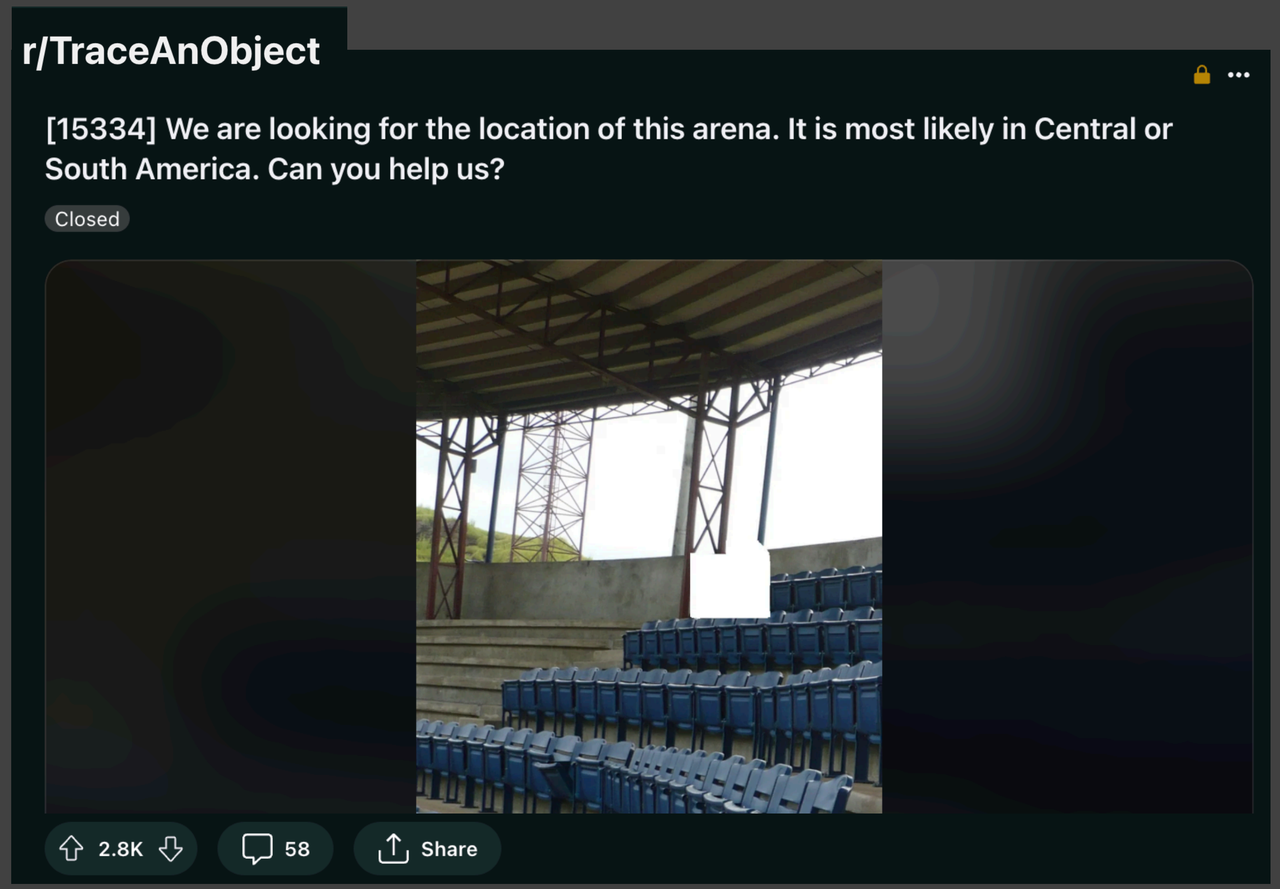
“Peer Production” example submitted by Isabelle L., Vyoma R.



In open source investigation (OSINT) communities, professionals and amateurs work together to investigate, verify, and geolocate pieces of evidence. For instance, people help Europol identify objects on the subreddit r/TraceAnObject.

Attendance

0.5% extra credit for examples relevant to recent or upcoming lectures. Submit on Ed under the “Lectures” category

Anti-Social Computing

CS 278 | Stanford University | Michael Bernstein

content warning for the last part of lecture: online rape, bullying, doxing, revenge porn, intimate partner violence

Announcements

Assignment 3 Part 2 (Remixes) due by 11:59pm tonight

- Part 3, Votes, due Friday — attention check questions are included
- Reflection (Part 4) due after the exam

Project milestone due 11:59pm next Tuesday

No reading this week

Exam question bank out next Tuesday

Exam one week later (no class, evening exam)

Last time: peer production

Shifting from simple wisdom-of-the-crowd tasks requires much more than just a scaling up of ambition: it requires designing for interdependence.

Peer production — the term encompassing shared open work (e.g., Wikipedia, open source) is one powerful method for volunteer coordination. Workflows and algorithms offer another approach. Both have their issues.

Aiming higher means we will need to solve issues of convergence and coordinated adapatation.

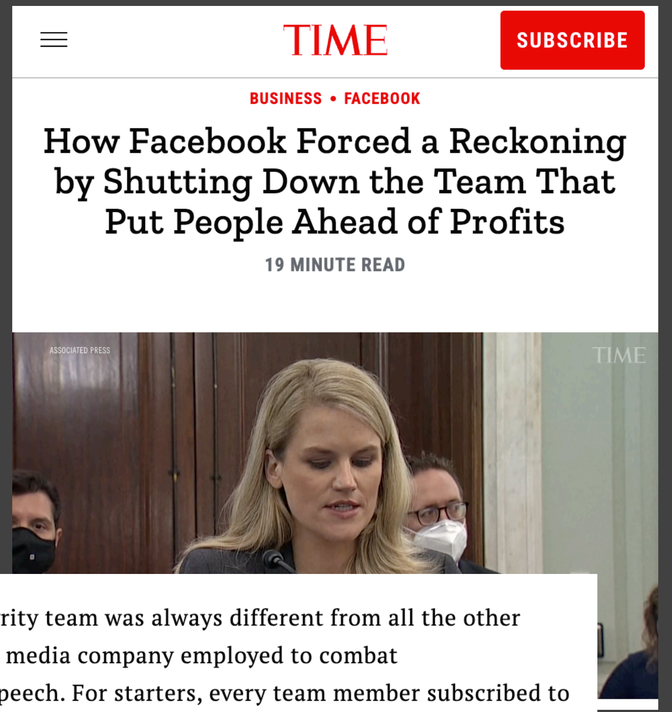
Thursday's guest visitor

Samidh Chakrabarti, creator of the Civic Integrity team at Meta

> *"Widely seen by employees as the spiritual leader of the push to make sure the platform had a positive influence on democracy and user safety"*

Pushing changes to newsfeed ranking, tracking civil society issues, advocating within the company

Samidh was the manager of whistleblower Frances Haugen



Facebook's civic-integrity team was always different from all the other teams that the social media company employed to combat misinformation and hate speech. For starters, every team member subscribed to an informal oath, vowing to "serve the people's interest first, not Facebook's."

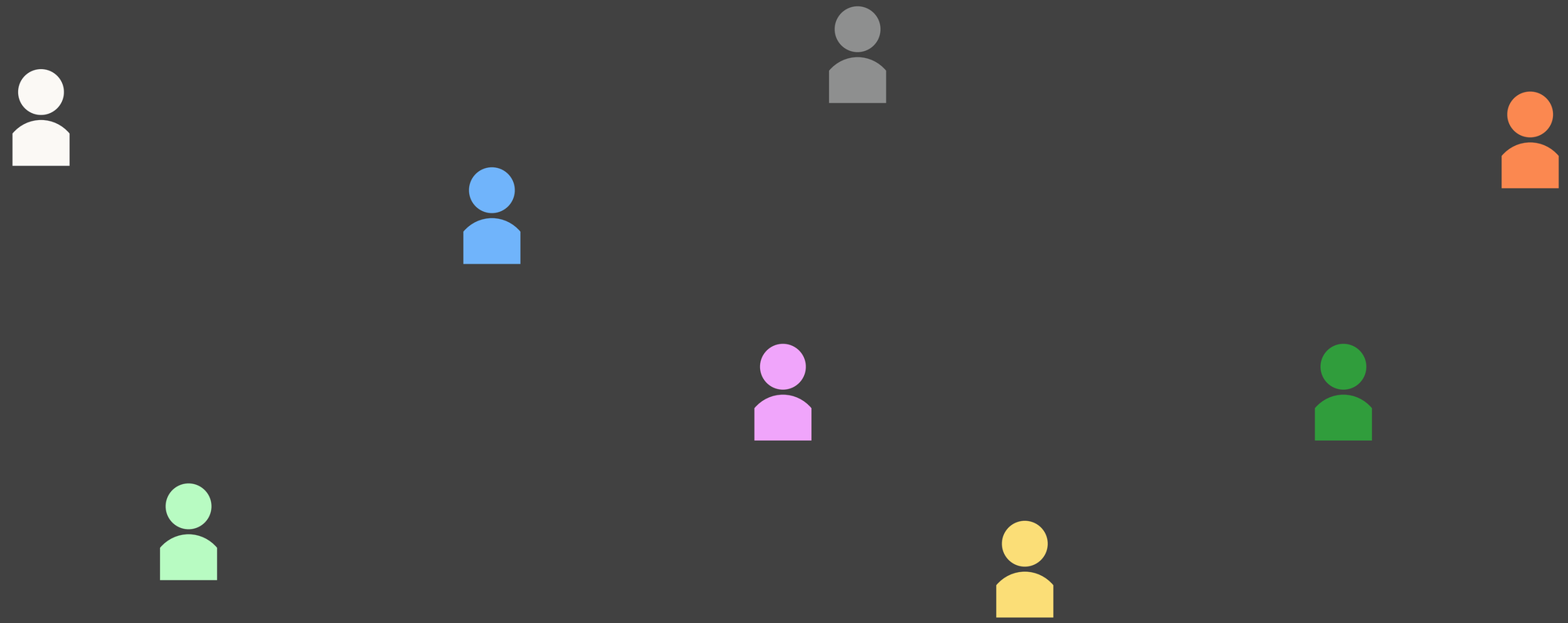
The "civic oath," according to five former employees, charged team members to understand Facebook's impact on the world, keep people safe and defuse angry polarization. Samidh Chakrabarti, the team's leader, regularly referred to this oath—which has not been previously reported—as a set of guiding principles behind the team's work, according to the sources.

