

CS 4873-A: Computing and Society

Munmun De Choudhury | Associate Professor | School of Interactive Computing



Week 11: Freedom of Speech
March 28, 2021





Implications of Internet Technologies



Censorship

(Warf 2010)

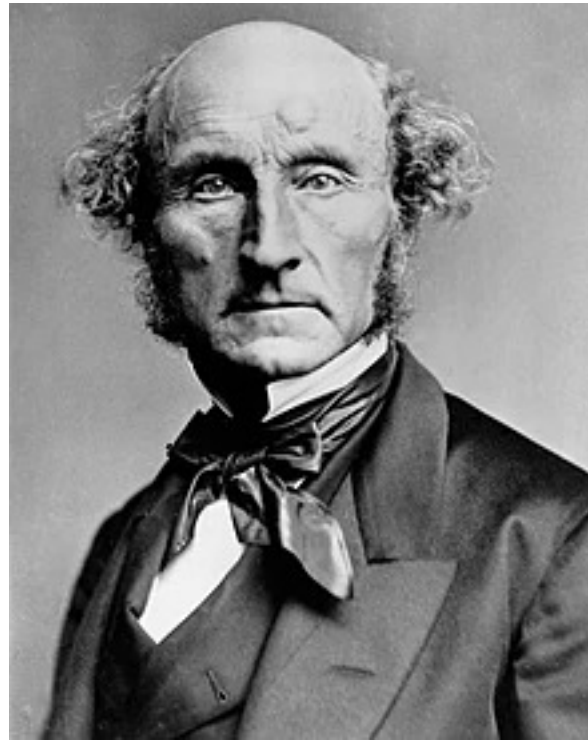
Censorship: Definition and History

- *Censorship is the attempt to suppress or regulate public access to material considered offensive or harmful*
- Forms of censorship
 - Direct censorship
 - Gov't monopoly, e.g., former USSR
 - Prepublication review; e.g., can't publish classified material
 - Licensing & registration, e.g., TV stations must comply with decency laws or lose license
 - Self-censorship
 - CNN suppressed negative reports on Iraqi gov't to keep Bagdad Bureau open
 - Publishers wanting to maintain good relationship with the government
 - Voluntary rating systems, like the mature label on games



Is Censorship Ethical?

John Stuart Mill



Kant's vs. Mill's Views on Censorship

- Radically different ethical theories, but had similar views on censorship

Kant's View

- Kant asked: “Why don't people think for themselves?”
- He replied rhetorically: “Laziness and cowardice are the reason why so great a portion of mankind, after nature has long since discharged them from external direction, nevertheless remain under lifelong tutelage, and why it is so easy for others to set themselves up as their guardians”
- Kant believed he lived in a time in which many obstacles prevented people exercising their own reason

Mill's View

- Mill championed freedom of expression
- He offered four reasons
 - Preventing someone from voicing their concern could be silencing truth
 - A person can be erroneous, but all opinions need to be heard to assess the whole truth
 - Truth needs to be rationally tested and validated
 - An opinion that has been tested through open discourse is likely to have a “vital effect on the character and conduct”

Mill's Principle of Harm

- “The only purpose for which power can be rightfully exercised over by any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant”
- Why use of adult porn by adults should not be censored by the government



Discussion Point 1:

Does the Internet pose new challenges to censorship? How?



