# Designing Restorative Approaches to Moderating Adversarial Online Interactions

Cliff Lampe - @clifflampe - cacl@umich.edu
University of Michigan School of Information

# Cliff Lampe

Old Slashdot mod and researcher.

Human-Computer Interaction, Social Computing, Computer-Mediated Communication.

Deep into organizing the SIGCHI conferences.

How do we increase the good and decrease the bad?

# IMPORTANT POSITIONS

This presentation will have content that includes misogyny, hate speech and other sensitive topics.

I’m on the moderate left in U.S. politics.

My brand of research is interventionist.

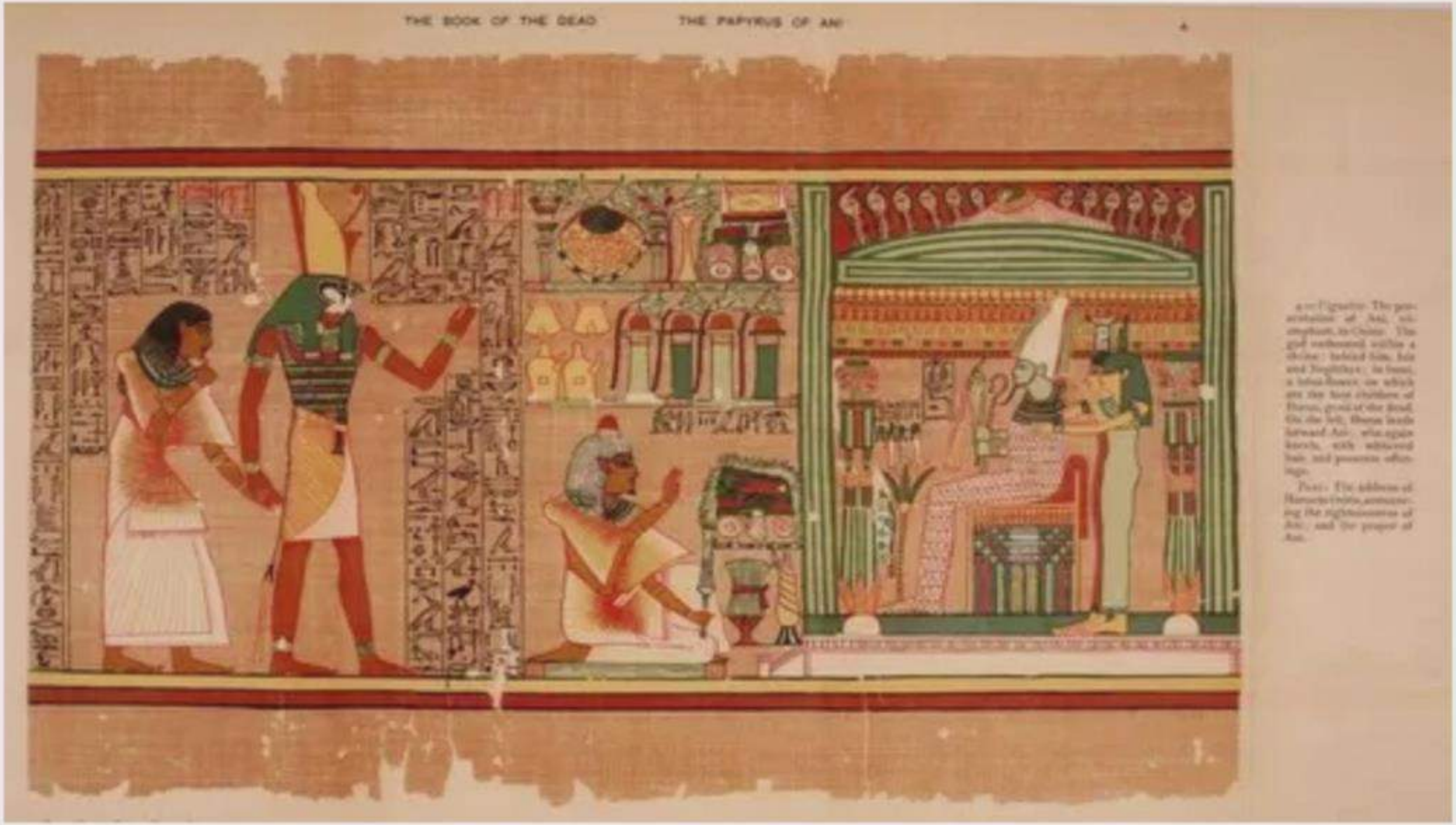
Sarita Yardi Schoenebeck

JJ Prescott

# NSF: Drawing from Theories of Justice to Respond to Online Harassment

People do bad things.

# Book of the Dead - Papyrus of Ani (1250 BCE)



THE BOOK OF THE DEAD    THE PAPYRUS OF ANI

4

4—Vignette: The presentation of Ani, triumphant, to Osiris. The god enthroned within a shrine; behind him, Isis and Nephthys; in front, a lotus-flower on which are the four children of Horus, gods of the dead. On the left, Horus leads forward Ani, who again kneels, with whitened hair and presents offerings.

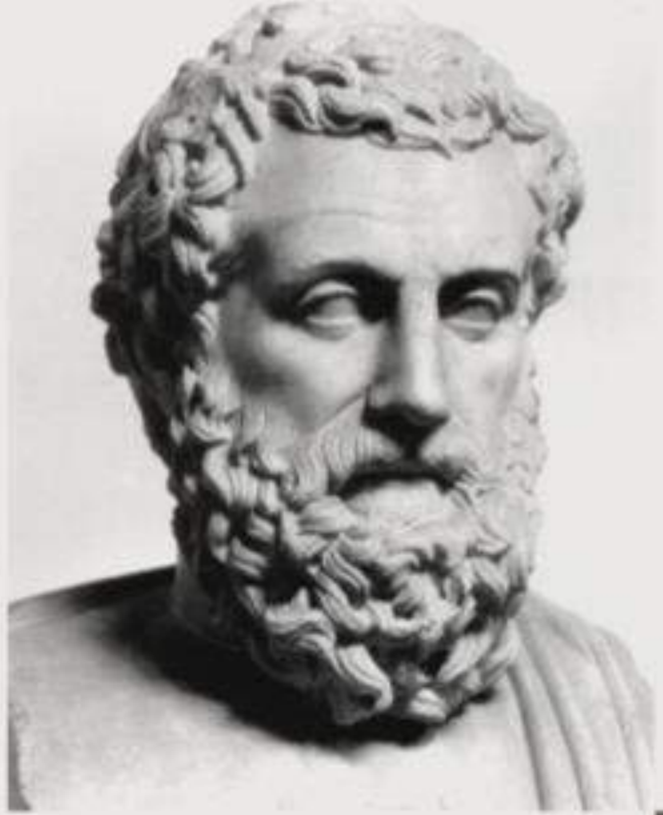
Text: The address of Horus to Osiris, announcing the righteousness of Ani; and the prayer of Ani.

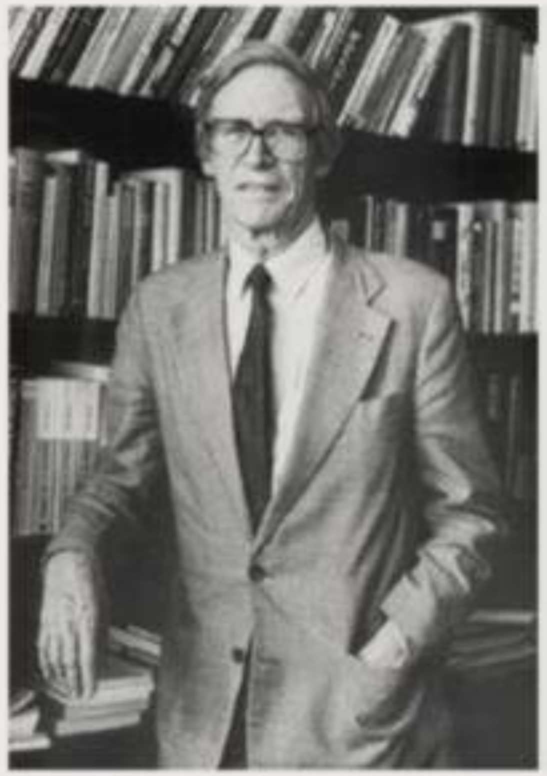
“Hail, Neb-abui, who comest forth from Sauti, I have not multiplied my words in speaking.”

# Defining “bad”

Norms - Flexible, elegant, imprecise, unfair.

Laws - Inflexible, clunky, mostly precise, fair





People do bad things online.

# “Affordances”

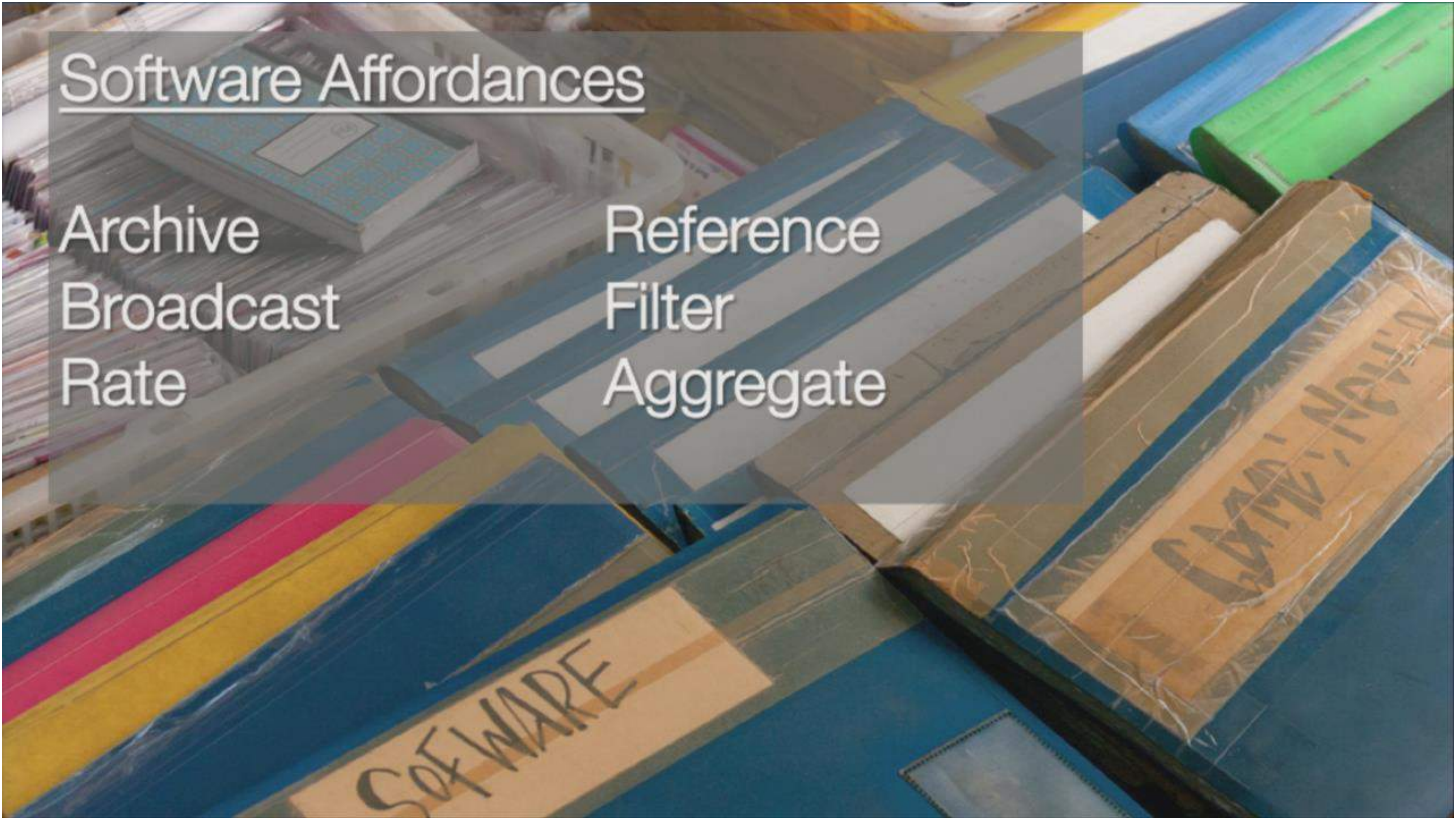
Software and hardware features that allow a user to accomplish an action

Software

Hardware

## Hardware affordances

Ubiquity
Bundle devices
Comparatively cheap



## Software Affordances

| | |
|---|---|
| Archive | Reference |
| Broadcast | Filter |
| Rate | Aggregate |

Different bundles of “affordances” create different genres of social media.

# People experience “adversarial interactions” online

Harm done to someone on social media.

“Bad” groups who use social media.

