Moderation

CS 278 | Stanford University | Michael Bernstein

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

Announcements

Project final deliverable due Friday June 6

Last time

Anti-social behavior is a fact of life in social computing systems. Trolling is purposeful; flaming may be due to a momentary lack of self-control.

The environment and mood can influence a user's propensity to engage in anti-social behavior: but (nearly) anybody, given the wrong circumstances, can engage in flaming.

Changing the environment, allowing mood to pass, and allowing face-saving can help reduce anti-social behavior.

Dark behavior exists: be prepared to respond.

A story of Facebook's content moderation

[Adler, Abumrad, Krulwich 2018]

No pornography.

What counts as pornography?

Fine. 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.

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

@marksweney Email

Tue 30 Dec 2008 08.17 EST

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 profile by protestors - called "lactivists" in some circles - called "Hey Facebook, breast feeding is not obscene".

Fine. Nudity means 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,

Fine. Nudity means 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

Hold up. So, it's not a picture of me at college as soon as I get handed my degree at graduation?

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

ZDNet MENU US

MUST READ: SHA-1 collision attacks are now actually practical and a looming danger

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

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

FINE. Age cap: only infants.

OK, then 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.

Right, but now 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.

…

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

Tarleton Gillespie, in his book Custodians of the Internet [2018]:

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

Today

Approaches to moderation

Does moderation work?

Regulation and Safe Harbor

Moderation

“Three imperfect solutions”

h/t Gillespie [2018]

Three imperfect solutions

[Gillespie 2018]

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

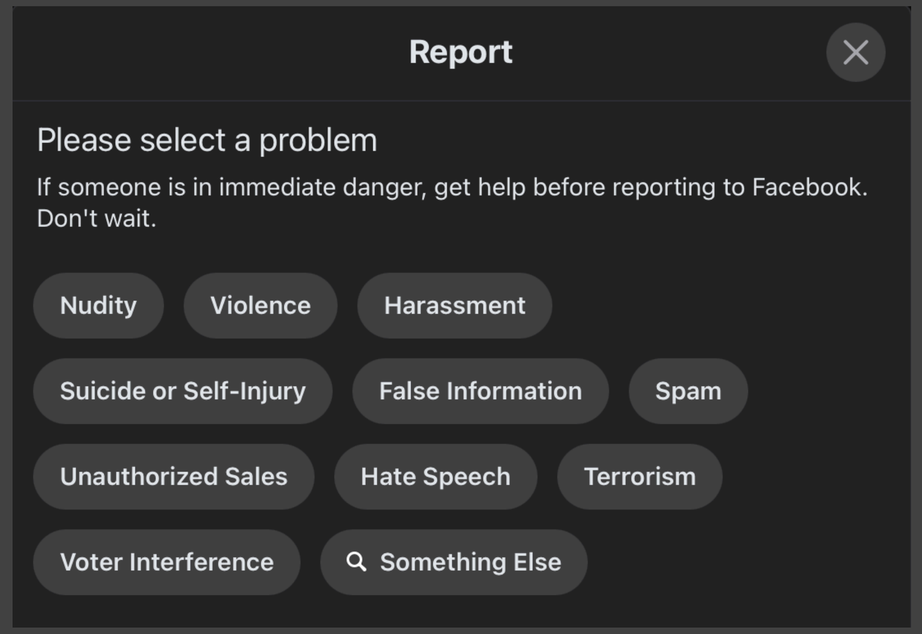
Each with their pros and cons

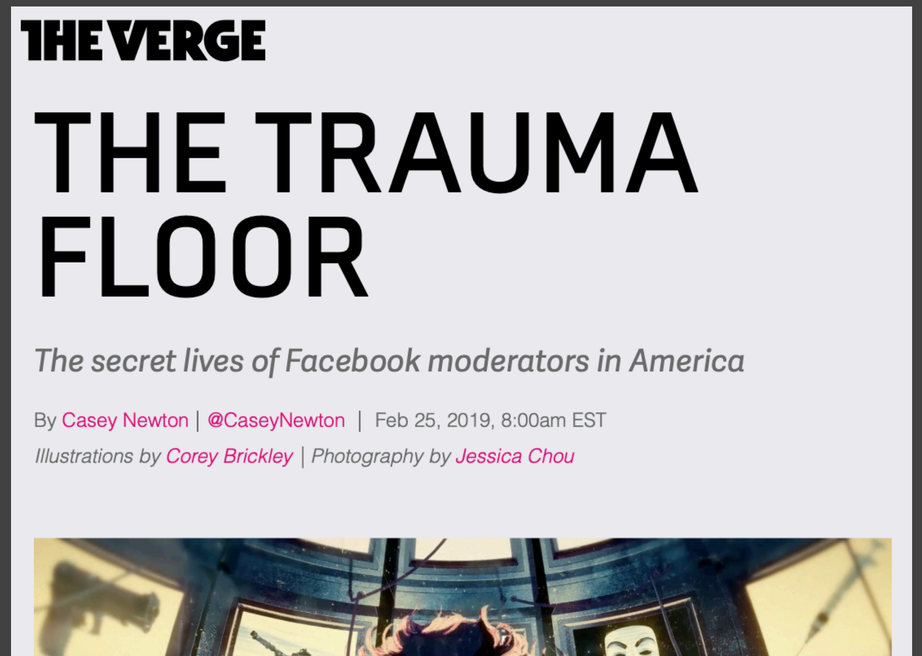
Paid moderation

Rough estimates:

- ~15,000 contractors on Facebook and Instagram [Newton 2019]
- ~40,000 contractors on TikTok [Reuters 2025]

Moderators at Meta are trained on over 100 manuals, spreadsheets and flowcharts to make judgments about flagged content.





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 [Chen 2017]

Paid moderation

Strengths

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

Weaknesses

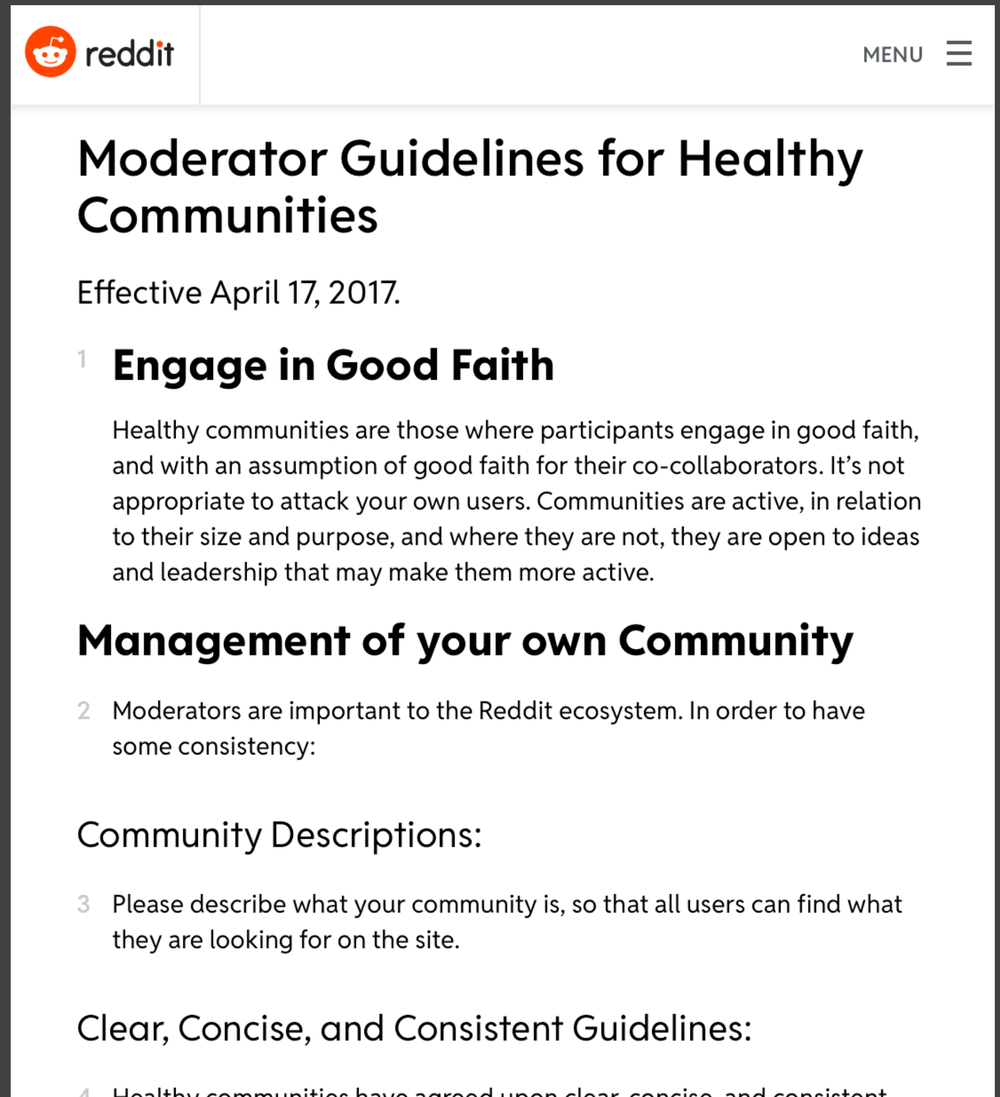
- Major emotional trauma and PTSD for moderators
- Evaluators may have only seconds to make a snap judgment