Platform measures



Children and Inappropriate Content



Child Internet Protection Act

Censorship and Internet

- Warf (2010) mapped the severity of censorship worldwide and assesses the numbers of people affected, and used the Freedom House index to correlate political liberty with penetration rates.
- Many governments employ filtering of or restricting access to certain Internet content
 - North Korea
 - Middle East
 - China
 - Germany
 - United States

Censorship and Internet (Warf 2010)

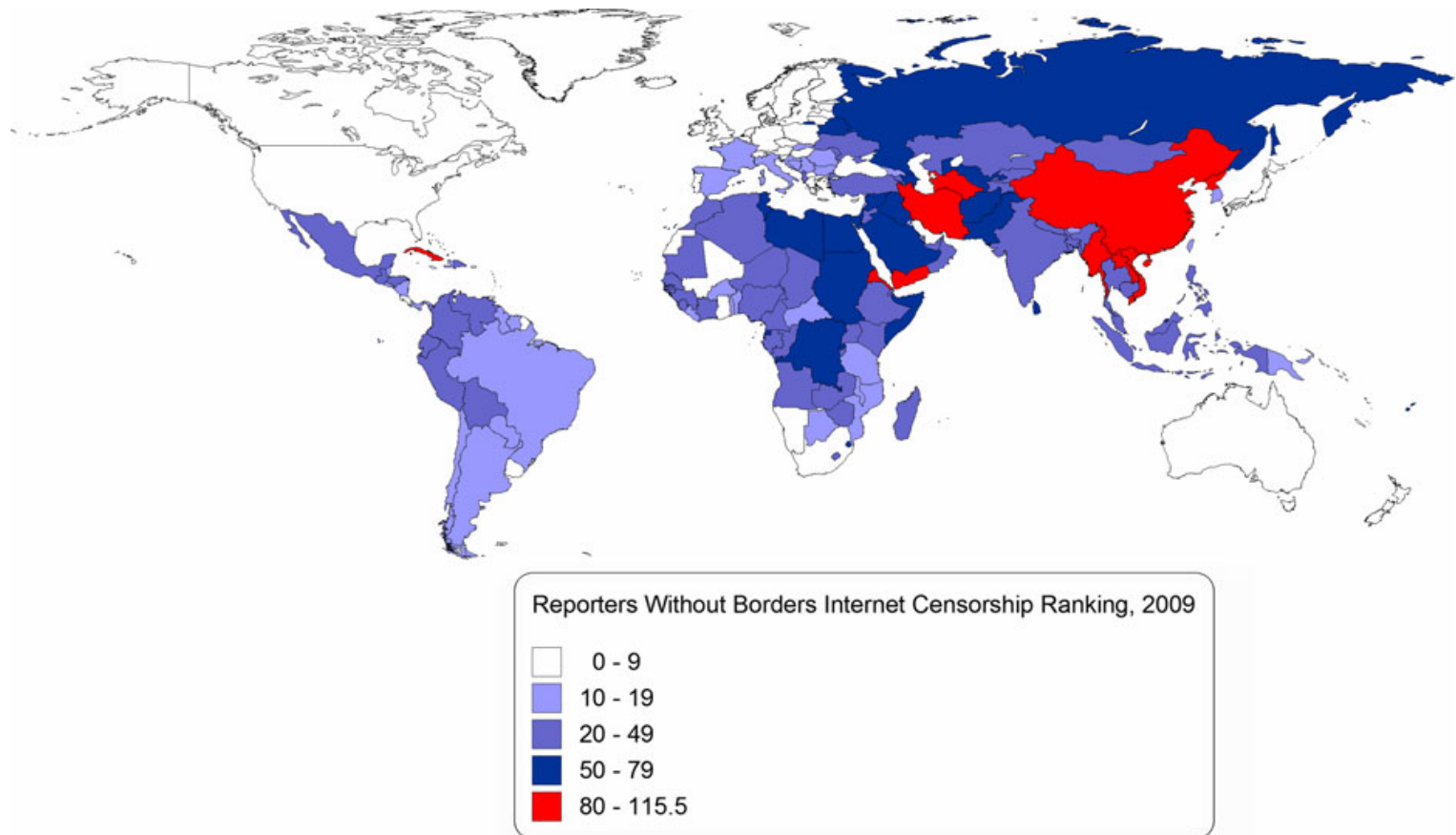


Fig. 2 Reporters Without Borders Internet Censorship Ranking 2009. *Source:* data drawn from <http://www.rsf.org/en-classes/ment1003-2009.html>



Freedom of Expression




First Amendment




First Amendment



First Amendment



Discussion point: In the US, television commercials for cigarettes are banned. Should there be a ban on commercials for violent video games too?



