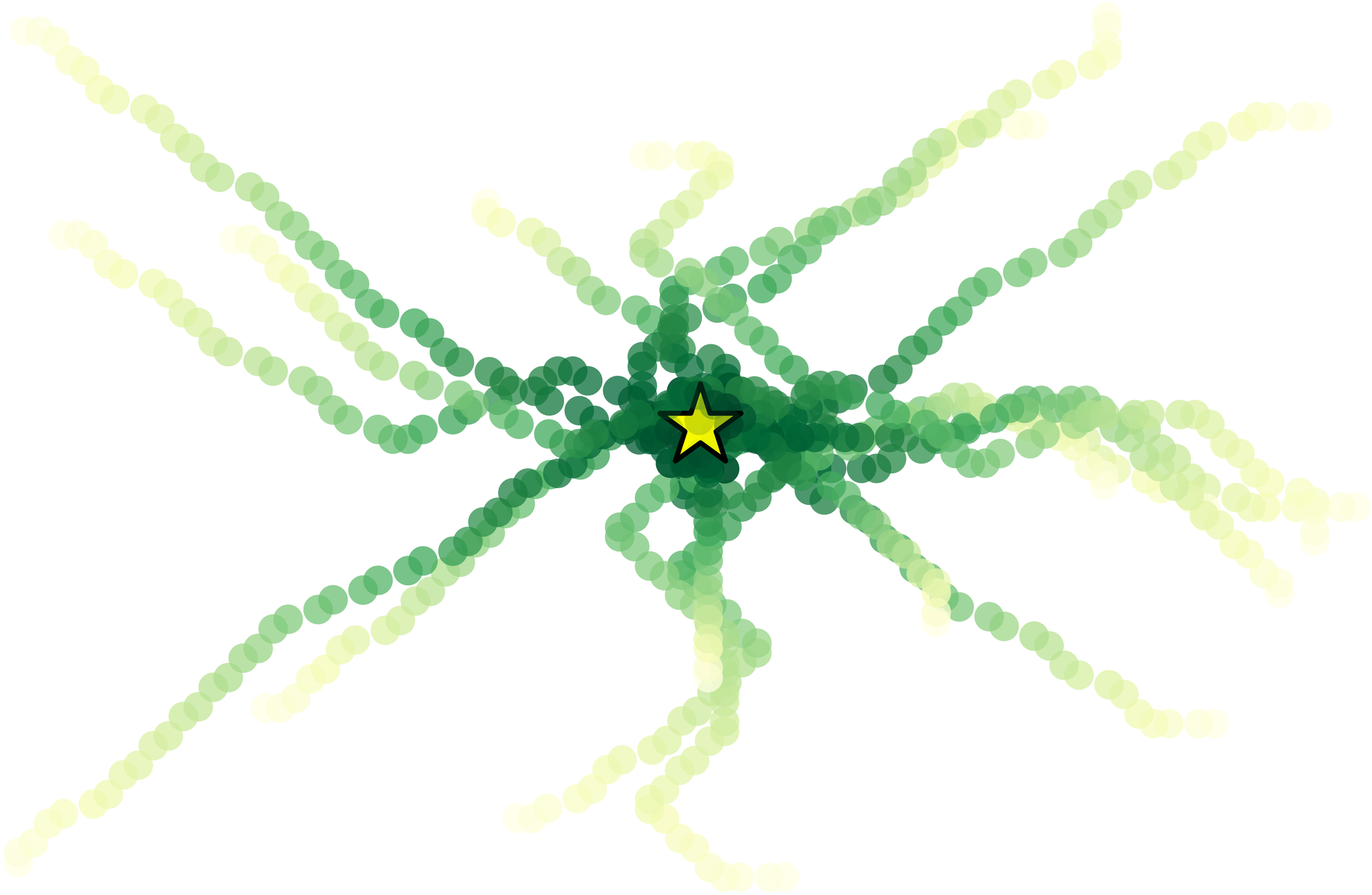


$$P(u_H \mid x; \theta) \propto e^{Q(x, u_H; \theta)}$$

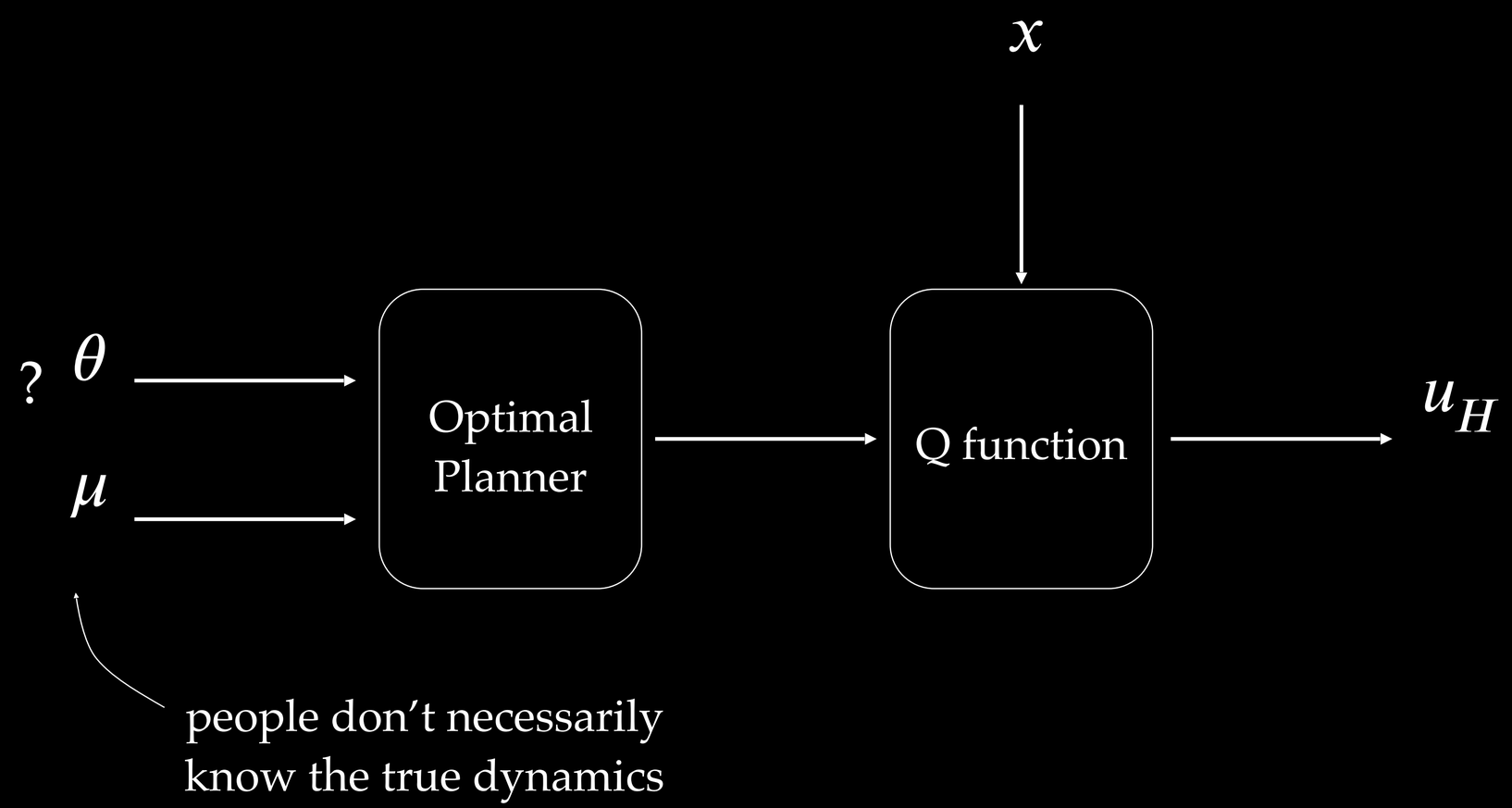




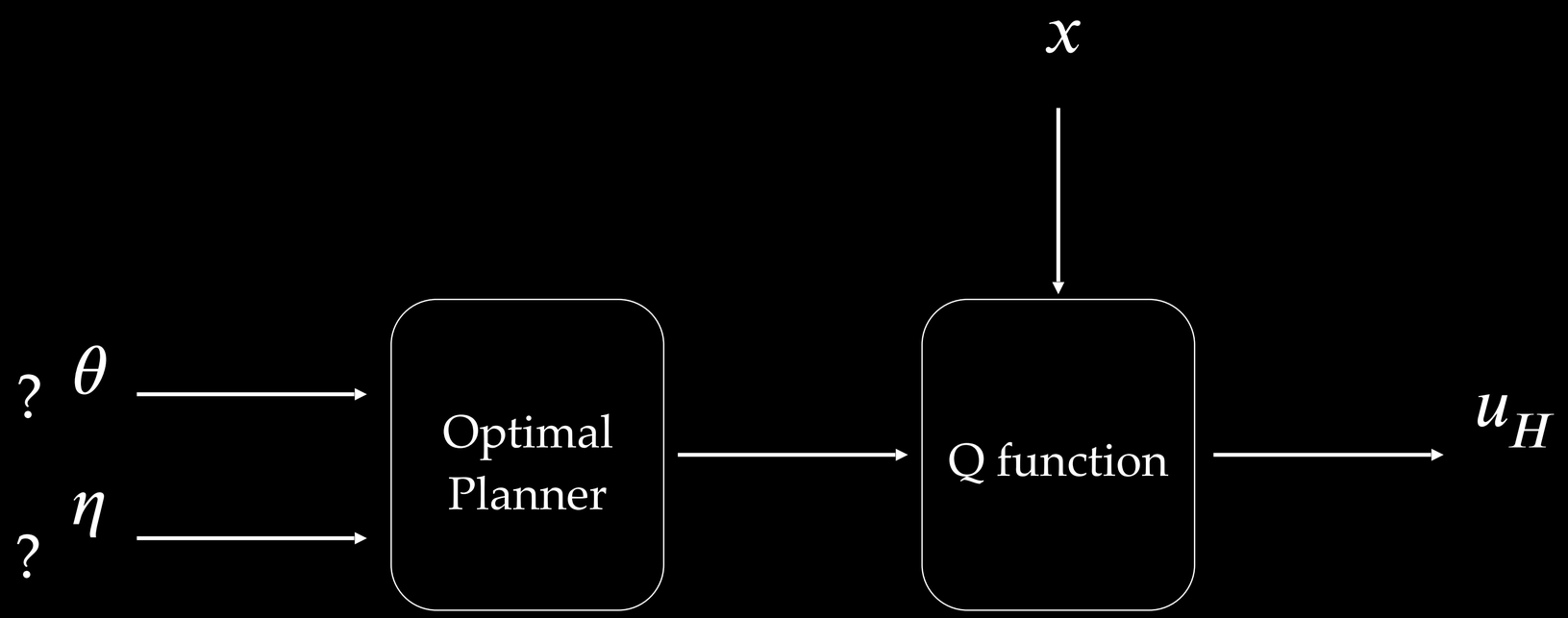




$$P(u_H \mid x; \theta) \propto e^{Q(x, u_H; \theta)}$$

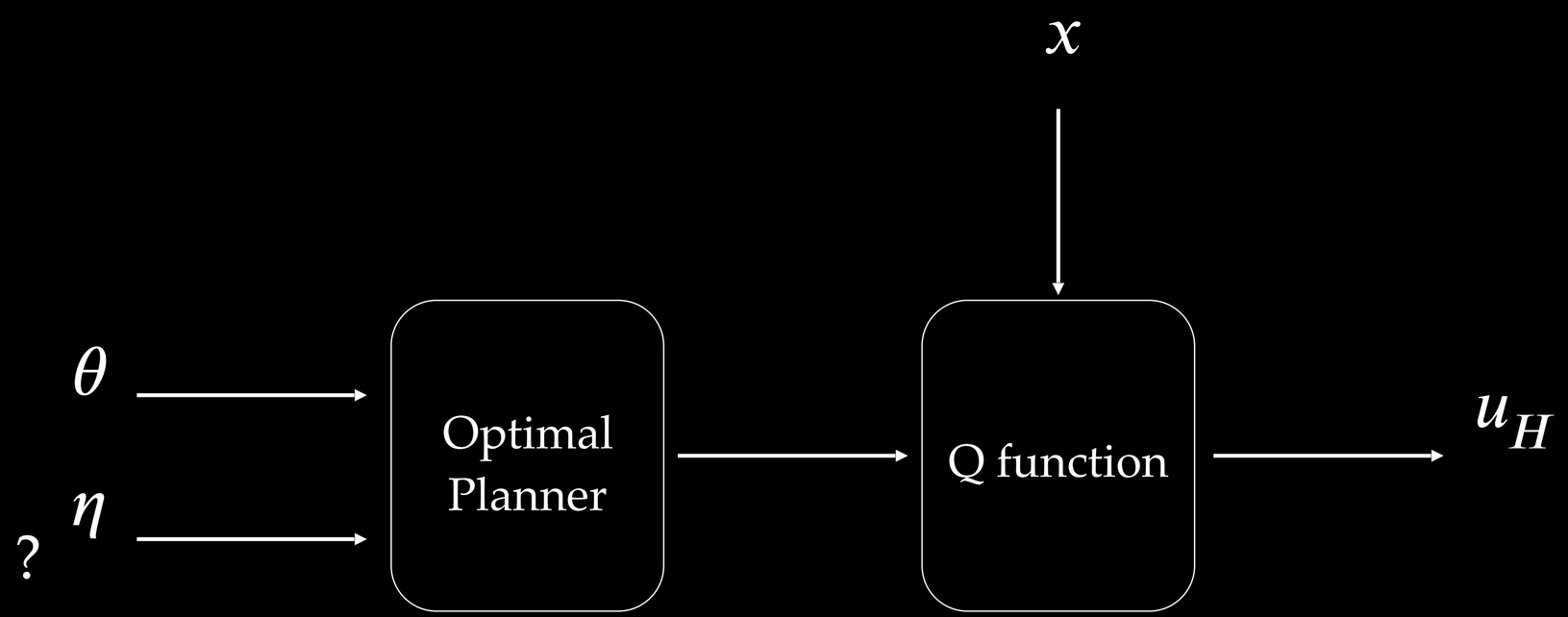


$$P(u_H \mid x; \theta, \eta) \propto e^{Q(x, u_H; \theta, \eta)}$$



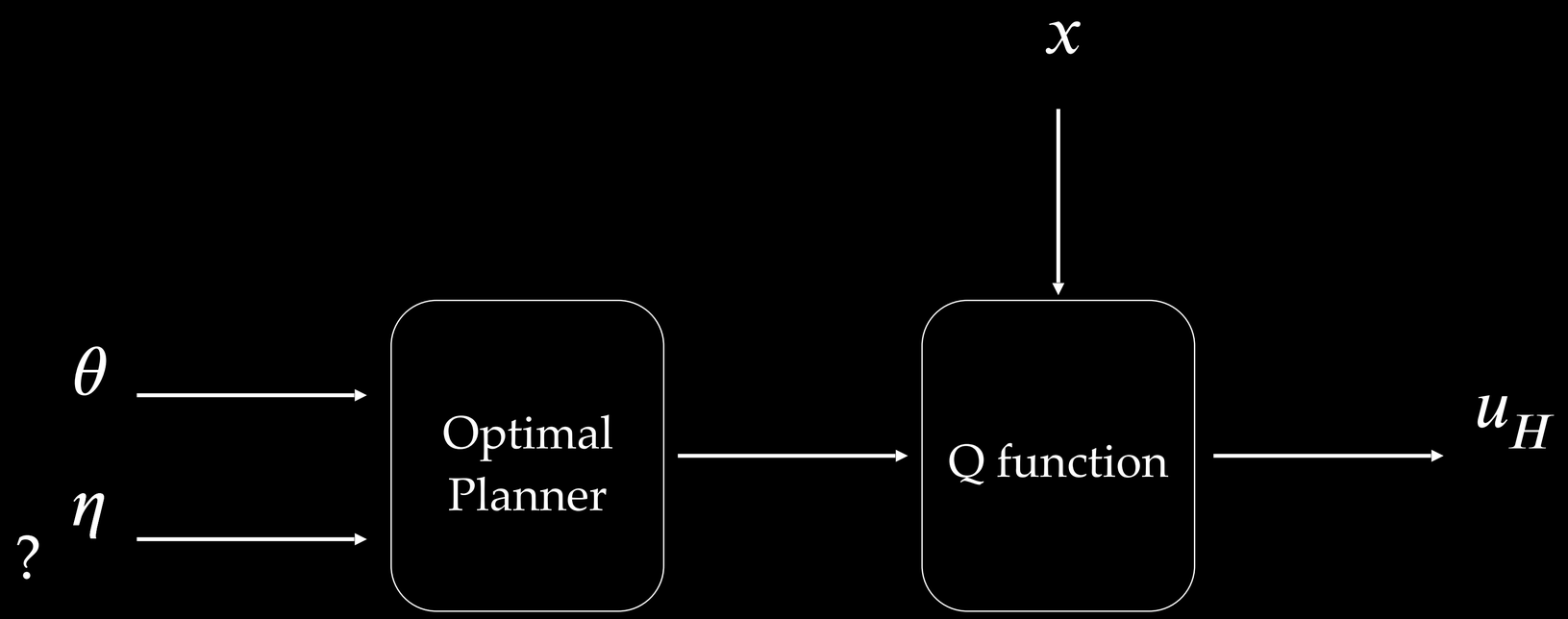
$$b'(\theta, \eta) \propto b(\theta, \eta) P(u_H \mid x; \theta, \eta)$$

$$P(u_H \mid x; \theta, \eta) \propto e^{Q(x, u_H; \theta, \eta)}$$



