full paper

SCAN ME

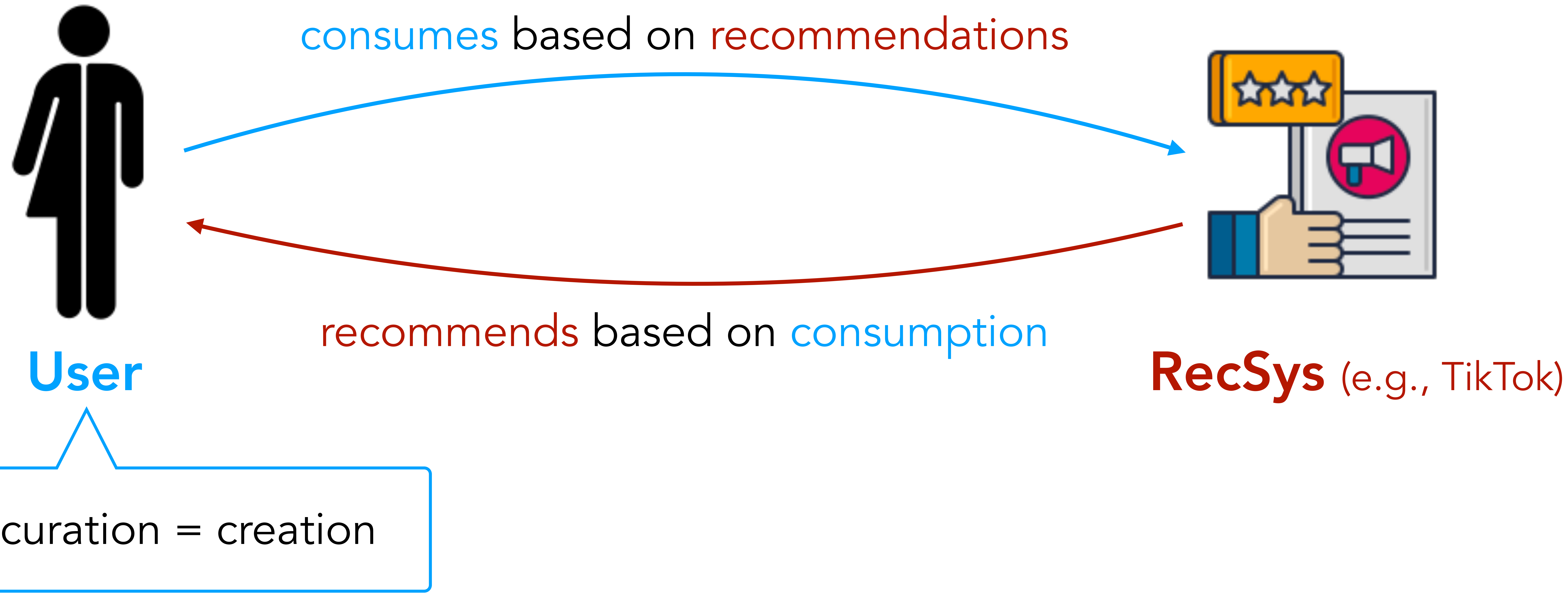
The Disparate Effects of Recommending to Strategic Users

Chara Podimata (MIT)

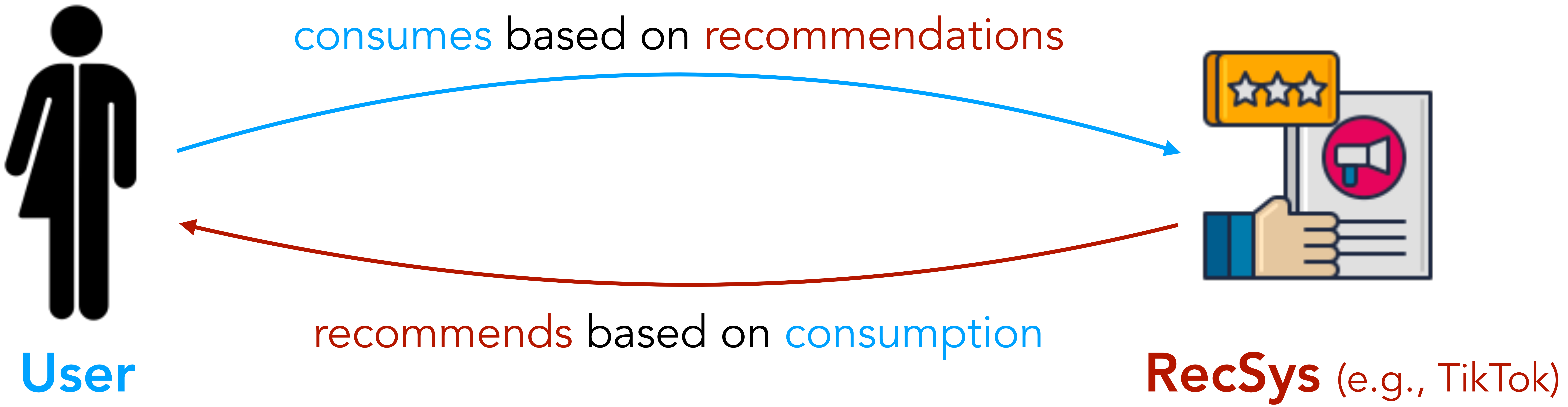
joint work with ***Andy Haupt (MIT)*** and ***Dylan Hadfield-Menell (MIT)***

