

IN4MATX 153: CSCW

Class 11:
Content Moderation

Professor Daniel Epstein
TA Dennis Wang
Reader Weijie Du

Announcements

- No class next Monday, May 13
 - Guests needed to reschedule, will come on May 20 instead
 - I'm traveling, more on that next time
- A2 due Wednesday
 - I need to leave my office hours after 30 minutes, so come early or come to Dennis' hours today
- A3 will probably be posted on Thursday

Today's goals

By the end of today, you should be able to...

- Articulate what makes moderating online content challenging
- Differentiate the advantages and challenges around algorithmic, community-led, and paid approaches to moderation
- Explain the nuances of when and for whom moderation works, and does not work

A story of Facebook's content moderation

<https://radiolab.org/podcast/post-no-evil/transcript>

Facebook's content moderation



No pornography.

What counts as pornography?

Hmm. No nudity.

But then... what's actually nudity?
And what's not? What's the rule?

No visible male or female genitalia. And no exposed female breasts.

Facebook's content moderation



Sign in
[Contribute →](#)

The Guardian

News | Opinion | Sport | Culture | Lifestyle

US World Environment Soccer US Politics Business Tech Science More

Facebook

Mums furious as Facebook removes breastfeeding photos

Mark Sweney

Twitter @marksweney Email
Tue 30 Dec 2008 08.17 EST

[f](#) [Twitter](#) [Email](#)

6 130

[Facebook](#) has become the target of an 80,000-plus protest by irate mothers after banning breastfeeding photographs from online profiles.

Facebook's policy, which bans any breastfeeding images uploaded that show nipples, has led an online protest by lactivists - called "Hey Facebook, breast feeding is not obscene".

Facebook's content moderation



<https://radiolab.org/podcast/post-no-evil/transcript>

Ok, fine. Nudity means that the nipple and areola are visible. Breastfeeding blocks those.

PARENTS 16/03/2015 10:12 GMT | Updated 30/03/2015 11:59 BST

Facebook Clarifies Nudity Policy: Breastfeeding Photos Are Allowed (As Long As You Can't See Any Nipples)

Rachel Moss
The Huffington Post UK

As the public breastfeeding debate rages on, Facebook have updated their nudity policy to clarify their stance on breastfeeding photos.

'Brelfies' (that's breastfeeding selfie for the uninitiated) are permitted on the site, as long as they do not show the mother's nipple,

Facebook's content moderation



Ok, fine. Nudity means that the nipple and areola are visible. Breastfeeding blocks those.

Moms still pissed: their pictures of them holding their sleeping baby after breastfeeding get taken down.

Wait but that's not breastfeeding!

Facebook's content moderation



Nevermind. It's nudity and disallowed unless the baby is actively nursing.

A screenshot of a ZDNet news article. At the top, there is a dark navigation bar with the ZDNet logo, a search icon, a 'MENU' link, a user icon, and a 'US' link. Below the navigation, a light gray sidebar contains the text: 'MUST READ: SHA-1 collision attacks are now actually practical and a looming danger'. The main content area has a black header with the title 'Facebook clarifies breastfeeding photo policy' in white text. The main text of the article discusses Facebook's policy change regarding breastfeeding photos.

Facebook clarifies breastfeeding photo policy

Facebook has clarified its policy when it comes to photos of breastfeeding: only photos of babies actively nursing are allowed. Everything else is considered nudity and will be taken down if reported.



By Emil Protalinski for [Friending Facebook](#) | February 7, 2012 -- 11:54
GMT (03:54 PST) | Topic: [Social Enterprise](#)

Facebook's content moderation



OK, here's a picture of a woman in her twenties breastfeeding a teenage boy.

Wait, what?!? Ok, an age cap: only infants.

But, what's the line between an infant and a toddler?

If it looks big enough to walk on its own, then it's too old.

But the WHO says to breastfeed at least partially until two years old.

NOPE. Can't enforce it.

Facebook's content moderation



Ok, here's a new one: I've got this photo of a woman breastfeeding a goat.

...What?

It's a traditional practice in Kenya. If there's a drought, and a lactating mother, the mother will breastfeed the baby goat to help keep it alive.

... oh.

Facebook's content moderation

Radiolab quote on the rulebook:

- “*This is utilitarian document. It’s not about being right one hundred percent of the time, it’s about being able to execute effectively.*”

“Moderation is the most important commodity of any social computing system.”

Moderation and acceptability

- We will primarily focus on questions of whether content is harmful, vulgar, etc., and how moderators handle those decisions
- There are other decisions that moderators are forced to make, like relevance of content to the specific community
 - You post in the UCI subreddit about UCSD
 - You post an advertisement, which is expressly not allowed
 - You'll have to make these decisions, as well as others, in A3

**Discuss: any run-ins with moderation
when you posted your memes? Are any
of you mods on club Discords, etc.?**

What's been your experience with content moderation?