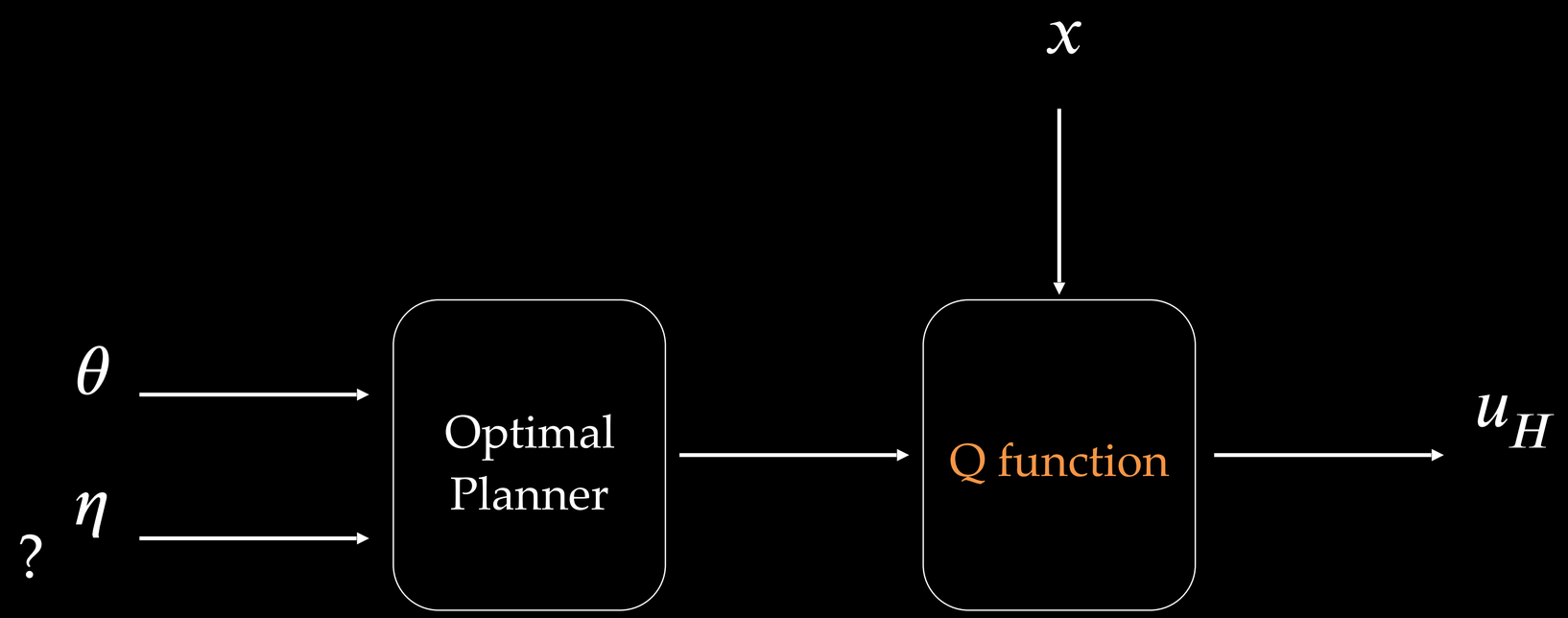
$$b'(\eta) \propto b(\eta) P(u_H \mid x; \theta, \eta)$$

$$P(u_H \mid x; \theta, \eta) \propto e^{Q(x, u_H; \theta, \eta)}$$



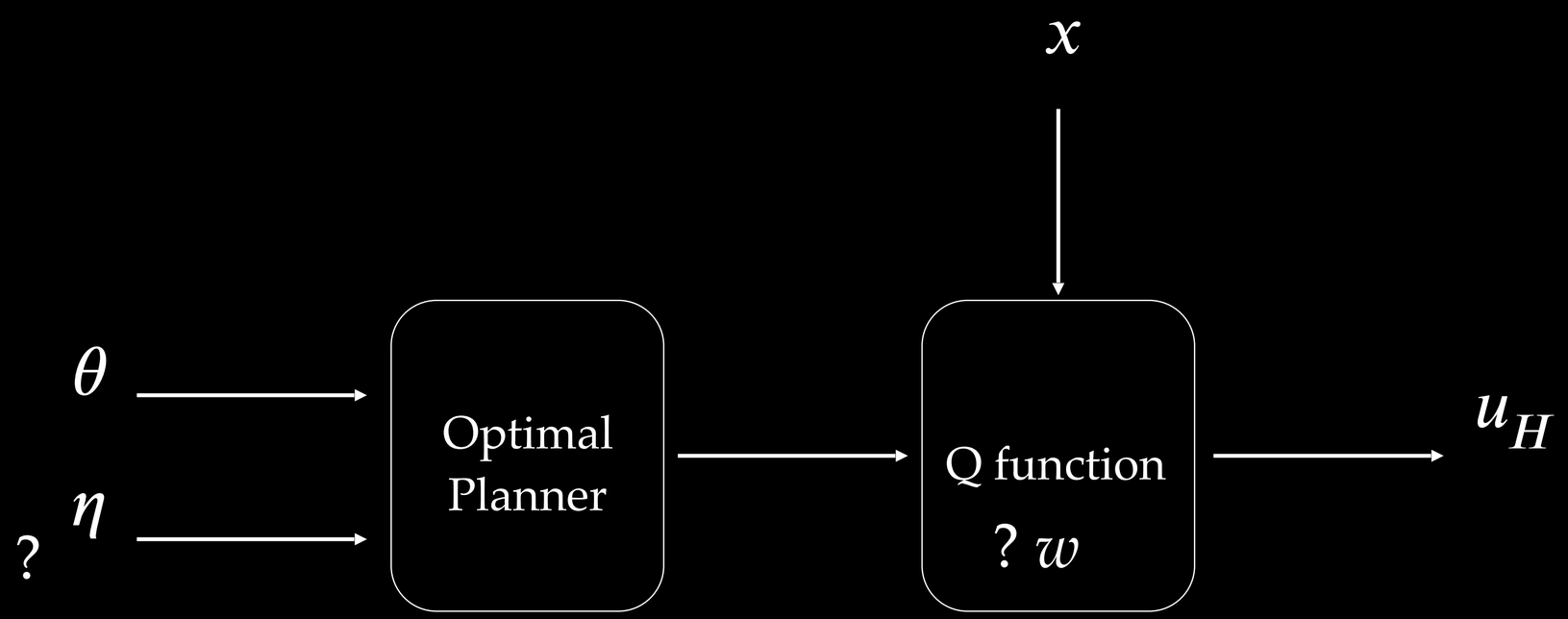
$$\max_{\eta} P(u_H \mid x; \theta, \eta)$$

$$P(u_H \mid x; \theta, \eta) \propto e^{Q(x, u_H; \theta, \eta)}$$



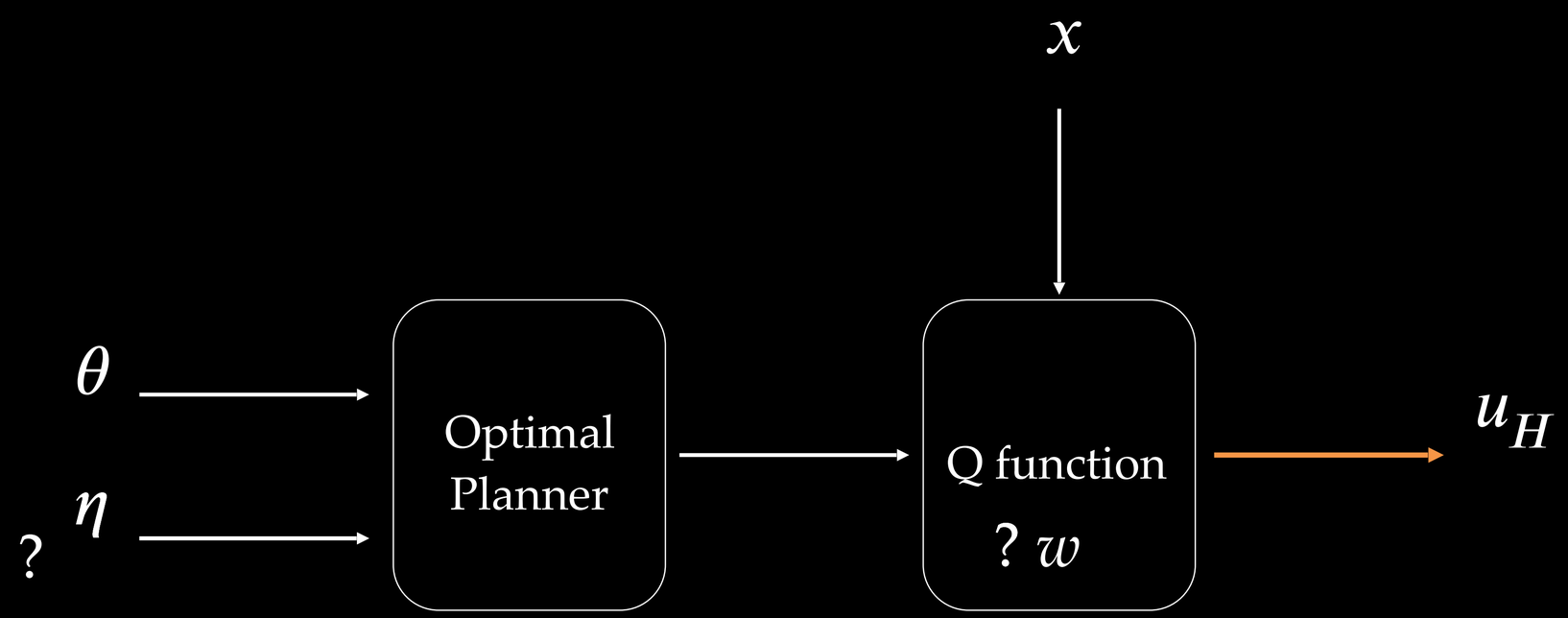
$$\max_{\eta} P(u_H \mid x; \theta, \eta)$$

$$P(u_H \mid x; \theta, \eta) \propto e^{Q(x, u_H; \theta, \eta)}$$



$$\max_{\eta, w} P\left(u_H \middle| x; w\right) - \lambda \delta\left(w, \eta; \theta\right)^2$$