We Work

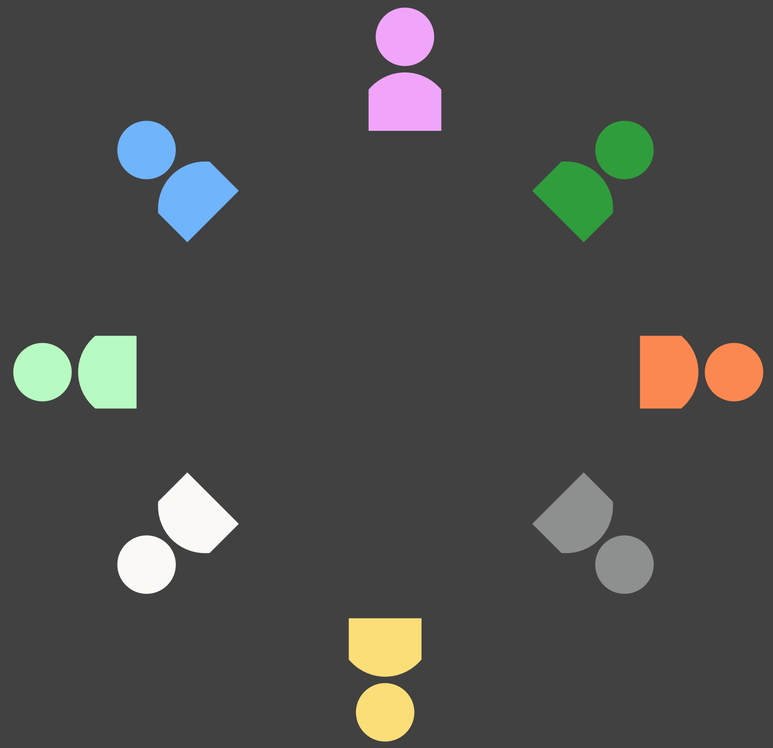
Unit 3

Don’t Feed The Trolls

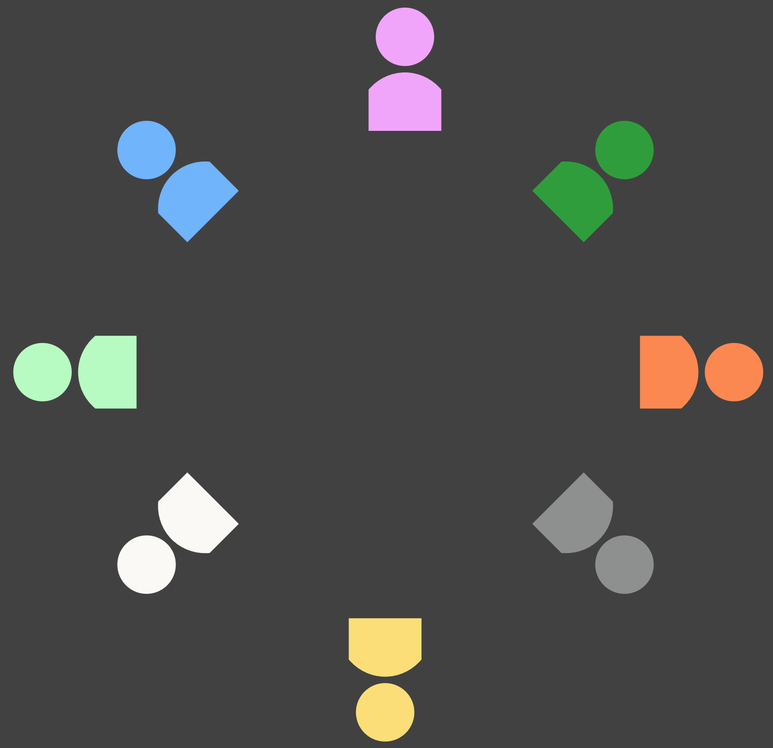
Unit 4



Once upon a time, the people were disconnected online.



So, we formed social computing systems to connect us to each other.



We met new friends, created online culture, and shared our ideas with the world. Life was happy.

But then, anti-social behavior arose. It grew and grew until it threatened to destroy the people and the platforms.

The people were trolled and flamed. Their communities fractured. Had the internet lost its way?

Today: anti-social computing

How can a community manage anti-social behavior?

(ノ °□°)ノ ⌒┻━┻

Premeditated anti-social behavior: trolling

Non-premeditated anti-social behavior: flaming

The darkest of the dark: beyond trolling [content warning]

Premeditated anti-social behavior

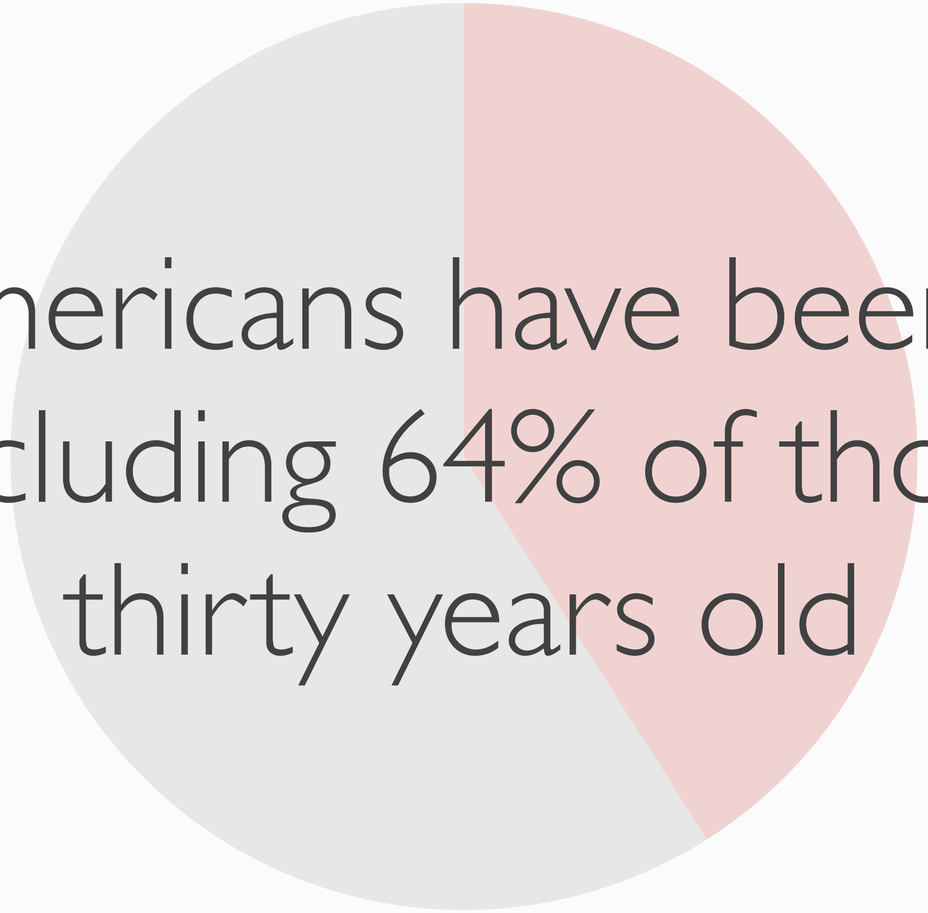
a.k.a., trolling 

How Trolls Are Ruining the Internet

When Will the Internet Be Safe for Women?

FAKE NEWS IS ABOUT TO GET EVEN SCARIER THAN YOU EVER DREAMED

Examples via Justin Cheng [Time 2016; The Atlantic 2016; Vanity Fair 2017]



41% of Americans have been harassed online, including 64% of those under thirty years old

[Vogels 2021; Thomas et al. 2022]

One in twenty comments left up on Reddit violate Reddit's own norms

[Park, Seering, and Bernstein 2022]

From YouTube to Reddit to Facebook to Usenet to Twitter to Telegram, 5-10% of comments are toxic

[Avalle et al. 2024]

WHY WE'RE SHUTTING OFF OUR COMMENTS

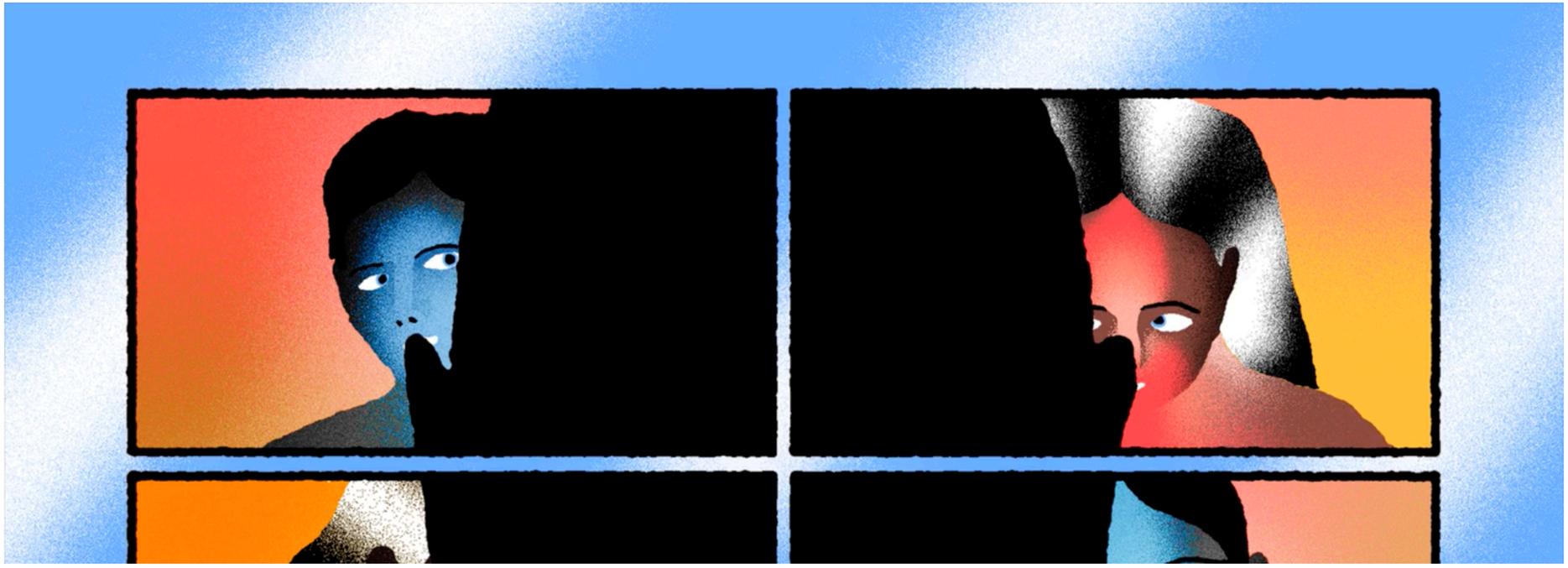
We're turning comments off for a bit

Sick of Internet comments? Us, too - here's what we're doing about it

Examples via Justin Cheng [Popular Science 2013; The Verge 2015; Chicago Sun-Times 2014]

'Zoombombing' Becomes a Dangerous Organized Effort

Zoom, the videoconferencing app, has become a target for harassment and abuse coordinated in private off-platform chats.



What is trolling?

