

# Social Computing Capstone

## Day 12: Community Moderation

CSE 481p | Winter 2024

**Amy X. Zhang**

Instructor

Assistant Professor | University of Washington, Allen School of Computer Science & Engineering

**Ruotong Wang**

TA

Ph.D. Student | University of Washington, Allen School of Computer Science & Engineering

# Schedule for today's class

- Presentation and Q&A with Sanjay Kairam (20 min)
- Discussion of reading and lecture on community moderation (15 min)
- Work time on G5: Midterm Presentation on Thursday (40 min)

# Sanjay Kairam



I'm a Data/Research Scientist with deep academic and industry experience analyzing how users and information interact in social platforms.

Currently: Reddit  
Formerly: Twitch, Stanford (PhD)

# Community Moderation

# Last time:

**Commercial content moderation:** thousands of paid contractors who work for the platform reviewing content

vs.

**Algorithmic moderation:** AI systems trained on previously removed comments predict whether new comments should be removed



# Commercial content moderation

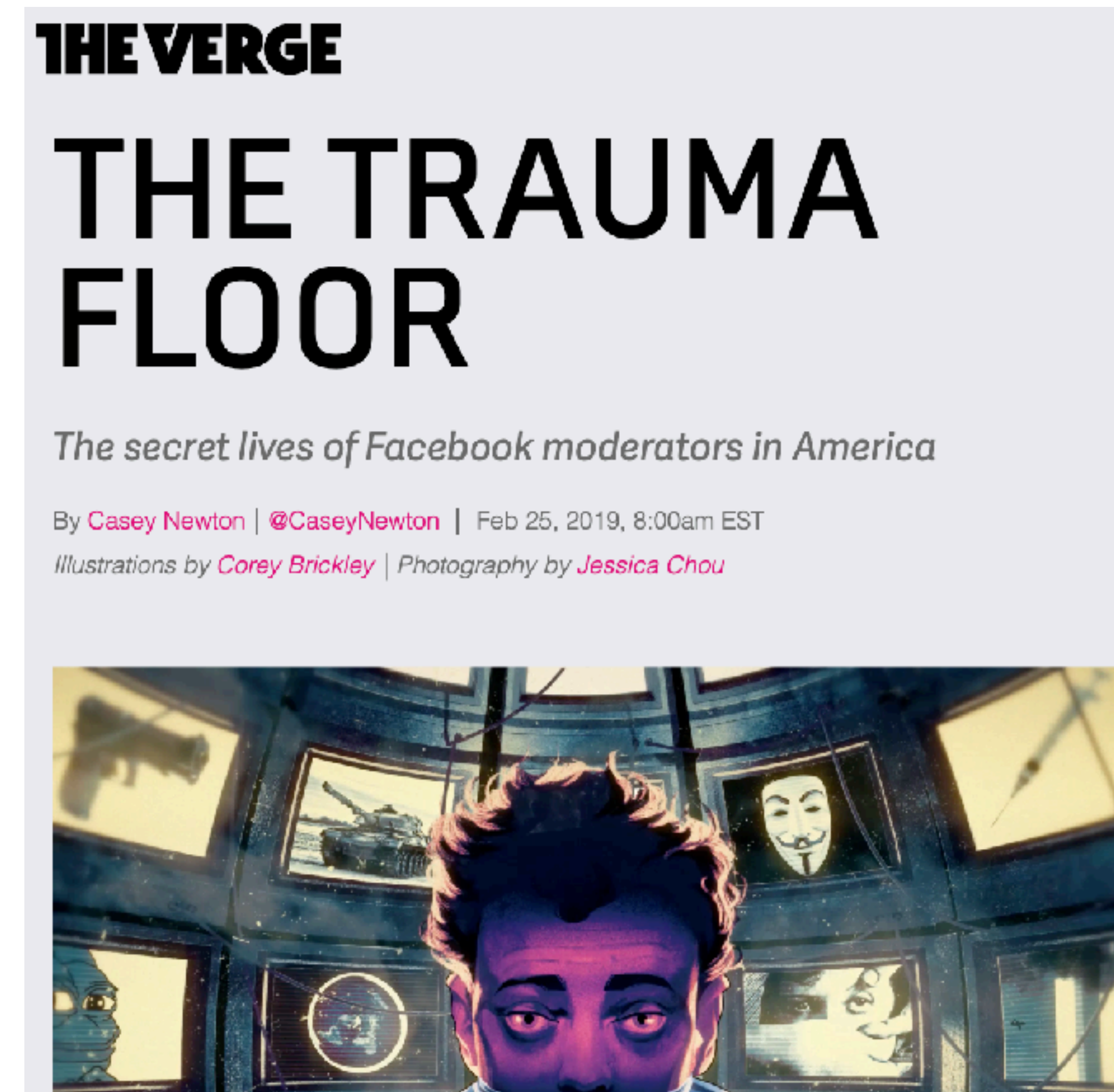
## Strengths

Trained reviewers check claims, which helps avoid brigading and supports more calibrated and consistent outcomes.