Future of AI and Economics, 07/31/2023

RecSys create a feedback loop



Main Questions



Main Questions

Q1: Are users aware of feedback loop? Do they act in response?

consumes based on recommendations

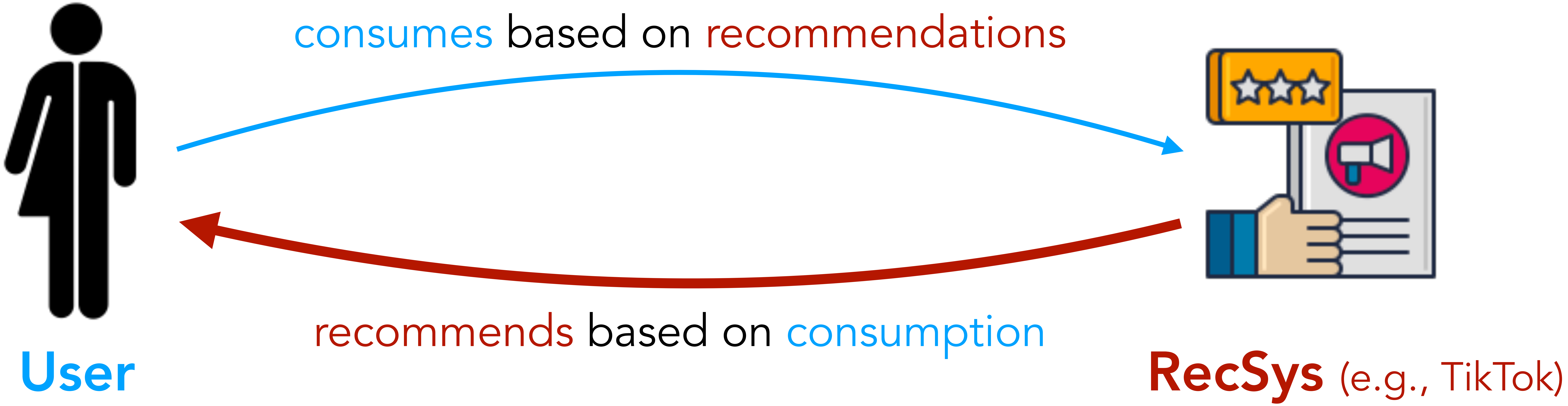
recommends based on consumption

User

RecSys (e.g., TikTok)

Main Questions

Q1: Are users aware of feedback loop? Do they act in response?



Q2: Harms to users if RecSys does not adapt? Interventions?

Contributions

Q1: Are users aware of feedback loop? Do they act in response?

1. **Survey** on user consumption patterns on TikTok.

2. Introduction of theoretical **model** about recommending to strategic agents.

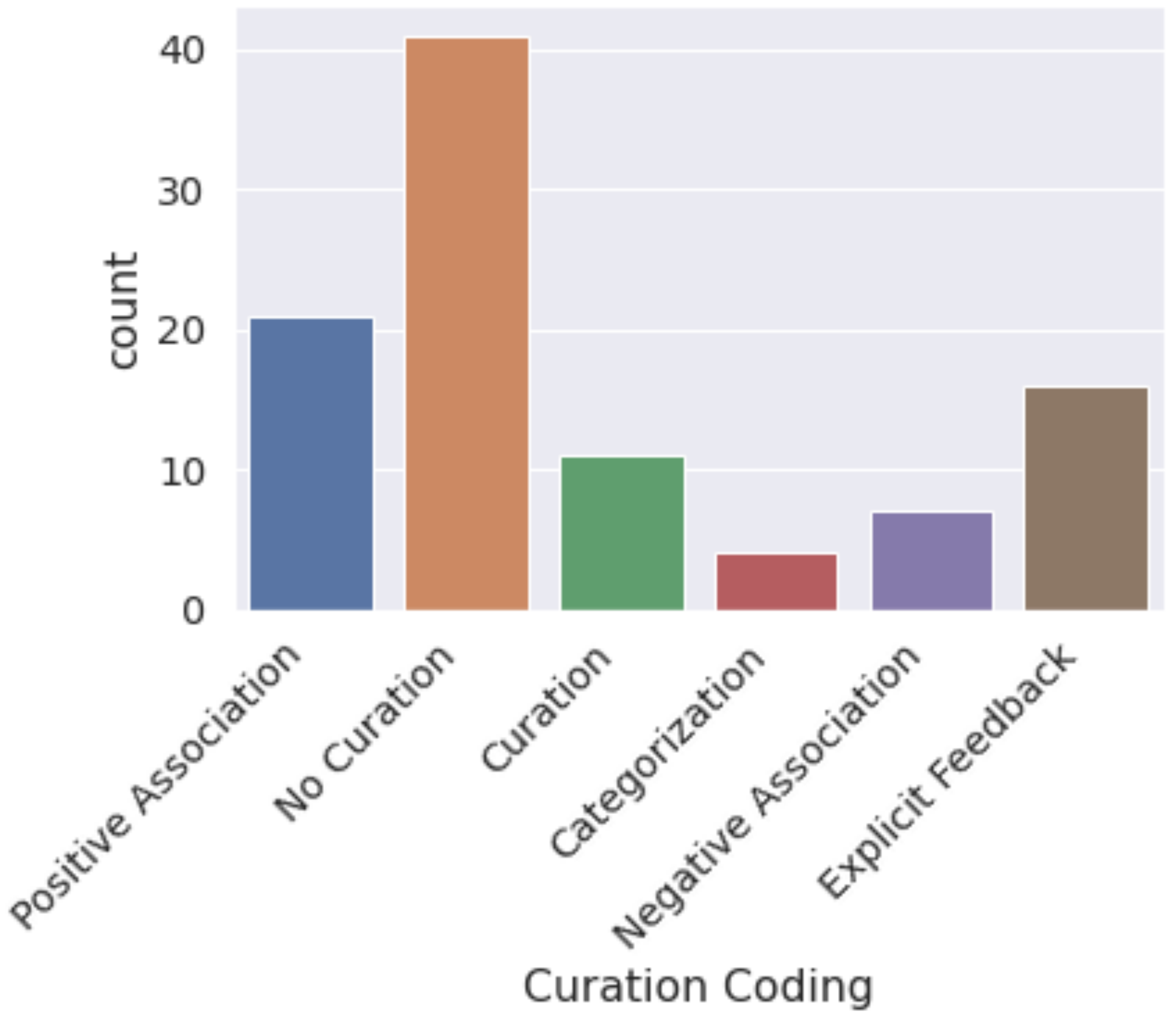
3. **Disparate impact** for **minority** in equilibrium (proof of concept).