Community moderation

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

Examples: Reddit, Twitch, Discord, Fizz

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



Community moderation

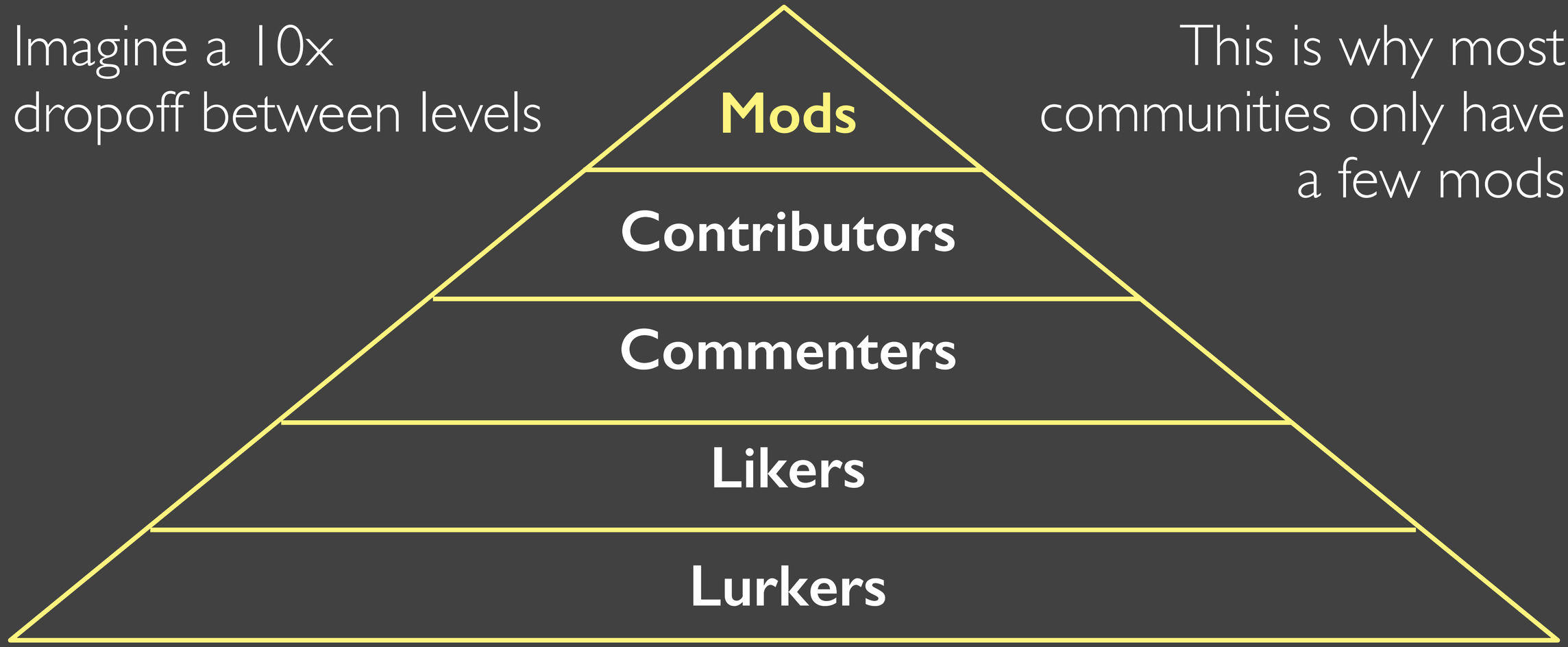
"I really enjoy being a gardener and cleaning out the bad weeds and bugs in subreddits that I'm passionate about. Getting rid of trolls and spam is a joy for me. When I'm finished for the day I can stand back and admire the clean and functioning subreddit, something a lot of people take for granted.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

- /u/noeatnosleep, moderator on 60 subreddits

[Hutch 2019; Seering, Kaufman and Chancellor 2020; Matias 2019]

Contribution pyramid redux

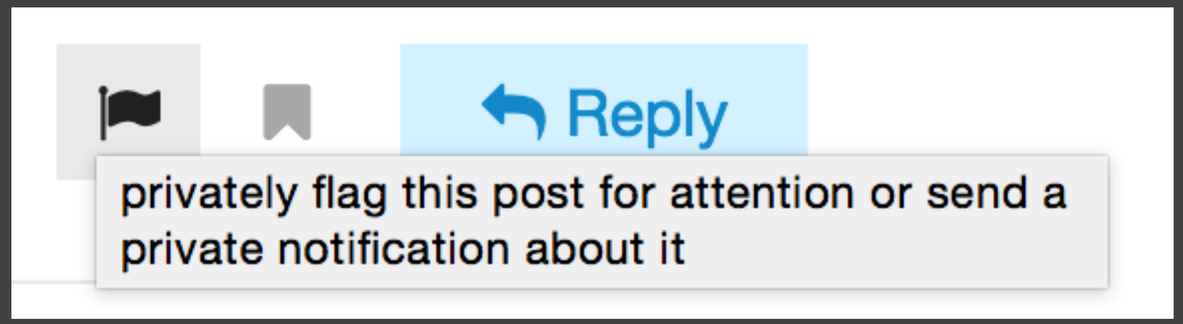
Imagine a 10x dropoff between levels



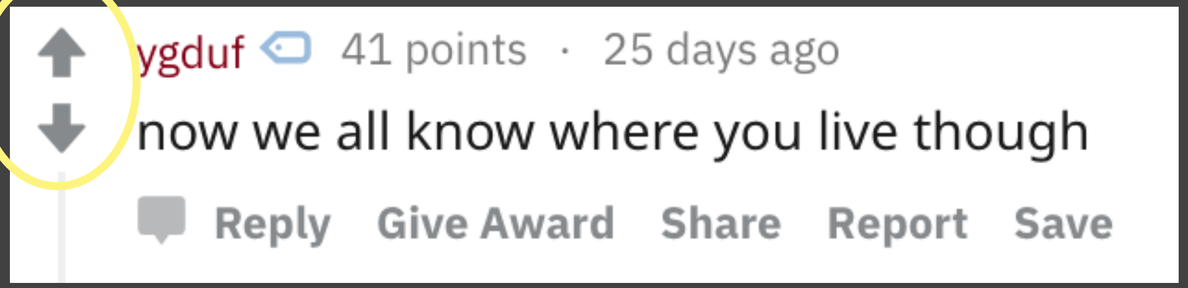
This is why most communities only have a few mods

Community member roles in moderation

Community feedback beyond moderators



Flagging



Voting [Lampe and Resnick 2004]

Community moderation

Strengths:

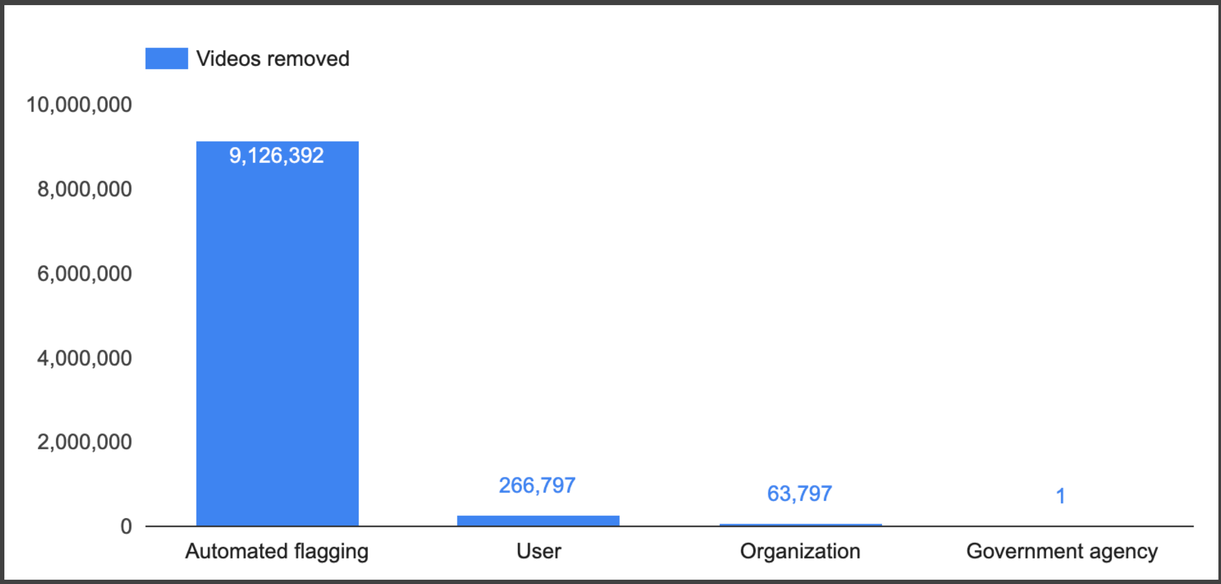
- Leverages intrinsic motivation
- Local experts are more likely to have context to make hard calls

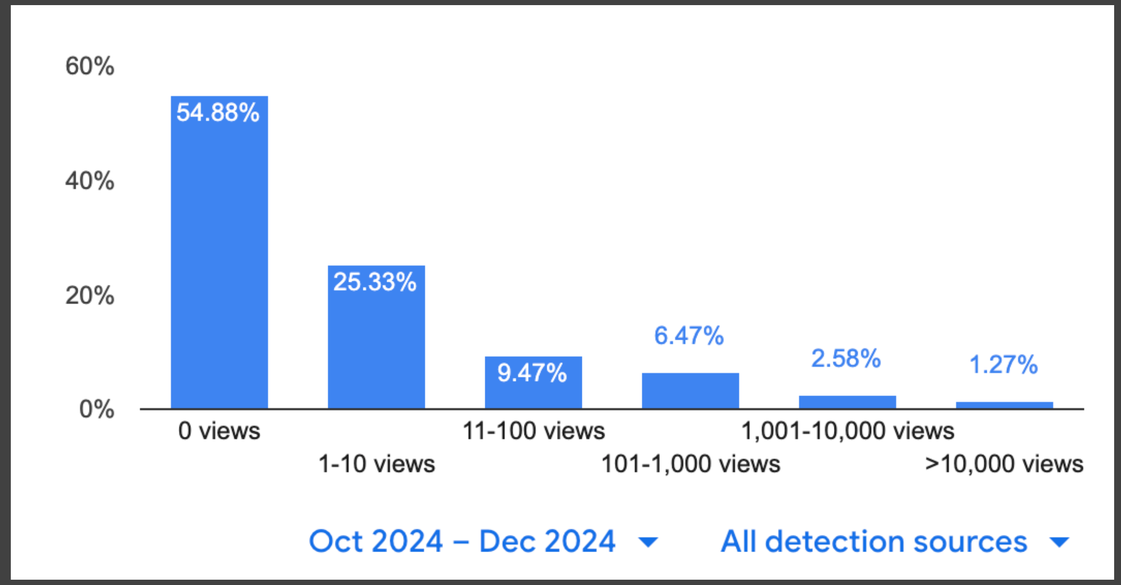
Weaknesses:

- Mods don't feel they get the recognition they deserve
- Can be inconsistent—local fiefdoms
- Without oversight, mods can grow anti-social communities

Algorithmic moderation

Train an algorithm to automatically flag or take down content that violates rules (e.g., nudity). Example via YouTube:





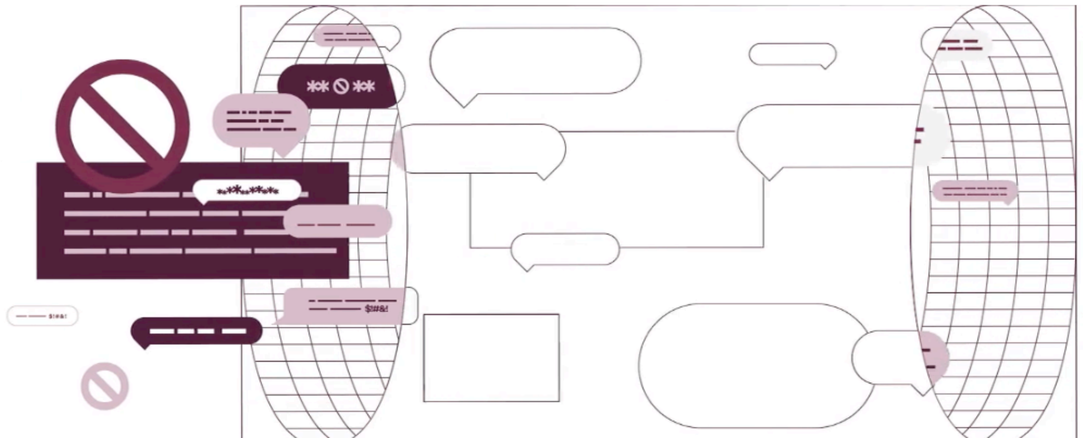
https://transparencyreport.google.com/youtube-policy/removals

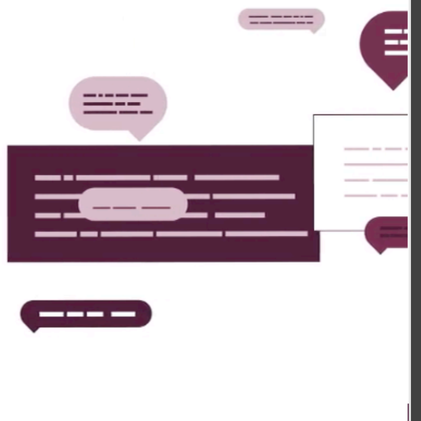
Perspective

Using machine learning to reduce toxicity online

Perspective API can help mitigate toxicity and ensure healthy dialogue online.