## Weaknesses

Major emotional trauma and PTSD for moderators.

Evaluators may have only seconds to make a snap judgment.



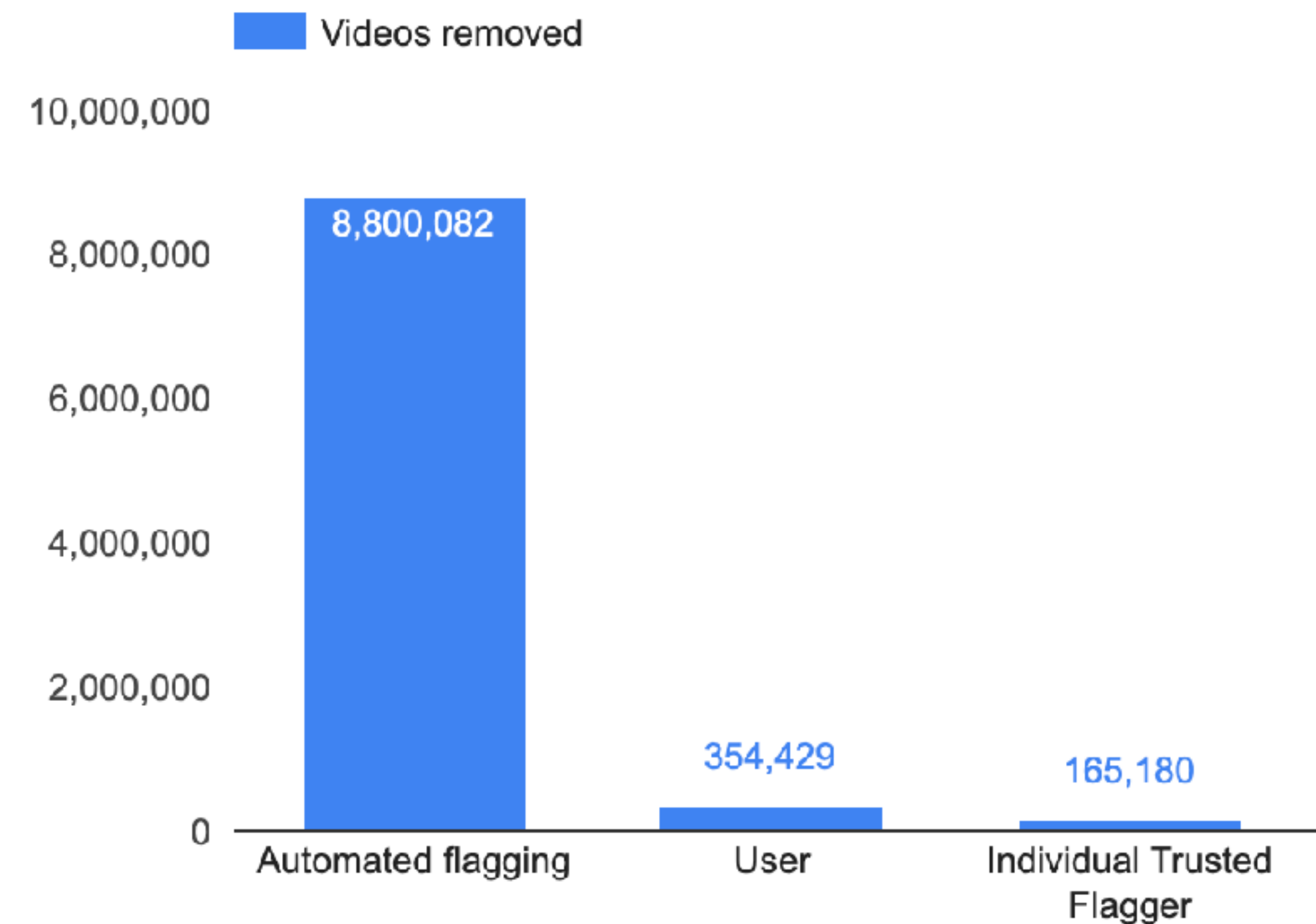
# Algorithmic moderation

Strengths: Can act quickly, before people are hurt by the content. Scales.

Weaknesses:

These systems make embarrassing errors or exhibit biases, often ones that the creators didn't intend. Errors are often interpreted as intentional platform policy.