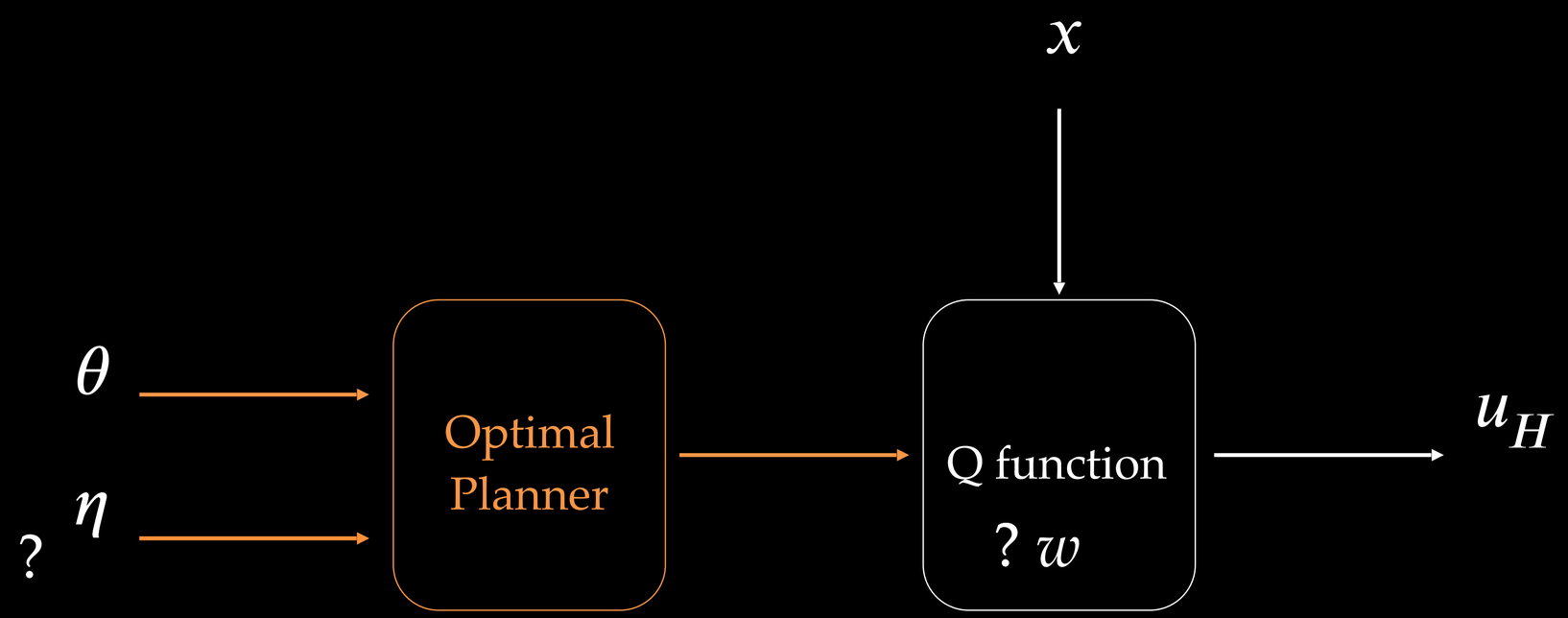
$$P(u_H \mid x; \theta, \eta) \propto e^{Q(x, u_H; \theta, \eta)}$$



$$\max_{\eta, w} P(u_H \mid x; w) - \lambda \delta(w, \eta; \theta)^2$$

give u high Q-value

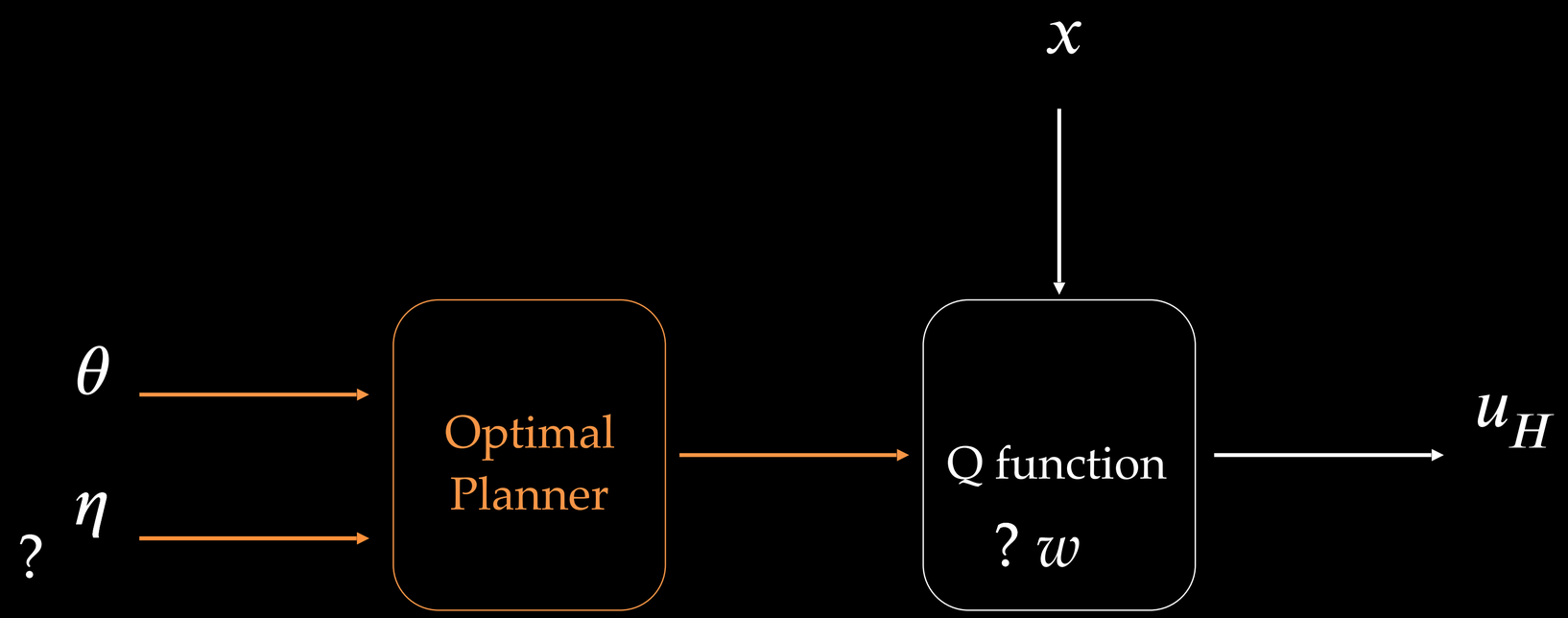
$$P(u_H \mid x; \theta, \eta) \propto e^{Q(x, u_H; \theta, \eta)}$$



$$\max_{\eta, w} P(u_H \mid x; w) - \lambda \delta(w, \eta; \theta)^2$$

have low Bellman residual on sample states

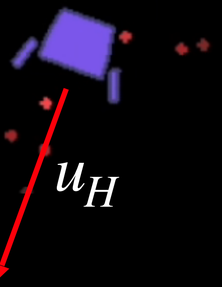
$$P(u_H \mid x; \theta, \eta) \propto e^{Q(x, u_H; \theta, \eta)}$$

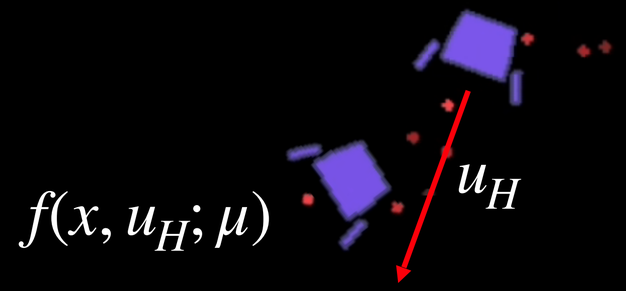


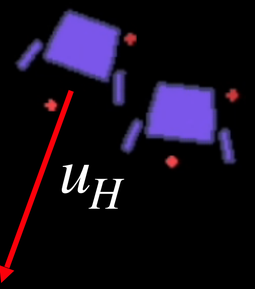
$$\max_{\eta, w} P(u_H \mid x; w) - \lambda \delta(w, \eta; \theta)^2$$

have low Bellman residual on sample states

$$\sum_x \sum_u Q(x_i, u; w) - (R(x_i, u; \theta) + \gamma \max_{u'} Q(f(x_i, u; \eta), u'; w))$$

