Anti-social computing example

Julia R: Video example of trolling a class —
Content moderation
CS 278 | Stanford University | Michael Bernstein
content warning: moderation policy documents describing revenge porn, hate speech, and harassment of minority groups
Announcements

Exam question bank is published! You may study in groups, but notes must be taken individually.

Exam is closed-book, closed-notes, next Tuesday. Bring a pencil!

OAE rooms have been sent out

Assignment 3 reflection due a week from Friday, after the exam

No reading this week or next week
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

[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.
Facebook

Mums furious as Facebook removes breastfeeding photos

Mark Sweney

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,
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?

Fine. Nudity means the nipple and areola are visible. Breastfeeding blocks those.
Forget it. It's nudity and disallowed unless the baby is actively nursing.

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 Friendoing 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 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],

~10,000 contractors on YouTube [Perez 2017]

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

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
- Not necessarily consistent
- Without oversight, mods can grow problematic communities
Algorithmic moderation

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

Perspective API can help mitigate toxicity and ensure healthy dialogue online.
Feed algorithms are an instrument of moderation

[Gillespie 2022]
Algorithmic errors

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

[Alkhatib and Bernstein 2019]

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

[Motherboard]

[The Verge]

[Deceiving Google’s Perspective API Built for Detecting Toxic Comments]
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
Example: Facebook

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

Moderators are responsible for:
Removing comments, banning users in real time
Example: Reddit

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

Unpaid and abused: Moderators speak out against Reddit
Keeping Reddit free of racism, sexism and spam comes with a mental health risk.

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

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.
Even in systems like Archive of Our Own that are light on moderation, content debates rage. [Fiesler, Morrison, and Bruckman 2016]
Example: Email

[SQUADBOX](#)

**SQUADBOX**

Fight back against harassment.

Try it out!

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.

[Mahar, Zhang, and Karger 2018]

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

*With Squadbox, friends moderate harassing messages in your email*

MIT researchers developed it as a way to mitigate online harassment.
So...what do we do?

Many social computing 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 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.

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

Alex is the Director of the Stanford Internet Observatory, previously head of site integrity at Facebook, and teaches CS 152
Tools help facilitate moderator decisions by automatically flagging problematic posts, and providing relevant information.

Moderators often script tools if the platform API allows it.
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?
Are we correct? Does moderation work?

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.

[Vogels 2021]
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]
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 [Jhaver et al. 2021]

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]
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]
Moderation’s unequal impact

Three groups that tend to get their content moderated at higher rates: [Haimsan 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)
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)
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.
Is this solution good?

(And what does “good” even mean in this context?) [2min]

For later: you can check out examples to think through at https://trustandsafetyfoundation.org/case-studies/
On rules and regulations
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
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.
We await possible changes…
“History Doesn’t Repeat Itself, but It Often Rhymes”
— Mark Twain

Facebook
Mums furious as Facebook removes breastfeeding photos

Mark Sweeney
@marksweney
Tue 30 Dec 2008 08:17 EST

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".

Tumblocalypse: Where Tumblr And Its Users Are Headed After The Ban

Mason Sands
Former Contributor
Social Media & Digital Entertainment

Dec 20, 2018, 09:45am EST
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


References


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


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.

Slide content shareable under a Creative Commons Attribution-NonCommercial 4.0 International License.