Need to constantly be re-training and tweaking the models. Doesn't actually take commercial content moderation out of the picture (still need humans to label content).

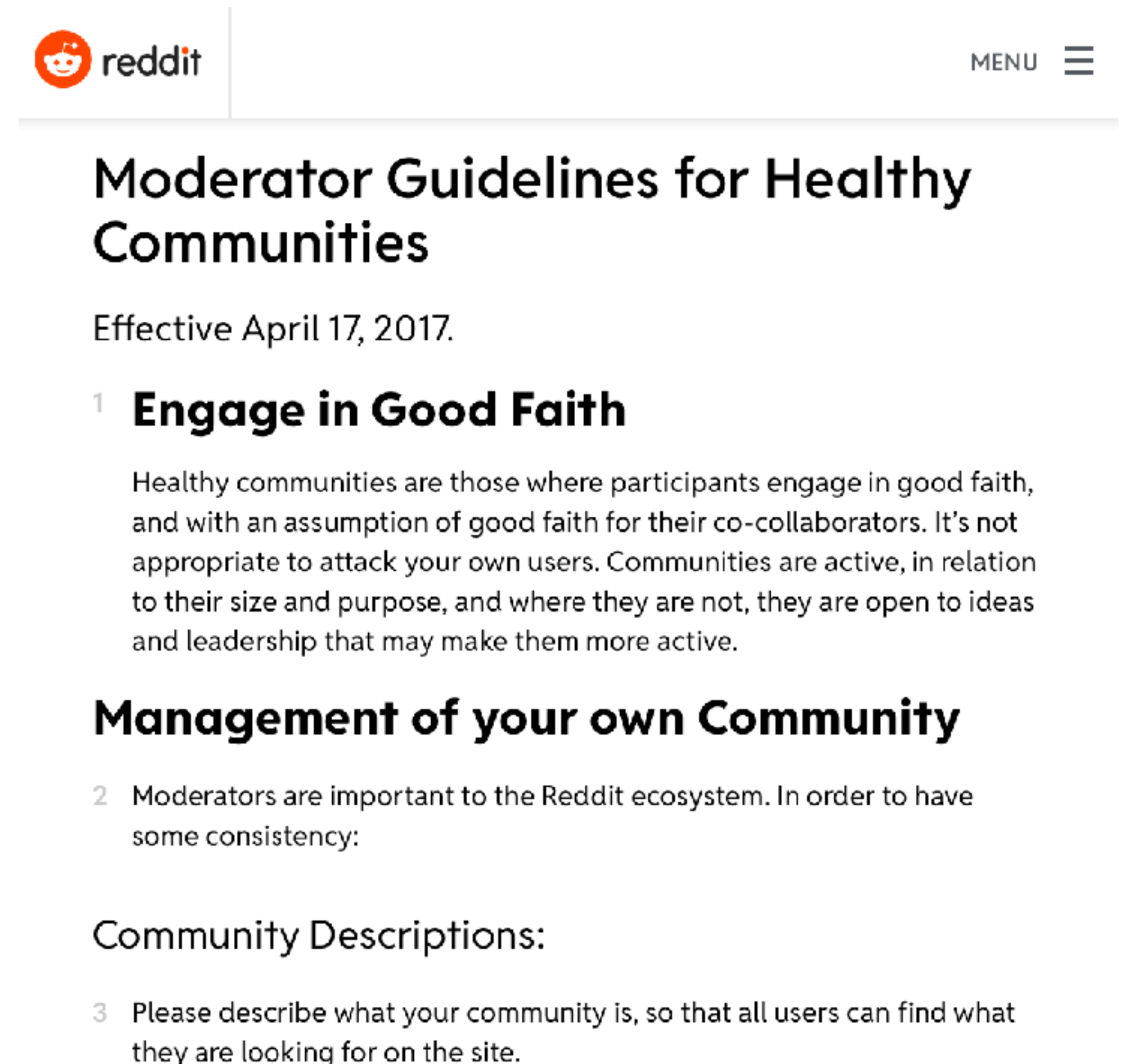


# Third way: Community moderation

Members of the community, or moderators who run the community, handle reports and proactively remove comments

Examples: Reddit, Twitch, Discord, HackerNews

It's best practice for the moderator team to publish their rules, rather than let each moderator act unilaterally



The screenshot shows the top of a Reddit page. The Reddit logo and name are in the top left, and a 'MENU' button with a hamburger icon is in the top right. The main heading is 'Moderator Guidelines for Healthy Communities', followed by the date 'Effective April 17, 2017.' The first section is '1 Engage in Good Faith', which explains that healthy communities engage in good faith and do not attack their own users. The second section is 'Management of your own Community', which states that moderators are important for consistency. The third section is 'Community Descriptions:', which asks users to describe their community for better findability.

reddit

MENU

## Moderator Guidelines for Healthy Communities

Effective April 17, 2017.

