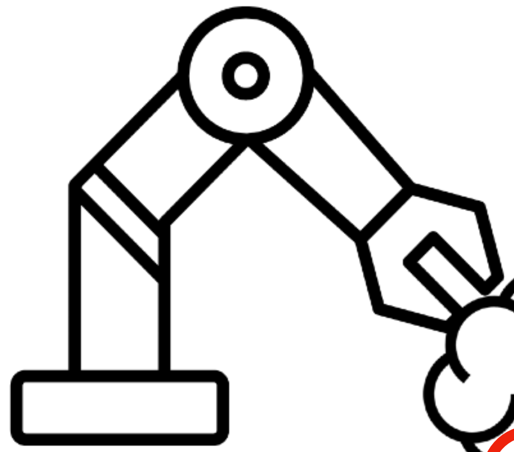
Estimating and penalizing preference shifts induced by recommender systems

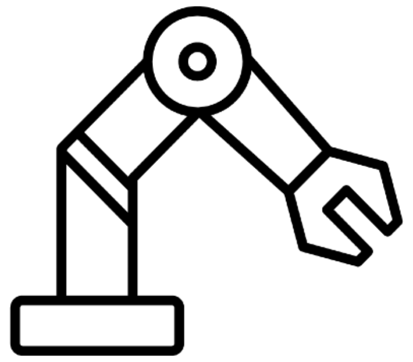
Micah Carroll, Dylan Hadfield-Menell, Stuart Russell, Anca Dragan

Preference influence? Preliminaries

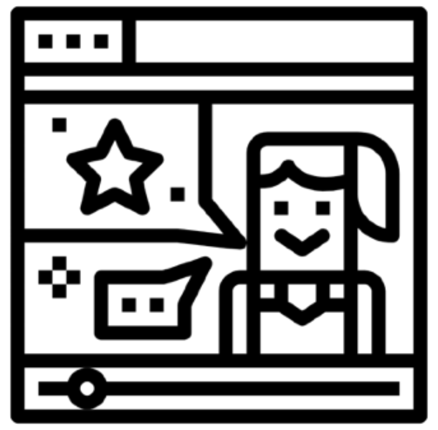


- Behavior on the platform
- User’s psychological state
- Polarization? Fake news / conspiracy content?
- Preferences?

User dynamics

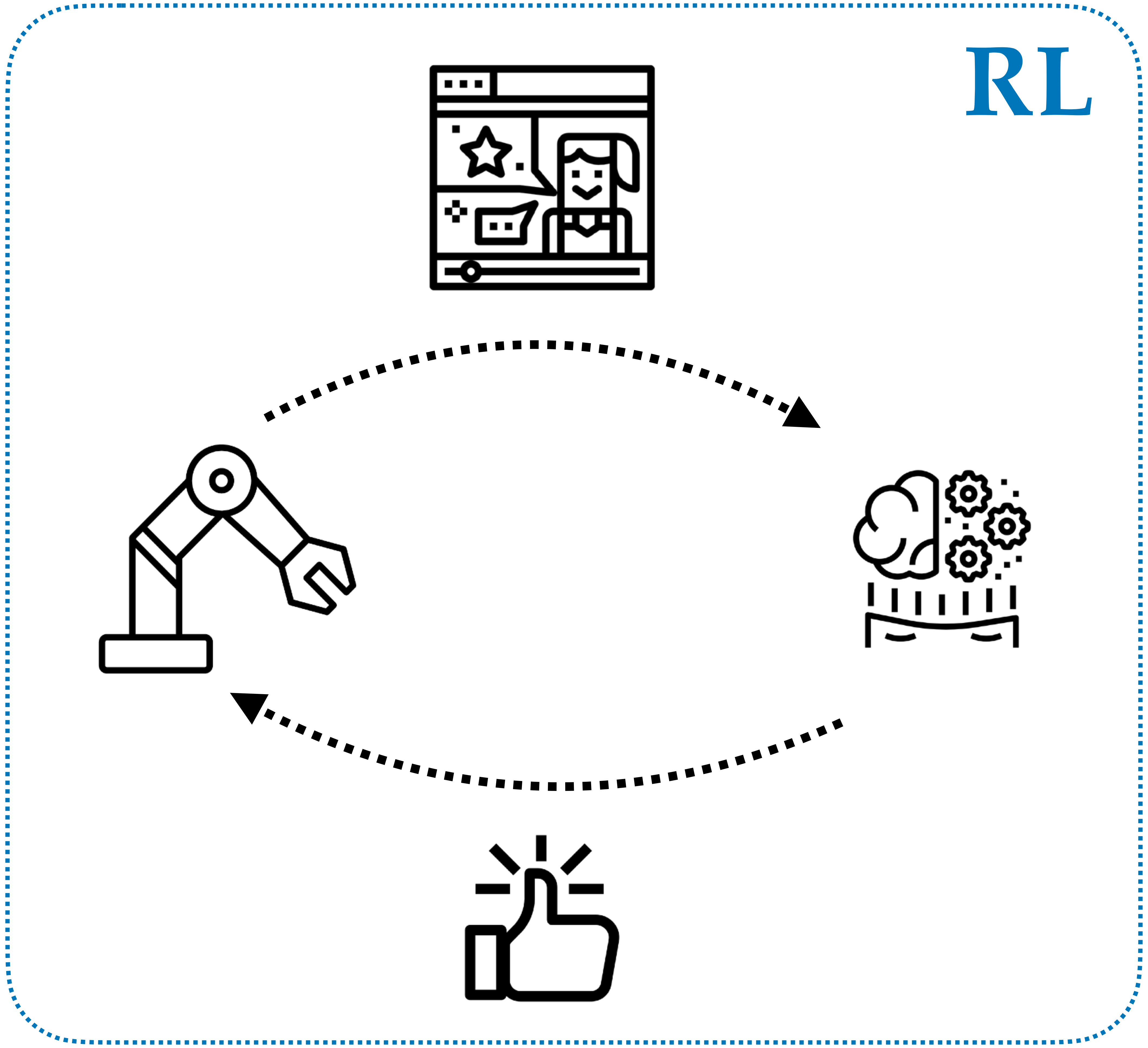


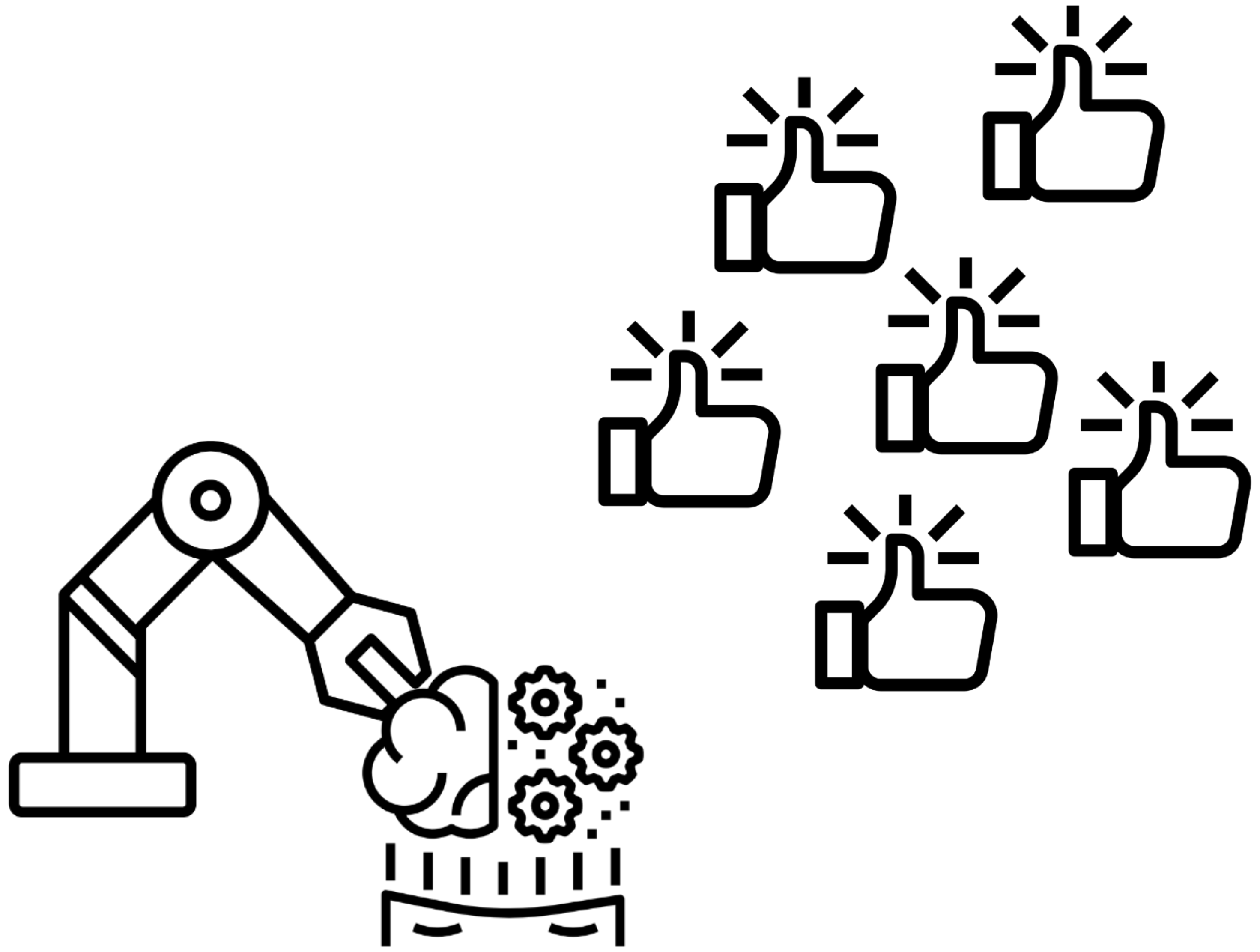
RS policy (i.e. the algorithm)



Other factors (UI, content-creators etc.)

Long-term-"value" systems

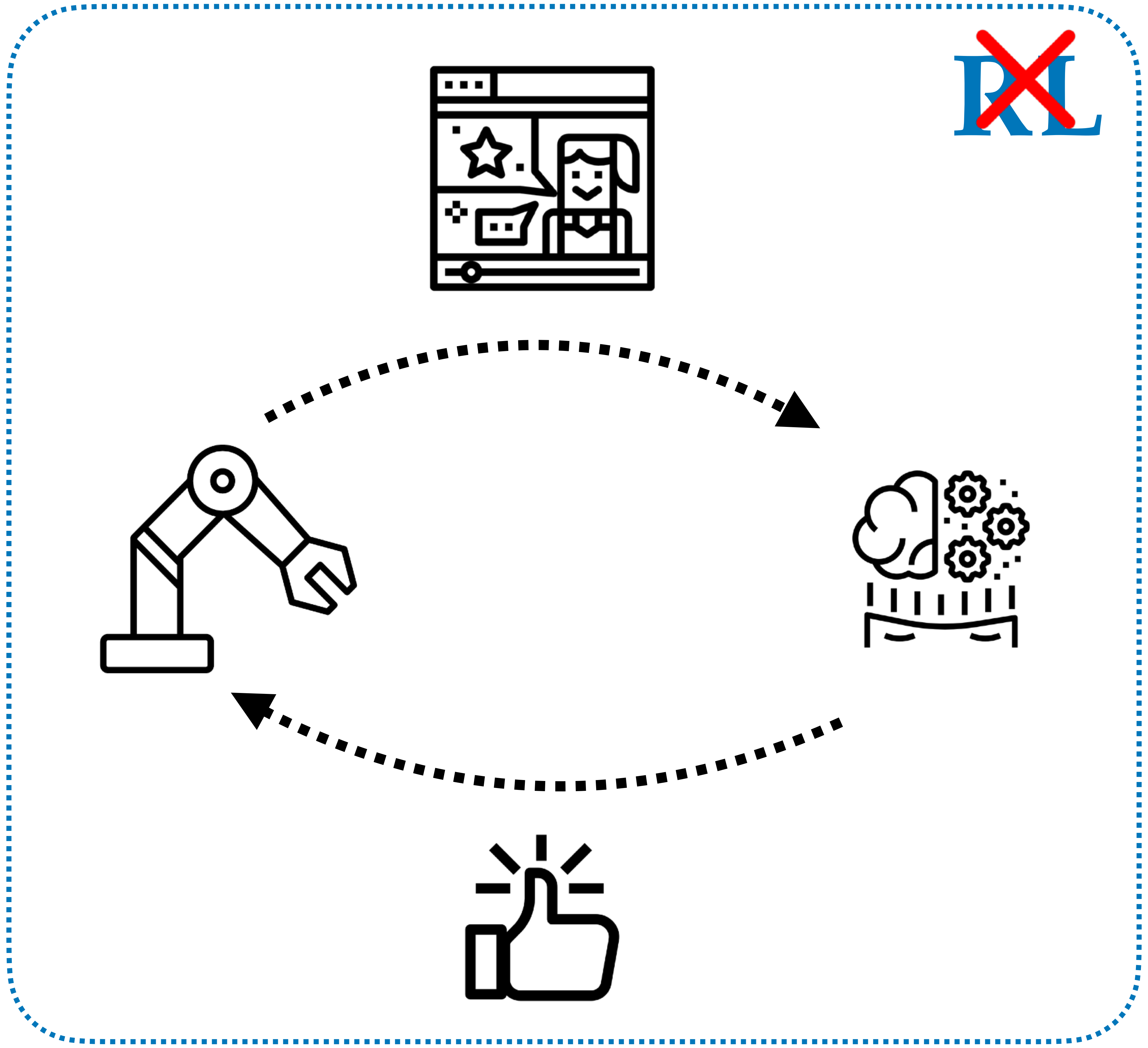




System will actively try to change the user by default!

[Krueger et. al, 2020] Hidden Incentives for Auto-Induced Distributional Shift

Long-term-“value” systems



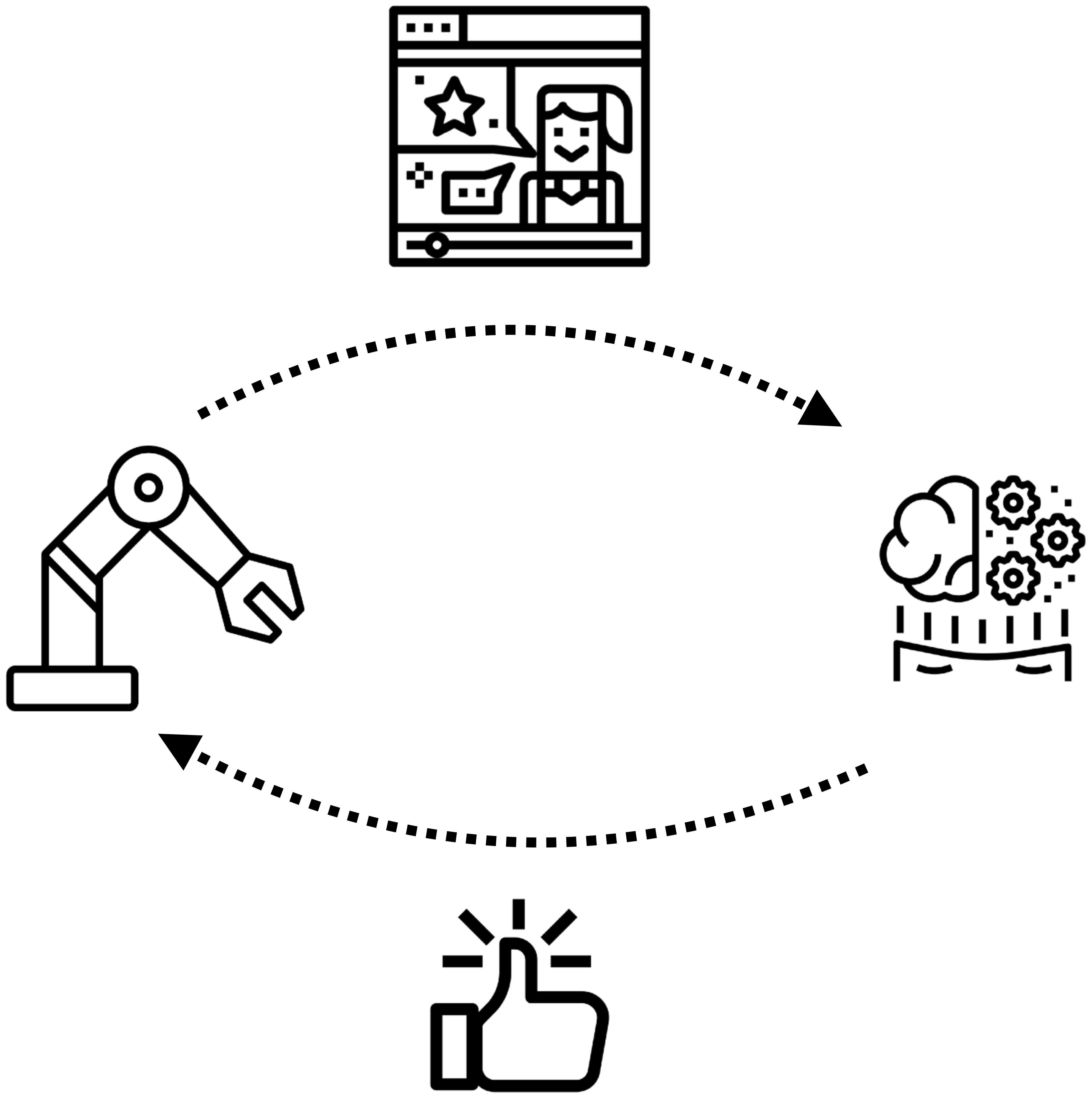
Preference shifts are assumed to be intrinsically value-less → *Misalignment!*

Make you easier to satisfy:

- Make you like common things
- Make you more predictable
  - Stabilize your preferences
  - Lead your choices to be less exploratory

[Krueger et. al, 2020] Hidden Incentives for Auto-Induced Distributional Shift

Myopic systems can still cause preference shifts



While myopia guarantees no active manipulation, it can still cause unwanted influence.

(Myopia is misaligned too)

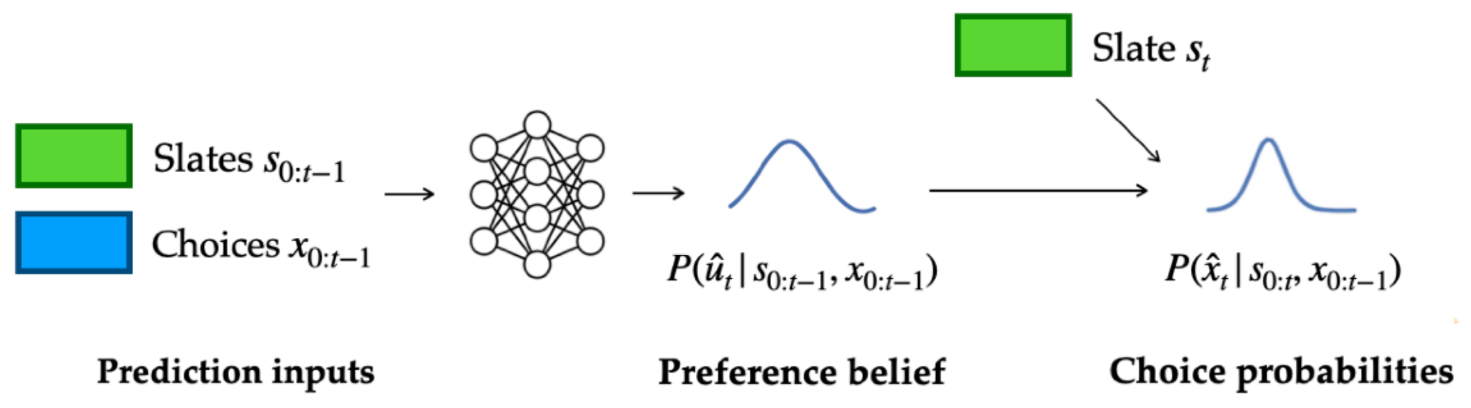
[Chaney et. al, 2017] How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility

[Jiang et. al, 2019] Degenerate Feedback Loops in Recommender Systems

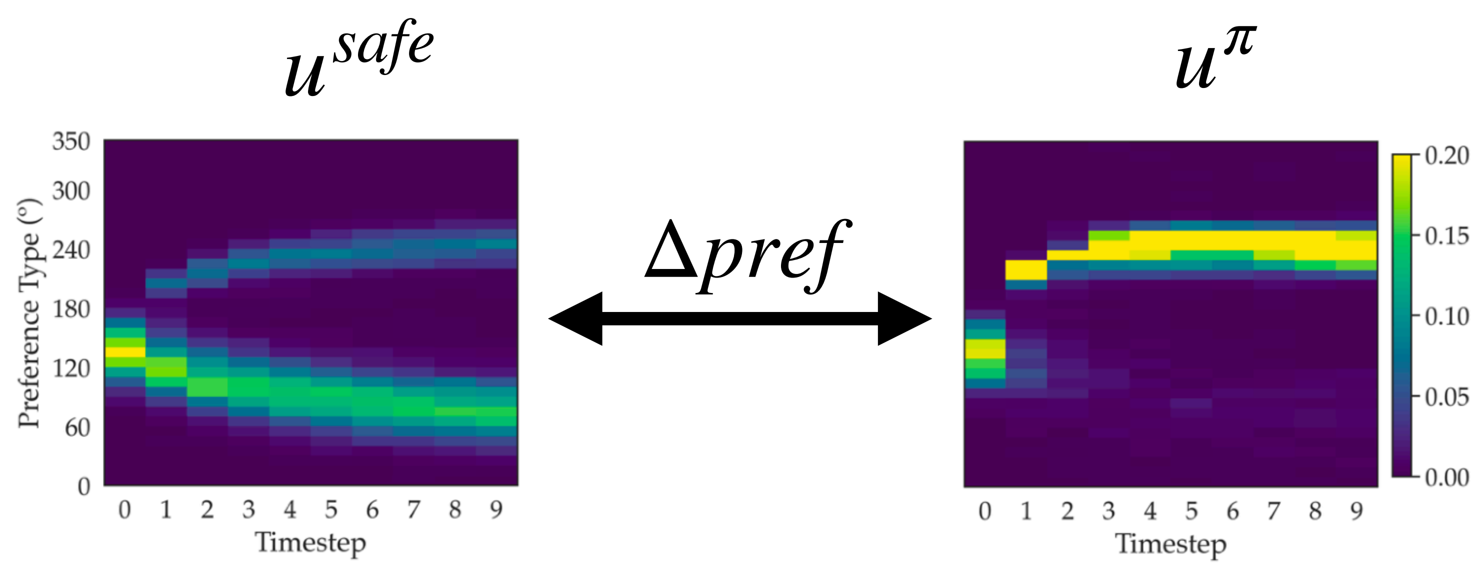
[Mansoury et. al, 2020] Feedback Loop and Bias Amplification in Recommender Systems

