Estimating and Penalizing Induced Preference Shifts in Recommender Systems

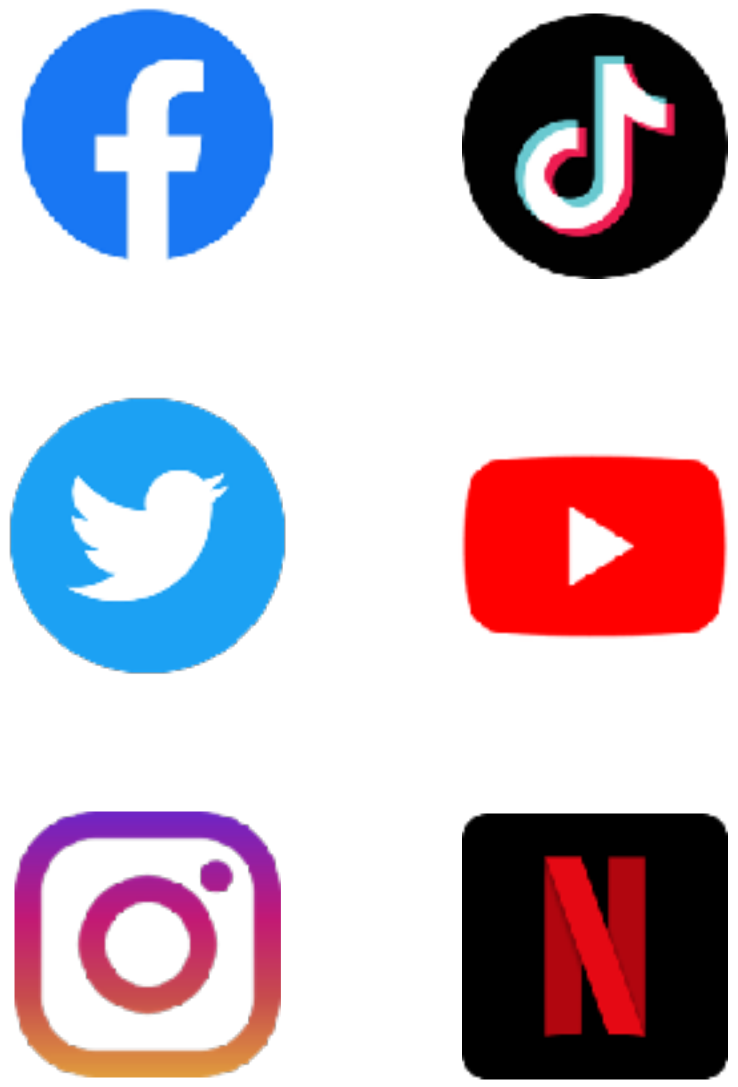
Micah Carroll, Anca Dragan, Stuart Russell, Dylan Hadfield-Menell

Berkeley
UNIVERSITY OF CALIFORNIA

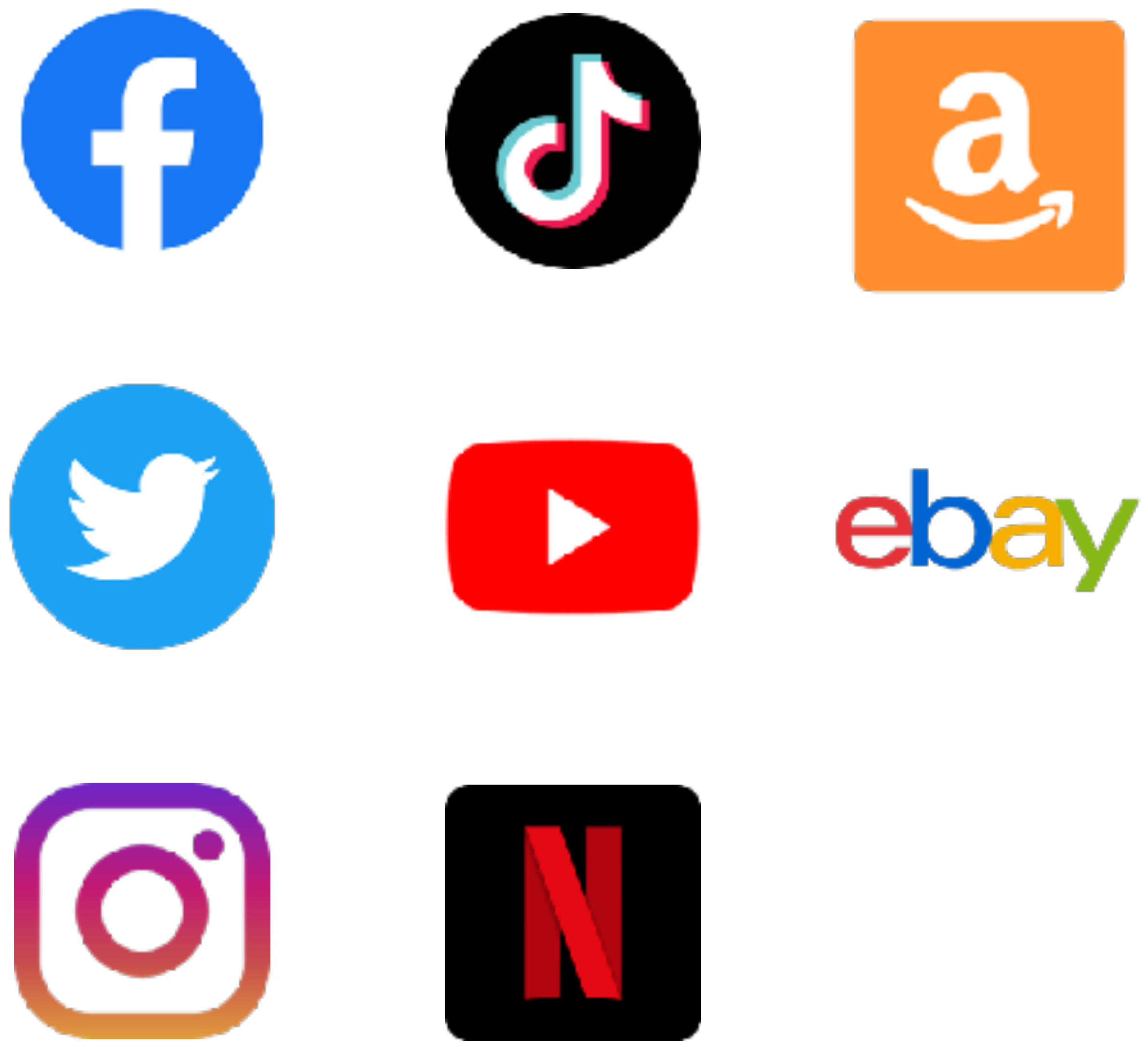
Massachusetts
Institute of
Technology

statista.com/statistics/433871/daily-social-media-usage-worldwide/

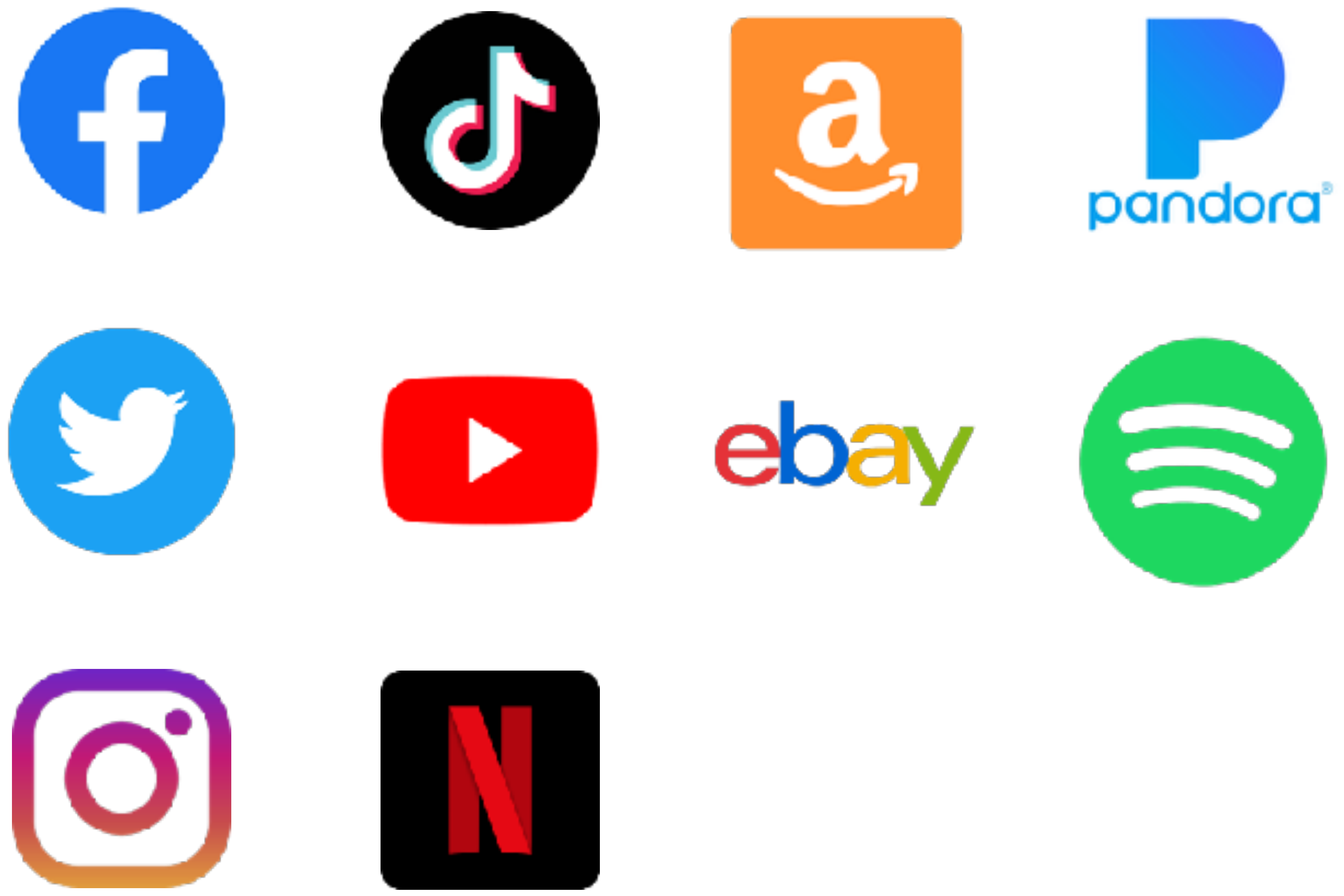
statista.com/statistics/433871/daily-social-media-usage-worldwide/



statista.com/statistics/433871/daily-social-media-usage-worldwide/



statista.com/statistics/433871/daily-social-media-usage-worldwide/

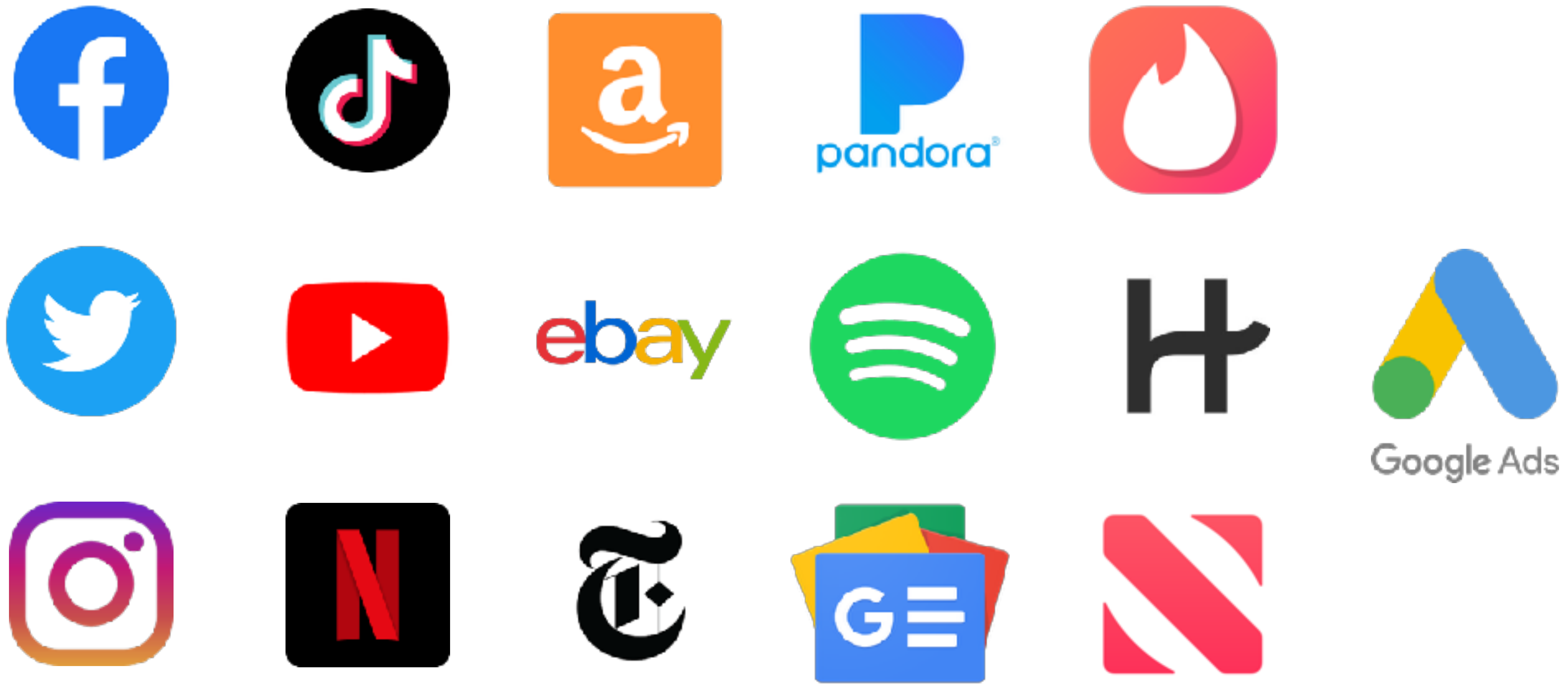


statista.com/statistics/433871/daily-social-media-usage-worldwide/

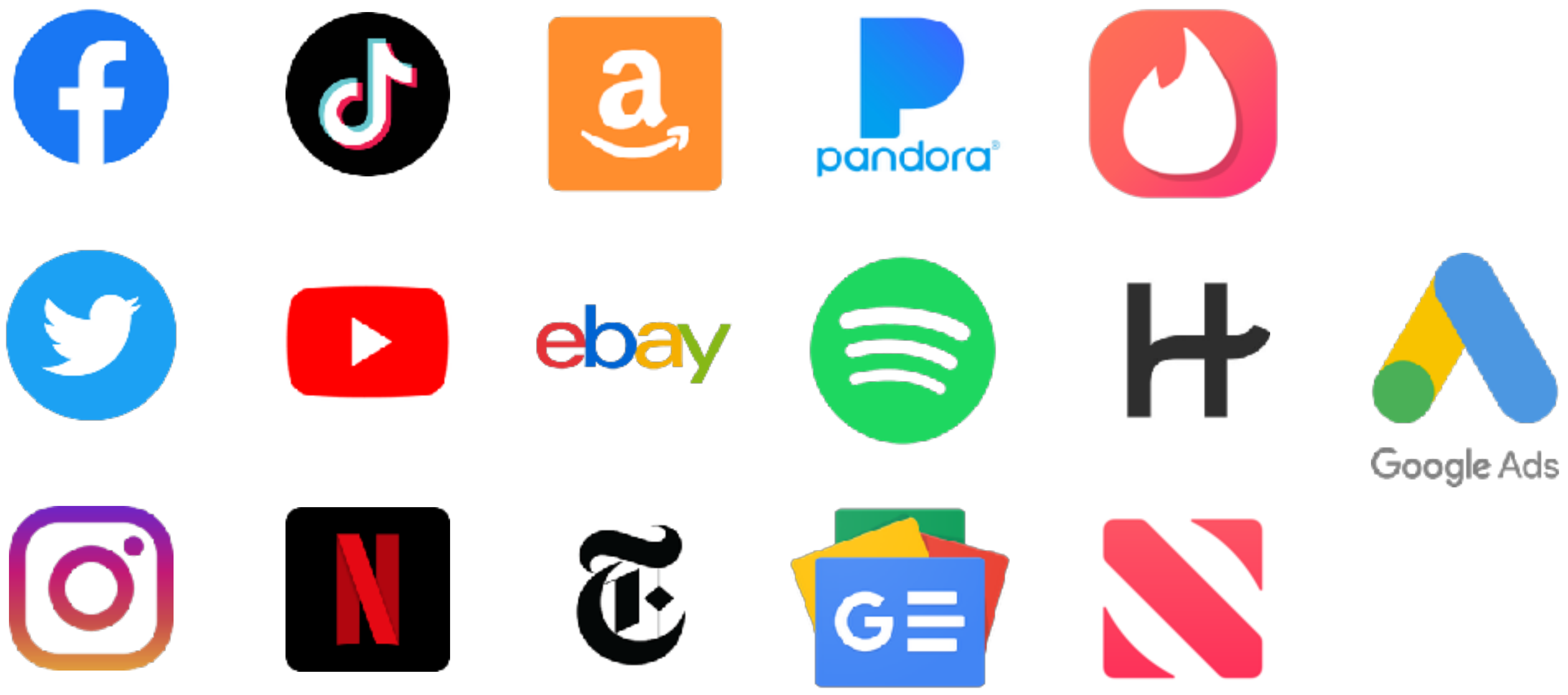
statista.com/statistics/433871/daily-social-media-usage-worldwide/