HOW IT WORKS →

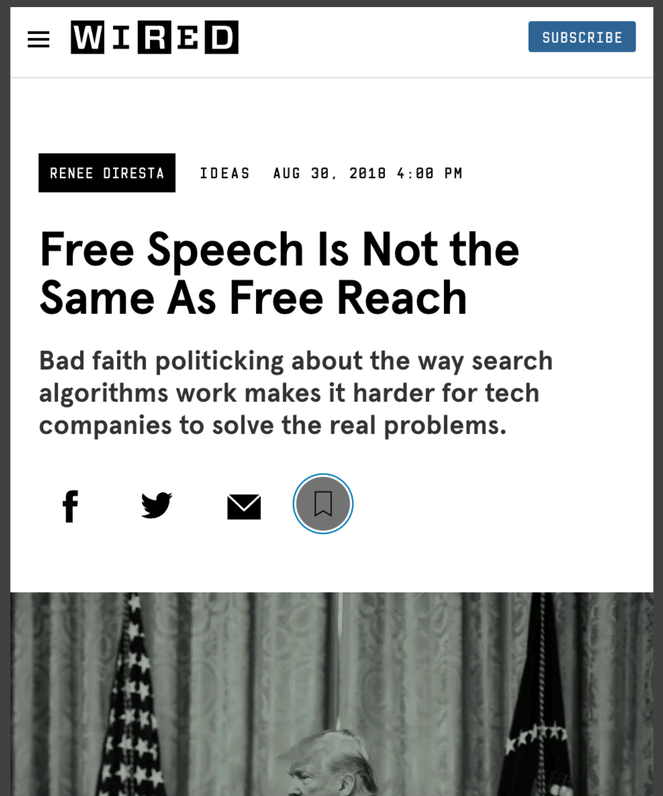


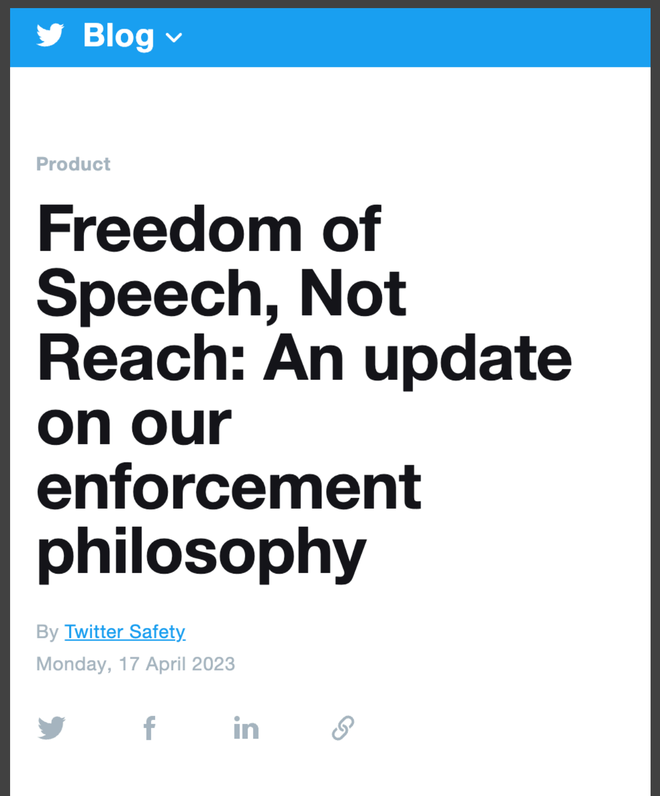


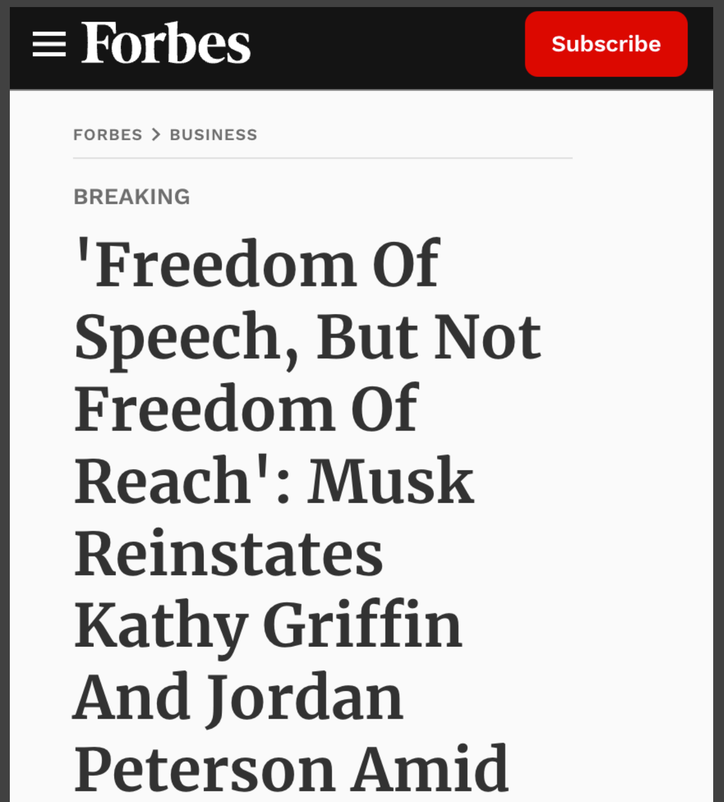

```
{
   "attributeScores":{
      "TOXICITY":{
         "spanScores":[
            {
               "score":{
                  "value":0.4445836,
                  "type":"PROBABILITY"
               }
            }
         ],
         "summaryScore":{
            "value":0.4445836,
            "type":"PROBABILITY"
         }
      }
   },
   "languages":[
      "en"
   ]
```

Feed algorithms are an instrument of moderation

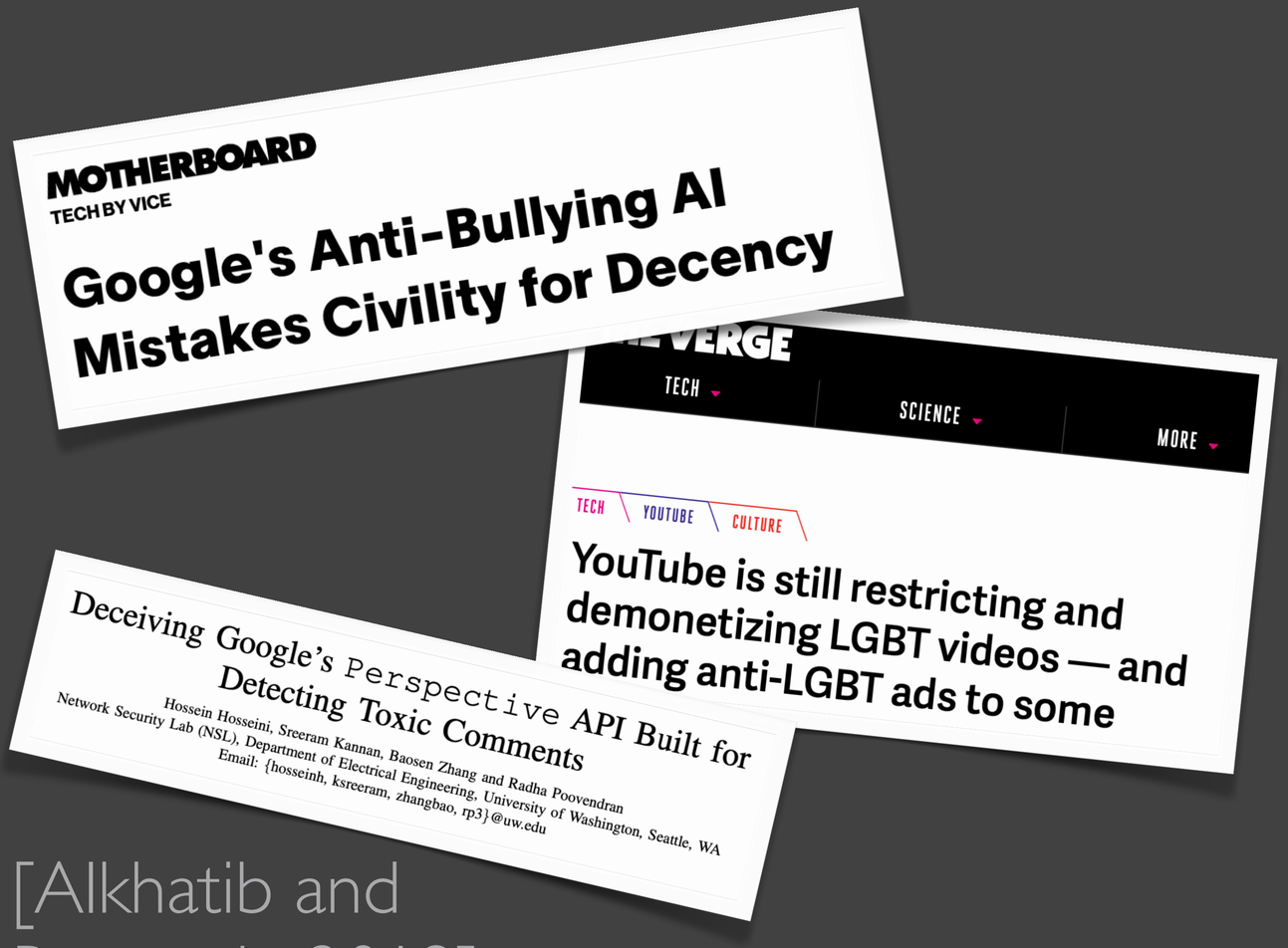
[Gillespie 2022]







Algorithmic errors



[Alkhatib and Bernstein 2019]

Why such a problem?

1. Errors are especially likely to hit minoritized groups, who are less represented in the training data
2. "Ground truth" labels for these tasks are a fallacy: when society has no consensus on what qualifies as harassment, even a "perfect" ML model will anger a substantial number of users [Gordon et al. 2021]

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 (FAccT) algorithm were possible, culture would evolve and training data would become out of date.

Deploying moderation

So…what do we do?

Many social computing systems use multiple tiers:

Tier I: 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 II: 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.