What I'll be talking about

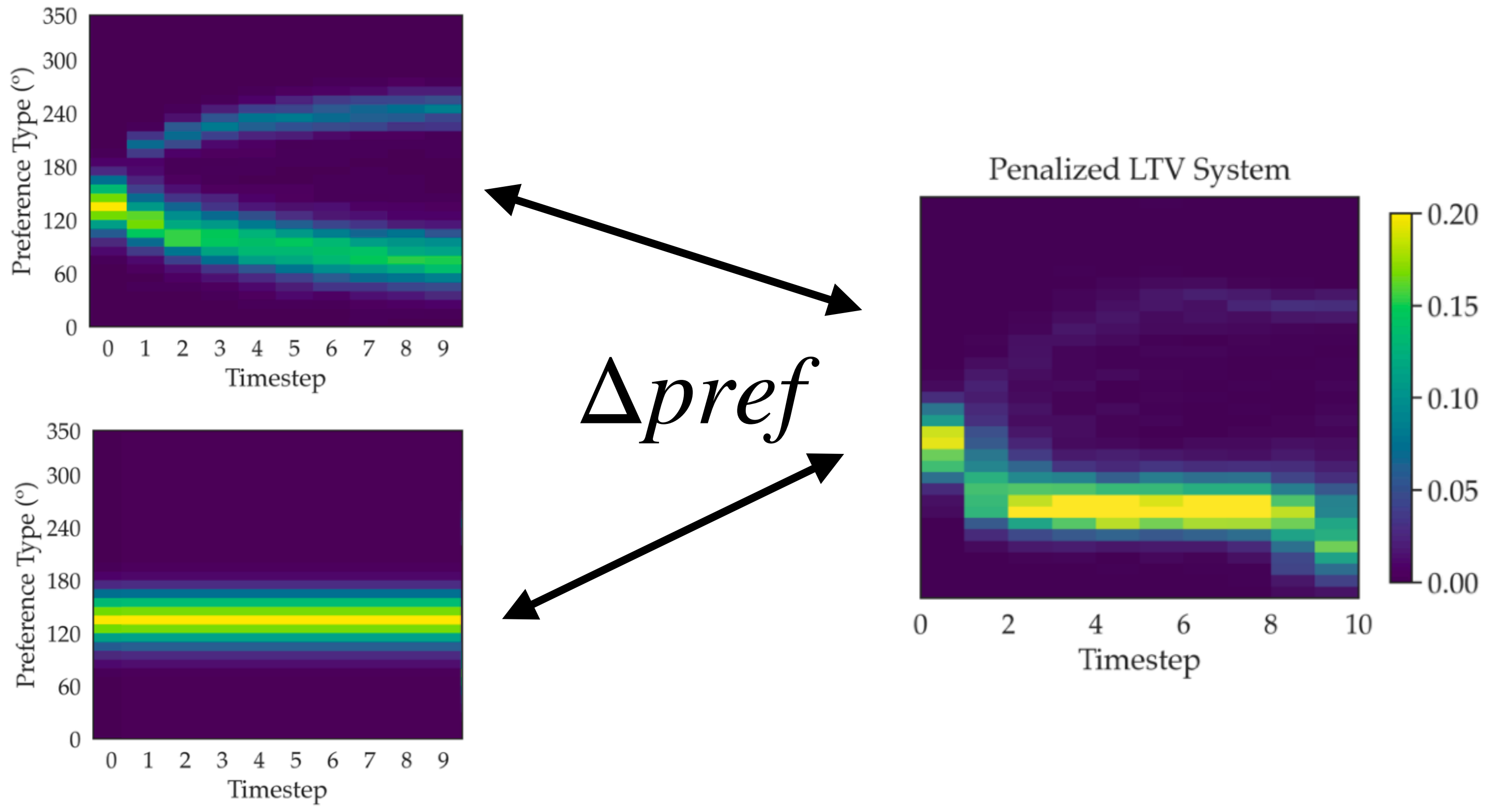
1. Method for estimating preference shifts that would be induced by a policy



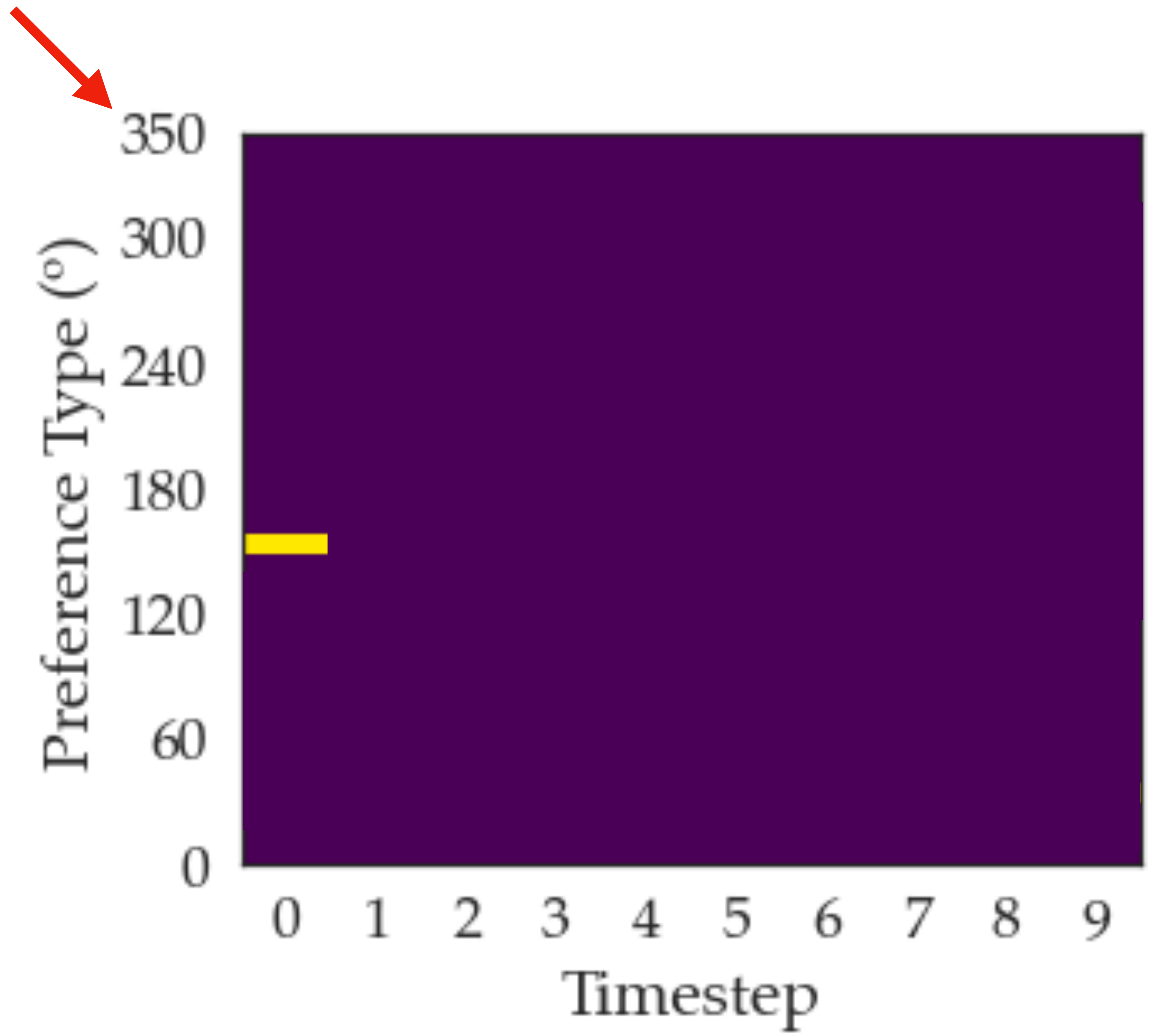
2. Framework for comparing induced shifts to "safe shifts"…

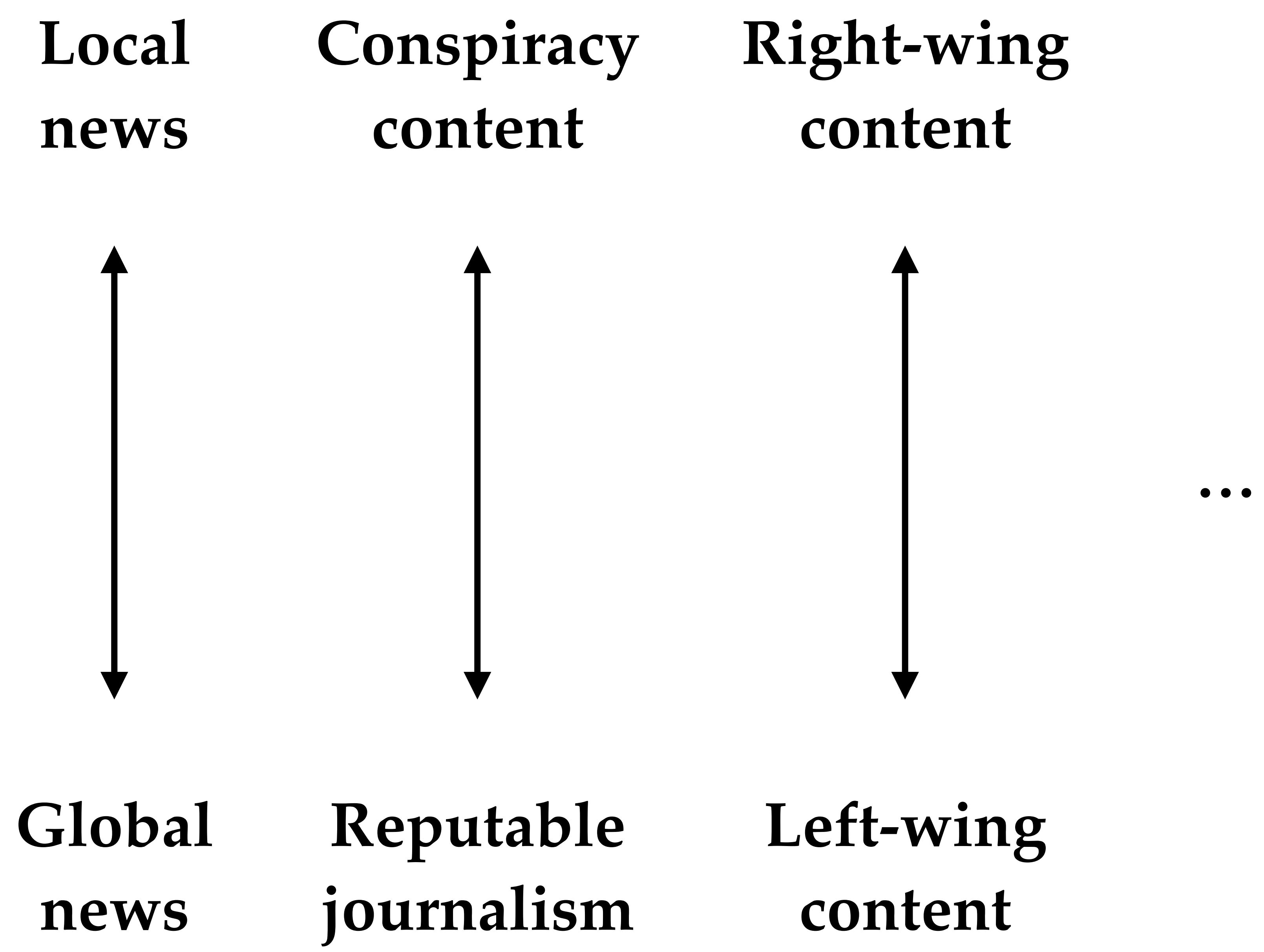


…which can be used to penalize RL training to actively avoid unwanted shifts

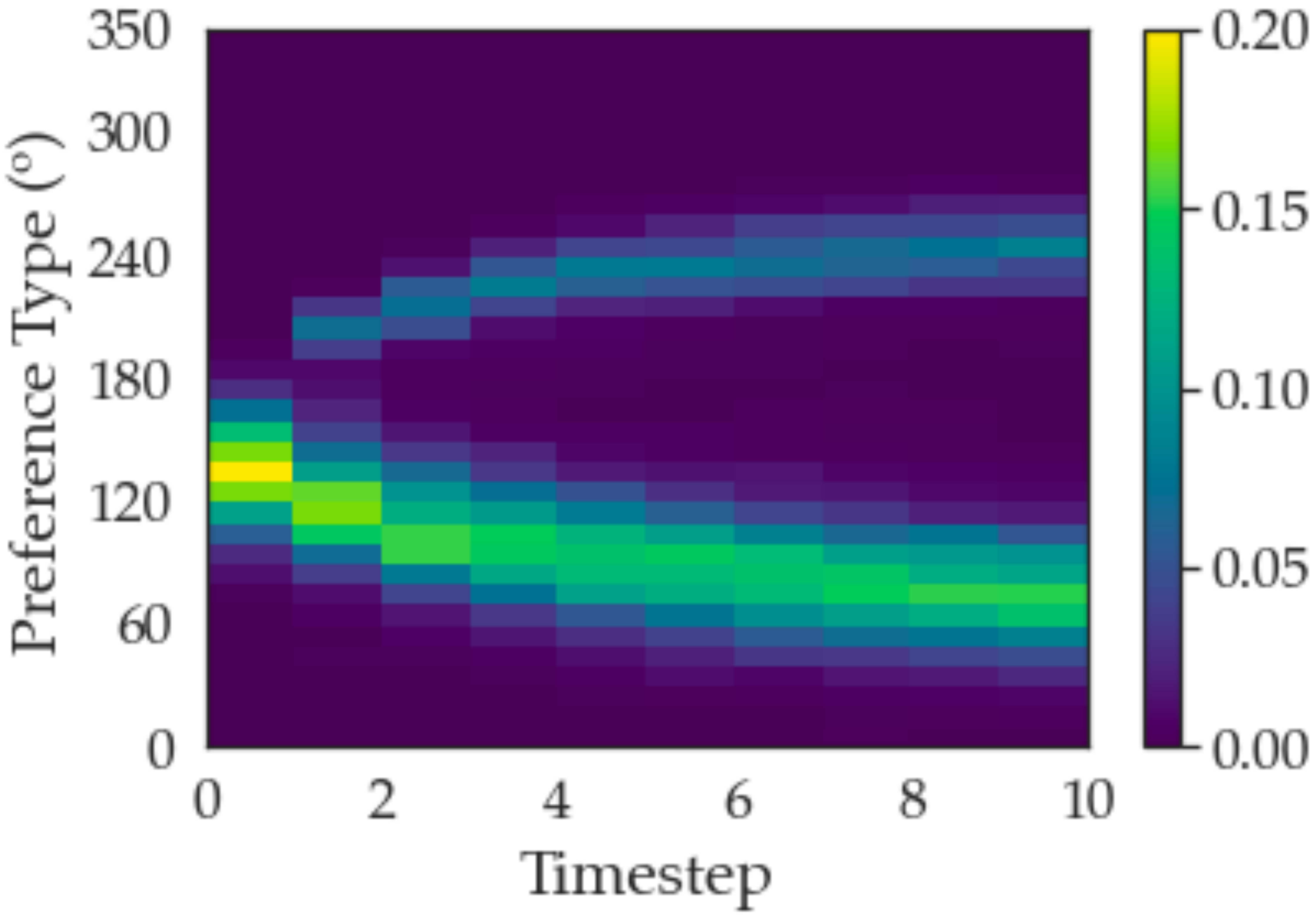


Preference influence? Preliminaries





Preference influence? Preliminaries



Sensationalized news

Reputable journalism

Policy-induced preference shifts

Learning model of human preference dynamics

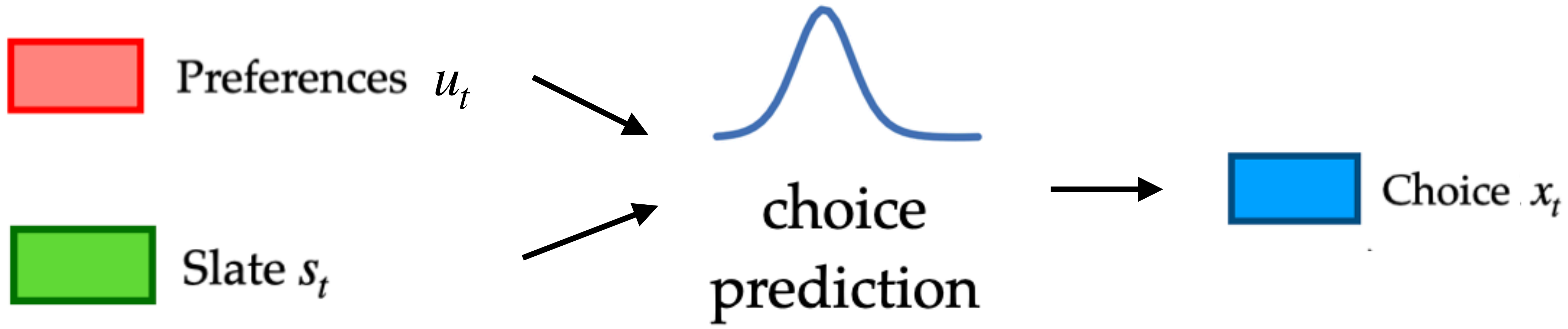
Preferences $u_t \in \mathbb{R}^d$

Slate s_t

Choice $x_t \in \mathbb{R}^d$

True Value r_t^* ?

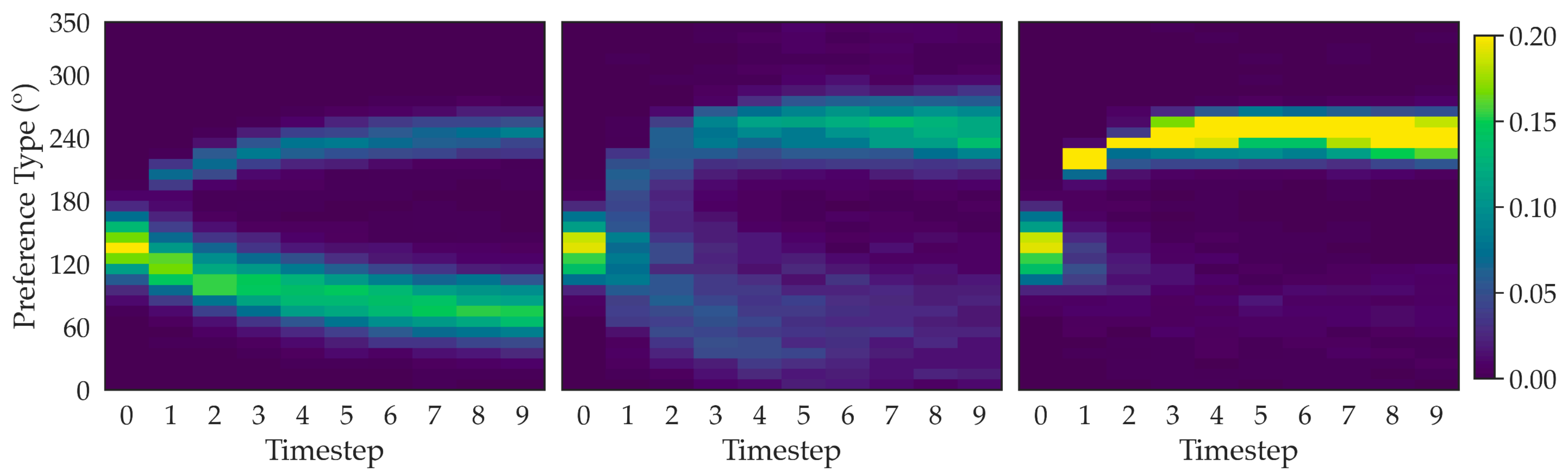
Assume human *choice model* is known

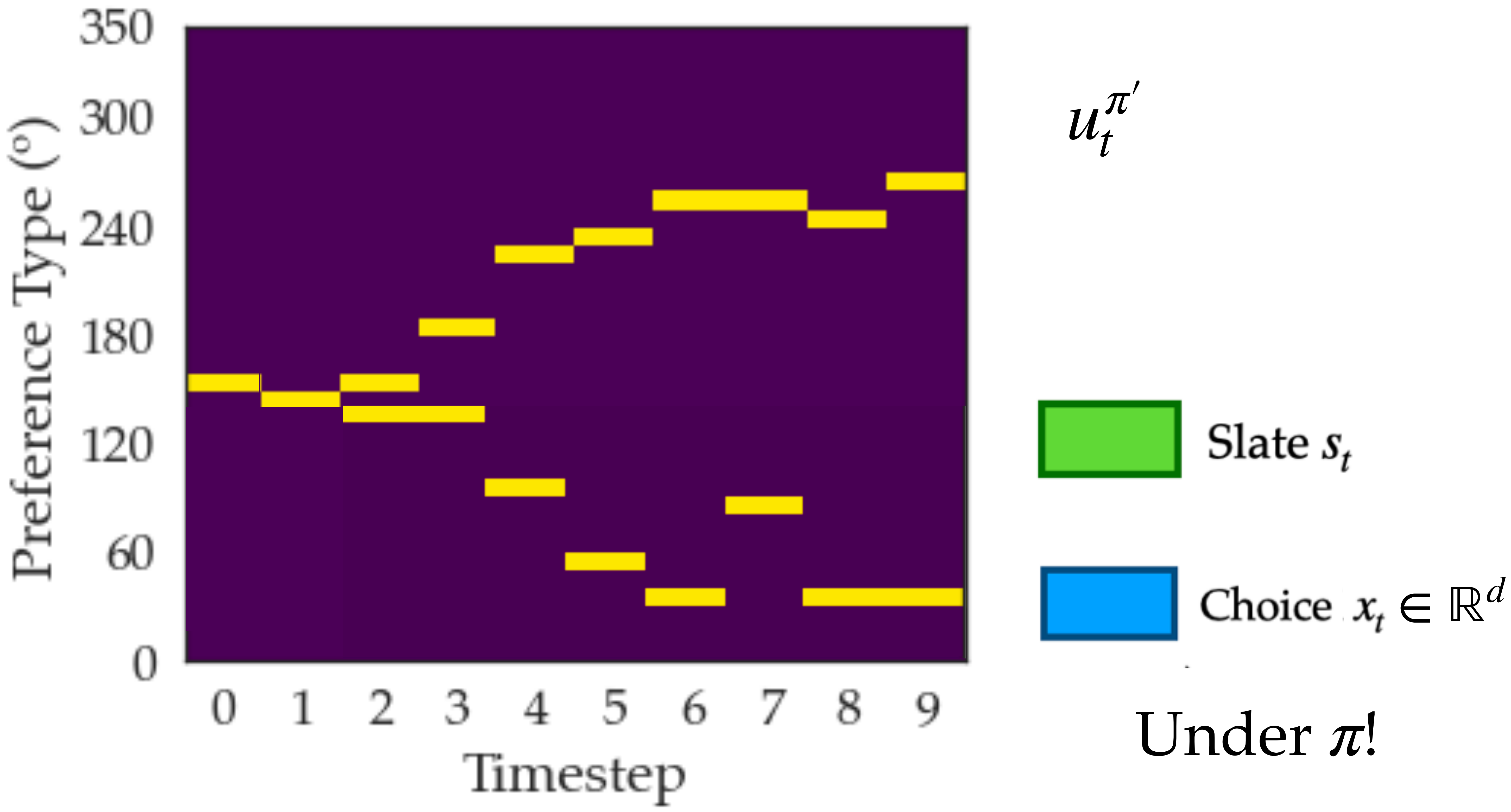


(Boltzmann rational)

Engagement Value $\hat{r}_t^{u_t} = u_t^T x_t$

Estimating counterfactual internal states





Under π!