pandora
ebay

statista.com/statistics/433871/daily-social-media-usage-worldwide/

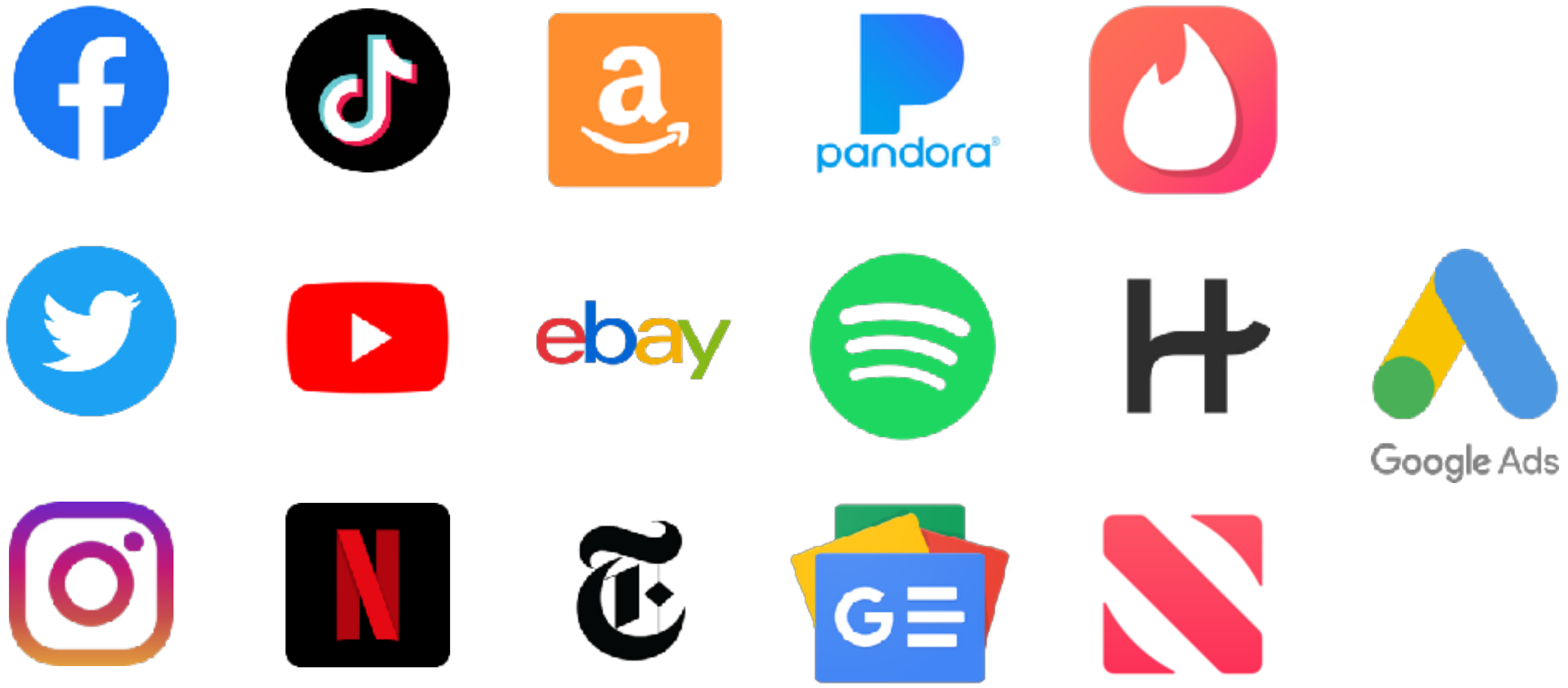


statista.com/statistics/433871/daily-social-media-usage-worldwide/



User preferences change

statista.com/statistics/433871/daily-social-media-usage-worldwide/



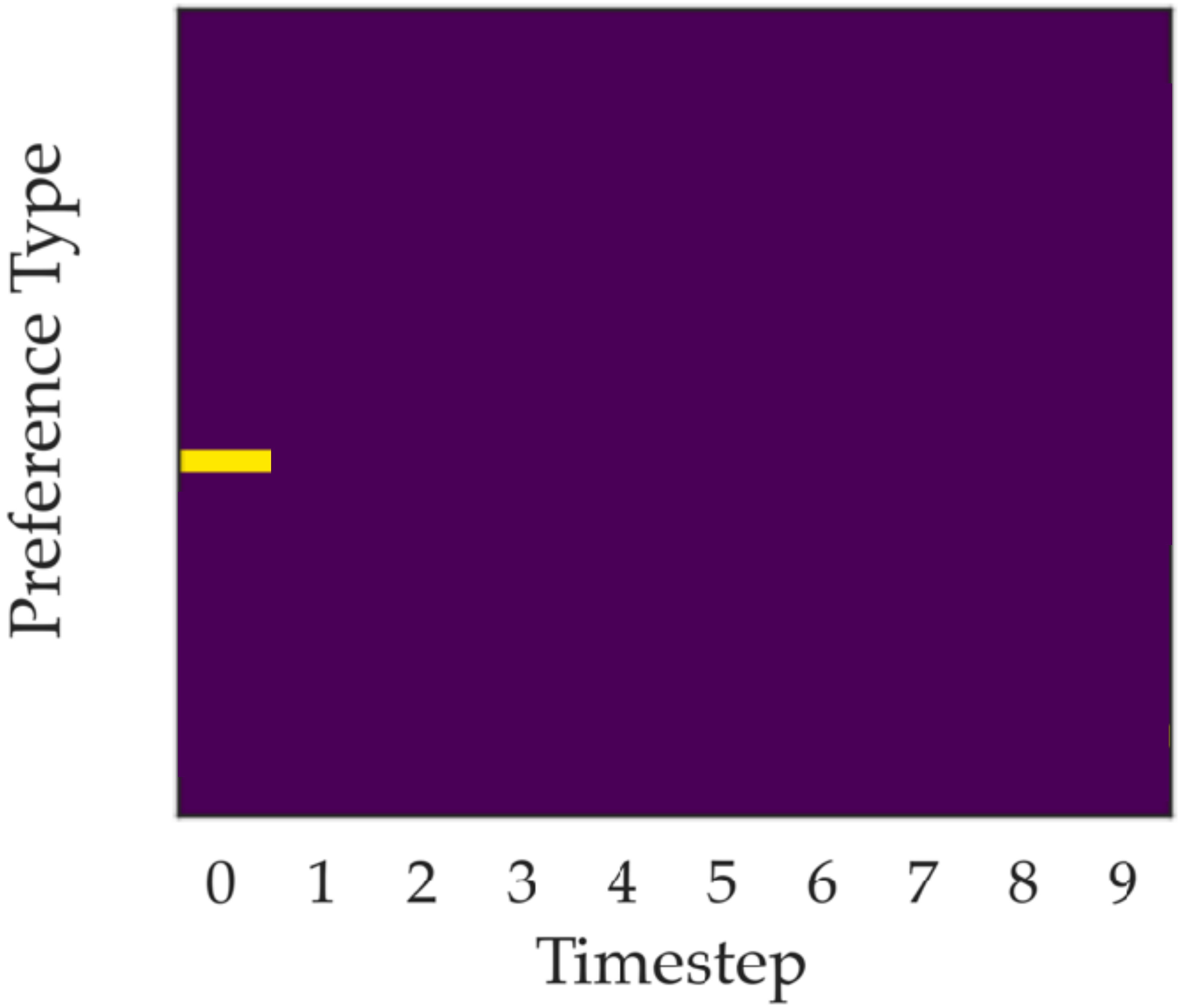
User preferences change, and recommenders will affect them