Discussion point: Should people publishing accusations against others on their blogs or Facebook pages be held responsible if they disseminate false information (e.g., fake news, anti-vax content, COVID-19 infodemic)?




Spam

Spam

- What is spam?
- With ease of internet access, businesses looked for ways to capitalize on market opportunities associated with Internet communications – easier/cheaper to send emails than physical mails
 - How to find email addresses though?
 - Crawling the web; scrape address books with viruses; listen to chatroom conversations; sneaky way to sign up; dictionary attacks on ISPs
- This entrepreneurial behavior has given rise to a new set of legal and ethical problems



Spam Case Study



Discussion point: Why is “cold calling” considered to be an acceptable sales practice, but spamming isn’t?

CS 4873-A: Computing and Society

Munmun De Choudhury | Associate Professor | School of Interactive Computing



Week 11: Regulating Online Speech
March 28, 2021



- 1 44 Any women have experience with having a baby while in grad school? (self.GradSchool)
submitted 12 hours ago by HigHog
29 comments share
- 2 10 Hitler doesn't get a postdoc (youtube.com)
submitted 5 hours ago by Epistaxis (PhD, genetics)
5 comments share
- 3 6 I think it's time to quit (self.GradSchool)
submitted 5 hours ago * by Amelorn
11 comments share
- 4 Advice: Masters or go straight for PhD (self.GradSchool)
submitted 12 minutes ago by Cagedcrab
comment share
- 5 For Profit or Not University (self.GradSchool)
submitted 19 minutes ago by lilferncat
comment share
- 6 First time writing an state of purpose !! pls some help (self.GradSchool)
submitted 25 minutes ago by nautiluz92
comment share
- Does it make sense to take a mortgage (if I have the downpayment) than rent out a studio for my 5 year PhD program? (self.GradSchool)

MY SUBREDDITS FRONT - ALL - RANDOM | ASKREDDIT - FUNNY - PICS - WORLDNEWS - GIFS - VIDEOS - TODAYILEARNED - NEWS - MOVIES - CREEPY - GAMING - AWW - SHOWERTHOUGHTS - IAMA - MILDLYINTERESTING



From the SW Mods | If you see abuse, trolling, or guideline violations, click here to message us!

- 156 Update on PM trolls, self-appointed enforcers, and SW Moderation Practices. If you got a PM that was abusive, read, thx! (self.SuicideWatch)
submitted 11 months ago * by SQLwitch - - stickied post
39 comments share
- 45 Something New - an automod message to helpers (don't panic!) (self.SuicideWatch)
submitted 3 months ago * by skyqween - [M] - stickied post
57 comments share
- 1 No one will want to read the note, so I'm leaving it here (self.SuicideWatch)
submitted 2 hours ago * by spacedoughnut
5 comments share
- 2 I need some help here (self.SuicideWatch)
submitted an hour ago * by imightneedhelphere
2 comments share
- 3 Tired of the pain (self.SuicideWatch)