Intentional disruption of an online community [Schwarz 2008]

Behavior that falls outside the acceptable bounds of the community [Binns 2012; Hardaker 2010]

People who habitually engage in trolling are known as trolls, as in the grumpy monsters who hide under bridges.

Is trolling worse online?

It's certainly more well-publicized. There are two reasons we might run into it more online than offline:

(1) Scale: a single troll can impact many communities, or a single highly visible community, in ways that their reach would otherwise be limited.

(2) People troll more online than they do offline.

Are these true? [1 min]

(1) Scale: a single troll can impact many communities, or a single highly visible community, in ways that their reach would otherwise be limited.

People are typically equally hostile offline as they are online [Bor and Peterson 2021]

But, status-driven individuals who are drawn to hot-button topics such as politics have large audiences online

So, we experience hostility more often online than offline

Why do trolls troll?

"Trolls are born that way"

Inveterate trolls do, on average, register strong personality dispositions such as high self-report scores in three of the four Dark Tetrad of personality traits: especially sadism, but also psychopathy and Machiavellianism [Buckels, Trapnell, and Paulhus 2014]

> And unfortunately, one in fourteen people internationally fall in the Dark Triad [Neumann et al. 2020]

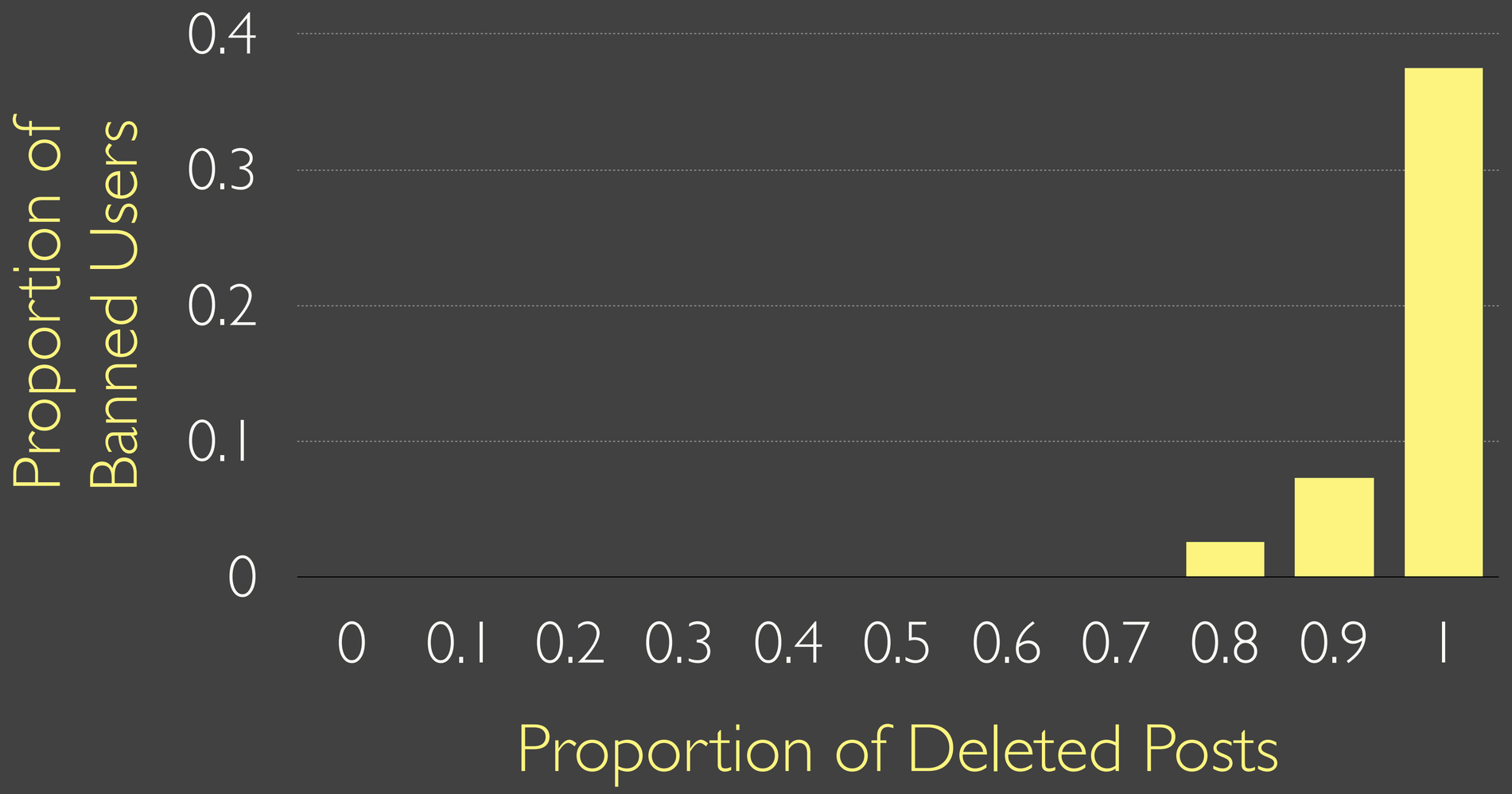
Reasons given range from boredom [Varjas et al. 2010], to doing it for fun [Shachaf and Hara 2010], to venting [Lee and Kim 2015].

Blocking

If a few people are responsible for the majority of the deleterious content, then blocking them should silence most of the negative behavior.

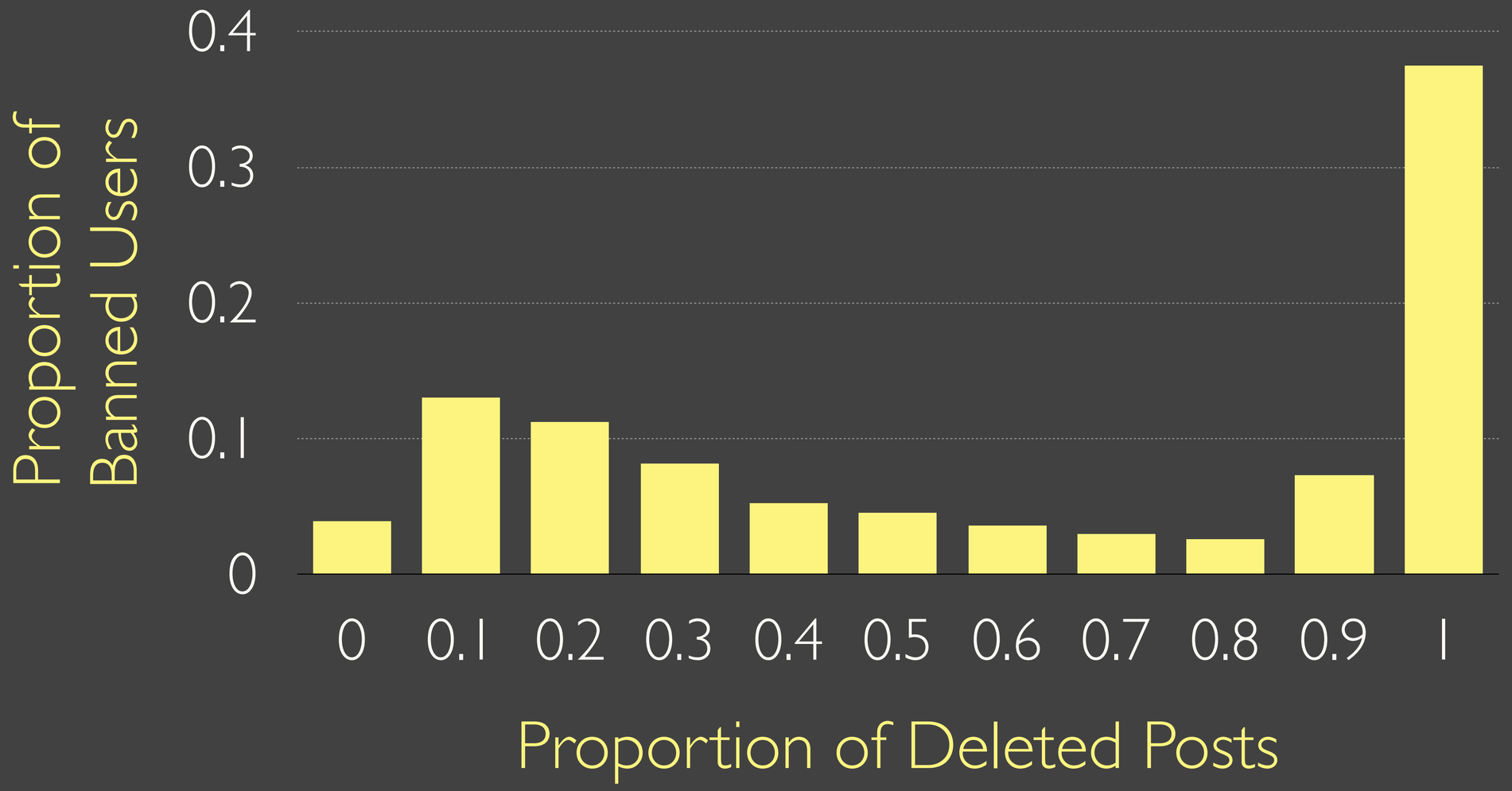
However, blocking just becomes whack-a-mole if it's easy for participants to create another account. So, this only works if identities are expensive to create.

How much do trolls troll?