I've seen content that I would call inappropriate or offensive

0%

I've been a moderator

0%

I've had things that I've posted removed, whether inappropriate or not

0%

None of the above

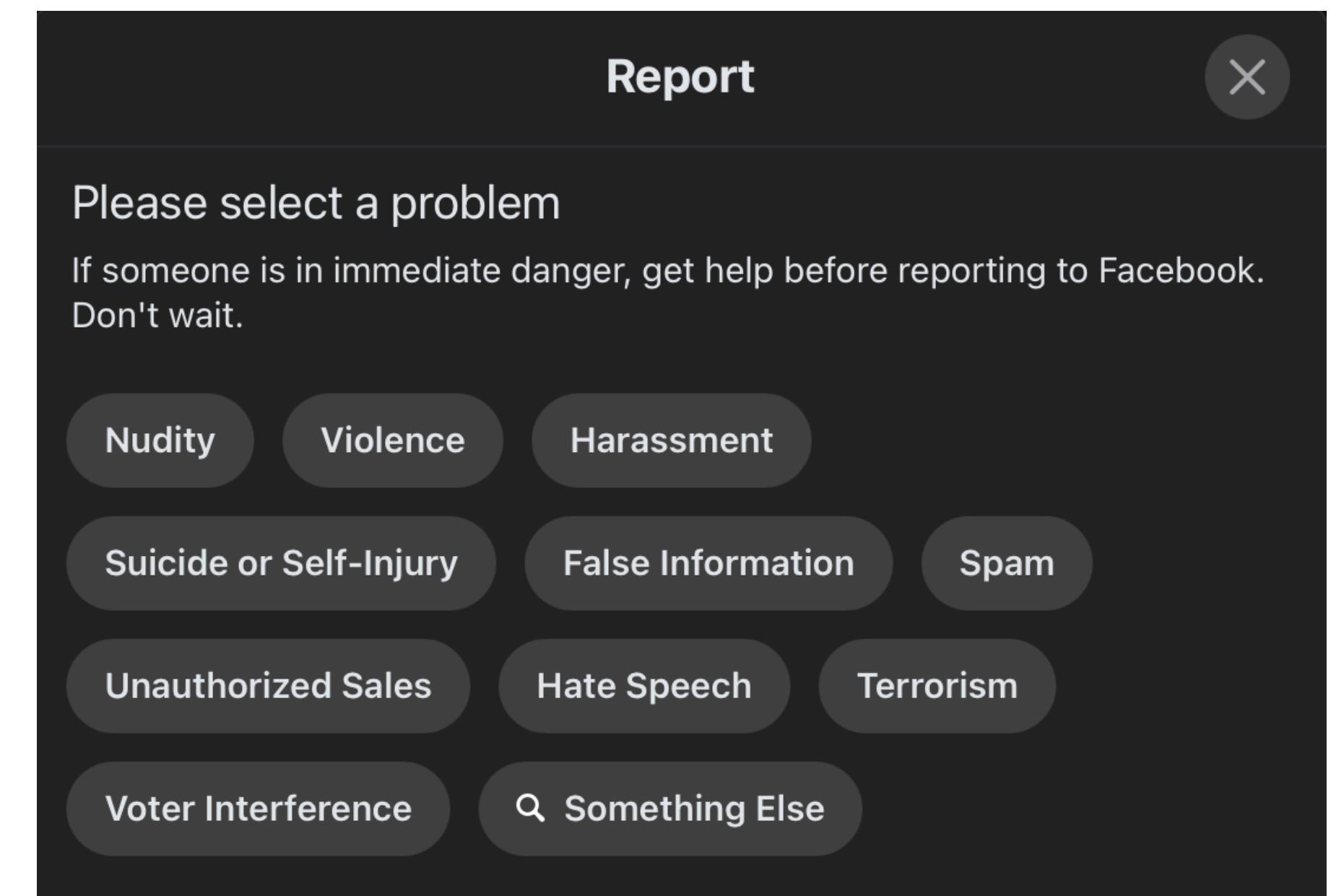
0%

Three imperfect solutions

- Paid moderation: thousands of paid contractors who work for the platform reviewing claims
- Community moderation: volunteers in the community take on the role of mods, remove comments, and handle reports
- Algorithmic moderation: AI systems trained on previously removed comments predict whether new comments should be removed
- Each with their pros and cons

Paid moderation

- Rough estimates:
 - ~15,000 contractors on Facebook and Instagram
 - ~10,000 contractors on YouTube
- Moderators at Meta are trained on over 100 manuals, spreadsheets and flowcharts to make judgments about flagged content



THE VERGE

THE TRAUMA FLOOR

The secret lives of Facebook moderators in America

By Casey Newton | @CaseyNewton | Feb 25, 2019, 8:00am EST
Illustrations by Corey Brickley | Photography by Jessica Chou



<https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona>

Paid moderation

- “*Think like that there is a sewer channel and all of the mess/dirt/waste/shit of the world flow towards you and you have to clean it.*” - Paid Facebook Moderator
- Trauma is not part of the compensation

The image shows a screenshot of the BBC News website. At the top, there is a navigation bar with the BBC logo, a search icon, and links for Home, News, Sport, More, and a search bar. Below the navigation bar is a red header with the word "NEWS" in white capital letters. Underneath the red header, the word "Tech" is written in a smaller font. The main headline is "Microsoft staff 'suffering from PTSD'", attributed to "Dave Lee" (North America technology reporter) and dated "12 January 2017". Below the headline is a red share button. At the bottom of the screenshot, there is a blue banner featuring the Microsoft logo.

<https://www.newyorker.com/tech/annals-of-technology/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity>

Paid moderation

- Strengths
 - Trained reviewers check claims, which can help avoid community brigading (think coordinated effort to report a post a group finds unpopular)
 - Supports more calibrated and consistent outcomes
- Weaknesses
 - Major emotional trauma and PTSD for moderators
 - Evaluators have only seconds to make a judgment

Community moderation

- Members of the community, or moderators who run the community, handle reports and proactively remove comments
- Examples: Reddit, Twitch, Discord
- It's best practice for the moderator team to publish their rules, rather than let each moderator act unilaterally



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.

Community moderation

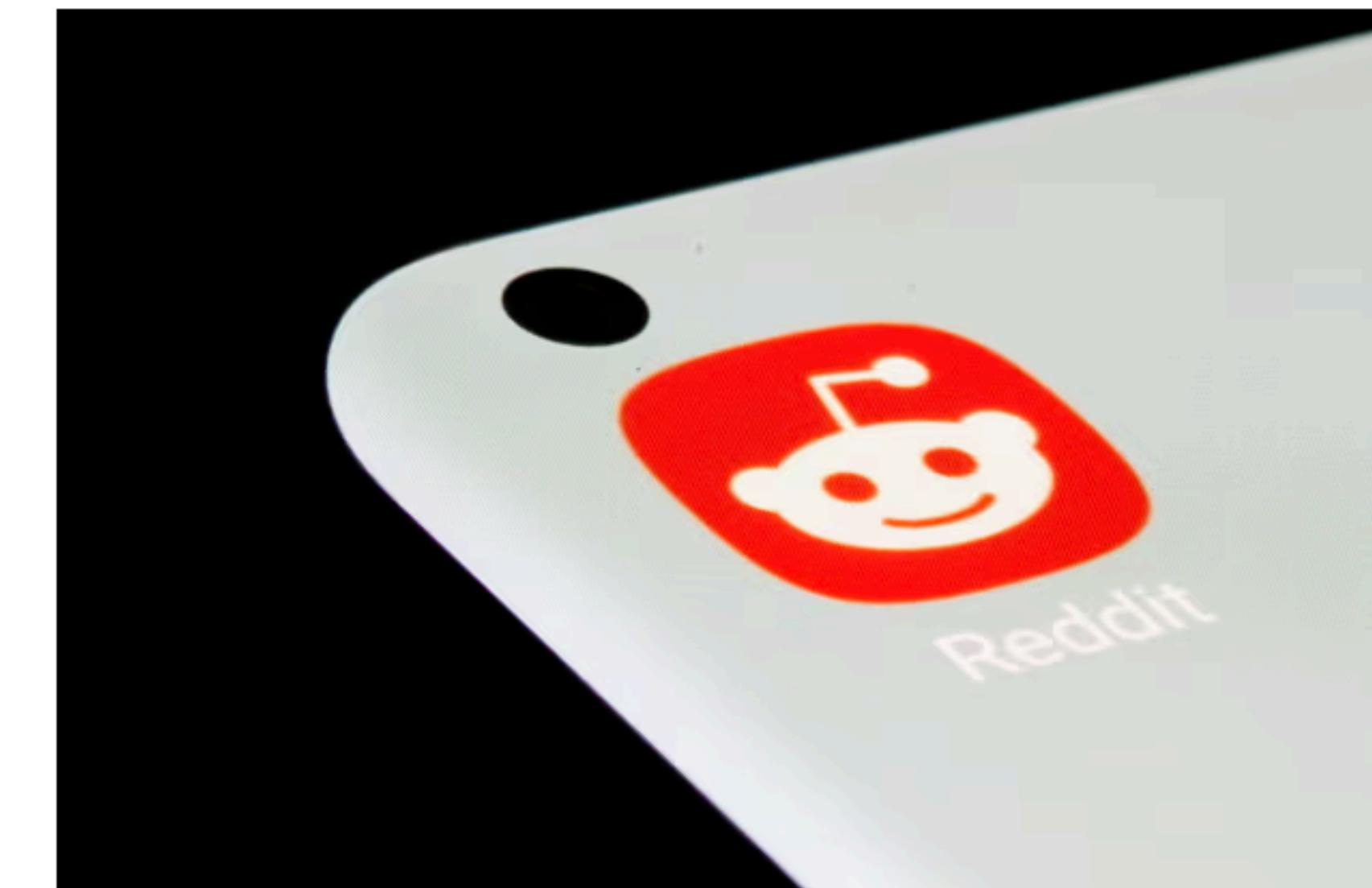
- “*I really enjoy being a gardener and cleaning out the bad weeds and bugs in subreddits that I’m passionate about. Getting rid of trolls and spam is a joy for me. When I’m finished for the day I can stand back and admire the clean and functioning subreddit, something a lot of people take for granted. I consider moderating a glorified janitor’s job, and there is a unique pride that janitors have.*”
- /u/noeatnosleep, moderator on 60 subreddits

Community moderation

- “We... estimate that Reddit moderators worked a minimum of 466 hours per day in 2020. These hours amount to 3.4 million USD a year based on the median hourly wage for comparable content moderation services in the U.S.”

