Yes: Affirmative Consent as a Theoretical Framework for Understanding and Imagining Social Platforms

Jane Im @im__jane

Jill Dimond, Melody Berton, Una Lee,

Katherine Mustelier, Mark Ackerman, Eric Gilbert

Project website: https://consentful.systems/

SCHOOL OF INFORMATION
UNIVERSITY OF MICHIGAN

EECS ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

sassafras
tech collective

AND
ALSO
TOO

Content Warning:

Mention of sexual abuse

The current social internet is plagued with problematic interactions that violate people's consent.

And they disproportionately impact marginalized groups.

Online harassment

Cyberstalking

Hate speech

Online image abuse

How can we protect people’s consent on the social internet?

Online harassment

Cyberstalking

Hate speech

Online image abuse

Illustration by Cynthia Zhang

For decades, feminist scholars and activists have defined consent to prevent sexual assaults.

Illustration by Cynthia Zhang

How can the theoretical framework of ***affirmative consent*** can be used to understand and build a safer social internet?

Affirmative Consent

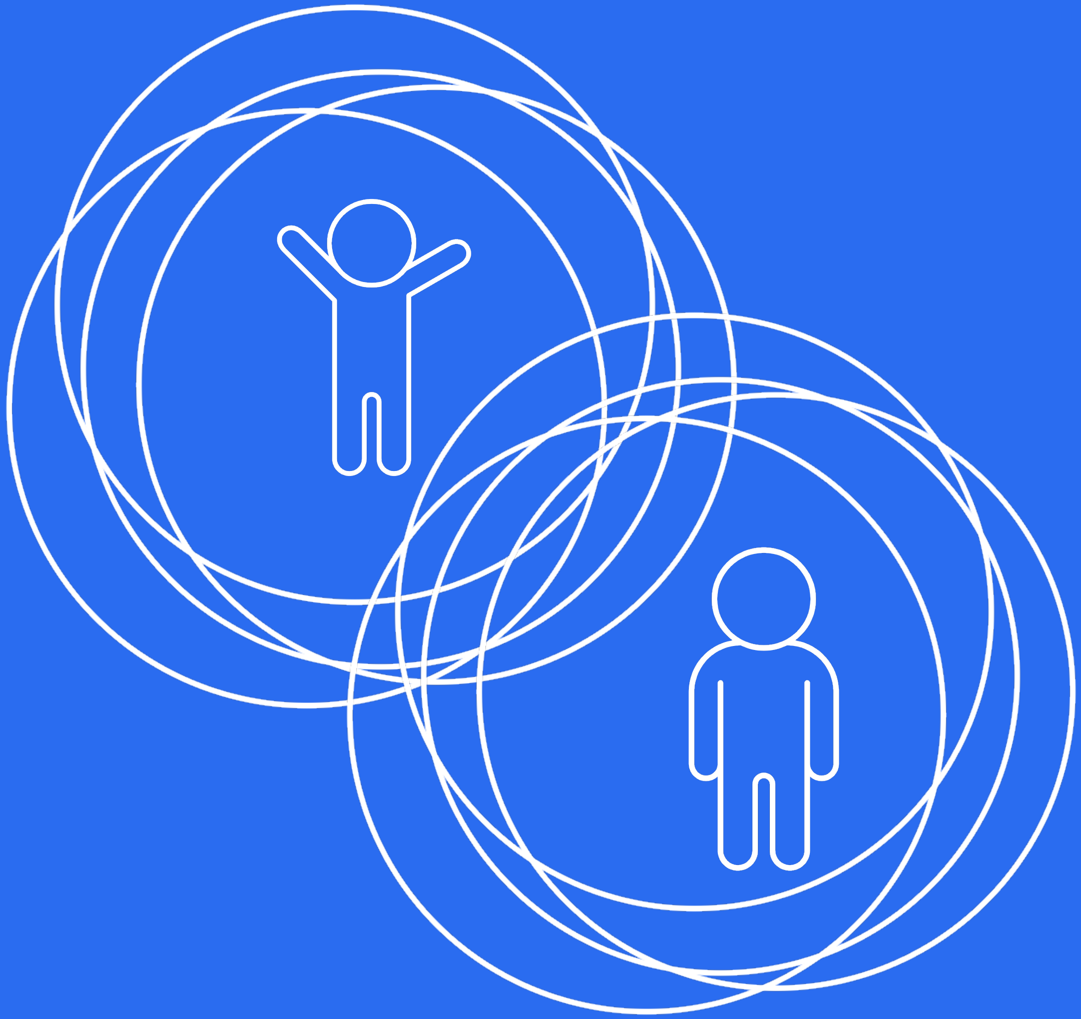
- An important feminist movement in the U.S. to prevent sexual assaults (“Yes means yes!”)