CNN.com data: some trolls will troll almost exclusively

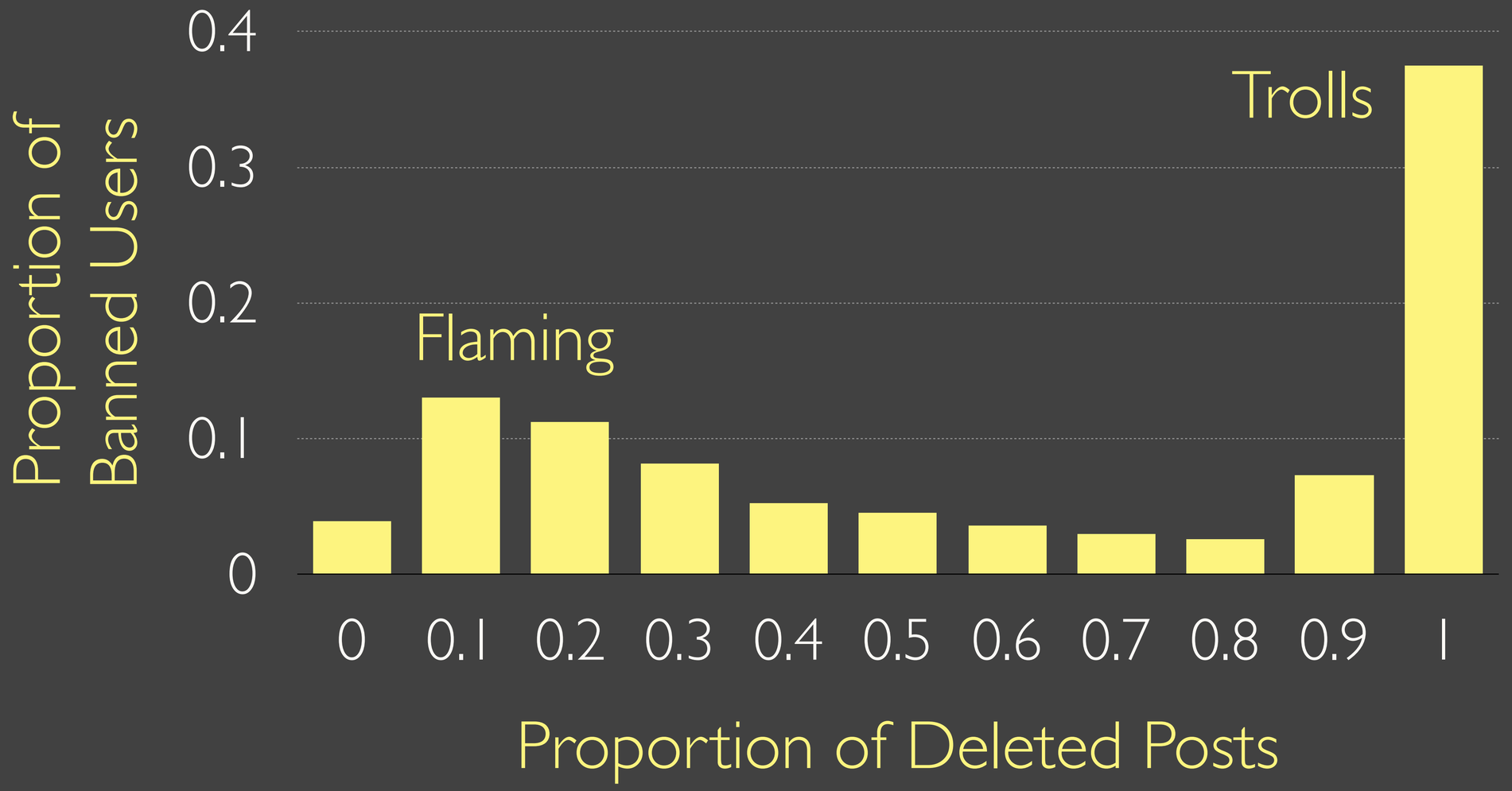
How much do trolls troll?



CNN.com data: some trolls will troll almost exclusively… and some only rarely.

Why? [1 min]

How much do trolls troll?



CNN.com data: some trolls will troll almost exclusively… and some only rarely.

Why? [1 min]

Unpremeditated anti-social behavior

a.k.a., flaming 🔥

What is flaming?

Flaming: uninhibited hostile behavior directed at another person or group [Kayany 1998, Kiesler 1986]

Common examples: swearing, calling names, ridiculing, insulting

While trolling usually refers to someone intentionally riling people up, flaming usually refers to someone who lost self-control.

Online disinhibition effect

[Suler 2004]

When we interact online, we say and do things that we would not do offline and in-person. We self-disclose more, and we act out more.

This is known as the online disinhibition effect: we have less inhibition when online.

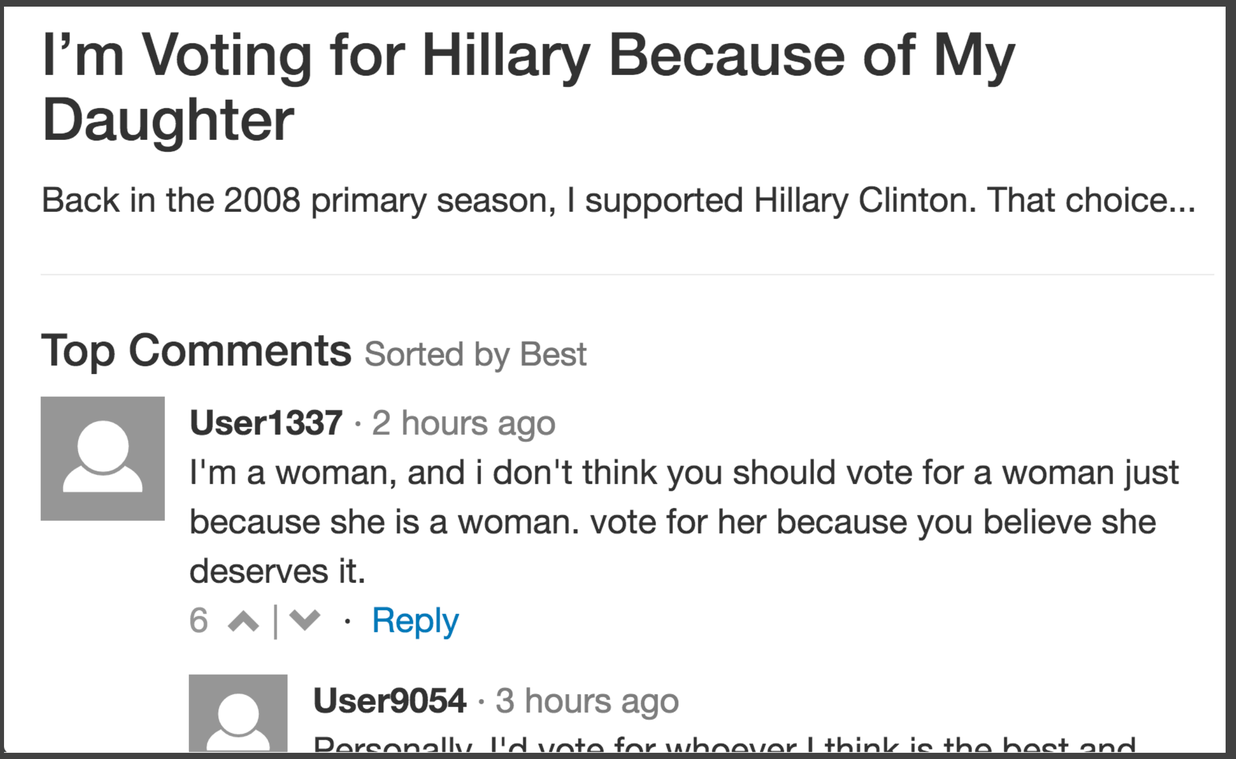
Online disinhibition would imply that we do troll more online than offline.

“Well, that escalated quickly.”

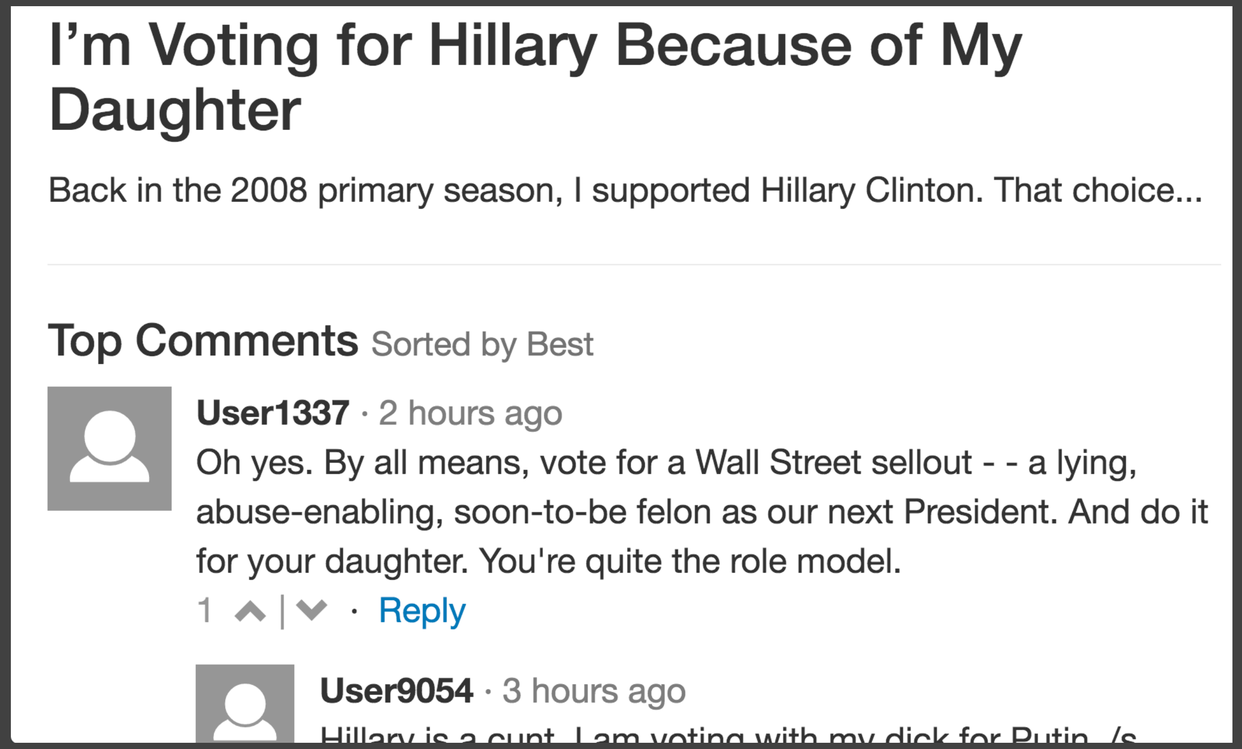
We are not good at predicting how others will read our comments

> We overperceive moral outrange online: readers perceive more moral outrage in social media content than the content author actually feels [Brady et al. 2023]

Recall: environment matters



(Real comments)



Positive comments

Result: 35% antisocial comments

[Cheng et al. 2017]