Broad categories of bad behavior

# Two examples of adversarial interactions.

# Online Radicalization: “Redpilling”

# Redpilling

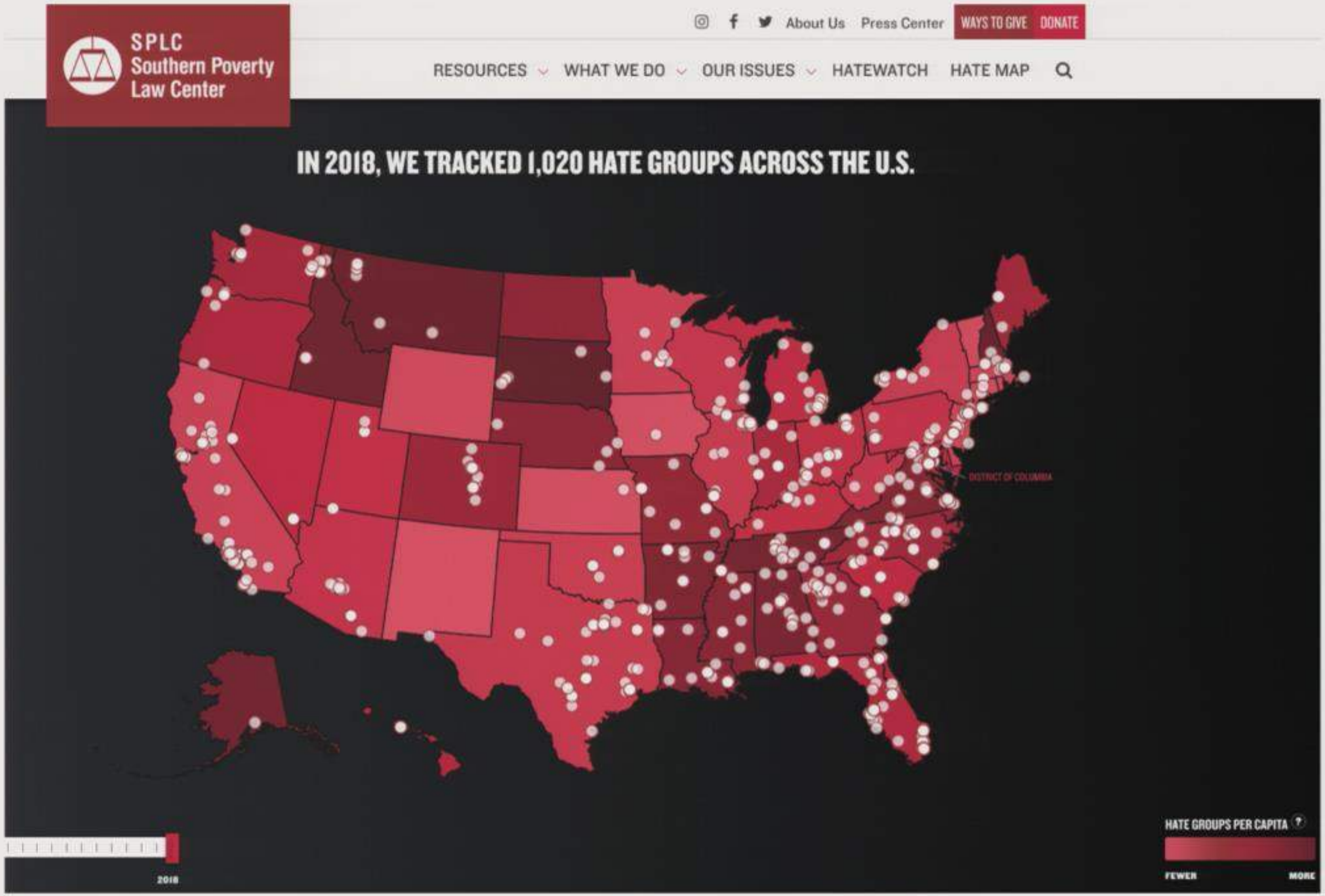
"You take the blue pill, the story ends. You wake up in your bed and believe whatever you want to believe.

You take the red pill, you stay in Wonderland, and I show you how deep the rabbit hole goes."

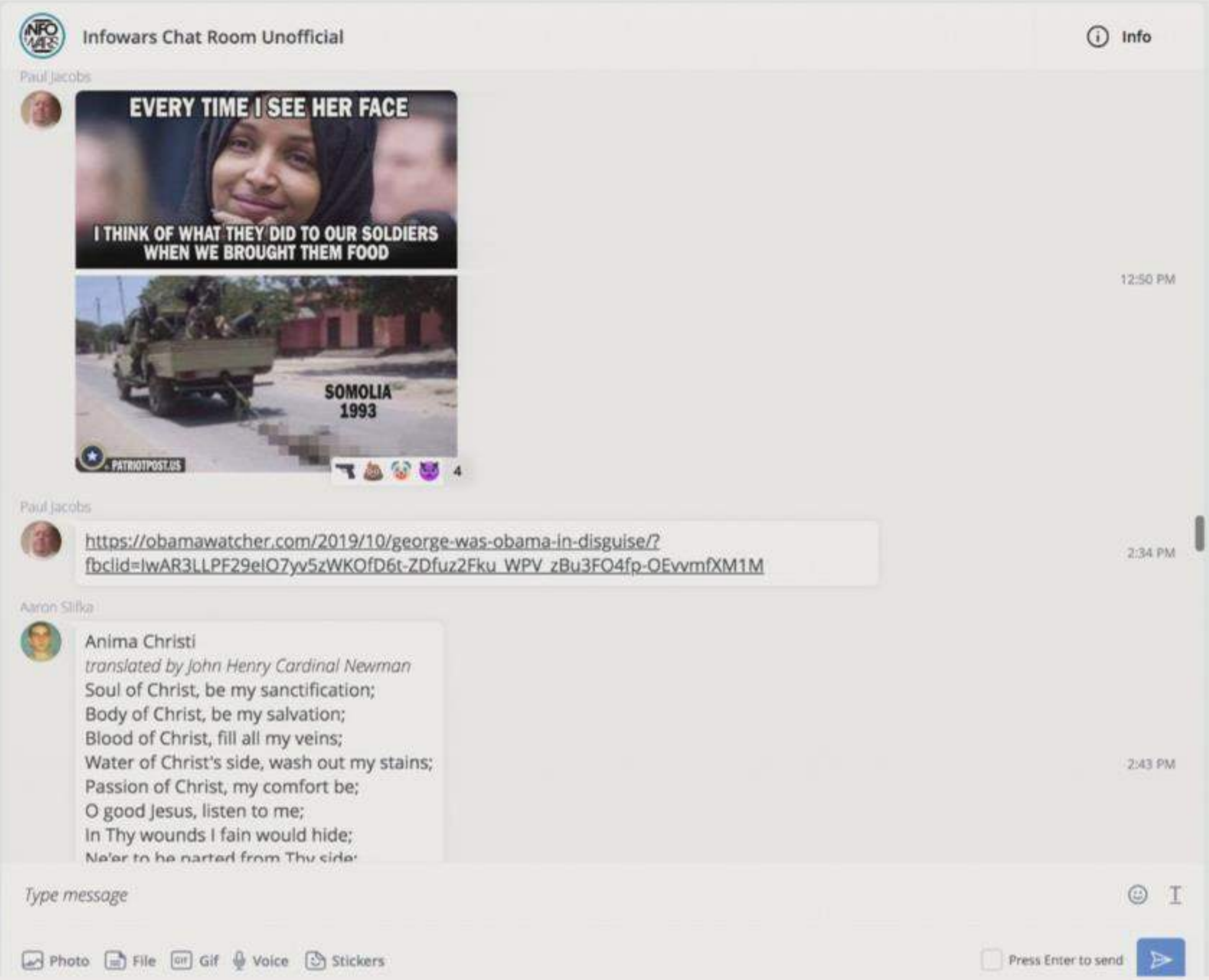
Redpilling:

Converting someone to a radical point of view.

Most often converting a liberal or moderate (Normie) to an extreme conservative POV.



IN 2018, WE TRACKED 1,020 HATE GROUPS ACROSS THE U.S.


Infowars Chat Room Unofficial
Info
EVERY TIME I SEE HER FACE
I THINK OF WHAT THEY DID TO OUR SOLDIERS WHEN WE BROUGHT THEM FOOD
SOMOLIA 1993
PATRIOTPOST.US
🔫 💩 🤡 😈 4
Type message
Photo
File
Gif
Voice
Stickers
Press Enter to send

Paul Jacobs

12:50 PM

Paul Jacobs

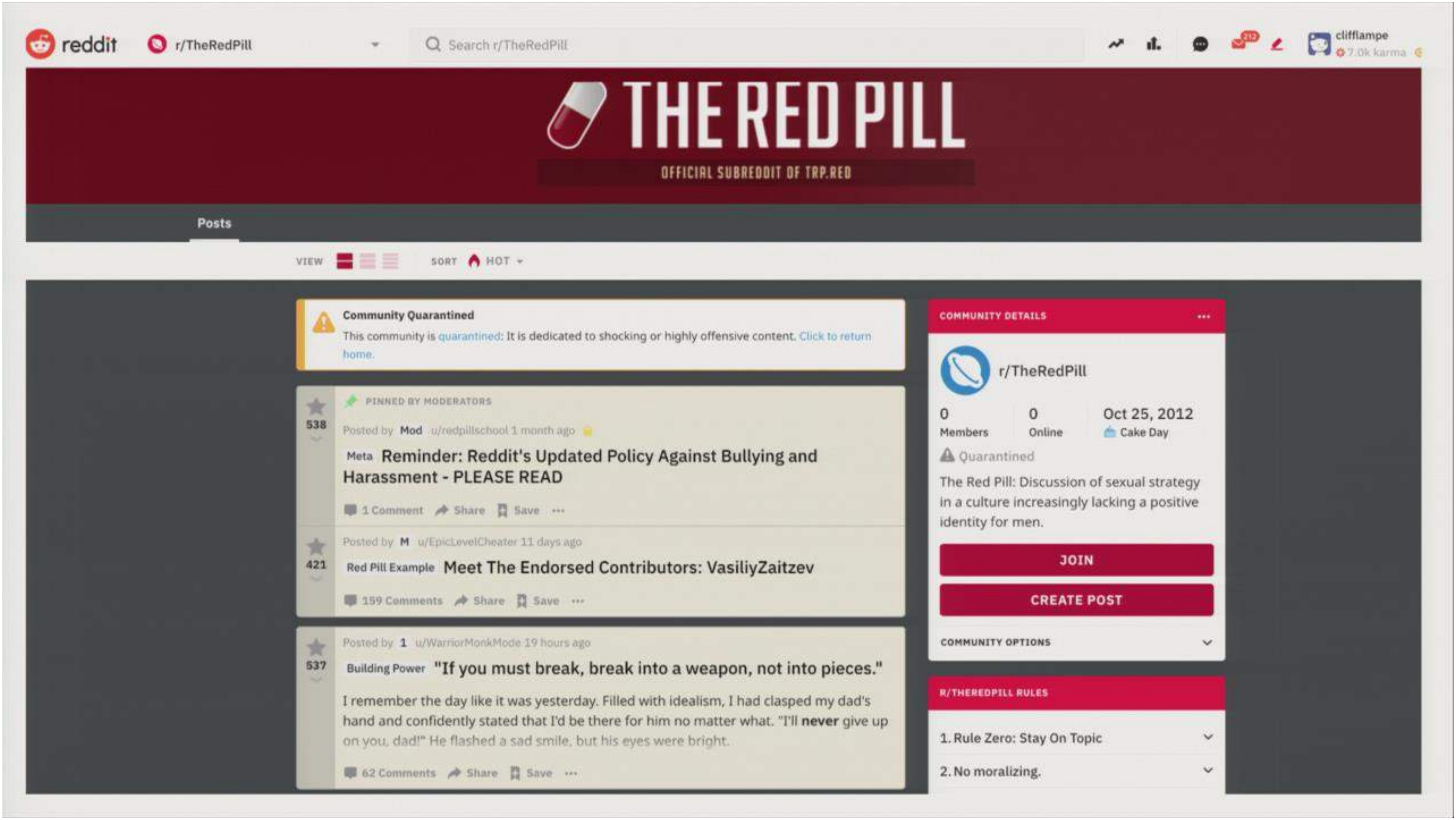
https://obamawatcher.com/2019/10/george-was-obama-in-disguise/?fbclid=IwAR3LLPF29eIO7yv5zWKOfD6t-ZDfuz2Fku_WPV_zBu3FO4fp-OEvvmfXM1M

2:34 PM

Aaron Slifka

Anima Christi
*translated by John Henry Cardinal Newman*
Soul of Christ, be my sanctification;
Body of Christ, be my salvation;
Blood of Christ, fill all my veins;
Water of Christ's side, wash out my stains;
Passion of Christ, my comfort be;
O good Jesus, listen to me;
In Thy wounds I fain would hide;
Ne'er to be parted from Thy side;

2:43 PM



**Community Quarantined**

This community is quarantined: It is dedicated to shocking or highly offensive content. Click to return home.