### 1 Engage in Good Faith

Healthy communities are those where participants engage in good faith, and with an assumption of good faith for their co-collaborators. It's not appropriate to attack your own users. Communities are active, in relation to their size and purpose, and where they are not, they are open to ideas and leadership that may make them more active.

### Management of your own Community

2 Moderators are important to the Reddit ecosystem. In order to have some consistency:

Community Descriptions:

3 Please describe what your community is, so that all users can find what they are looking for on the site.



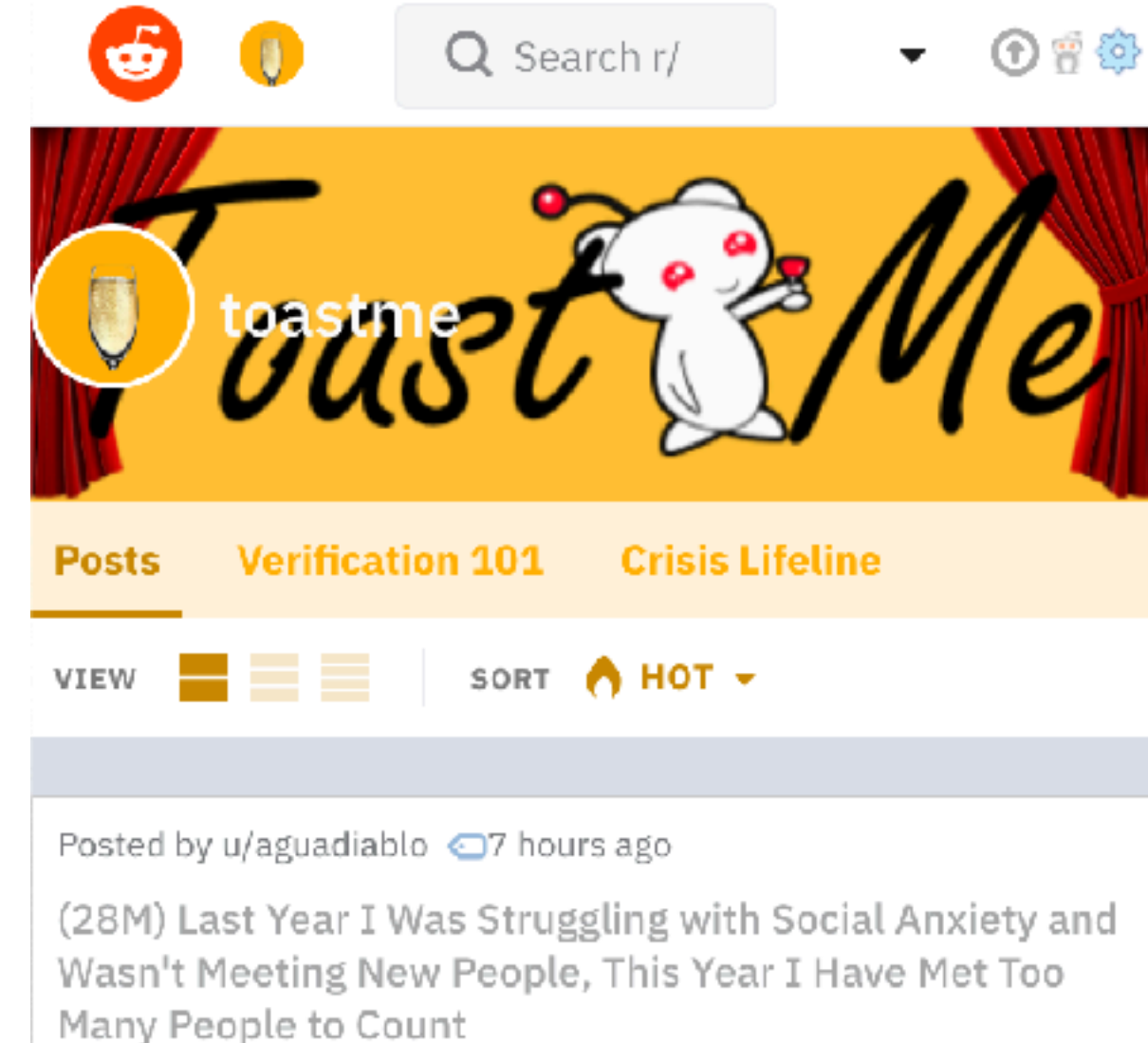
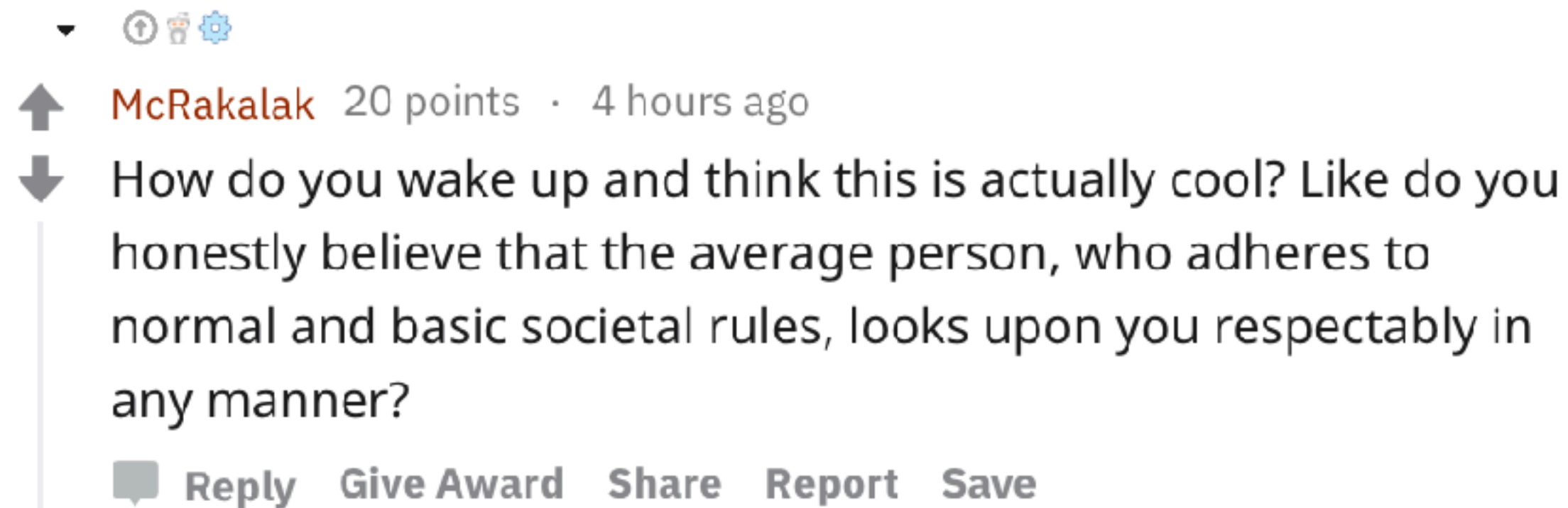
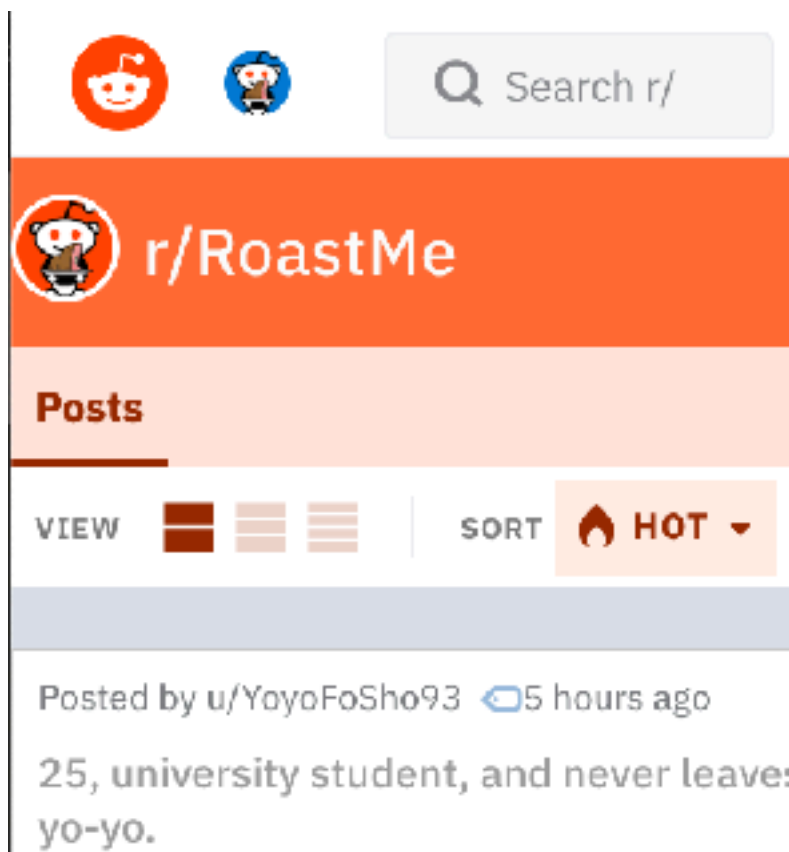
“I really enjoy being a **gardener** and cleaning out the bad weeds and bugs in subreddits that I’m passionate about. Getting rid of trolls and spam is a joy for me. When I’m finished for the day I can stand back and admire the clean and functioning subreddit, something a lot of people take for granted. I consider moderating a glorified **janitor’s** job, and there is a unique pride that janitors have.”