- Someone must ask for, and earn, *enthusiastic* approval *before* interacting with another person

Jaclyn Friedman and Jessica Valenti. 2019. *Yes means yes!: Visions of female sexual power and a world without rape.* Seal Press.

Noah Hilgert. 2016. The Burden of Consent: Due Process and the Emerging Adoption of the Affirmative Consent Standard in Sexual Assault Laws. *Ariz. L. Rev.* 58 (2016), 867

Deriving from feminist, legal, and HCI literature in the context of social platforms, we defined the concepts of affirmative consent as...



Affirmative consent is...

Voluntary:

Consent is an agreement that is 1) freely given and 2) enthusiastic

Informed

Revertible

Specific

Unburdensome

Affirmative consent is...

Voluntary

Informed:

People can only consent to an interaction after being given correct information about it—in an accessible way.

Revertible

Specific

Unburdensome

Affirmative consent is...

Voluntary

Informed

Revertible:

Consent can be revoked at any time.

Specific

Unburdensome

Affirmative consent is...

Voluntary

Informed

Revertible

Specific:

People should be able to consent to a particular action (or a particular person), and not a series of actions or people.

Unburdensome

Affirmative consent is...

Voluntary

Informed

Revertible

Specific

Unburdensome:

The costs associated with refusing to consent should not be so high that a person gives in and says “yes” when they would rather say “no.”

How can this theoretical framework of affirmative consent be used to understand and design the social internet?

1) Affirmative consent is explanatory

- Lets us pinpoint *which property* is being violated in design (e.g., specific)

- Provides a way for researchers and designers to *systematically dissect* a wide range of social computing problems

1) Affirmative consent is explanatory

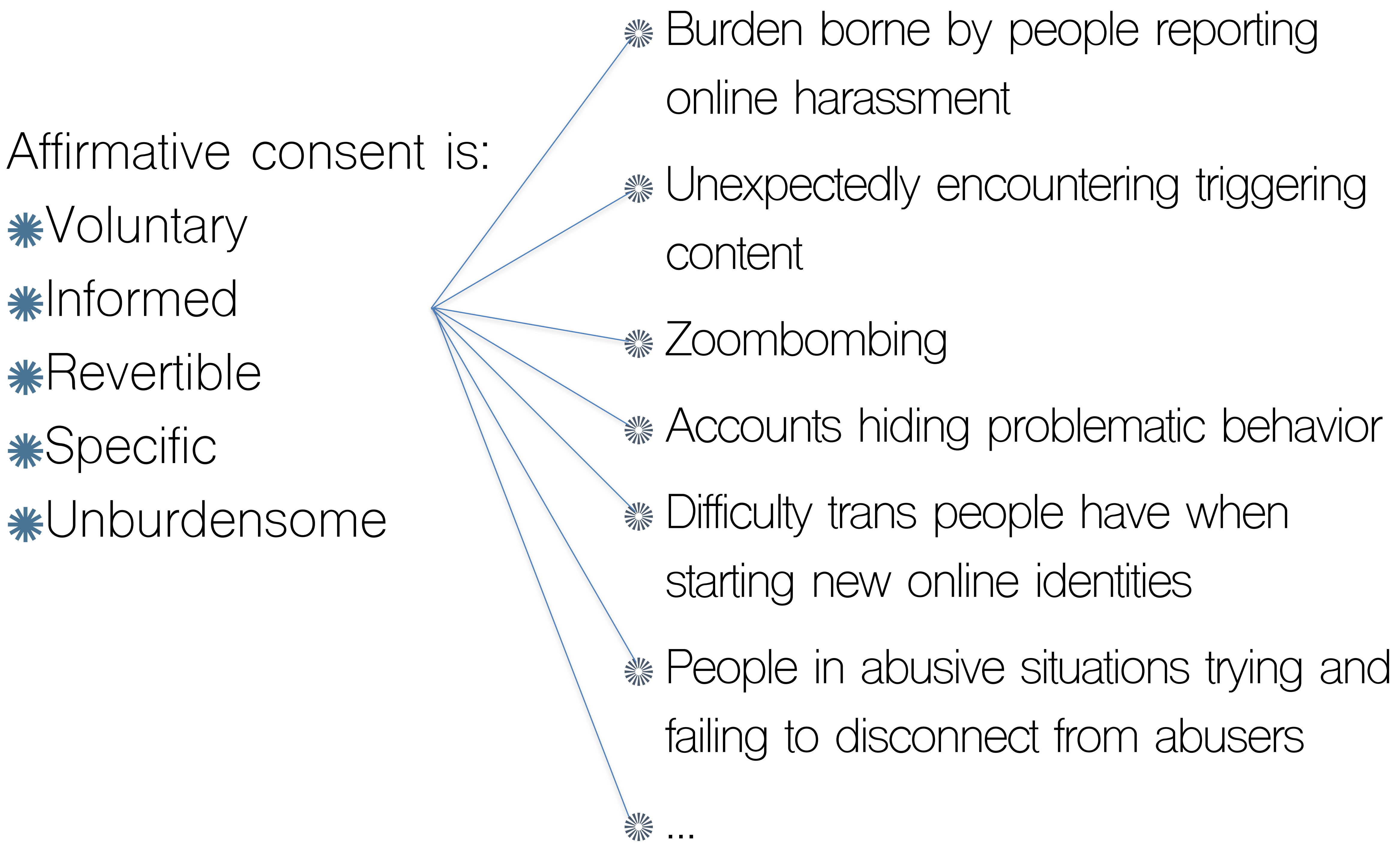
Example:

Problem with content feed algorithms

Eslami, Motahhare, et al. "First I 'like' it, then I hide it: Folk Theories of Social Feeds." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016.

Core issue: Users cannot signal their enthusiastic agreement to posts/accounts showing up in feeds (voluntary).

1) Affirmative consent is explanatory



Affirmative consent is:

- Voluntary
- Informed
- Revertible
- Specific
- Unburdensome

- Burden borne by people reporting online harassment
- Unexpectedly encountering triggering content
- Zoombombing
- Accounts hiding problematic behavior
- Difficulty trans people have when starting new online identities
- People in abusive situations trying and failing to disconnect from abusers
- ...

2) Affirmative consent is generative

Helps us generate design ideas grounded in consent.

If we build a totally new social platform grounded in consent, how would it look like?

	Voluntary	Informed	Revertible	Specific	Unburdensome
DM + group chat					
Profile					
Friend + Follow					
Post + comment					
Feed					
Tag					
Share + retweet					

	Voluntary	Informed	Revertible	Specific	Unburdensome
DM + group chat					
Profile					
Friend + Follow					
Post + comment					
Feed					
Tag					
Share + retweet					

	Voluntary	Informed	Revertible	Specific	Unburdensome
DM + group chat	Users are asked if they want to join when invited to group chat.	Platform visualizes topics discussed in group chat before a person decides to enter.	Users can revert message read status to unread.	Different online status by group: would love to chat for friends; online, but busy for others.	Classify DMs from strangers using sender's content and behavior.
Profile	Users can control profile visibility by audience: only show selfies to friends & friends' friends.	Platform shows how many people that viewed the profile are strangers.	Users can query and delete, en masse, tags and comments from their profile related to account (e.g., ex-partner).	Some profile fields are only shown to accounts that have been friends for > *t* time.	Platform periodically reminds user how their profile looks to other people: "This is how your profile looks to Jake."
Friend + Follow	Users can accept a friend request but can isolate it, sending it to a separate queue. (e.g., if acceptance is coerced).	Platform alerts if friend request comes from account with history of posting toxic content.	Requests from people previously unfriended are sent to a queue. — ensuring revert.	Assign people to "circles" at follow time with rules: no tags from this circle.	Periodic reviews of followers/friends with new risk scores (e.g. toxicity level).
Post + comment	*most platforms already support voluntary posting and commenting	Users receive reports of how many post viewers are strangers.	Users can query and delete posts/comments at large scale.	Users can apply audience rules to hashtags: e.g, creator can restrict who can use it.	Users can rate limit comments per post.
Feed	Feed asks what users want to see today (or this week).	Content feed makes algorithms visible and salient.	Users can bookmark feed settings to easily revert to prior settings.	Users can set different types of content feeds per social circle.	Users can annotate posts in feed, from which the system can learn what posts the person wants to see (or not see).
Tag	By default, platform always asks user if they consent to being tagged when another user initiates tagging.	Platform provides high-level summary of audience, outside friends, that sees tagged post.	If user unfriends, the system asks if they also want to delete tags of the person.	Users set tagging rules by content type: disallow tags in photos of people.	Users can timebox tag frequency: Jake can only tag once a month.
Share + retweet	Users can limit how many hops shares are allowed to travel.	Users are notified if post is shared to a new network "neighborhood."	When user deactivates post's sharing, or deletes the post, existing shares disappear. *twitter partially implements this	Leveraging data of past interactions, users can decide who can share each post: Only people who I have messaged 5 times can share.	Platform alerts user if their post starts being shared rapidly by strangers.