Given $s^{\pi}_{0:T}, x^{\pi}_{0:T}$, estimate $u^{\pi'}_{0:T}$

Oracle access to internal state dynamics

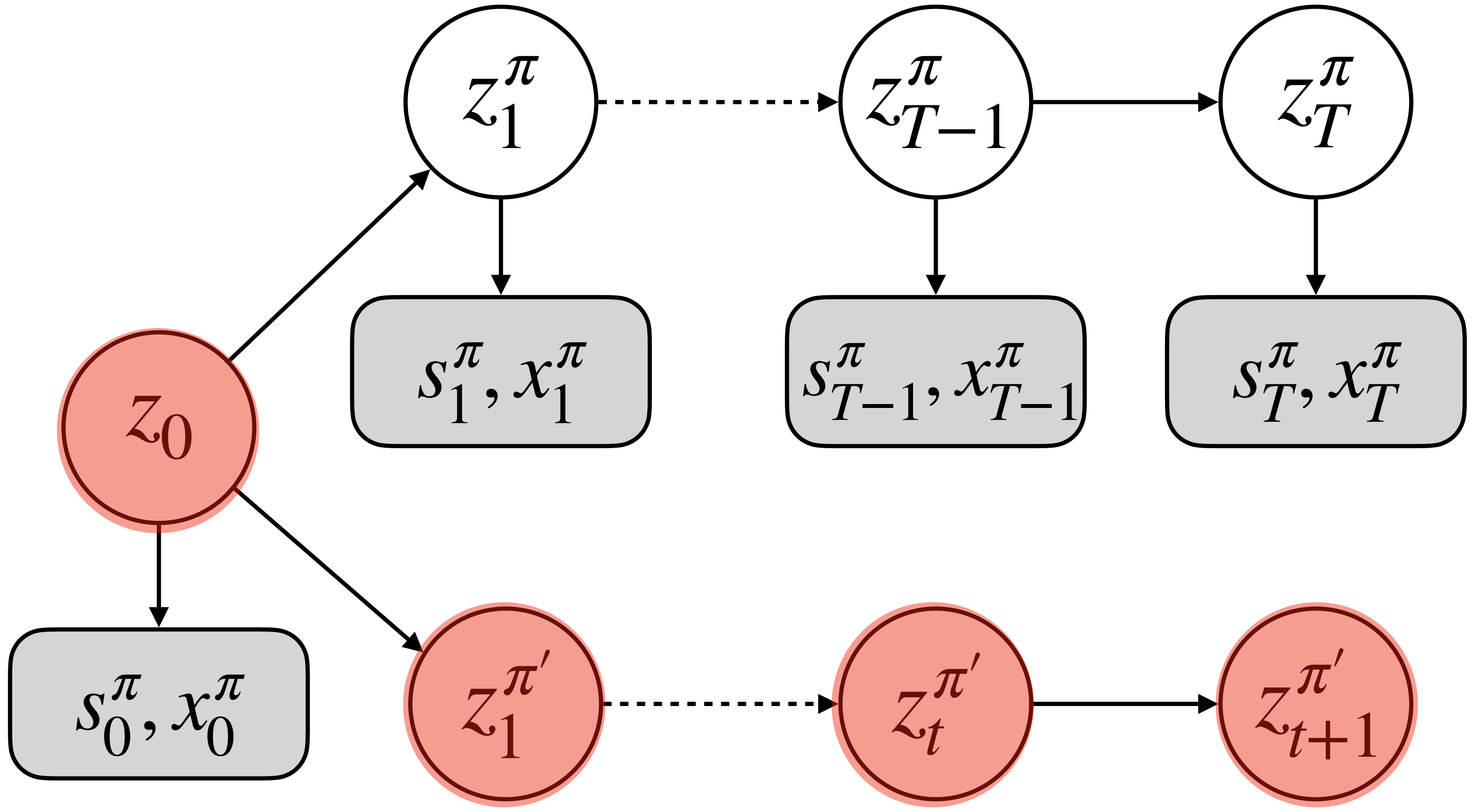
Preferences u_t

Beliefs

Psychological state

…

Internal state z_t

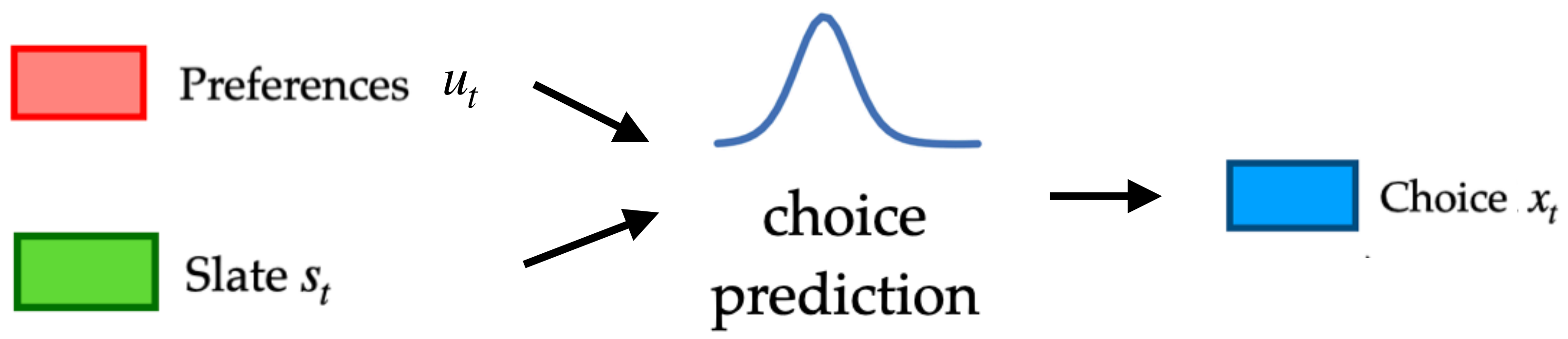


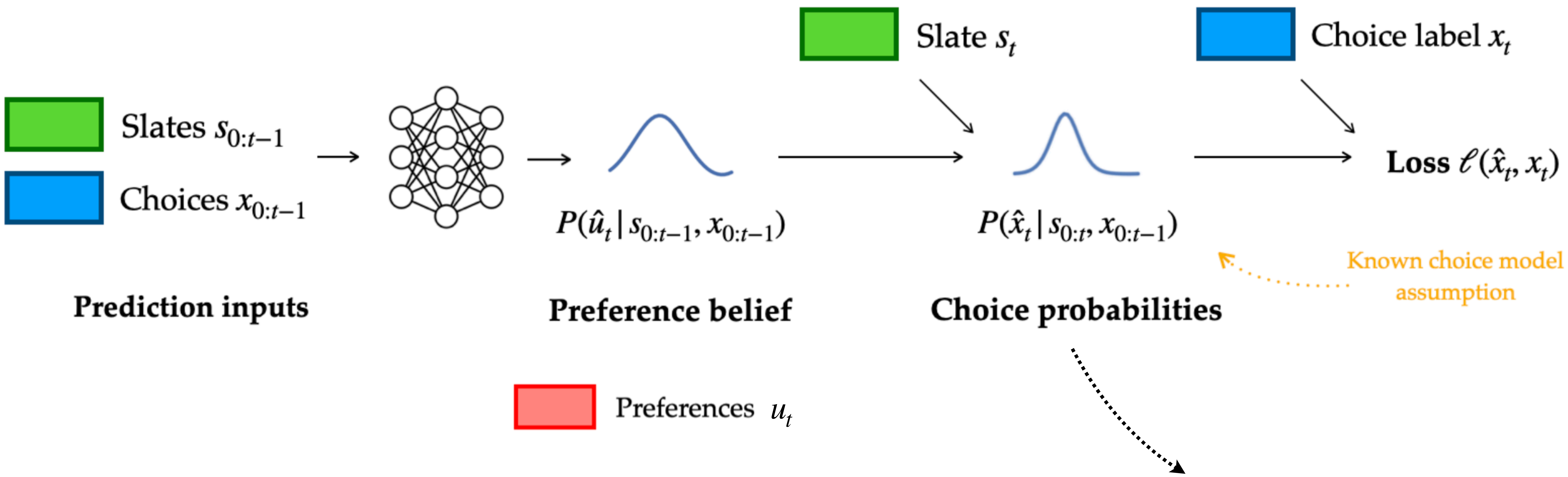
Given $s_{0:T}^{\pi}, x_{0:T}^{\pi}$, estimate $z_{0:T}$ $z_{0:T}^{\pi''}$

- Smoothing $P(z_0 \mid s_{0:T}^{\pi}, x_{0:T}^{\pi})$
- Forward prediction $P(z_t^{\pi'} \mid z_0)$

We don't have access to internal state dynamics!

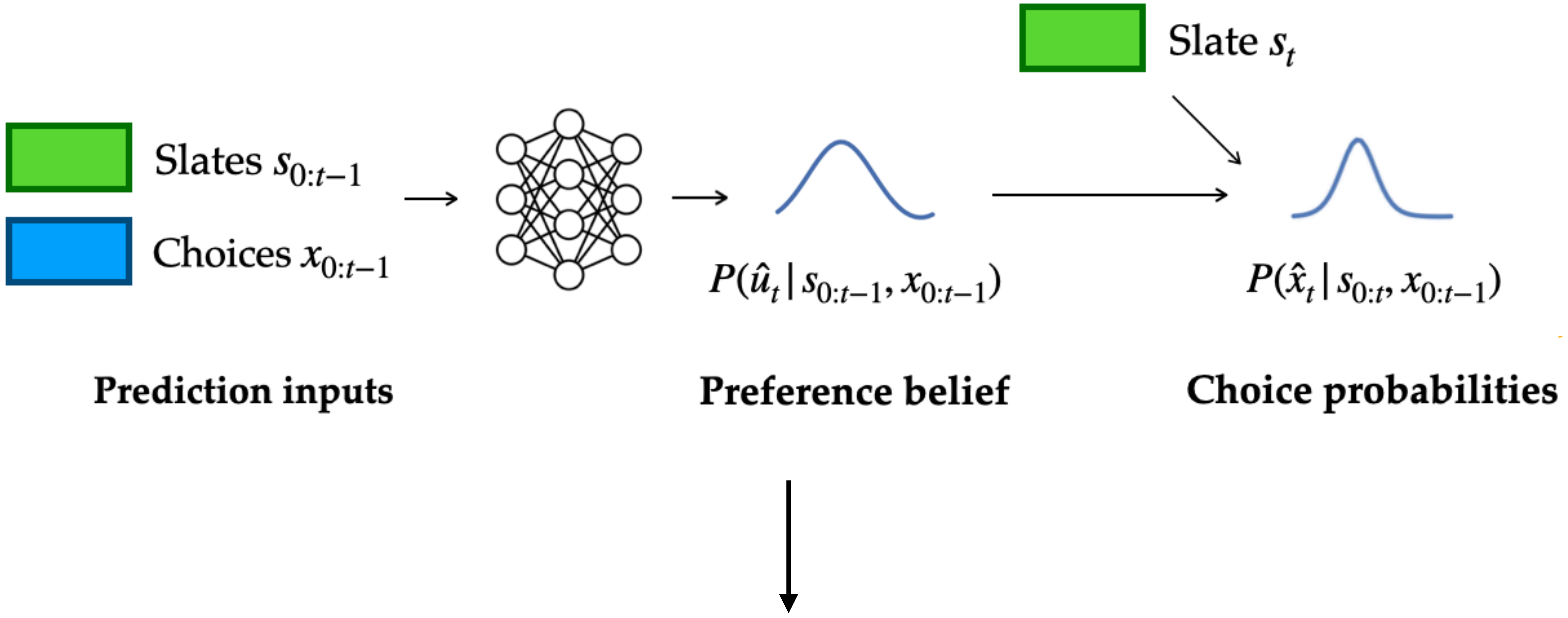
Approximating NHMM tasks without known dynamics





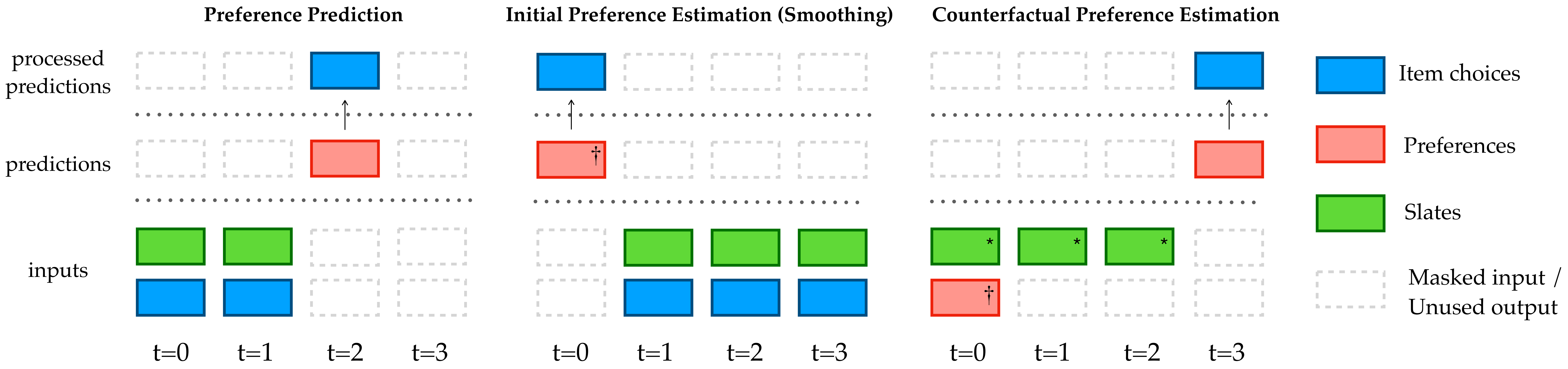
$$P(x_t|s_{0:t}, x_{0:t-1}) = \int_{u_t} P(x_t|u_t, s_t)P(u_t|s_{0:t-1}, x_{0:t-1})$$

Approximating NHMM tasks without known dynamics



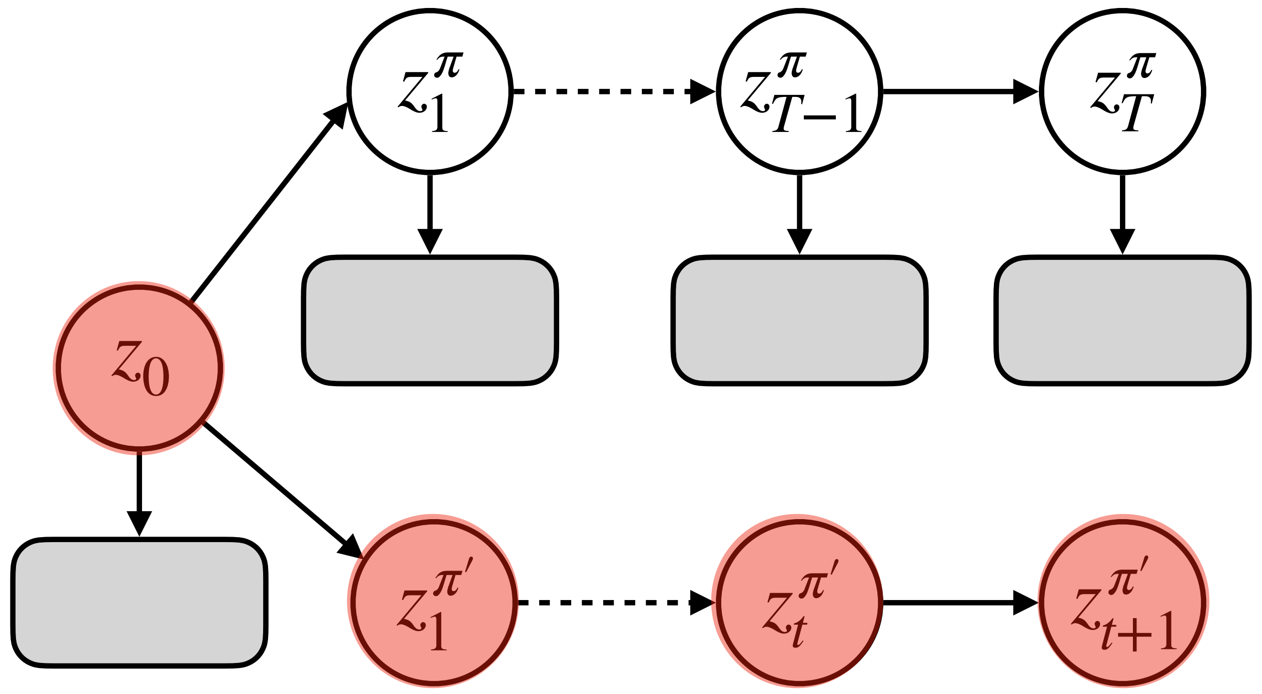
- **Future preferences** u_t
- **Initial preferences** u_0
- **Counterfactual preferences** (under different recsys policy π'): $u_t^{\pi'}$

Learning model of human preference dynamics



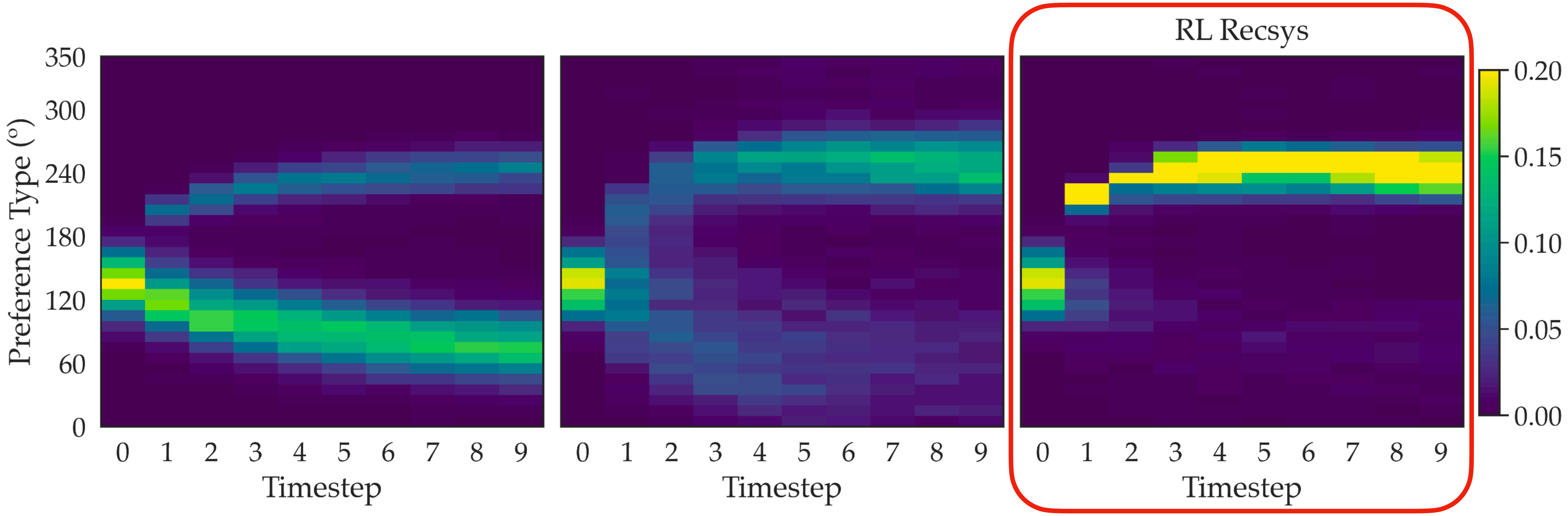
What we get $P(u_0|x_{1:t}, s_{1:t})$

What we want $P(u_0|x_{0:t}, s_{0:t}) \propto P(x_0, s_0|u_0)P(u_0|x_{1:t}, s_{1:t})$



Policy-induced preference shifts

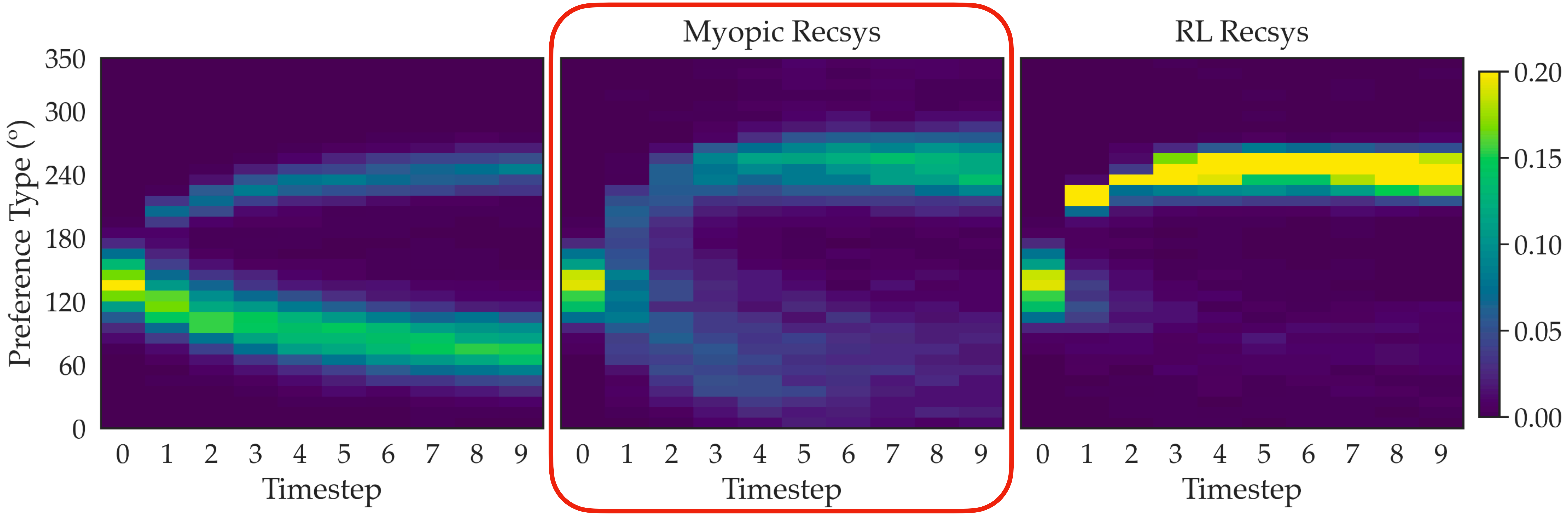
RL RS-induced preference shifts



High engagement

Likely
undesirable shifts?

Myopic RS-induced preference shifts



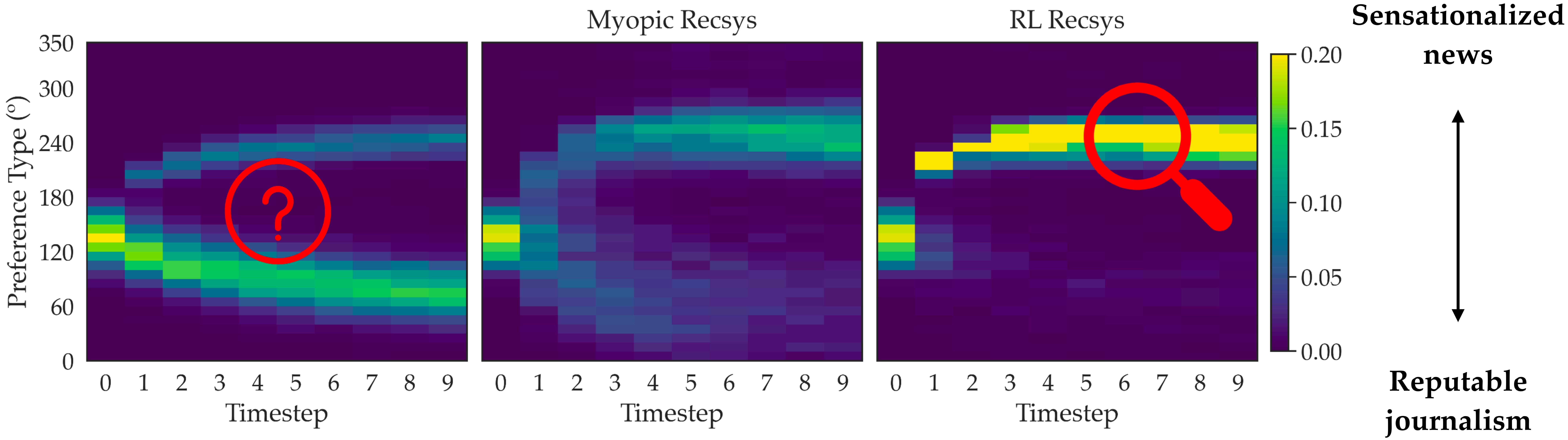
Medium engagement

Potentially undesirable shifts?

