Reply in Zoom chat:
If you’re a moderator for any communities: are you light touch or heavy handed?
I fine-tuned some AI language models on your submitted midterm questions.
“What are some of the differences between Social Proof and Unproof Proof?”
“Why is a hyperpersonal mode so appealing? What are the potential harms and benefits of this type of system?”
“Ask the following question on your next goographer: Which group collaboration tool of your choice is more effective?”
“How do you calculate the overhead of a commons-like system and a communism-like system on the same system?”
“Ask a University of California, Berkeley, student if they were to create a platform that rewarded each undergraduate with a goldfish and each cup of coffee, how would they achieve this goal? Please explain your answer(s) using reference 7 black and white and explain why goldfish and coffee are not rewarded in the trophy explaining why goldfish and purple coffee do not.”
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.
Mums furious as Facebook removes breastfeeding photos

Mark Sweney

@marksweney  Email
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".
Fine, fine. Nudity is when you can see the nipple and areola. The baby will block those.

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 punching someone if the person is currently recoiling from the hit?
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.
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.”
Moderation is the actual commodity of any social computing system.

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

Moderation as invisible labor and classification

Does moderation work?

Regulation and moderation
Invisible labor and moderation
Invisible labor [Star and Strauss 1999]

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

Examples of invisible labor in social computing systems:

- Moderation
- Paid data annotation [Irani and Silberman 2013; Gray and Suri 2019]
- Server administration
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.
Even in systems like Archive of Our Own that are light on moderation, content debates rage.
Example: Email

[SQUADBOX]
Fight back against harassment.

[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.

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.
Why is the labor invisible?

Because all that most people see when they arrive is the results of the curation, not the curation happening. When was the last time you saw Facebook’s army of moderators change the content of your feed?

The invisible nature of this labor makes moderation feel thankless, and the content that mods face can prompt PTSD and emotional trauma. <3 your mods.
Does moderation work?
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]
Stronger actions work too.

Reddit’s ban of two subreddits 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]

🤔 Studies of police surges into IRL neighborhoods just shift crime elsewhere. Why the different outcome here?
However, it can 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]
So how do we walk this line?

For moderation to set and maintain norms, it’s best if the lines are drawn clearly up-front and enforced clearly and visibly from the beginning.

Trying to change the rules later is essentially changing the social contract, so you get far more pushback (e.g., #thyghgap)

What do you think — should Facebook/Instagram change their policies? [2min]
Moderation and classification

Content warning: definitions of revenge porn, hate speech
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)
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)*
If it’s dehumanizing, delete it.

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.
Is this solution good?

(What does “good” mean in this context?)

[3min]
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.

**Facebook**

Mums furious as Facebook removes breastfeeding photos
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. [Alkhatib and Bernstein 2019]

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.

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.
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.