ADRIAN CHEN BUSINESS 10.23.14 6:30 AM

THE LABORERS WHO KEEP DICK PICS AND BEHEADINGS OUT OF YOUR FACEBOOK FEED

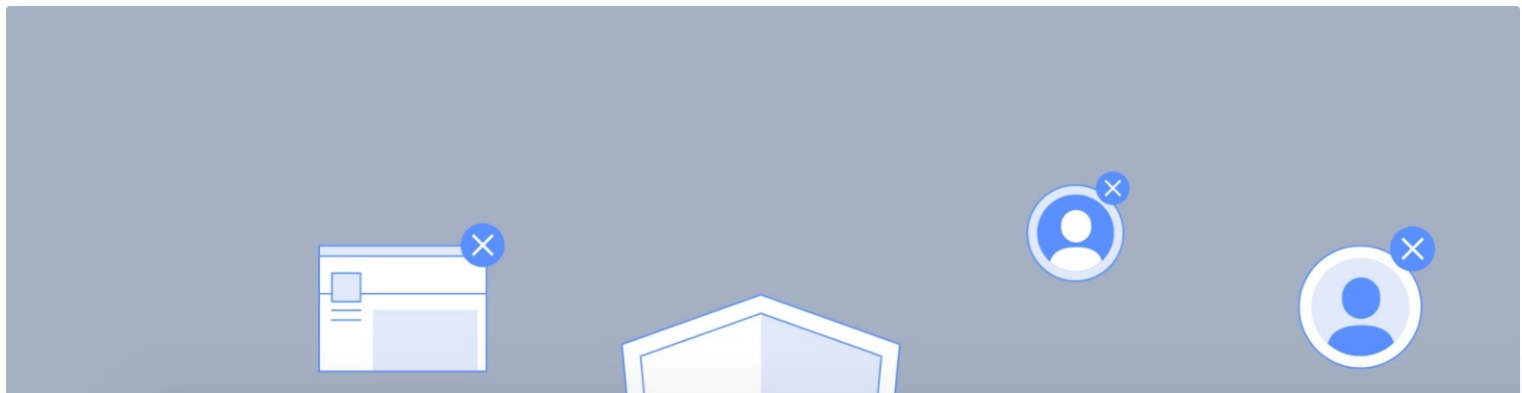
WIRED

 [Back to Newsroom](#)

Facebook


Banning More Dangerous Organizations from Facebook in Myanmar

February 5, 2019



Deviance

- Behaviors that violate the norms of a group
 - Akers, 1977; Suler and Phillips 1998.
- Sociological concept
 - Classically comes from Durkeim's *Anomie* book
- Online content moderation and the connection to deviance



You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech

MEDIA

JANUARY 17, 2021

Deplatforming Trump Is Already Having a Huge Impact

A new report finds election misinformation online has fallen 73 percent since the president's ban from Twitter.



MADISON PAULY

Reporter


[Bio](#) | [Follow](#)





The Fallacy of Deplatforming

Stringent moderation such as deplatforming works. But does it always?



#thyghgapp: Instagram
content moderation and lexical
variation in pro-eating disorder
communities

But deviant behavior subverts attempts to intervene

anorexic, anorexie, anoressia, anorexi, anorexia, anorexique, anorexica, anorectic, anorexia, anorectic

eatingdisorders, eatingdissorder, eatingdisoder, eatingdis, eatingdisorter, eatingdisoreder, eatingdisorde,
eatingdisorderrr, eatingdisordered, eating_disorder

thighgaps, thygap, thighgapp, thigh_gap, thightgap, thyhgap, thighgappp, thegap, thigap, thighgapss