statista.com/statistics/433871/daily-social-media-usage-worldwide/

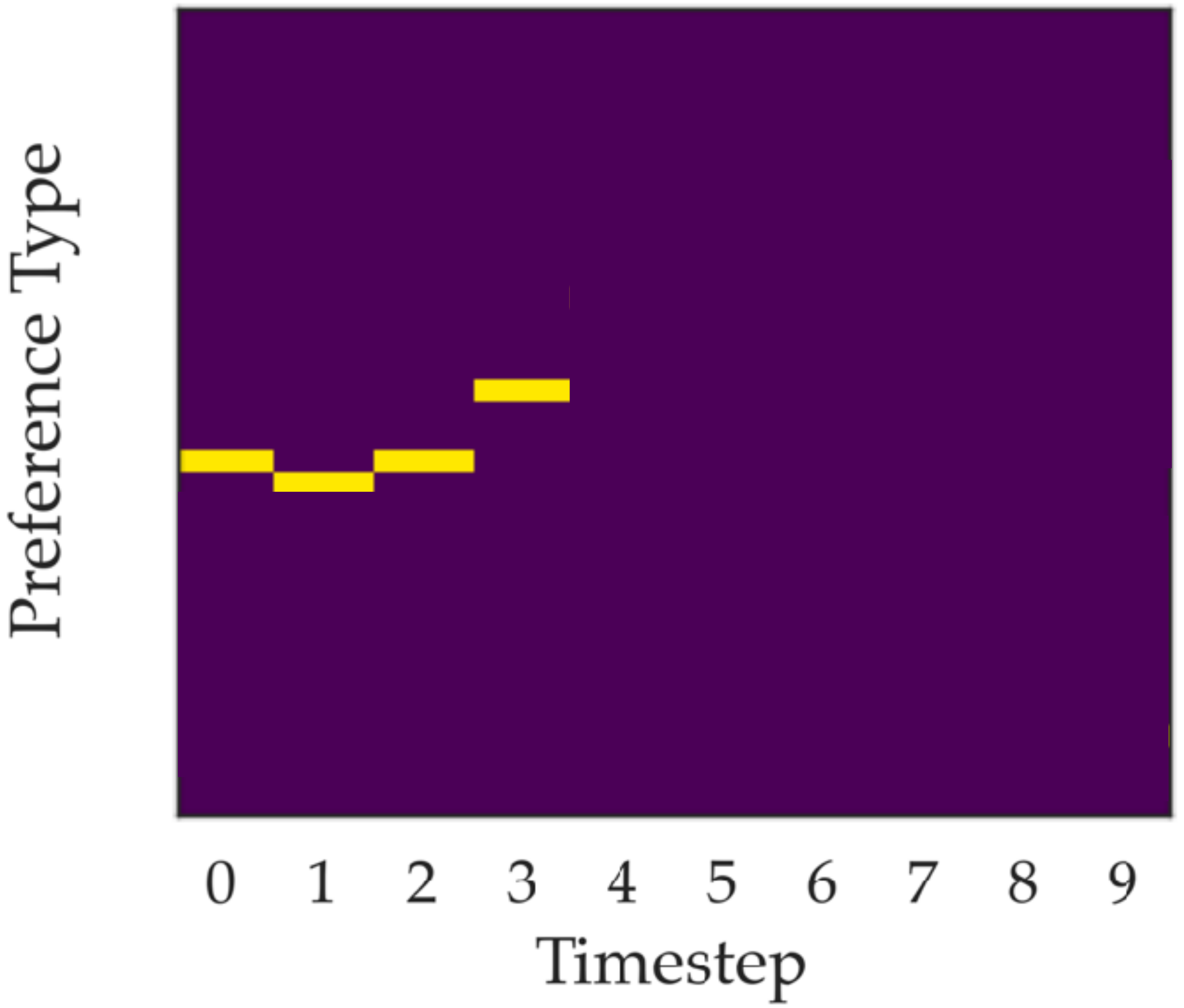
Policy-induced preference shifts

Policy-induced preference shifts

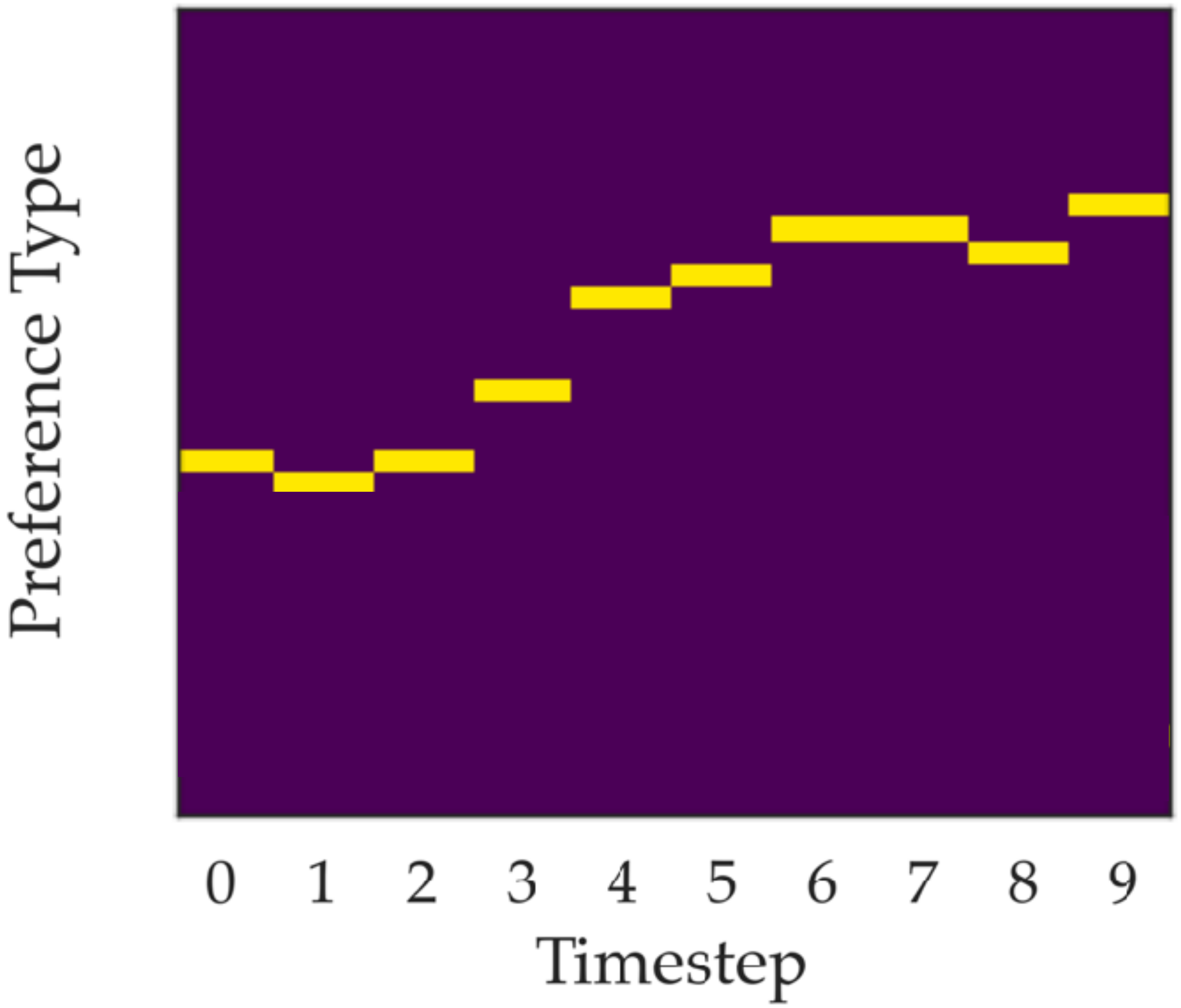
Policy-induced preference shifts



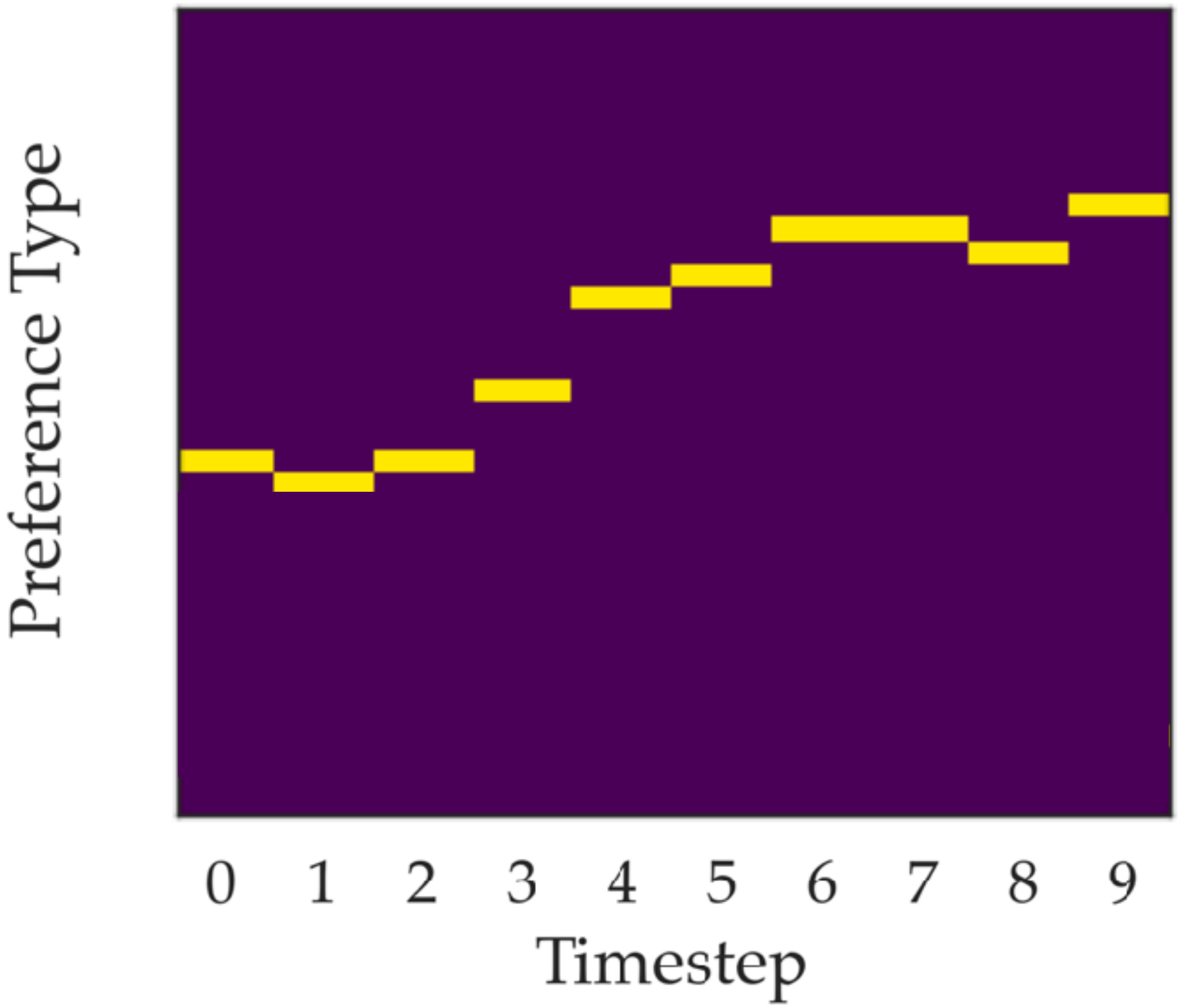
Policy-induced preference shifts



Policy-induced preference shifts



Policy-induced preference shifts

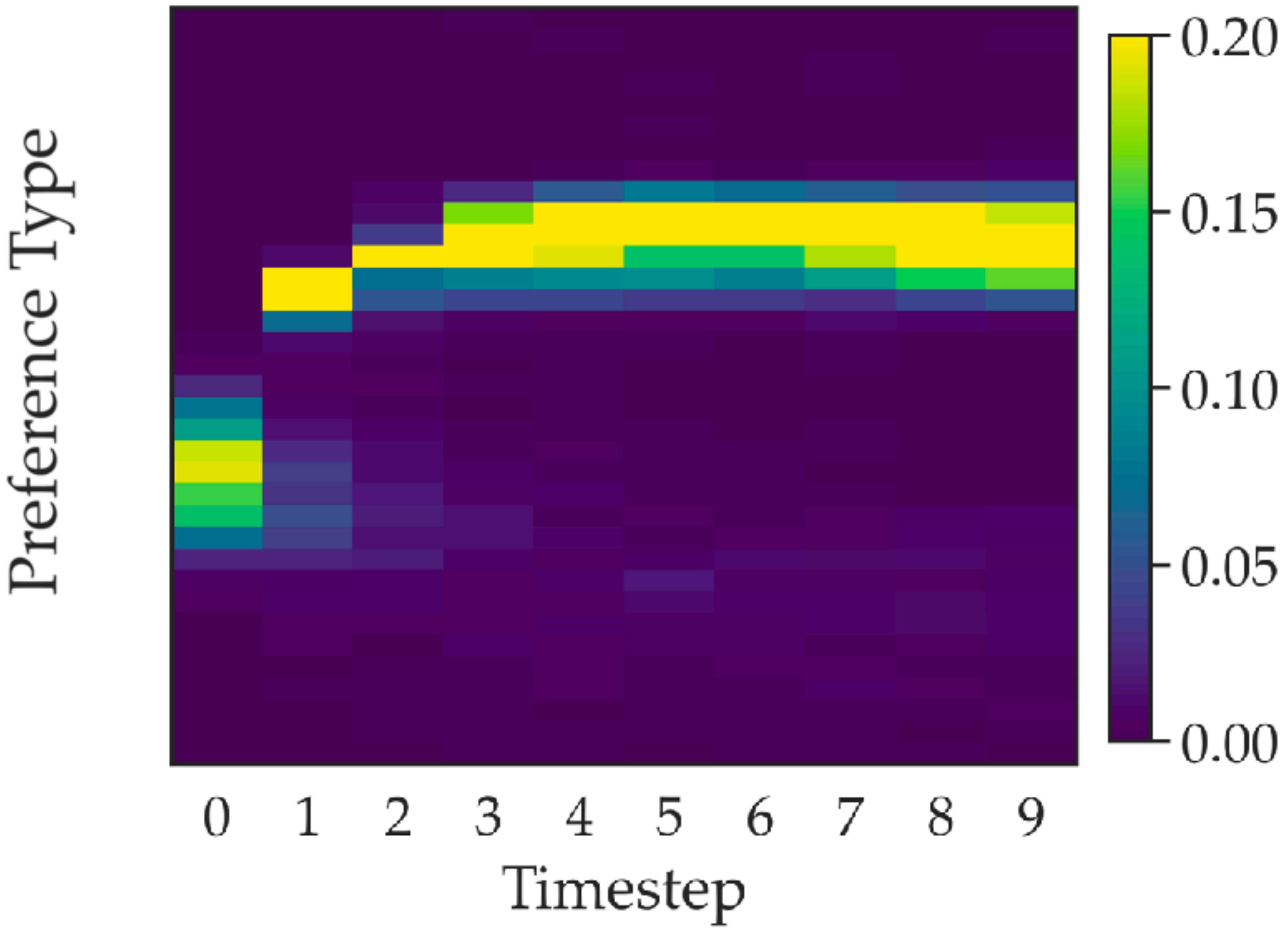


Sensationalized news

↕

Dry journalism

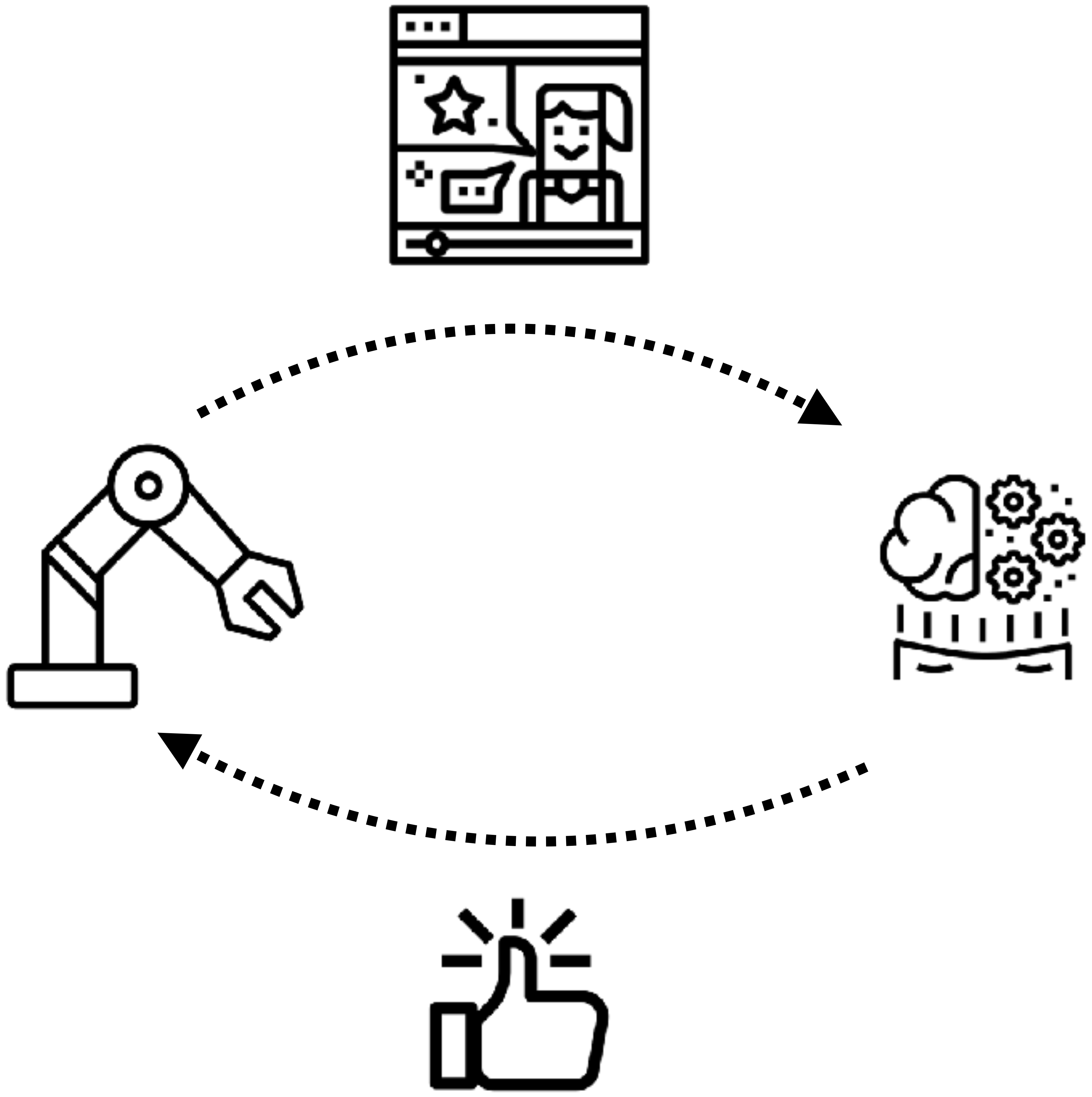
Policy-induced preference shifts



Sensationalized news

Dry journalism

Incentives for user manipulation



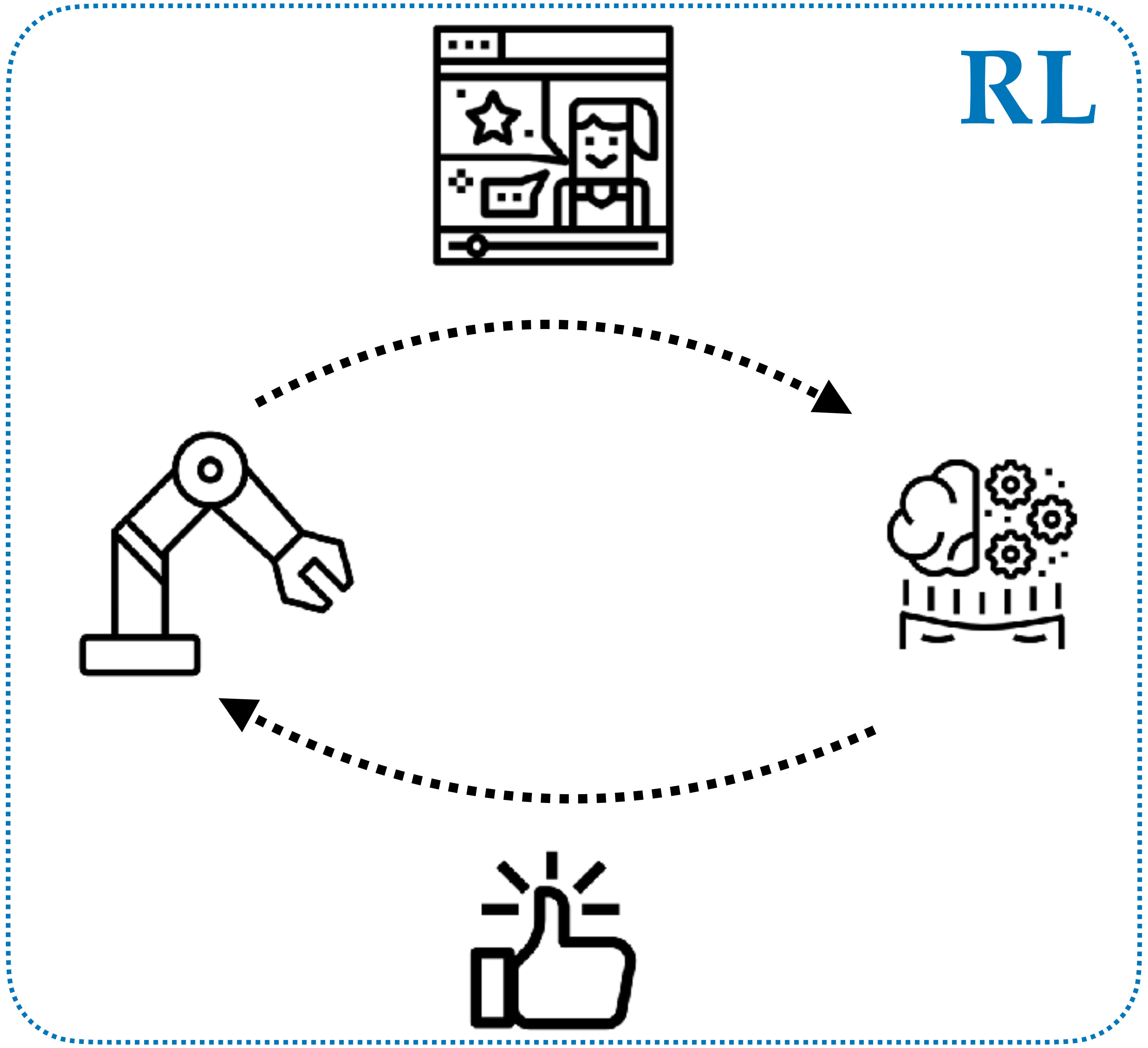
[Krueger et. al, 2020] Hidden Incentives for Auto-Induced Distributional Shift

