Content Moderation

CS 278 | Stanford University | Michael Bernstein

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 become a troll.

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

For more, listen to Radiolab's excellent "Post No Evil" episode

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, fine. Nudity is when you can see the nipple and areola. The baby will block 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, fine. Nudity is when you can see the nipple and areola. The baby will block 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 kicking a ball if the ball was kicked and is now midair?

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.

…

Radiolab quote on Facebook's moderation 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."

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

Moderation is the actual commodity of any social computing system.

Today

How do platforms moderate?

How should they moderate?

Recall: moderation's effects

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

Reddit's ban of /r/CoonTown and /r/fatpeoplehate due to violations of anti-harassment policy succeeded: accounts either left entirely, or migrated to other subreddits and drastically reduced their hate speech. [Chandrasekharan et al. 2017]

Today: how do we do it?

“Three imperfect solutions”

h/t Gillespie [2018]

Paid moderation

Rough estimates:

- ~15,000 contractors on Facebook [Statt 2018, theverge.com],
- ~10,000 contractors on YouTube [Popper 2017, theverge.com]

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

Bloomberg Subscribe

Technology

Facebook to Raise Pay for Thousands of Contract Workers, Including Content Moderators

Social-media giant says current minimum wage isn't enough in expensive areas such as Silicon Valley

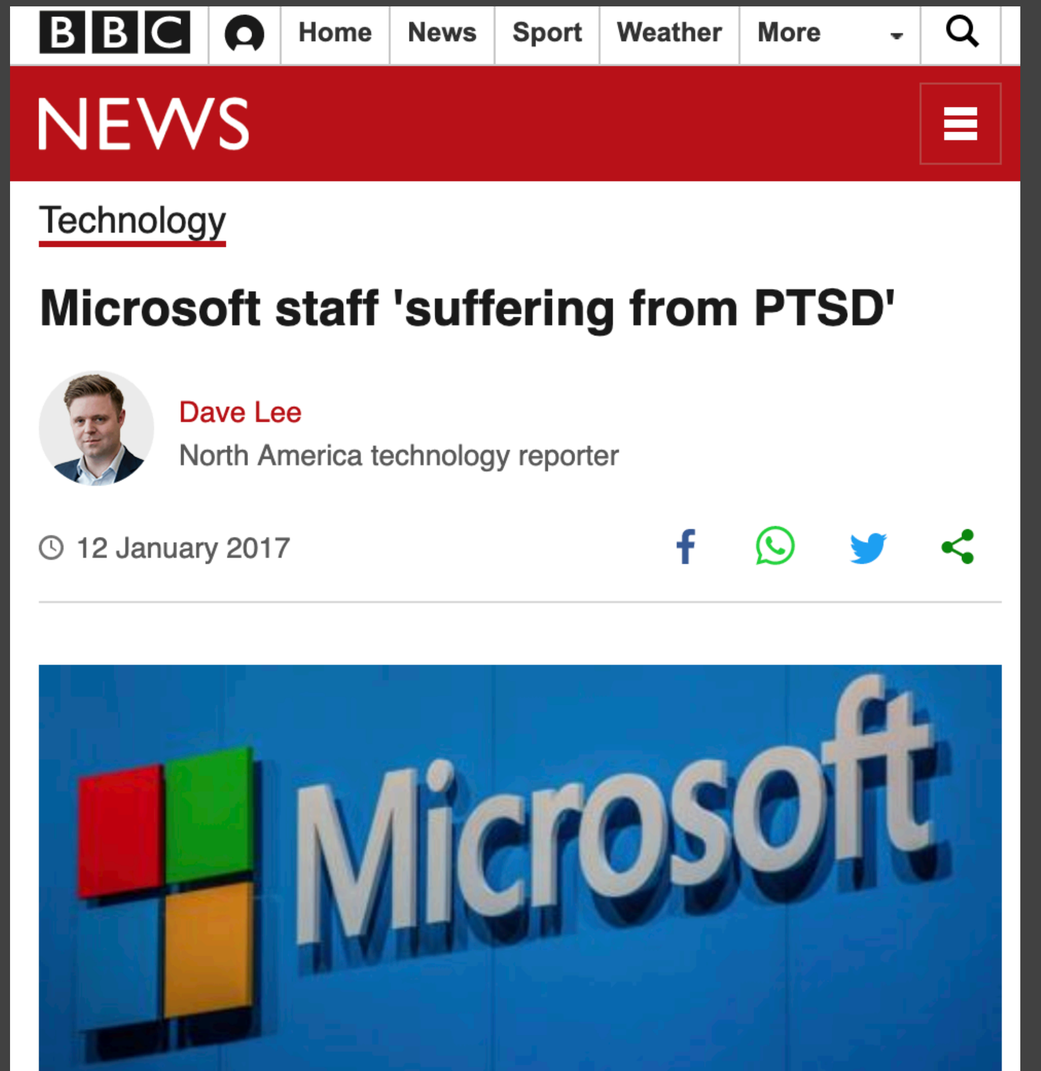
By Kurt Wagner
May 13, 2019, 9:15 AM PDT
Updated on May 13, 2019, 11:21 AM PDT