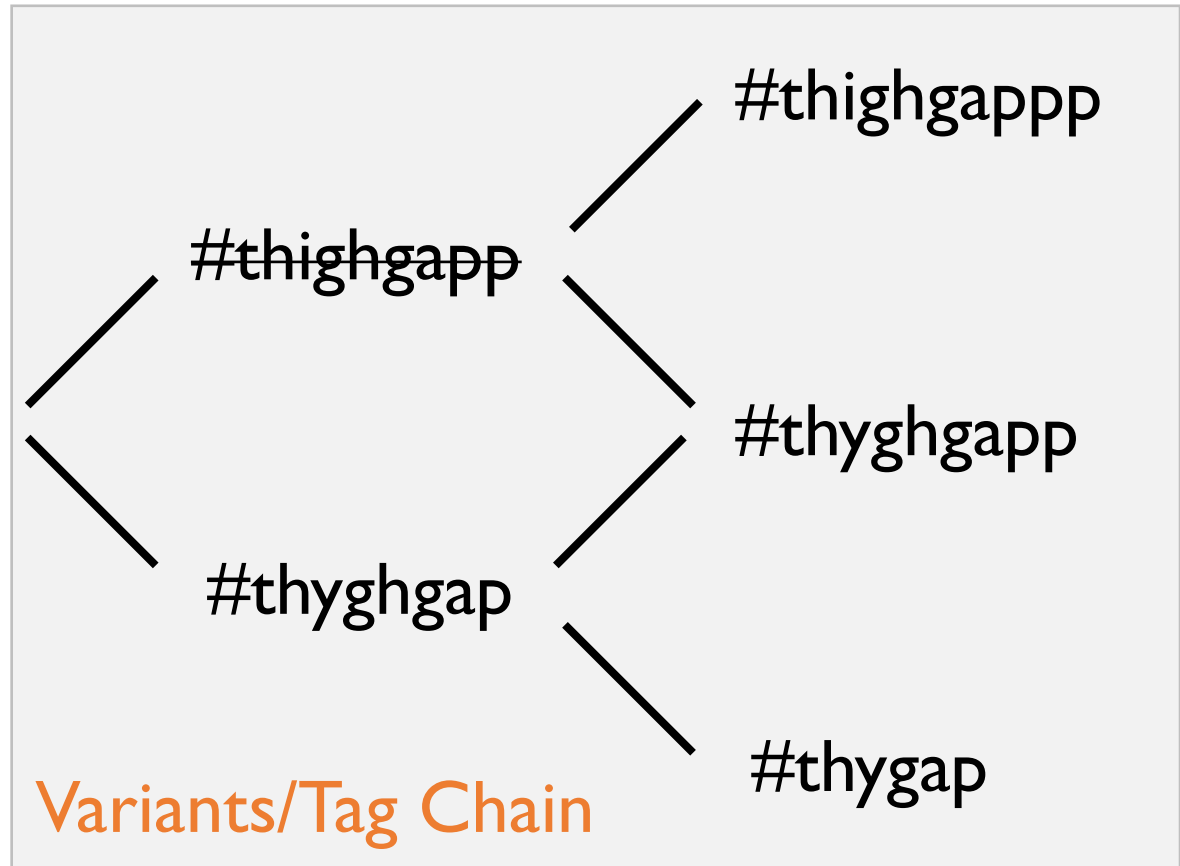
thinspoooo, thynspo, thynspoo, thynspooo, thinspoo, thinspooo, thynspoooo, thinnsपो, thinspoooooo

Increasing and more complex lexical variations have emerged since Instagram enforced moderation of pro eating disorder content in 2012

Automatically detect lexically variant tags that emerged out of moderated tags using edit distance computation + regular expression matching + semantic annotations via crowdsourcing (Amazon's mechanical turk)

“thighgap” > “th*g*p*”