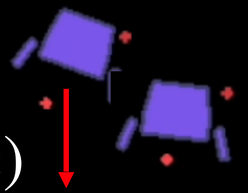

u_H

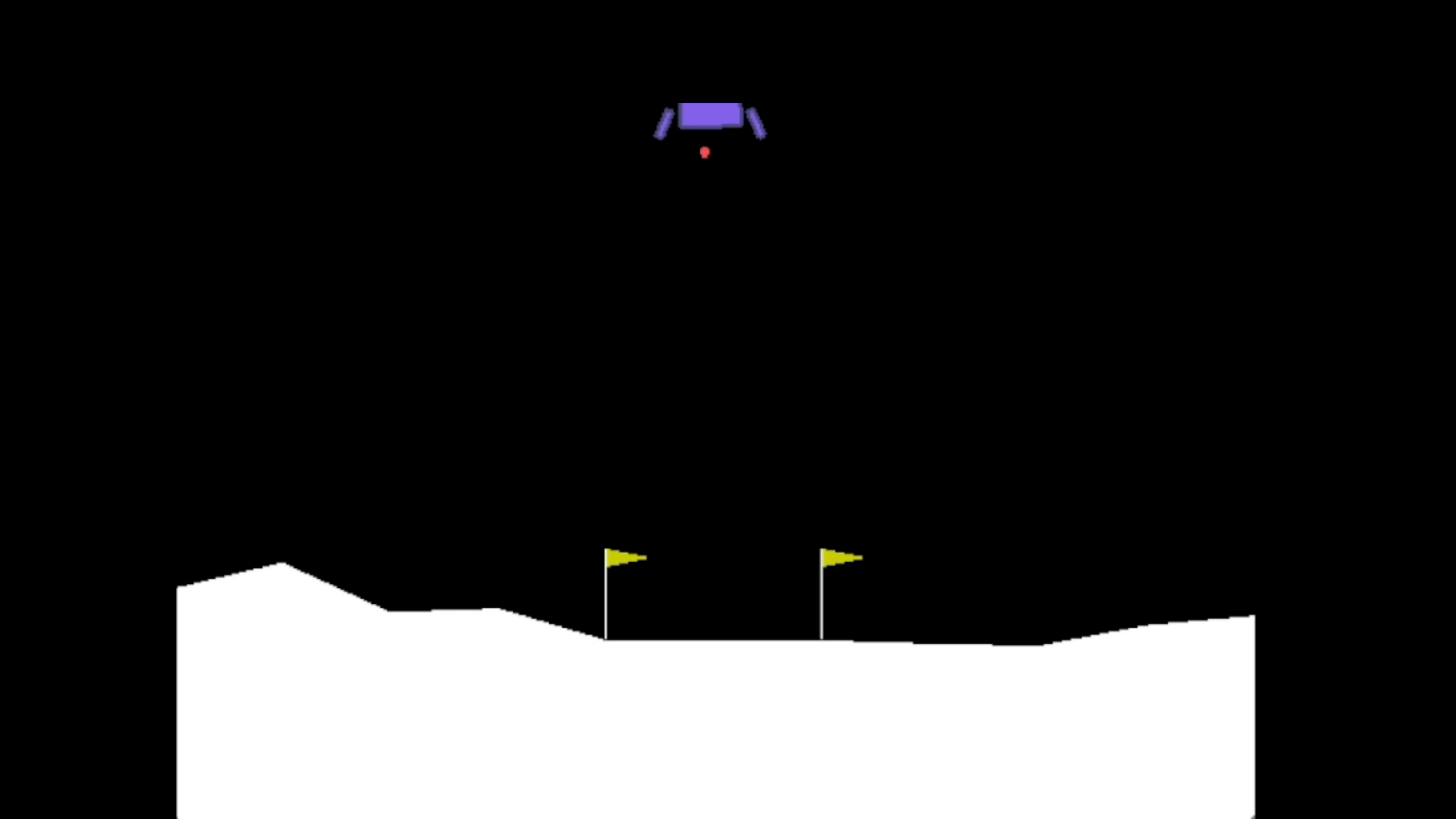

$f(x, u_H; \mu)$
u_H

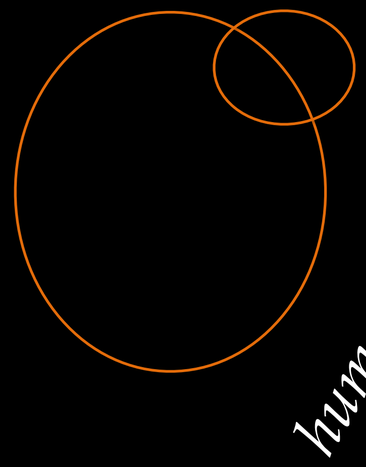


$f(x, u_H; \hat{\eta})$

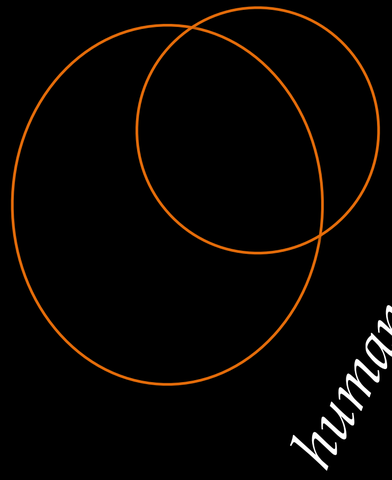
$f^{-1}(x, f(x, u_H; \hat{\eta}; \mu)$







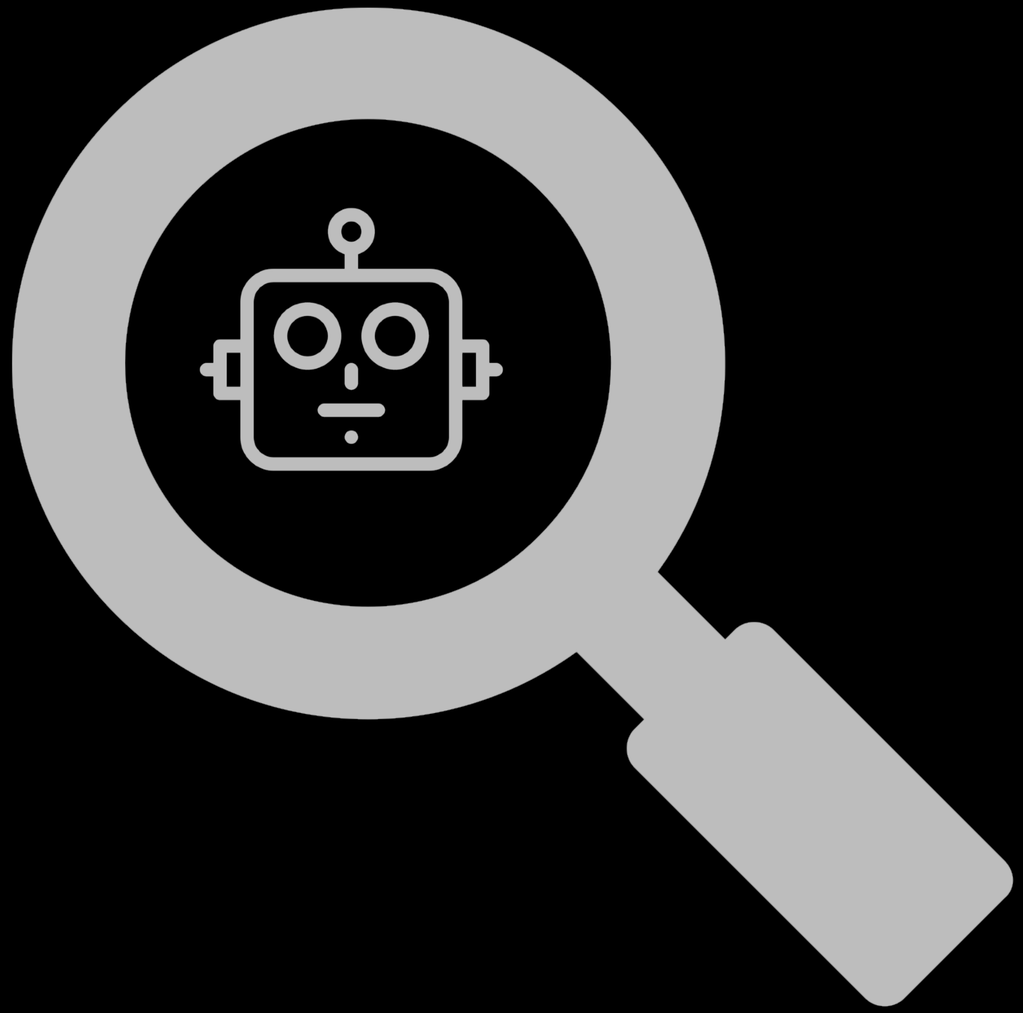
all policies



noisy rationality under internal dynamics

all policies

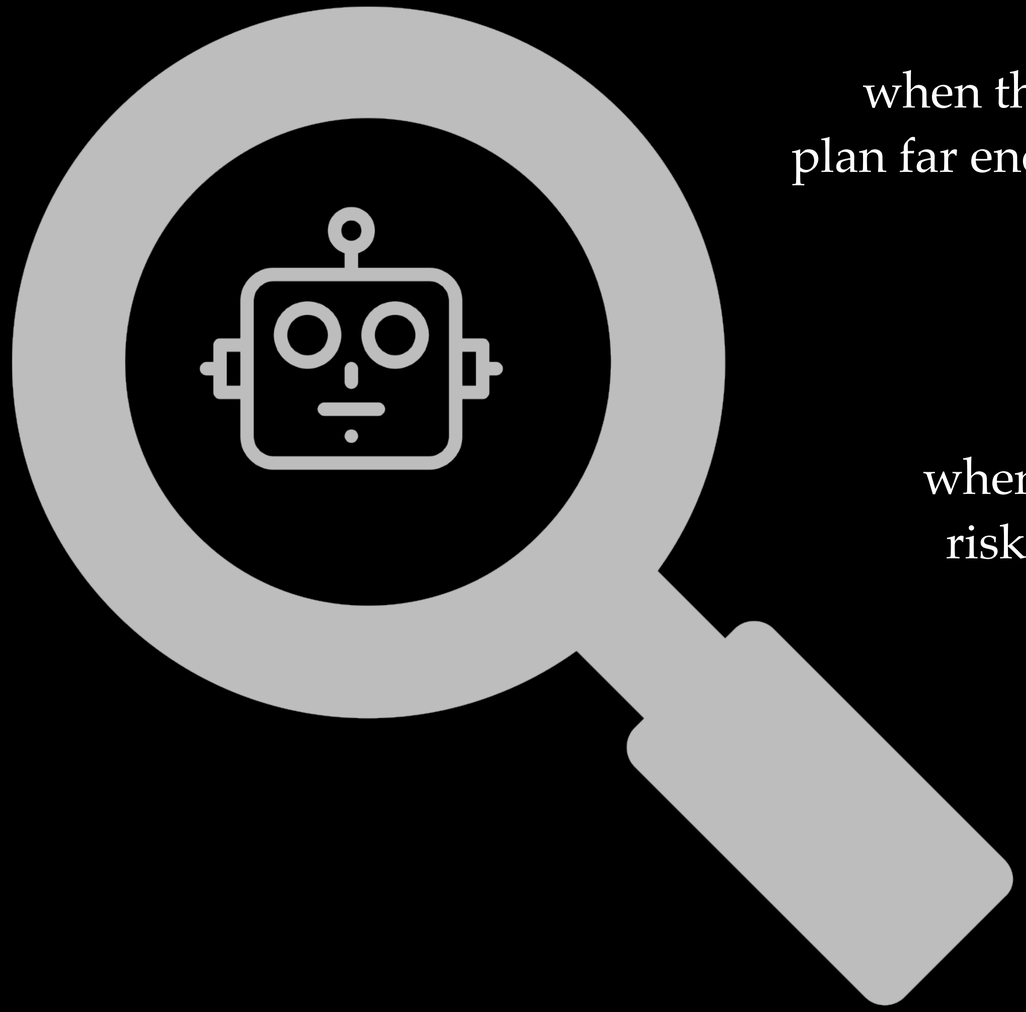
When are robots not rational?



When are robots not rational?

when they don’t
know the dynamics

when they can’t
plan far enough ahead



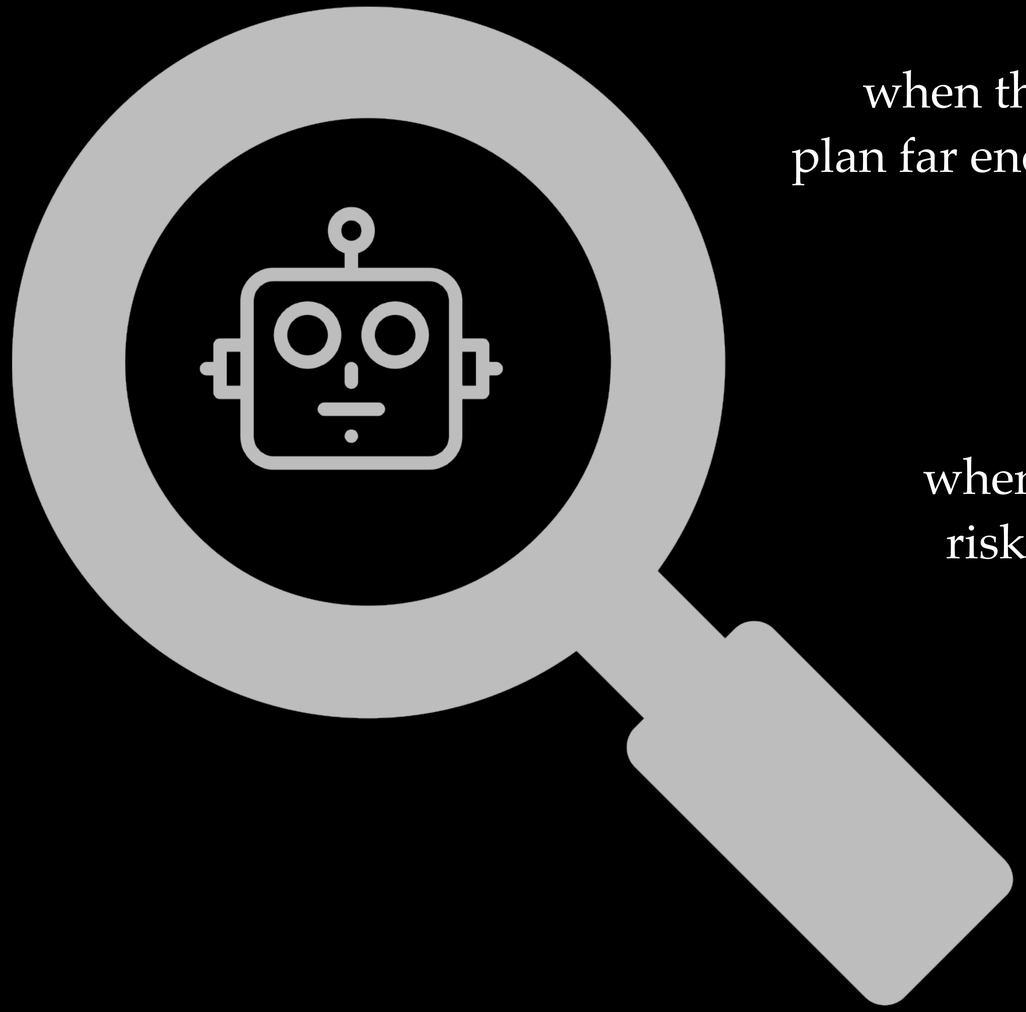
when they’re
risk-averse

when they’re
still learning

When are humans not rational?

when they don’t
know the dynamics

when they can’t
plan far enough ahead

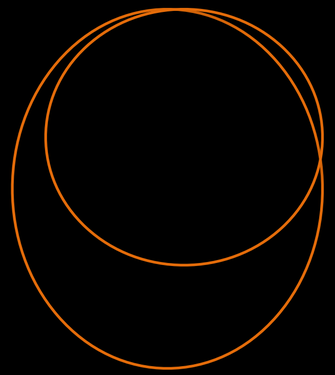


when they
are optimistic

when they’re
risk-averse

when they’re
still learning

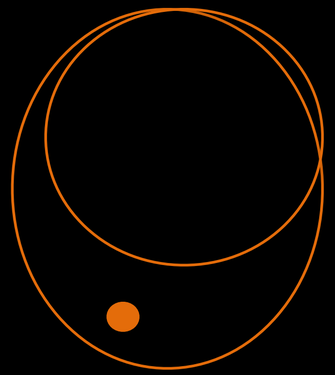
…



broader rationality

humans

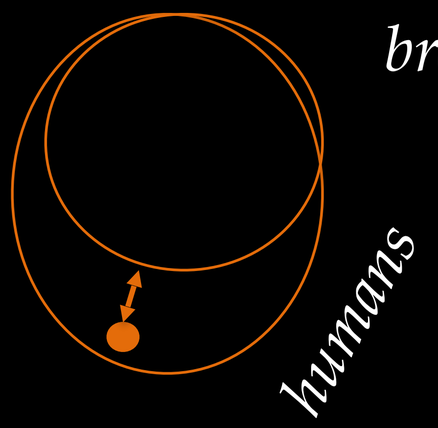
all policies



broader rationality

humans

all policies

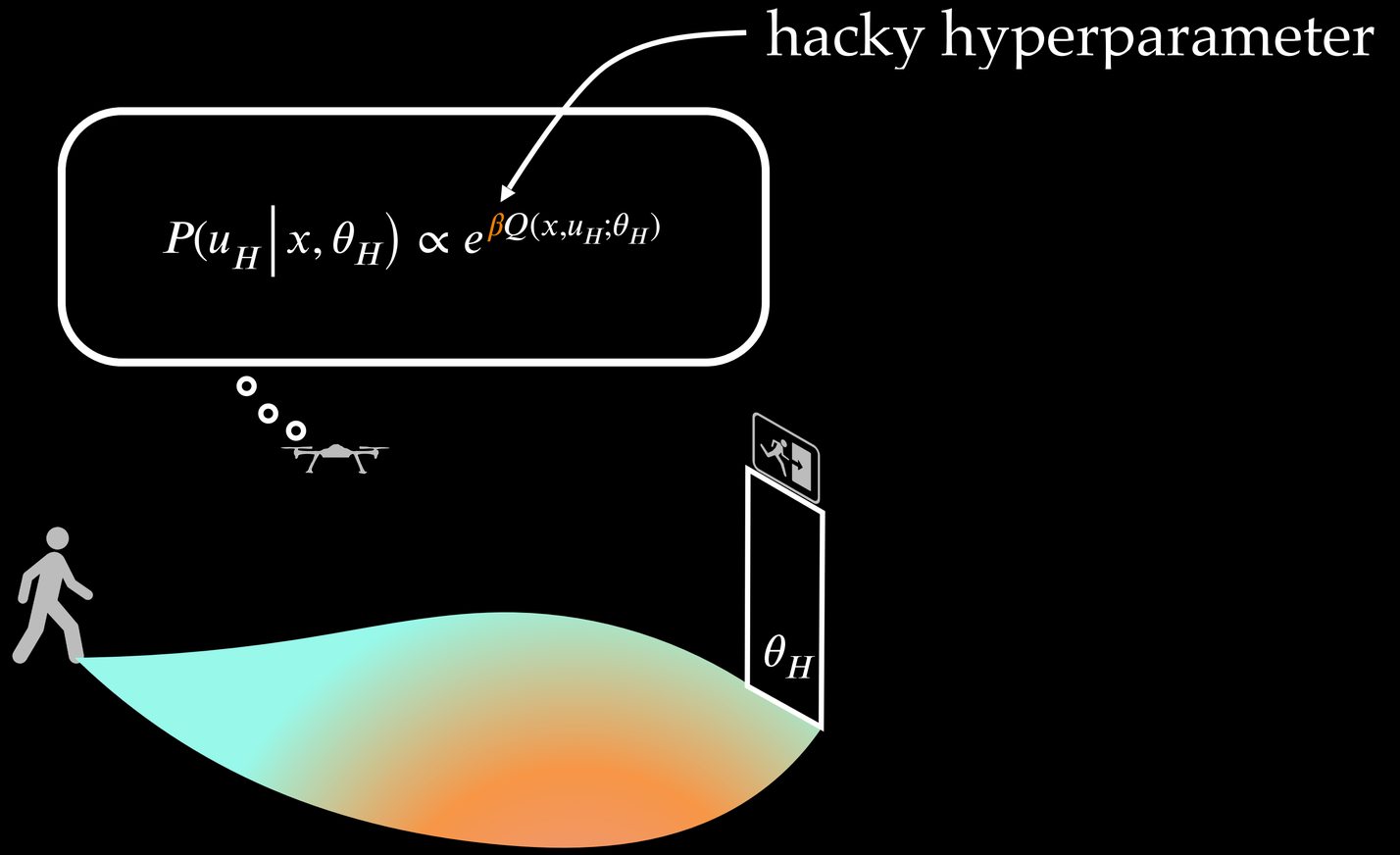


Can we detect that we have the wrong hypothesis space?

all policies


MIND THE GAP

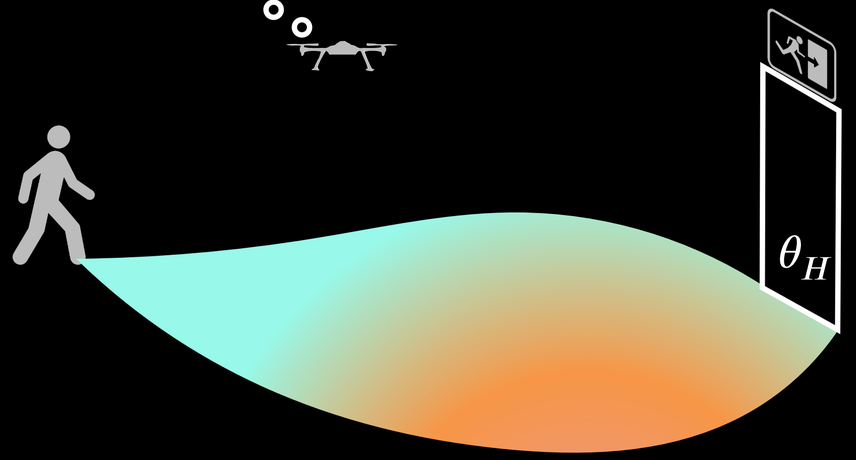
The rationality coefficient



The rationality coefficient

make it part of the inference!