[Carroll et. al, 2021] Estimating and Penalizing Induced Preference Shifts in Recommender Systems

[Evans et. al, 2021] User Tampering in Reinforcement Learning Recommender Systems

[Farquhar, Carey, Everitt, 2022] Path-Specific Objectives for Safer Agent Incentives

Incentives for user manipulation



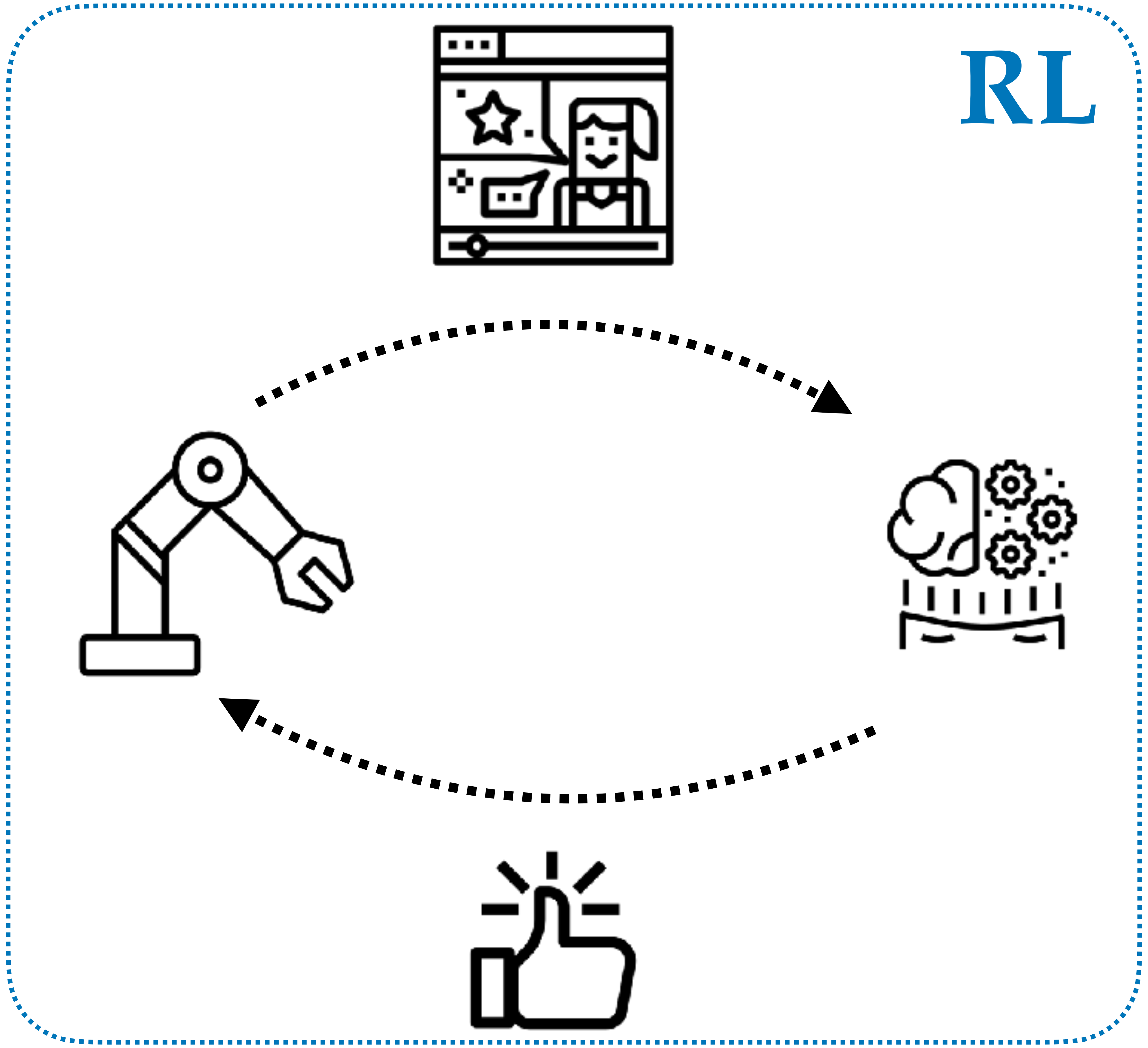
[Krueger et. al, 2020] Hidden Incentives for Auto-Induced Distributional Shift

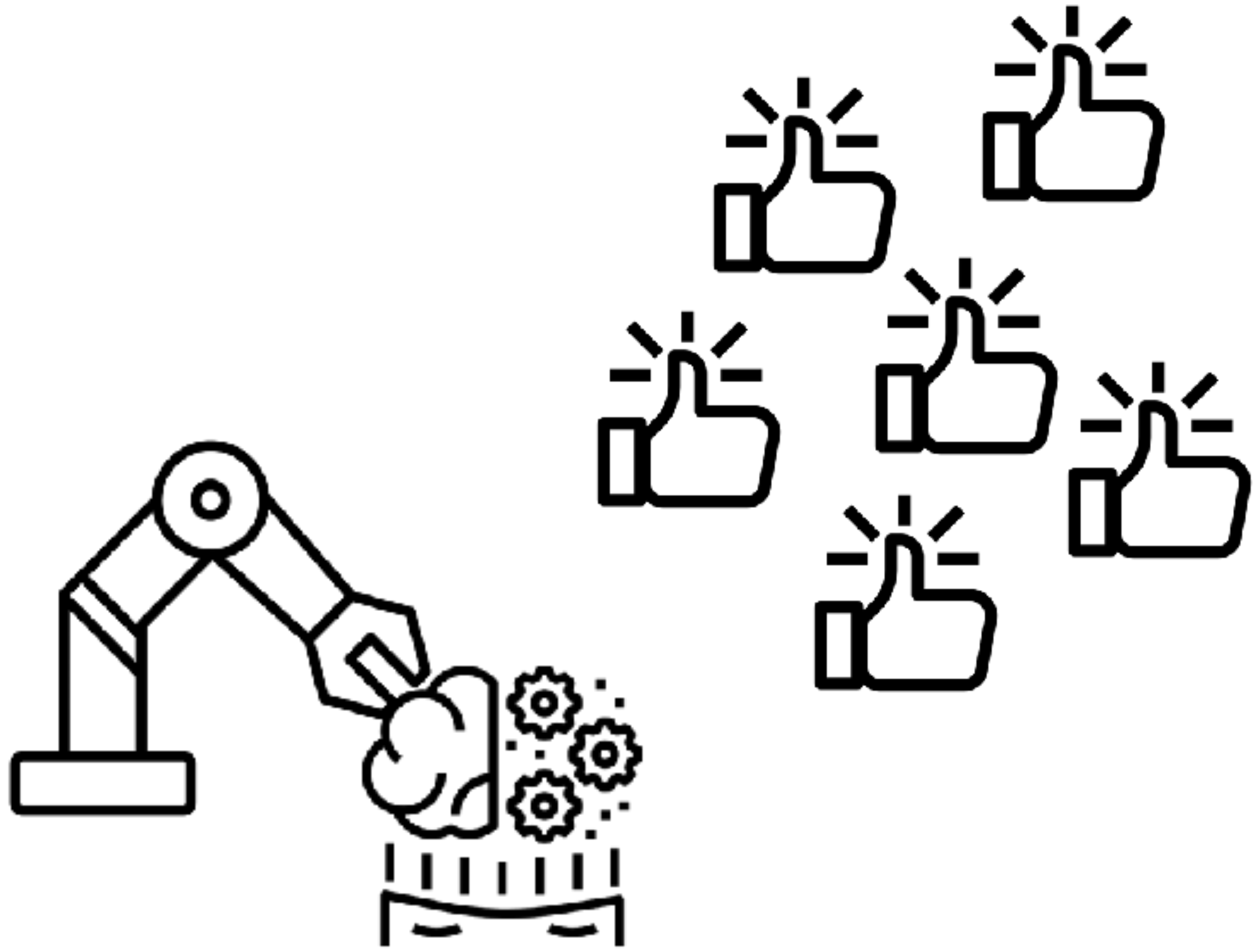
[Carroll et. al, 2021] Estimating and Penalizing Induced Preference Shifts in Recommender Systems

[Evans et. al, 2021] User Tampering in Reinforcement Learning Recommender Systems

[Farquhar, Carey, Everitt, 2022] Path-Specific Objectives for Safer Agent Incentives

Incentives for user manipulation





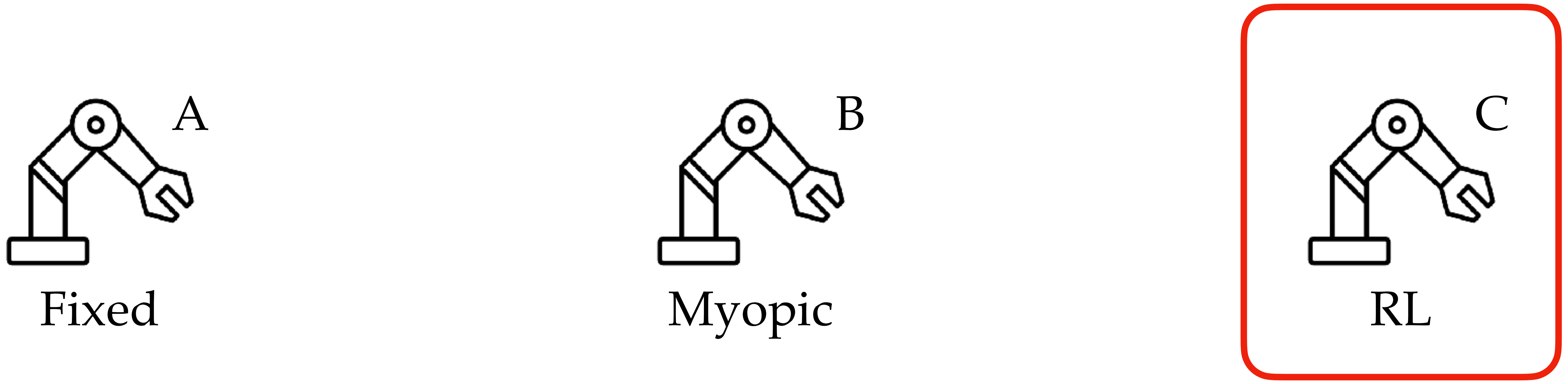
[Krueger et. al, 2020] Hidden Incentives for Auto-Induced Distributional Shift

[Carroll et. al, 2021] Estimating and Penalizing Induced Preference Shifts in Recommender Systems

[Evans et. al, 2021] User Tampering in Reinforcement Learning Recommender Systems

[Farquhar, Carey, Everitt, 2022] Path-Specific Objectives for Safer Agent Incentives