PINNED BY MODERATORS

538 Posted by Mod u/redpillschool 1 month ago

Meta **Reminder: Reddit's Updated Policy Against Bullying and Harassment - PLEASE READ**

1 Comment Share Save ···

421 Posted by M u/EpicLevelCheater 11 days ago

Red Pill Example **Meet The Endorsed Contributors: VasiliyZaitzev**

159 Comments Share Save ···

537 Posted by 1 u/WarriorMonkMode 19 hours ago

Building Power **"If you must break, break into a weapon, not into pieces."**

I remember the day like it was yesterday. Filled with idealism, I had clasped my dad's hand and confidently stated that I'd be there for him no matter what. "I'll **never** give up on you, dad!" He flashed a sad smile, but his eyes were bright.

62 Comments Share Save ···

COMMUNITY DETAILS

r/TheRedPill

| 0 | 0 | Oct 25, 2012 |
|---|---|---|
| Members | Online | Cake Day |

Quarantined

The Red Pill: Discussion of sexual strategy in a culture increasingly lacking a positive identity for men.

R/THEREDPILL RULES

1. Rule Zero: Stay On Topic
2. No moralizing.

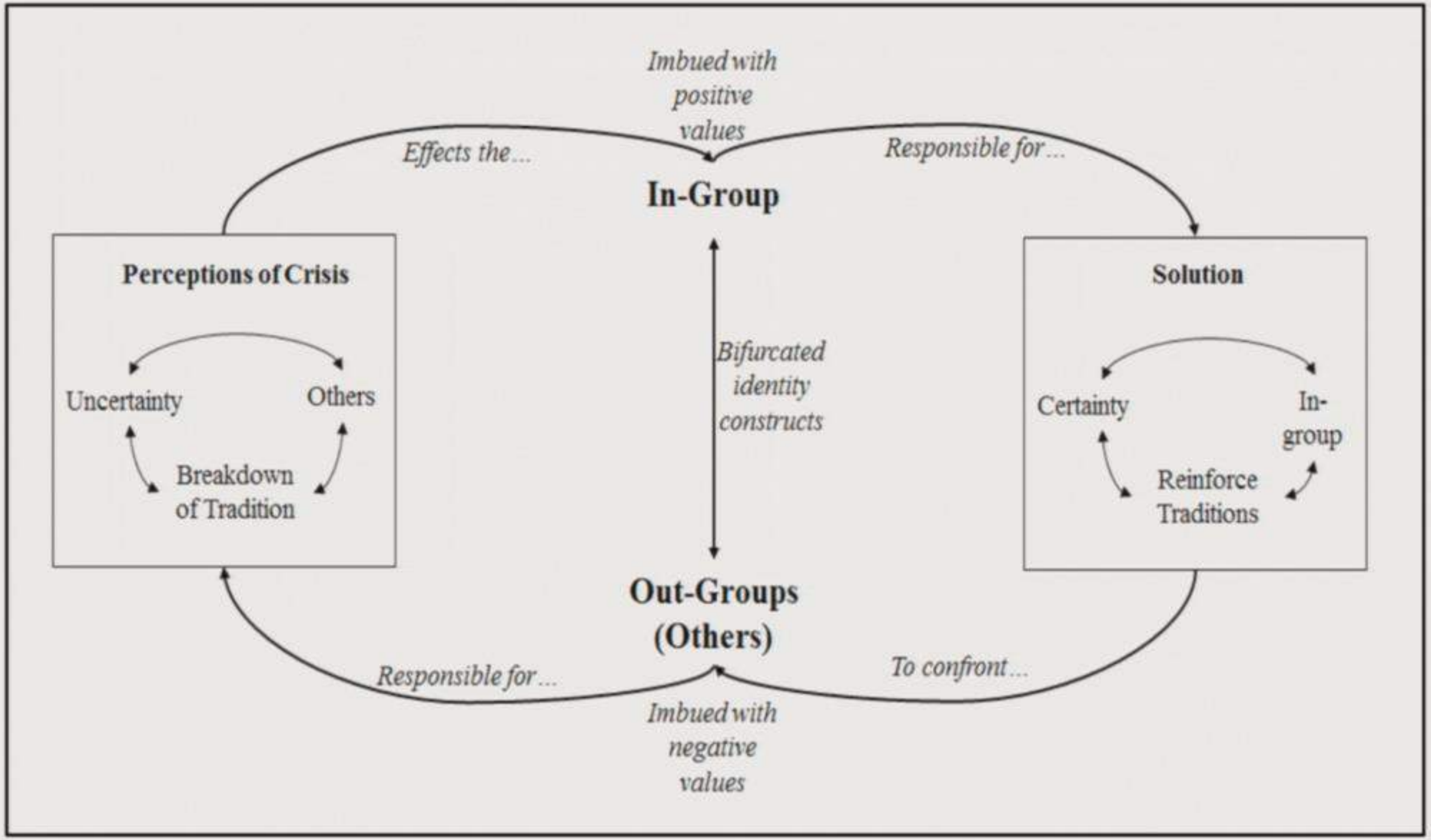
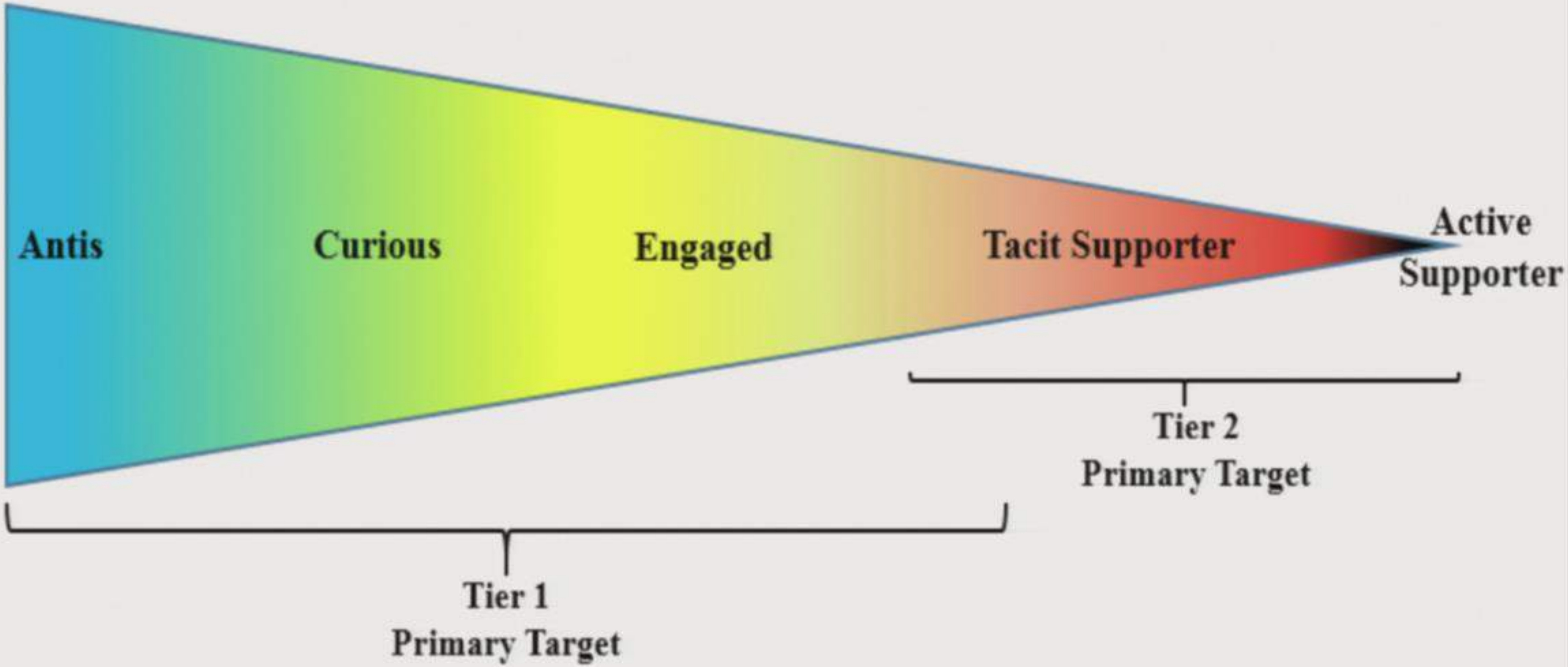


**Figure 1: Violent extremist "system of meaning" and its self-reinforcing dynamic**

Ingram, H. J. (2016). A "Linkage-Based" Approach to Combating Militant Islamist Propaganda: A Two-Tiered Framework for Practitioners.