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

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



Paid moderation

Strengths

- A third party reviews any claims, which helps avoid brigading and supports more calibrated and neutral evaluation.

Weaknesses

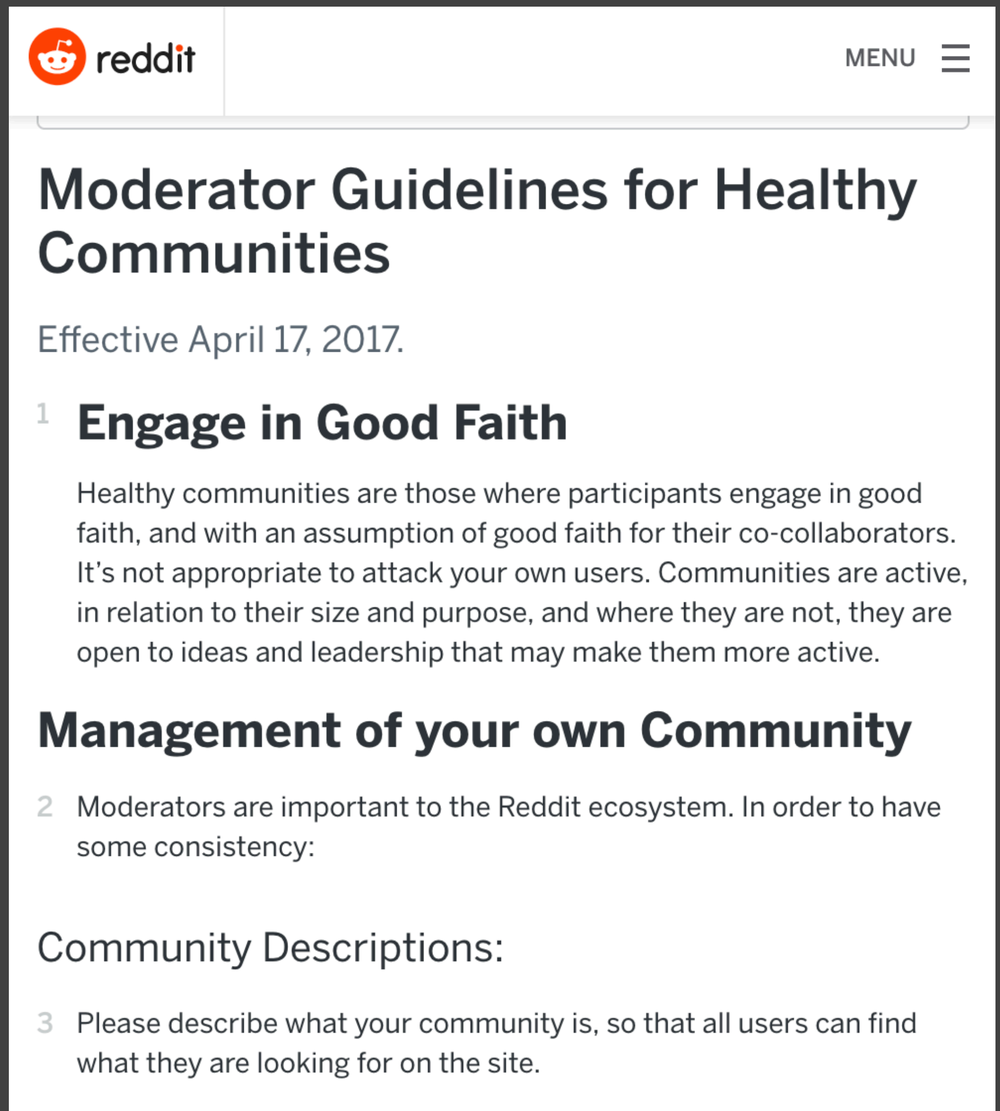
- Major emotional trauma and PTSD for moderators.
- Evaluators may have only seconds to make a snap judgment.