High engagement

Likely undesirable shifts?

What preference shifts do we want?



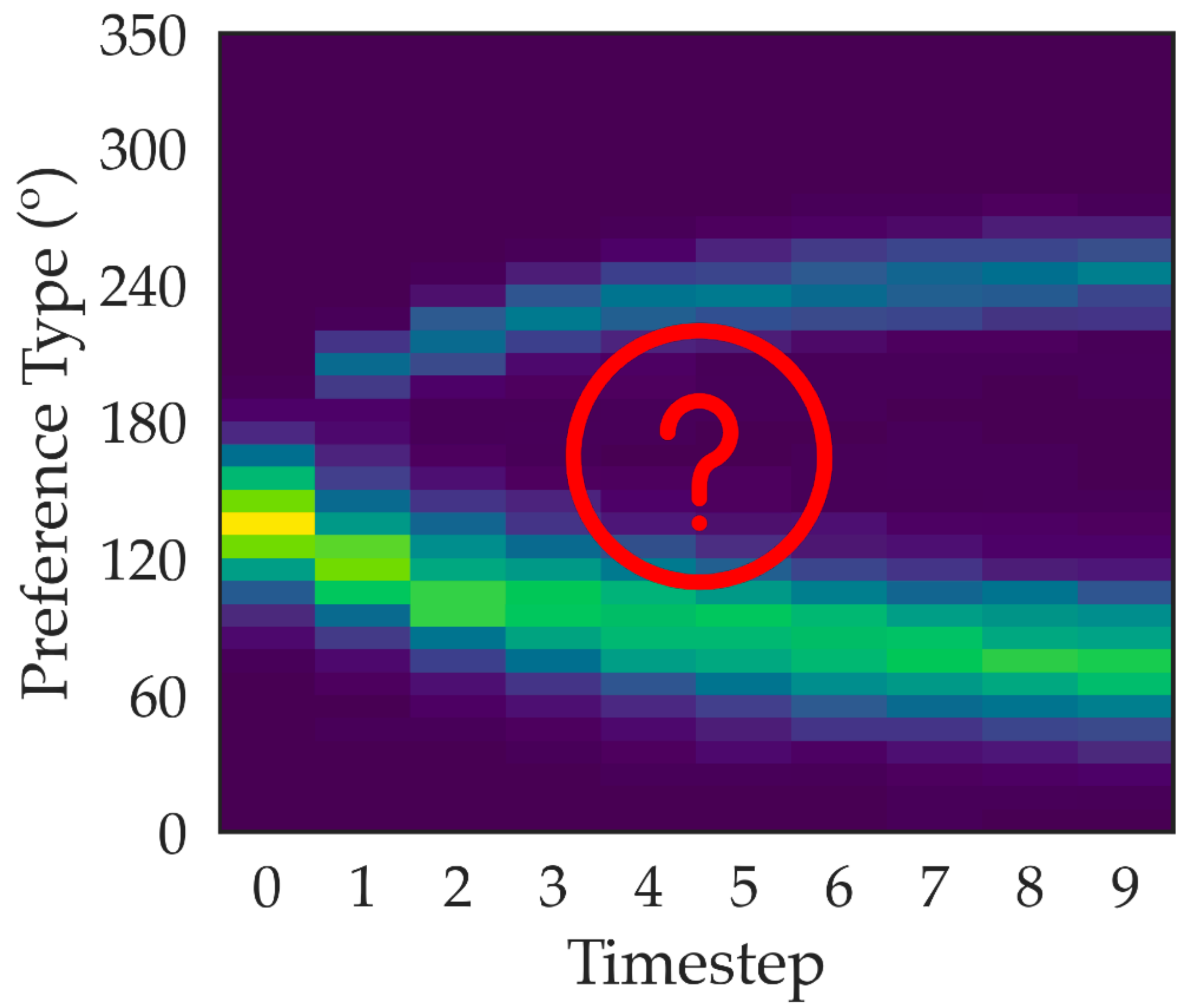
Medium engagement **High engagement**

"Desirable" shifts?

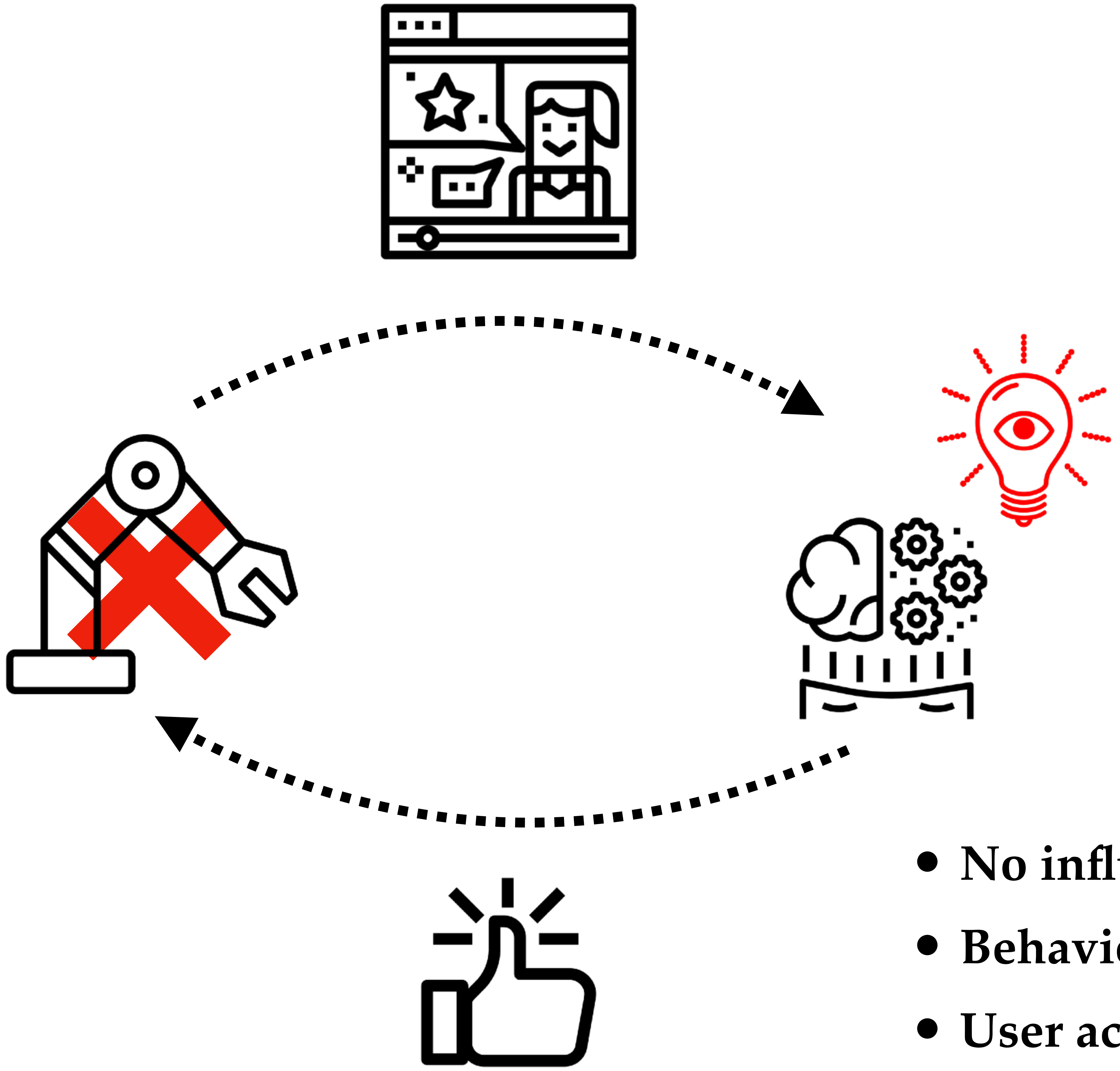
Potentially undesirable shifts?

Likely undesirable shifts?

Ideally, what would we want?

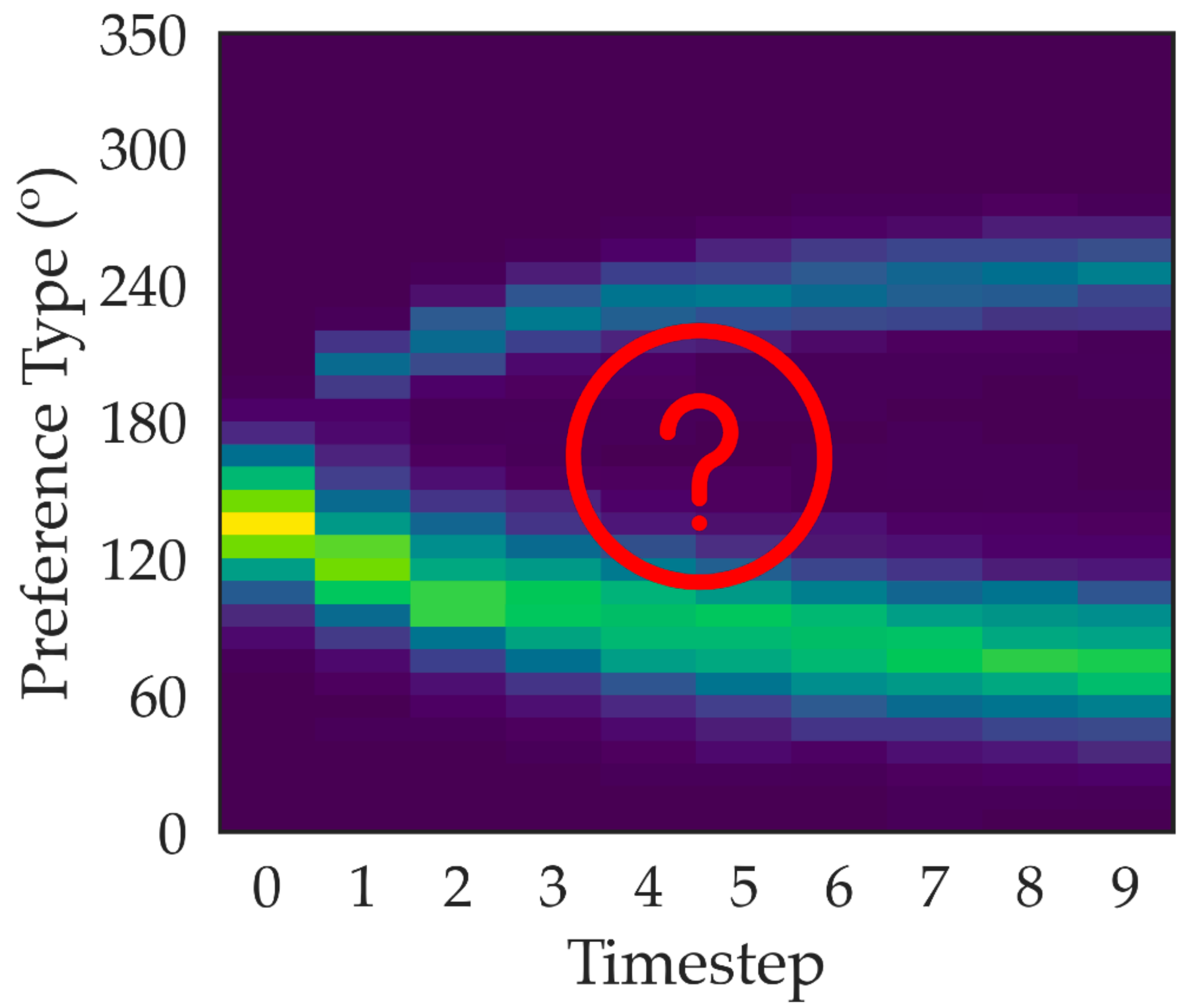


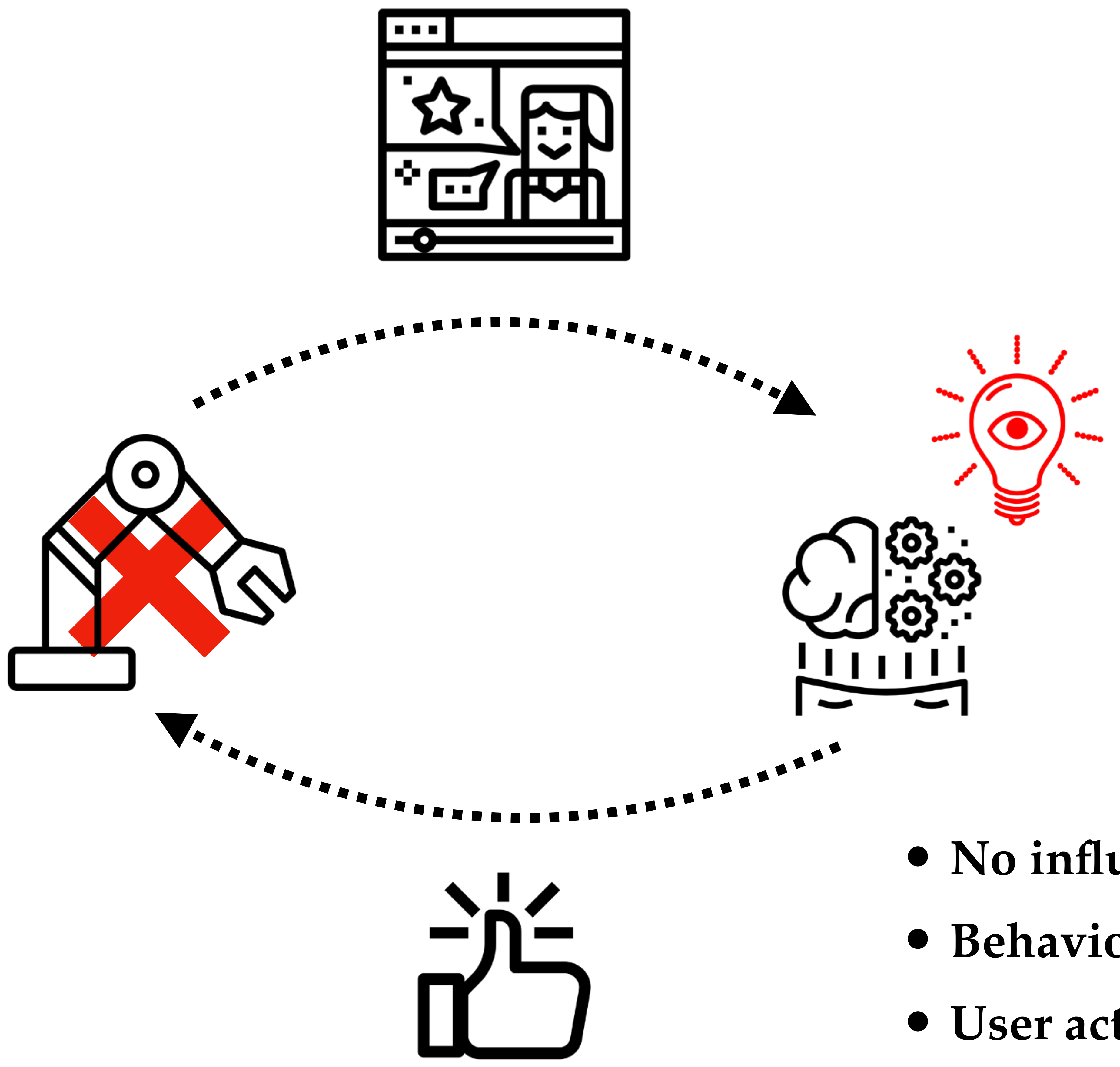
"Desirable" shifts?



- **No influence from RS**
- **Behavior under full information**
- **User acting in the interest of their "best-self"**

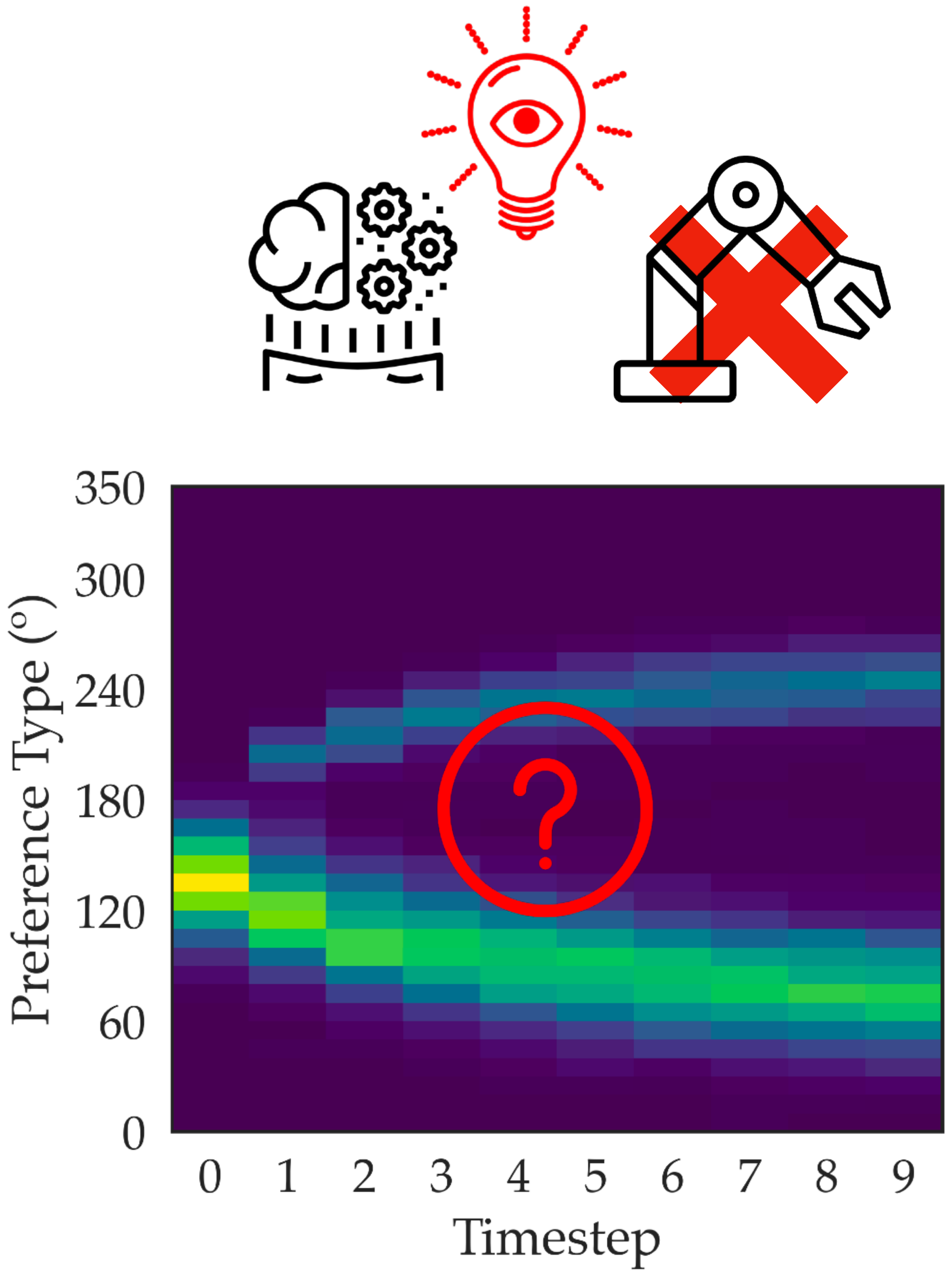
Ideally, what would we want?



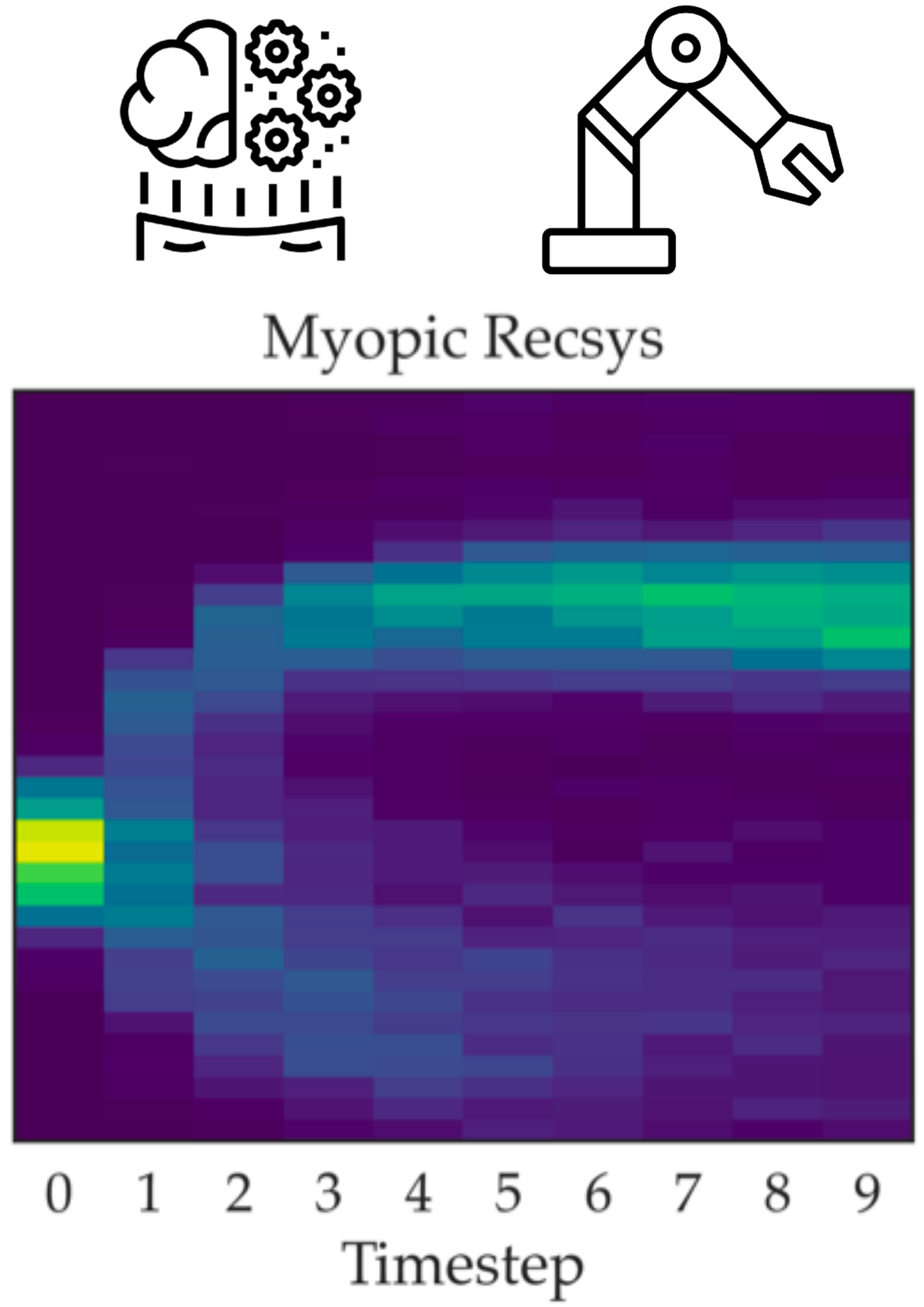


- No influence from RS
- Behavior under full information
- User acting in the interest of their "best-self"

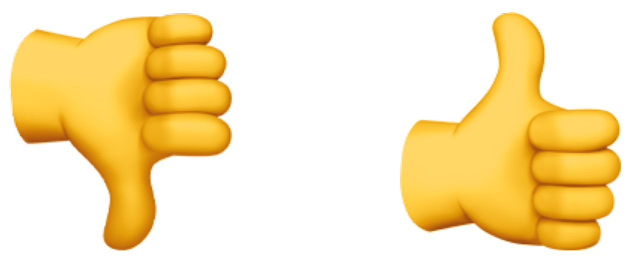
Ideally, what would we want?