#thighgap
Root



But deviant behavior subverts attempts to intervene

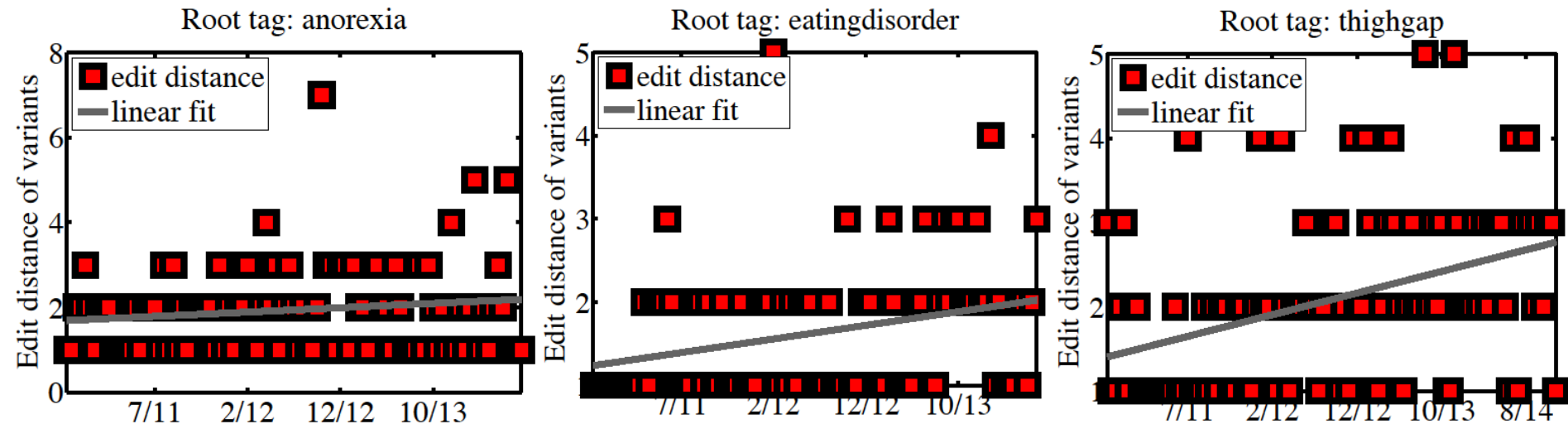
anorexic, anorexie, anoressia, anorexi, anorexia, anorexique, anorexica, anorectic, anorexia, anoretic