Community moderation

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

Examples: Reddit, Twitch, Steam

It's best practice for the moderator team to publish their rules, rather than let each moderate 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 including /r/politics, /r/history, /r/futurology, and /r/listentothis

[https://thebetterwebmovement.com/interview-with-reddit-moderator-unoeatnosleep/]

Community moderation

Strengths:

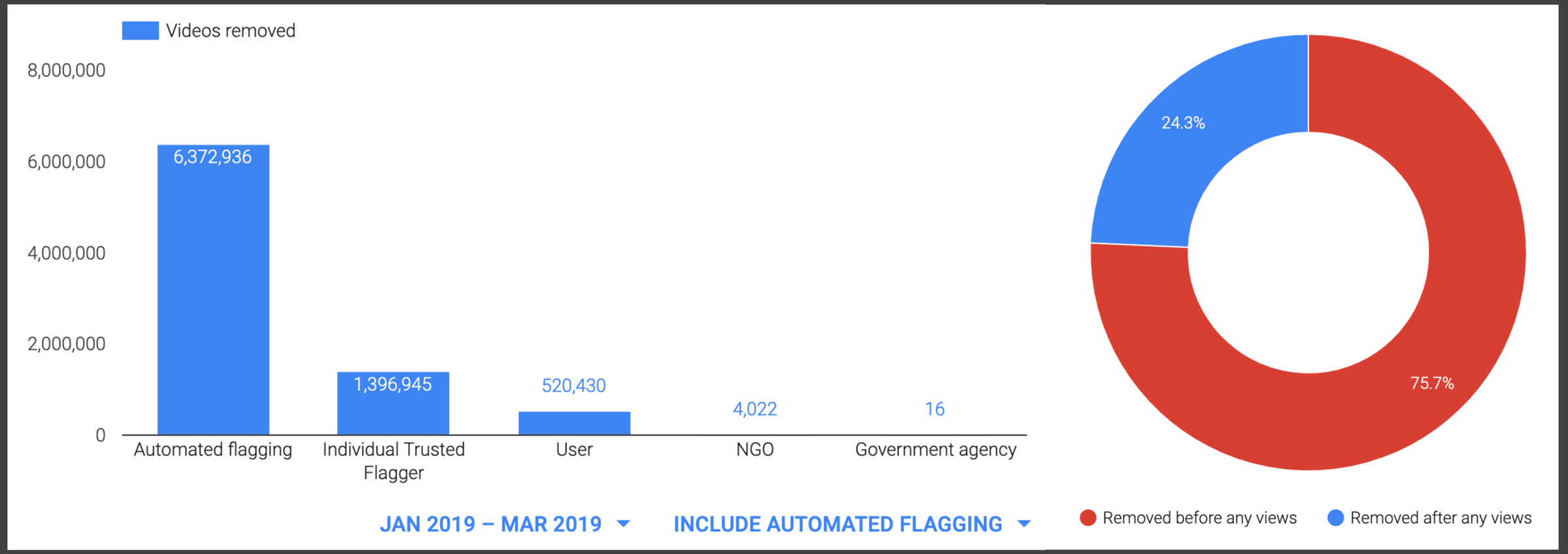
- Leverages intrinsic motivation
- Local experts are more likely to have context to make hard calls

Weaknesses:

- Mods don't feel they get the recognition they deserve
- Resentment that the platform makes money off free labor
- Not necessarily consistent, fair, or just

Algorithmic moderation

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



Algorithmic moderation

Examples of errors via Ali Alkhatib [2019, al2.in/street]

Google's Anti-Bullying AI Mistakes Civility for Decency

by **Jillian York**, motherboard.vice.com
August 18, 2017 12:00 PM

Something is wrong on the internet

by **James Bridle**, medium.com
November 6, 2017 10:09 AM

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

by **Megan Farokhmanesh**, theverge.com
June 4, 2018 01:46 PM

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, transparent and accountable (FAT*) algorithm were possible, culture would evolve and training data would become out of date [Alkhatib 2019].