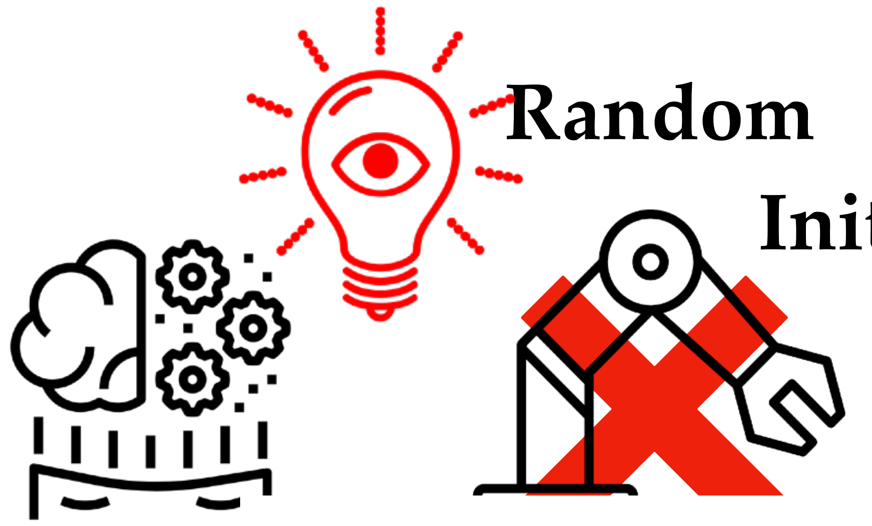
$\Delta pref$



Desirable shifts



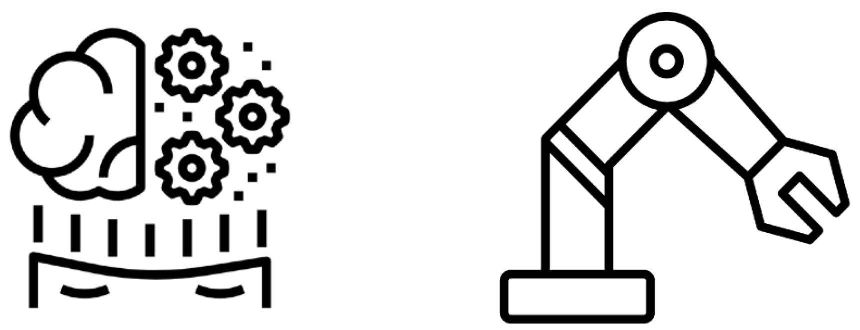
Realistically, what can we do?

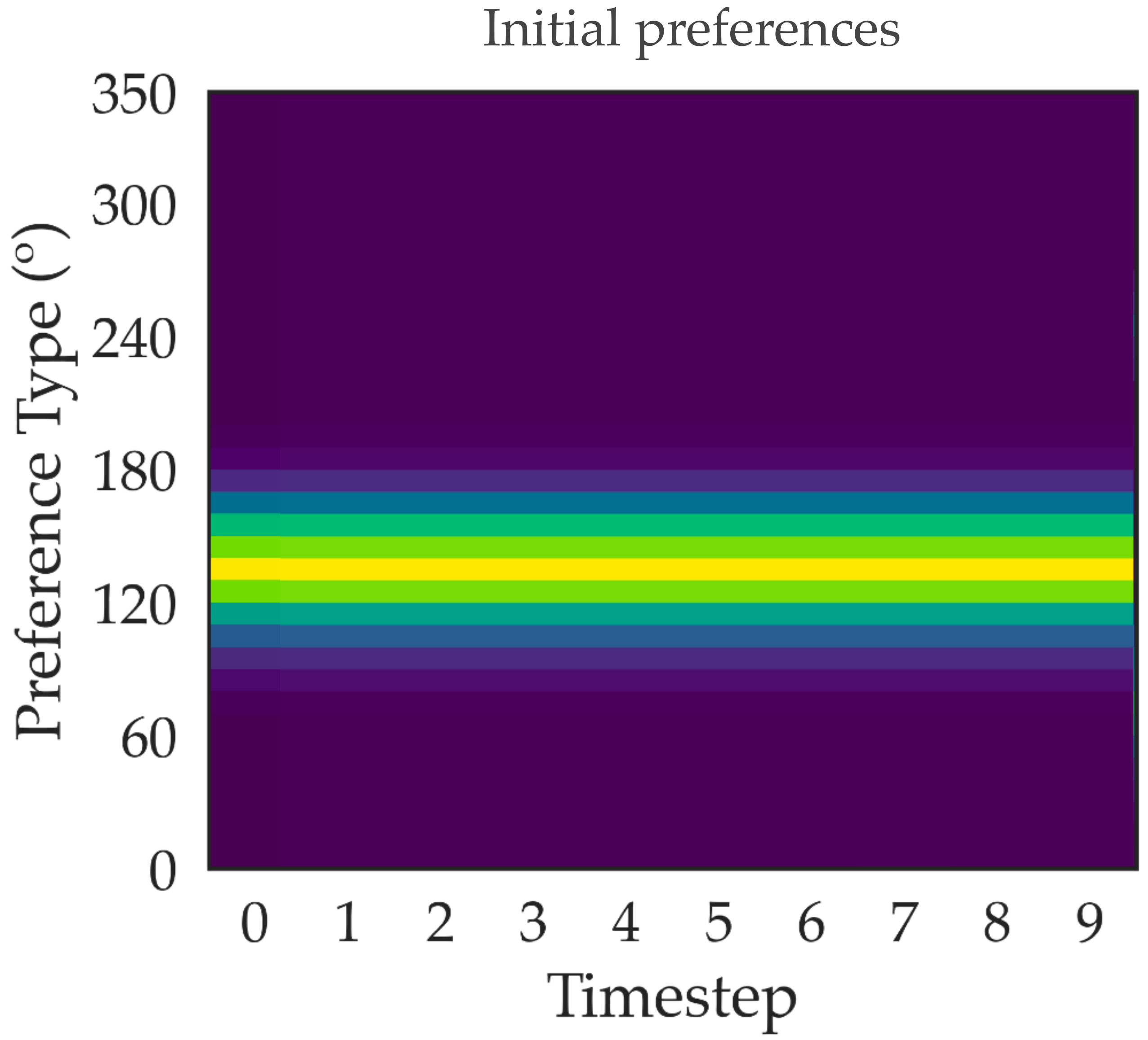


Random

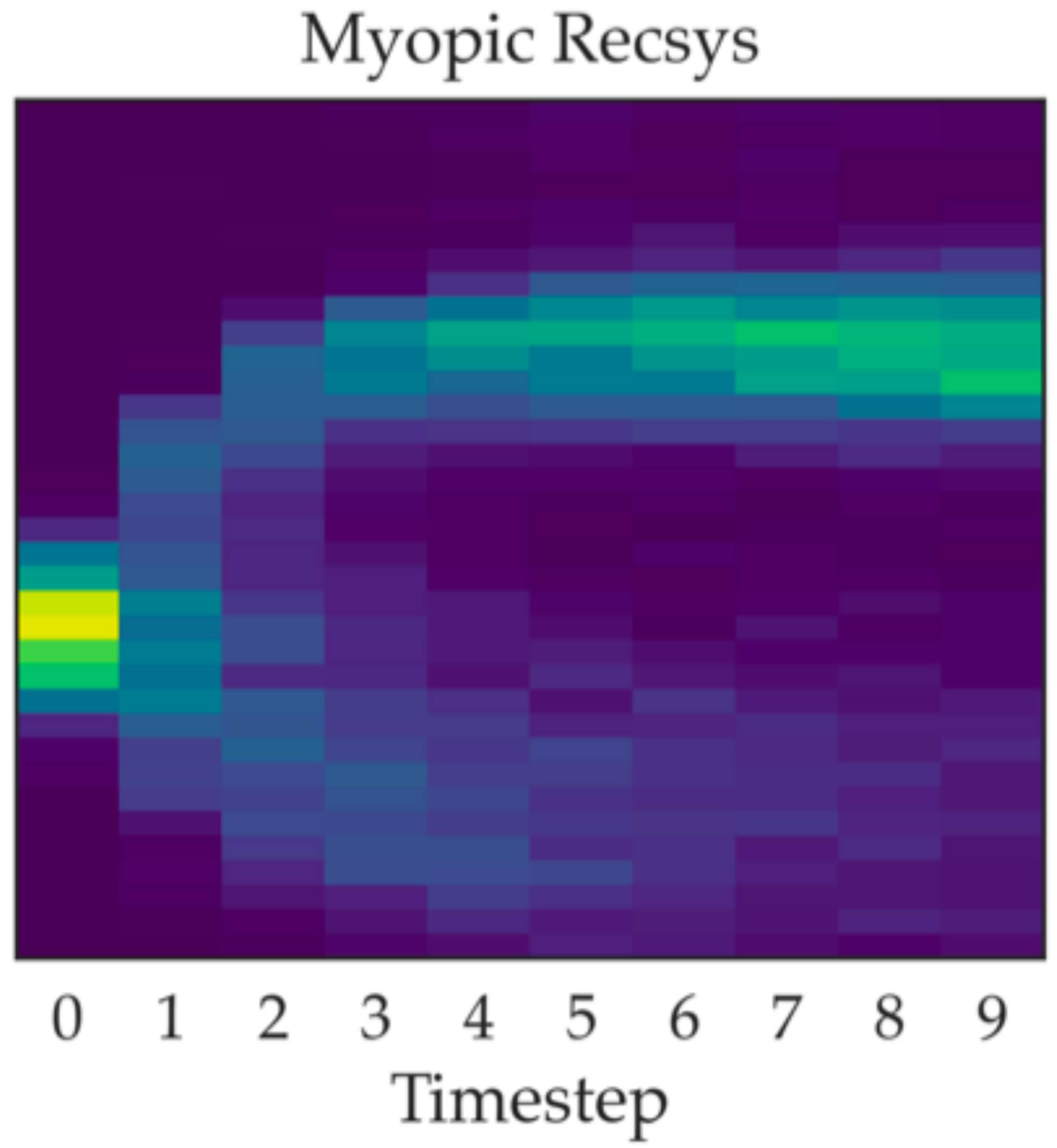
Initial pref. - don't change anything!

Collective / political decision?



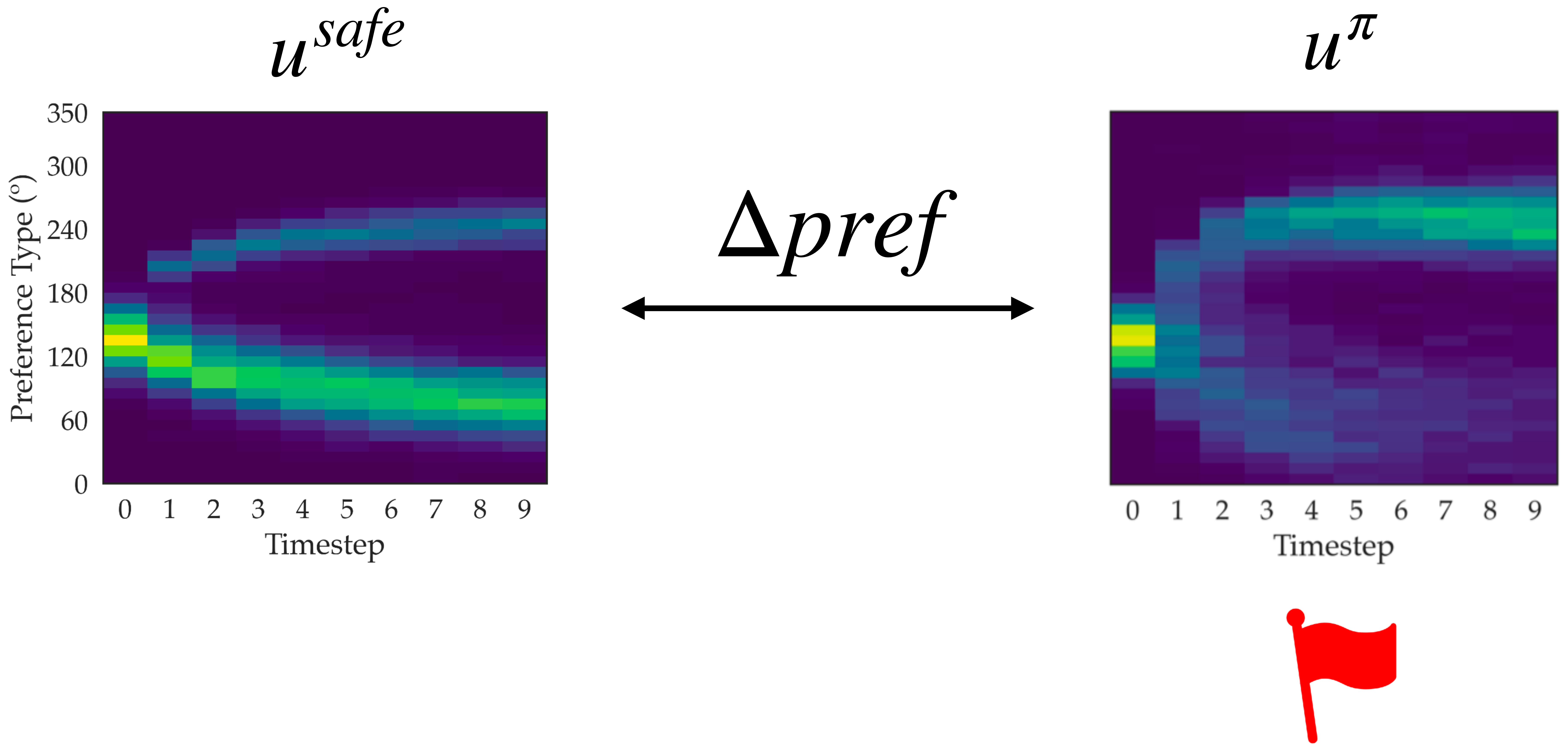


$\Delta pref$



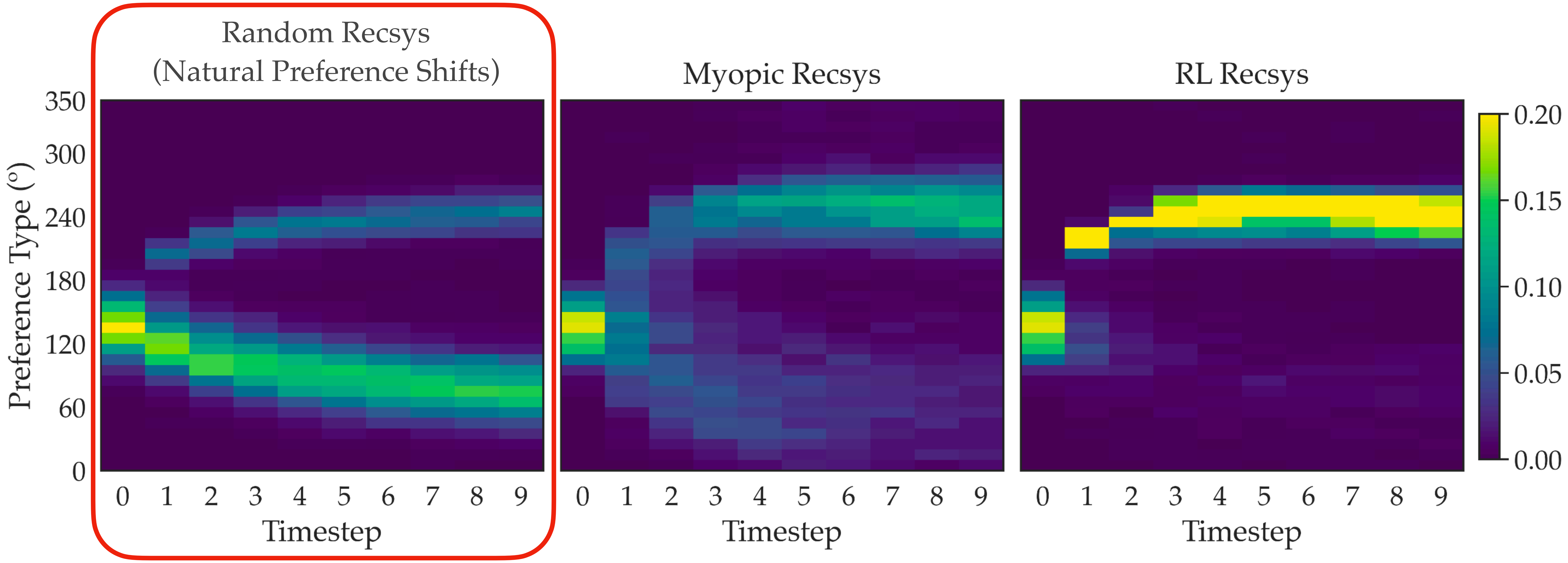
"Natural preference shifts" (NPS)

Realistically, what can we do?



A framework for creating <u>conservative metrics</u> for a policy π's degree of unwanted preference shift

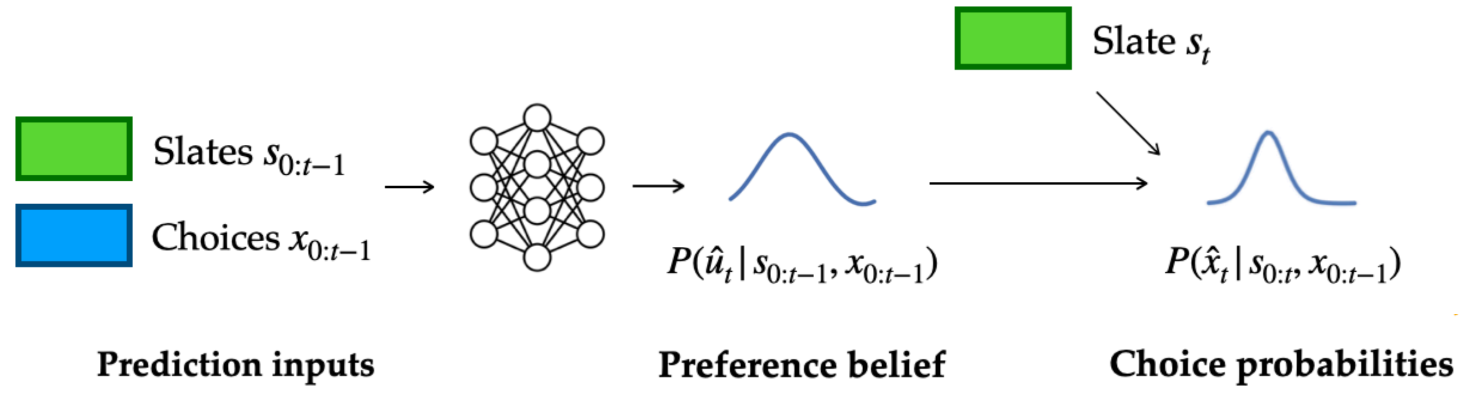
What preference shifts do we want?



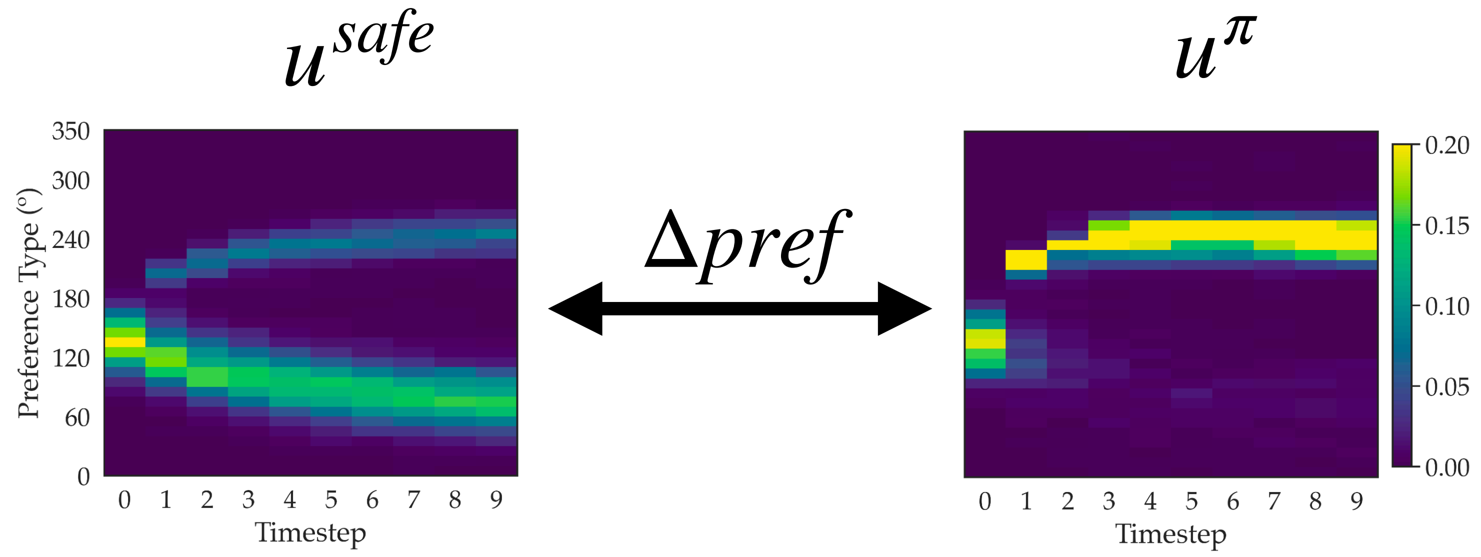
Low engagement	Medium engagement	High engagement
Safe shifts	Potentially undesirable shifts?	Likely undesirable shifts?