   Q2: Harms to users if RecSys does not adapt? Interventions?

Survey on TikTok consumption

* Survey on MTurk 12/22 - 01/23 (100 participants)
* Mostly ***free-form text responses*** → qualitative data analysis

Why do you think TikTok recommends these categories?



*"I feel that TikTok continues to put these in my feed because I almost always get sucked into watching them. That **tells the algorithm I like them**, even though I am mostly **just using them for background noise** and have seen most of them before"*

"I think because I liked a video once of this type of content. I believe by me liking the video the algorithm thought I would like to see more videos like that one."

Actions you take to curate your feed?

*"I also like stuff just to see more of that type stuff evn though I don't like it. LIke soemtimes if my content gets to dark I try to like animal videos and **comedy more to get off the darker content for a bit**." [sic]*

*"Currently, I am **cognizant of what category** of video I think material falls under. I am careful to watch completely videos that fall under the correct category (**even if I am not interested in that particular video**). I am careful to **skip** over videos from the "**wrong**" categories."*

Contributions

Q1: Are users aware of feedback loop? Do they act in response?

Q2: Harms to users if RecSys does not adapt? Interventions?

1. **Survey** on user consumption patterns on TikTok.

   amazon mechanical turk

2. Introduction of theoretical **model** about recommending to strategic agents.

3. **Disparate impact** for **minority** in equilibrium (proof of concept).

Model: Strategic Recommendation as a Stackelberg Game

Players Leader: ***RS*** Follower: ***User***

Timeline of play

RS commits to a **strategy** g. Users **observe** g.

Exploration phase (Cold Start)

Exploitation phase (Recommendation)

time

RS randomly presents contents to users

Users **consume** content ~ **preferences + strategy**

From consumption pattern, ***RS* infers user type** (e.g., sporty spice).

RS implements **policy** g to **map** inferred type to recommended content

Model: Strategic Recommendation

How does ***RS*** choose the **recommendation policy** g?

Exploration phase data at **face value** → choose g to **max welfare**

How does the ***User*** choose the **consumption plan** a?

$$\Big(\text{consumption} \sim \text{Poisson}(\text{exposure\_rate} \cdot \text{consumption\_plan})\Big)$$

$$\text{user\_utility}(\delta, a, g) = \underbrace{u^{\text{CS}}(a)}_{\text{Exploration phase utility}} + \underbrace{\frac{\delta}{1-\delta}}_{\text{future discount factor}} \underbrace{u^{\text{Rec}}(g(a))}_{\text{Exploitation phase utility}}$$

Contributions

Q1: Are users aware of feedback loop? Do they act in response?

Q2: Harms to users if RecSys does not adapt? Interventions?

1. **Survey** on user consumption patterns on TikTok.

2. Introduction of theoretical **model** about recommending to strategic agents.

3. **Disparate impact** for **minority** in equilibrium (proof of concept).

   Sources:

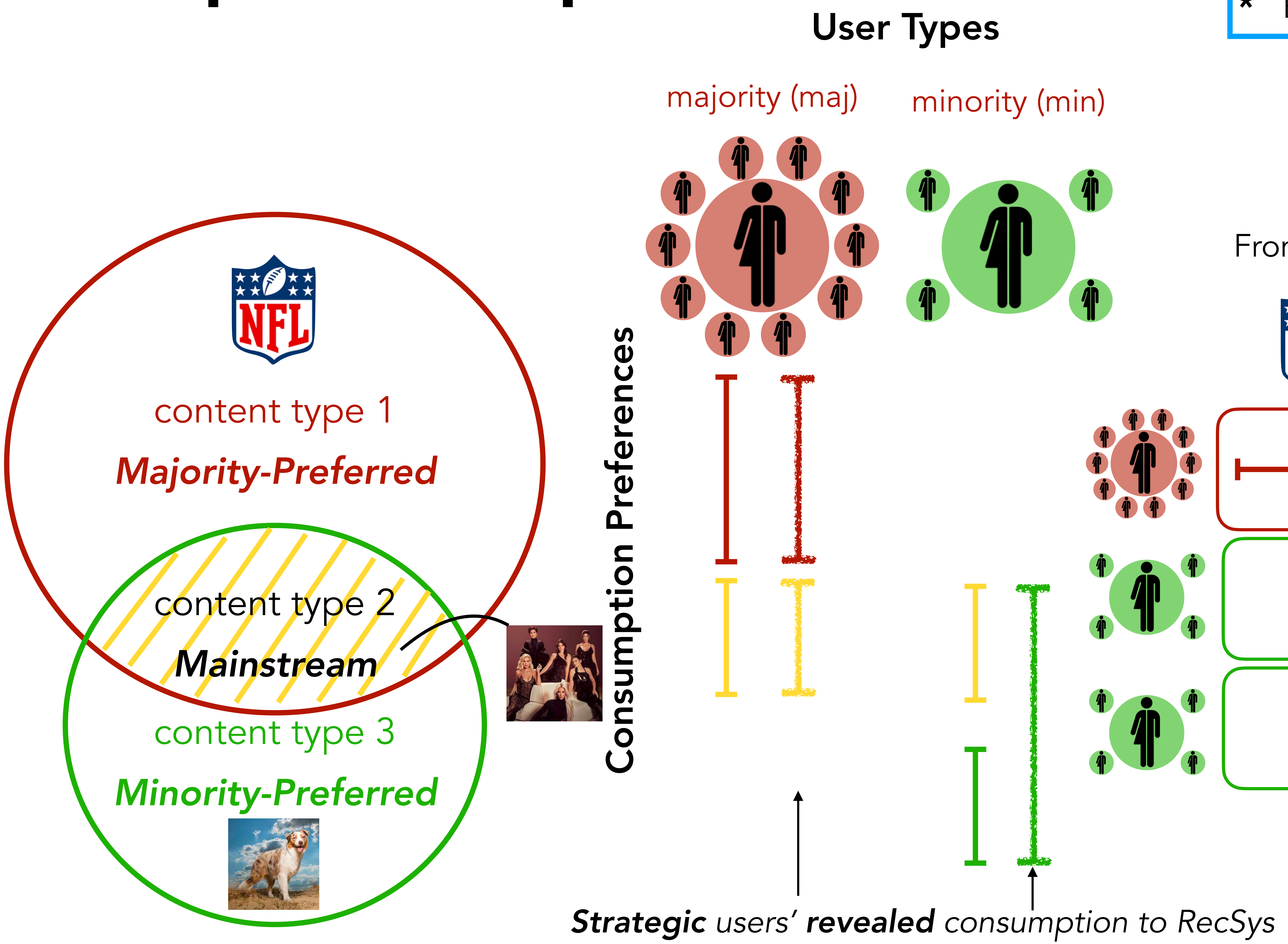
   i. cognitive burden

   ii. utility under strategizing vs under truthtelling

Disparate impact

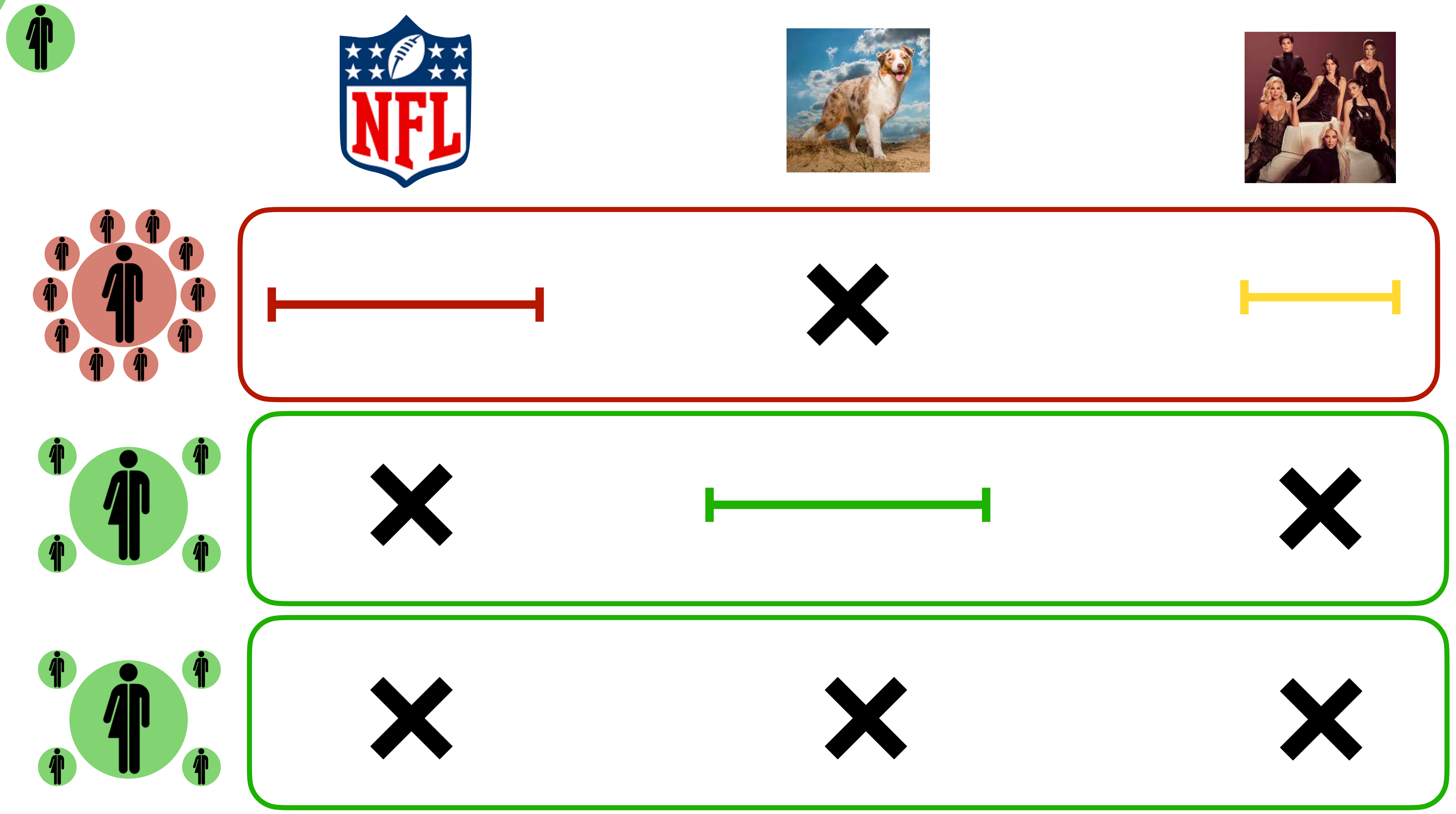
RecSys equilibrium:

* Recommend **min-pref**, if no preference for content type 1, 2 or some preference for type 3.
* Else, **maj-pref**.



Why equilibrium?

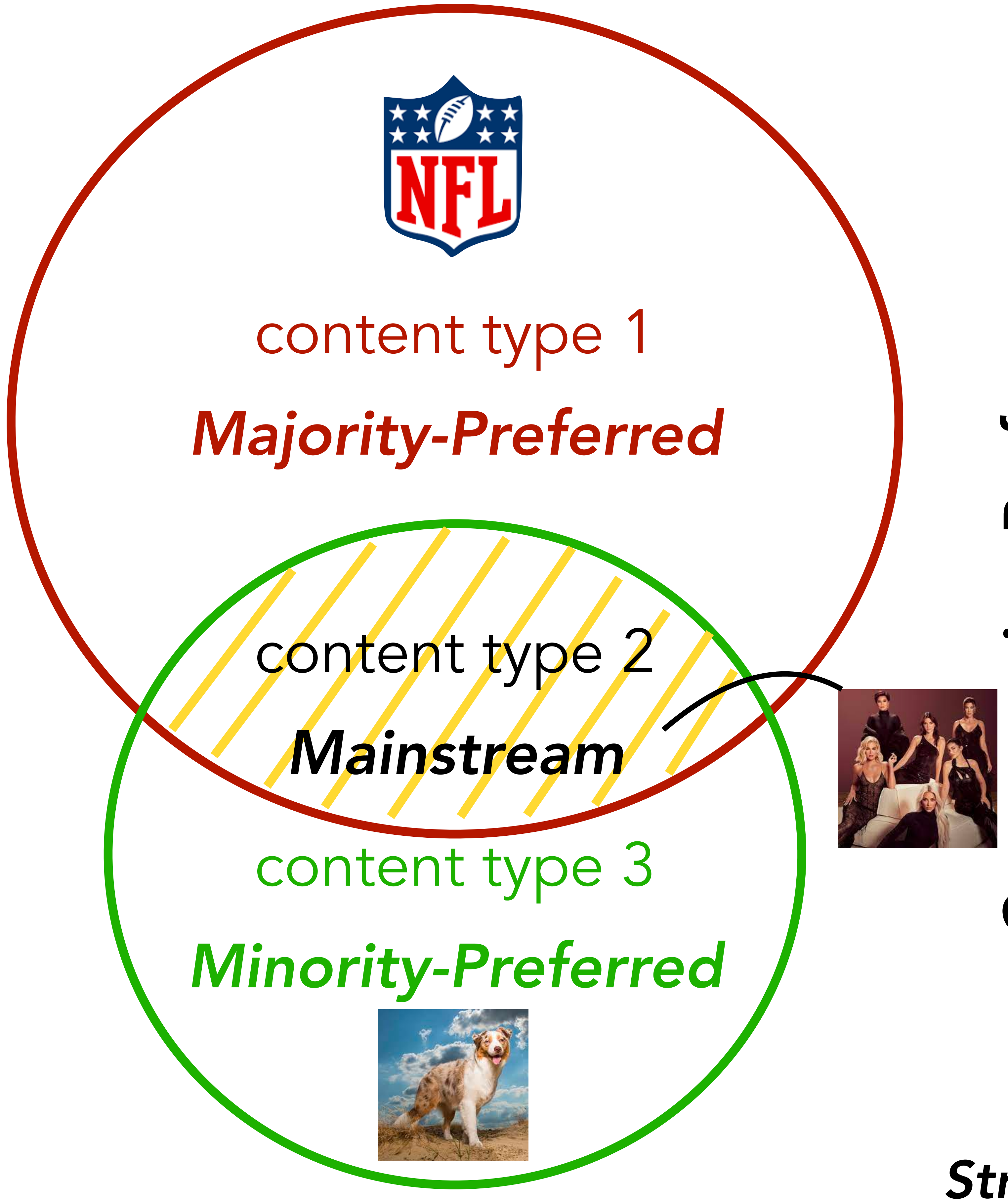
From *RS* perspective: possible consumption profiles