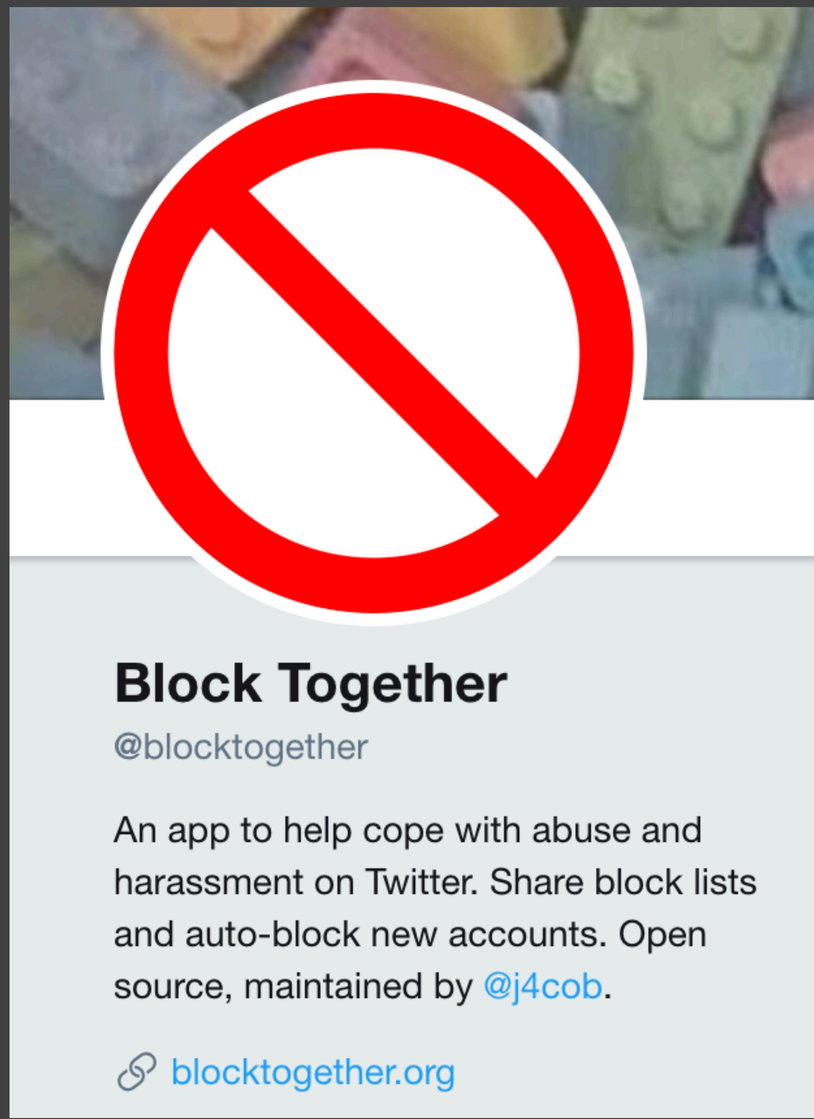
Fourth option: blocklists

When the platform can't provide, users take it into their own hands

Blocklists are lists of users who a community has found are toxic and should be blocked. These lists are shared amongst community members. [Geiger 2016]

Strengths: can succeed when platforms don't

Weaknesses: no due process, so many feel blocked unfairly [Jhaver et al. 2018] (…not that other approaches have due process either.)



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.