Antis

Curious

Engaged

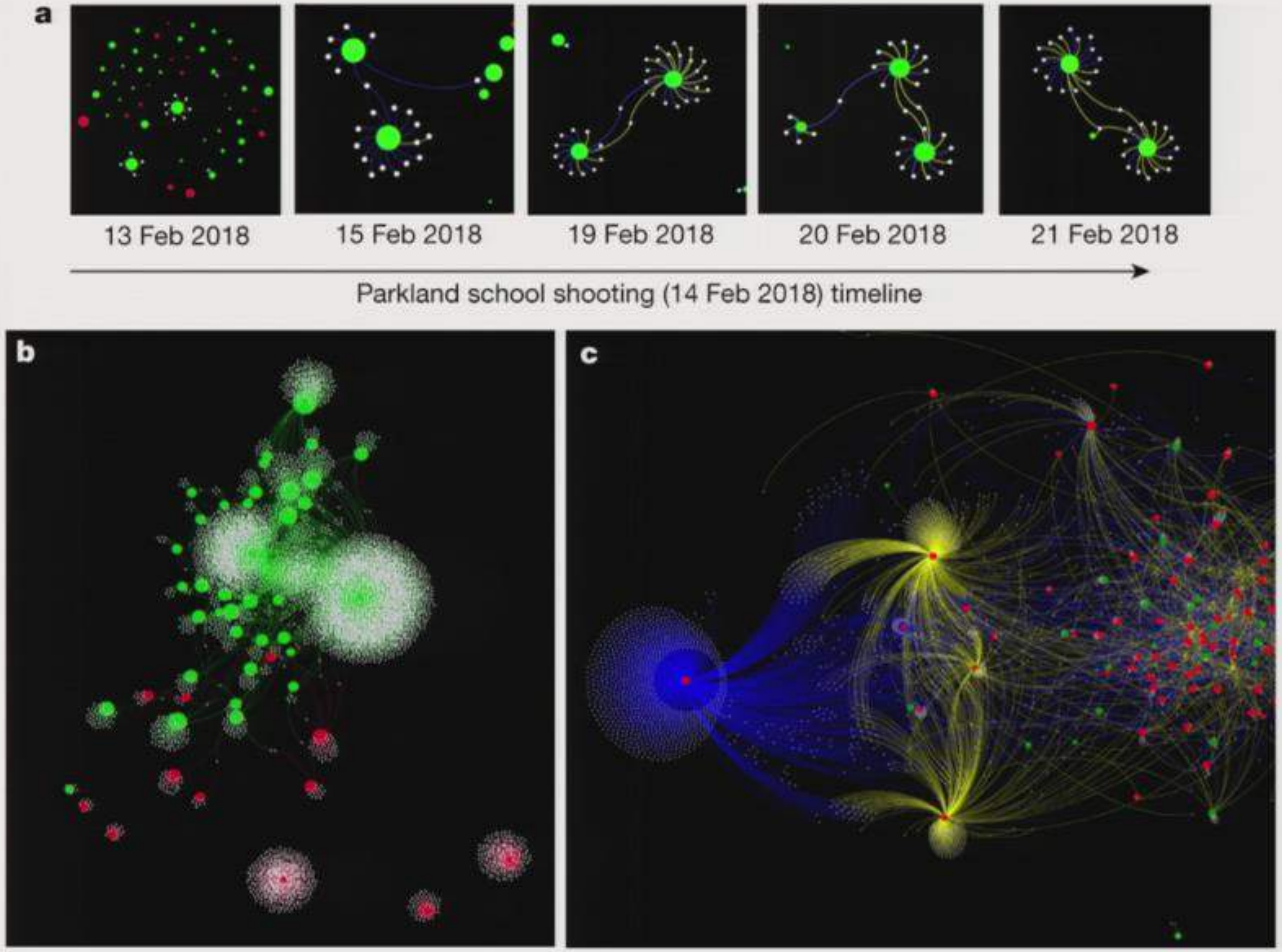
Tacit Supporter

Active Supporter

Tier 2
Primary Target

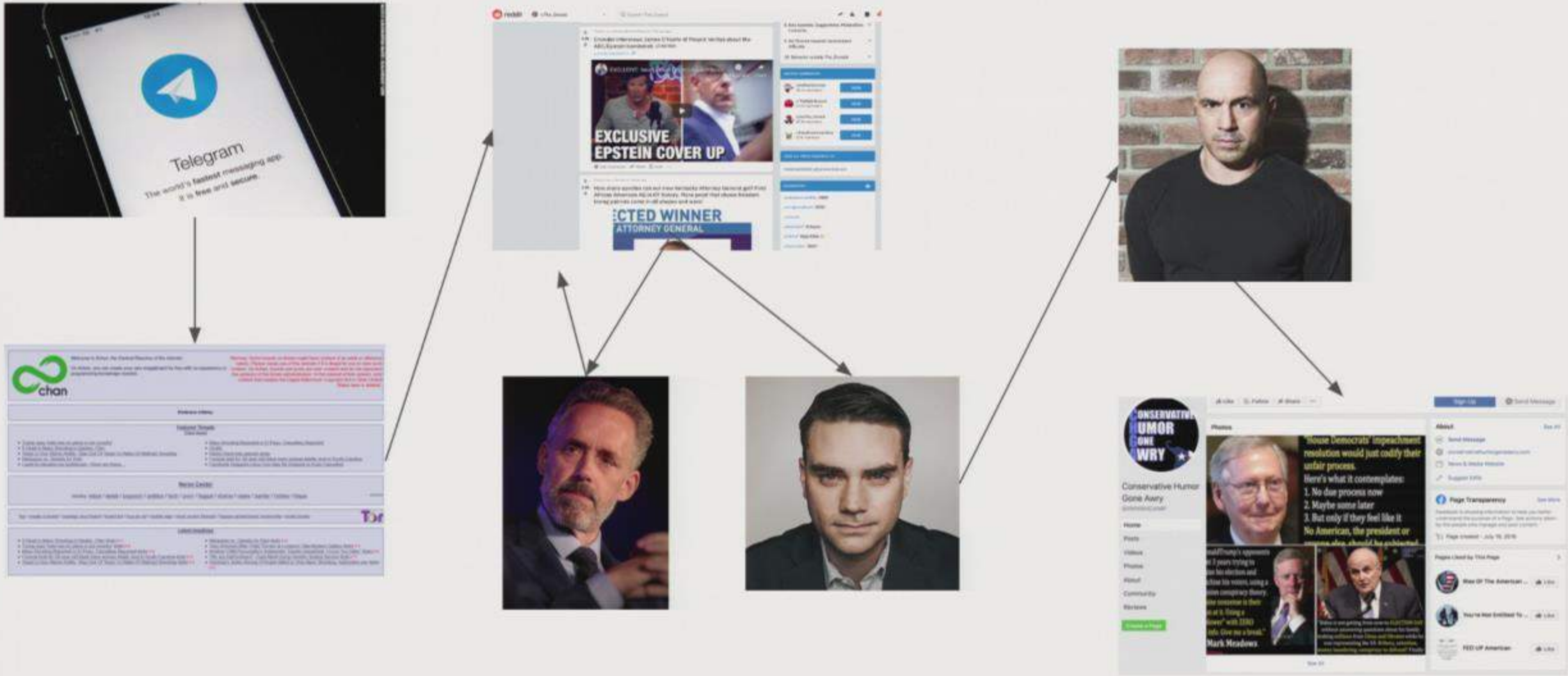
Tier 1
Primary Target

# Hate spreads virally, and in a resilient way.

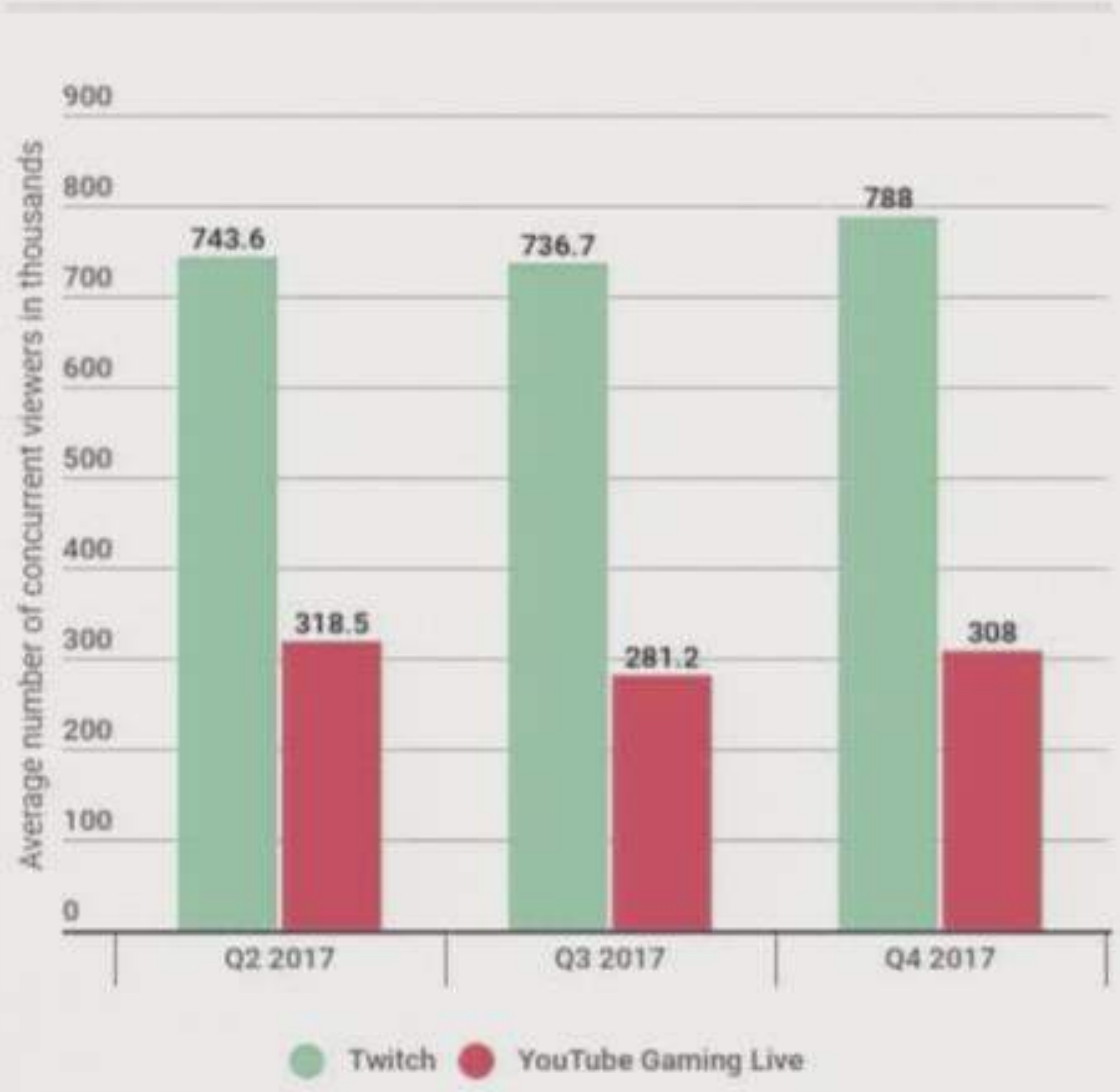


N. F. Johnson, R. Leahy, N. Johnson Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota & S. Wuchty. Hidden resilience and adaptive dynamics of the global online hate ecology. Nature, 2019 DOI: 10.1038/s41586-019-1494-7

# Pathways of radicalization


Telegram
The world's fastest messaging app.
It is free and secure.
8chan
Tor
reddit
EXCLUSIVE EPSTEIN COVER UP
ECTED WINNER
ATTORNEY GENERAL
CONSERVATIVE HUMOR GONE AWRY
Conservative Humor Gone Awry
Home
Photos
About
Page Transparency
"House Democrats' impeachment resolution would just codify their unfair process.
Here's what it contemplates:
1. No due process now
2. Maybe some later
3. But only if they feel like it
No American, the president or
Mark Meadows

# Example: Pinkpilling in games



2017 Average Number of Concurrent Viewers on Twitch, Youtube Gaming Live in 1000s (Q2 to Q4 of 2017)

Pinkpilling: using language that is derogatory of progressive terms in order to attempt to radicalize children.

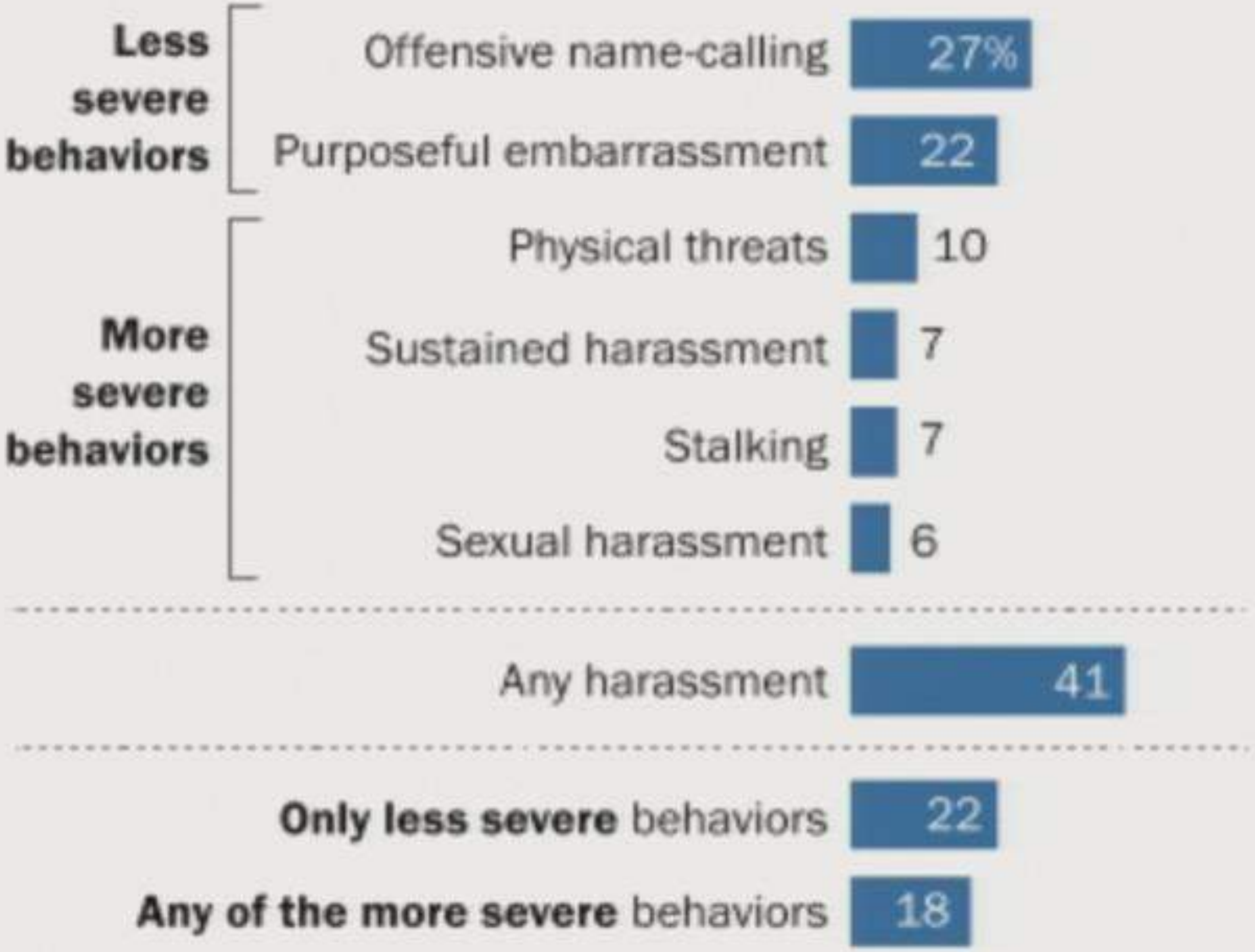
“Watch this snowflake run for their safe space! What a cuck move.”