What I'll be talking about

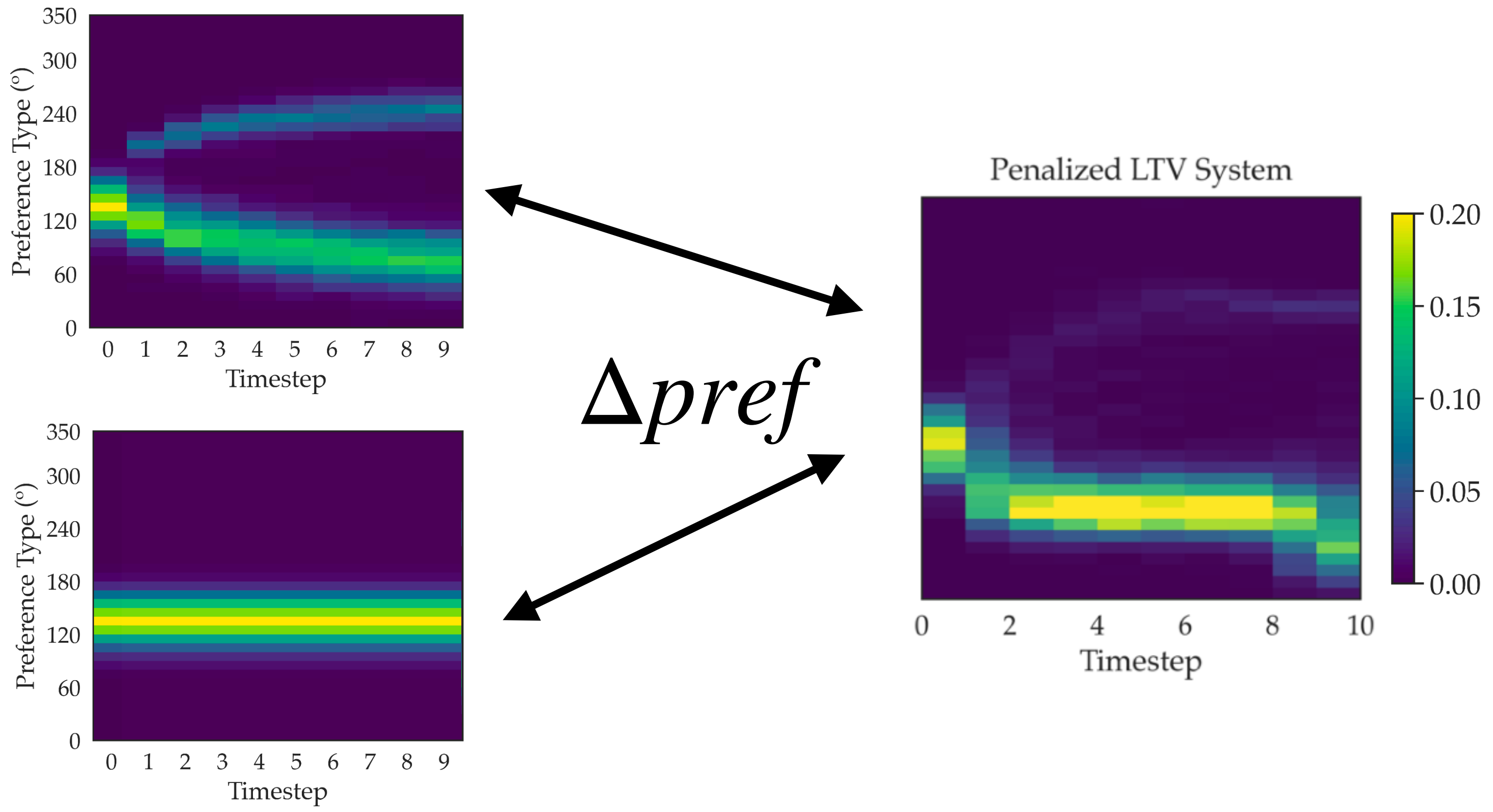
1. Method for estimating preference shifts that would be induced by a policy



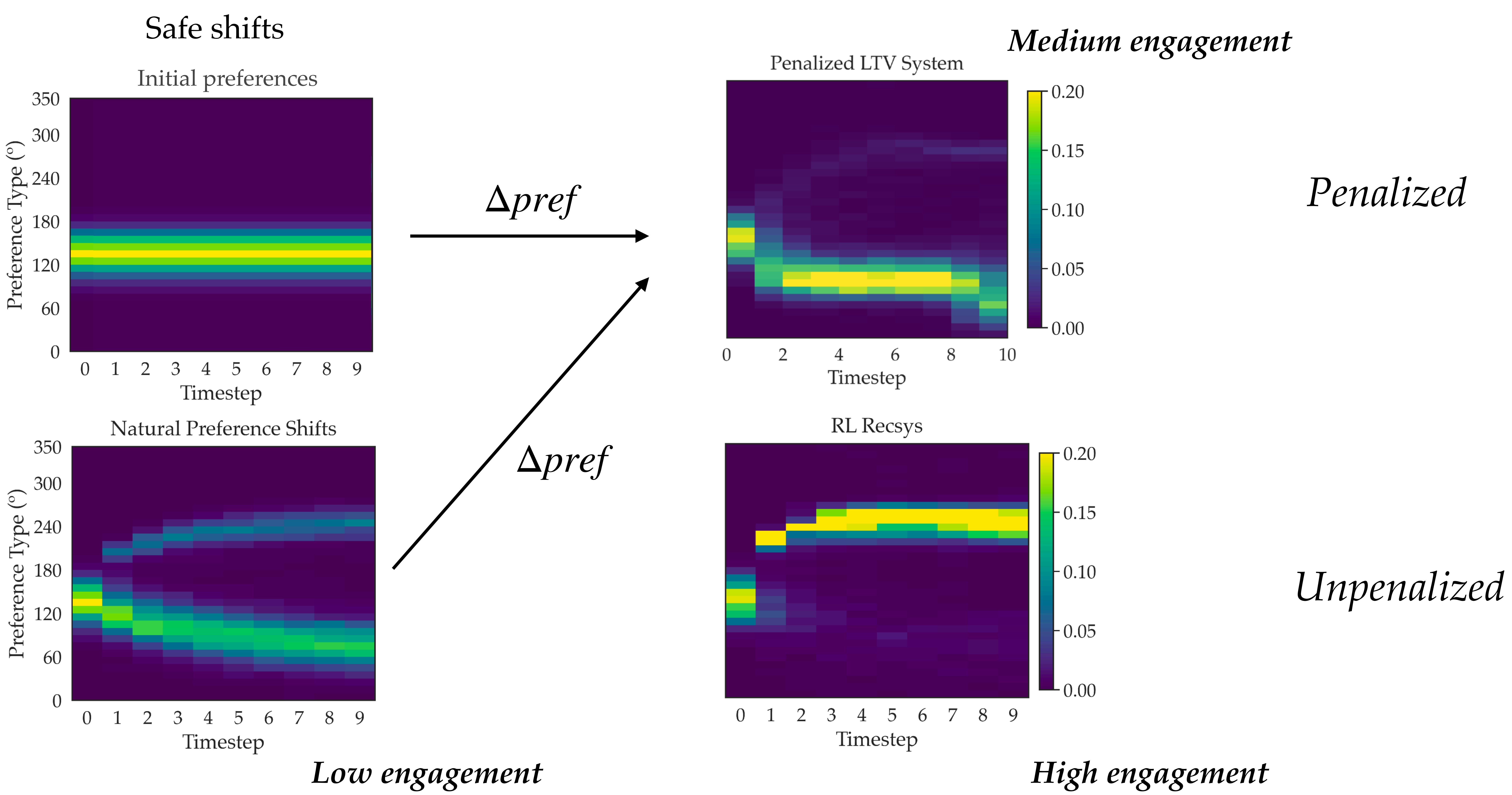
2. Framework for comparing induced shifts to "safe shifts"…



…which can be used to penalize RL training to actively avoid unwanted shifts



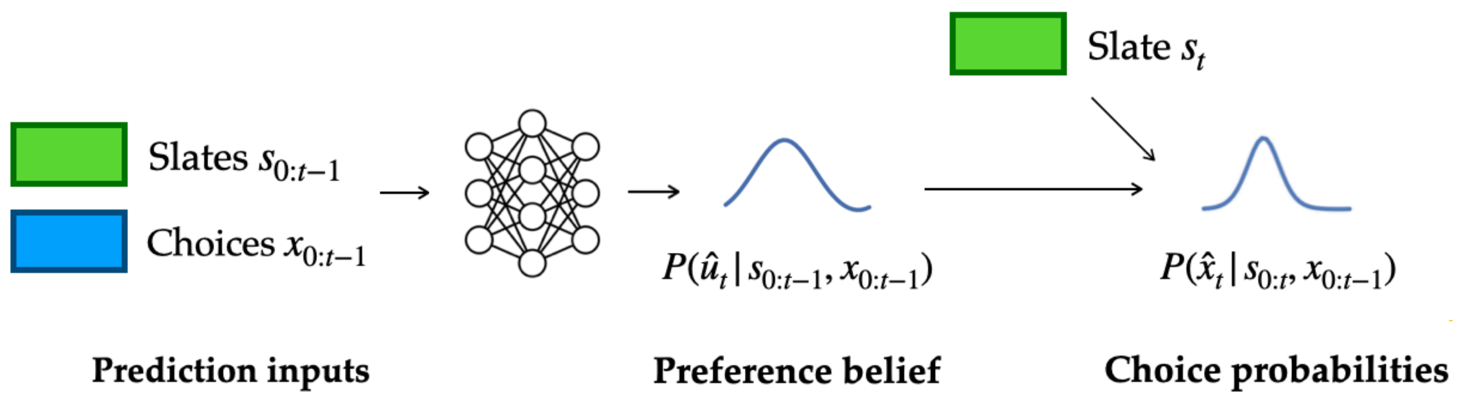
Using proxies to obtain manipulation-penalized RL system



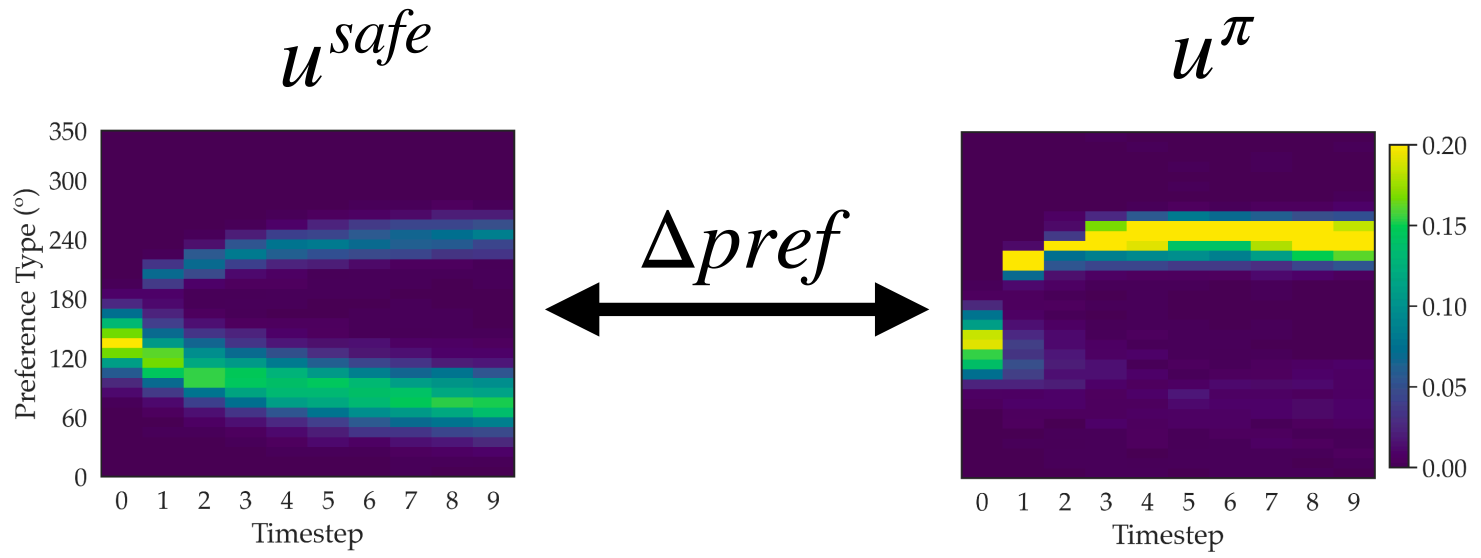
Experiments

0. Setup

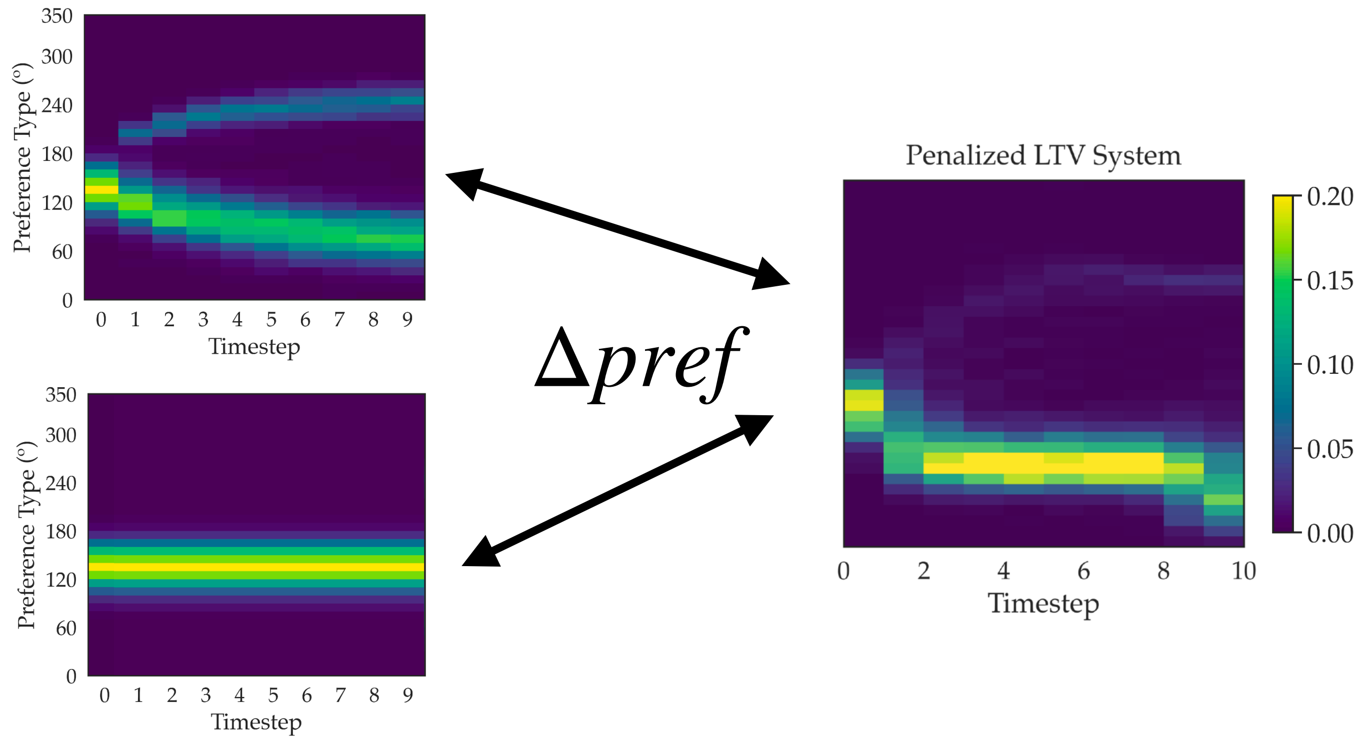
1. Method for estimating policy-induced preference shifts



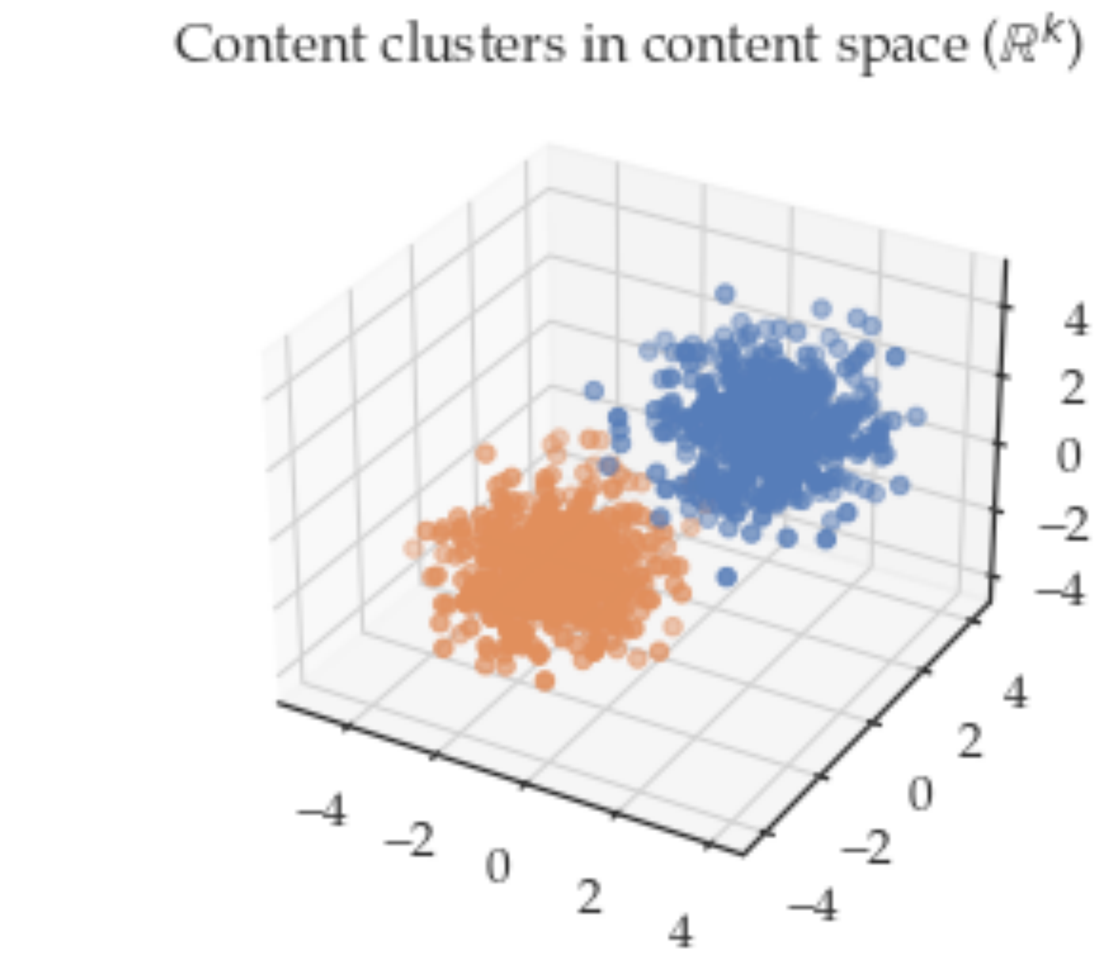
2. Framework for comparing induced shifts to "safe shifts"…