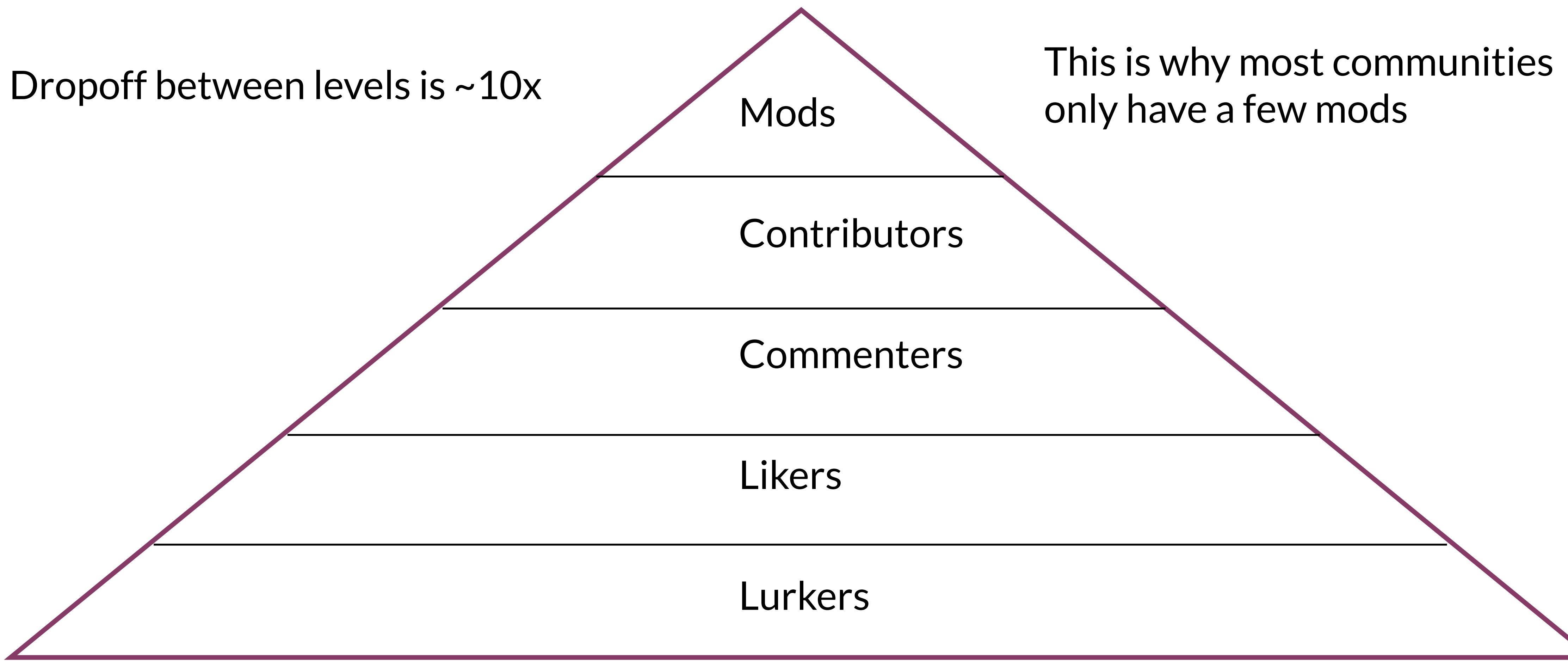
How social media's biggest user protest rocked Reddit

A mass user protest six months ago over technical tweaks had big downstream effects, and now the ‘front page of the internet’ is changed for ever



Reddit experienced a revolt by its volunteer moderators after the social media site decided to charge for access to its API. Photograph: Dado Ruvic/Reuters

Contribution period



Community member roles in moderation

Community feedback beyond moderators

The image shows a screenshot of a social media post interface. At the top, there are three icons: a flag, a bookmark, and a blue 'Reply' button. A tooltip for the flag icon says: "privately flag this post for attention or send a private notification about it". Below this, a user named 'ygduf' has posted a comment with 41 points and 25 days ago. The comment text is: "now we all know where you live though". The upvote and downvote arrows for this comment are circled in purple. At the bottom of the post are five interaction buttons: Reply, Give Award, Share, Report, and Save.

privately flag this post for attention or send a private notification about it

ygduf 41 points · 25 days ago

now we all know where you live though

Reply Give Award Share Report Save

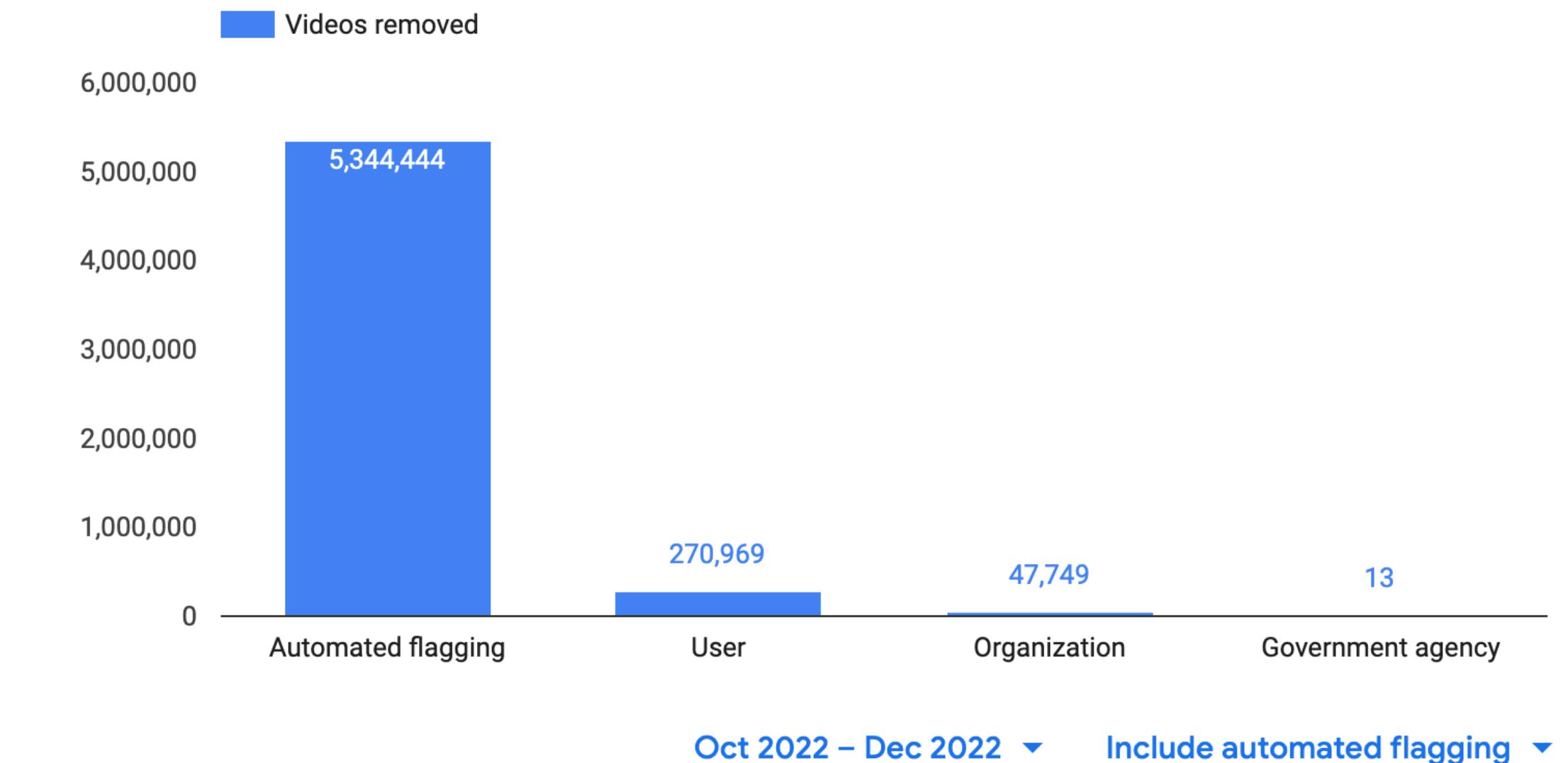
Lampe, C., & Resnick, P. (2004, April). Slash (dot) and burn: distributed moderation in a large online conversation space. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 543-550).