# Online harassment

# 41% have experienced harassment (up from 35% in 2014)

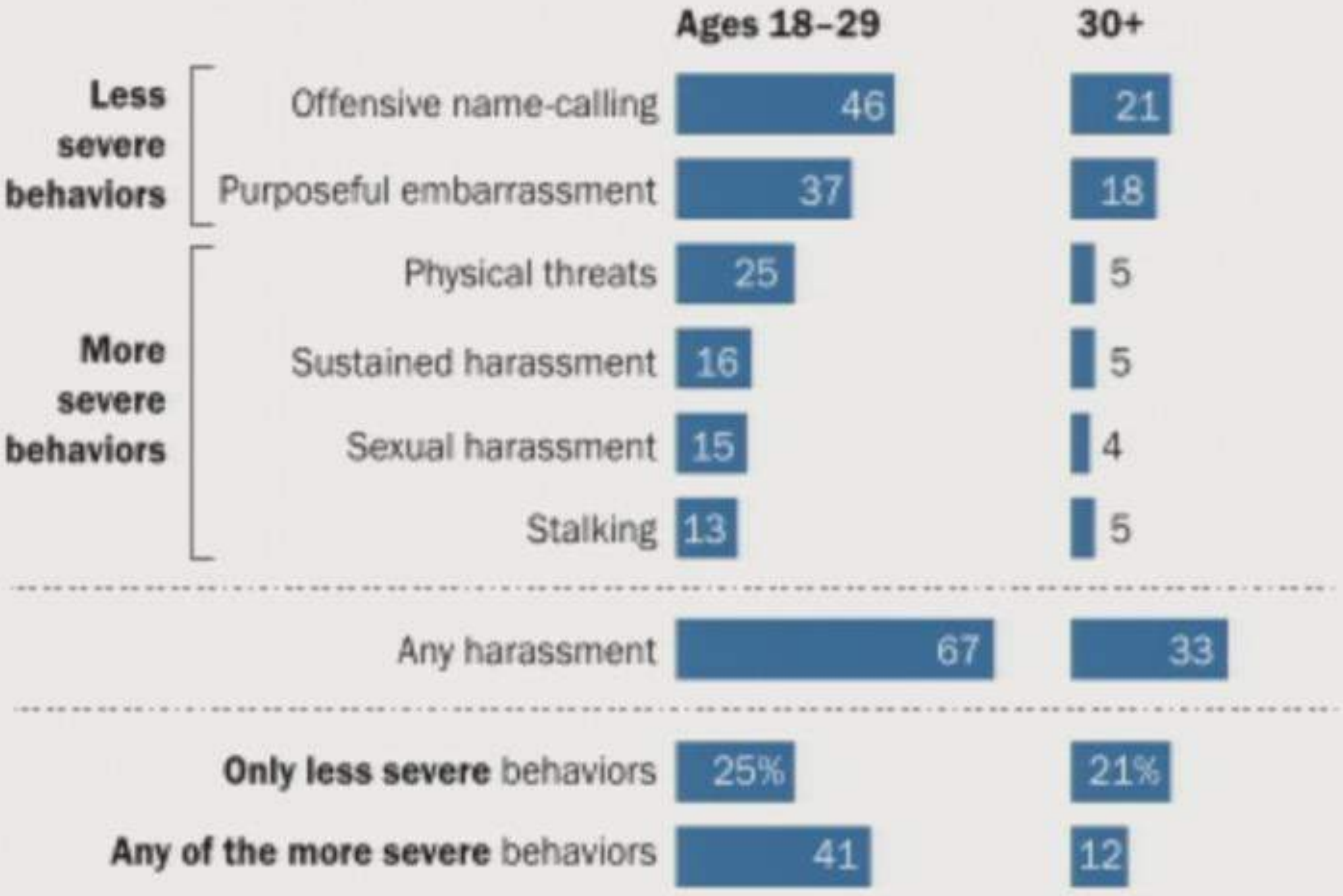
## 66% have witnessed it

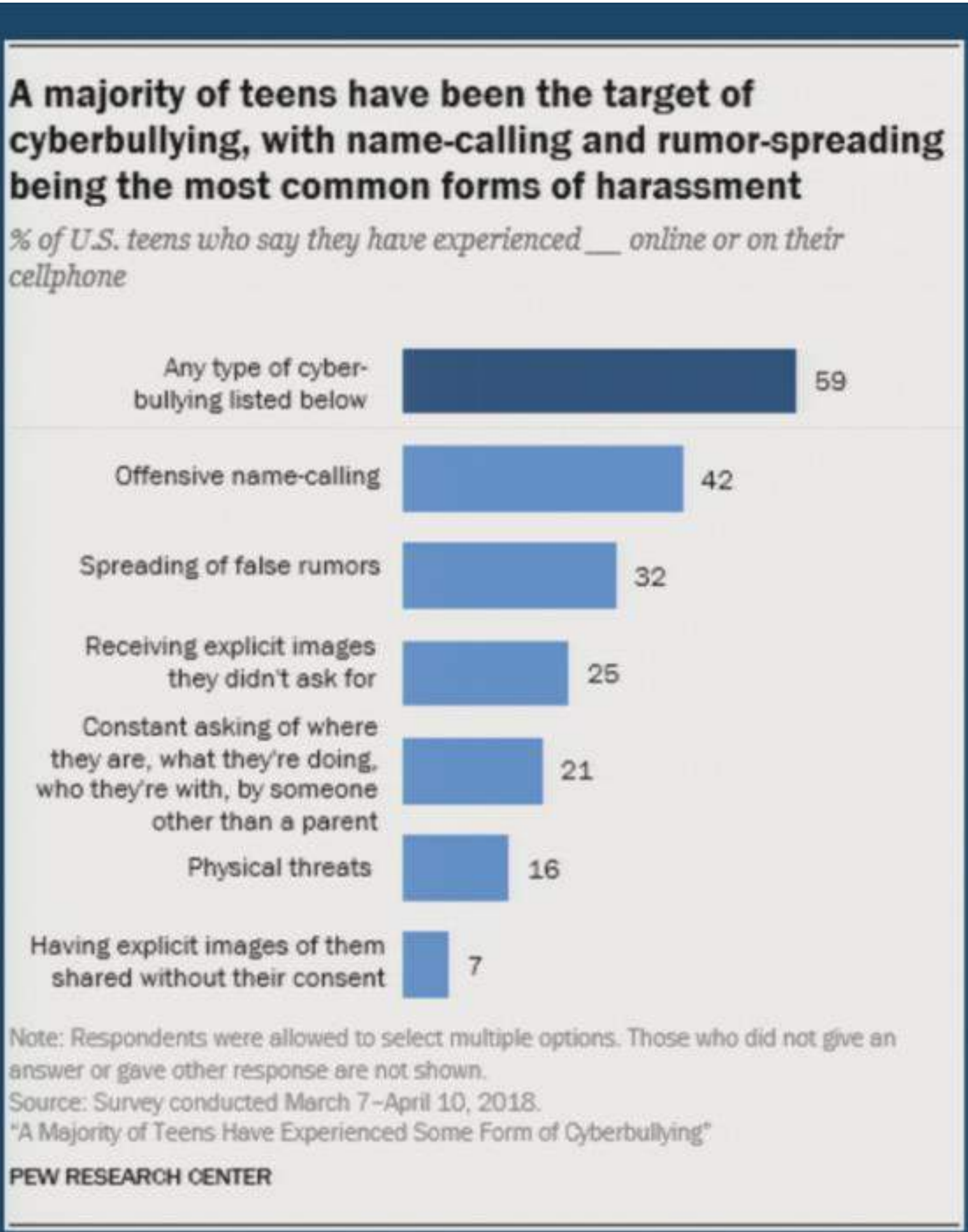
*% of U.S. adults who have experienced ____ online*



# Younger adults especially likely to encounter severe forms of online harassment

*% of U.S. adults who say they have experienced the following types of harassment online, by age*





## A majority of teens have been the target of cyberbullying, with name-calling and rumor-spreading being the most common forms of harassment

*% of U.S. teens who say they have experienced __ online or on their cellphone*

Note: Respondents were allowed to select multiple options. Those who did not give an answer or gave other response are not shown.

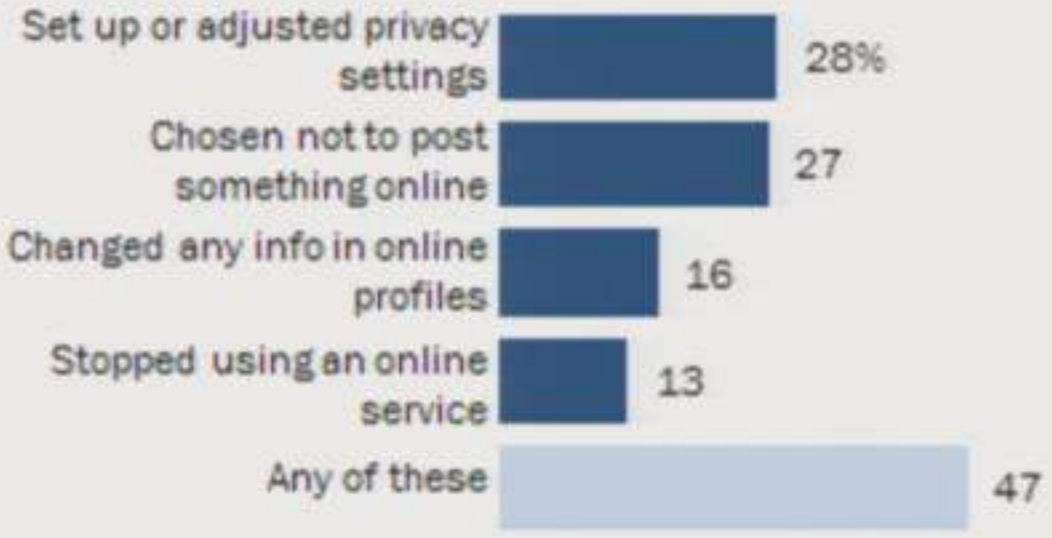
Source: Survey conducted March 7–April 10, 2018.

"A Majority of Teens Have Experienced Some Form of Cyberbullying"

PEW RESEARCH CENTER

## More than a quarter of Americans have chosen to not post something online after seeing harassment of others

*% of U.S. adults who have ____ after witnessing harassing behaviors directed toward others online*

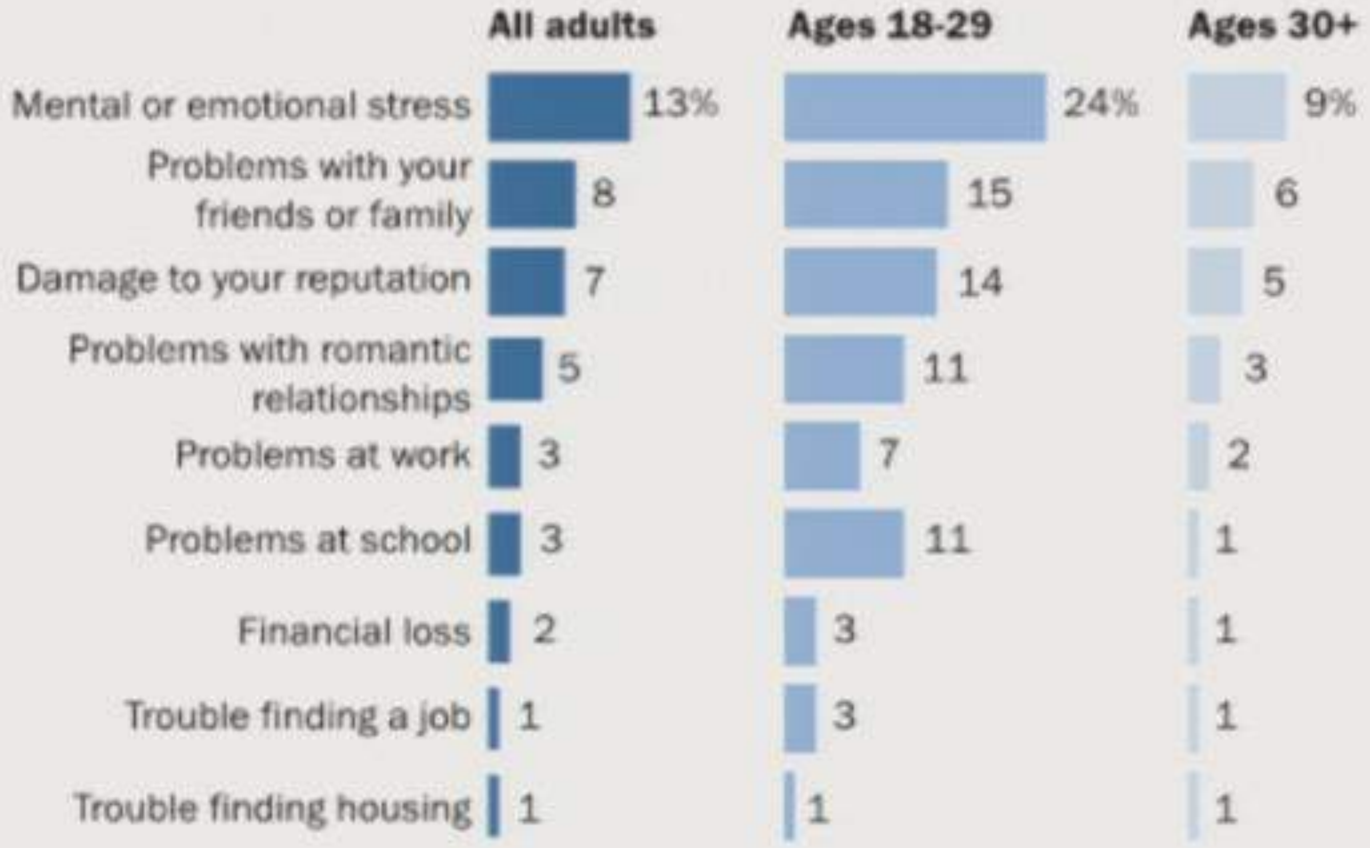


Note: Total may not add to 100% because respondents could select multiple options.
Source: Survey conducted Jan. 9-23, 2017
"Online Harassment 2017"

PEW RESEARCH CENTER

## Online harassment has caused mental or emotional stress for 13% of Americans overall – including 24% of young adults

*% of U.S. adults who say they have experienced the following due to online harassment*