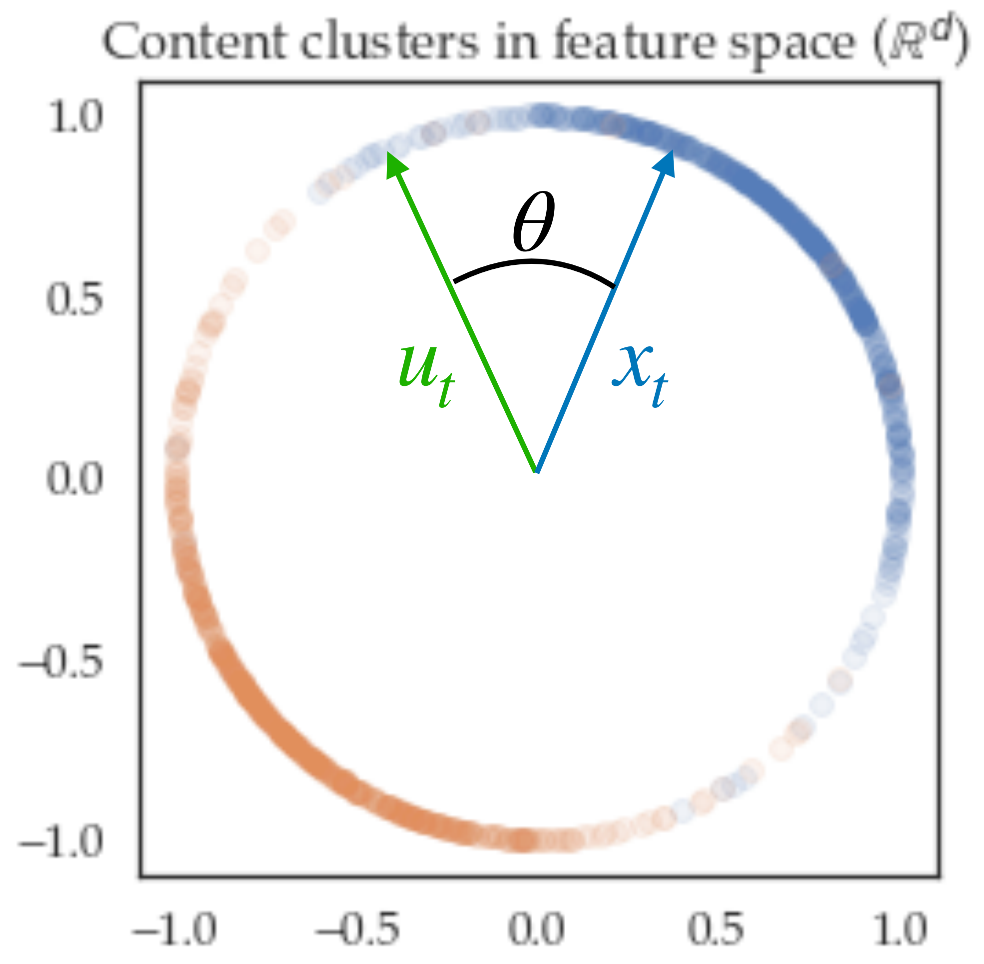


…which can be used to penalize RL training to actively avoid unwanted shifts

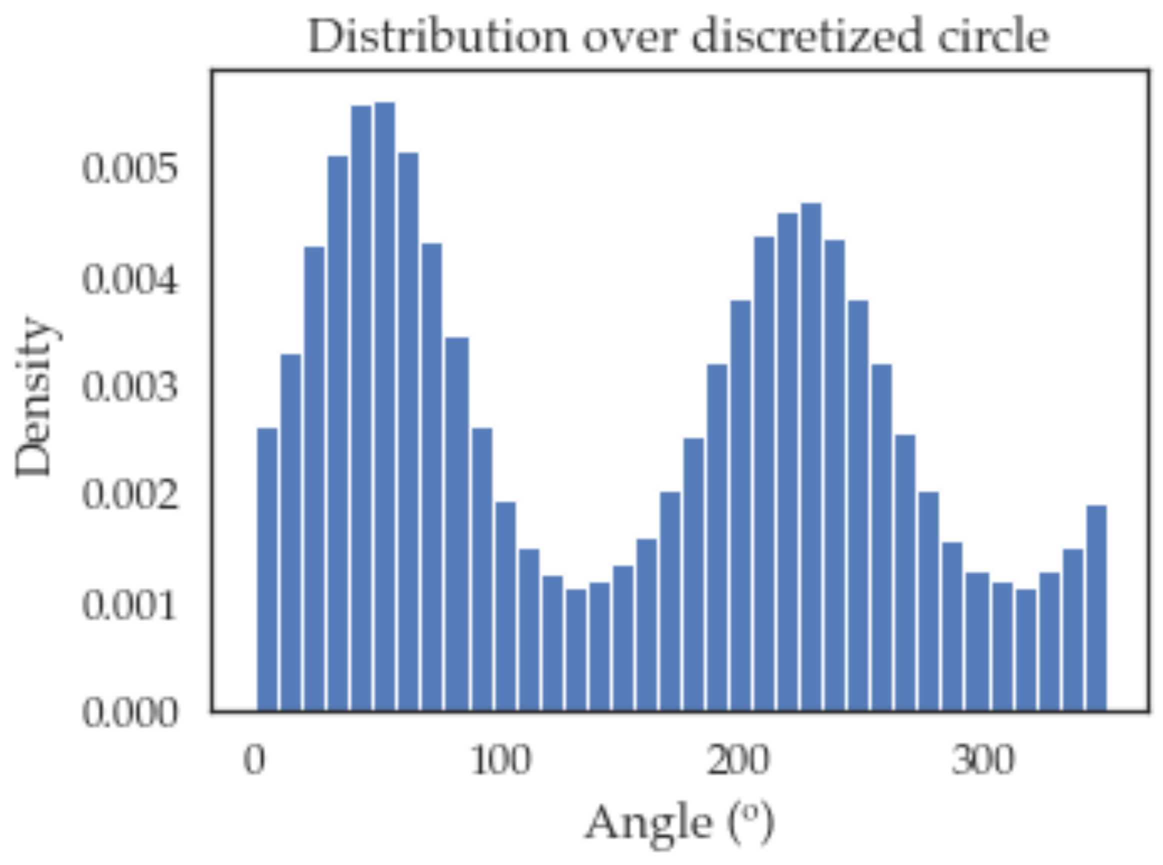


Experimental setup assumptions

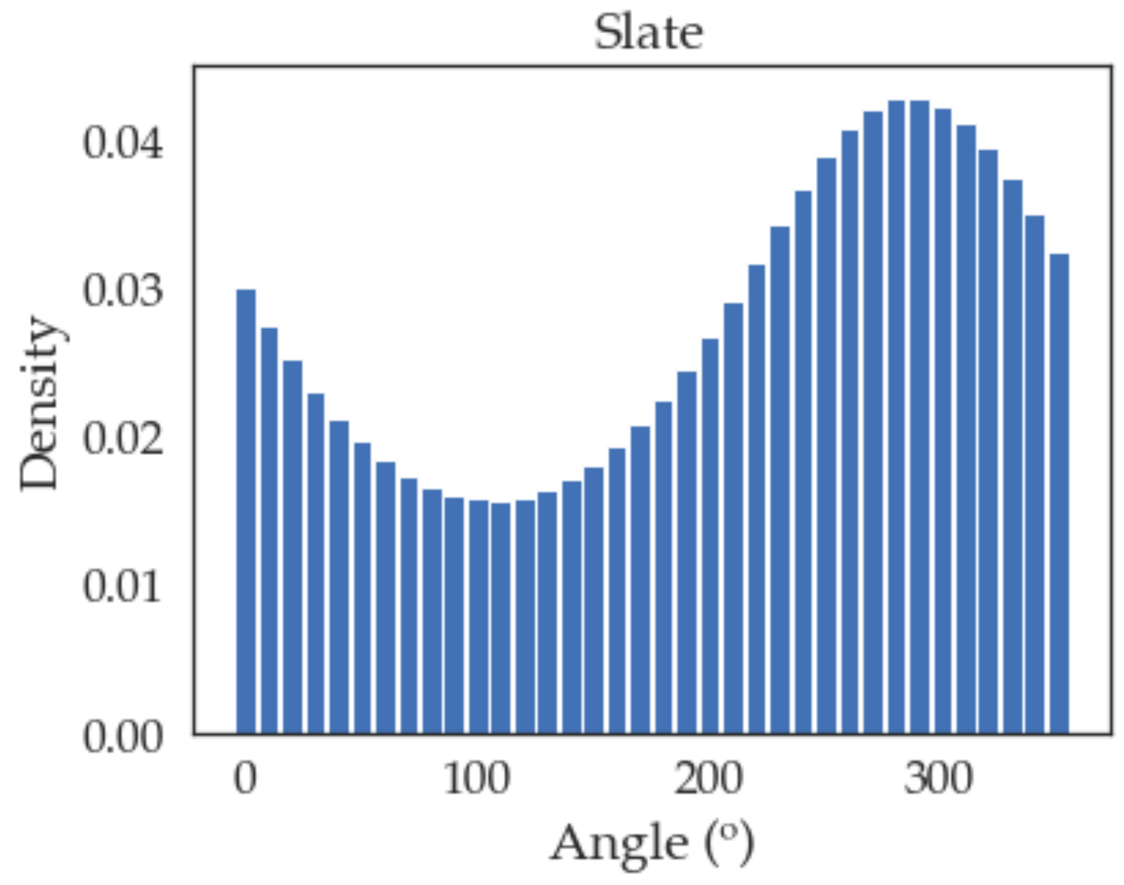




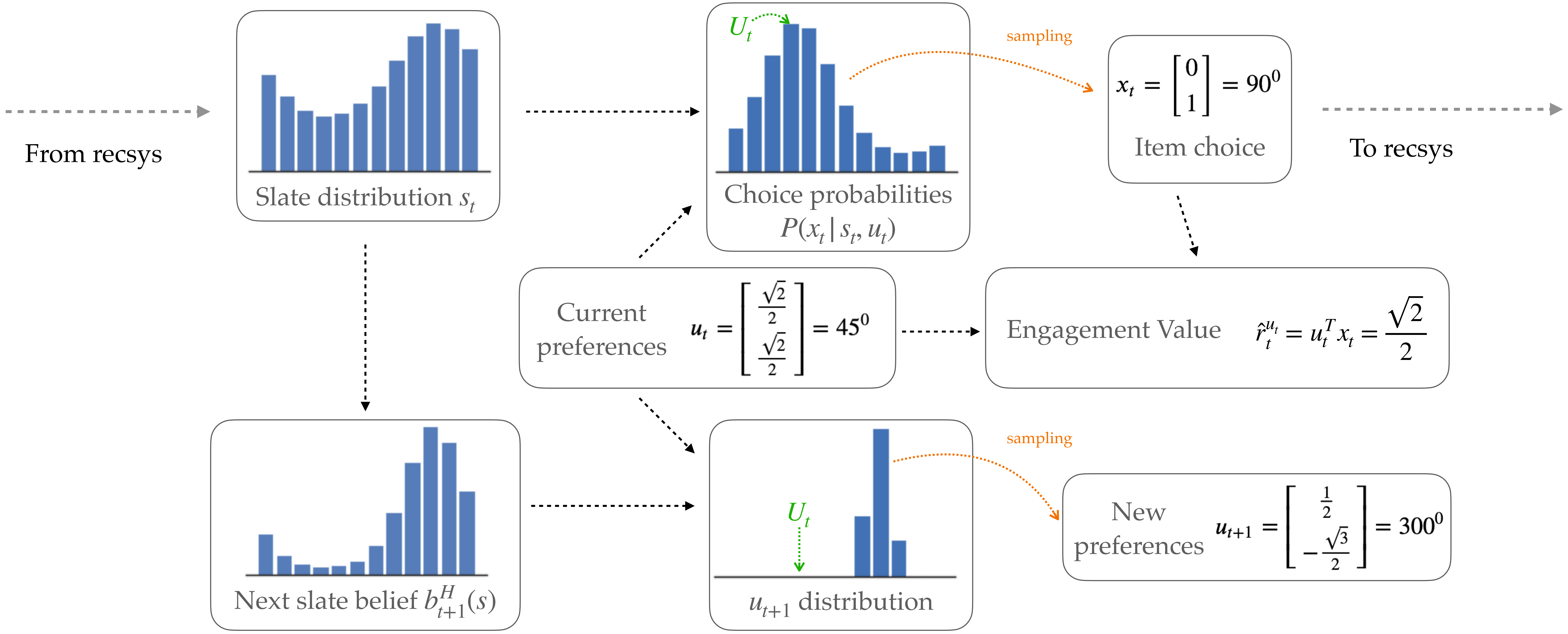
Engagement Value $\hat{r}_t^{u_t} = u_t^T x_t$



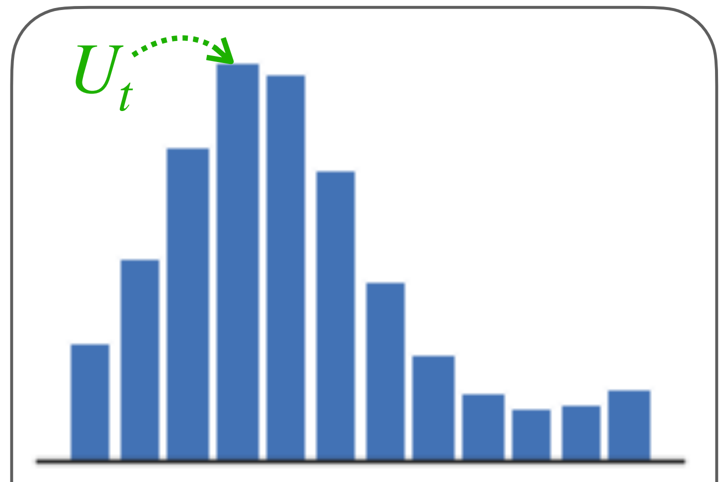
RS
subsampling



Simulated Human Model



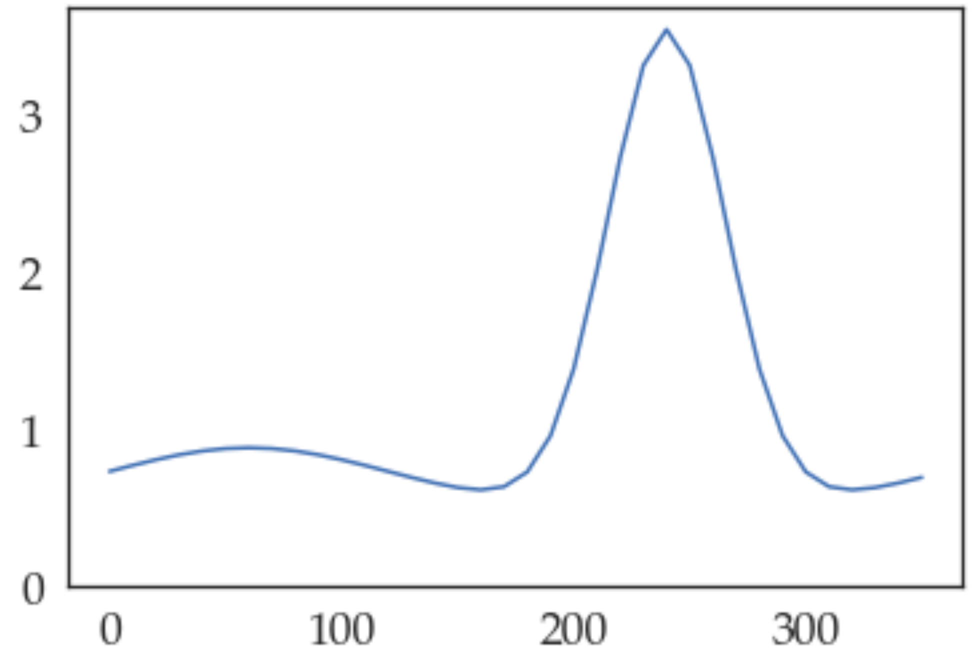
Simulated Human Model

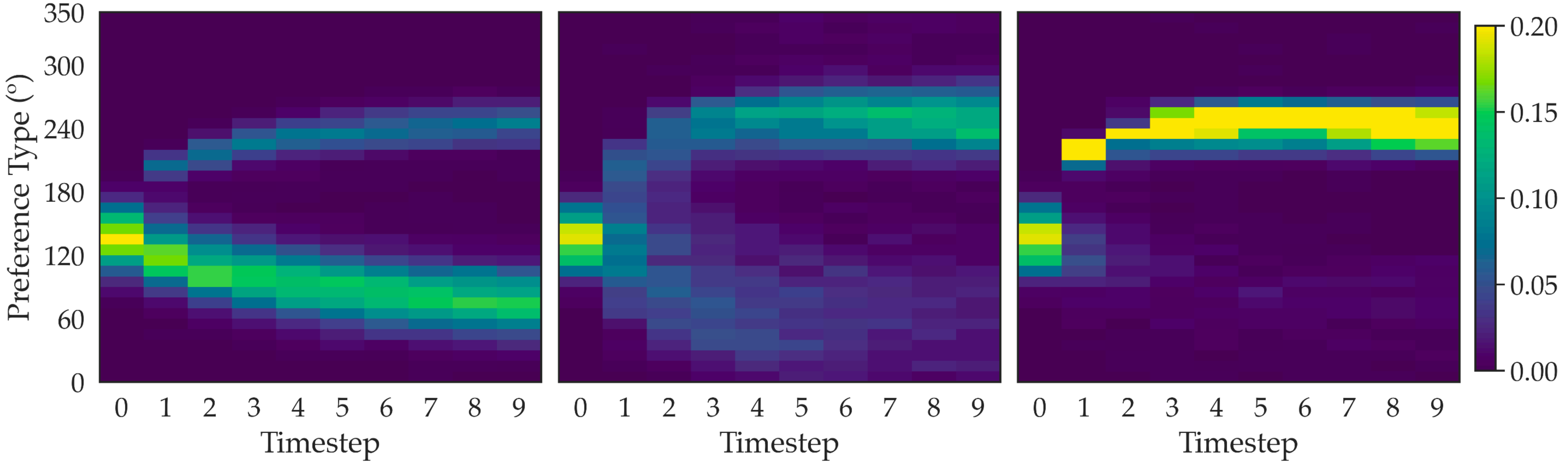


Choice probabilities
$P(x_t \mid s_t, u_t)$

Engagement Value $\hat{r}_t^{u_t} = u_t^T x_t$

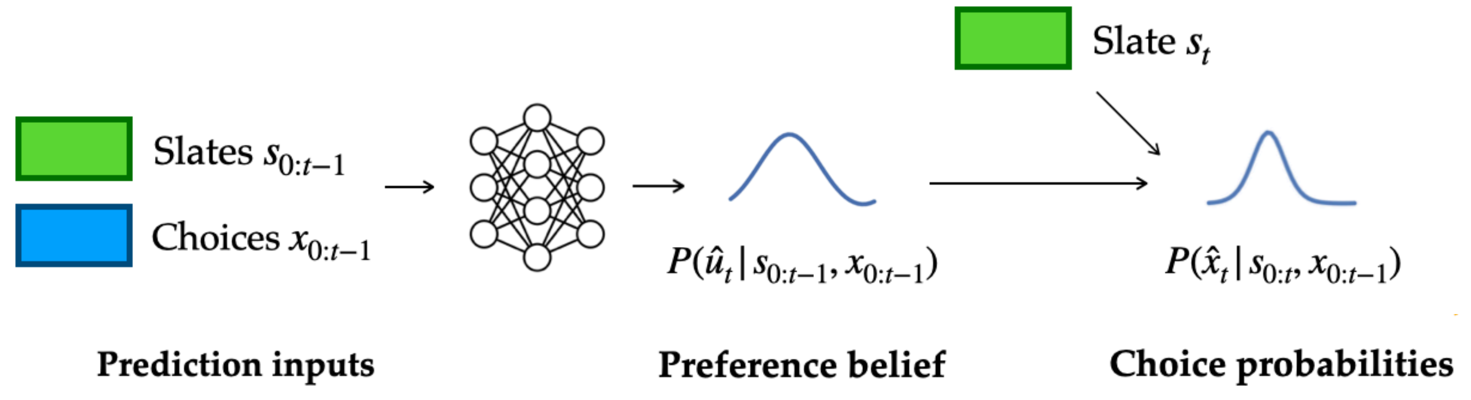
$$P(x_t = x \mid s_t, u_t) = \frac{P(s_t{=}x)\, e^{\boxed{\beta_c} x^T u_t}}{\sum_x^n P(s_t{=}x)\, e^{\beta_c x^T u_t}}$$



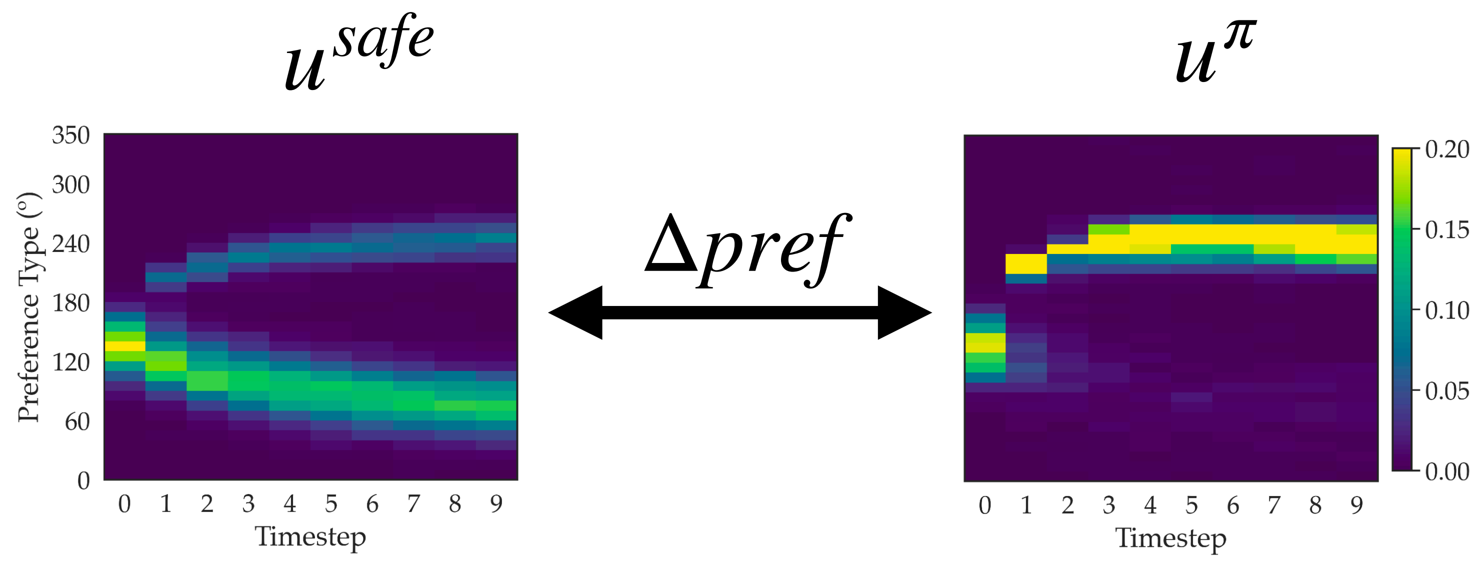


Experiments

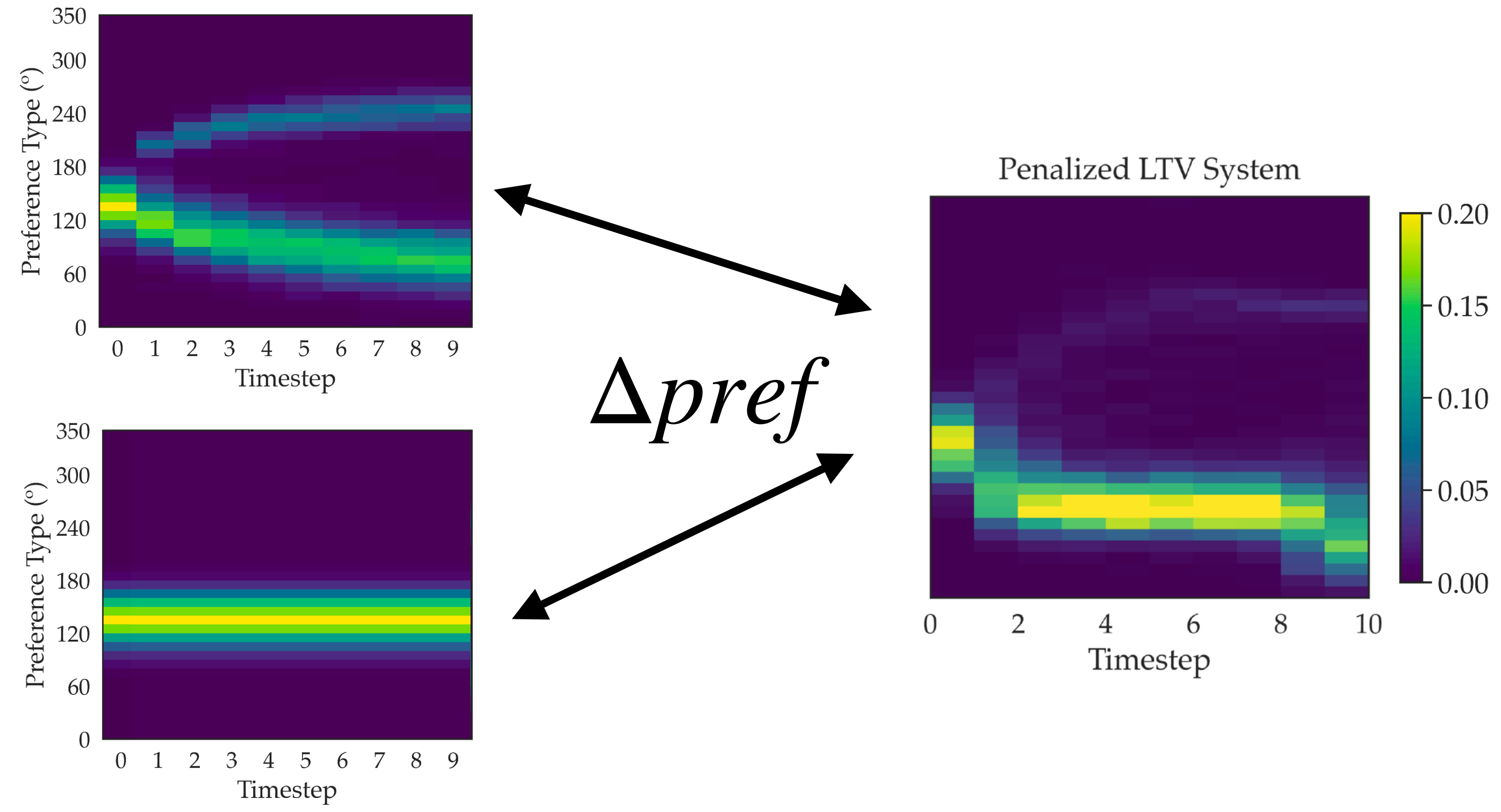
1. Method for estimating policy-induced preference shifts