Negative comments

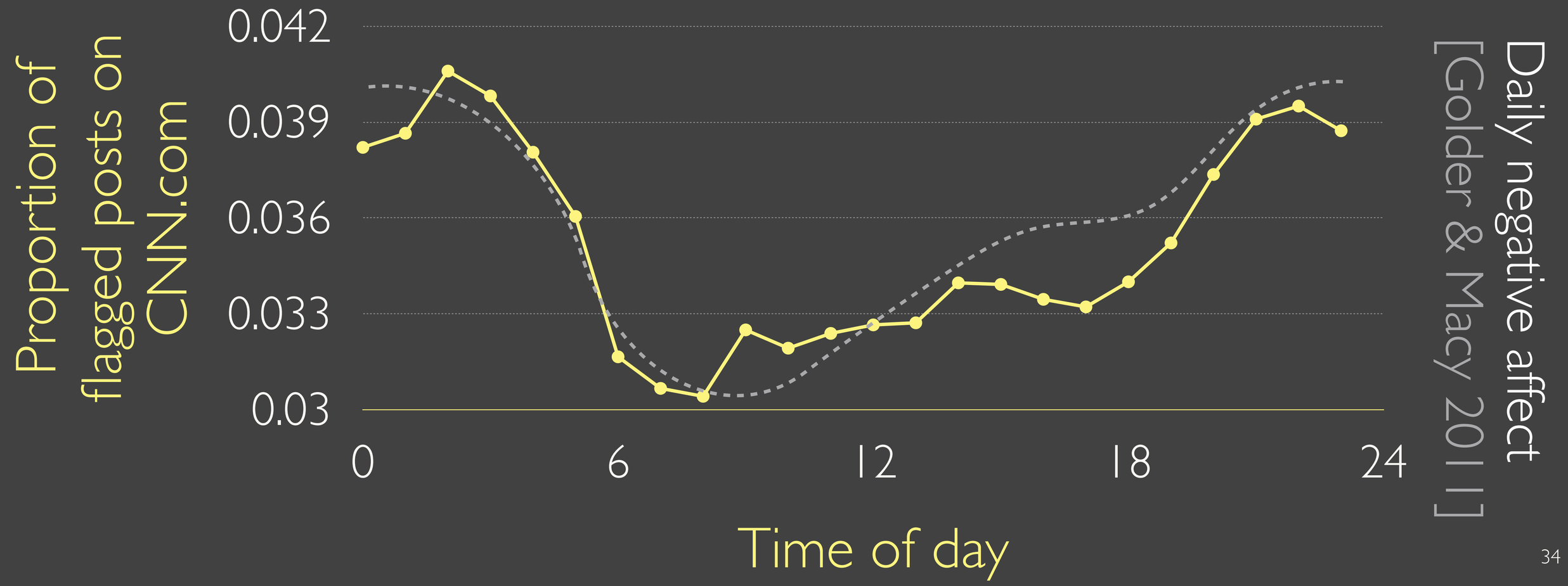
Result: 47% antisocial comments (Relative increase of one third compared to the 35% baseline)

Mood 😡

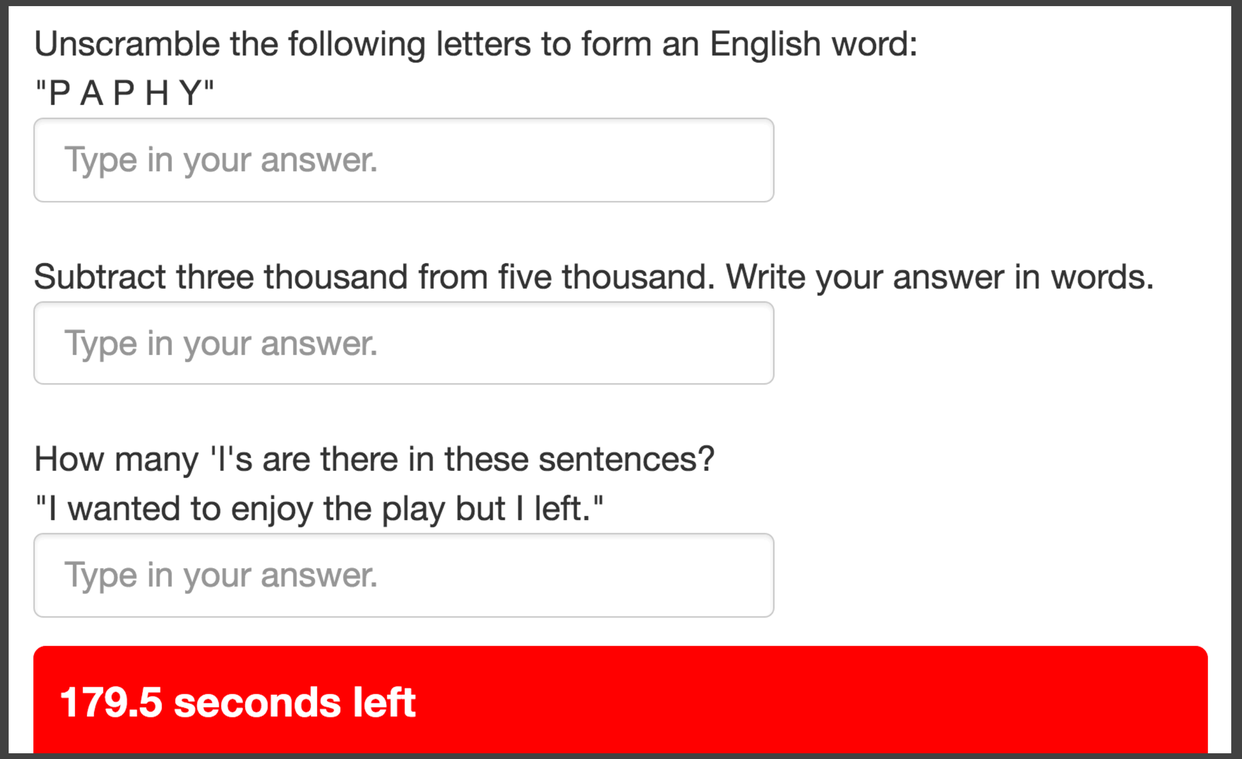
The environment is something the designer has control of. But people also bring their own affective state to a social computing system.

Being in a bad mood reduces self-regulation [Leith and Baumeister 1996] and results in less favorable impressions of others [Forgas and Bower 1987].

Antisocial behavior tracks human diurnal mood patterns

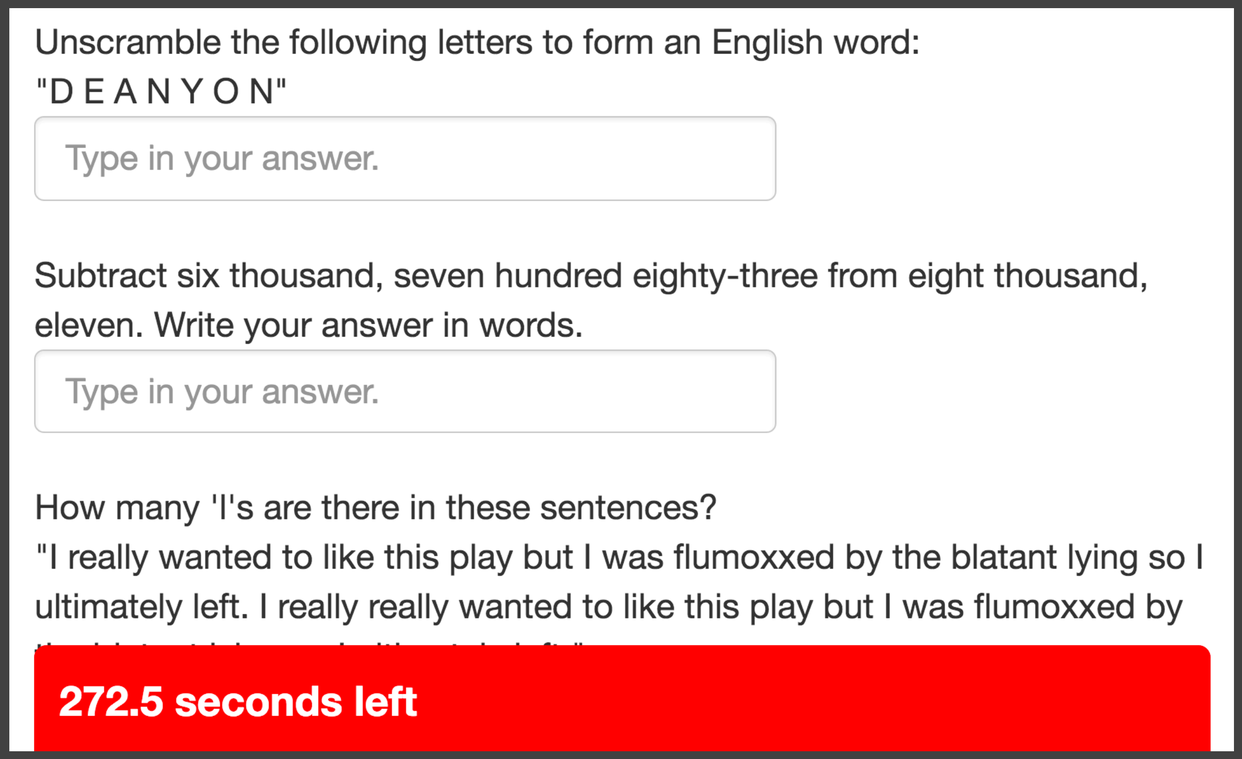


Mood influences behavior [Cheng et al. 2017]



Placed in a good mood by doing well on an easy test

Result: 35% troll comments



Placed in a negative mood by doing poorly on a difficult test

Result: 49% troll comments

(Same effect as seeing troll comments!)

Positive Mood

Negative Mood

Positive Norm

Unscramble the following letters to form an English word:
"P A P H Y"

Type in your answer.

User1337 · 2 hours ago
I'm a woman, and i don't think you should vote for a woman just because she is a woman. vote for her because you believe she deserves it.

35% antisocial comments

Unscramble the following letters to form an English word:
"D E A N Y O N"

Type in your answer.

User1337 · 2 hours ago
I'm a woman, and i don't think you should vote for a woman just because she is a woman. vote for her because you believe she deserves it.

49% antisocial comments

The effects compound: "Anybody can become a troll."

Negative Norm

Unscramble the following letters to form an English word:
"P A P H Y"

Type in your answer.

User1337 · 2 hours ago
Oh yes. By all means, vote for a Wall Street sellout - - a lying, abuse-enabling, soon-to-be felon as our next President. And do it for your daughter. You're quite the role model.

47% antisocial comments

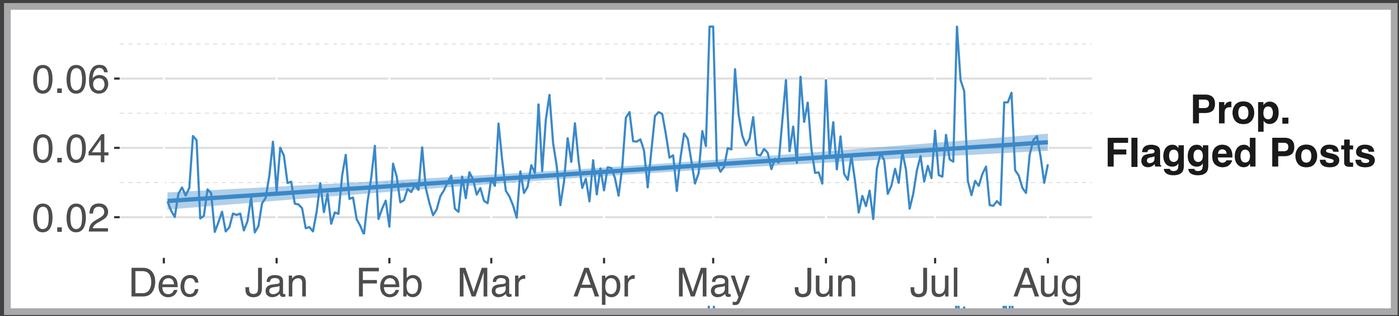
Unscramble the following letters to form an English word:
"D E A N Y O N"

Type in your answer.