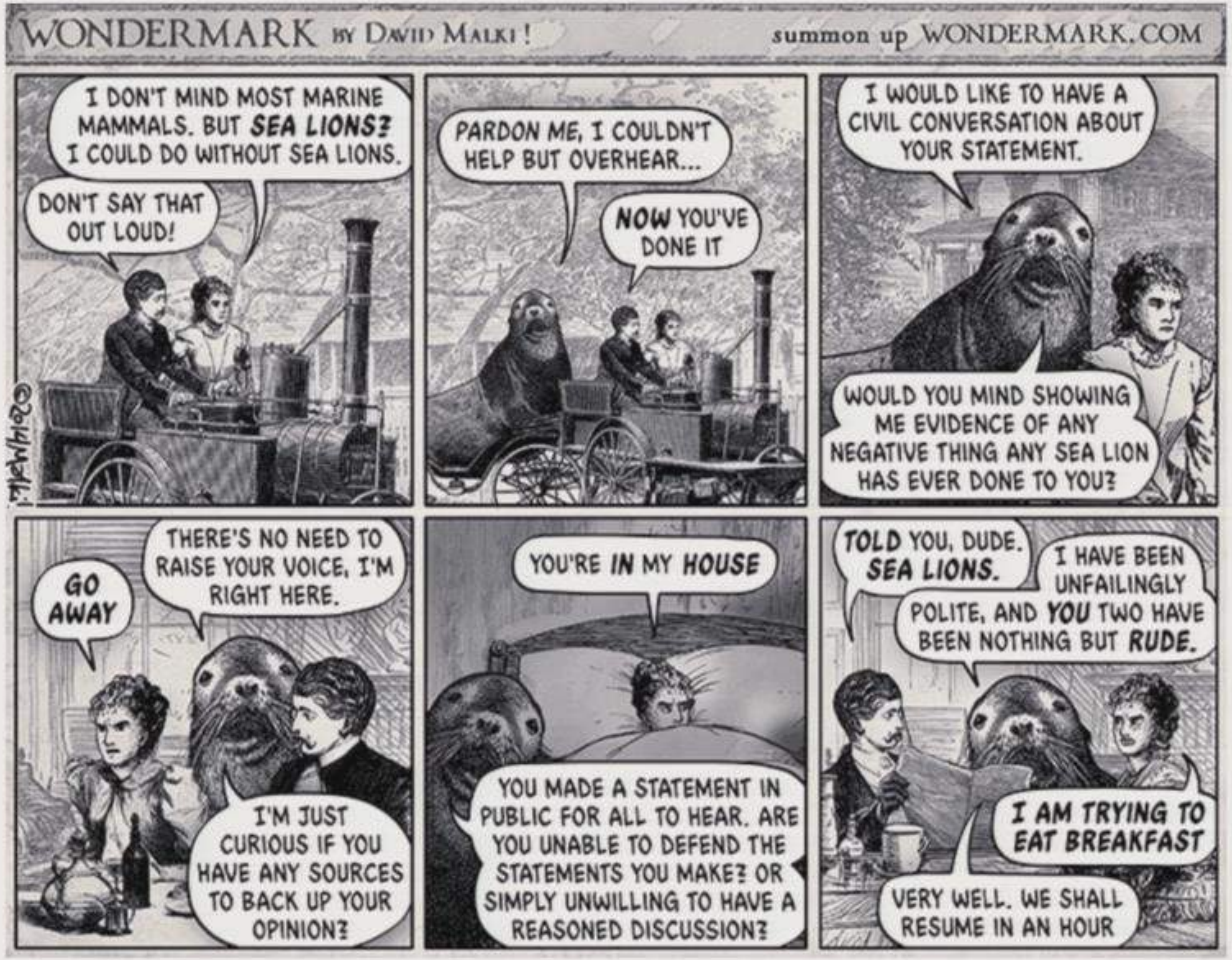
Note: Total may not add to 100% because respondents could select multiple options.
Source: Survey conducted Jan. 9-23, 2017.
"Online Harassment 2017"

PEW RESEARCH CENTER

# Example: Sealioning on Wikipedia

# Sealioning

**Sealioning** (also spelled sea-lioning and sea lioning) is a type of trolling or harassment which consists of pursuing people with persistent requests for evidence or repeated questions, while maintaining a pretense of civility.


WONDERMARK BY DAVID MALKI !
summon up WONDERMARK.COM
I DON'T MIND MOST MARINE MAMMALS. BUT SEA LIONS? I COULD DO WITHOUT SEA LIONS.
DON'T SAY THAT OUT LOUD!
PARDON ME, I COULDN'T HELP BUT OVERHEAR...
NOW YOU'VE DONE IT
I WOULD LIKE TO HAVE A CIVIL CONVERSATION ABOUT YOUR STATEMENT.
WOULD YOU MIND SHOWING ME EVIDENCE OF ANY NEGATIVE THING ANY SEA LION HAS EVER DONE TO YOU?
GO AWAY
THERE'S NO NEED TO RAISE YOUR VOICE, I'M RIGHT HERE.
I'M JUST CURIOUS IF YOU HAVE ANY SOURCES TO BACK UP YOUR OPINION?
YOU'RE IN MY HOUSE
YOU MADE A STATEMENT IN PUBLIC FOR ALL TO HEAR. ARE YOU UNABLE TO DEFEND THE STATEMENTS YOU MAKE? OR SIMPLY UNWILLING TO HAVE A REASONED DISCUSSION?
TOLD YOU, DUDE. SEA LIONS.
I HAVE BEEN UNFAILINGLY POLITE, AND YOU TWO HAVE BEEN NOTHING BUT RUDE.
I AM TRYING TO EAT BREAKFAST
VERY WELL. WE SHALL RESUME IN AN HOUR

# Women are especially victims of sealioning in Wikipedia Talk pages

Aaron Halfaker

Elizabeth Whittaker

"So, if someone accuses you of violating a neutral point of view, that immediately is a red flag for Wikipedia. And of course it's very difficult, probably close to impossible in some ways, to prove that you haven't done that. So you have to have diffs, you have to have examples of things that do represent what happened."

# What do we do when someone has done a “bad” thing?

Offenses need to be both detected and punished.

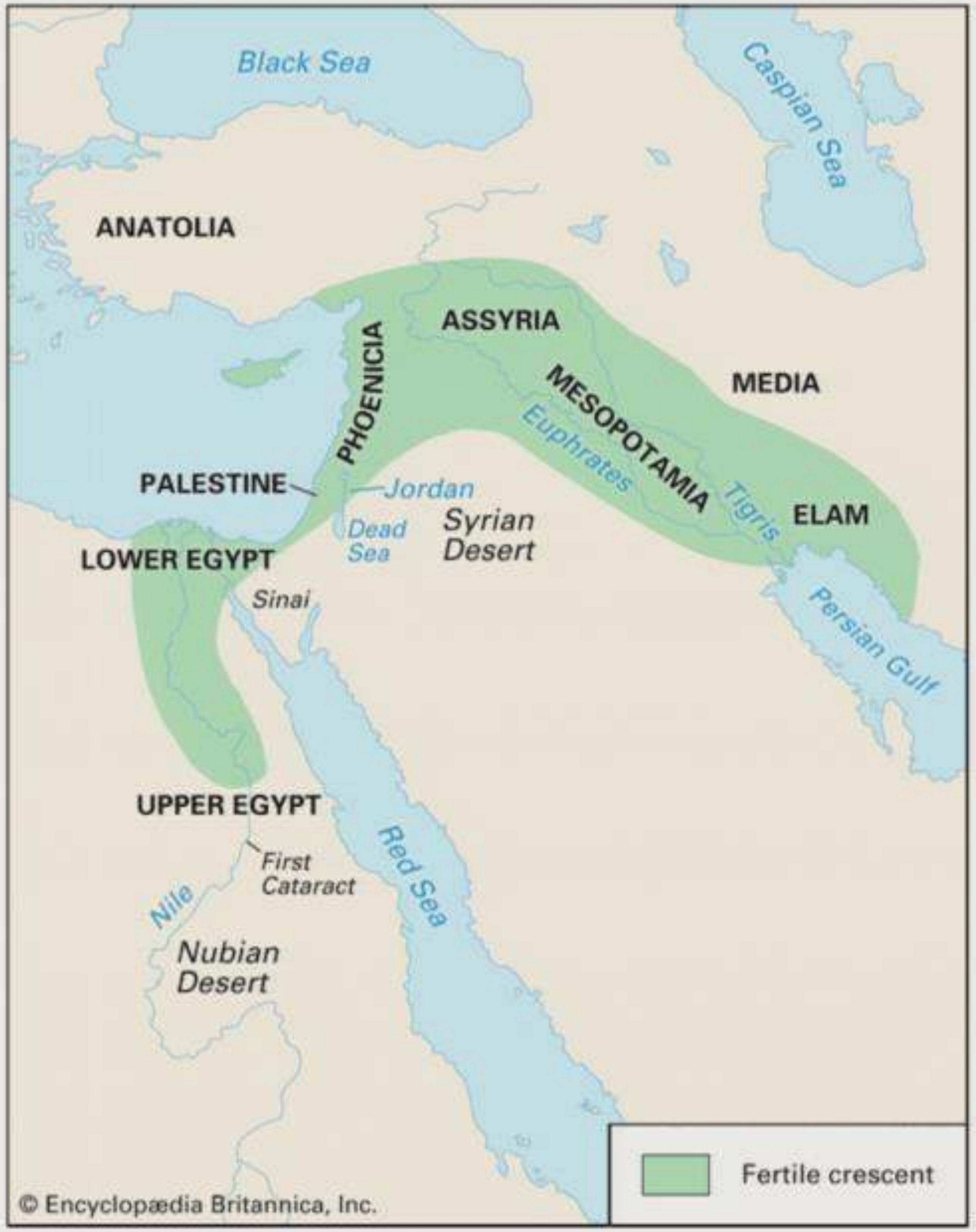
# What is the goal of justice?

# 1754 BCE Code of the Hammurabi

"If a man destroy the eye of another man, they shall destroy his eye. If one break a man's bone, they shall break his bone. If one destroy the eye of a freeman or break the bone of a freeman he shall pay one gold mina. If one destroy the eye of a man's slave or break a bone of a man's slave he shall pay one-half his price."

The Code of the Hammurabi became the template for Western forms of justice.

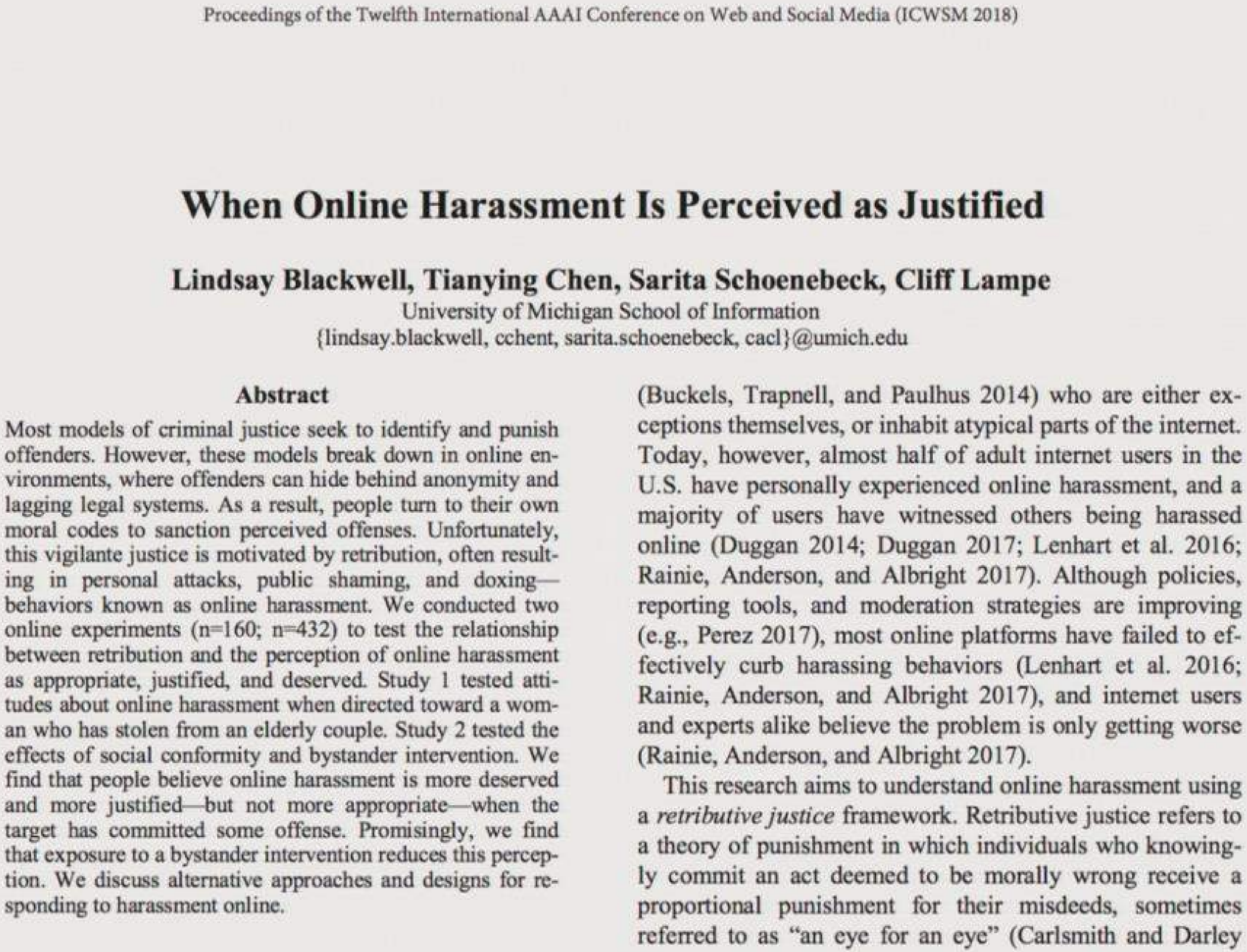
*What is our epistemology of justice? How often do we challenge that?*



© Encyclopædia Britannica, Inc.