2. Framework for comparing induced shifts to "safe shifts"…

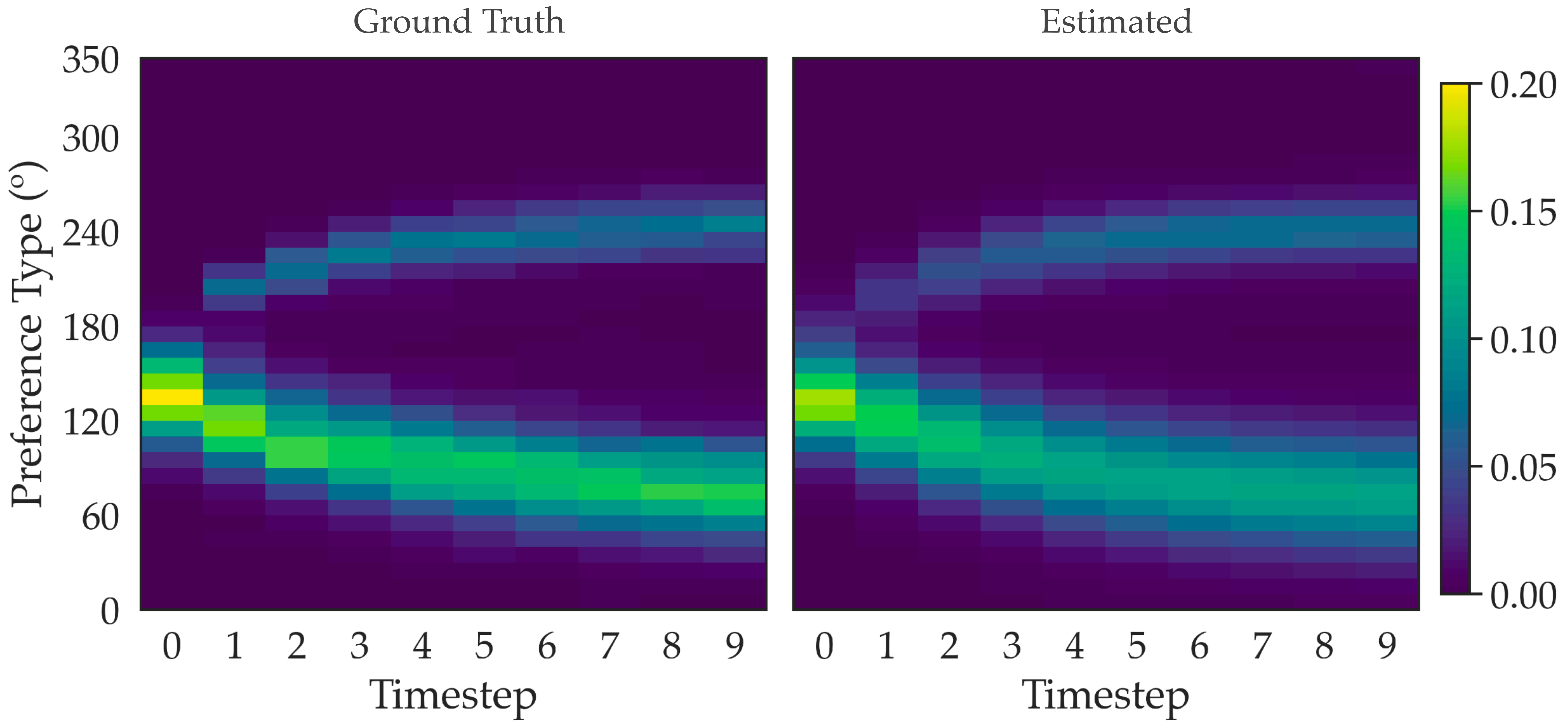


…which can be used to penalize RL training to actively avoid unwanted shifts



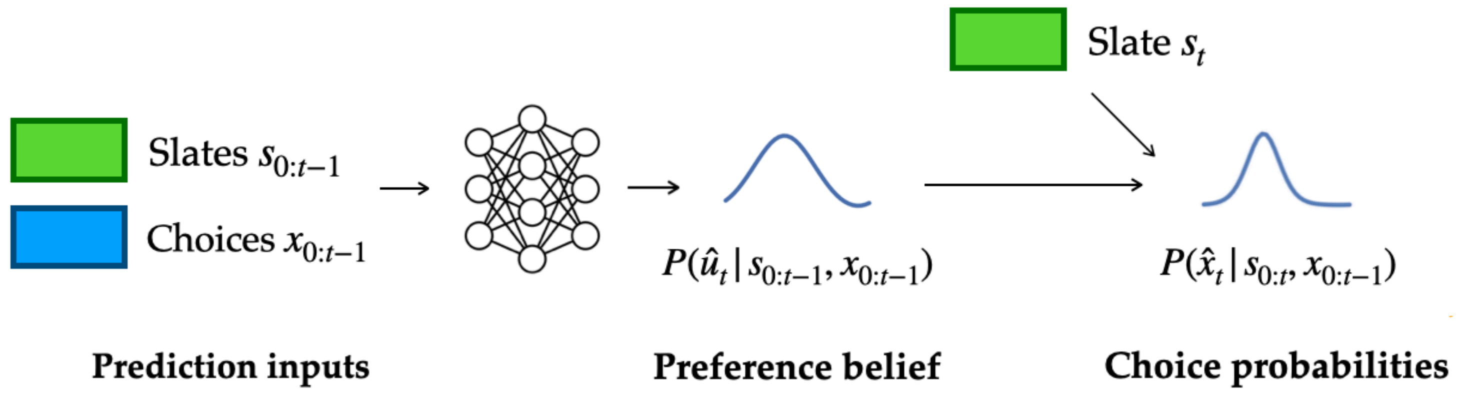
Estimation of policy-induced shifts

Can we predict preference shifts under never-deployed policy π' from historical data?



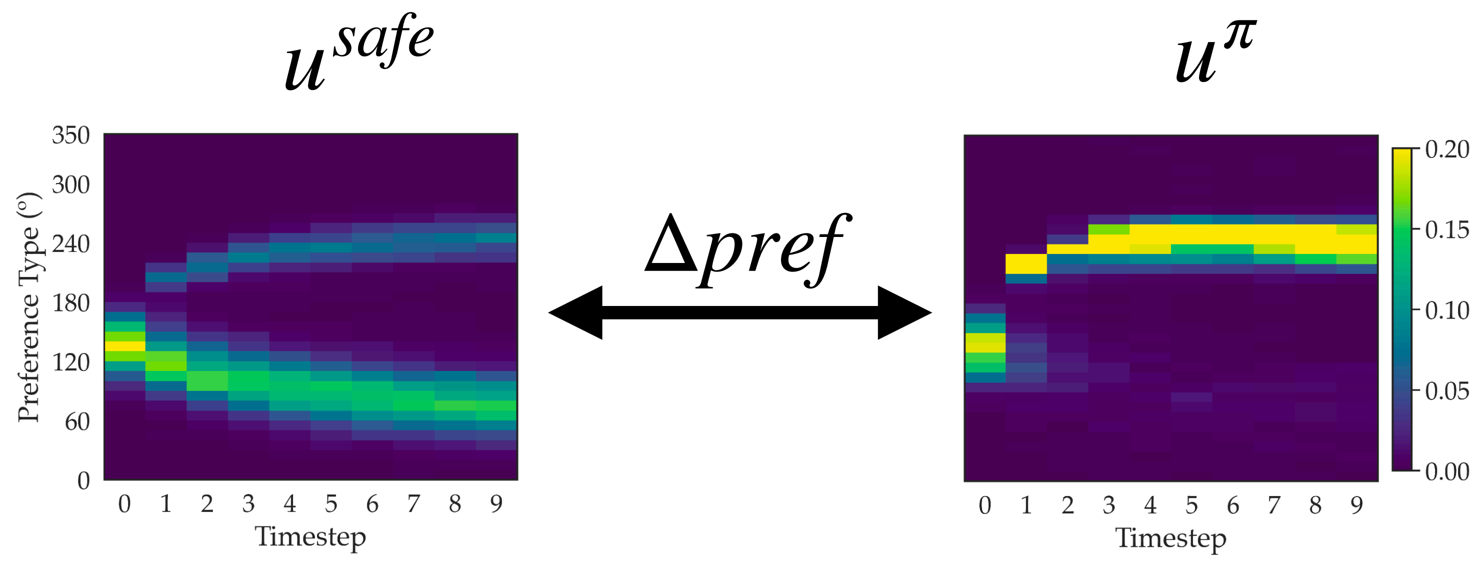
Experiments

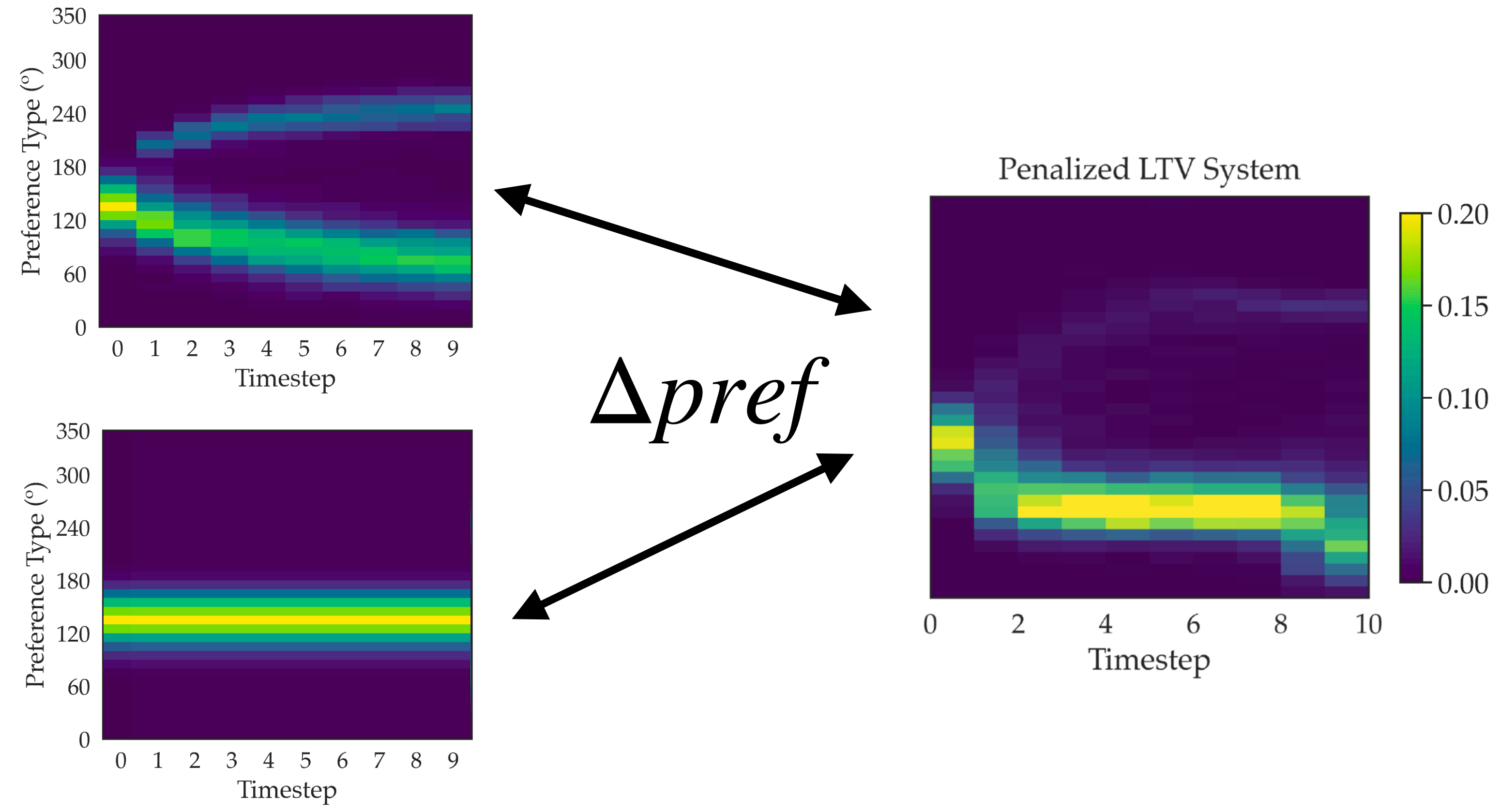
1. Method for estimating policy-induced preference shifts



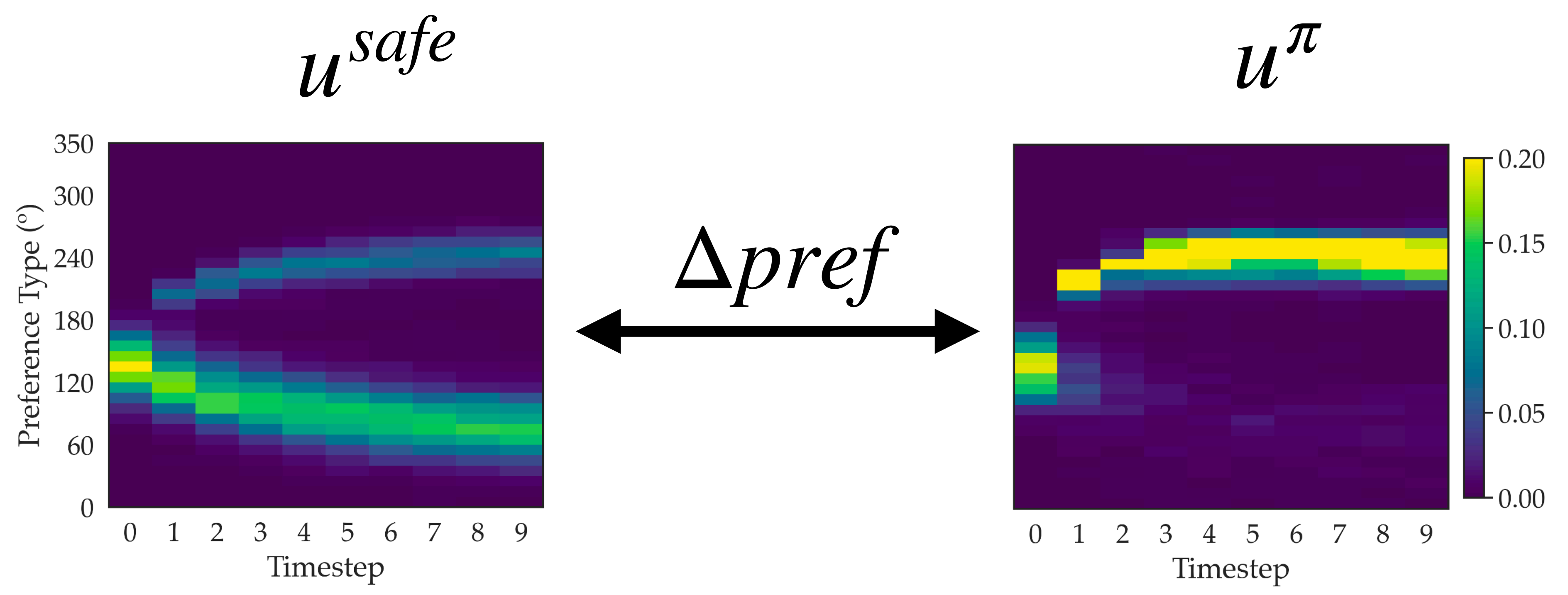
2. Framework for comparing induced shifts to "safe shifts"…

…which can be used to penalize RL training to actively avoid unwanted shifts





Preference distances and RL training



$$D(u^{safe}_{0:T}, u^{\pi}_{0:T}) = \sum_t \mathbb{E}\left[\hat{r}_t^{u^{safe}} - \hat{r}_t^{u^{\pi}}\right]$$

Unpenalized $\displaystyle\sum_t^T \mathbb{E}[\hat{r}_t^{u_t^{\pi}}]$

Metrics **estimated** or **computed with GT dynamics**

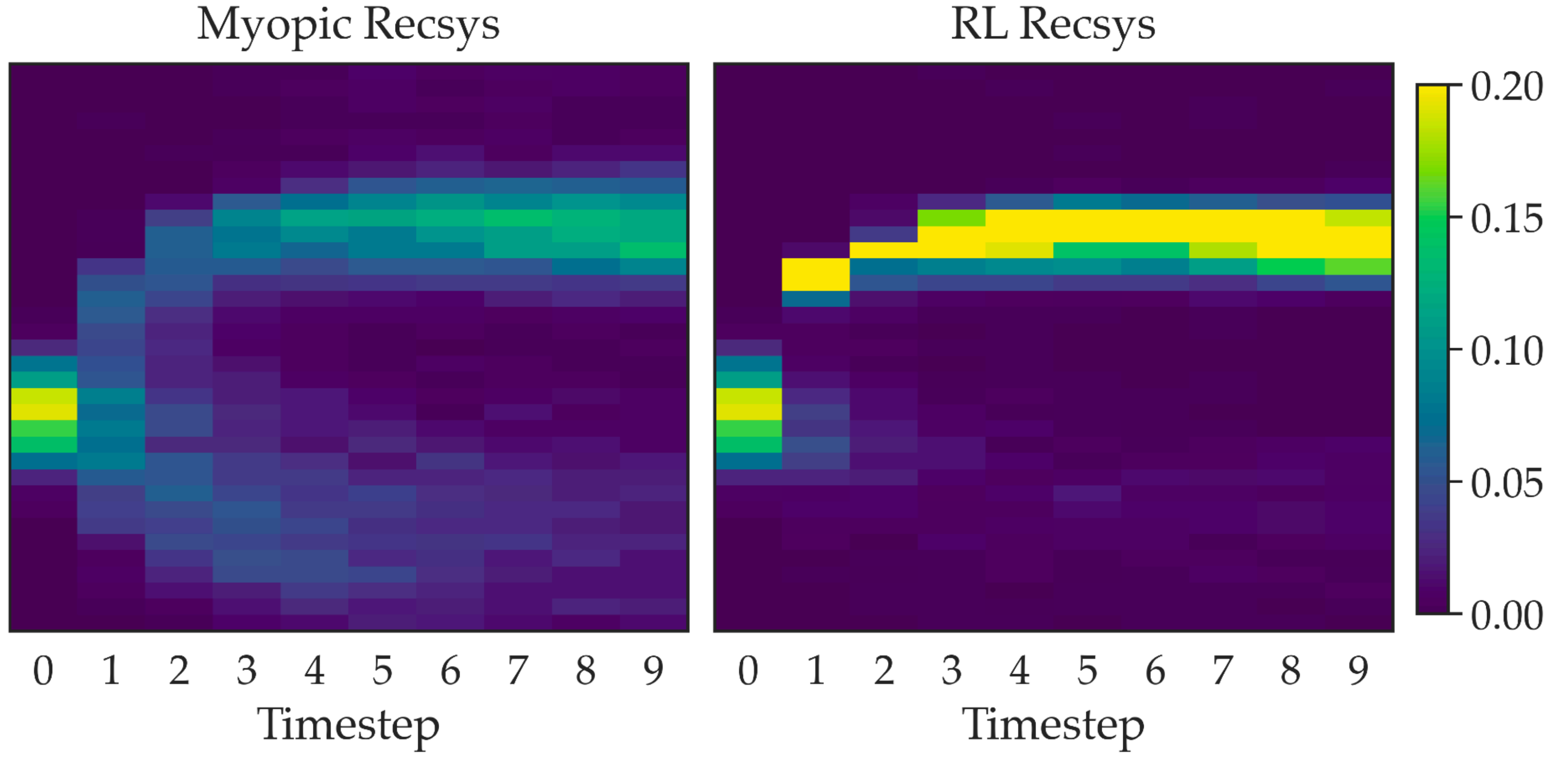
Training is with **simulated** or **on-policy** samples

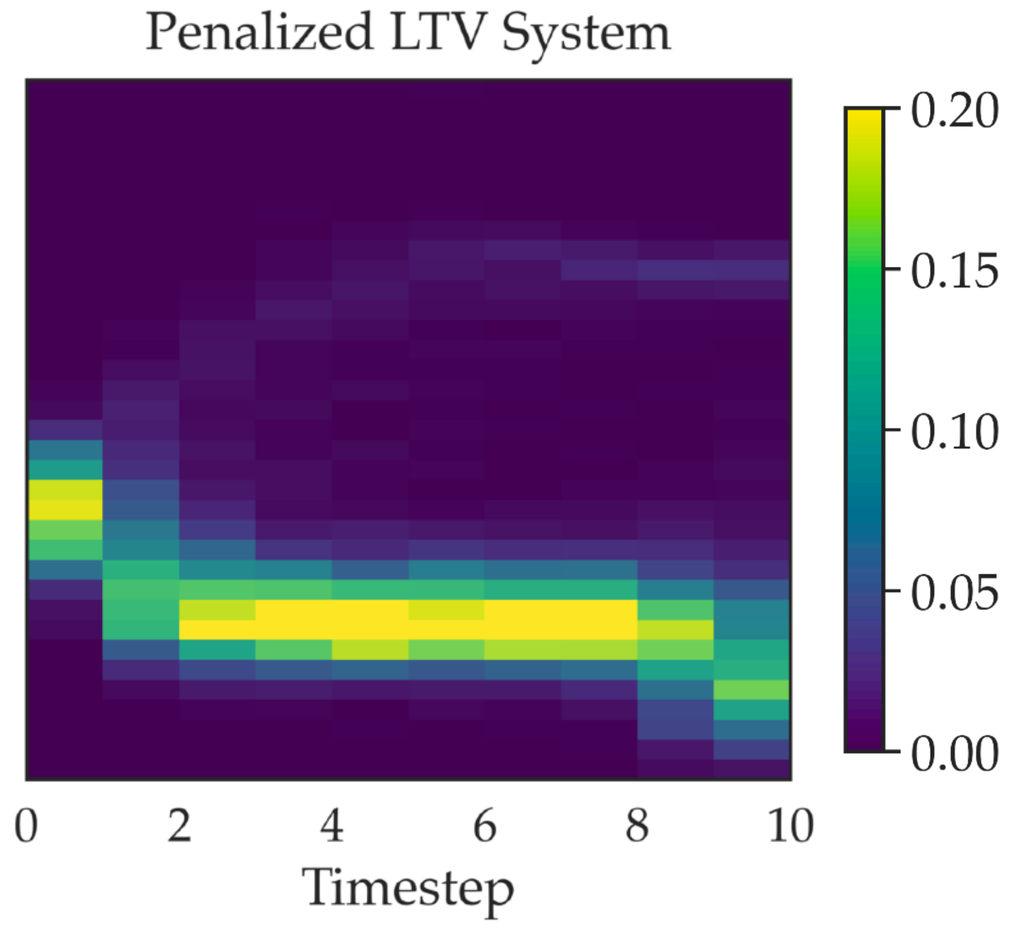
Penalized $\displaystyle\sum_t^T \mathbb{E}[\hat{r}_t^{u_t^{\pi}}] + \lambda_1 \mathbb{E}[\hat{r}_t^{u_t^{*}}] + \lambda_2 \mathbb{E}[\hat{r}_t^{u_0}]$

Results

		Oracle Training			
		Unpenalized		*Penalized*	
		Myopic	RL	Myopic	RL
Oracle Eval	$\hat{r}^{u_t^{\pi}}$	5.71	7.49	6.20	5.28
	$\hat{r}^{u_0}$	1.99	-0.08	3.61	6.21
	$\hat{r}^{u_t^*}$	2.01	-1.09	3.10	4.57
	Avg.	3.23	2.11	4.30	5.35

		Oracle Training		Training in Simulation	
		Unpen.	Penal.	Unpen.	Penal.
Oracle Eval	$\hat{r}^{u_t^{\pi}}$	7.49	5.28	6.40	5.48
	$\hat{r}^{u_0}$	-0.08	6.21	-1.24	5.61
	$\hat{r}^{u_t^*}$	-1.09	4.57	-1.83	4.43
	Avg.	2.11	5.35	1.12	5.84
Estimated E.	$\hat{r}^{u_t^{\pi}}$	5.58	5.42	6.49	5.78
	$\hat{r}^{u_0}$	1.28	5.57	-0.80	4.94
	$\hat{r}^{u_t^*}$	2.05	3.88	1.48	4.41
	Avg.	2.97	4.95	2.39	5.05





Key takeaway

in order to ethically use recommender systems at scale, we may need to take active steps to ***measure*** *and* ***penalize*** *how such systems shift users' internal states*

Thank you!