Policy-induced preference shifts



Fixed

Myopic

RL

Policy-induced preference shifts

A

Fixed

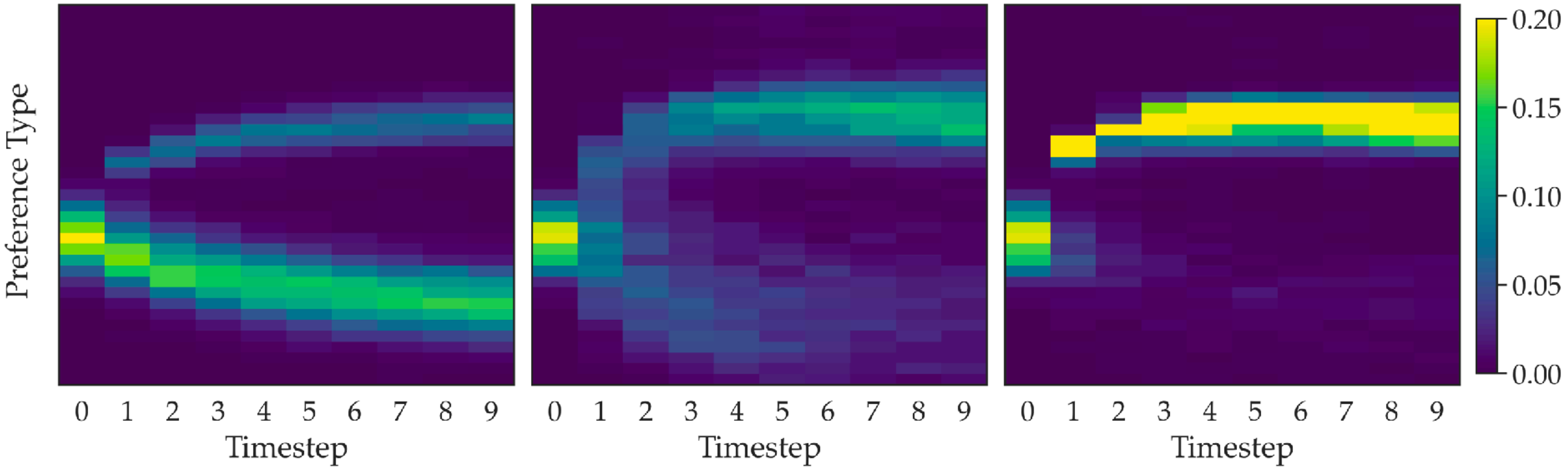
B

Myopic

C

RL

Same user cohort



Policy-induced preference shifts

A

Fixed

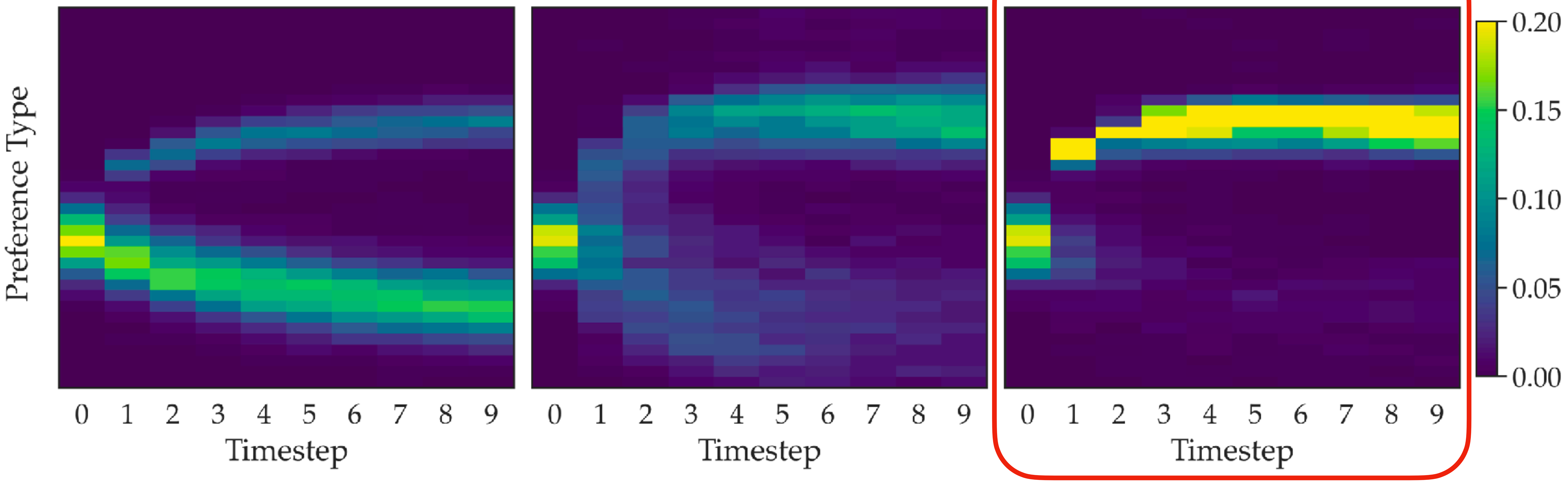
B

Myopic

C

RL

Same user cohort



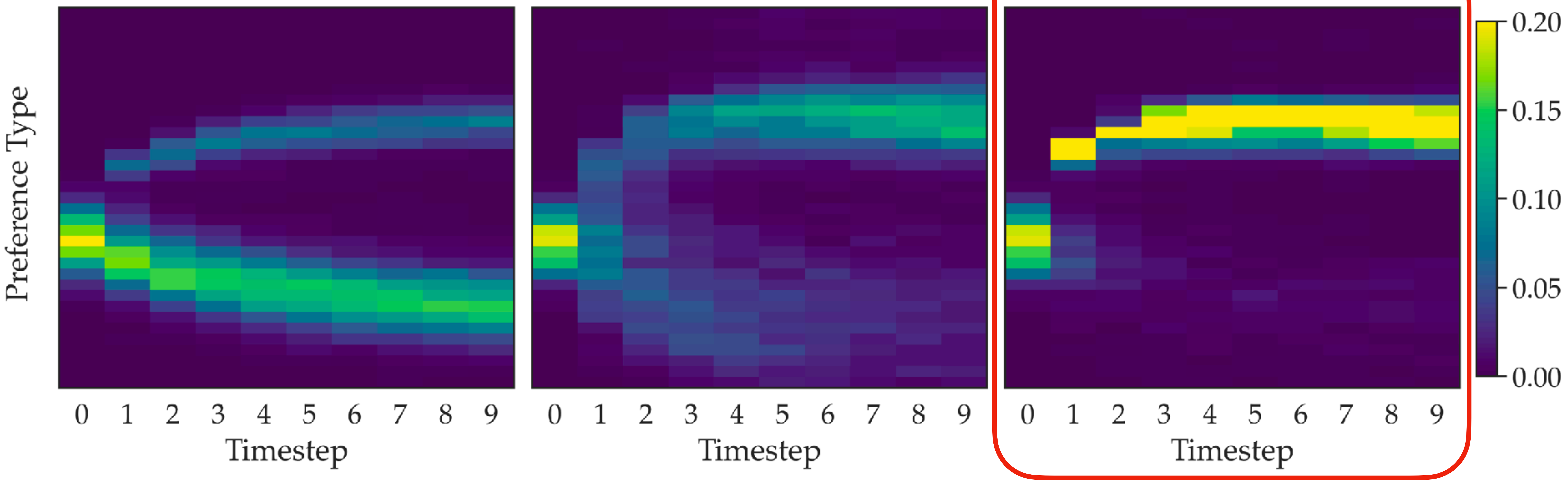
Policy-induced preference shifts

A Fixed

B Myopic

C RL

Same user cohort



Sensationalized news

Dry journalism

How to cope with recommender-induced effects on users?

How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

2) Quantifying: flagging unwanted preference-shifts

How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

2) Quantifying: flagging unwanted preference-shifts

3) Optimizing: penalizing unwanted preference-shifts

How to cope with recommender-induced effects on users?

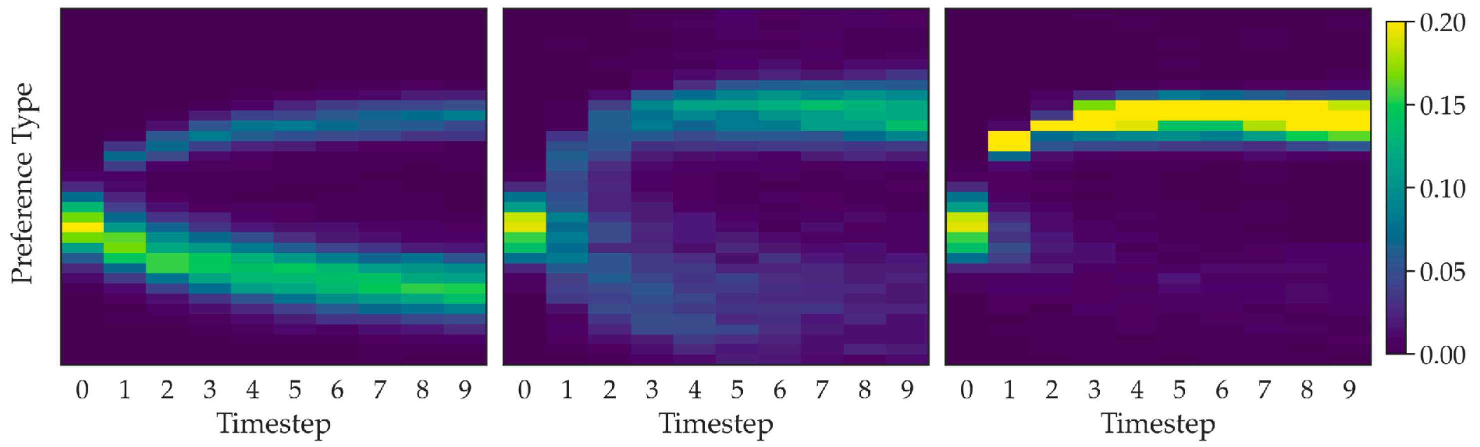
1) Monitoring: estimating preference-shifts

2) Quantifying: flagging unwanted preference-shifts

3) Optimizing: penalizing unwanted preference-shifts

How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

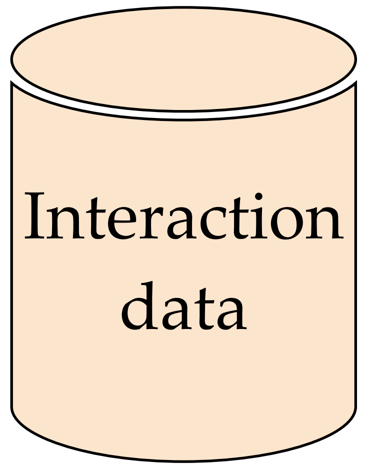


2) Quantifying: flagging unwanted preference-shifts

3) Optimizing: penalizing unwanted preference-shifts

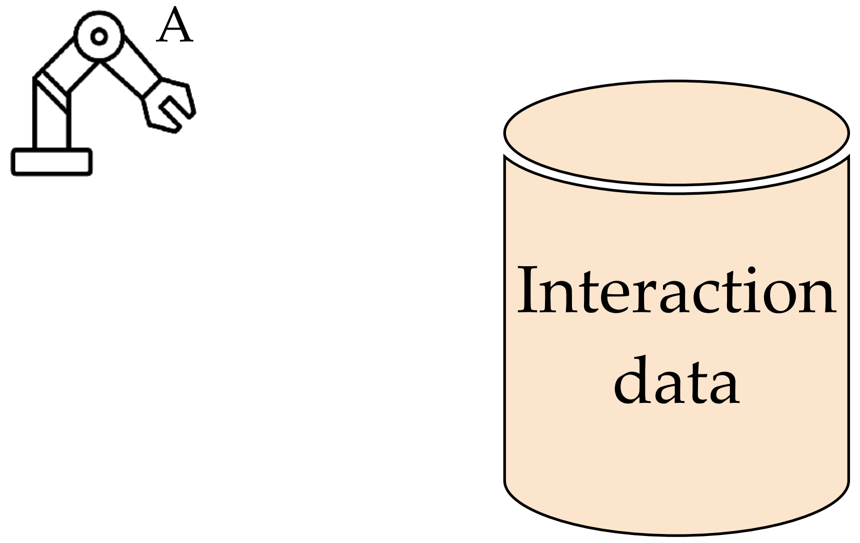
Estimating policy-induced preference shifts