Retributive Justice is primarily concerned with delivering a "just desert" for a morally wrong act.

People have a strong appetite for retributive justice in online spaces.



Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

# When Online Harassment Is Perceived as Justified

**Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, Cliff Lampe**
University of Michigan School of Information
{lindsay.blackwell, cchent, sarita.schoenebeck, cacl}@umich.edu

## Abstract

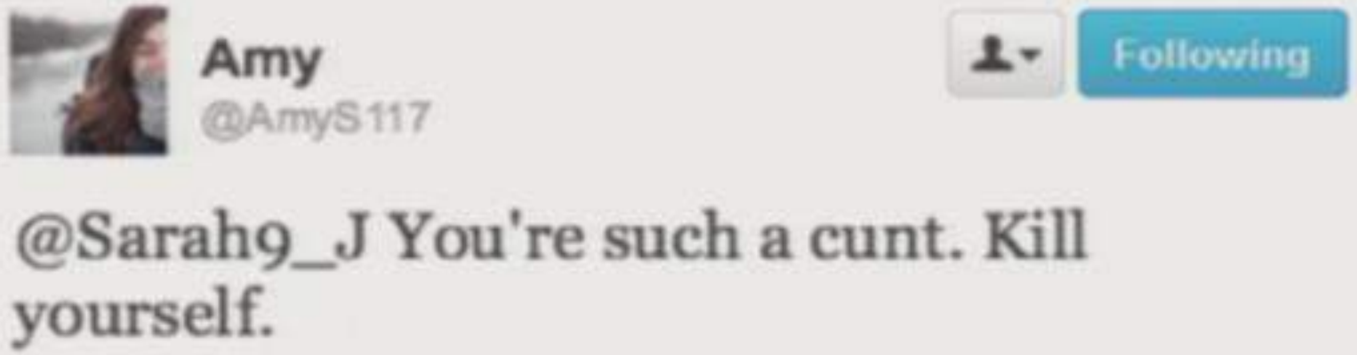
Most models of criminal justice seek to identify and punish offenders. However, these models break down in online environments, where offenders can hide behind anonymity and lagging legal systems. As a result, people turn to their own moral codes to sanction perceived offenses. Unfortunately, this vigilante justice is motivated by retribution, often resulting in personal attacks, public shaming, and doxing—behaviors known as online harassment. We conducted two online experiments (n=160; n=432) to test the relationship between retribution and the perception of online harassment as appropriate, justified, and deserved. Study 1 tested attitudes about online harassment when directed toward a woman who has stolen from an elderly couple. Study 2 tested the effects of social conformity and bystander intervention. We find that people believe online harassment is more deserved and more justified—but not more appropriate—when the target has committed some offense. Promisingly, we find that exposure to a bystander intervention reduces this perception. We discuss alternative approaches and designs for responding to harassment online.

(Buckels, Trapnell, and Paulhus 2014) who are either exceptions themselves, or inhabit atypical parts of the internet. Today, however, almost half of adult internet users in the U.S. have personally experienced online harassment, and a majority of users have witnessed others being harassed online (Duggan 2014; Duggan 2017; Lenhart et al. 2016; Rainie, Anderson, and Albright 2017). Although policies, reporting tools, and moderation strategies are improving (e.g., Perez 2017), most online platforms have failed to effectively curb harassing behaviors (Lenhart et al. 2016; Rainie, Anderson, and Albright 2017), and internet users and experts alike believe the problem is only getting worse (Rainie, Anderson, and Albright 2017).

This research aims to understand online harassment using a *retributive justice* framework. Retributive justice refers to a theory of punishment in which individuals who knowingly commit an act deemed to be morally wrong receive a proportional punishment for their misdeeds, sometimes referred to as "an eye for an eye" (Carlsmith and Darley

Condition 1: “Sarah stole $100 from an elderly couple.”

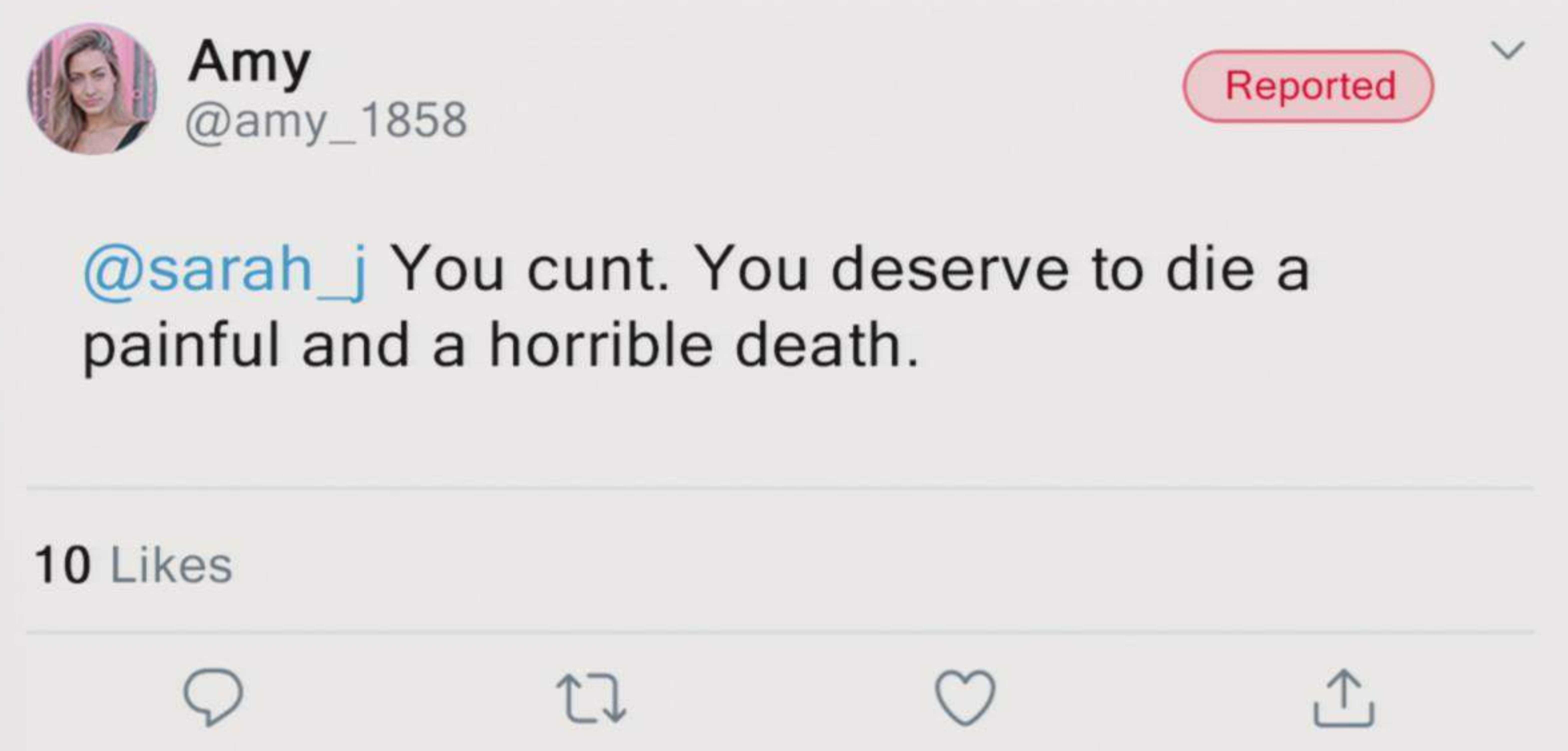
Condition 2: “Sarah stole $10,000 from an elderly couple.”



@Sarah9_J You're such a cunt. Kill yourself.

People rated the harassment as more deserved for the larger amount. No difference in appropriateness.

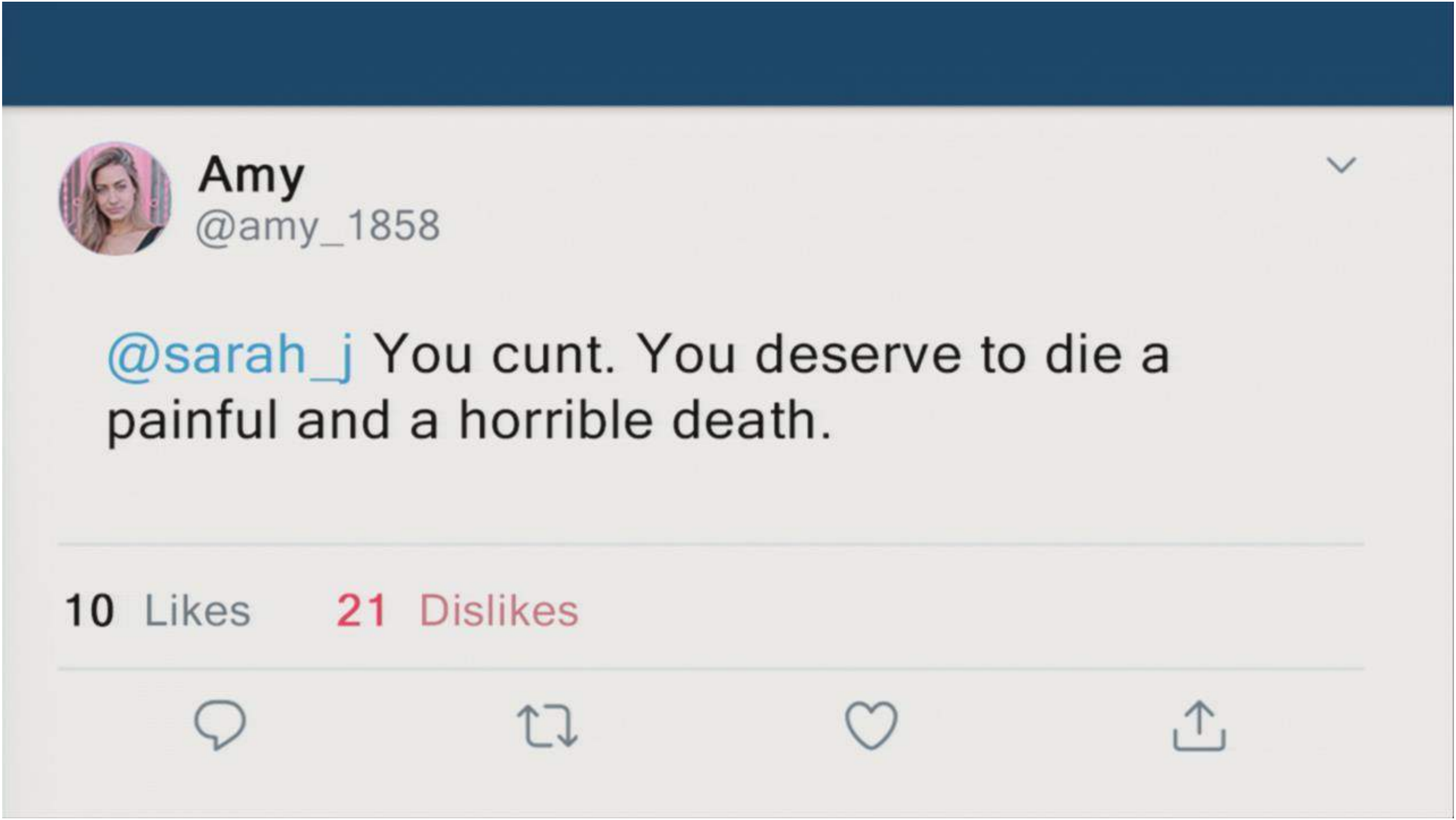
# Extending this work



Amy
@amy_1858

@sarah_j You cunt. You deserve to die a painful and a horrible death.