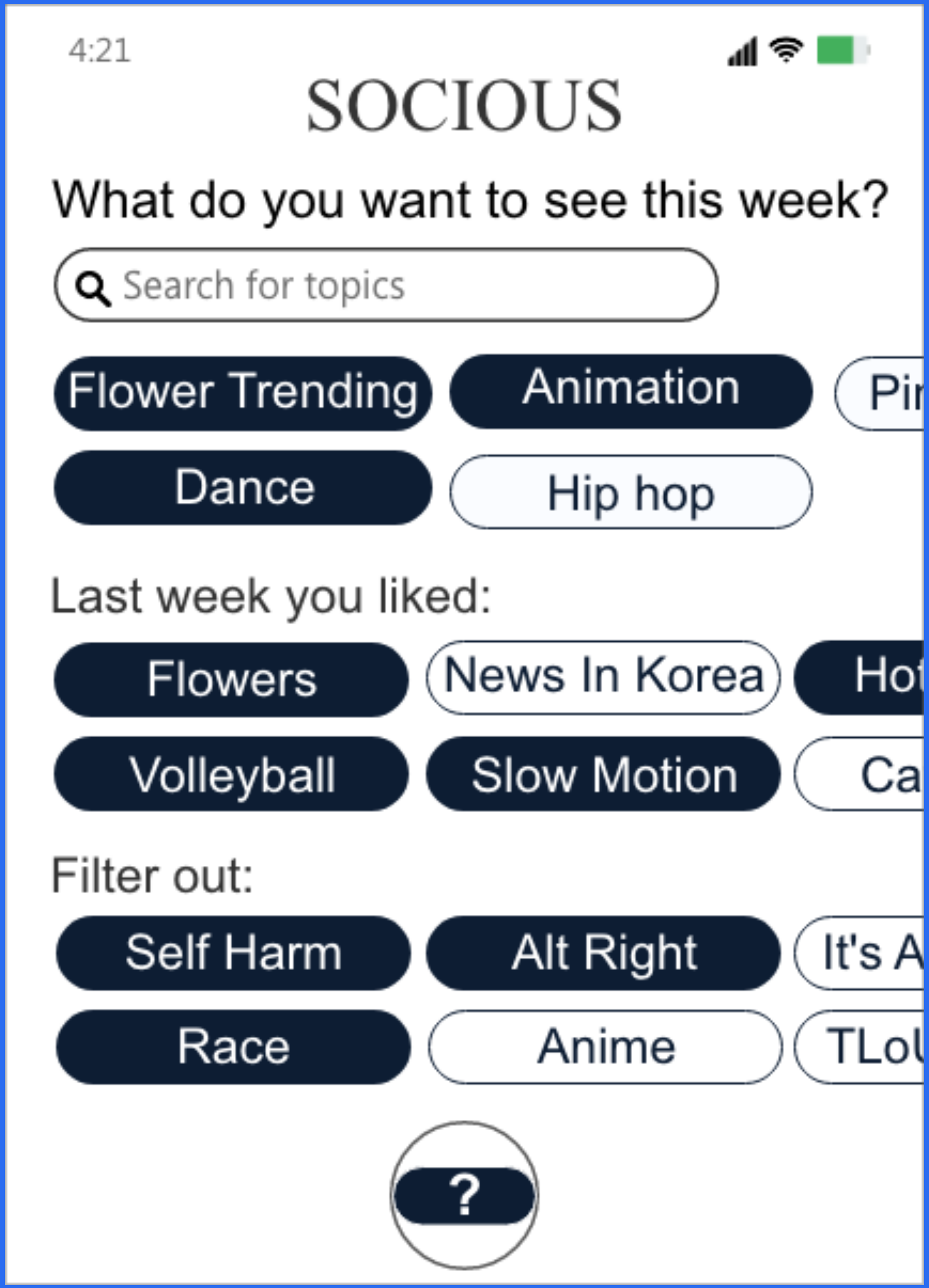
	Voluntary	Informed	Revertible	Specific	Unburdensome
DM + group chat	Users are asked if they want to join when invited to group chat.	Platform visualizes topics discussed in group chat before a person decides to enter.	Users can revert message read status to unread.	Different online status by group: would love to chat for friends; online, but busy for others.	Classify DMs from strangers using sender's content and behavior.
Profile	Users can control profile visibility by audience: only show selfies to friends & friends' friends.	Platform shows how many people that viewed the profile are strangers.	Users can query and delete, en masse, tags and comments from their profile related to account (e.g., ex-partner).	Some profile fields are only shown to accounts that have been friends for > *t* time.	Platform periodically reminds user how their profile looks to other people: "This is how your profile looks to Jake."
Friend + Follow	Users can accept a friend request but can isolate it, sending it to a separate queue. (e.g., if acceptance is coerced).	Platform alerts if friend request comes from account with history of posting toxic content.	Requests from people previously unfriended are sent to a queue. — ensuring revert.	Assign people to "circles" at follow time with rules: no tags from this circle.	Periodic reviews of followers/friends with new risk scores (e.g. toxicity level).
Post + comment	*most platforms already support voluntary posting and commenting	Users receive reports of how many post viewers are strangers.	Users can query and delete posts/comments at large scale.	Users can apply audience rules to hashtags: e.g, creator can restrict who can use it.	Users can rate limit comments per post.
Feed	Feed asks what users want to see today (or this week).	Content feed makes algorithms visible and salient.	Users can bookmark feed settings to easily revert to prior settings.	Users can set different types of content feeds per social circle.	Users can annotate posts in feed, from which the system can learn what posts the person wants to see (or not see).
Tag	By default, platform always asks user if they consent to being tagged when another user initiates tagging.	Platform provides high-level summary of audience, outside friends, that sees tagged post.	If user unfriends, the system asks if they also want to delete tags of the person.	Users set tagging rules by content type: disallow tags in photos of people.	Users can timebox tag frequency: Jake can only tag once a month.
Share + retweet	Users can limit how many hops shares are allowed to travel.	Users are notified if post is shared to a new network "neighborhood."	When user deactivates post's sharing, or deletes the post, existing shares disappear. *twitter partially implements this	Leveraging data of past interactions, users can decide who can share each post: Only people who I have messaged 5 times can share.	Platform alerts user if their post starts being shared rapidly by strangers.