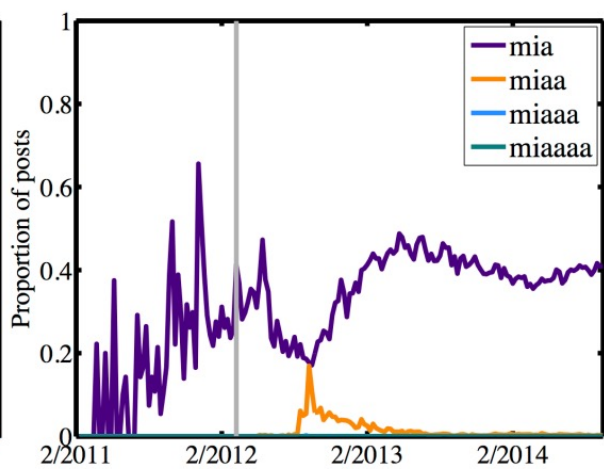
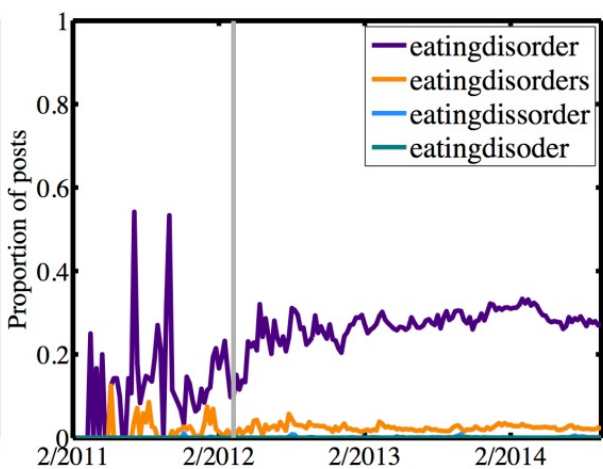
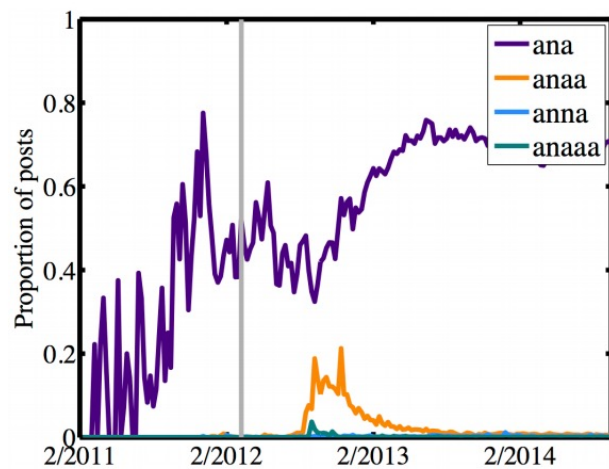
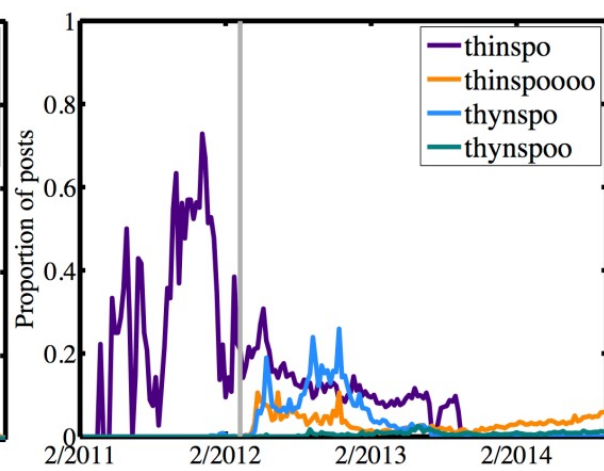
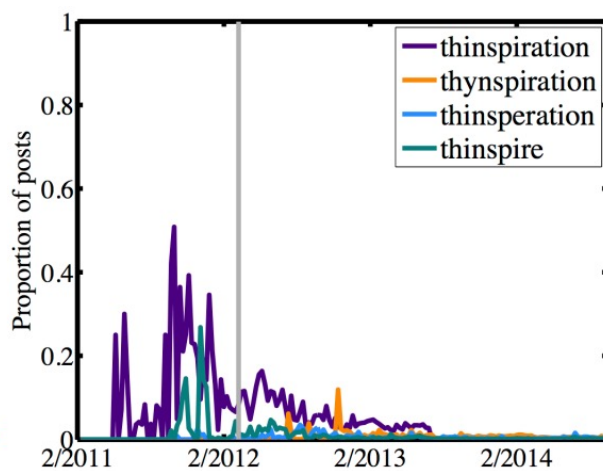
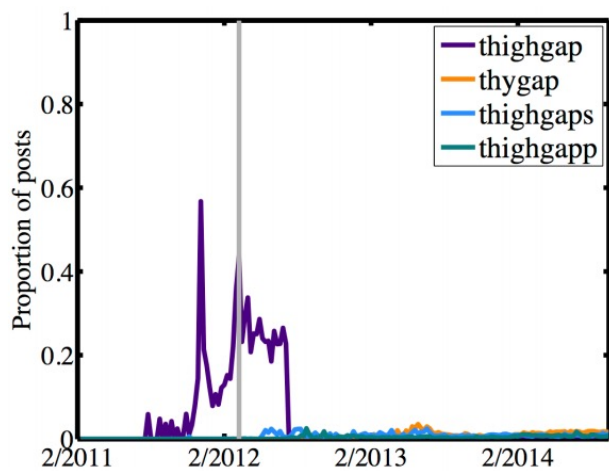
eatingdisorders, eatingdissorder, eatingdisoder, eatingdis, eatingdisorter, eatingdisoreder, eatingdisorde, eatingdisorderrr, eatingdisordered, eating_disorder

thighgaps, thygap, thighgapp, thigh_gap, thightgap, thyhgap, thighgapppp, thegap, thigap, thighgapss

thinspoooo, thynspo, thynspoo, thynspooo, thinspoo, thinspooo, thynspoooo, thinnsपो, thinspooooo

Increasing and more complex lexical variations have emerged since Instagram enforced moderation of pro eating disorder content in 2012





TECH

Why Eating Disorders Are So Hard For Instagram And Tumblr To Combat

Over the last four years, the social media platforms have done a lot to curb content that promotes self-injury. But they'll never fully succeed. Is it worth trying?