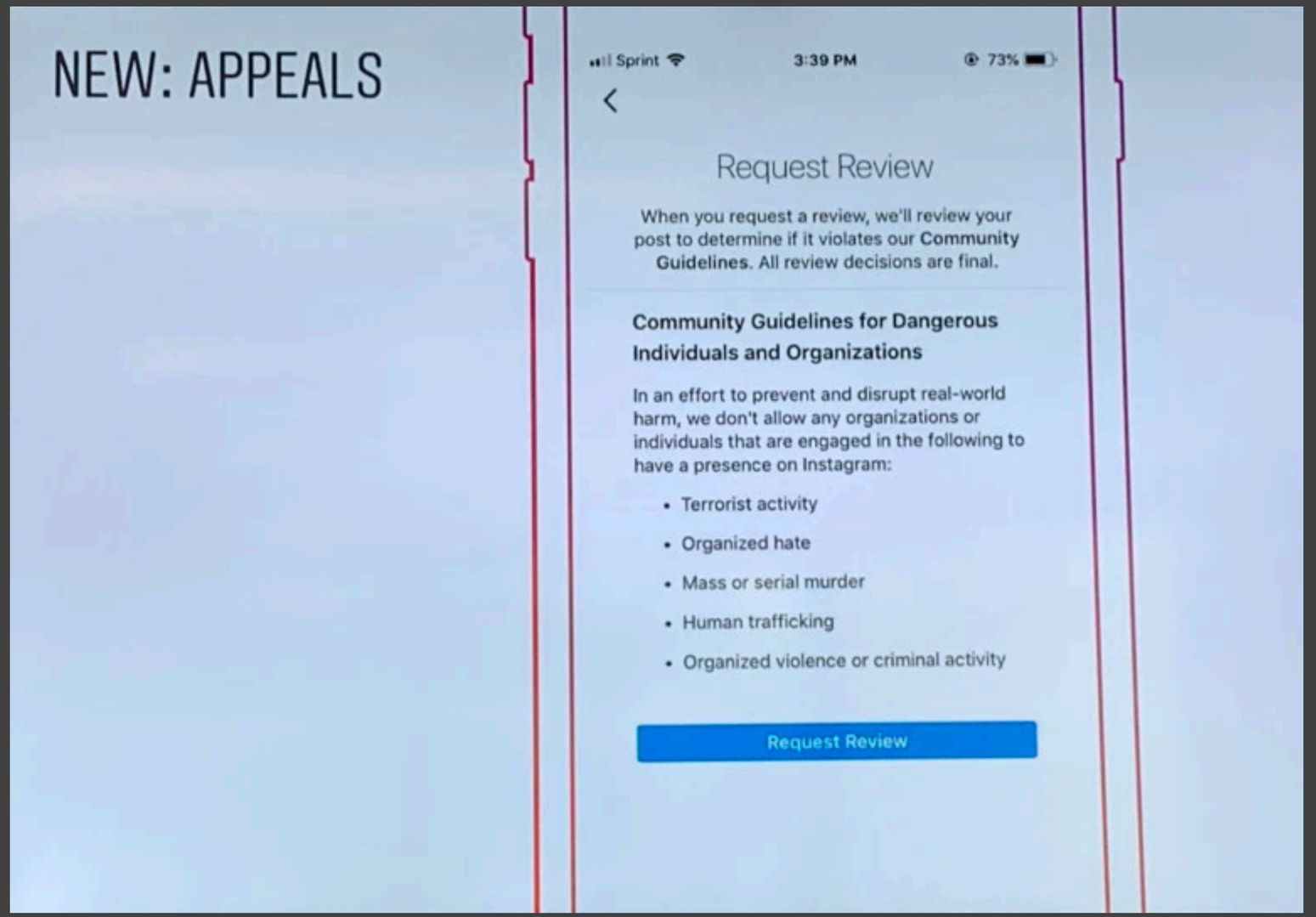
Tier II: Human moderation, paid or community depending on the platform. Moderators monitor flagged content, review an algorithmically curated queue, or monitor all new content, depending on platform.

Appeals

Most modern platforms allow users to appeal unfair decisions.

If the second moderator disagrees with the first moderator, the post goes back up.

Instagram, last week →



Moderation and classification

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…

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.

Classification and its consequences [Bowker and Star 1999]

We live in a world where ideas get classified into categories.