	Voluntary	Informed	Revertible	Specific	Unburdensome
DM + group chat	Users are asked if they want to join when invited to group chat.	Platform visualizes topics discussed in group chat before a person decides to enter.	Users can revert message read status to unread.	Different online status by group: would love to chat for friends; online, but busy for others.	Classify DMs from strangers using sender's content and behavior.
Profile	Users can control profile visibility by audience: only show selfies to friends & friends' friends.	Platform shows how many people that viewed the profile are strangers.	Users can query and delete, en masse, tags and comments from their profile related to account (e.g., ex-partner).	Some profile fields are only shown to accounts that have been friends for > *t* time.	Platform periodically reminds user how their profile looks to other people: "This is how your profile looks to Jake."
Friend + Follow	Users can accept a friend request but can isolate it, sending it to a separate queue. (e.g., if acceptance is coerced).	Platform alerts if friend request comes from account with history of posting toxic content.	Requests from people previously unfriended are sent to a queue. — ensuring revert.	Assign people to "circles" at follow time with rules: no tags from this circle.	Periodic reviews of followers/friends with new risk scores (e.g. toxicity level).
Post + comment	*most platforms already support voluntary posting and commenting	Users receive reports of how many post viewers are strangers.	Users can query and delete posts/comments at large scale.	Users can apply audience rules to hashtags: e.g, creator can restrict who can use it.	Users can rate limit comments per post.
Feed	Feed asks what users want to see today (or this week).	Content feed makes algorithms visible and salient.	Users can bookmark feed settings to easily revert to prior settings.	Users can set different types of content feeds per social circle.	Users can annotate posts in feed, from which the system can learn what posts the person wants to see (or not see).
Tag	By default, platform always asks user if they consent to being tagged when another user initiates tagging.	Platform provides high-level summary of audience, outside friends, that sees tagged post.	If user unfriends, the system asks if they also want to delete tags of the person.	Users set tagging rules by content type: disallow tags in photos of people.	Users can timebox tag frequency: Jake can only tag once a month.
Share + retweet	Users can limit how many hops shares are allowed to travel.	Users are notified if post is shared to a new network "neighborhood."	When user deactivates post's sharing, or deletes the post, existing shares disappear. *twitter partially implements this	Leveraging data of past interactions, users can decide who can share each post: Only people who I have messaged 5 times can share.	Platform alerts user if their post starts being shared rapidly by strangers.