User1337 · 2 hours ago
Oh yes. By all means, vote for a Wall Street sellout - - a lying, abuse-enabling, soon-to-be felon as our next President. And do it for your daughter. You're quite the role model.

68% antisocial comments

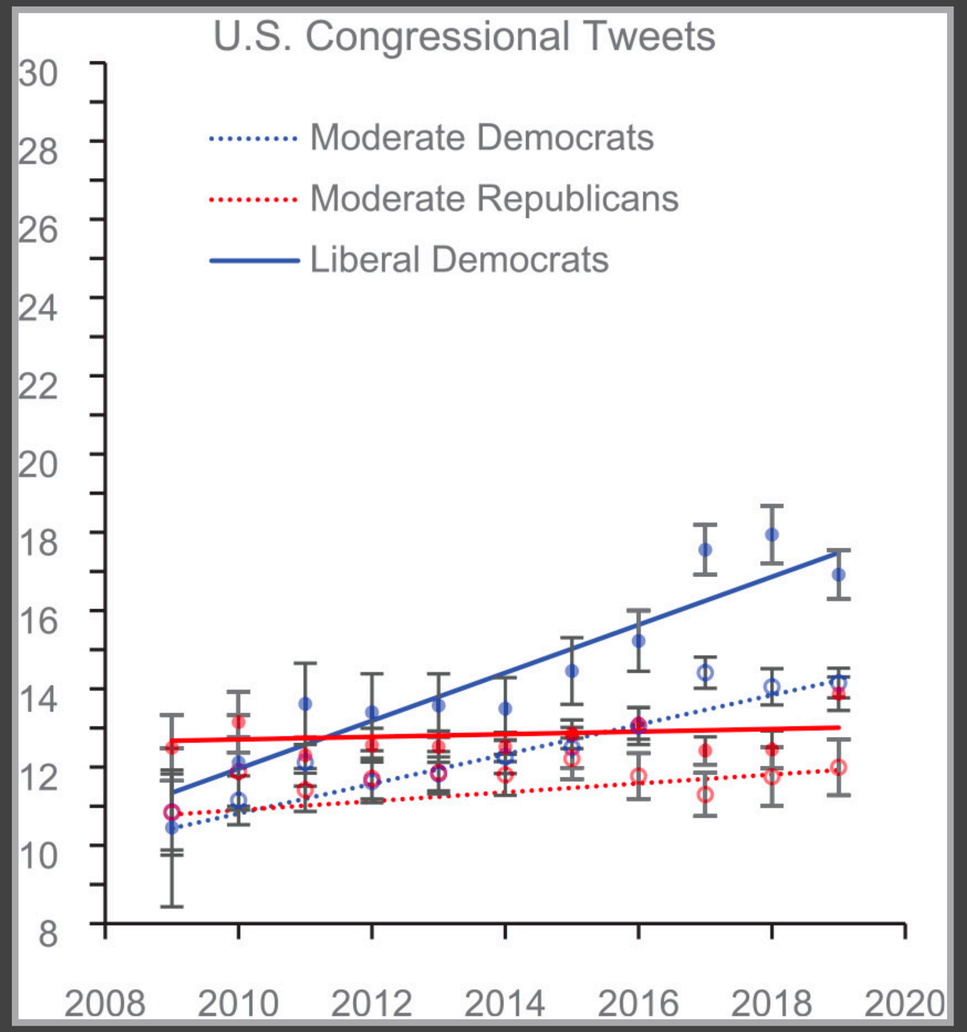
Does it get worse over time?



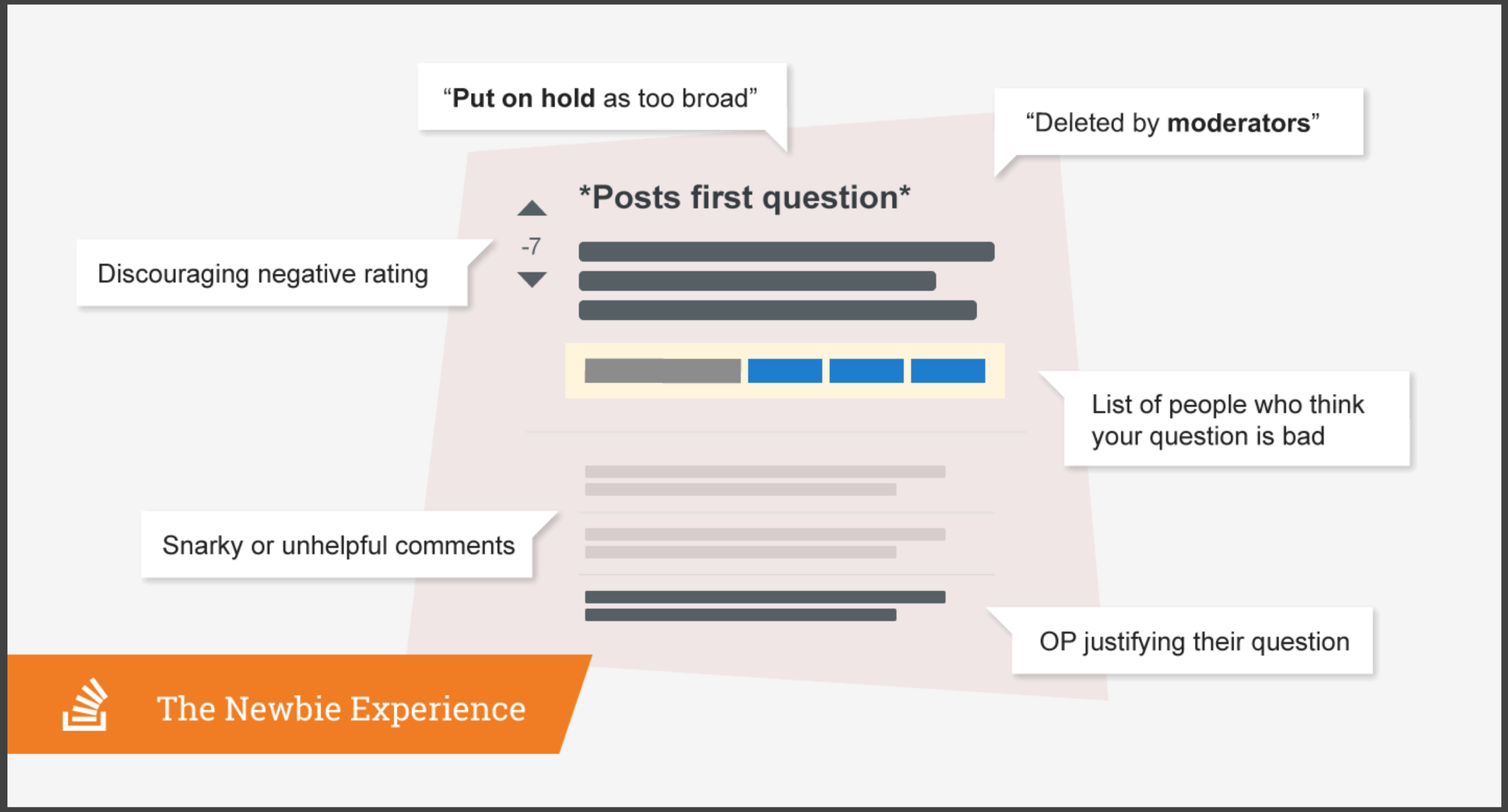
Flagged posts ~double on CNN.com over six months [Cheng et al 2017]

Toxicity of tweets by US Congress members up by 22% over a decade [Frimer et al 2022]

Cause: more engagement for anti-social behavior reinforces the behavior



Many independent signals can combine to create a hostile or negative environment



[https://stackoverflow.blog/2019/07/18/building-community-inclusivity-stack-overflow/]

Design responses

Dealing with disinhibition

Reasons include:

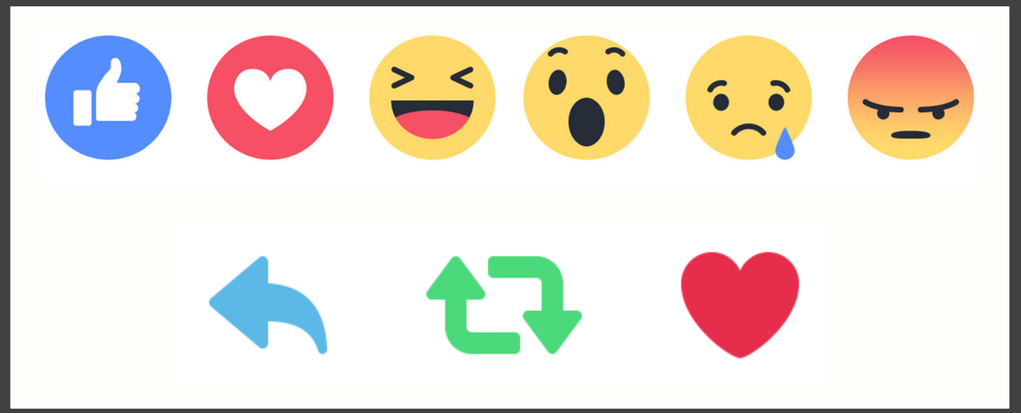
- Anonymity: dissociation from my real identity, so fewer consequences
- Few social cues: no facial expressions, reactions, etc.; a socially opaque system
- Asynchronicity: conversations never cool off

So, design interventions might be:

- Re-individuate by associating actions with an identity that I care about
- Re-introduce social cues: e.g., when I reply to a mean comment, they often soften up
- Take it offline: don't try to manage fights online if possible

What does the design encourage?

Systems that reward short-term engagement are likely to produce snark and flame, since these activities are the most likely ones to raise an affective response.



[Image from Niloufar Salehi]

Early detection of off-the-rails conversations

[Chang and Danescu-Niculescu-Mizil 2019]

Two threads on the Wikipedia discussion for the Dyadlov Pass incident:

A1: Why there's no mention of it here? Namely, an altercation with a foreign intelligence group? True, by the standards of sources some require it wouln't even come close, not to mention having some really weak points, but it doesn't mean that it doesn't exist.

A2: So what you're saying is we should put a bad source in the article because it exists?

B1: Is the St. Petersberg Times considered a reliable source by wikipedia? It seems that the bulk of this article is coming from that one article, which speculates about missile launches and UFOs. I'm going to go through and try and find corroborating sources and maybe do a rewrite of the article. I don't think this article should rely on one so-so source.