Also, 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. [Norms lecture]

Ask yourself today:

[Seering 2021]

What's your plan for dealing with teenagers who like spamming "every word they could think of that meant shitting or fucking" [Stone 1993]

What's your plan for dealing with people who harass each other, whether publicly and or DMs? [for drama, see Slack's failed attempt to have open DMs]

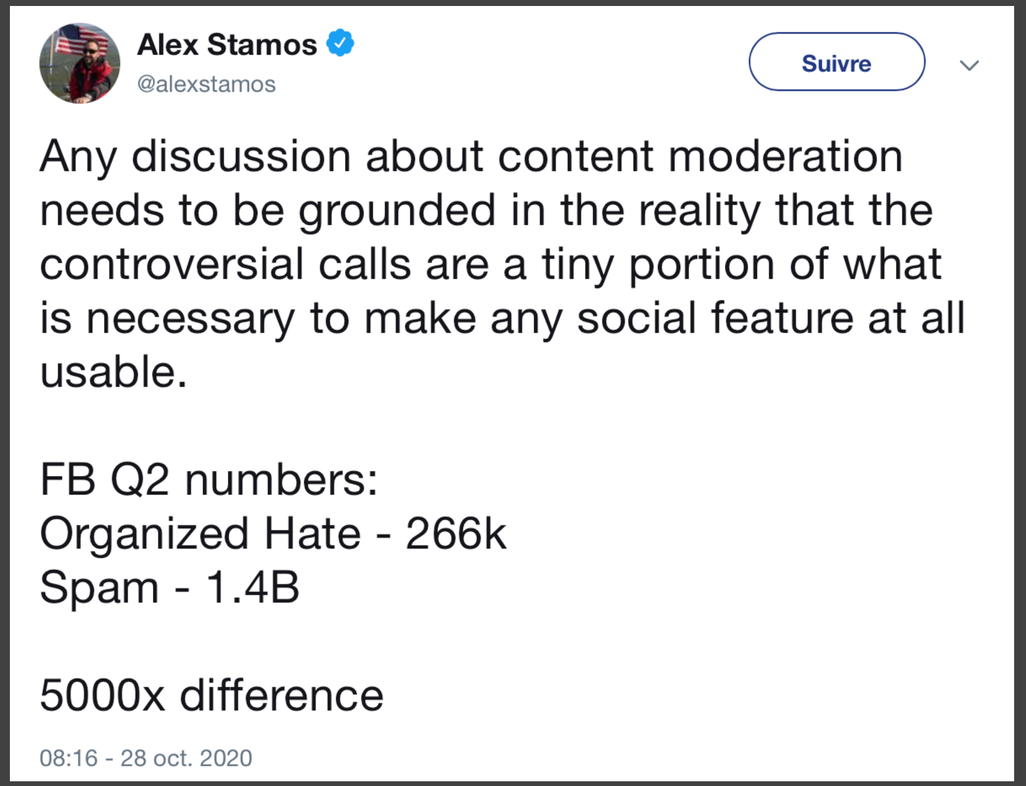
What's your plan for allowing or not allowing adult content: what's your line?

What does your reporting system look like? What types of things are you going to allow users to report? How do you plan to deal with people abusing the report feature?

What's your plan for dealing with content created in languages that you and your team don't speak?

Conflict and volume

There is a huge volume of moderation work to be done.

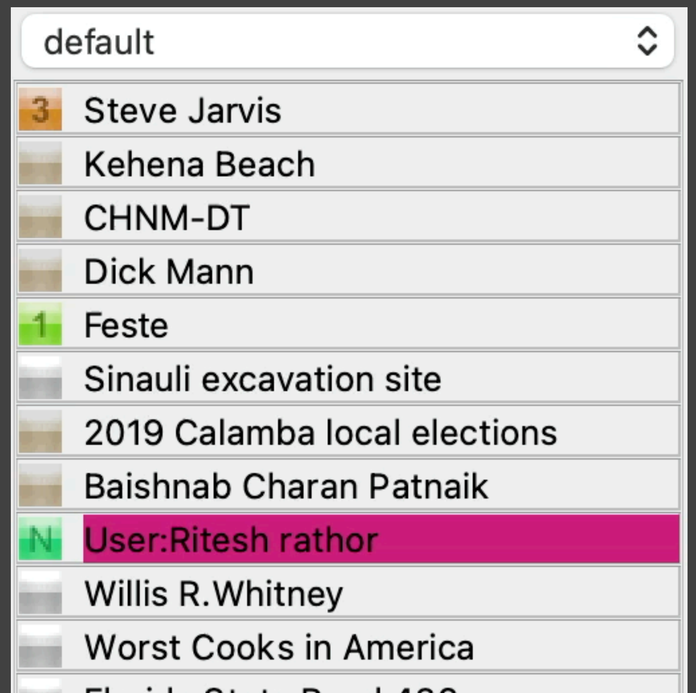


← Alex was the Director of the Stanford Internet Observatory, previously head of site integrity at Facebook, and teaches CS 152

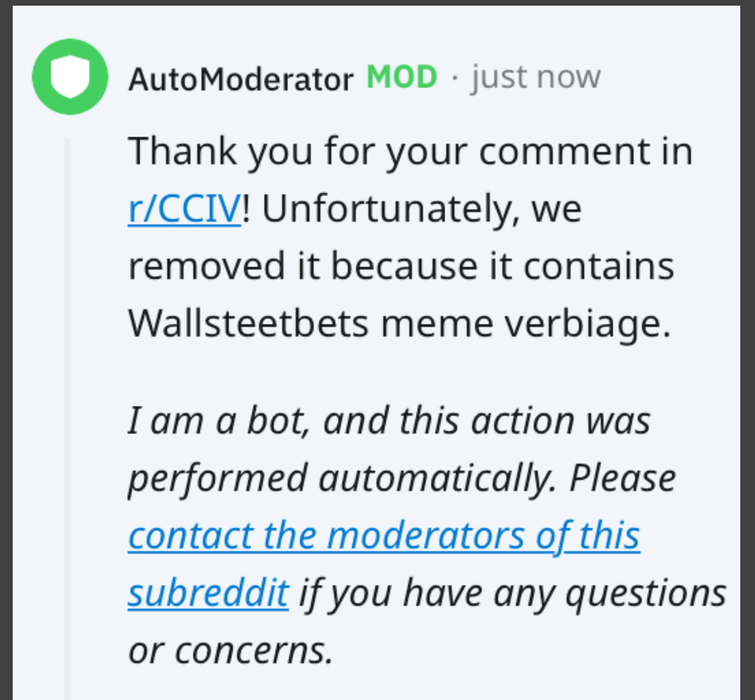
Human+AI moderation

Tools help facilitate moderator decisions by automatically flagging problematic posts, and providing relevant information.