	Voluntary	Informed	Revertible	Specific	Unburdensome
DM + group chat	Users are asked if they want to join when invited to group chat.	Platform visualizes topics discussed in group chat before a person decides to enter.	Users can revert message read status to unread.	Different online status by group: would love to chat for friends; online, but busy for others.	Classify DMs from strangers using sender's content and behavior.
Profile	Users can control profile visibility by audience: only show selfies to friends & friends' friends.	Platform shows how many people that viewed the profile are strangers.	Users can query and delete, en masse, tags and comments from their profile related to account (e.g., ex-partner).	Some profile fields are only shown to accounts that have been friends for > t time.	Platform periodically reminds user how their profile looks to other people: "This is how your profile looks to Jake."
Friend + Follow	Users can accept a friend request but can isolate it, sending it to a separate queue. (e.g., if acceptance is coerced).	Platform alerts if friend request comes from account with history of posting toxic content.	Requests from people previously unfriended are sent to a queue. — ensuring revert.	Assign people to "circles" at follow time with rules: no tags from this circle.	Periodic reviews of followers/friends with new risk scores (e.g. toxicity level).
Post + comment	*most platforms already support voluntary posting and commenting	Users receive reports of how many post viewers are strangers.	Users can query and delete posts/comments at large scale.	Users can apply audience rules to hashtags: e.g, creator can restrict who can use it.	Users can rate limit comments per post.
Feed	Feed asks what users want to see today (or this week).	Content feed makes algorithms visible and salient.	Users can bookmark feed settings to easily revert to prior settings.	Users can set different types of content feeds per social circle.	Users can annotate posts in feed, from which the system can learn what posts the person wants to see (or not see).
Tag	By default, platform always asks user if they consent to being tagged when another user initiates tagging.	Platform provides high-level summary of audience, outside friends, that sees tagged post.	If user unfriends, the system asks if they also want to delete tags of the person.	Users set tagging rules by content type: disallow tags in photos of people.	Users can timebox tag frequency: Jake can only tag once a month.
Share + retweet	Users can limit how many hops shares are allowed to travel.	Users are notified if post is shared to a new network "neighborhood."	When user deactivates post's sharing, or deletes the post, existing shares disappear. *twitter partially implements this	Leveraging data of past interactions, users can decide who can share each post: Only people who I have messaged 5 times can share.	Platform alerts user if their post starts being shared rapidly by strangers.

	Voluntary	Informed	Revertible	Specific	Unburdensome
DM + group chat	Users are asked if they want to join when invited to group chat.	Platform visualizes topics discussed in group chat before a person decides to enter.	Users can revert message read status to unread.	Different online status by group: would love to chat for friends; online, but busy for others.	Classify DMs from strangers using sender's content and behavior.
Profile	Users can control profile visibility by audience: only show selfies to friends & friends' friends.	Platform shows how many people that viewed the profile are strangers.	Users can query and delete, en masse, tags and comments from their profile related to account (e.g., ex-partner).	Some profile fields are only shown to accounts that have been friends for > *t* time.	Platform periodically reminds user how their profile looks to other people: "This is how your profile looks to Jake."
Friend + Follow	Users can accept a friend request but can isolate it, sending it to a separate queue. (e.g., if acceptance is coerced).	Platform alerts if friend request comes from account with history of posting toxic content.	Requests from people previously unfriended are sent to a queue. — ensuring revert.	Assign people to "circles" at follow time with rules: no tags from this circle.	Periodic reviews of followers/friends with new risk scores (e.g. toxicity level).
Post + comment	*most platforms already support voluntary posting and commenting	Users receive reports of how many post viewers are strangers.	Users can query and delete posts/comments at large scale.	Users can apply audience rules to hashtags: e.g, creator can restrict who can use it.	Users can rate limit comments per post.
Feed	Feed asks what users want to see today (or this week).	Content feed makes algorithms visible and salient.	Users can bookmark feed settings to easily revert to prior settings.	Users can set different types of content feeds per social circle.	Users can annotate posts in feed, from which the system can learn what posts the person wants to see (or not see).
Tag	By default, platform always asks user if they consent to being tagged when another user initiates tagging.	Platform provides high-level summary of audience, outside friends, that sees tagged post.	If user unfriends, the system asks if they also want to delete tags of the person.	Users set tagging rules by content type: disallow tags in photos of people.	Users can timebox tag frequency: Jake can only tag once a month.
Share + retweet	Users can limit how many hops shares are allowed to travel.	Users are notified if post is shared to a new network "neighborhood."	When user deactivates post's sharing, or deletes the post, existing shares disappear. *twitter partially implements this	Leveraging data of past interactions, users can decide who can share each post: Only people who I have messaged 5 times can share.	Platform alerts user if their post starts being shared rapidly by strangers.