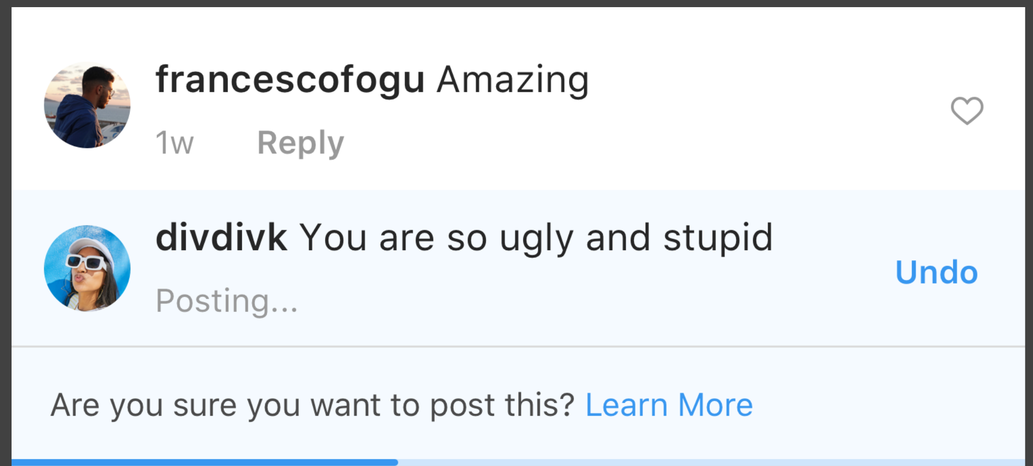
B2: I would assume that it's as reliable as any other mainstream news source.

Reducing flaming

Assuming the environment and norms aren't changeable, then one possible mechanism is to manage mood. Moods pass, so consider a cool-off period before posting a flame post.

Examples: Twitter, Tinder, Instagram

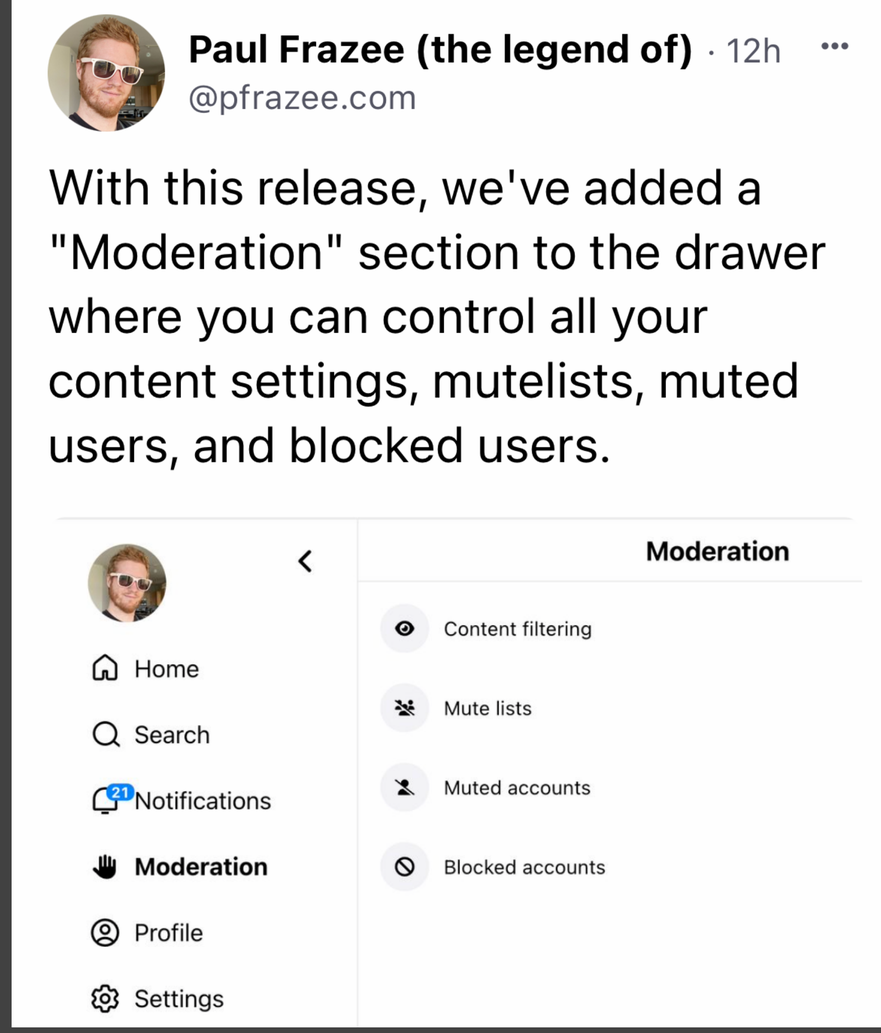
Or: "we will ask you again in 20 minutes if you really want to post this"

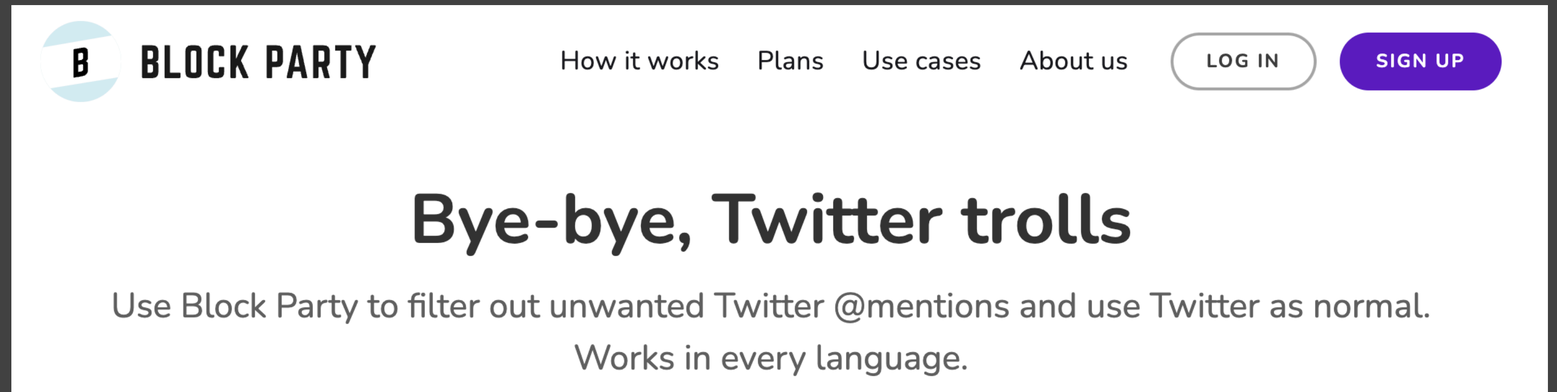


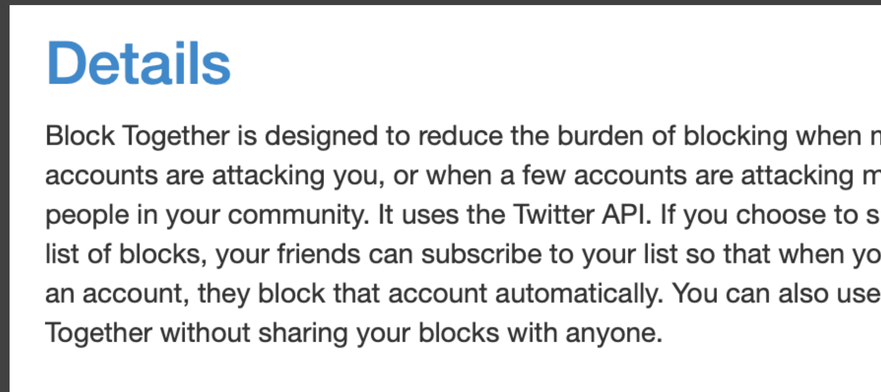
Blocklists

[Geiger 2016]

Community-maintained block lists: harassers can get added to the blocklist, then are automatically blocked from any user's account that subscribes to the blocklist

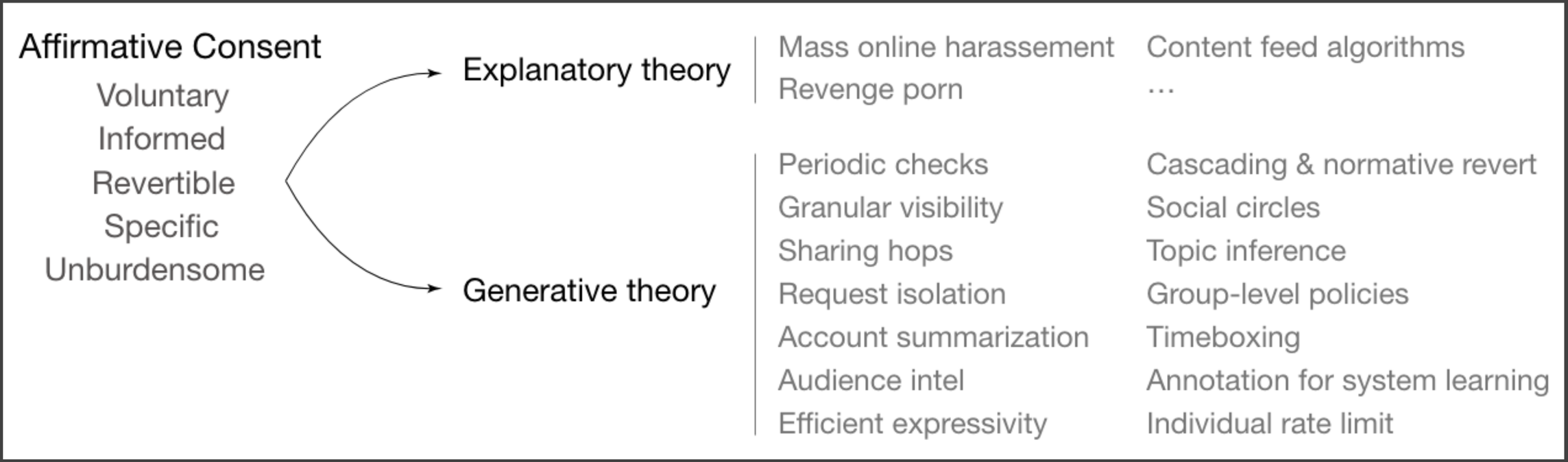






Affirmative consent

[Im et al. 2021]



Rather than defaults that focus on repair after the fact, what if designs aimed for affirmative consent?

Dealing with norm breakers

Imagine you were a Stanford admin/RF/RA/etc. and it was brought to your attention that someone was engaging in substantial amounts of antisocial behavior in your community's online spaces.

What would you do? [2min]

> COMING UP NEXT: "darkest of the dark" (content warning: online rape, bullying, doxing, revenge porn, partner abuse). This is a good moment to step out if you prefer. This material will not be on the exam.

Face-saving

[Kiesler et al. 2012]

People self-regulate if they can do so without having to admit that they deliberately violated norms. MIT's warning email to students:

> Someone using your account did [whatever the offense is]. Account holders are responsible for the use of their accounts. If you were unaware that your account was being used in this way, it may have been compromised. User Accounts can help you change your password and re-secure your account.