Moderators often script tools if the platform API allows it

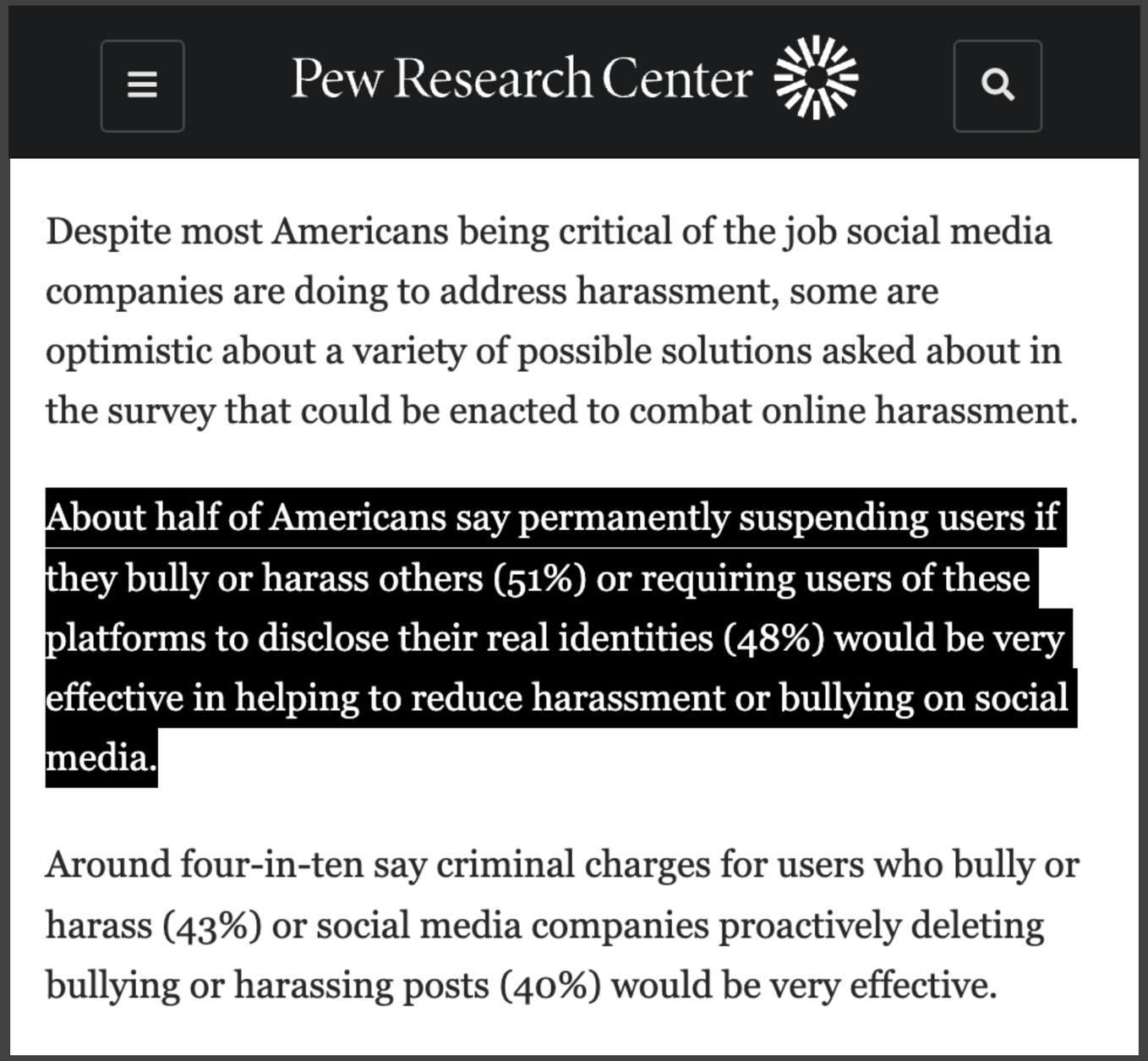


Wikipedia
Huggle



Reddit
AutoModerator

Does moderation work?



[Vogels 2021]

Are we correct? Does moderation work?

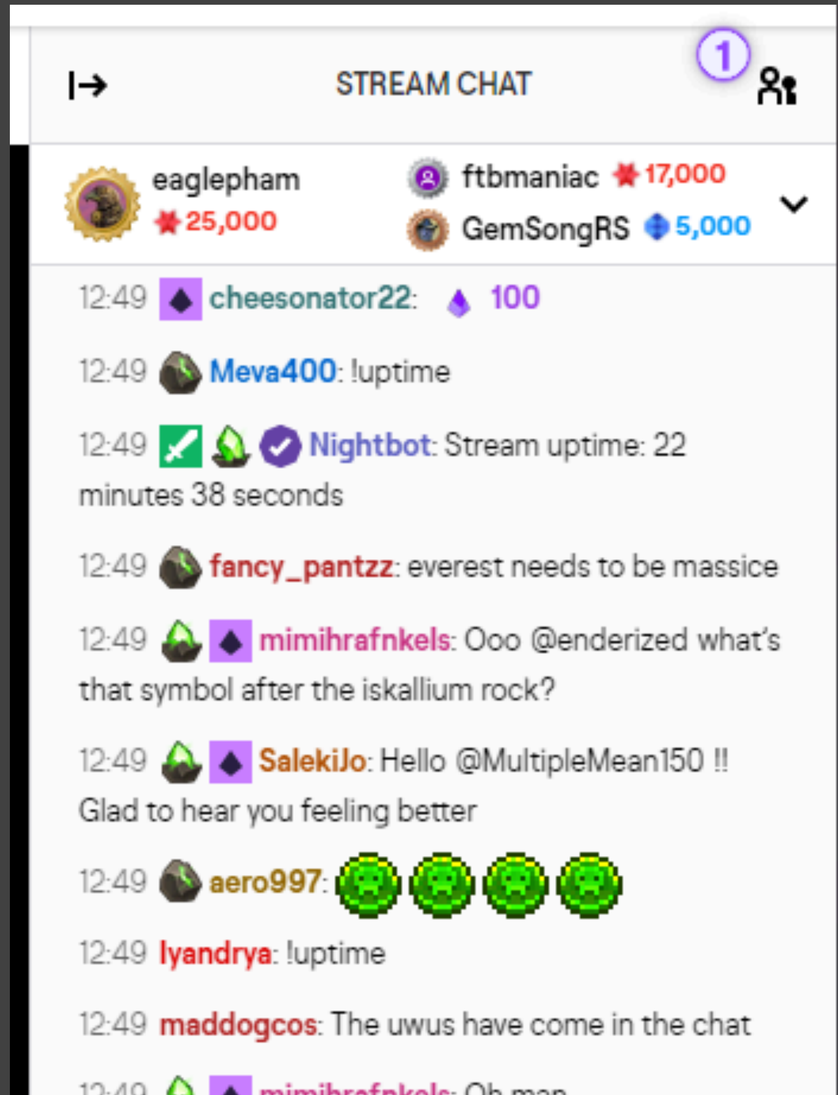
[1 min]

Yes, for short periods.

Moderation shifts descriptive norms and reinforces injunctive norms by making them salient.

Moderating content or banning substantially decreases negative behaviors in the short term on Twitch. [Seering et al. 2017]

Even algorithmic moderation reduces subsequent rule-breaking behavior [Horta Ribeiro, Cheng, and West 2023]



Deplatforming works.

After a toxic community is deplatformed, its members leave entirely, or migrate and drastically reduce their hate speech [Chandrasekharan et al. 2017]

Discussion reduces about the deplatformed individuals in mainstream spaces by 40-60% [Jhaver et al. 2021; Horta Ribeiro et al. 2024]

User activity reduces, not increases, on the deplatformed communities' new alt sites [Horta Ribeiro et al. 2021]

Deplatforming works. *

* However, those who migrate and stay active tend to further radicalize, albeit to smaller audiences [Horta Ribeiro et al. 2021]

* More recent work has found that, when Parler was taken off the App Store, other fringe platforms such as Gab and Rumble more than compensated [Horta Ribeiro et al. 2023]

It can also backfire.

Moderation can drive away newcomers, who don't understand the community's norms yet. [Growth lecture]

Users circumvent algorithmic controls

> Instagram hides #thighgap as as promoting unhealthy behavior…and users create #thygap instead [Chancellor et al. 2016]

Negative community feedback leads people to produce more negatively-reviewed content, not less. [Cheng et al. 2014]

BANNING WORDS ON INSTAGRAM TOTALLY BACKFIRED

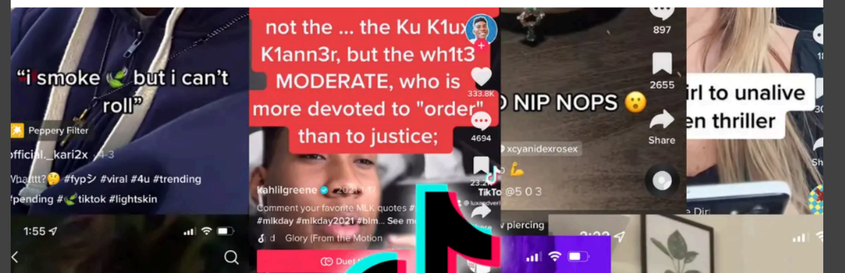
The New York Times

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.