Voluntary Content Feed

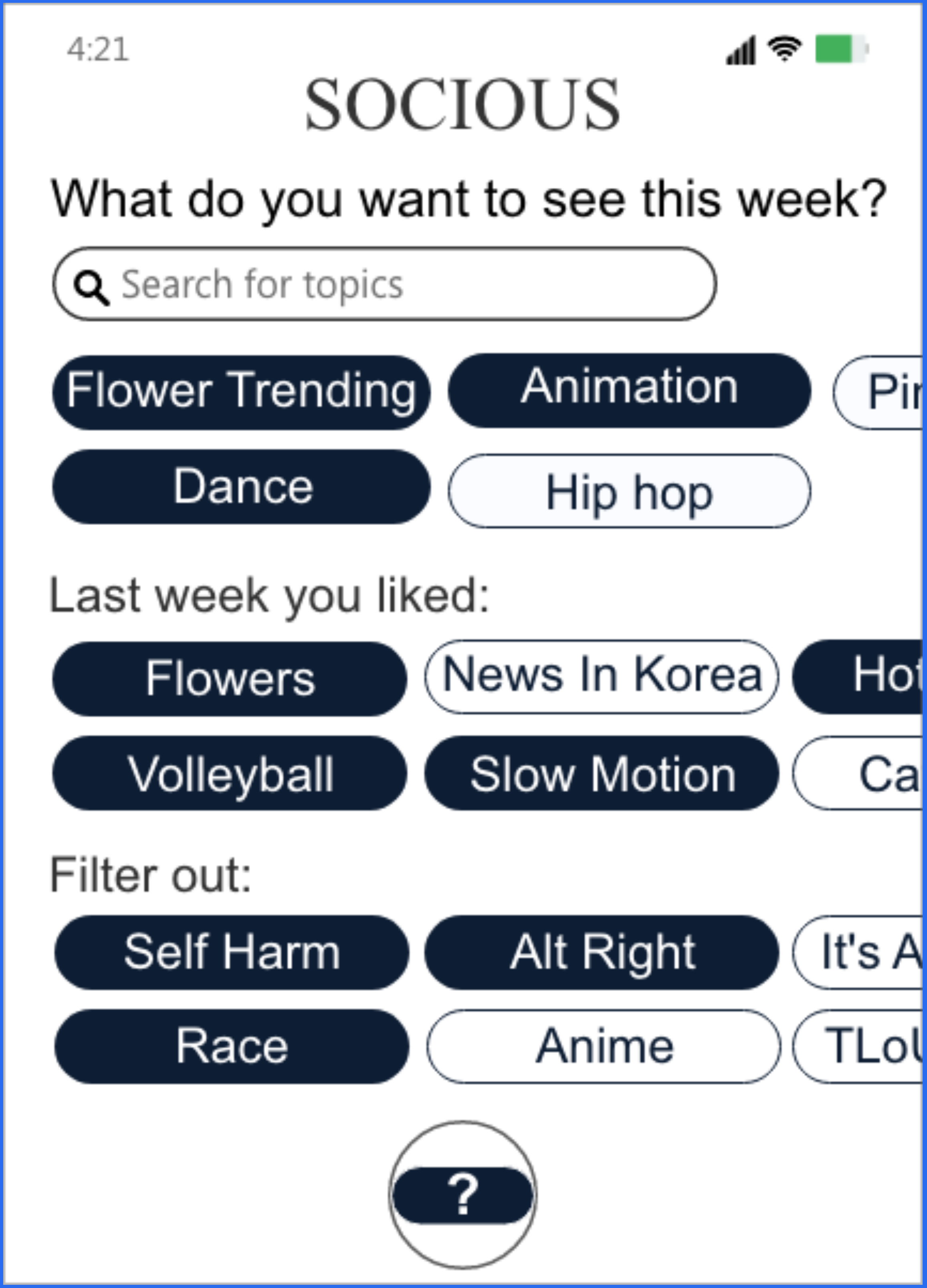


What do you want to see this week?

Last week you liked:

Filter out:

Voluntary Content Feed



4:21

SOCIOUS

What do you want to see this week?