These classifications have import:

- Which conditions are classified as diseases and thus eligible for insurance
- Which content is considered hate speech and removed from a platform
- Which gender options are available in the profile dropdown
- Which criteria enable news to be classified as misinformation

Classification + moderation

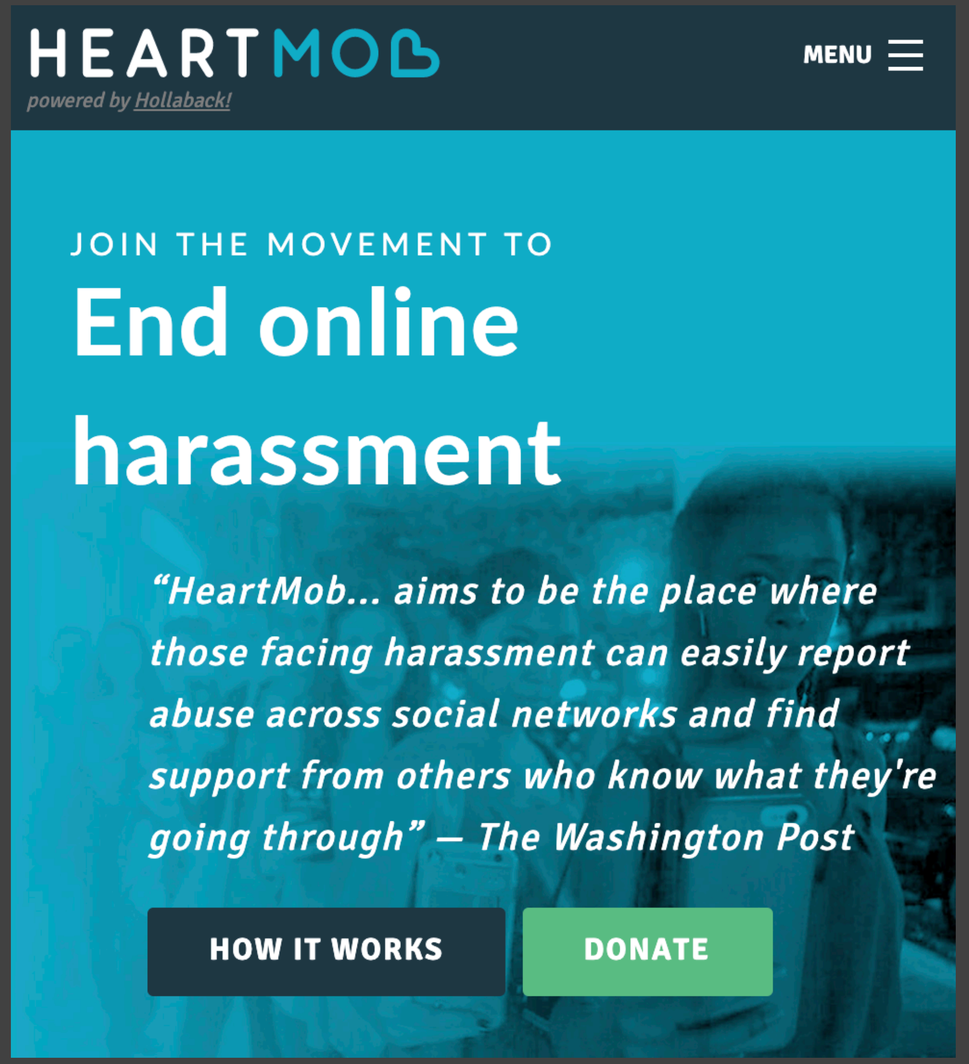
Specifics of classification rules in moderation have real and tangible effects on users' lives, and of the norms that develop on the platform.

Typically, we observe the negative consequences: a group finds that moderation classifications are not considerate of their situation, especially if that group is rendered invisible or low status in society.

Classification + moderation

To consider a bright side: classification can also be empowering if used well.

On HeartMob, a site for people to report harassment experiences online, the simple act of having their experience classified as harassment helped people feel validated in their experiences. [Blackwell et al. 2017]



Design implications

When developing moderation rules, think about which groups your classification scheme is rendering invisible or visible.

Even if it's a "utilitarian document" (vis a vis Facebook earlier), it's viewed by users as effective platform policy.

But, remember that not moderating is itself a classification decision and a design decision. Norms can quickly descend into chaos without it.

On rules and regulations

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 Twitter:

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

Could I sue Twitter?

Suppose I saw this in the New York Times:

> Michael Bernstein is a [insert your favorite libel or threat 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.

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.

The three common approaches to moderation today are paid labor, community labor, and algorithmic. Each brings tradeoffs.

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

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.