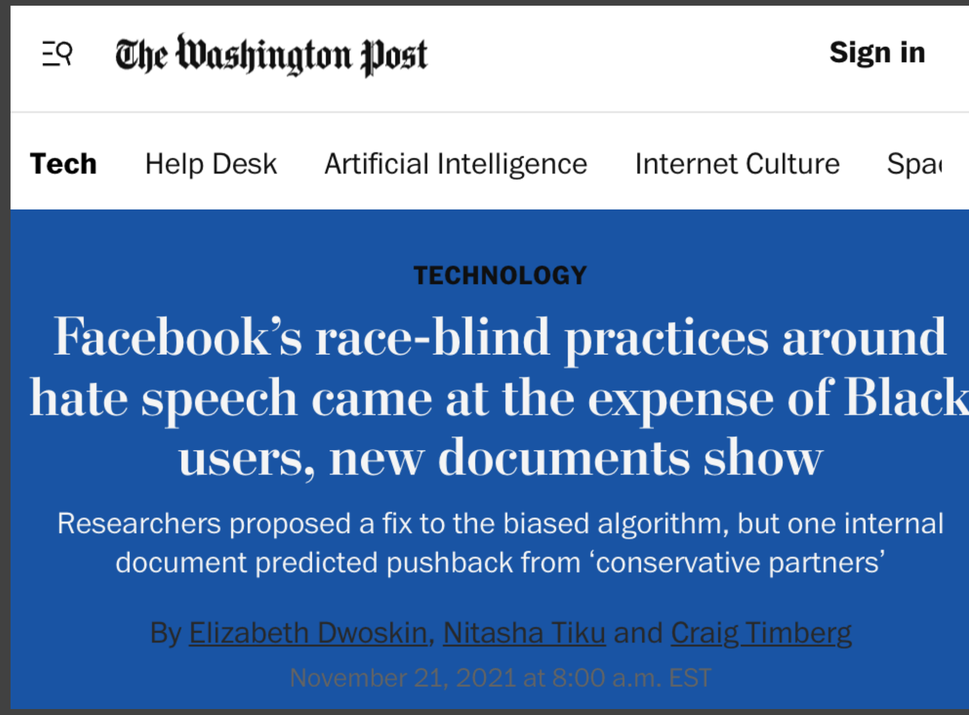
Give this article



Moderation's unequal impact

Three groups that tend to get their content moderated at higher rates: [Haimsen et al. 2021; Marshall 2021]

- 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



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?

How do you define which content constitutes…

- Nudity?
- Harassment?
- Cyberbullying?
- A threat?
- Suicidal ideation?

Recall:

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

A glimpse into the process

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…

(We would likely draw our lines 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)

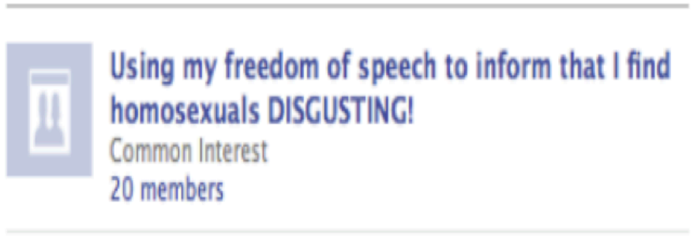
ANDing of three conditions

Hate Speech

REMOVE

What do we protect?

- Protected
  - Individuals
  - Groups
  - Humans



Hate Speech

ALLOWED

What do we NOT protect?

Not protected

- Concepts
- Institutions
- Beliefs

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

Thus, "I hate Christians" is banned, but "I hate Christianity" Facebook allows.

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 ethical and policy algebra to handle cases in this category

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.

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.

If it's dehumanizing, delete it.

Dismissing is different than dehumanizing.

On rules and regulations

The Washington Post

Sign in

Tech | Help Desk | Artificial Intelligence | Internet Culture | Spac

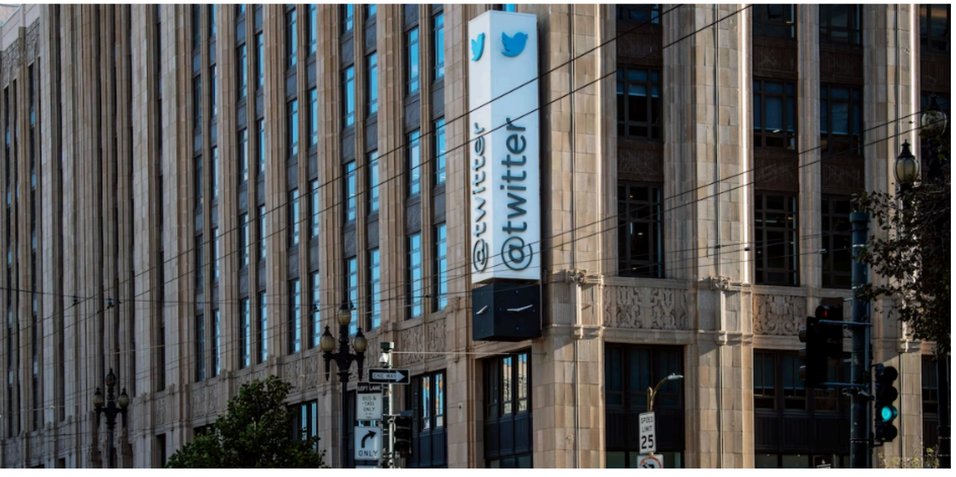
TECHNOLOGY

Musk's 'free speech' agenda dismantles safety work at Twitter, insiders say

Amid the wider turmoil since his takeover last month, Musk has moved rapidly to undermine Twitter's deliberative content moderation system

By Cat Zakrzewski, Faiz Siddiqui and Joseph Menn

November 22, 2022 at 8:07 a.m. EST



THE NEW YORKER

NEW YORKER FAVORITES | The Hitmakers Behind Rihanna | Where Prince Ch

OUR COLUMNISTS

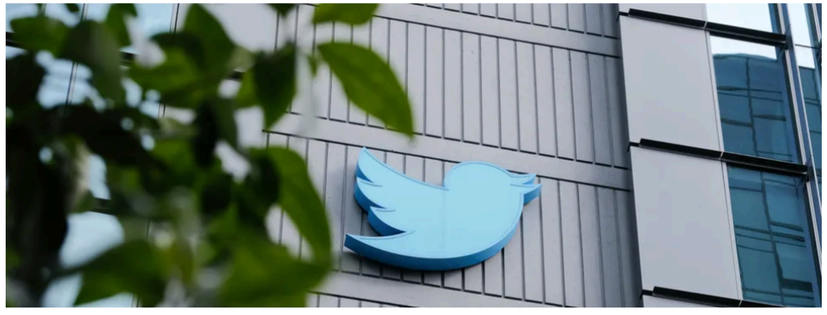
WHAT ELON MUSK DOESN'T KNOW ABOUT FREE SPEECH

The First Amendment does not protect one's right to have a social-media account, but today's dissent has mostly moved online, and, as a result, is privately owned.

Menu

By Jay Caspian Kang

December 6, 2022



The Atlantic

Subs

IDEAS

Elon Musk Is Right That Twitter Should Follow the First Amendment

A long history of free-speech jurisprudence backs him up.

By Jeffrey Rosen

Getty; The Atlantic

Why are we discussing this?

In the particular case of content moderation, legal policy has had a large impact on how social computing systems' manage their moderation approaches.

I hate Michael Bernstein.

Suppose I saw this on social media:

> Michael Bernstein is a [insert your favorite libel here]

Could I sue the social media platform?

Suppose I saw this in the New York Times:

> Michael Bernstein is a [insert your favorite libel here]

Could I sue the NYT?

Safe harbor

U.S. law provides what is known as safe harbor to platforms with user-generated content. This law has two intertwined components:

1. Platforms are not liable for the content that is posted to them. (You can't sue Discord for a comment posted to Discord, and I can't sue Piazza if someone posts a flame there.)
2. Platforms can choose to moderate content if they wish without becoming liable.

In other words, platforms have the right, but not the responsibility, to moderate. [Gillespie 2018]

Free speech

But don't we have this thing called the first amendment?

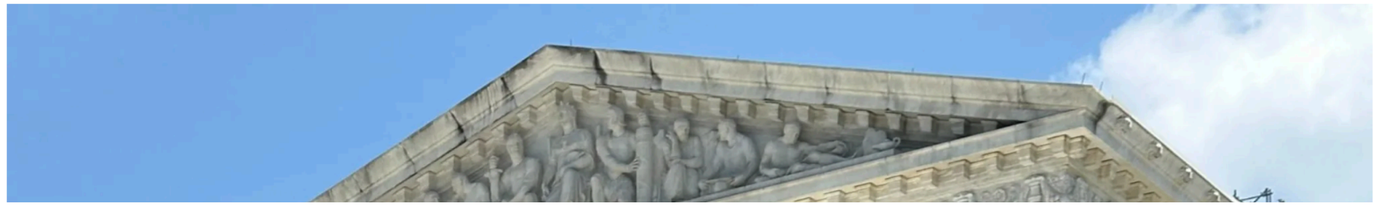
> *Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the government for a redress of grievances.*

Social computing platforms are not Congress. By law, they are not required to allow all speech. Even further: safe harbor grants them the right (but, again, not the responsibility) to restrict speech.