- /u/noeatnosleep, moderator on 60 subreddits

[<https://thebetterwebmovement.com/interview-with-reddit-moderator-unoeatnosleep>; Seering, Kaufman and Chancellor 2020; Matias 2019]

# Why community moderation?

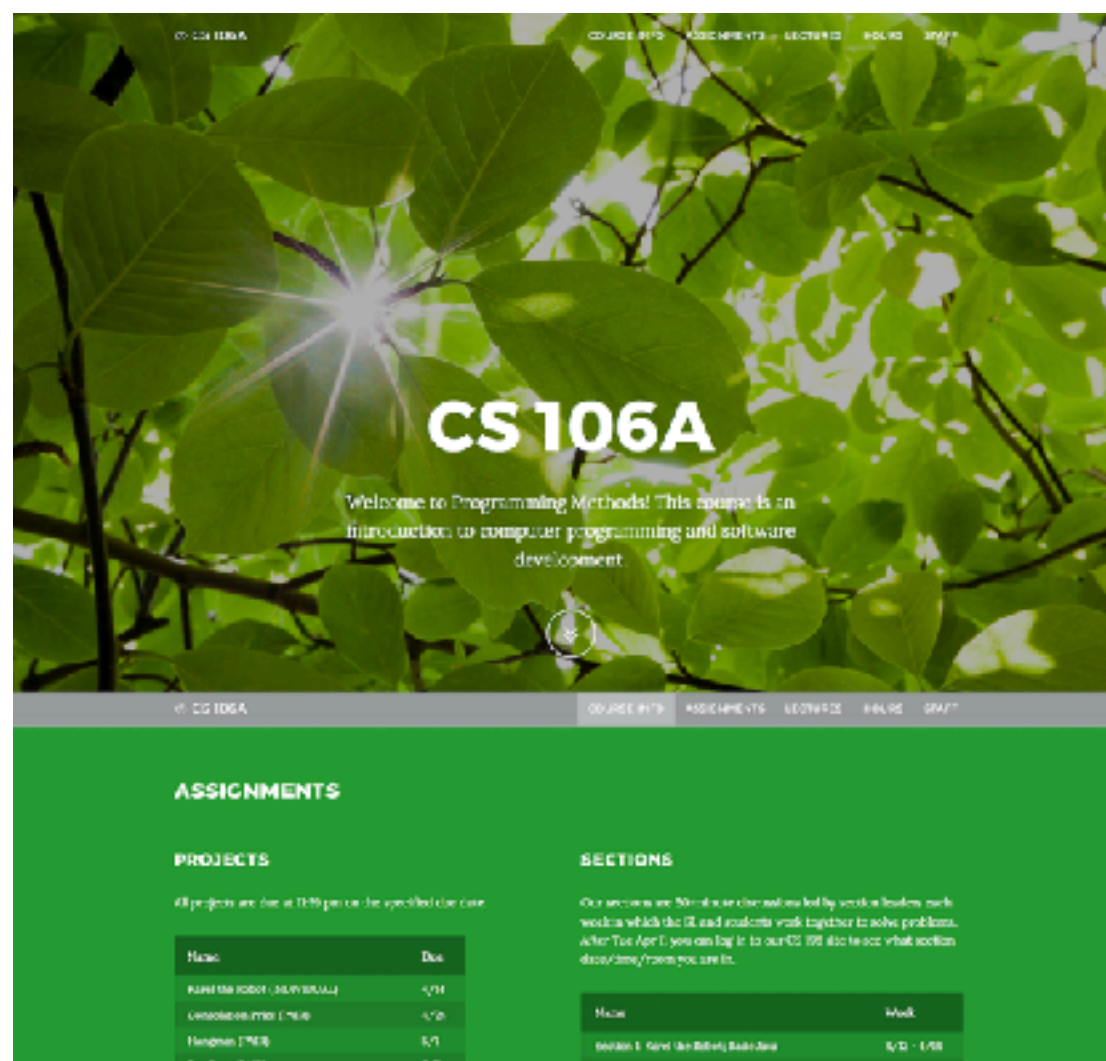
- Moderators can do a lot more than taking down content retroactively. They get to influence **community norms**.