Community moderation

- Strengths:
 - Leverages intrinsic motivation
 - Local experts are more likely to have context to make hard calls
- Weaknesses:
 - Mods don't feel they get the recognition they deserve
 - Not necessarily consistent
 - Without oversight, mods can grow problematic communities
 - Lots of labor, hard to scale!

Algorithmic moderation

- Train an algorithm to automatically flag or take down content that violates rules (e.g., nudity, copyright)
- **Discuss:** why might AI not be as prevalent on TikTok as YouTube, Facebook, and Instagram?



Meta also highlights the efficiency of its artificial intelligence (AI) moderation tools, which TikTok does not emphasize. The rate of moderation automation is very high at Meta: At Facebook and Instagram, respectively, 94% and 98% of decisions are made by machines – far more than the 45% reported by TikTok.

https://www.lemonde.fr/en/pixels/article/2023/11/14/content-moderation-key-facts-to-learn-from-facebook-instagram-x-and-tiktok-transparency-reports_6252988_13.html

Algorithmic moderation and feeds

- Debate primarily focuses on removal, suspending users, and the option not to do so
- But, reducing the visibility of problematic content is also effective moderation
 - If no one sees it, it can't have impact

= WIRED

SUBSCRIBE

 Blog ▾

RENEE DIRESTA IDEAS AUG 30, 2018 4:00 PM

Free Speech Is Not the Same As Free Reach

Bad faith politicking about the way search algorithms work makes it harder for tech companies to solve the real problems.



Product

Freedom of Speech, Not Reach: An update on our enforcement philosophy

By [Twitter Safety](#)

Monday, 17 April 2023



Algorithmic errors

- Errors are especially likely to hit minoritized groups, who are less represented in the training data
- “Ground truth” labels for these tasks are a fallacy
 - There’s no consensus on what qualifies as harassment, so even a “perfect” ML model will anger a substantial number of users



TECH YOUTUBE CULTURE

YouTube is still restricting and demonetizing LGBT videos — and adding anti-LGBT ads to some

Deceiving Google’s Perspective API Built for Detecting Toxic Comments

Hossein Hosseini, Sreeram Kannan, Baosen Zhang and Radha Poovendran
Network Security Lab (NSL), Department of Electrical Engineering, University of Washington, Seattle, WA
Email: {hosseinh, ksreeram, zhangbao, rp3}@uw.edu

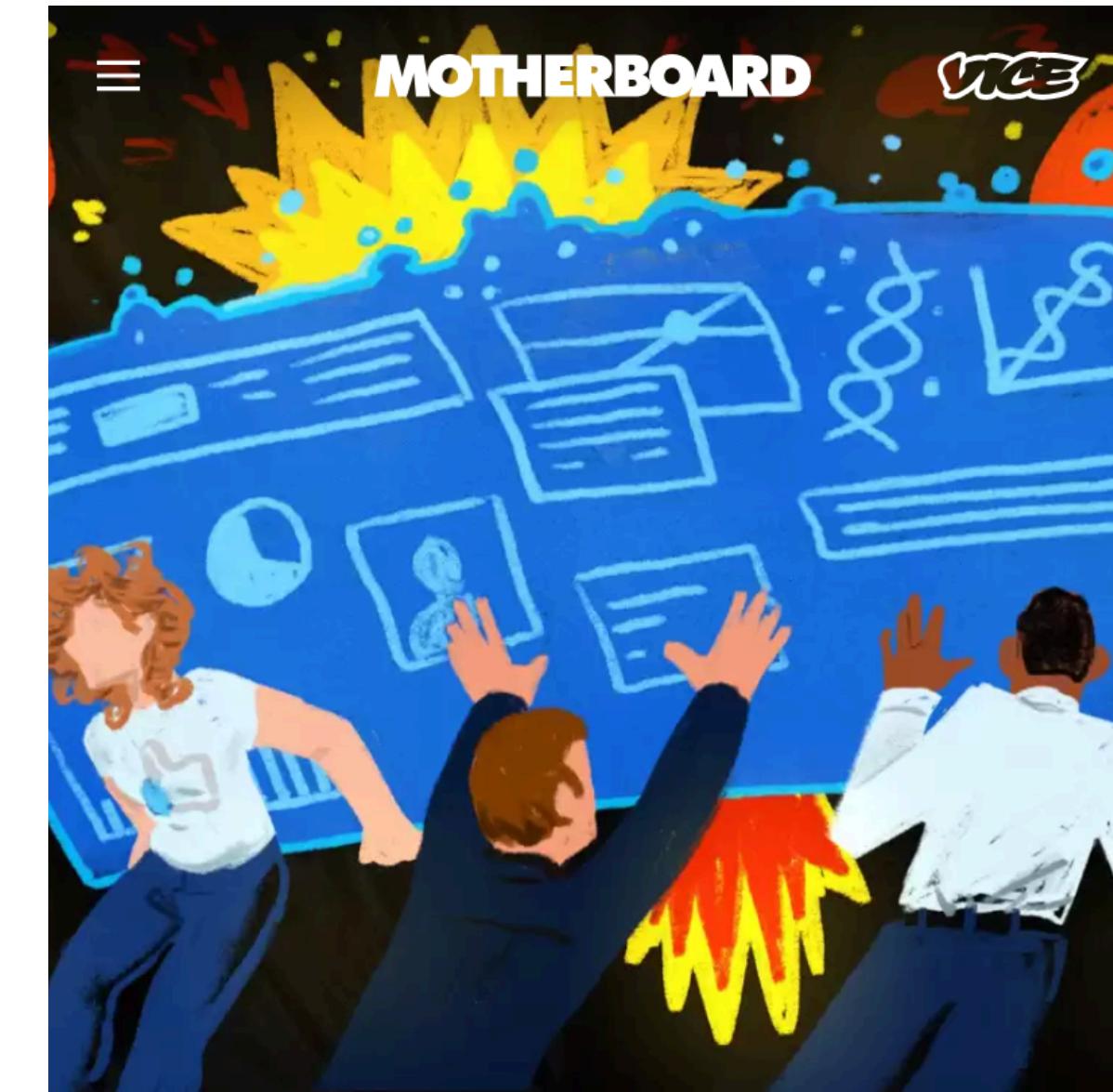
Algorithmic moderation

- Strengths:
 - Can act quickly, before people are hurt by the content
- Weaknesses:
 - These systems make embarrassing errors, often ones that the creators didn't intend
 - Errors are often interpreted as intentional platform policy
 - Even if a perfectly fair, accountable, and transparent algorithm were possible, culture would evolve and training data would become out of date

Deploying moderation

Deploying moderation on Facebook

- Moderators are responsible for:
 - Removing violent content, threats, nudity, and other content breaking TOS



FACEBOOK

The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People

Moderating billions of posts a week in more than a hundred languages has become Facebook's biggest challenge.