Search for topics

Flower Trending | Animation | Pi
Dance | Hip hop

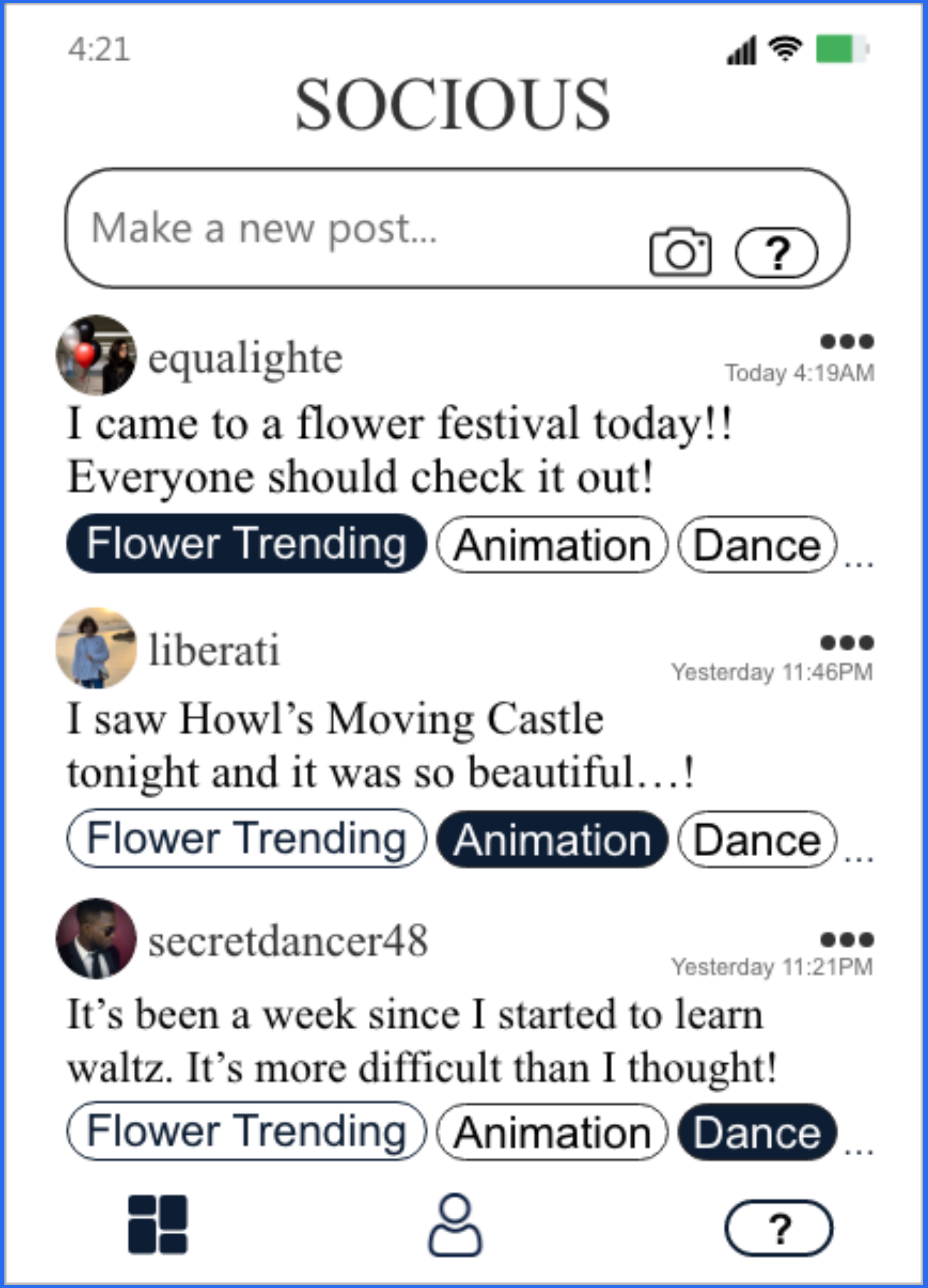
Last week you liked:

Flowers | News In Korea | Ho
Volleyball | Slow Motion | Ca

Filter out:

Self Harm | Alt Right | It's A
Race | Anime | TLo

?



4:21

SOCIOUS

Make a new post... ?

equalighte
Today 4:19AM
I came to a flower festival today!!
Everyone should check it out!
Flower Trending | Animation | Dance ...

liberati
Yesterday 11:46PM
I saw Howl's Moving Castle
tonight and it was so beautiful…!
Flower Trending | Animation | Dance ...

secretdancer48
Yesterday 11:21PM
It's been a week since I started to learn
waltz. It's more difficult than I thought!
Flower Trending | Animation | Dance ...

?

Reimagining and building new social platforms grounded in consent

Project website: https://consentful.systems/

Jane Im 🐦 @im__jane

Co-authors: Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark Ackerman, Eric Gilbert