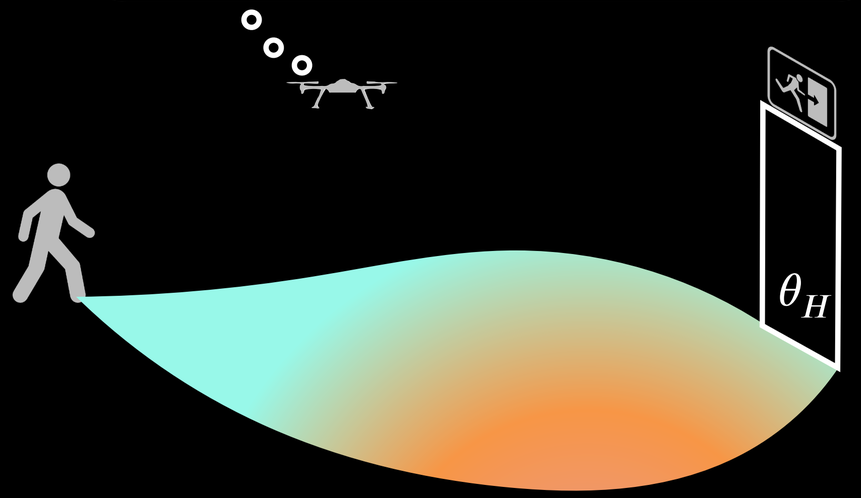
$$P(u_H \mid x, \theta_H) \propto e^{\beta Q(x, u_H; \theta_H)}$$



The rationality coefficient

$$P(u_H \mid x, \theta_H, \beta) \propto e^{\beta Q(x, u_H; \theta_H)}$$

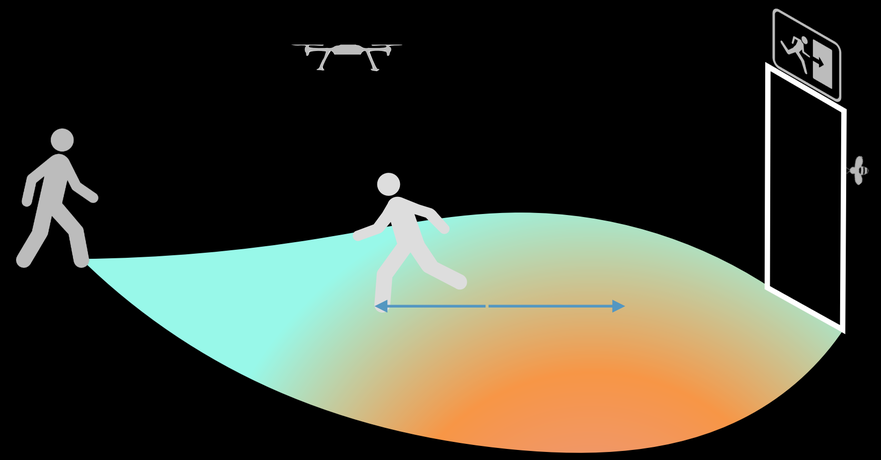
$$b'(\theta_H, \beta) \propto b(\theta_H, \beta)\, P(u_H \mid x, \theta_H, \beta)$$

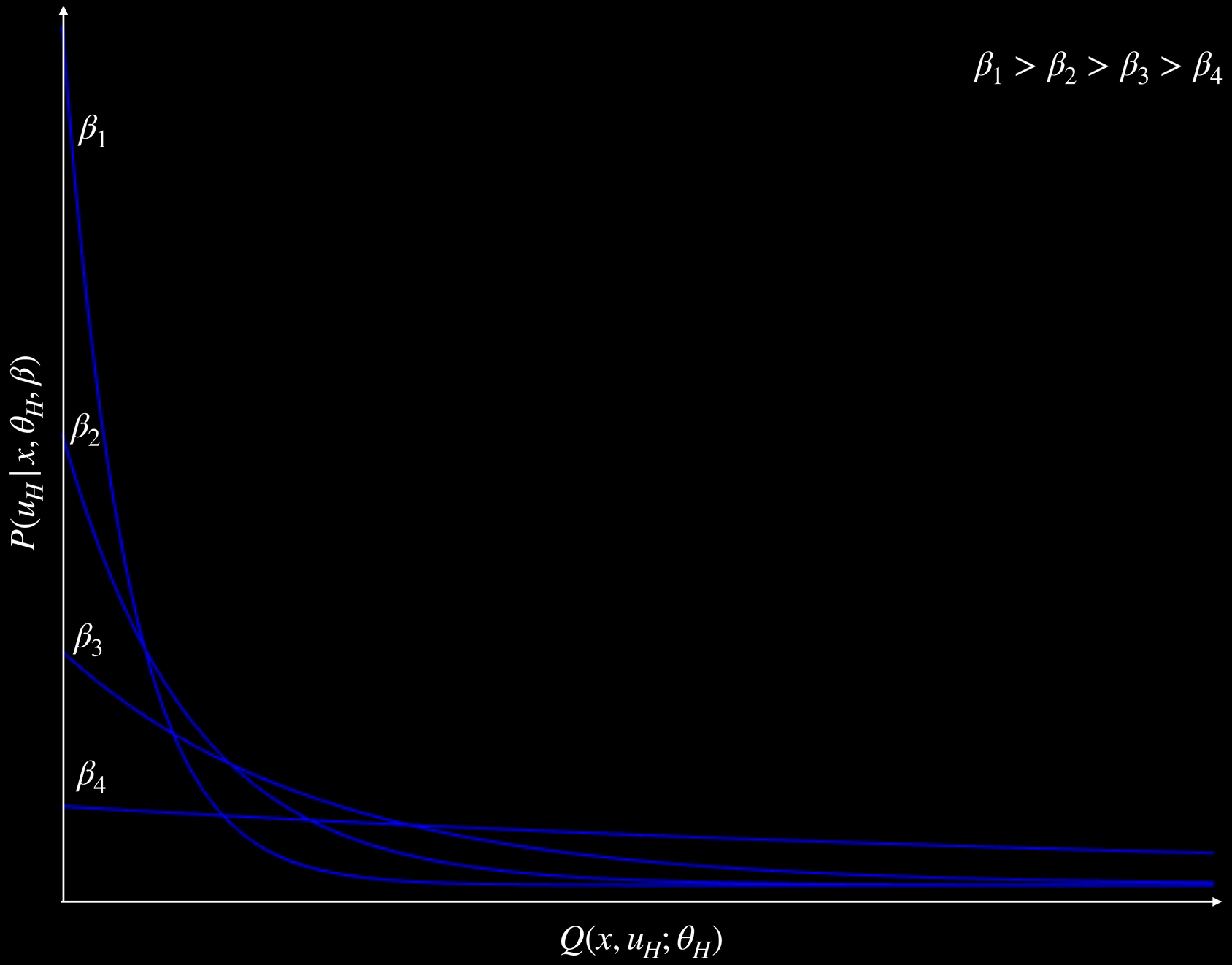


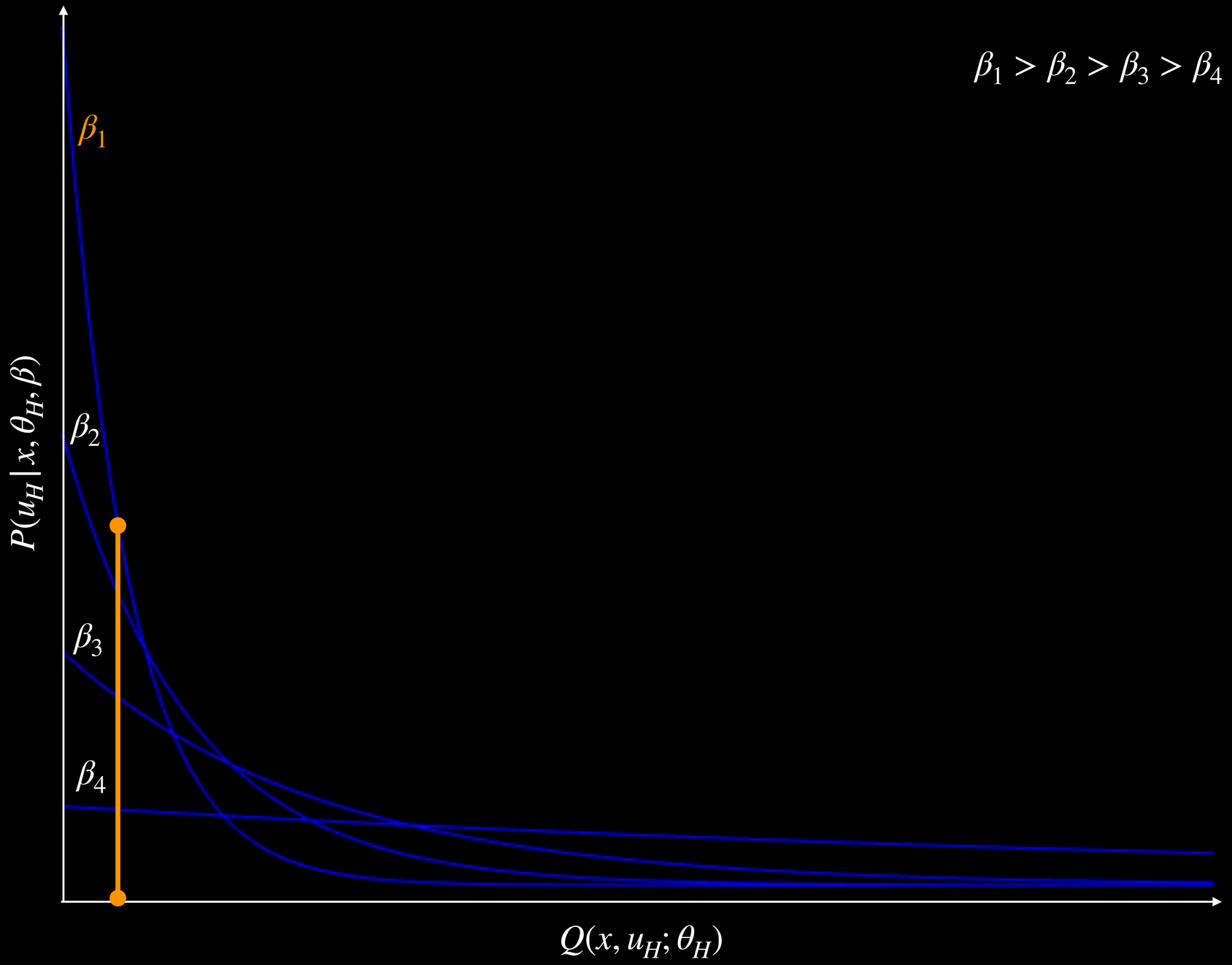
If the human appears
too suboptimal to the model,
be skeptical of the model.

