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?
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”
• “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)

Authorities that invoke diverse strategies of suppression of various groups and individuals for a broad array of reasons and motivations. Adding to this complexity is the rapidity with which the Internet has grown and changed technologically; often government censors have difficulty keeping up-to-date with changing technologies (e.g., text messaging) or slang terms used to communicate hidden meanings.

The degree and type of Internet censorship obviously varies widely and reflects how democratic and open to criticism different political systems are. Reporters Without Borders, an NGO headquartered in Paris and one of the world’s preeminent judges of censorship, ranks governments across the planet in terms of the severity of their Internet censorship (Fig. 2; see also Quirk 2006). Their index of Internet censorship is generated from surveys of 50 questions sent to legal experts, reporters, and scholars in each country. Thus, countries in northern Europe, the US and Canada, Australia and New Zealand, and Japan exhibit minimal or no censorship (scores less than 10). Conversely, a rogue’s list of the world’s worst offenders, including China, Vietnam, Burma/Myanmar, Iran, and Turkmenistan, exhibit the planet’s most severe and extensive restrictions (scores greater than 80). In North Korea, Internet access is illegal, although the government uses it to send messages to the outside world (Hachigian 2002). In between these extremes lies a vast array of states with modest to moderate forms of Internet censorship that reflect their diverse systems of governance, the presence or absence of civil liberties, and the ability of various groups to resist limitations on their ability or right to use the Internet in whatever manner they so prefer.

Using the categories of Fig. 2, Table 1 summarizes the distribution of the world’s population and Internet users according to the level of severity of censorship. Thus, only 13% of the world’s people, but a third of Internet users, live in countries with minimal censorship; conversely, roughly one-quarter of the world’s people and Internet users live under governments that engage in very heavy censorship (the vast bulk of whom are located in China).

Internet penetration rates—the proportion of the population with regular access to cyberspace at home, school, or work—also shape the contours of censorship geography (Fig. 3). Rates vary from as low as 0.2% (Myanmar) to 100% (Falkland Islands).

Fig. 2 Reporters Without Borders Internet Censorship Ranking 2009. Source: data drawn from http://www.rsf.org/en-classement1003-2009.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
• 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?
Week 11: Regulating Online Speech
March 28, 2021
Any women have experience with having a baby while in grad school? (self.GradSchool)

Hitler doesn't get a postdoc (youtube.com)

I think it's time to quit (self.GradSchool)

Advice: Masters or go straight for PhD (self.GradSchool)

For Profit or Not University (self.GradSchool)

First time writing an state of purpose!! pls some help (self.GradSchool)

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)

Update on PM trolls, self-appointed enforcers, and SW Moderation Practices. If you got a PM that was abusive
read, thx! (self.SuicideWatch)

Something New - an automod message to helpers (don't panic!) (self.SuicideWatch)

No one will want to read the note, so I'm leaving it here (self.SuicideWatch)

I need some help here (self.SuicideWatch)

Tired of the pain (self.SuicideWatch)
THE LABORERS WHO KEEP DICK PICS AND BEHEADINGS OUT OF YOUR FACEBOOK FEED
Banning More Dangerous Organizations from Facebook in Myanmar

February 5, 2019
Deviance

- Behaviors that violate the norms of a group
- 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
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, anoretic |
| eatingdisorders, eatingdisorder, eatingdisoder, eatingdis, eatingdisorter, eatingdisoreder, eatingdisorde, eatingdisorderrr, eatingdisordered, eating_disorder |
| thighgaps, thygap, thighgapp, thigh_gap, thightgap, thyghgap, thighgapppp, thegap, thigap, thighgapss |
| thinspoooo, thynspo, thynspoo, thynspooo, thinspoo, thinspooo, thynspoooo, thinspooooo |

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*”
Variants/Tag Chain

#thighgap

Root
But deviant behavior subverts attempts to intervene

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

eatingdisorders, eatingdisorder, eatingdisoder, eatingdis, eatingdisor, eatingdisorderer, eatingdisorde, eatingdisorderrr, eatingdisordered, eating_disorder

thigthaps, thygap, thigapp, thigh_gap, thightgap, thyghgap, thigapppp, thegap, thigap, thigapsss

thinspoooo, thy upheld, thyynspoo, thyynspo, thynspoo, thinspoo, thynspoo, thynspooooo, thinspooooo

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

5,770,983 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.
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 removal explanation; user behavior.
Platform Governance outside of the US
Decentralized platform governance