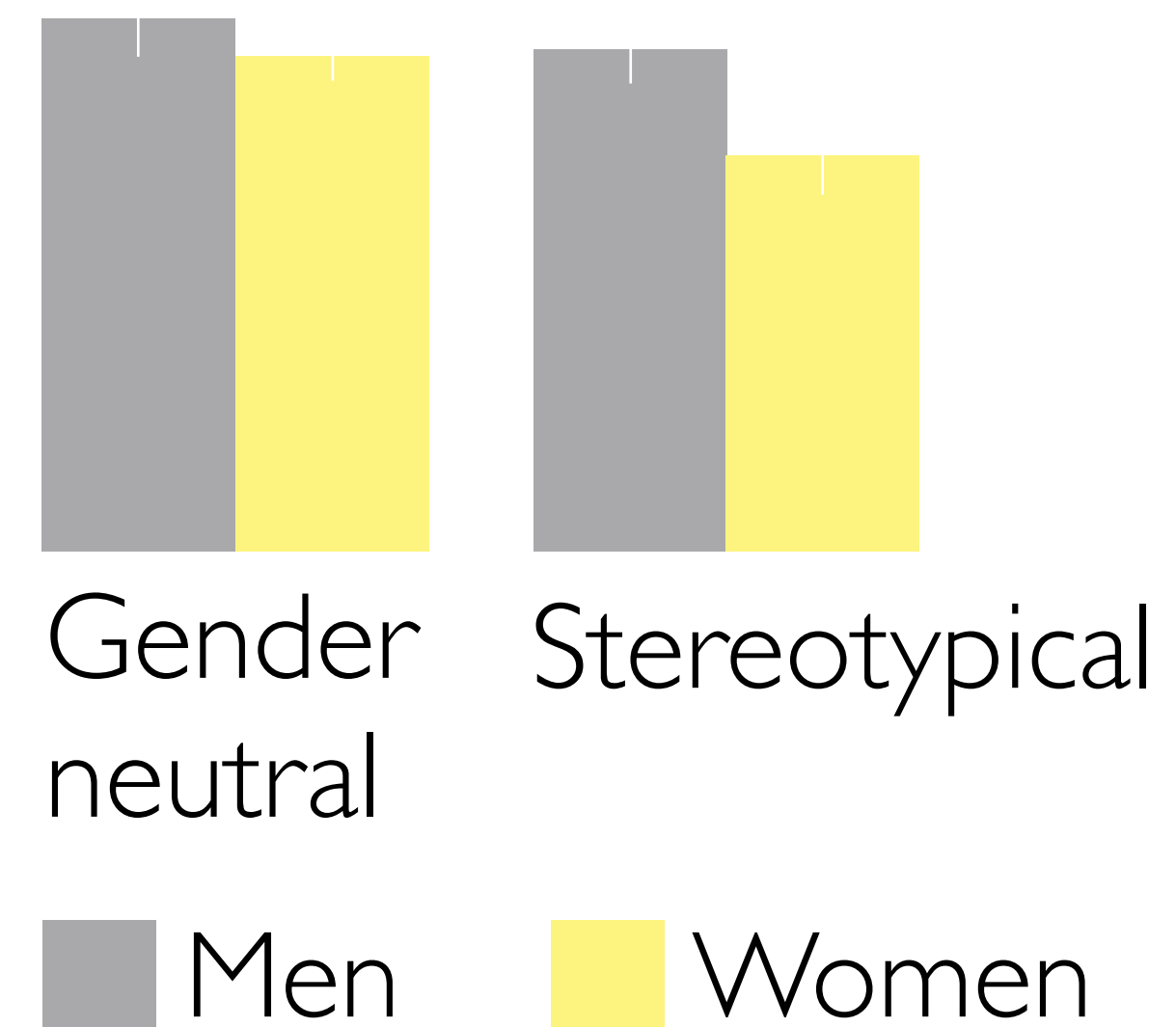


# Norms are incredibly important!

- Norms are the informal rules that govern behavior in groups and societies [Stanford Encyclopedia of Philosophy]
- They can be intuited quickly.
- They can be influenced by design.



Intent to enroll



[Metaxa et al. 2018]

# Descriptive Norms

- Norms can also be influenced by common behavior (descriptive norms).
- This is particularly the case for behavior by high status members of the community (moderators). Mods can model what is good behavior in a community.