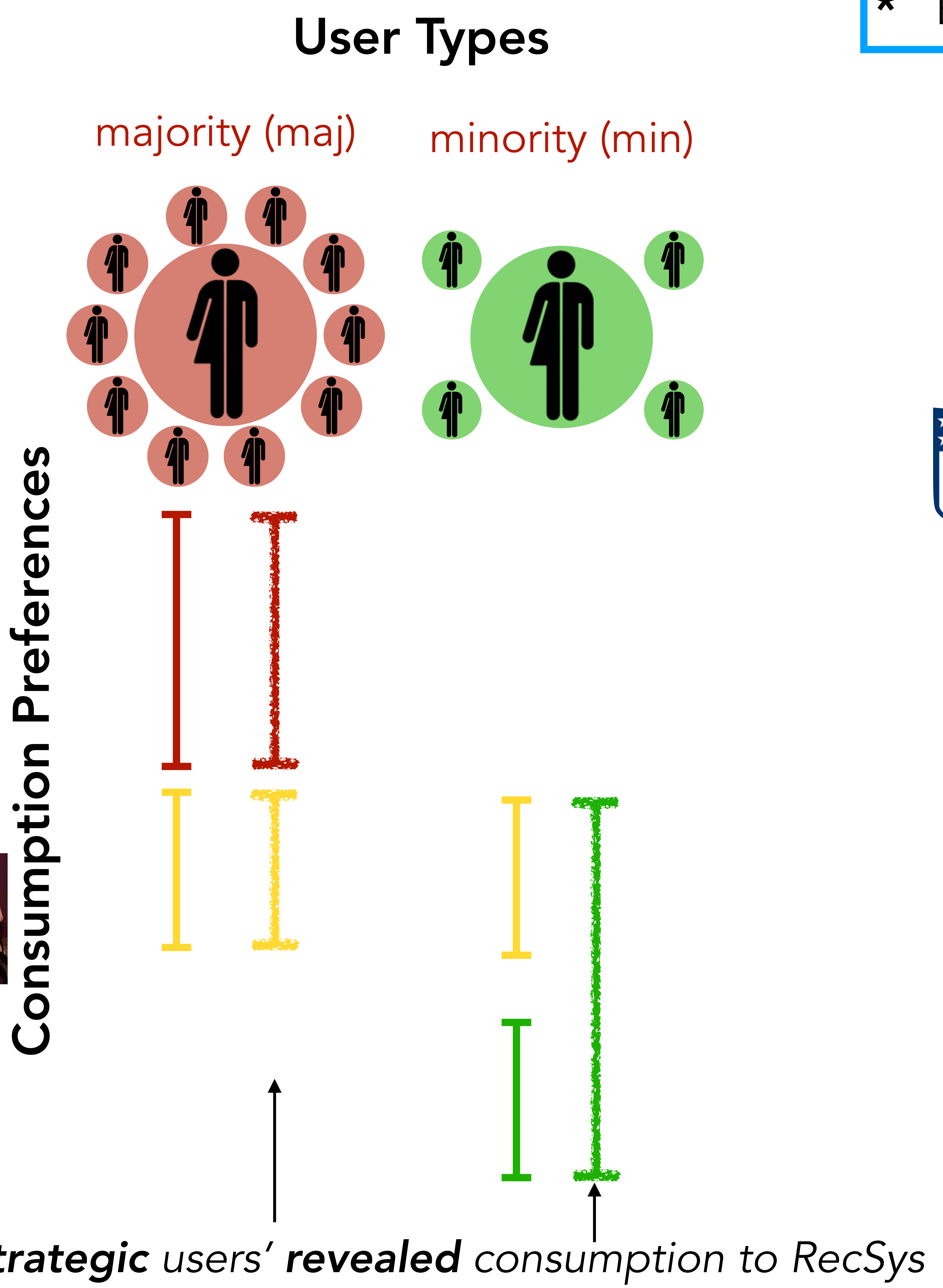


Disparate impact

RecSys equilibrium:

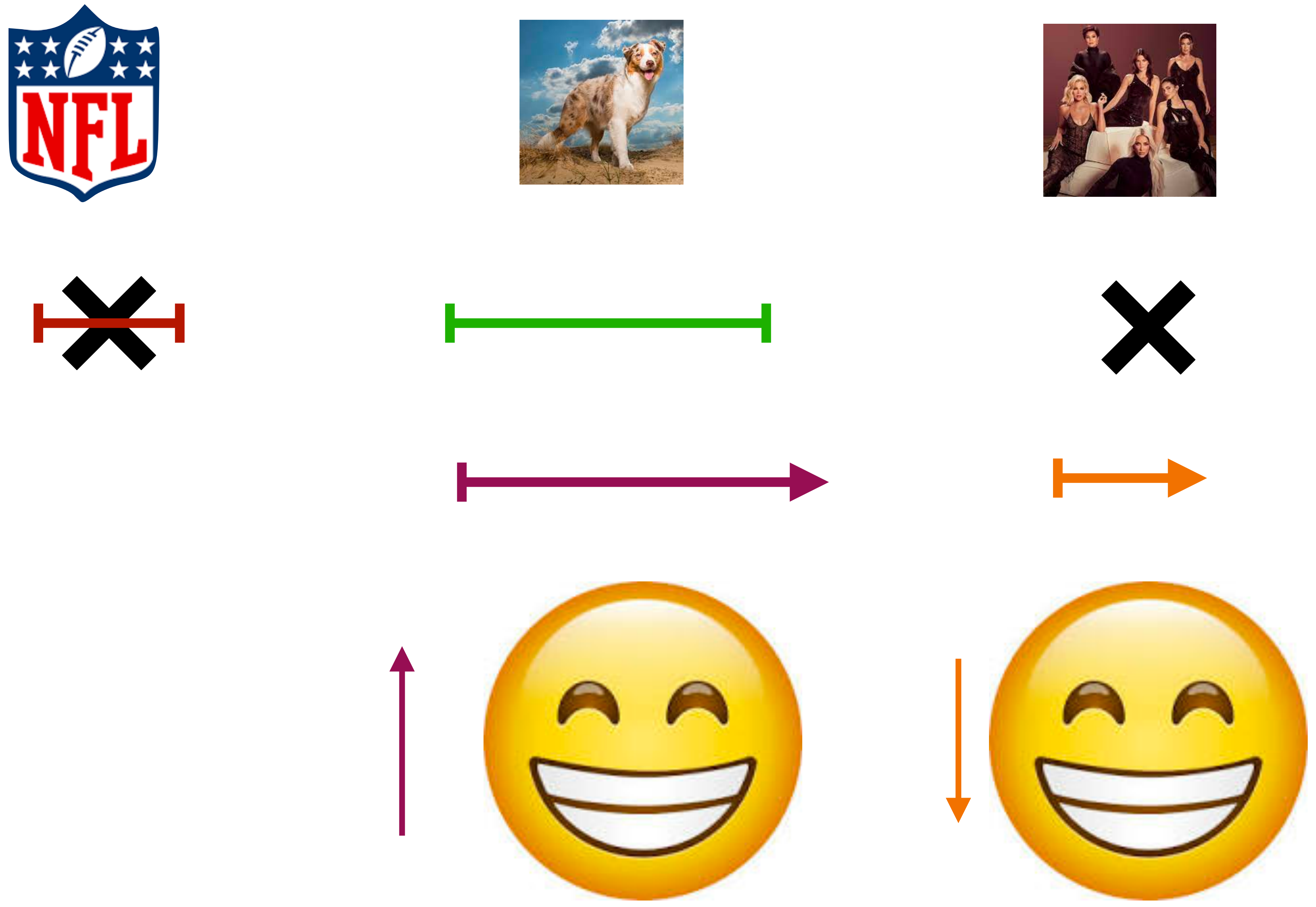
* Recommend **min-pref**, if no preference for content type 1, 2 or some preference for type 3.
* Else, **maj-pref**.





Why equilibrium?

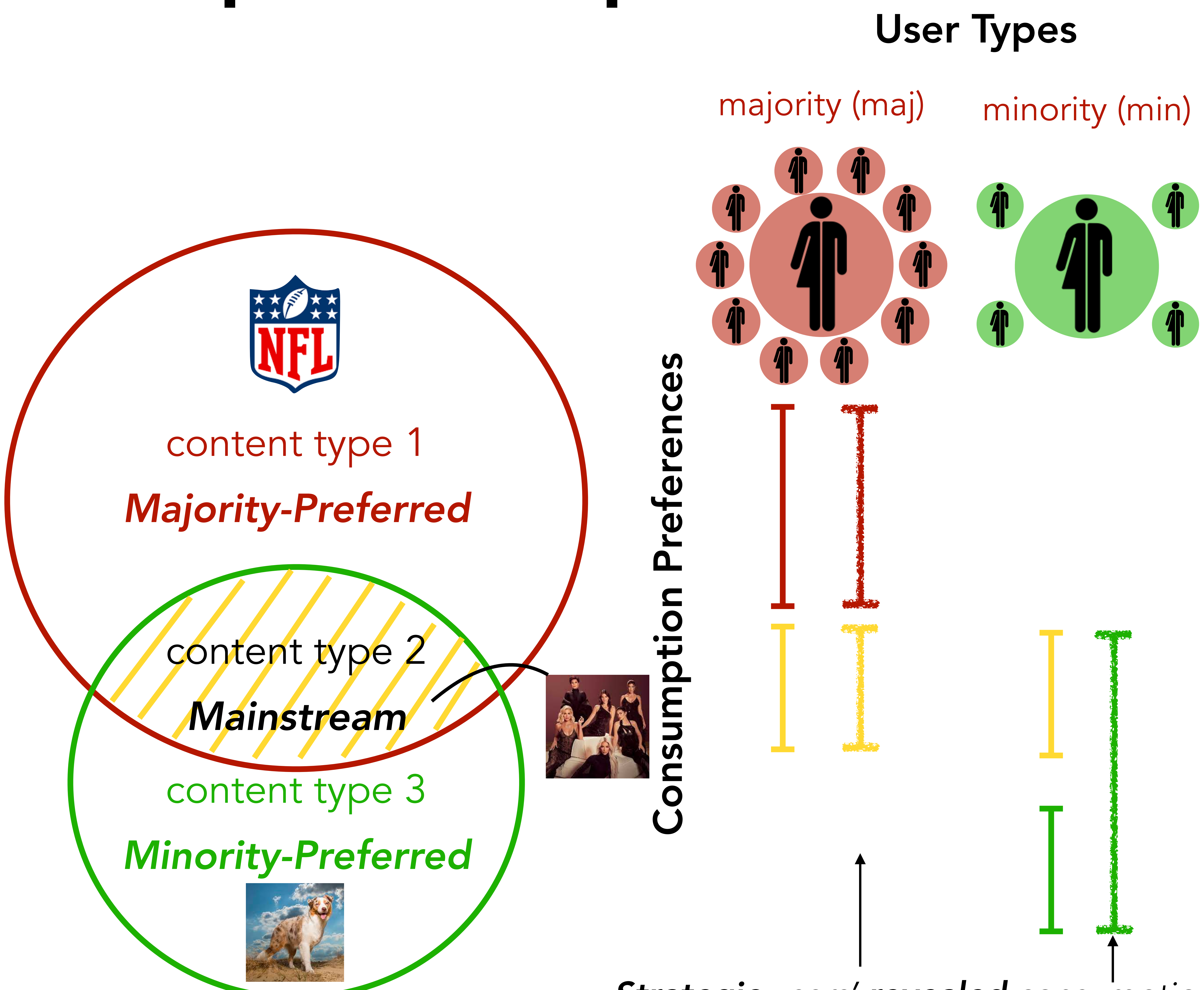
From ***User*** perspective: minority user utility