Many would change their password and the practice would stop, even if MIT knew from eyewitnesses that they had done it.

Calling them out instead prompted people to assert the behavior as within their rights and continue doing it to challenge authority.

The darkest of the dark

Content warning: online rape, bullying, doxing, revenge porn, partner abuse

Anti-social behavior gets (even more) personal

We typically think of anti-social behavior as hurling insults at each other, but this behavior escalates.

In this final section, we'll survey some troubling behaviors and what we know about them.

For more detail on this content, take CS 152: Trust & Safety Engineering

A Rape in Cyberspace

[Dibbell 1993]

In LambdaMOO, a text-based online spaces (a la a textual MMO), a character named Mr. Bungle developed a piece of software that allowed him to command other characters to perform actions.

He then forced two other avatars to perform sexual acts on him and on each other, and to violate their own bodies.

Mr. Bungle was eventually banned from the server after a community meeting, but the damage had been done.

HUMANS AND TECHNOLOGY

The metaverse has a groping problem already

A woman was sexually harassed on Meta's VR social media platform. She's not the first—and won't be the last.

By Tanya Basu

December 16, 2021

Cyberbullying

Adolescent bullies on social media — for example in public or private messages — are particularly problematic because they follow you home, not just at school. [Li 2007]

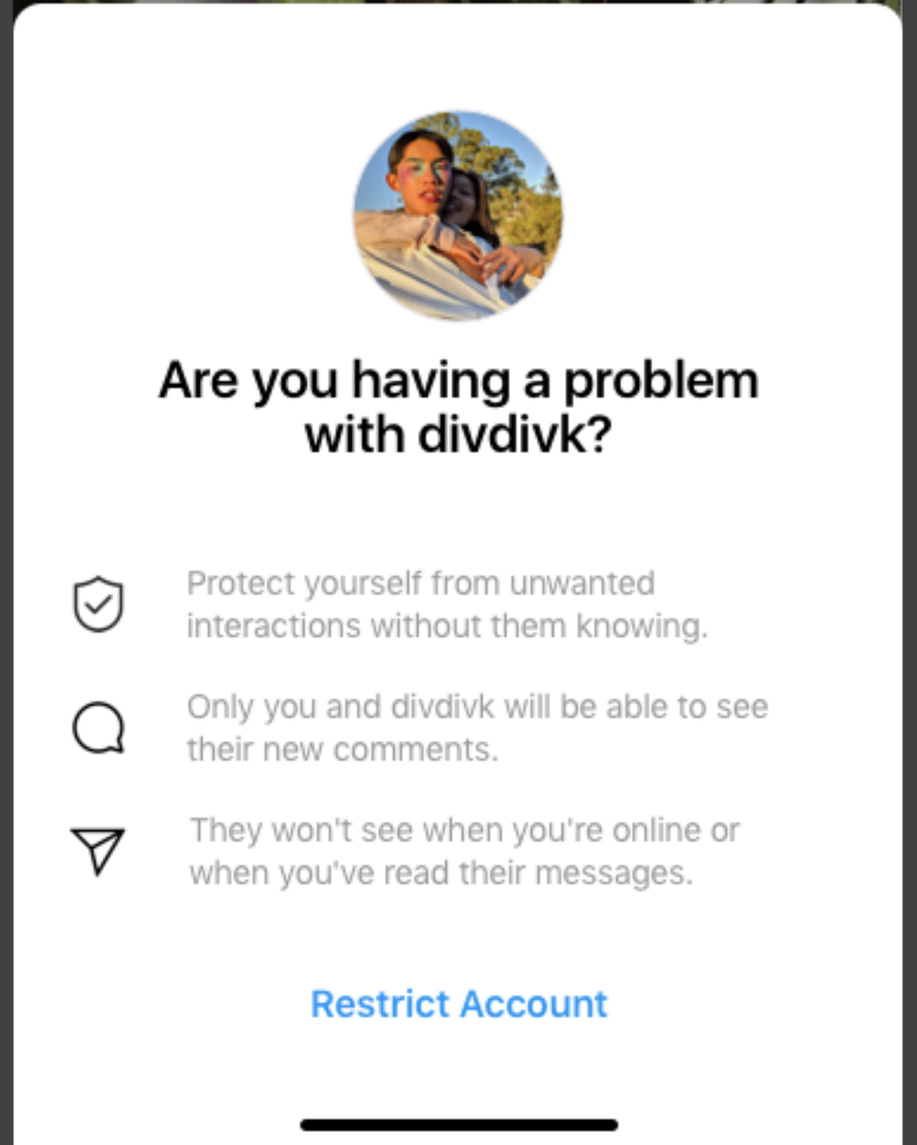
Women and LGBTQ are more likely to be victims, and perpetrators are more likely to be male. [Aboujaoude et al. 2015]

Being cyberbullied increases suicidal ideation and the probability of attempting suicide. [Hinduja and Patchin 2010]

Cyberbullying

Designs typically focus on blocking, but this doesn't erase the issues

Example from Instagram



Revenge porn

Revenge porn: nonconsensual distribution of sexual photos

Typically, the "revenge" in revenge porn means that the pair used to be partners and the photos were initially shared consensually, but a breakup or other event prompted one party to release the other person's sexual photos to hurt the other party.

Doxxing

Doxxing refers to releasing the personal information (e.g., address, name, photo) of an individual online against their will. The term originated from the idea of sharing "docs", or documents, of someone.

One shared, the information cannot easily be taken back.

Intimate partner violence

[Freed et al. 2018]

Abusers in intimate partner violence utilize technology to intimidate, monitor, impersonate, and harass. Often, the victims are married to their abusers and share social networks and physical space.

- Owning the device or paying for the family plan, threatening to remove it
- Installing or authorizing software to track the victim (e.g., Apple's Find My)
- Forcing victims to disclose social media passwords, monitoring messages

Generally, security approaches are not designed to combat attackers who know the victim intimately.

What do we do?

There is no permanent solution here. New behaviors arise over time. People manipulate the system to achieve their goals.

Step one: ensure that there are serious consequences for this kind of behavior — possibly legal ones.

Step two: find confidential and trusted means for users to report abuse. Develop a trusted adjudication system.

Summary

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 transform into 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.

References

Aboujaoude, Elias, et al. "Cyberbullying: Review of an old problem gone viral." Journal of adolescent health 57.1 (2015): 10-18.

Avalle, Michele, et al. "Persistent interaction patterns across social media platforms and over time." Nature 628.8008 (2024): 582-589.

Binns, Amy. "DON'T FEED THE TROLLS! Managing troublemakers in magazines' online communities." Journalism practice 6.4 (2012): 547-562.

Bor, Alexander, and Michael Bang Petersen. "The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis." American political science review 116.1 (2022): 1-18.

Brady, William J., et al. "Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility." Nature human behaviour (2023): 1-11.

Buckels, Erin E., Paul D. Trapnell, and Delroy L. Paulhus. "Trolls just want to have fun." Personality and individual Differences 67 (2014): 97-102.

Chang, Jonathan P., and Cristian Danescu-Niculescu-Mizil. "Trouble on the Horizon: Forecasting the Derailment of Online Conversations as they Develop." Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019.

Chang, Jonathan P., Justin Cheng, and Cristian Danescu-Niculescu-Mizil. "Don't let me be misunderstood: Comparing intentions and perceptions in online discussions." Proceedings of the Web Conference 2020. 2020.

References

Cheng, Justin, et al. "Anyone can become a troll: Causes of trolling behavior in online discussions." Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. 2017.

Dibbell, Julian. "A rape in cyberspace or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society." Ann. Surv. Am. L. (1994): 471.

Forgas, Joseph P., and Gordon H. Bower. "Mood effects on person-perception judgments." Journal of personality and social psychology 53.1 (1987): 53.

Freed, Diana, et al. "'A Stalker's Paradise' How Intimate Partner Abusers Exploit Technology." Proceedings of the 2018 CHI conference on human factors in computing systems. 2018.

Frimer, Jeremy A., et al. "Incivility is rising among American politicians on Twitter." Social Psychological and Personality Science 14.2 (2023): 259-269.

Geiger, R. Stuart. "Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space." Information, Communication & Society 19.6 (2016): 787-803.

Golder, Scott A., and Michael W. Macy. "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures." Science 333.6051 (2011): 1878-1881.

Hardaker, Claire. "Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions." (2010): 215-242.

Hinduja, Sameer, and Justin W. Patchin. "Bullying, cyberbullying, and suicide." Archives of suicide research 14.3 (2010): 206-221.

References

Im, Jane, et al. "Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021.

Kayany, Joseph M. "Contexts of uninhibited online behavior: Flaming in social newsgroups on Usenet." Journal of the American Society for Information Science 49.12 (1998): 1135-1141.

Kiesler, Sara, et al. "Regulating behavior in online communities." Building successful online communities: Evidence-based social design 1 (2012): 4-2.

Kiesler, Sara. The hidden messages in computer networks. Harvard Business Review Case Services, 1986.

Lee, So-Hyun, and Hee-Woong Kim. "Why people post benevolent and malicious comments online." Communications of the ACM 58.11 (2015): 74-79.

Leith, Karen Pezza, and Roy F. Baumeister. "Why do bad moods increase self-defeating behavior? Emotion, risk tasking, and self-regulation." Journal of personality and social psychology 71.6 (1996): 1250.

Li, Qing. "New bottle but old wine: A research of cyberbullying in schools." Computers in human behavior 23.4 (2007): 1777-1791.

Neumann, Craig S., et al. "Light and dark trait subtypes of human personality–A multi-study person-centered approach." Personality and Individual Differences 164 (2020): 110121.

References

Park, Joon Sung, Joseph Seering, and Michael S. Bernstein. "Measuring the Prevalence of Anti-Social Behavior in Online Communities." Proceedings of the ACM on Human-Computer Interaction 6.CSCW2 (2022): 1-29.

Schwartz, Mattathias. "The trolls among us." The New York Times 3.08 (2008). https://www.nytimes.com/2008/08/03/magazine/03trolls-t.html

Shachaf, Pnina, and Noriko Hara. "Beyond vandalism: Wikipedia trolls." Journal of Information Science 36.3 (2010): 357-370.

Suler, John. "The online disinhibition effect." Cyberpsychology & behavior 7.3 (2004): 321-326.

Thomas, Kurt, et al. "'It's common and a part of being a content creator': Understanding How Creators Experience and Cope with Hate and Harassment Online." Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. 2022.

Varjas, Kris, et al. "High school students' perceptions of motivations for cyberbullying: An exploratory study." Western Journal of Emergency Medicine 11.3 (2010): 269.

Vogels, Emily A. "The state of online harassment." Pew Research Center 13 (2021): 625.

Zhang, Justine, et al. "Conversations Gone Awry: Detecting Early Signs of Conversational Failure." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.. Vol. 1. 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.