Deploying moderation on Reddit

- Moderators are responsible for:
 - Removing content that breaks rules
 - Getting rid of spam, racism and other undesirable content

BUSINESS
INSIDER



For whom the troll trolls: A day in the life of a Reddit moderator



Kim Renfro, Tech Insider

Jan. 13, 2016, 12:27 PM

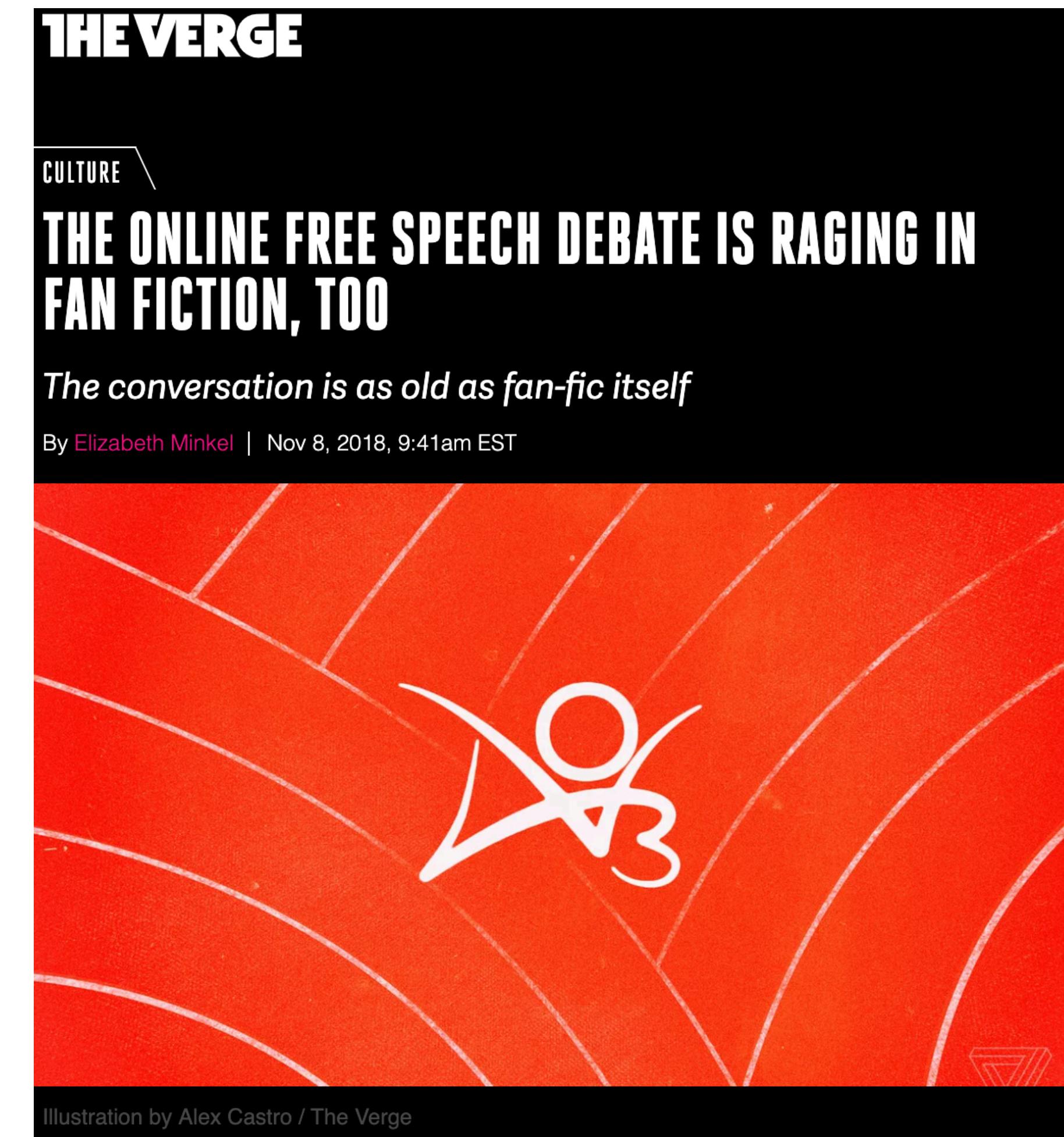


Kim Renfro/Tech Insider

Liz Crocker is a 33-year-old graduate student and the mother of a one-year-old. Her days are spent writing a dissertation, teaching a course to PhD students at Boston University, looking after her child, and — oh yeah — dealing with internet trolls.

Deploying moderation on AO3

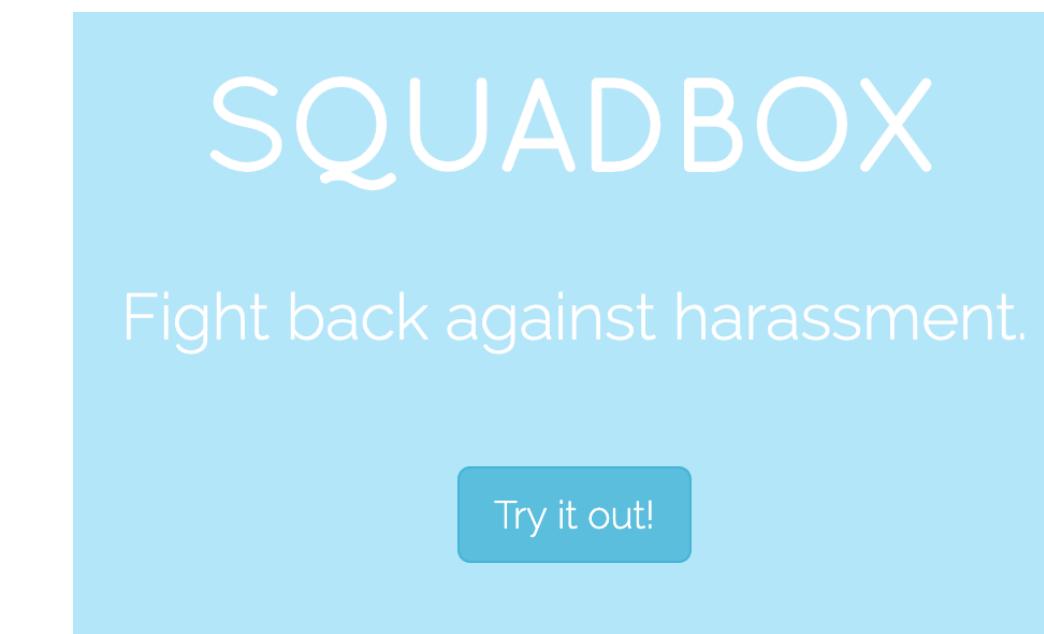
- Even in systems like Archive of Our Own that are light on moderation, people still debate moderation practices



Fiesler, C., Morrison, S., & Bruckman, A. S. (2016, May). An archive of their own: A case study of feminist HCI and values in design. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 2574-2585).

Deploying moderation on Email

- Friends intercept email before it makes its way to your inbox



Put a squad of trusted friends, volunteers, or paid moderators between the world and your inbox.

Messages only reach you if your squad approves it.

Together, the members of your squad can weather harassment so that you don't feel overwhelmed.



MIT researchers developed it as a way to mitigate online harassment.

Mahar, K., Zhang, A. X., & Karger, D. (2018, April). Squadbox: A tool to combat email harassment using friendsourced moderation. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

So... what do we do?

- Many social systems use multiple tiers:
 - **Tier 1:** Algorithmic moderation for the most common and easy-to-catch problems.
Tune the algorithmic filter conservatively to avoid false positives, and route uncertain judgments to human moderators
 - **Tier 2:** Human moderation, paid or community depending on the platform.
Moderators monitor flagged content, review an algorithmically curated queue, or monitor all new content, depending on platform
 - **Rarer, but: Tier 0:** Membership review, screening people who are allowed into the community in the first place

Don't wait until it becomes a problem

- Even if your community is small now, you should plan your moderation strategy. Young platforms run into moderation issues too, and it often catches them flat-footed
- Don't let it be obvious in hindsight that you needed moderation
- Establish the norm of expected conduct early, and enforce it early
- (See the norms lecture from a few weeks ago)

How to Protect Your Server from Raids 101



Discord Trust & Safety Team
9 months ago · Updated

Follow

On Discord, a raid is when a large number of users or bots join a server at once for disruptive or malicious intents and purposes. Raiding is against our Terms of Service [guidelines](#).