# Is it the norms or the people?

[Rajadesingan, Resnick and Budak 2020]

Are community norms influenced more by the people who choose to join them, or by what we see in the space?

Comparing people before and after they joined 56 political subreddits with different levels of toxicity: **it's the norms.** People match toxicity levels with their first post in the community, differing from their prior behavior in other political subreddits.





## News of the Day

### I'm Voting for Hillary Because of My Daughter

Back in the 2008 primary season, I supported Hillary Clinton. That choice...

#### Top Comments Sorted by Best



**User1337** · 2 hours ago

I'm a woman, and i don't think you should vote for a woman just because she is a woman. vote for her because you believe she deserves it.

6 ^ | v · [Reply](#)



**User9054** · 3 hours ago

Personally, I'd vote for whoever I think is the best and

(Real comments on the article)

## News of the Day

### I'm Voting for Hillary Because of My Daughter

Back in the 2008 primary season, I supported Hillary Clinton. That choice...

#### Top Comments Sorted by Best



**User1337** · 2 hours ago

Oh yes. By all means, vote for a Wall Street sellout - - a lying, abuse-enabling, soon-to-be felon as our next President. And do it for your daughter. You're quite the role model.

1 ^ | v · [Reply](#)



**User9054** · 3 hours ago

Hillary is a cunt. I am voting with my dick for Putin. /s

Positive comments

Result: 35% troll comments

[Cheng et al. 2017]

Negative comments

Result: 47% troll comments  
(Relative increase of one third compared to the 35% baseline)

# Why community moderation?

- Moderators can also make their own rules. Different communities can have different rules (inviting pluralism!).
- Rules can evolve into entire systems of governance.
  - EN Wikipedia has 100s of policy pages and >1000 admins. It has courts, mediation, bureaucrats, a Supreme Court...
  - Other language Wikipedias can be very...different

## Shock an aw: US teenager wrote huge slice of Scots Wikipedia

**Nineteen-year-old says he is 'devastated' after being accused of cultural vandalism**

📷 Scots, the language of Robert Burns, has been enjoying a resurgence. Photograph: S Vincent/Alamy

The Scots **Wikipedia** entry on the Canada goose - or "Canadae guiss" - was at first honest about its provenance. A tag warned: "The 'Scots' that wis uised in this airticle wis written bi a body that's mither tongue isna Scots. Please impruive this airticle gin ye can."

But, as the author grew in confidence, so he removed the caveat, and continued on his Scots-writing spree.

Now an American teenager - who does not speak Scots, the language of Robert Burns - has been revealed as responsible for almost half of the entries on the Scots language version of Wikipedia.

# What's not so great about community moderation?

**Invisible labor** is a term drawn from studies of women's unpaid work in managing a household, emphasizing that what the women do is labor in the traditional sense, but is not recognized or compensated as such.

[Star and Strauss 1999]

# Why is the labor invisible?

Because oftentimes people see just the **results** of the moderation, not the work involved behind the scenes.

The invisible nature of this labor makes moderation feel thankless, and the content that mods face can prompt **burnout**. A second component of this is **emotional** labor, or labor in which you must manage and perform emotions, which can also add to the burnout.

In addition, community moderators get little support from platforms despite enriching them.



# MODERATING ON DISCORD

Discord is a place where anyone can build and manage a community dedicated to the things they love, whether that is a favorite game, creating amazing art, or simply hanging out with friends and making new ones. Moderators are at the forefront of creating spaces where people feel safe and can find belonging. Moderators are a key part of making communities great and a place where people want to gather.

Moderation takes hard work, and a commitment to learning more about how to make communities better. A key part of learning how to moderate comes simply from building and managing a community, but also from sharing knowledge learned by others sharing their insights on how to moderate communities more effectively.

We built the Discord Moderator Academy as a comprehensive resource so that anyone, from first-time moderators to experienced veterans of massive online communities, can find resources to learn about moderation, community management, and more.





# Community moderation

## Strengths:

- Leverages intrinsic motivation

- Local experts are more likely to have context to make hard calls

- Mods have more levers, time, and social standing to influence norms

- A Plurality of different spaces with different norms and rules

## Weaknesses:

- Mods don't feel they get the recognition they deserve + burnout + unpaid labor

- Not necessarily consistent across a platform

- Without broader oversight, mods can grow problematic communities

# Assignment G5: Midterm Presentation (Thurs)

- Each group will have 5 minutes to present and 3-4 min for Q&A
- You should cover:
  - A short pitch of your project + your tentative product name
  - Plan for what you'll build, your current progress, and proposed timeline
  - Feedback you'd like to get and questions you have for the class