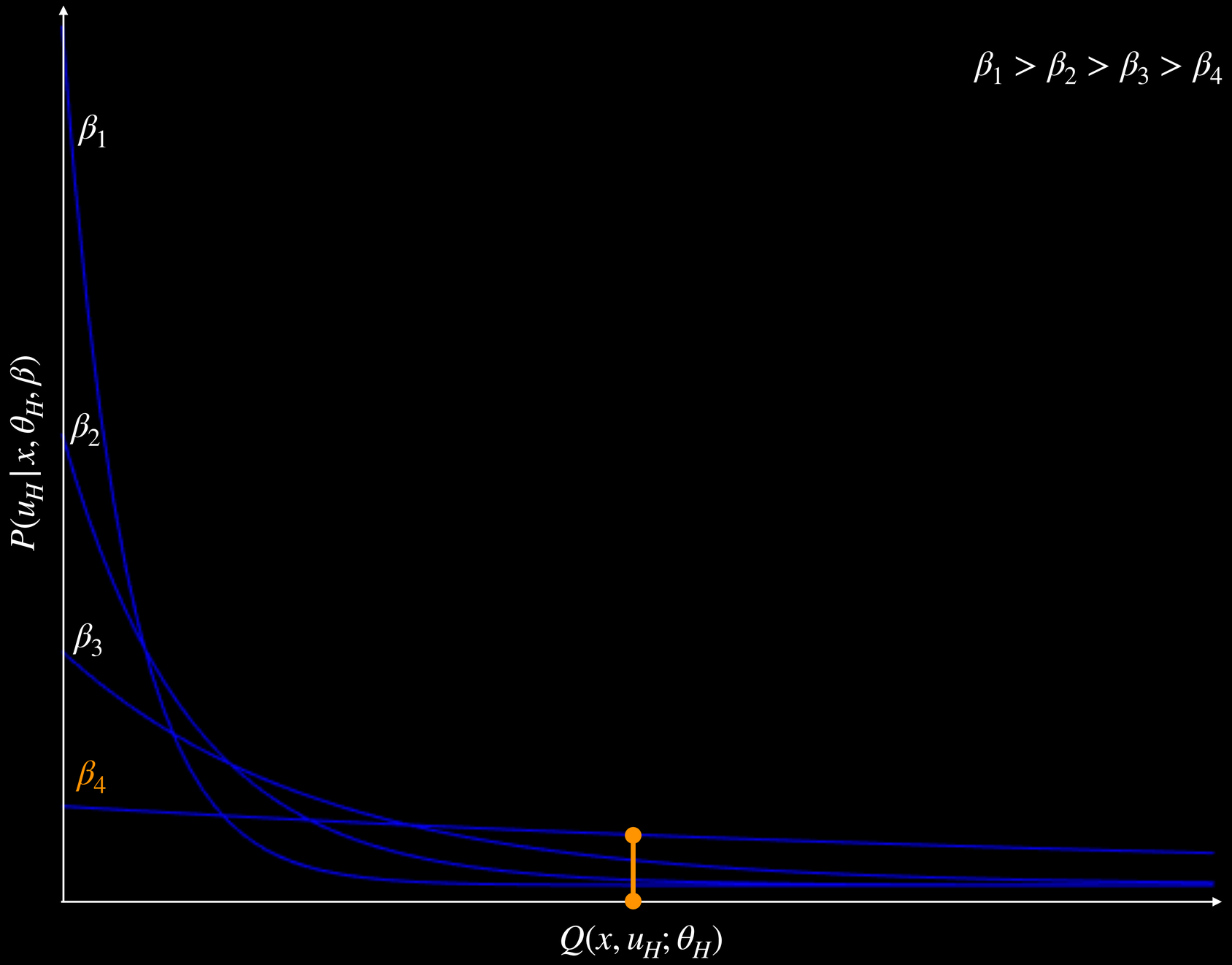
Human rationality = model confidence

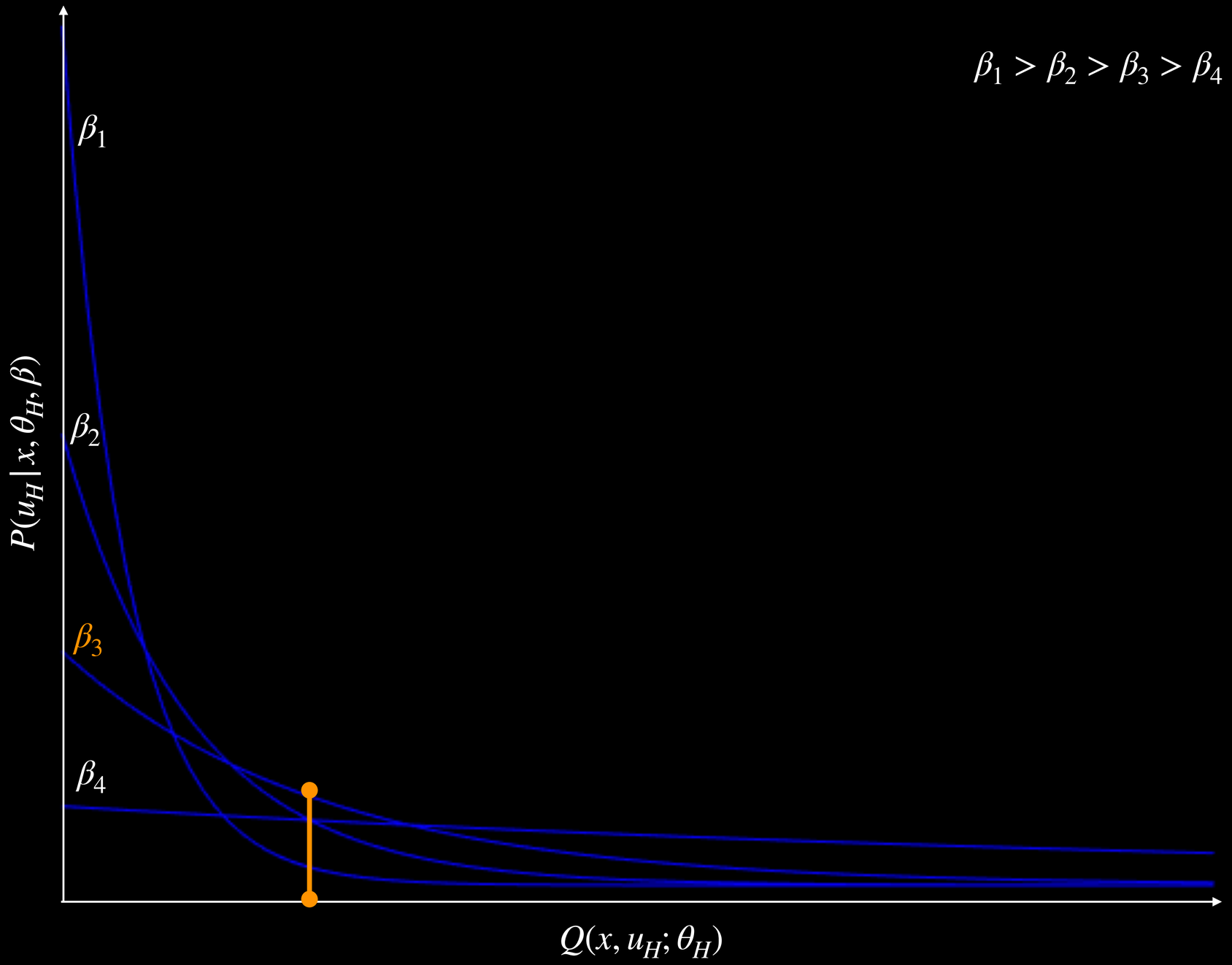
$$b'(\theta_H, \beta) \propto b(\theta_H, \beta) P(u_H \mid x, \theta_H, \beta)$$
$$\propto e^{\beta Q(x, u_H; \theta_H)}$$

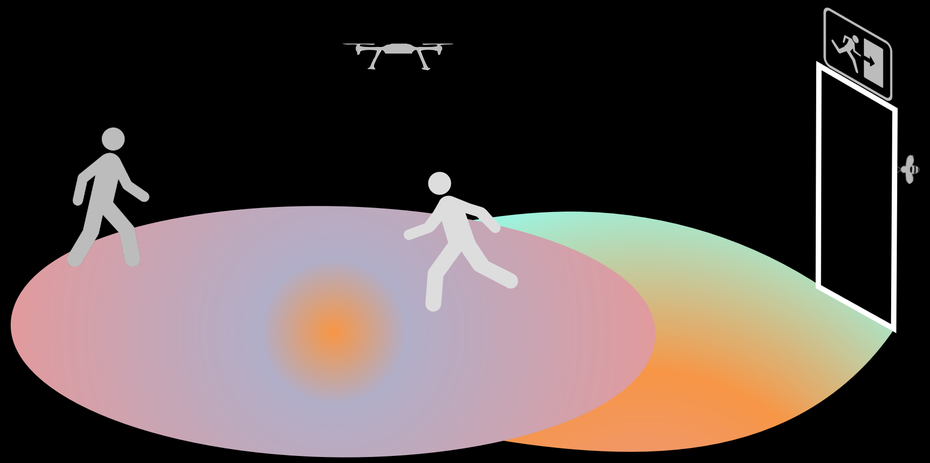


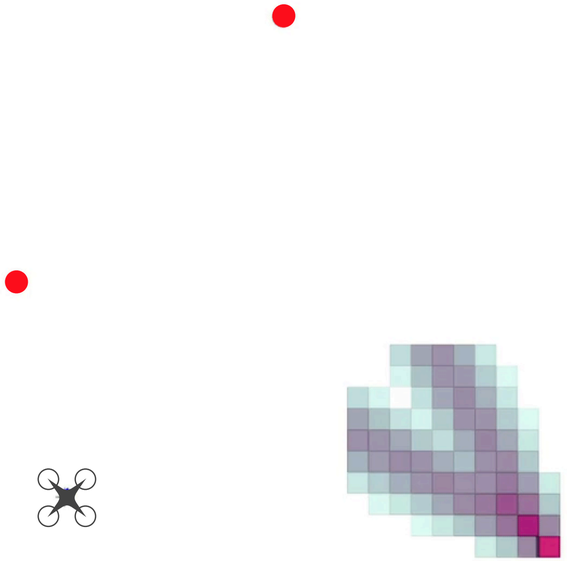

$\beta_1 > \beta_2 > \beta_3 > \beta_4$
β_1
β_2
β_3
β_4
$P(u_H \mid x, \theta_H, \beta)$
$Q(x, u_H; \theta_H)$


$\beta_1 > \beta_2 > \beta_3 > \beta_4$
β_1
β_2
β_3
β_4
$P(u_H \mid x, \theta_H, \beta)$
$Q(x, u_H; \theta_H)$

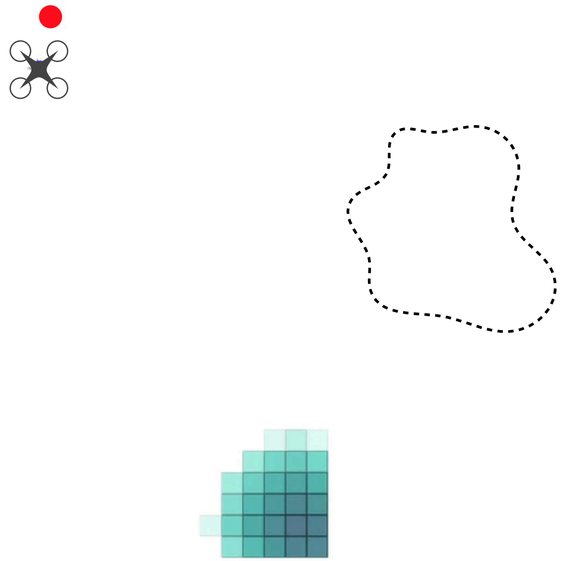

$\beta_1 > \beta_2 > \beta_3 > \beta_4$
β_1
β_2
β_3
β_4
$P(u_H \mid x, \theta_H, \beta)$
$Q(x, u_H; \theta_H)$


$\beta_1 > \beta_2 > \beta_3 > \beta_4$
β_1
β_2
β_3
β_4
$P(u_H \mid x, \theta_H, \beta)$
$Q(x, u_H; \theta_H)$