While there are different types of behavior and activities that raiders engage in, there are a number of steps you can take to prevent raids of all types before they occur, or minimize damage as they're happening.

<https://support.discord.com/hc/en-us/articles/10989121220631-How-to-Protect-Your-Server-from-Raids-101>

Conflict and volume

- There is a lot of moderation work that needs to be done



Alex Stamos 

@alexstamos

Suivre

Any discussion about content moderation needs to be grounded in the reality that the controversial calls are a tiny portion of what is necessary to make any social feature at all usable.

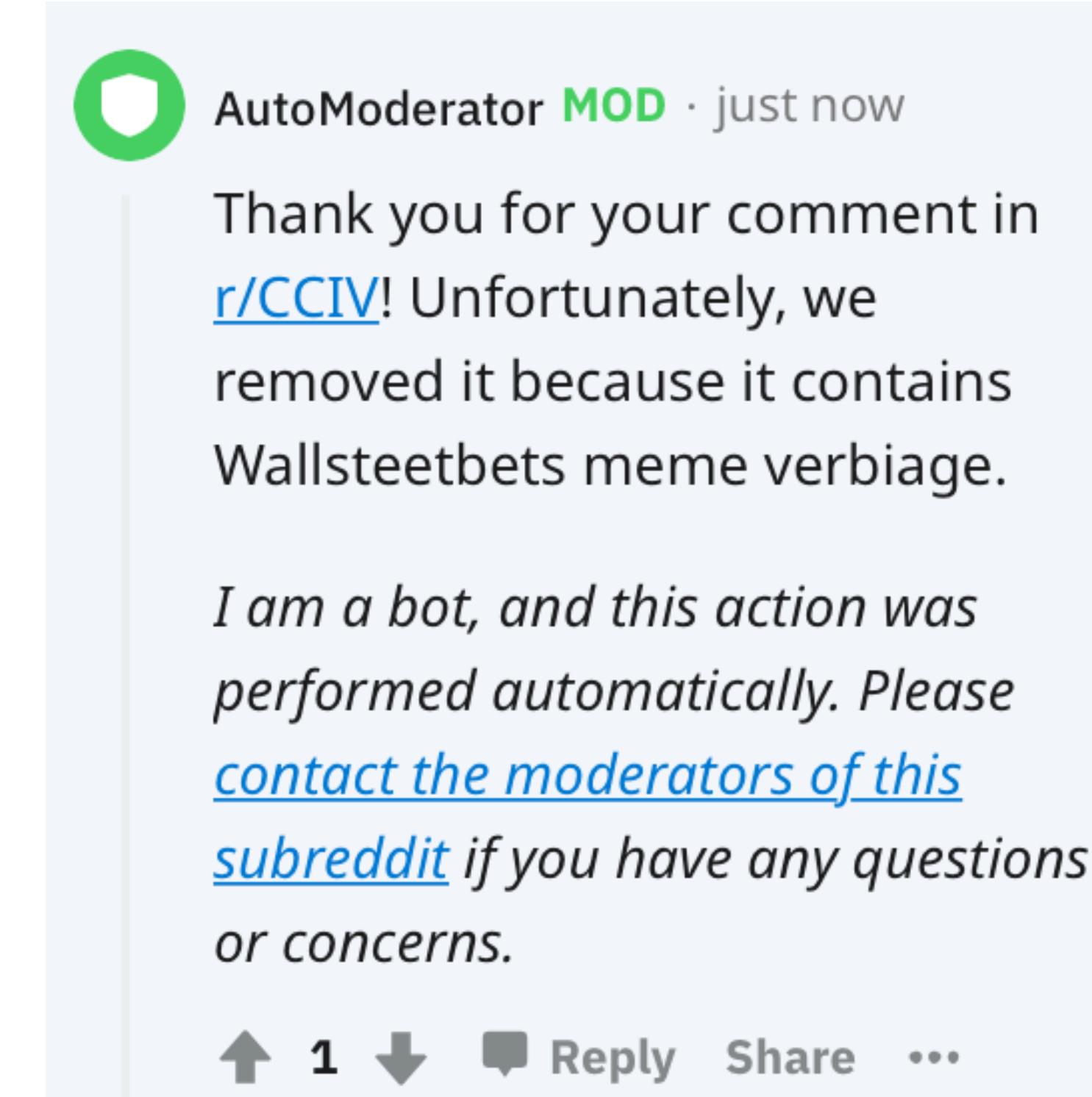
FB Q2 numbers:
Organized Hate - 266k
Spam - 1.4B

5000x difference

08:16 - 28 oct. 2020

Tools can help

- Tools help facilitate moderator decisions by automatically flagging problematic posts, and providing relevant information
- Moderators often script tools if the platform API allows it



AutoModerator MOD · just now

Thank you for your comment in [r/CCIV](#)! Unfortunately, we removed it because it contains Wallstreetbets meme verbiage.

I am a bot, and this action was performed automatically. Please [contact the moderators of this subreddit](#) if you have any questions or concerns.

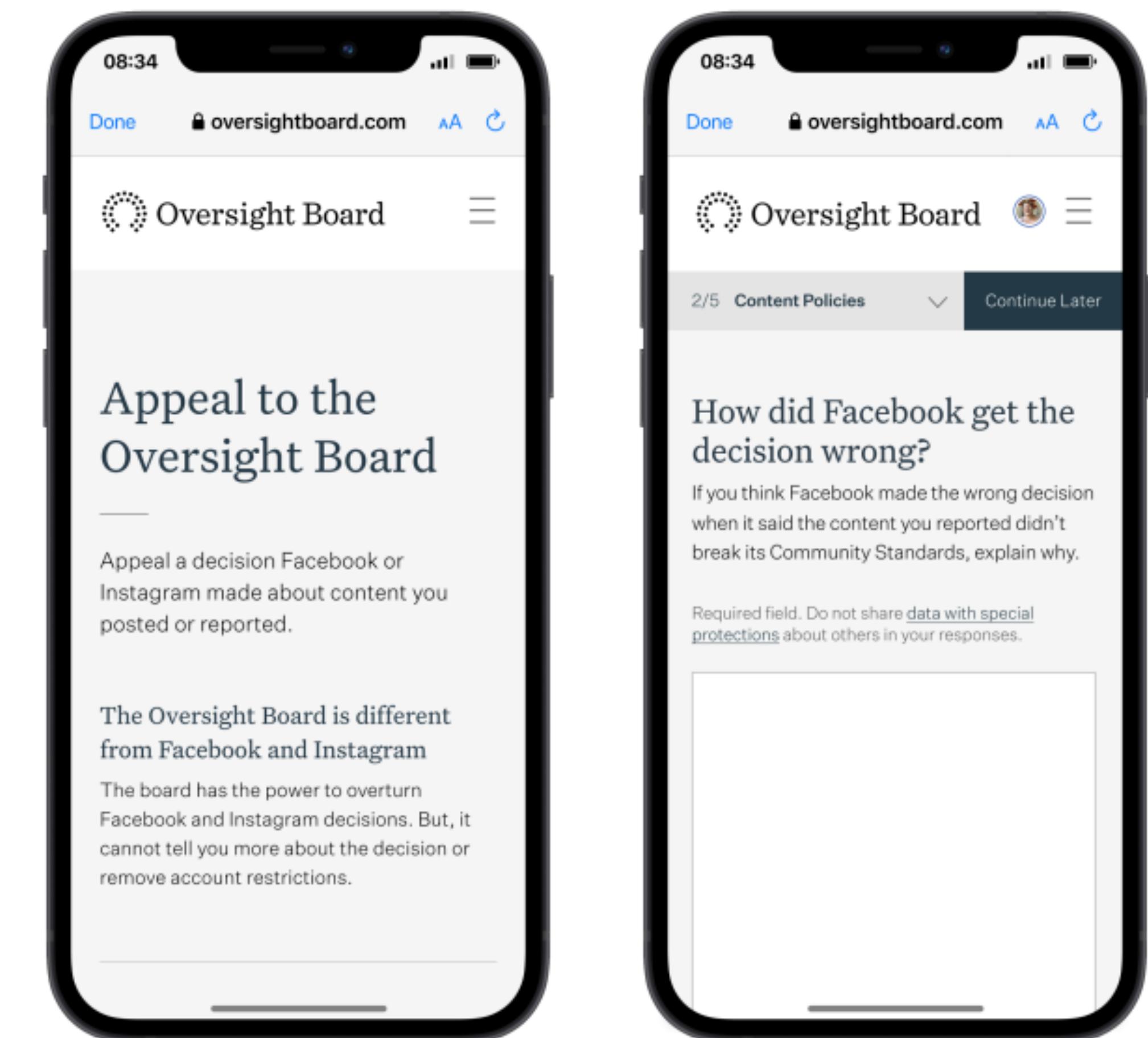
▲ 1 ▾ Reply Share ...