Disparate impact

RecSys equilibrium:

* Recommend **min-pref**, if no preference for content type 1, 2 or some preference for type 3.
* Else, **maj-pref**.



Strategic *users'* ***revealed*** *consumption to RecSys*

Implications for Minority

1) ***maj*** users: ***no need to strategize***!
2) ***min*** users: act more ***stereotypically***
3) ***min*** users: ***mainstream abstention***

> *"I make sure to interact things that are specific to content types I want to see,* ***even if I don't really like the content of that specific video****."*

> *"I would use that to search things that I wouldn't want recommended to me.* ***Stuff that I like, but stuff that I wouldn't want to clog my feed.****"*

Fundamental problem in recommending to strategic users

RS has **imperfect, coarse information** wrt user type.

in real life

"I'm mostly a sporty spice but I also really like dogs. The Kardashians are my guilty pleasure."

ideal information

Interventions guiding principle: fine tune learning priorities wrt user type inference

Open Questions / Directions

1. Understanding platforms' awareness of individuals' incentives.

2. Modeling **incentives** and system dynamics for **content creators**.

3. Understanding the **Price of Personalization** in RecSys.

Contributions

Q1: *Are users aware of feedback loop? Do they act in resp*

Q2: *Harms to users if RecSys does not adapt? Interventior*

1. **Survey** on user consumption patterns on TikTok.

2. Introduction of theoretical **model** about recommending to strategic agents.

3. **Disparate impact** for **minority** in equilibrium (proof of concept).

   Sources:

   i cognitive burden   ii utility under strategizing vs under truthtelling

Understanding user utility

$$u_U(q, x, a; \theta) = \boxed{u^{CS}(q, a; \theta)} + \boxed{\overset{\approx\, \theta^j = \Pr[\text{like content } j]}{u^{Rec}(x; \theta)}}$$

During the Cold Start Phase (assume duration n rounds)…

exposure probabilities (chosen independently from RecSys)

3 factors affecting user utility at Cold Start

1. exposure rates: p^j
2. round interaction: $1/n$
3. consumption plan: a^j

exposed to content j w. prob. p^j

online with prob. $1/n$

plan to consume with prob. a^j

x^1 p^1

x^2 p^2

x^3 p^3

Poisson Consumption

$q^j \sim \pi(a^j) = Pois(p^j a^j)$

i) If $a^j \leq \theta^j \Rightarrow$ consume only q^j & get utility $q^j \cdot 1$

ii) If $a^j > \theta^j \Rightarrow \Pr[\text{like } j \,|\, \text{consume } j] = \theta^j / a^j$. Utility = $q^j \cdot \left(1 \cdot \frac{\theta^j}{a^j}\right) + (-1) \cdot \left(1 - \frac{\theta^j}{a^j}\right)$

$$u^{CS} = \sum_{j \in [d]} p^j (2 \min\{\theta^j, a^j\} - a^j)$$

Interventions

1. Recommendation choice intervention: ***over-representing minorities***.

2. Information design intervention: ***automatic incognito mode***.

Would your behavior on TikTok change in an incognito mode that does not log responses?

Incognito Mode Coding	Participant Count
no change: no reason	45
no change: less personalization	14
change: click “avoided” content	10
change: click “feed-clogging” content	9
change: exploration increase	9
other	8

Interventions

1. Recommendation choice intervention: ***over-representing minorities***.

2. Information design intervention: ***automatic incognito mode***.

3. Information gathering intervention: ***Cold Start improvement***