Estimating policy-induced preference shifts



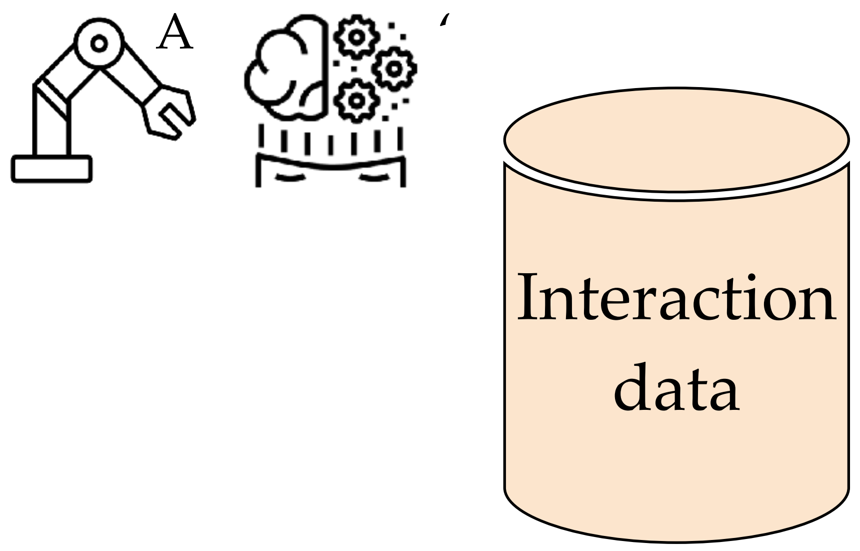
Interaction data

Estimating policy-induced preference shifts



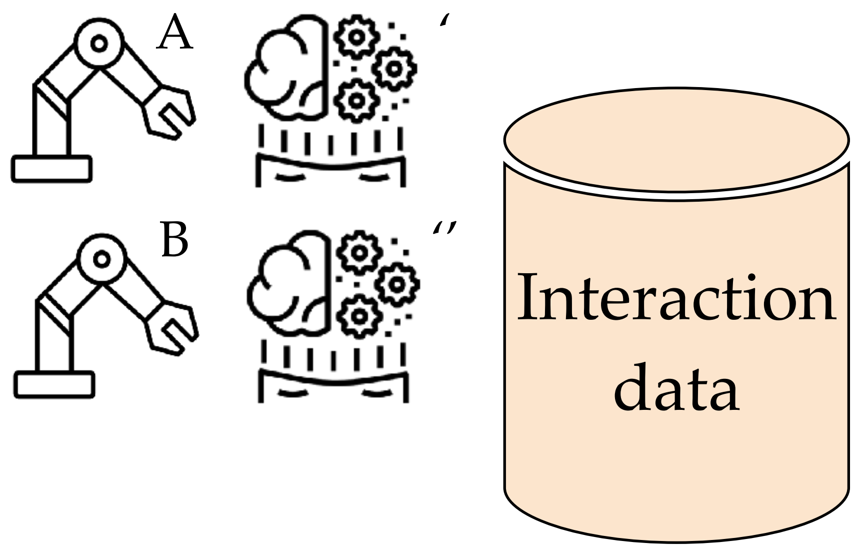
Interaction data

Estimating policy-induced preference shifts



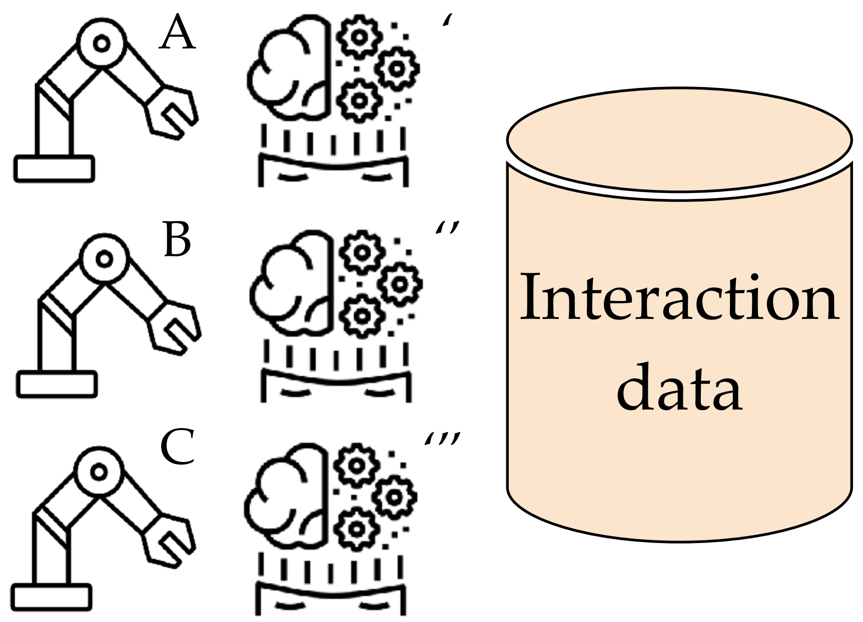
Interaction data

Estimating policy-induced preference shifts



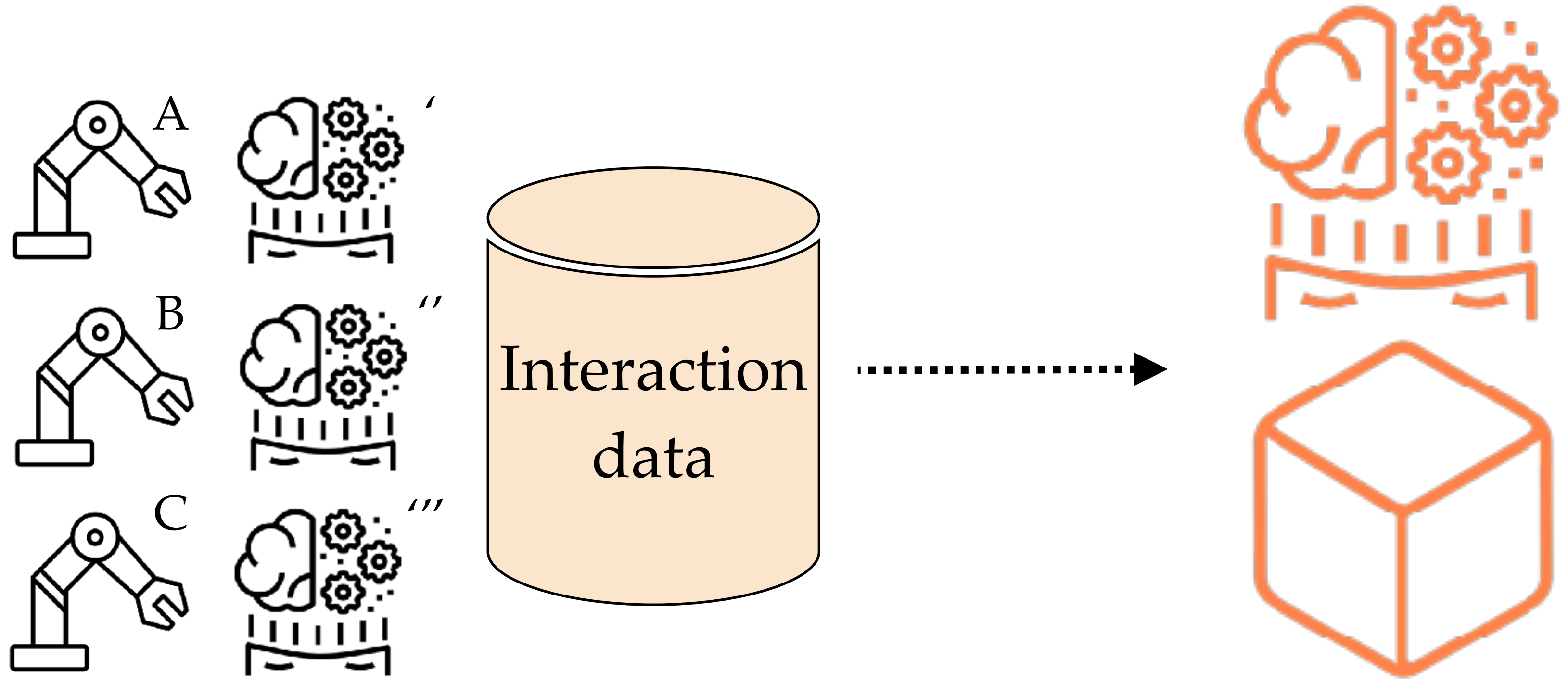
Interaction data

Estimating policy-induced preference shifts



Interaction data

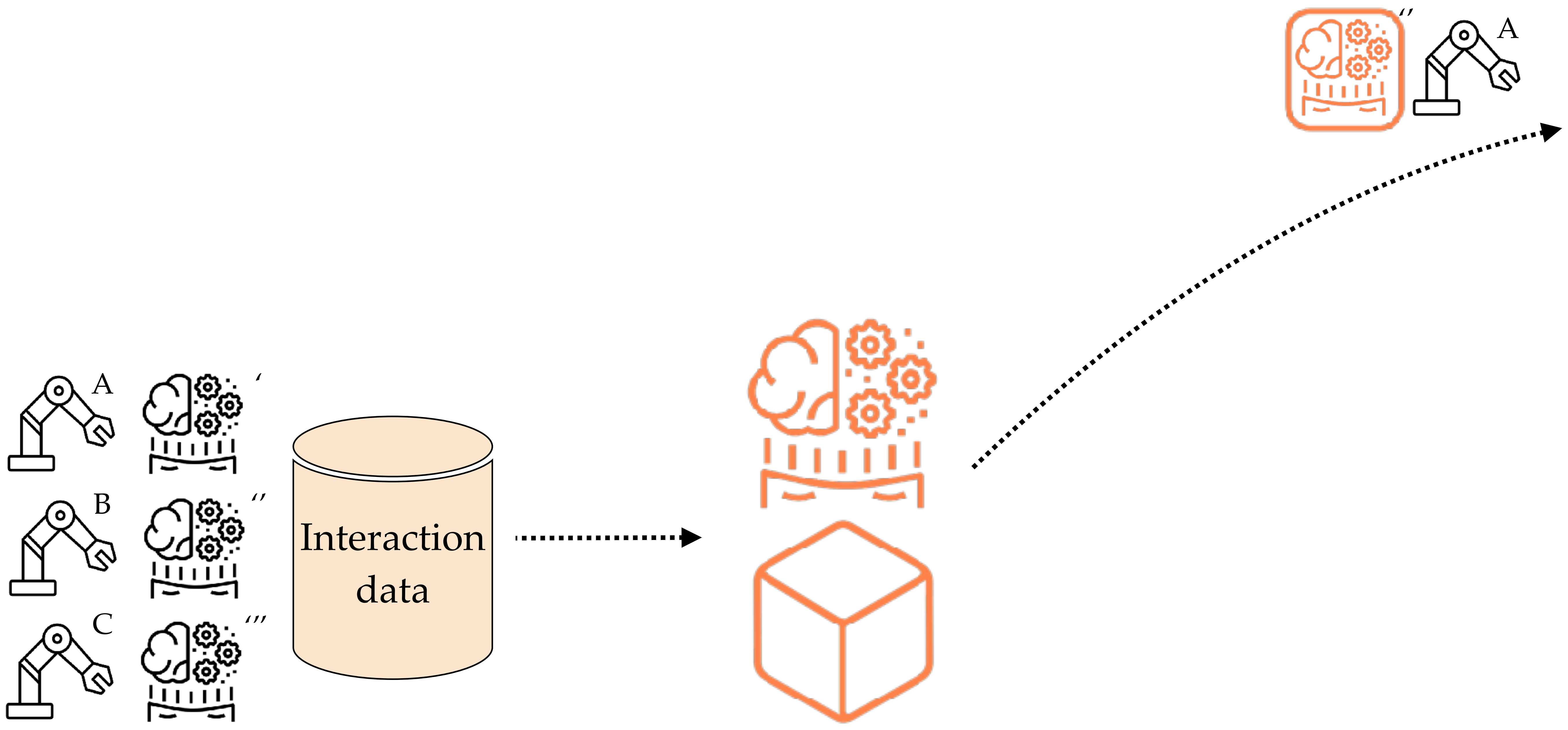
Estimating policy-induced preference shifts



Interaction data

Model of user preference dynamics

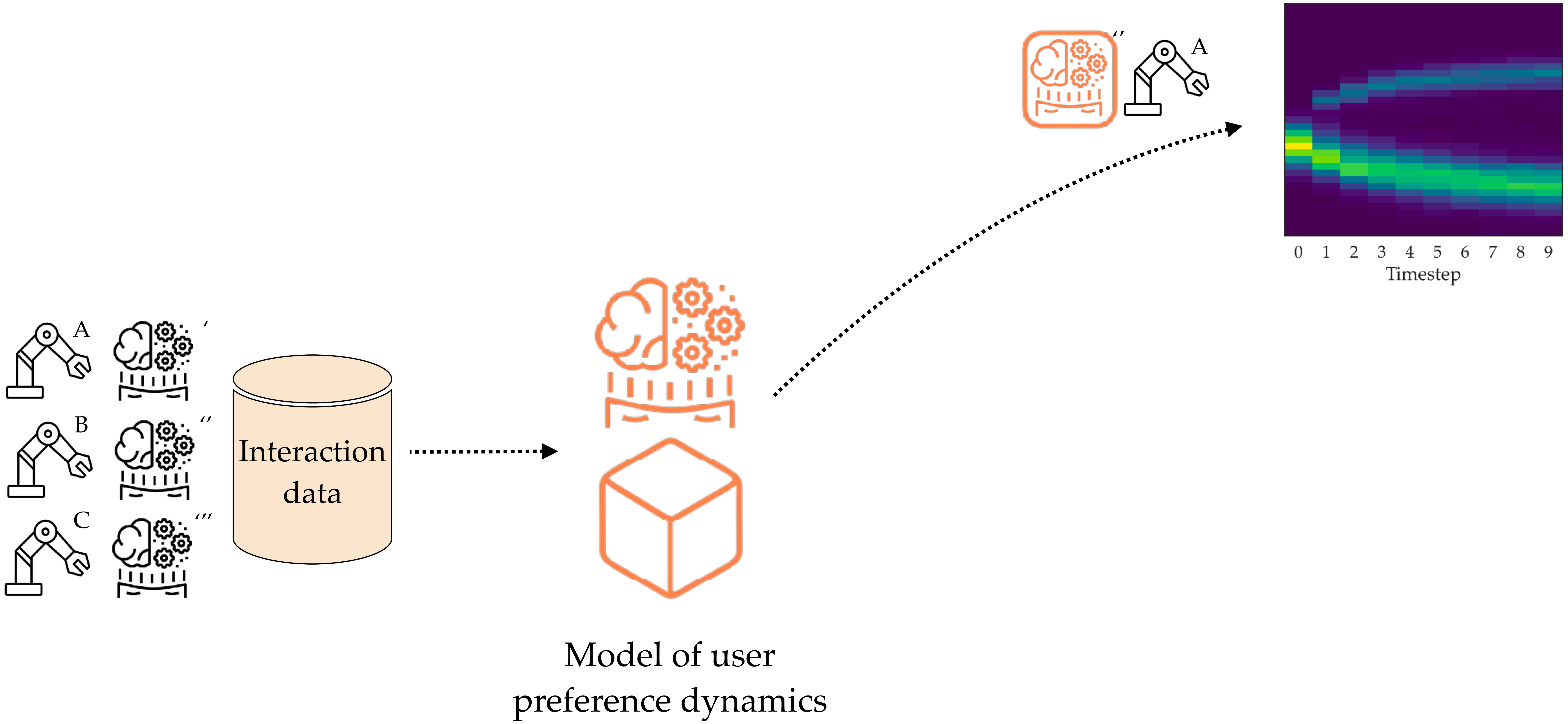
Estimating policy-induced preference shifts



Interaction data

Model of user preference dynamics

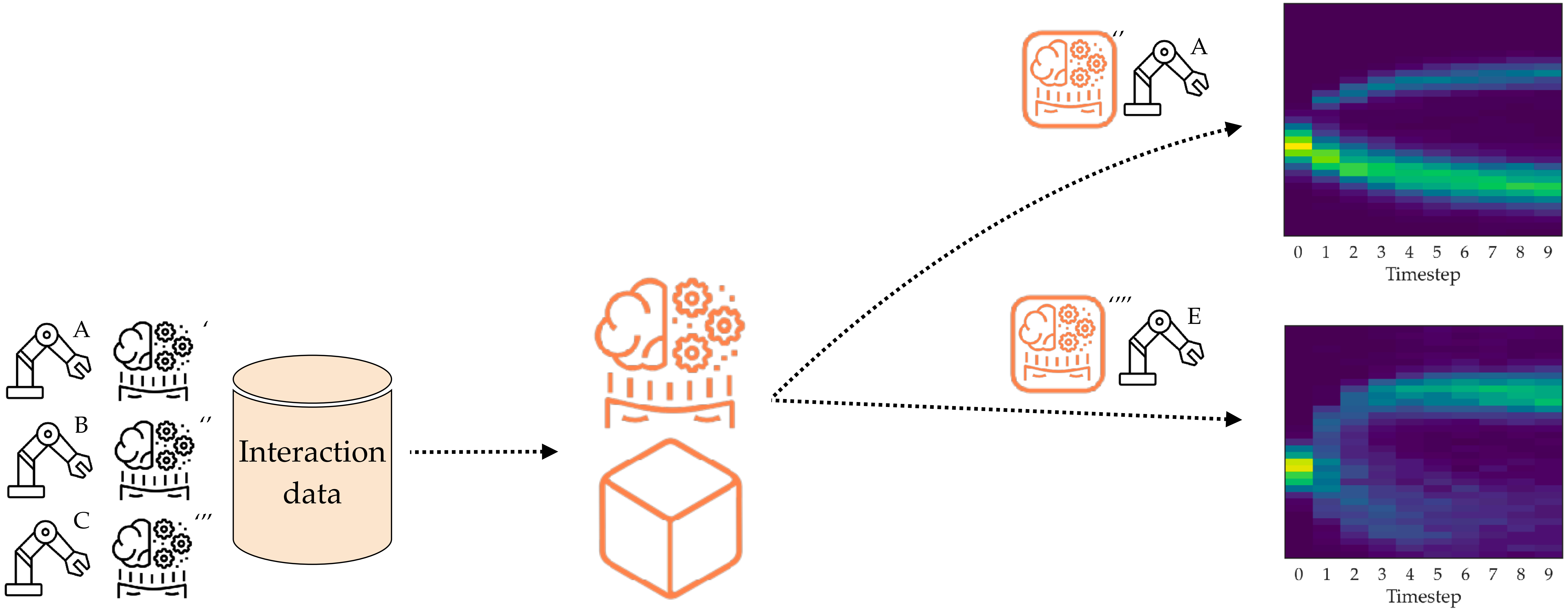
Estimating policy-induced preference shifts



Interaction data

Model of user preference dynamics

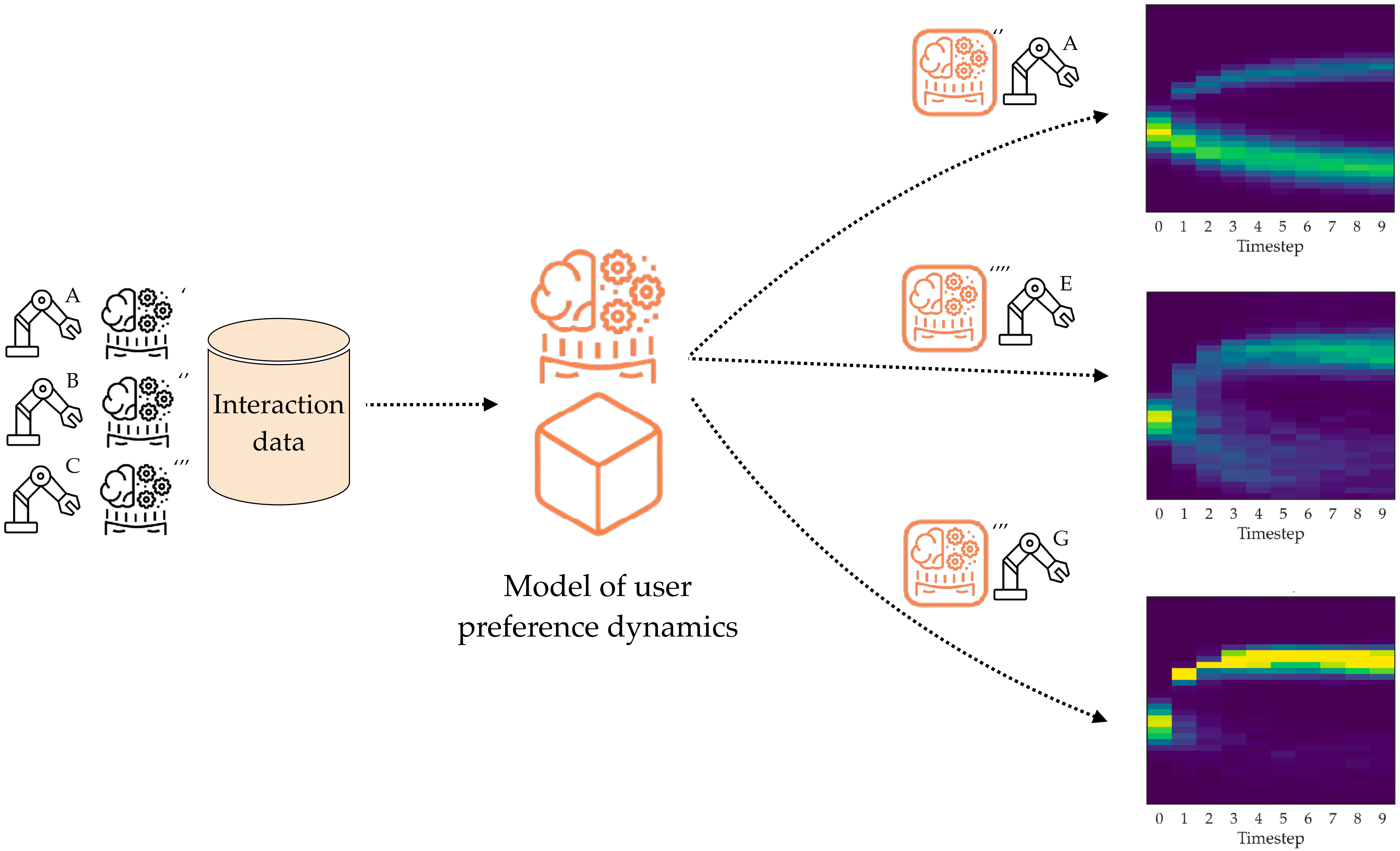
Estimating policy-induced preference shifts