Huggle is a [diff](#) browser intended for dealing with [vandalism](#) and other unconstructive edits on [Wikimedia](#) projects, written in [C++](#) using the [Qt](#) framework. It was originally developed in [.NET Framework](#) by [Gurch](#), who is no longer active on this project. Anyone can download Huggle, but [rollback](#) permissions are required to use the program without restrictions on the English Wikipedia.

The principal idea of Huggle as an anti-vandalism tool is to make it possible for Wikipedia to stay as open and free as possible (allowing everyone to edit without any restrictions), while also keeping it clean of any vandalism.

Appeals

- Most modern platforms allow users to appeal unfair decisions
- If the second moderator disagrees with the first moderator, the post goes back up



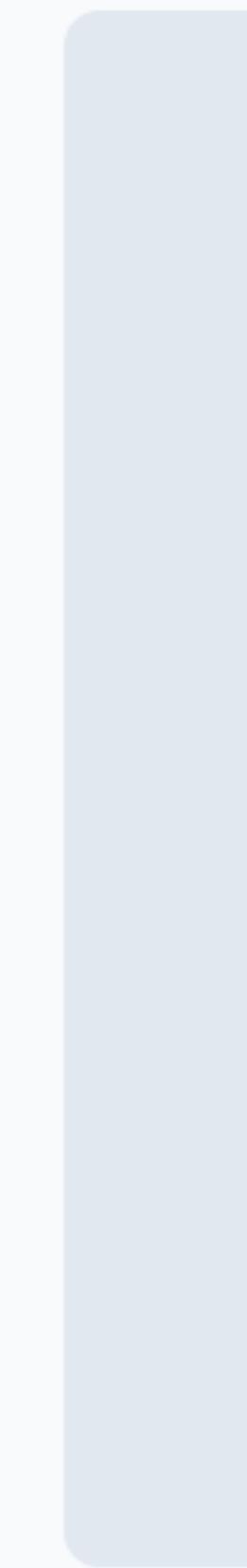
Does moderation work?

Does moderation work?

- What would it mean for moderation to work?
 - Less harassment
 - Less toxic or vulgar content
 - Only topically relevant content
 - People who harass and post toxic content either change their ways or disappear

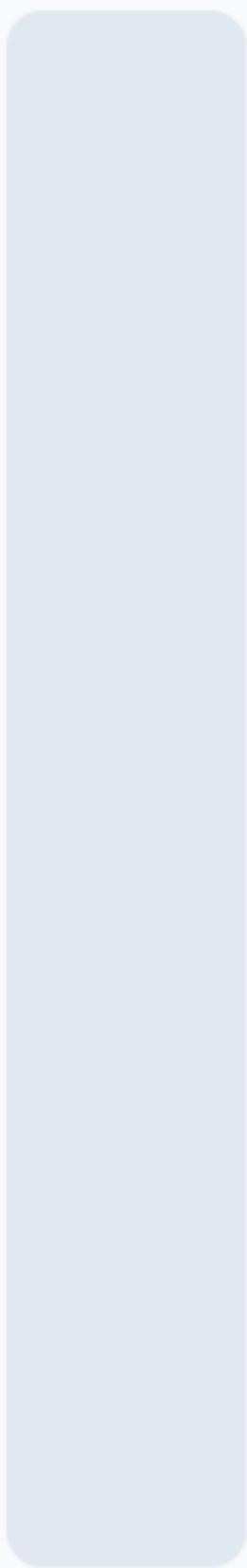
Do you think moderation "works"? Meaning, keeps harmful content off of online platforms in the long term.

0%



Yes

0%



No

- Does this mean moderation works?



Despite most Americans being critical of the job social media companies are doing to address harassment, some are optimistic about a variety of possible solutions asked about in the survey that could be enacted to combat online harassment.

About half of Americans say permanently suspending users if they bully or harass others (51%) or requiring users of these platforms to disclose their real identities (48%) would be very effective in helping to reduce harassment or bullying on social media.

Around four-in-ten say criminal charges for users who bully or harass (43%) or social media companies proactively deleting bullying or harassing posts (40%) would be very effective.

Does moderation work?

- Yes, for short periods
- Remember livestreaming lecture
- Moderation shifts descriptive norms and reinforces injunctive norms by making them salient

Behavior	% Increase after Event
Spam	43.8%
Question	55.3%
Smile	220.0%

Table 4: Percentage Increase after Event of Same Type

Behavior	User Type	% Increase after Event
Spam	Mod	67.8%
	Sub	46.4%
	Turbo	15.0%
	Regular User	38.6%
Question	Mod	76.3%
	Sub	45.2%
	Turbo	62.5%
	Regular User	52.1%
Smile	Mod	333.3%
	Sub	236.7%
	Turbo	166.7%
	Regular User	233.3%

Table 6: Percentage Increase after Event of Same Type Posted by Given User Type

Seering, J., Kraut, R., & Dabbish, L. (2017, February). Shaping pro and anti-social behavior on twitch through moderation and example-setting. In Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing (pp. 111-125).

Deplatforming works

- After a toxic community is deplatformed, its members leave entirely, or migrate and drastically reduce their hate speech
- Discussion reduces about the deplatformed individuals in mainstream spaces
- User activity reduces, not increases, on new alt sites

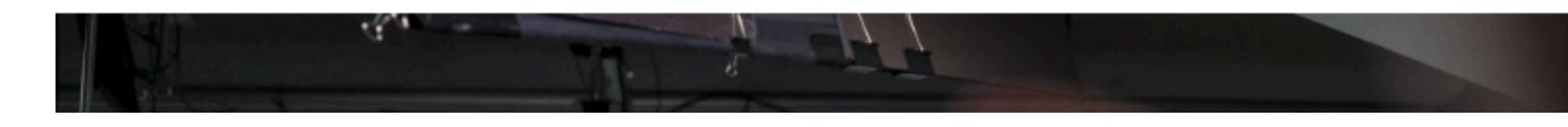
Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on human-computer interaction*, 1(CSCW), 1-22.

Jhaver, S., Boylston, C., Yang, D., & Bruckman, A. (2021). Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-30.

Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-24.