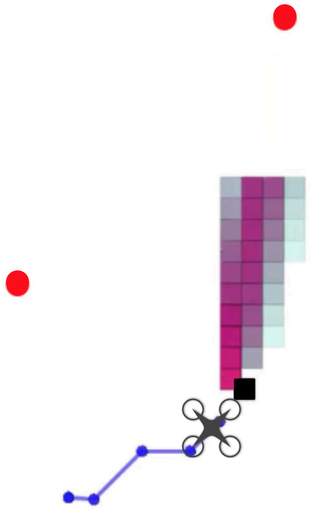


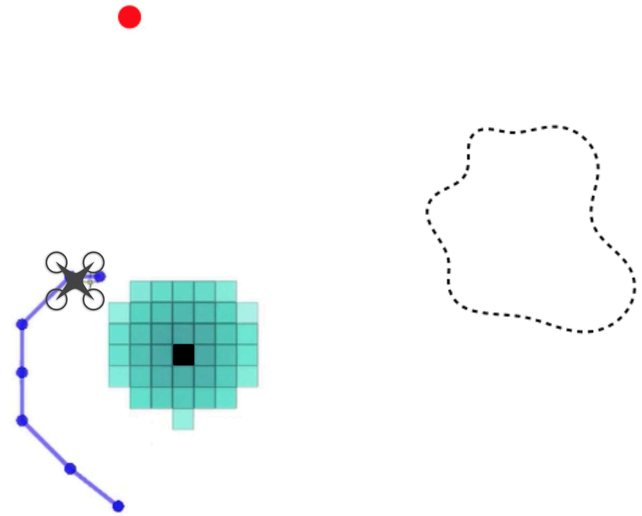


high confidence $\overline{\beta}$



low confidence $\underline{\beta}$



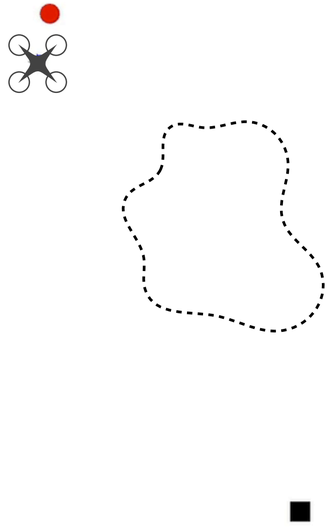


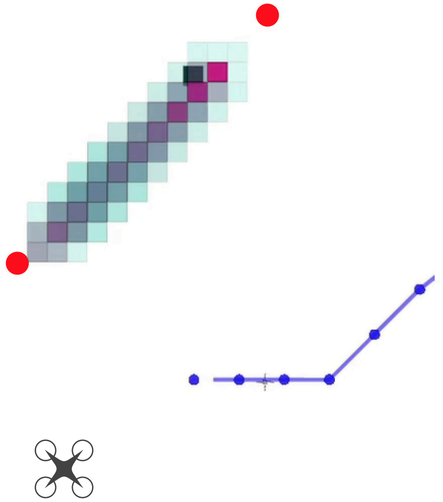
high confidence $\overline{\beta}$

low confidence $\underline{\beta}$

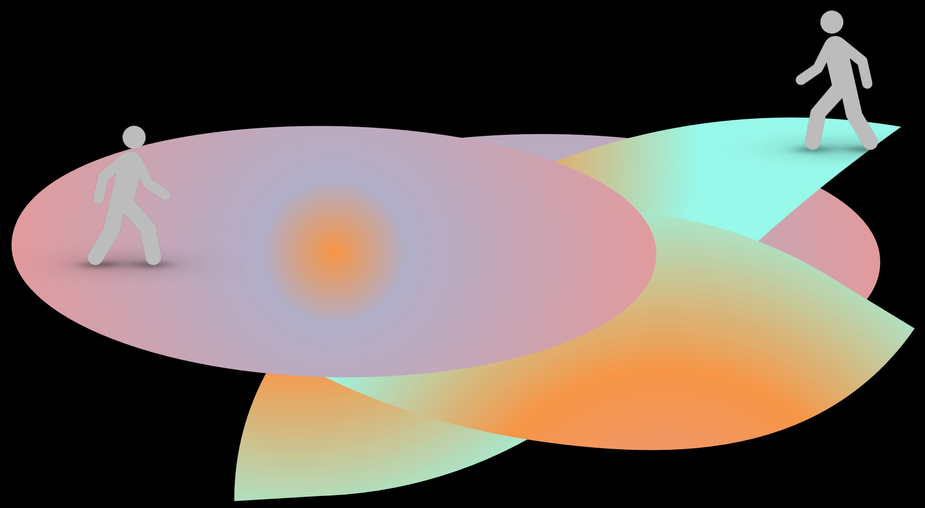
Bayesian confidence

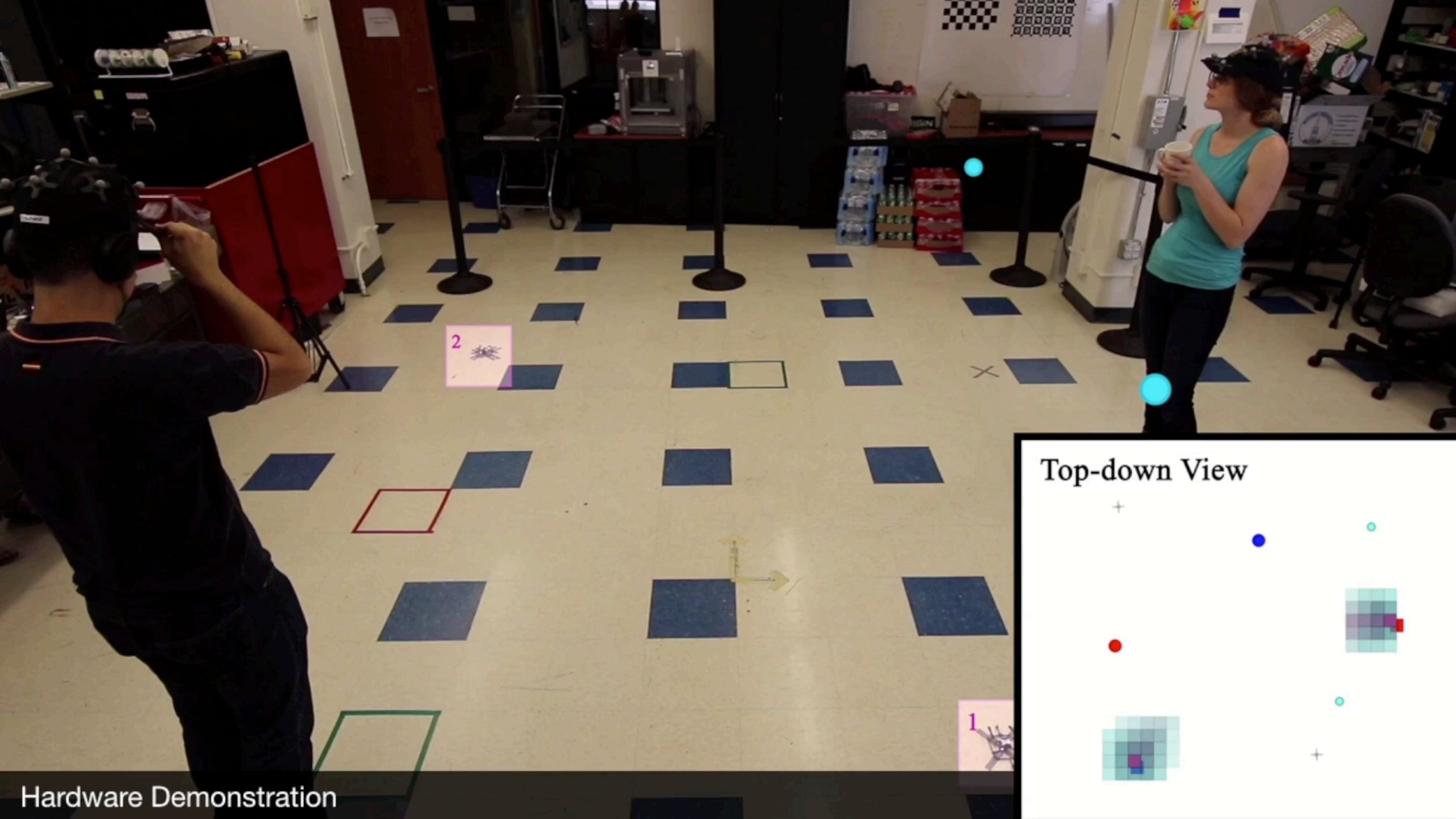
$b^t(\beta)$





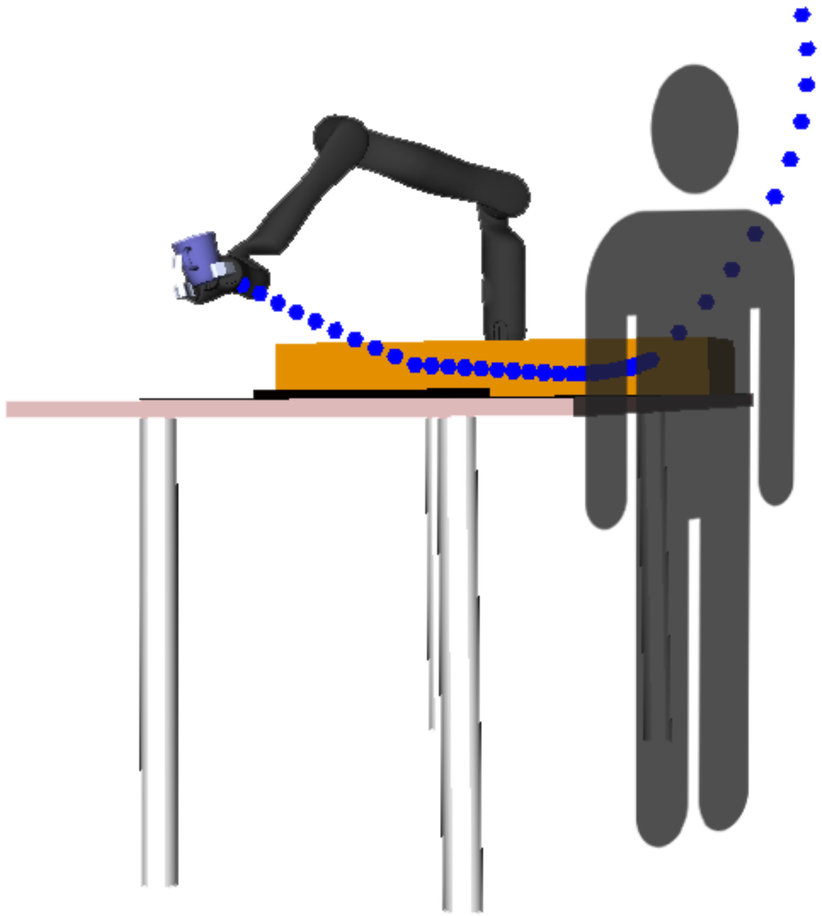
Multiple humans

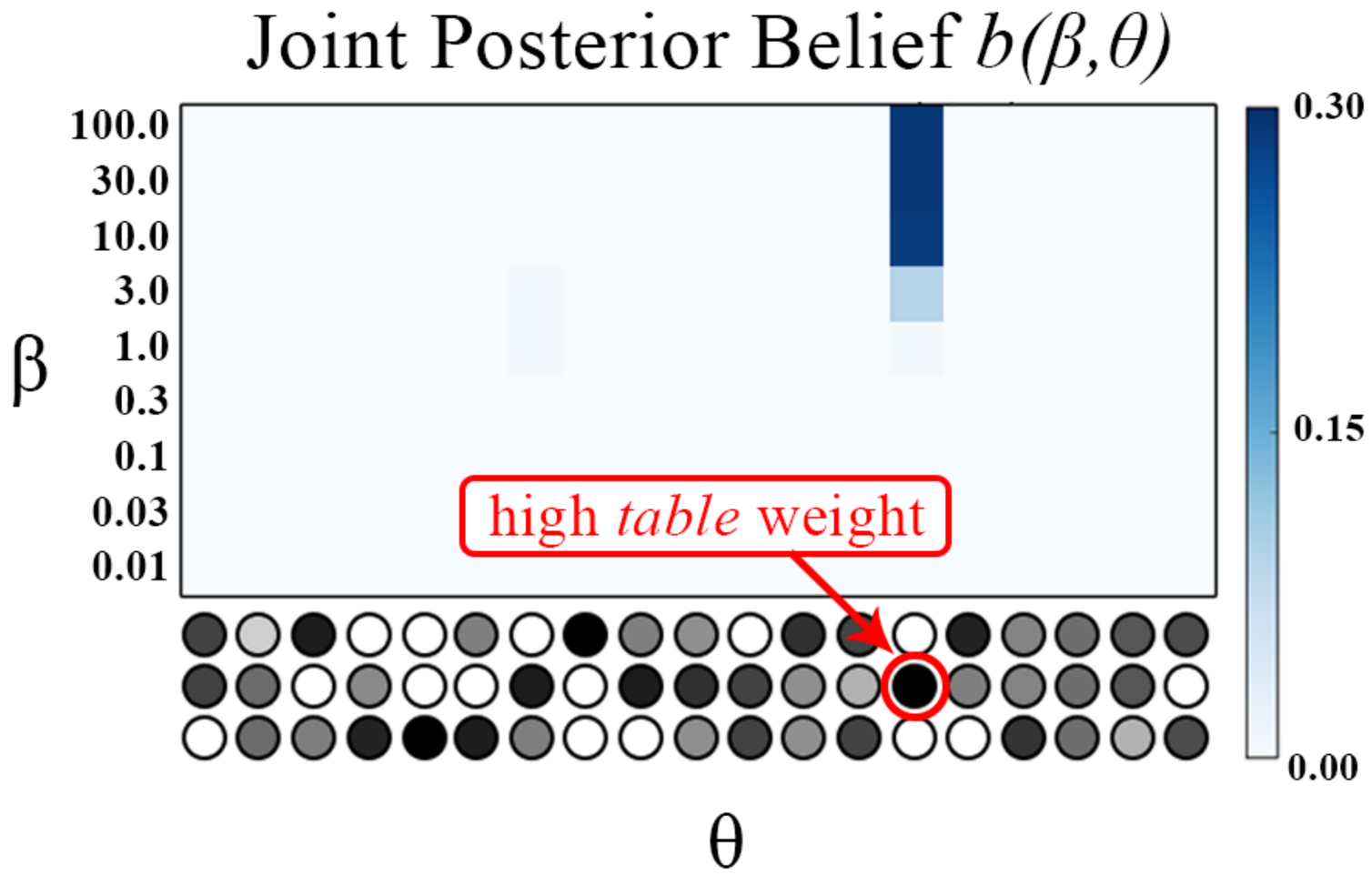




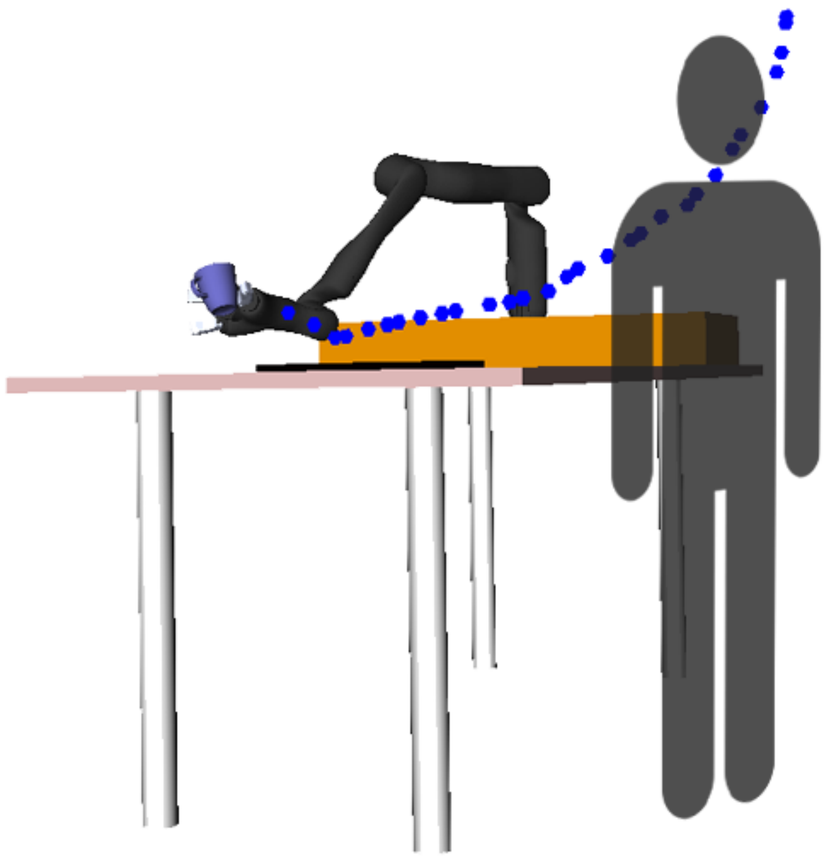
Hardware Demonstration

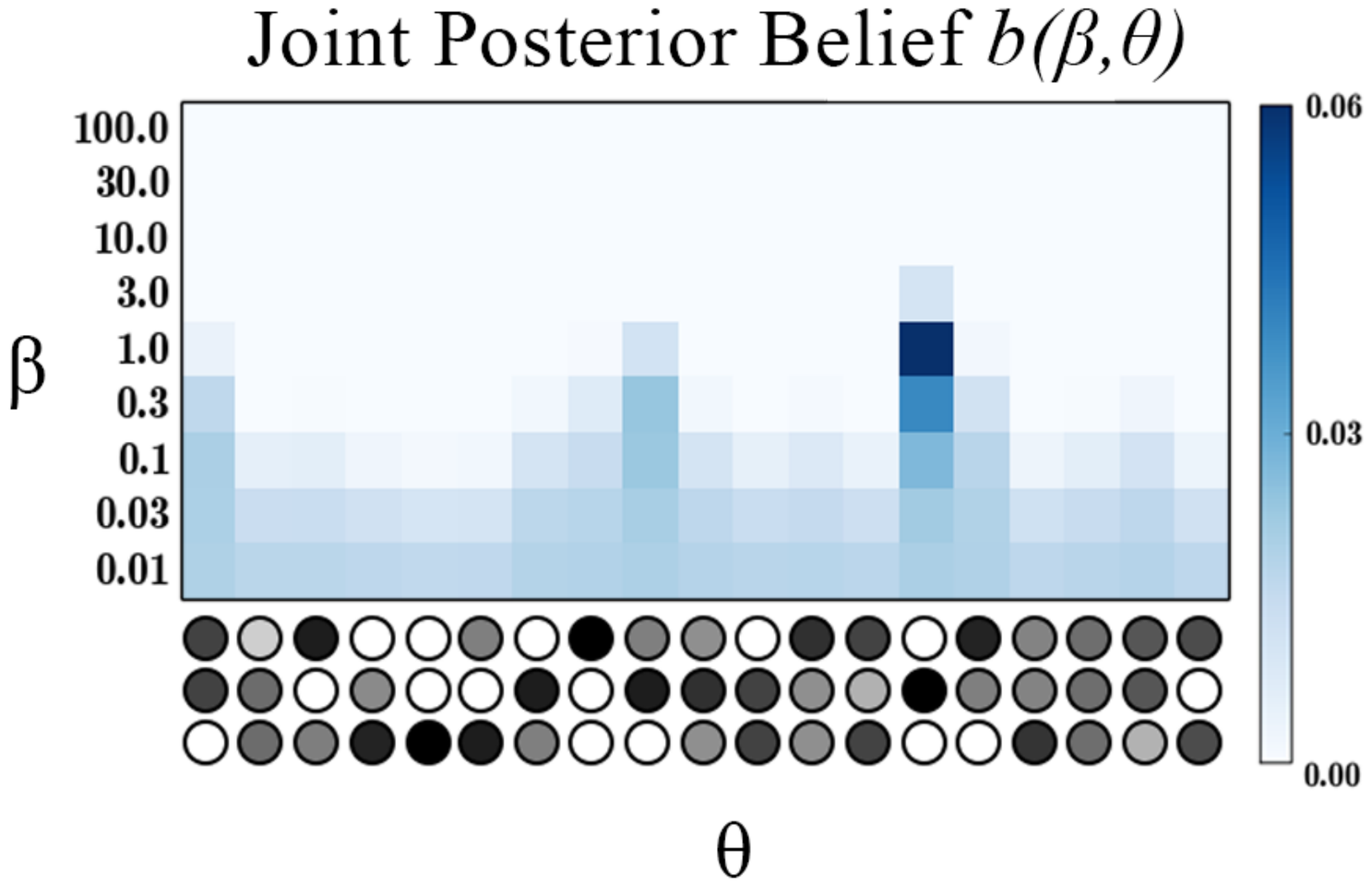
Detecting misspecification in demonstrations