Interaction data

Model of user
preference dynamics

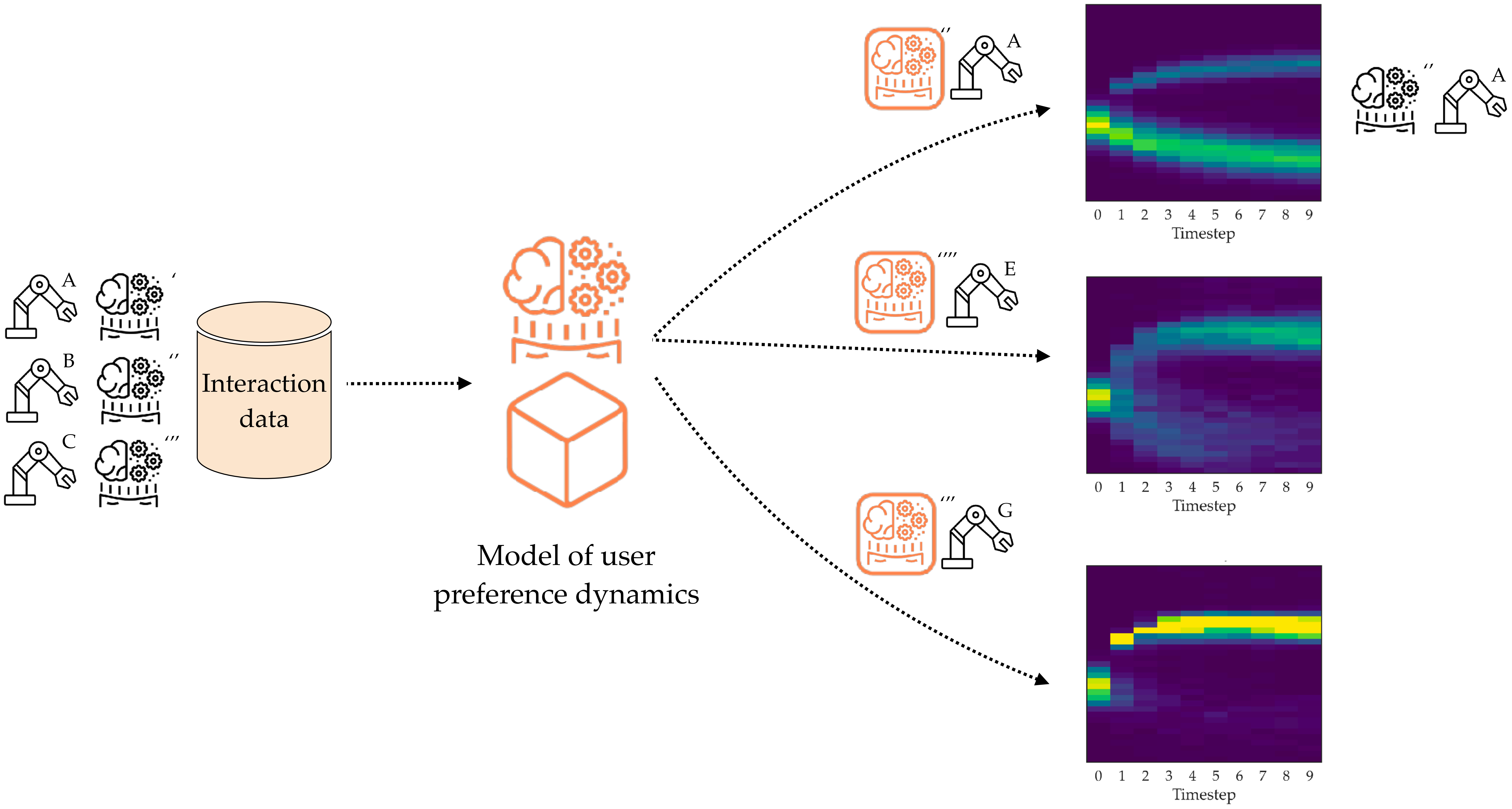
Estimating policy-induced preference shifts



Interaction data

Model of user preference dynamics

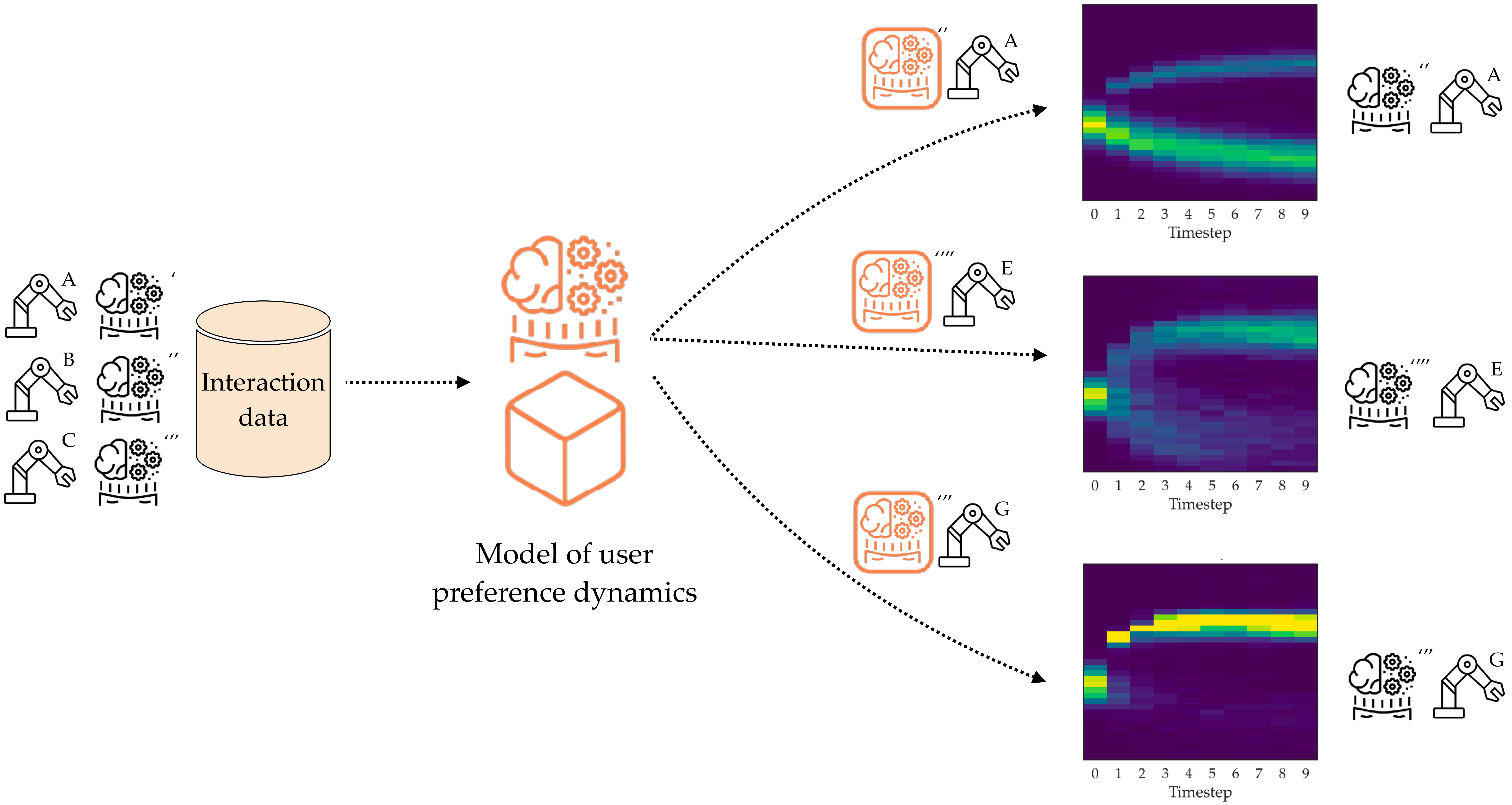
Estimating policy-induced preference shifts



Interaction data

Model of user preference dynamics

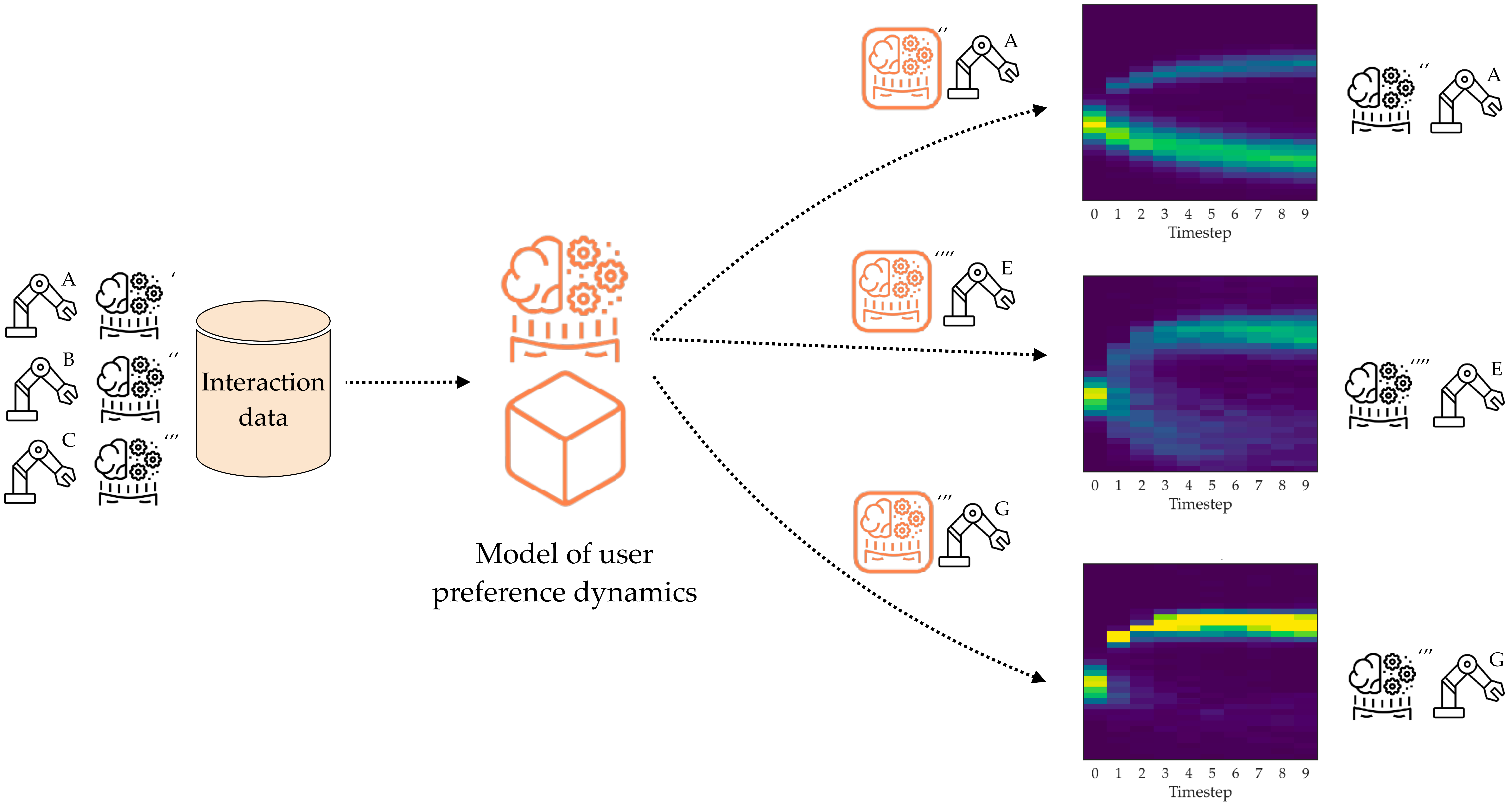
Estimating policy-induced preference shifts



Interaction data

Model of user preference dynamics

Estimating policy-induced preference shifts



Interaction data

Model of user preference dynamics

Estimate the effects of recommenders on user's preferences ***before deployment***

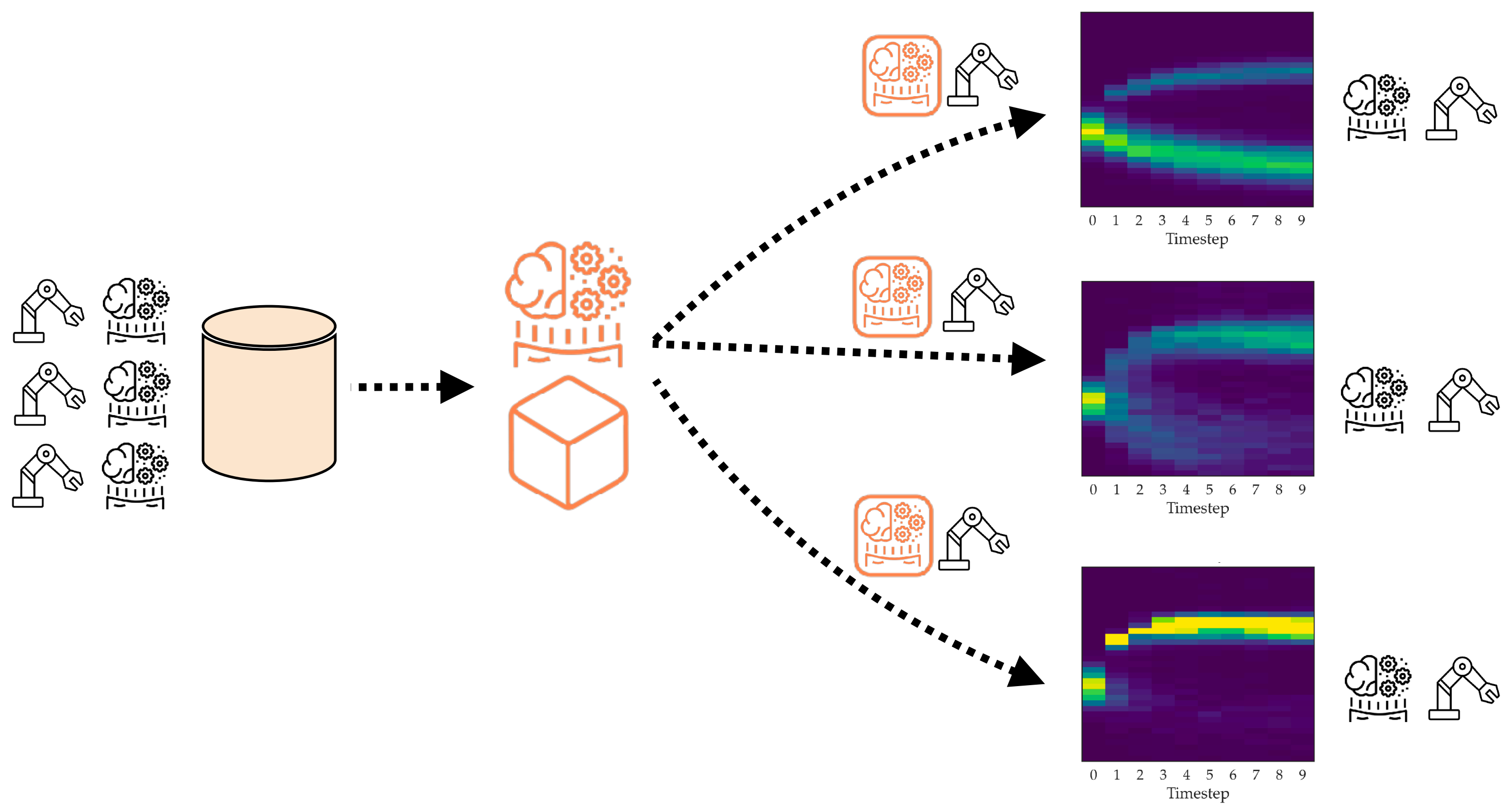
How to cope with recommender-induced effects on users?

How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts



How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

2) Quantifying: flagging unwanted preference-shifts

How to cope with recommender-induced effects on users?