THE SHIFT

Reddit Limits Noxious Content by Giving Trolls Fewer Places to Gather



DONATE

MEDIA

Twitter Bans Alex Jones And InfoWars; Cites Abusive Behavior

September 6, 2018 · 5:34 PM ET

By Avie Schneider

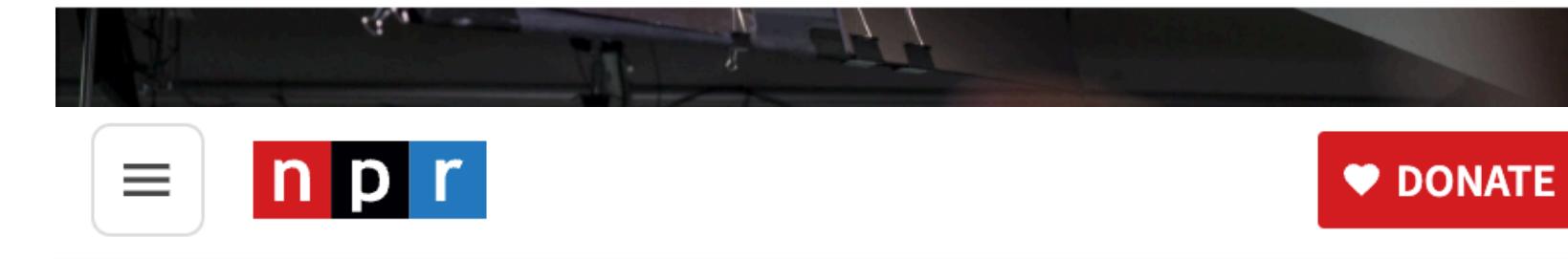


Deplatforming works*

- However, those who migrate and stay active tend to further radicalize, albeit to smaller audiences
- Use of other platforms is often sufficient

THE SHIFT

Reddit Limits Noxious Content by Giving Trolls Fewer Places to Gather



MEDIA

Twitter Bans Alex Jones And InfoWars; Cites Abusive Behavior

September 6, 2018 · 5:34 PM ET

By Avie Schneider



Horta Ribeiro, M., Jhaver, S., Zannettou, S., Blackburn, J., Stringhini, G., De Cristofaro, E., & West, R. (2021). Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-24.
Russo, G., Verginer, L., Ribeiro, M. H., & Casiraghi, G. (2023, June). Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 17, pp. 742-753).

Moderation can also backfire

- Moderation can drive away newcomers, who don't understand the community's norms yet
- Users circumvent algorithmic controls
 - Instagram hides #thighgap as as promoting unhealthy behavior...and users create #thygap instead [Chancellor et al. 2016]



BANNING WORDS ON
INSTAGRAM TOTALLY
BACKFIRED

The New York Times

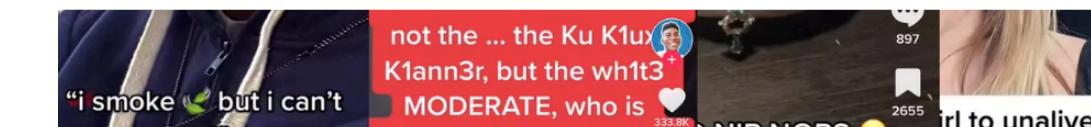


GIVE THE TIMES

*Leg Booty? Panoramic? Seggs? How
TikTok Is Changing Language*

A new vocabulary — a little fun, a little dystopian — has emerged on the social video platform, as creators try to get around algorithms and strict content moderation. They call it algospeak.

Give this article



Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016, February). # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing (pp. 1201-1213).

Moderation can have unequal impact

- Three groups that tend to get their content moderated at higher rates:
 - Political conservatives: content that violates rules; is misinformation, adult content, or hate speech
 - Transgender individuals: content that does not violate rules; is critical of dominant groups, or specific to transgender issues
 - Black individuals: content related to racial justice or racism

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1-35.

Marshall, B. (2021). Algorithmic misogyny in content moderation practice. Heinrich-Böll-Stiftung European Union.

The Washington Post

Sign in

Tech

Help Desk

Artificial Intelligence

Internet Culture

Spa

TECHNOLOGY

Facebook's race-blind practices around hate speech came at the expense of Black users, new documents show

Researchers proposed a fix to the biased algorithm, but one internal document predicted pushback from 'conservative partners'

By [Elizabeth Dwoskin](#), [Nitasha Tiku](#) and [Craig Timberg](#)

November 21, 2021 at 8:00 a.m. EST

Moderation policy enforcement

content warning: moderation policy documents describing revenge porn, hate speech, and harassment of minority groups, with examples

Why is moderation so hard?



Nevermind. It's nudity and disallowed unless the baby is actively nursing.

How do you define what content constitutes:
Nudity?
Harassment?
Cyberbullying?
A threat?
Suicidal ideation?

A glimpse into content moderation

- In 2017, The Guardian published a set of leaked moderation guidelines that Facebook was using at the time to train its paid moderators
- To get a sense for the kinds of calls that Facebook has to make and how moderators have to think about the content that they classify, let's inspect a few cases.
- (The lines might be different today, and this group would likely think differently)

Revenge Porn (1)

CURRENT POLICY

High-level: Revenge porn is sharing nude/near-nude photos of someone publicly or to people that they didn't want to see them in order to shame or embarrass them.

Abuse Standards:

6. Attempting to exploit intimate images by any of the following:

- Sharing imagery as "revenge porn" if it fulfills all three conditions:
 1. Image produced in a private setting. AND
 2. Person in image is nude, near nude, or sexually active. AND
 3. Lack of consent confirmed by:
 - Vengeful context (e.g. caption, comments, or page title), OR
 - Independent sources (e.g. media coverage, or LE record)

Note ANDing of
three conditions

Hate Speech

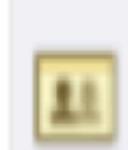
REMOVE

What do we protect?

- Protected
 - Individuals
 - Groups
 - Humans



Using my freedom of speech to inform that I find
homosexuals DISGUSTING!
Common Interest
20 members



i fuckin hate christians

Open Group

REMOVE

Hate Speech

ALLOWED

What do we NOT
protect?

Not protected

- Concepts
- Institutions
- Beliefs



Anti-homosexuality
Common Interest
15 members



I Hate Christianity

Request to Join

Info

Basic Info

Legalistic classification of what is protected: individuals, groups, and humans.
Concepts, institutions, and beliefs are not protected

“I hate Christians” is banned, but Facebook allows “I hate Christianity”

Quasi Protected Category (QPC)

People who cross an international border with intent to establish residency in a new country, regardless of whether their motivation is economic or political (defined as: migrants, refugees, immigrants, asylum seekers)

- Protected + Quasi protected = **Quasi protected**
 - “Muslim migrants ought to be killed” = **Quasi protected**
- Not Protected + Quasi protected = **not protected**
 - “Keep the horny migrant teenagers away from our daughters” = **allowed**
- Migrants are so filthy. (**Filthy** is an adjective not a noun, we consider this to be a description of their appearance rather than nature)

Creation of a new category to handle the case of migrants

Complicated math mixing policy and ethics to handle these cases

Hate Speech - Migrants

Examples: (DELETE)

Dehumanizing characteristics –

REMOVE

- Migrants are scum.
- Migrants are filthy cockroaches that will infect our country.
- The migrant rats have arrived in Berlin.
- Refugees? They're all rape-fugees!
- Refugees are state-financed child molesters.

If it's dehumanizing,
delete it.

EDGE CASE – “Dismissing” an entire QPC should be an IGNORE

- Migrants are lazy and just want to come here to feed off our social welfare benefits.
- Migrants are so filthy.
- Migrants are thieves and robbers.

But, dismissing is
different from
dehumanizing.

Today's goals

By the end of today, you should be able to...

- Articulate what makes moderating online content challenging
- Differentiate the advantages and challenges around algorithmic, community-led, and paid approaches to moderation
- Explain the nuances of when and for whom moderation works, and does not work

IN4MATX 153: CSCW

**Class 11:
Content Moderation**

Professor Daniel Epstein
TA Dennis Wang
Reader Weijie Du