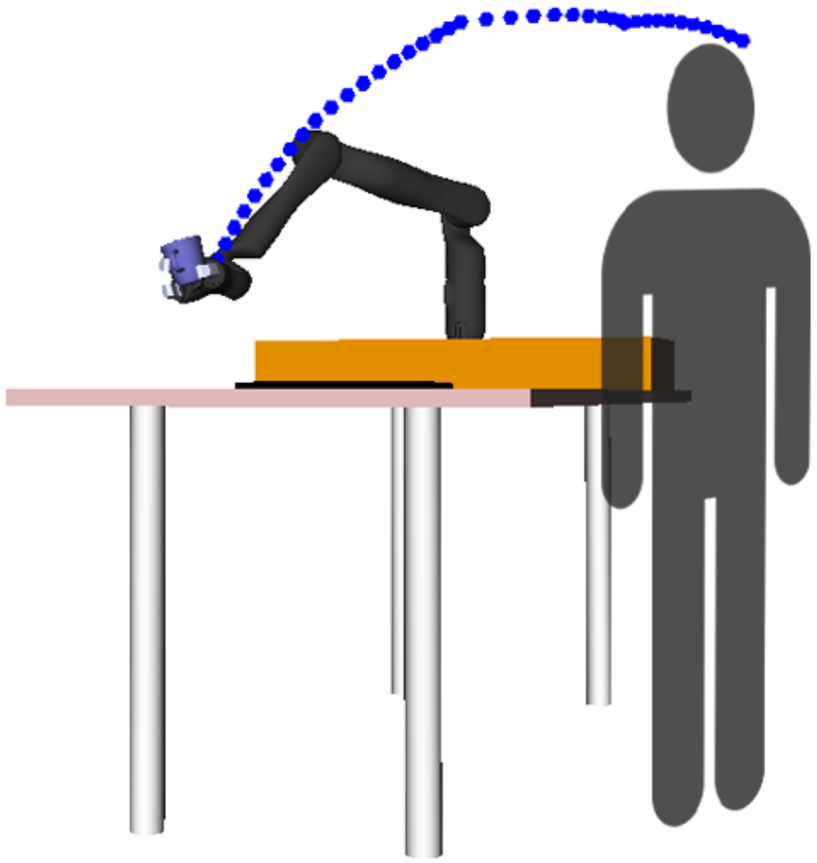


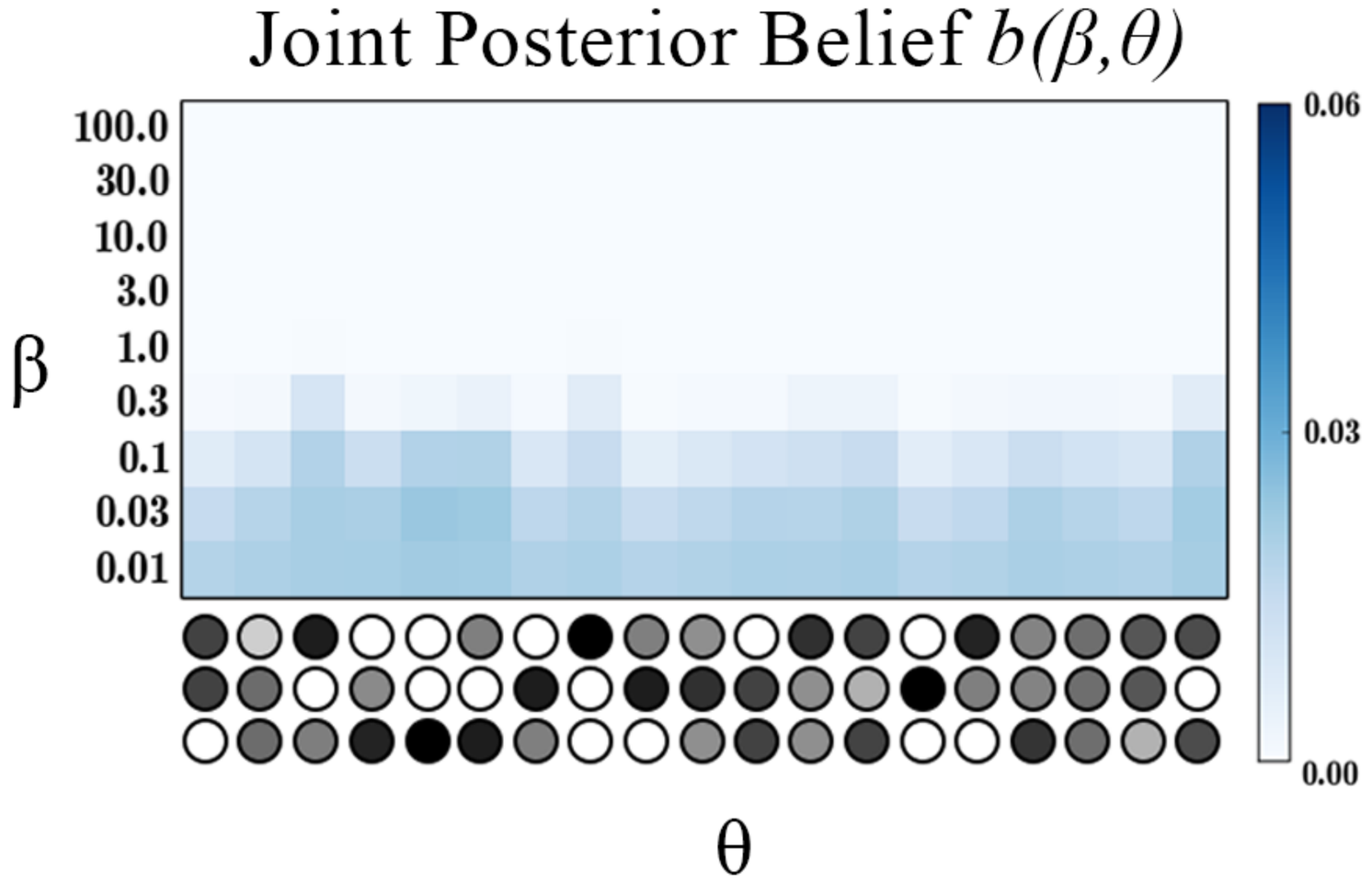
Detecting misspecification in demonstrations



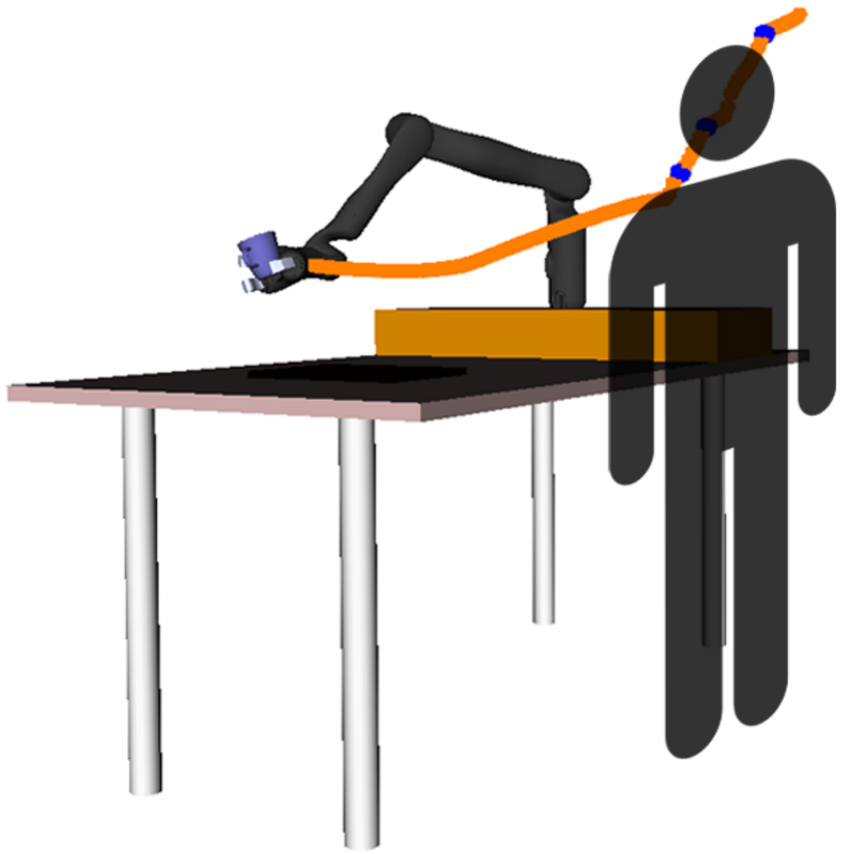


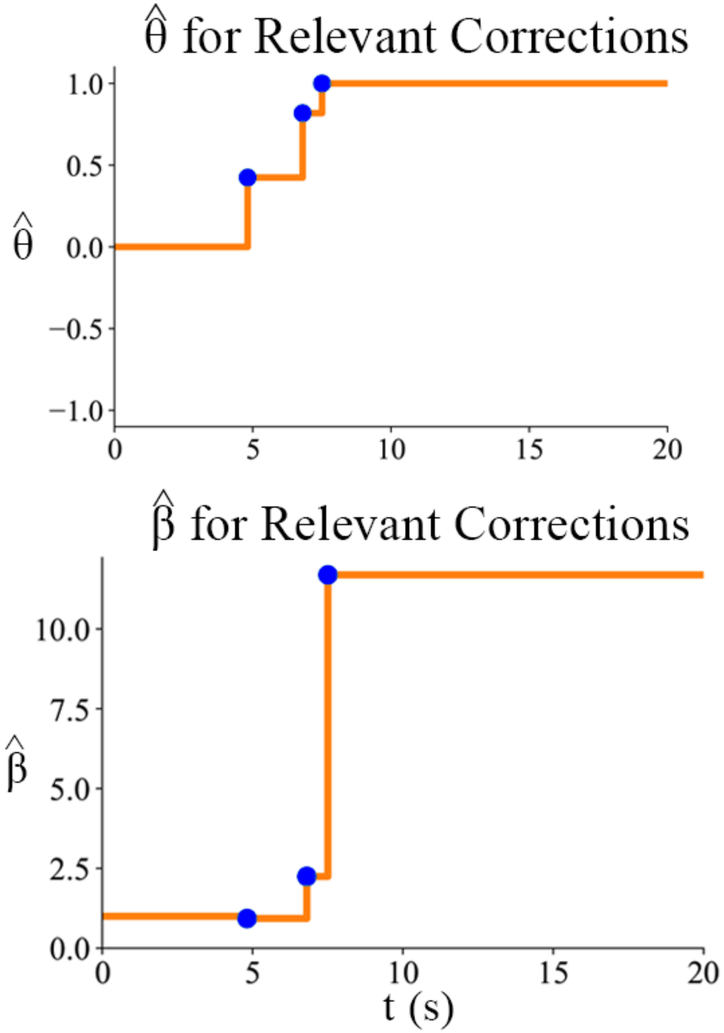
Detecting misspecification in demonstrations



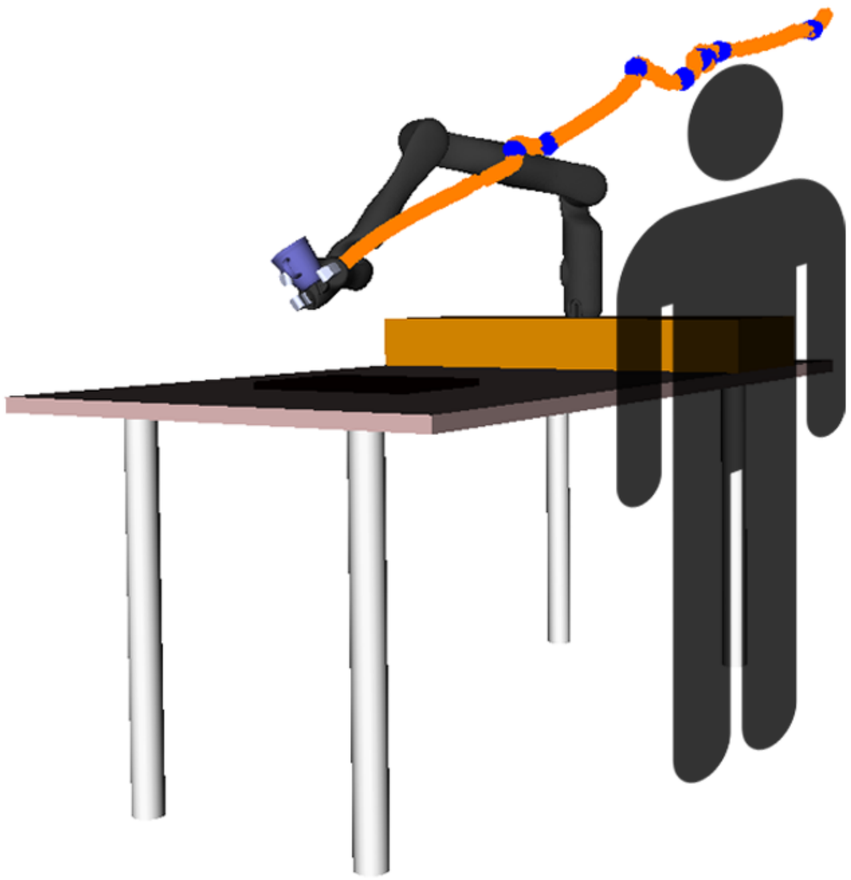


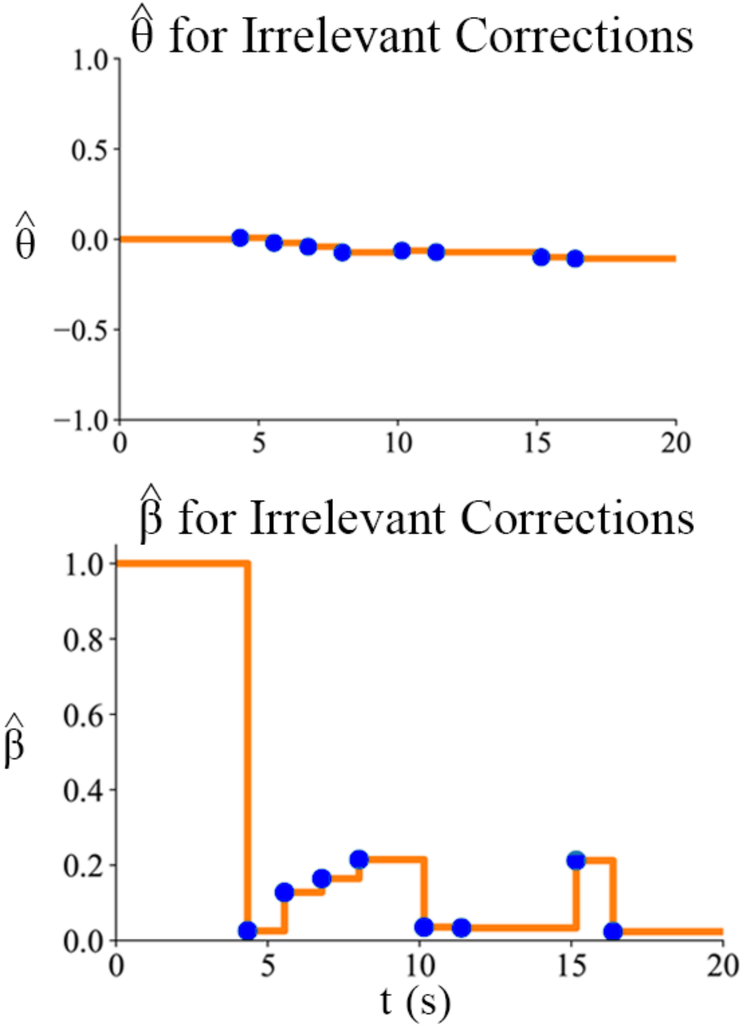
Detecting misspecification in physical corrections





Detecting misspecification in physical corrections





What is the right inductive bias for HRI?

Humans have intent

softmax

Bayes

Thanks!