Posted on April 14, 2016, at 2:01 p.m.



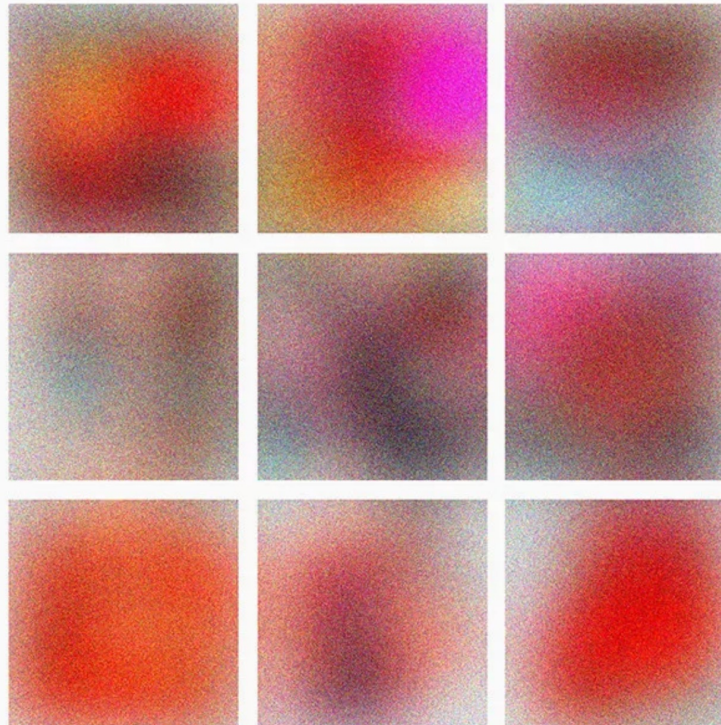
Stephanie M. Lee
BuzzFeed News Reporter



#anorexia

5,170,983 posts

TOP POSTS



Other examples where stringent content moderation and deplatforming didn't help



Reddit ran wild with Boston bombing conspiracy theories in 2013 and is now an epicenter for coronavirus misinformation. The site is doing almost nothing to change that.

Paige Leskin Mar 6, 2020, 9:14 AM



r/Coronavirus
u/wucaducadoo • 7d

+ JOIN

Challenges of reliance on AI moderation tools

Does Transparency in Moderation Really Matter?: User Behavior After Content Removal Explanations on Reddit

SHAGUN JHAVER, Georgia Institute of Technology

AMY BRUCKMAN, Georgia Institute of Technology

ERIC GILBERT, University of Michigan

When posts are removed on a social media platform, users may or may not receive an explanation. What kinds of explanations are provided? Do those explanations matter? Using a sample of 32 million Reddit posts, we characterize the removal explanations that are provided to Redditors, and link them to measures of subsequent user behaviors—including future post submissions and future post removals. Adopting a topic modeling approach, we show that removal explanations often provide information that educate users about the social norms of the community, thereby (theoretically) preparing them to become a productive member. We build regression models that show evidence of removal explanations playing a role in future user activity. Most importantly, we show that offering explanations for content moderation reduces the odds of future post removals. Additionally, explanations provided by human moderators did not have a significant advantage over explanations provided by bots for reducing future post removals. We propose design solutions that can promote the efficient use of explanation mechanisms, reflecting on how automated moderation tools can contribute to this space. Overall, our findings suggest that removal explanations may be under-utilized in moderation practices, and it is potentially worthwhile for community managers to invest time and resources into providing them.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: content moderation; content regulation; platform governance; post

Platform Governance outside of the US

Decentralized platform governance