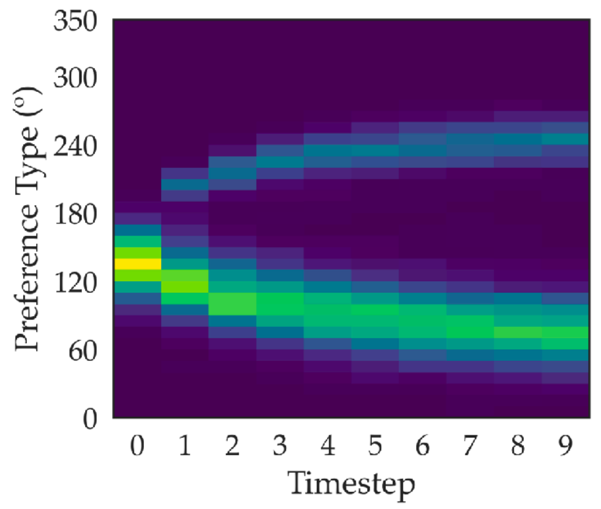
1) Monitoring: estimating preference-shifts

2) Quantifying: flagging unwanted preference-shifts

How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

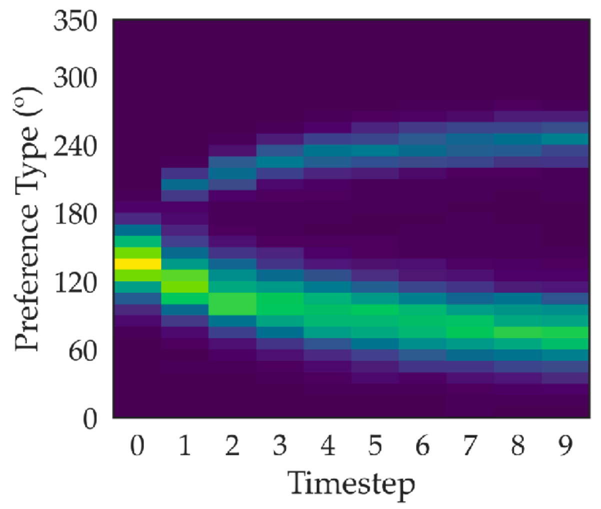
2) Quantifying: flagging unwanted preference-shifts



How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

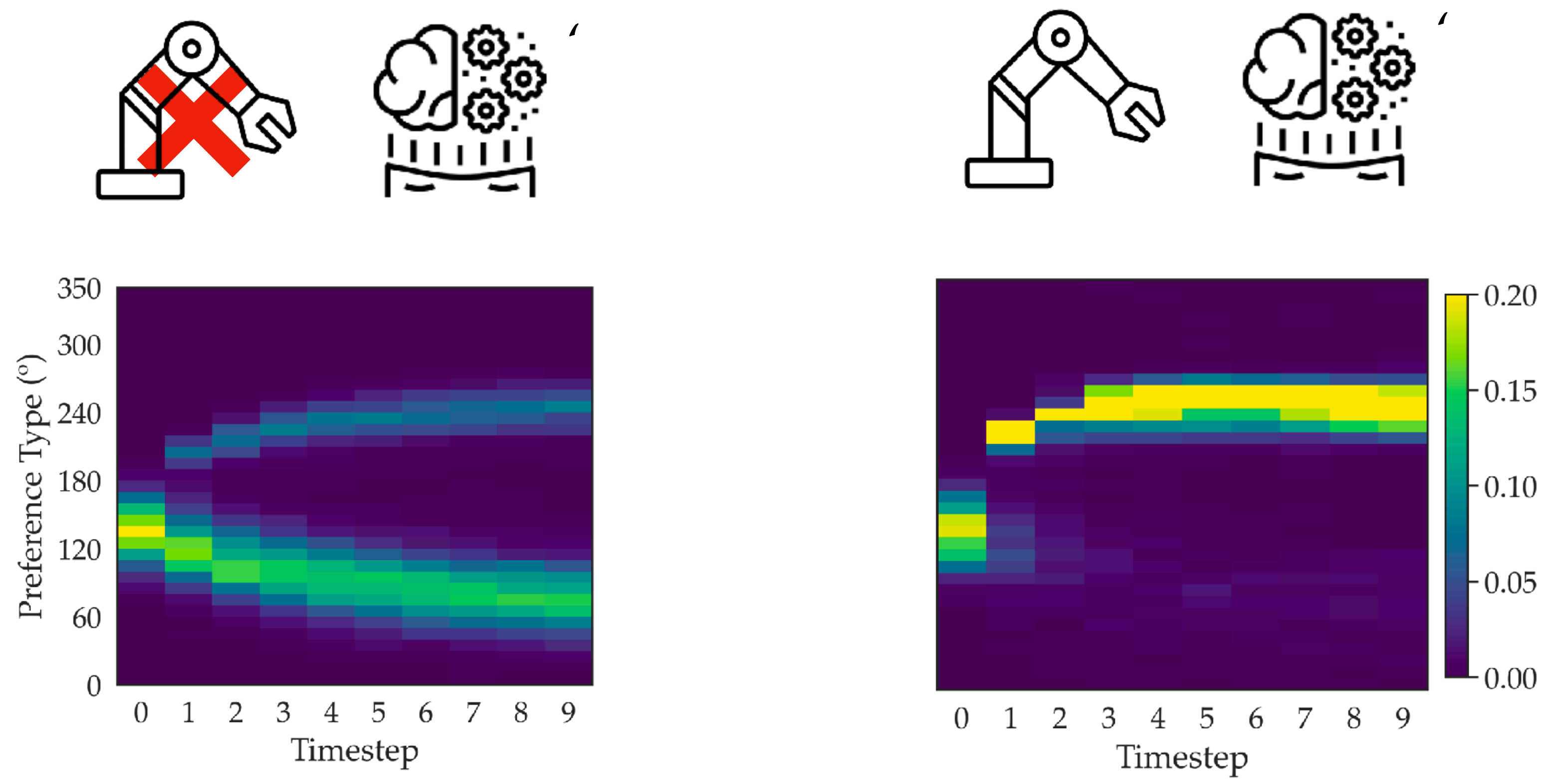
2) Quantifying: flagging unwanted preference-shifts



How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

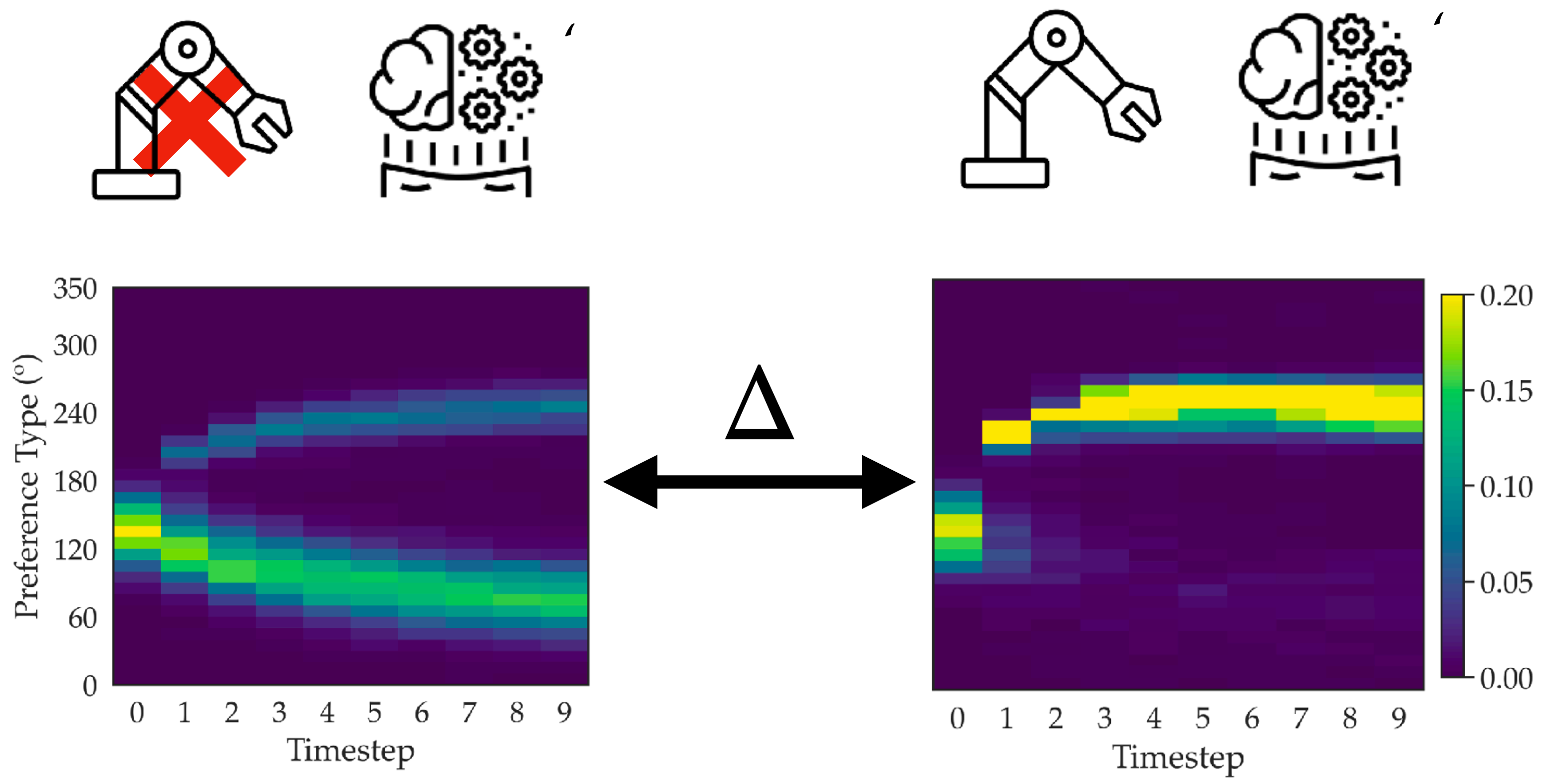
2) Quantifying: flagging unwanted preference-shifts



How to cope with recommender-induced effects on users?

1) Monitoring: estimating preference-shifts

2) Quantifying: flagging unwanted preference-shifts

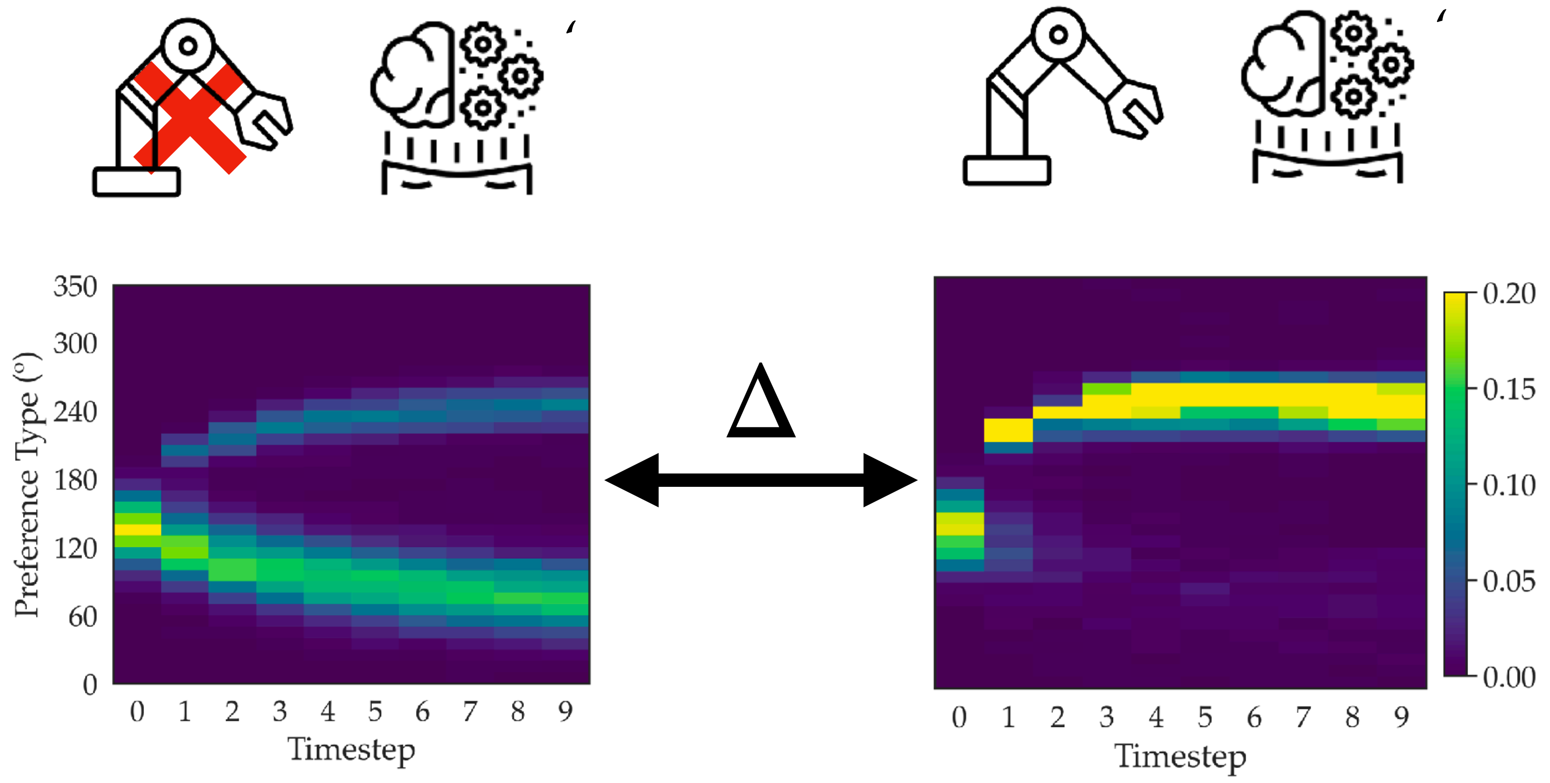


How to cope with recommender-induced effects on users?

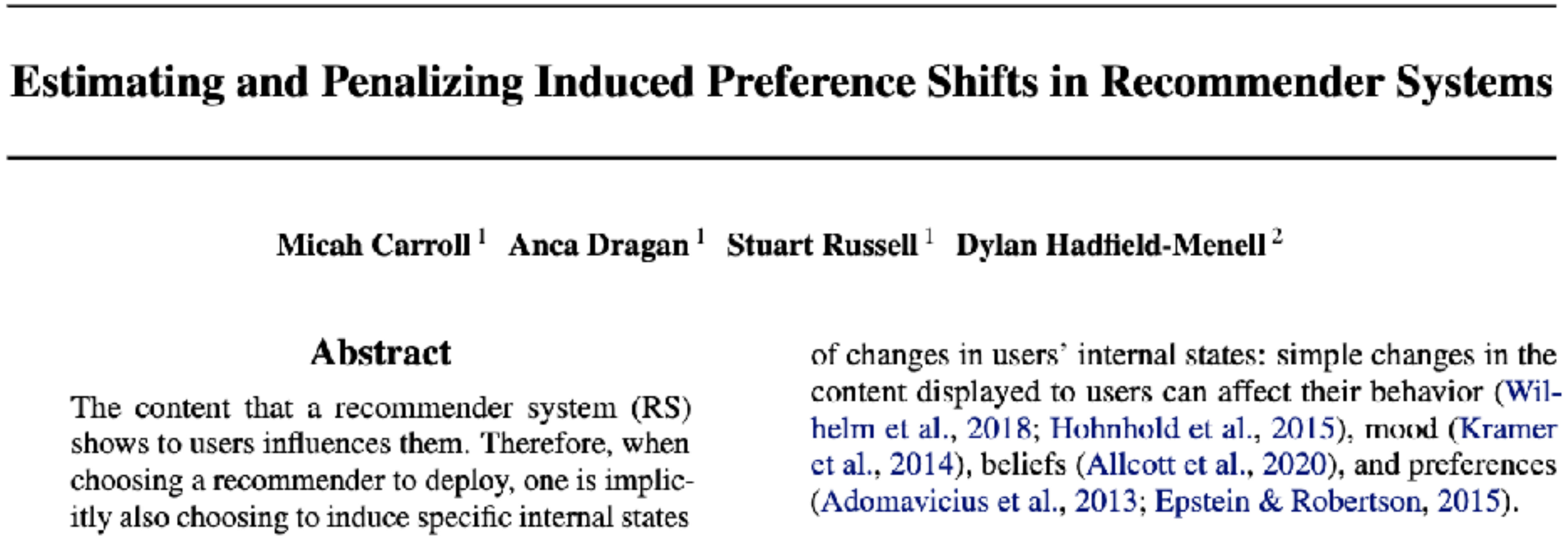
1) Monitoring: estimating preference-shifts

2) Quantifying: flagging unwanted preference-shifts

3) Optimizing: penalizing unwanted preference-shifts



See the paper for more details!



Estimating and Penalizing Induced Preference Shifts in Recommender Systems

Micah Carroll[1] Anca Dragan[1] Stuart Russell[1] Dylan Hadfield-Menell[2]

Abstract

The content that a recommender system (RS) shows to users influences them. Therefore, when choosing a recommender to deploy, one is implicitly also choosing to induce specific internal states

of changes in users' internal states: simple changes in the content displayed to users can affect their behavior (Wilhelm et al., 2018; Hohnhold et al., 2015), mood (Kramer et al., 2014), beliefs (Allcott et al., 2020), and preferences (Adomavicius et al., 2013; Epstein & Robertson, 2015).