OPINION ANALYSIS

Justices side with Biden over government's influence on social media content moderation

By Amy Howe
on Jun 26, 2024



Supreme Court Clarifies First Amendment and Standing Standards Applicable to Social Media Content Moderation Policy Challenges

JUL 18, 2024

By Brett W. Johnson, P.C. and Ian Joyce

Social media companies have long moderated the type of content that appears on a person's home page by, for instance, deleting explicit posts or "downgrading" posts containing misinformation. Based on the belief that these policies tend to favor one side of the political spectrum, state governments and private actors have increasingly sought to curtail these moderation policies through regulations or private actions. Until recently, it was not entirely

Based on a series of 2024 Supreme Court cases, the current state of content moderation legality is:

- Content moderation is the companies' expressive activity protected by the First Amendment
- As a result, the government cannot force "viewpoint neutrality" or other requirements on platform content moderation

That said...

NewScientist

Sign in | Enter search keywo

News Features Newsletters Podcasts Video Comment Culture Crosswords | This week's magazine

Health Space Physics Technology Environment Mind Humans Life Mathematics Chemistry Earth Society

Analysis and Technology

Are tech firms giving up on policing their platforms?

Social media companies have long struggled with moderating the behaviour of billions of users, and now it seems they are finally giving up policing their platforms in favour of a crowdsourced approach – but will it work?

Meta eliminating fact-checking to combat "censorship"

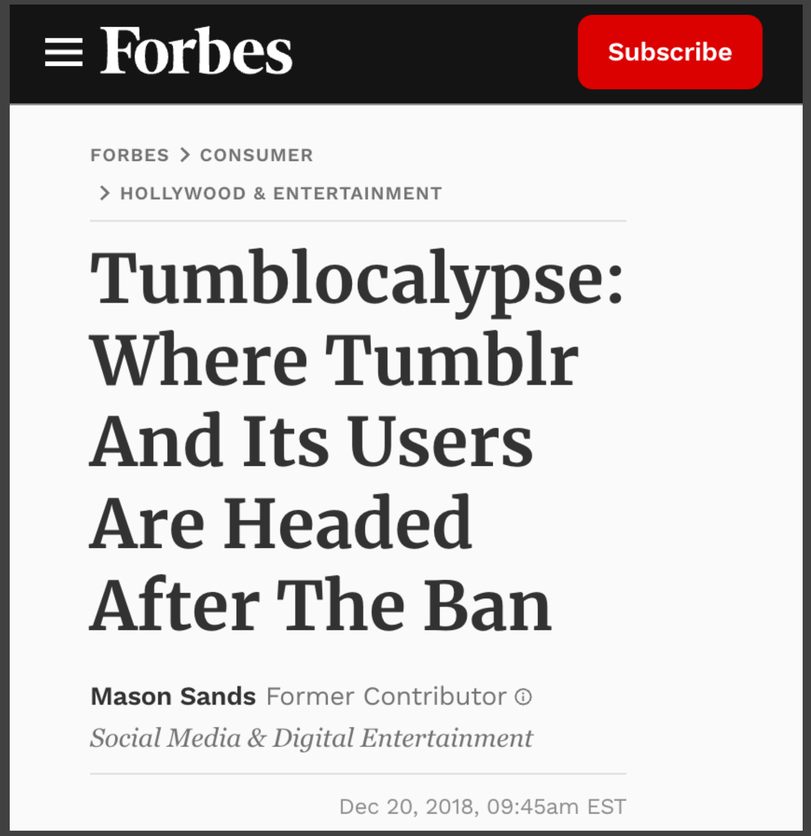
Avery Lotz, Sara Fischer

Takeaways

- Starting in the US, we are ending our third party fact-checking program and moving to a Community Notes model.
- We will allow more speech by lifting restrictions on some topics that are part of mainstream discourse and focusing our enforcement on illegal and high-severity violations.
- We will take a more personalized approach to political content, so that people who want to see more of it in their feeds can.

“History doesn’t repeat itself, but it often rhymes”

— Mark Twain



Summary

As Gillespie argues, moderation is the commodity of the platform: it sets apart what is allowed on the platform, and has downstream influences on descriptive norms.

Moderation works: it can change the community's behavior

Moderation classification rules are fraught and challenging — they reify what many of us carry around as unreflective understandings.

References

Adler, Simon, J. Abumrad, and R. Krulwich. "Post no evil." Episode. RadioLab. New York City, New York: WNYC Studios (2018).

Alkhatib, Ali, and Michael Bernstein. "Street-level algorithms: A theory at the gaps between policy and decisions." Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019.

Chancellor, Stevie, et al. "# thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities." Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing. 2016.

Chandrasekharan, Eshwar, et al. "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 (2017): 1-22.

Chen, Adrian. "The human toll of protecting the Internet from the worst of humanity." The New Yorker 28 (2017). https://www.newyorker.com/tech/annals-of-technology/the-human-toll-of-protecting-the-internet-from-the-worst-of-humanity

Cheng, Justin, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. "How community feedback shapes user behavior." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 8. No. 1. 2014.

Fiesler, Casey, Shannon Morrison, and Amy S. Bruckman. "An archive of their own: A case study of feminist HCI and values in design." Proceedings of the 2016 CHI conference on human factors in computing systems. 2016.

Gillespie, Tarleton. "Do Not Recommend? Reduction as a Form of Content Moderation." Social Media+ Society 8.3 (2022): 20563051221117552.

References

Gillespie, Tarleton. Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press, 2018.

Gordon, Mitchell L., et al. "The disagreement deconvolution: Bringing machine learning performance metrics in line with reality." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.

Haimson, Oliver L., et al. "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 (2021): 1-35.

Horta Ribeiro, Manoel, et al. "Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-24.

Horta Ribeiro, Manoel, et al. "Deplatforming did not decrease Parler users' activity on fringe social media." PNAS nexus 2.3 (2023): pgad035.

Horta Ribeiro, Manoel, Justin Cheng, and Robert West. "Automated content moderation increases adherence to community guidelines." Proceedings of the ACM web conference 2023. 2023.

Horta Ribeiro, Manoel, et al. "Deplatforming Norm-Violating Influencers on Social Media Reduces Overall Online Attention Toward Them." arXiv e-prints (2024): arXiv-2401.

Hutch. "The Mind of a Mod – An Interview With a Reddit Moderator." 2019. https://thebetterwebmovement.com/interview-with-reddit-moderator-unoeatnosleep/

References

Jhaver, Shagun, et al. "Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-30.

Lampe, Cliff, and Paul Resnick. "Slash (dot) and burn: distributed moderation in a large online conversation space." Proceedings of the SIGCHI conference on Human factors in computing systems. 2004.

Mahar, Kaitlin, Amy X. Zhang, and David Karger. "Squadbox: A tool to combat email harassment using friendsourced moderation." Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 2018.

Marshall, Brandeis. "Algorithmic misogynoir in content moderation practice." Heinrich-Böll-Stiftung European Union (2021).

Matias, J. Nathan. "The civic labor of volunteer moderators online." Social Media+ Society 5.2 (2019): 2056305119836778.

Newton, Casey. "The trauma floor: The secret lives of Facebook moderators in America." The Verge 25.02 (2019). https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Perez, S. "YouTube promises to increase content moderation and other enforcement staff to 10K in 2018." (2018). https://techcrunch.com/2017/12/05/youtube-promises-to-increase-content-moderation-staff-to-over-10k-in-2018

Reuters. "TikTok restructures trust and safety team, lays off staff in unit, sources say." (2025). https://www.reuters.com/technology/tiktok-restructures-trust-safety-team-lays-off-staff-unit-sources-say-2025-02-20/

References

Seering, Joseph, Geoff Kaufman, and Stevie Chancellor. "Metaphors in moderation." New Media & Society 24.3 (2022): 621-640.

Seering, Joseph, Robert Kraut, and Laura Dabbish. "Shaping pro and anti-social behavior on twitch through moderation and example-setting." Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. 2017.

Seering, Joseph. Personal communication. 2021.

Stone, Allucquere Rosanne. "What vampires know: Transsubjection and transgender in cyberspace." In Control: Mensch-Interface-Maschine (1993).

Vogels, Emily A. "The state of online harassment." Pew Research Center 13 (2021): 625.

Social Computing

CS 278 | Stanford University | Michael Bernstein

Creative Commons images thanks to Kamau Akabueze, Eric Parker, Chris Goldberg, Dick Vos, Wikimedia, MaxPixel.net, Mescon, and Andrew Taylor.