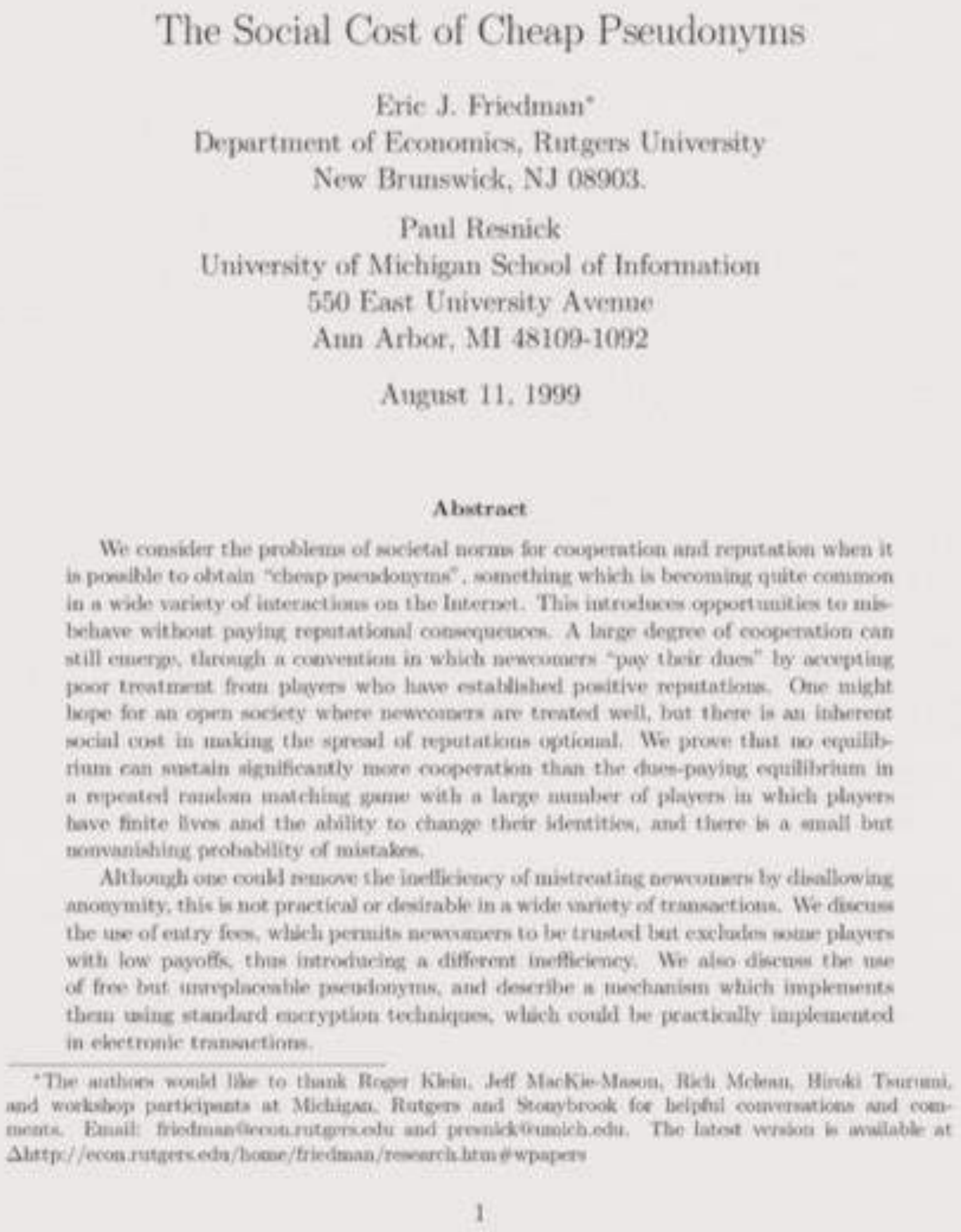
10 Likes



Amy
@amy_1858

@sarah_j You cunt. You deserve to die a painful and a horrible death.

Retributive Justice mechanisms often fail online.



# The Social Cost of Cheap Pseudonyms

Eric J. Friedman*
Department of Economics, Rutgers University
New Brunswick, NJ 08903.

Paul Resnick
University of Michigan School of Information
550 East University Avenue
Ann Arbor, MI 48109-1092

August 11, 1999

### Abstract

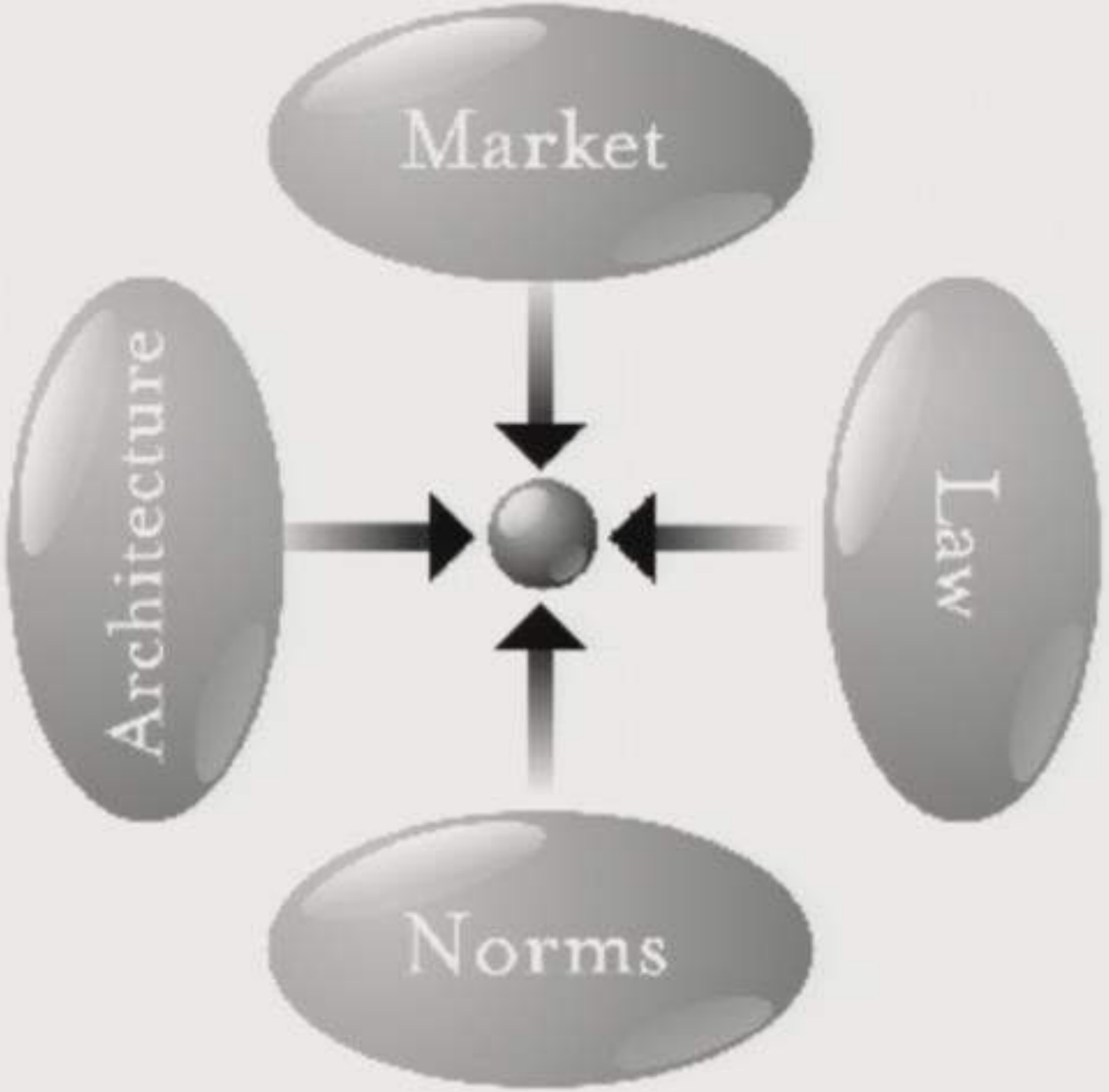
We consider the problems of societal norms for cooperation and reputation when it is possible to obtain "cheap pseudonyms", something which is becoming quite common in a wide variety of interactions on the Internet. This introduces opportunities to misbehave without paying reputational consequences. A large degree of cooperation can still emerge, through a convention in which newcomers "pay their dues" by accepting poor treatment from players who have established positive reputations. One might hope for an open society where newcomers are treated well, but there is an inherent social cost in making the spread of reputations optional. We prove that no equilibrium can sustain significantly more cooperation than the dues-paying equilibrium in a repeated random matching game with a large number of players in which players have finite lives and the ability to change their identities, and there is a small but nonvanishing probability of mistakes.

Although one could remove the inefficiency of mistreating newcomers by disallowing anonymity, this is not practical or desirable in a wide variety of transactions. We discuss the use of entry fees, which permits newcomers to be trusted but excludes some players with low payoffs, thus introducing a different inefficiency. We also discuss the use of free but unreplaceable pseudonyms, and describe a mechanism which implements them using standard encryption techniques, which could be practically implemented in electronic transactions.

[*] The authors would like to thank Roger Klein, Jeff MacKie-Mason, Rich Mclean, Hiroki Tsurumi, and workshop participants at Michigan, Rutgers and Stonybrook for helpful conversations and comments. Email: friedman@econ.rutgers.edu and presnick@umich.edu. The latest version is available at Δhttp://econ.rutgers.edu/home/friedman/research.htm#wpapers

Imagine if online punishment rules were the same in our justice system...

# Lawrence Lessig - forces that can shape behavior



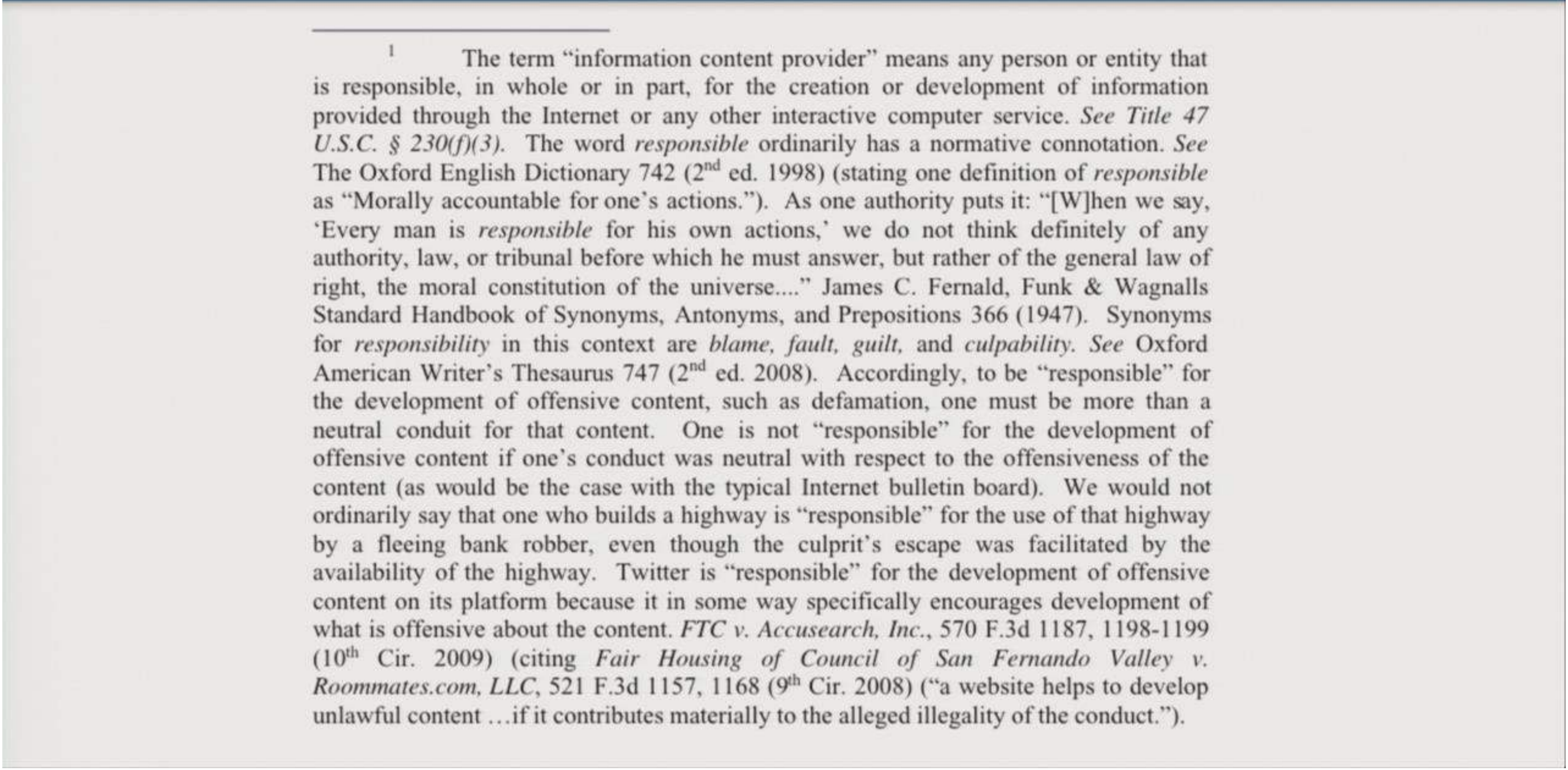
Machine Learning

New interfaces

Legal policies

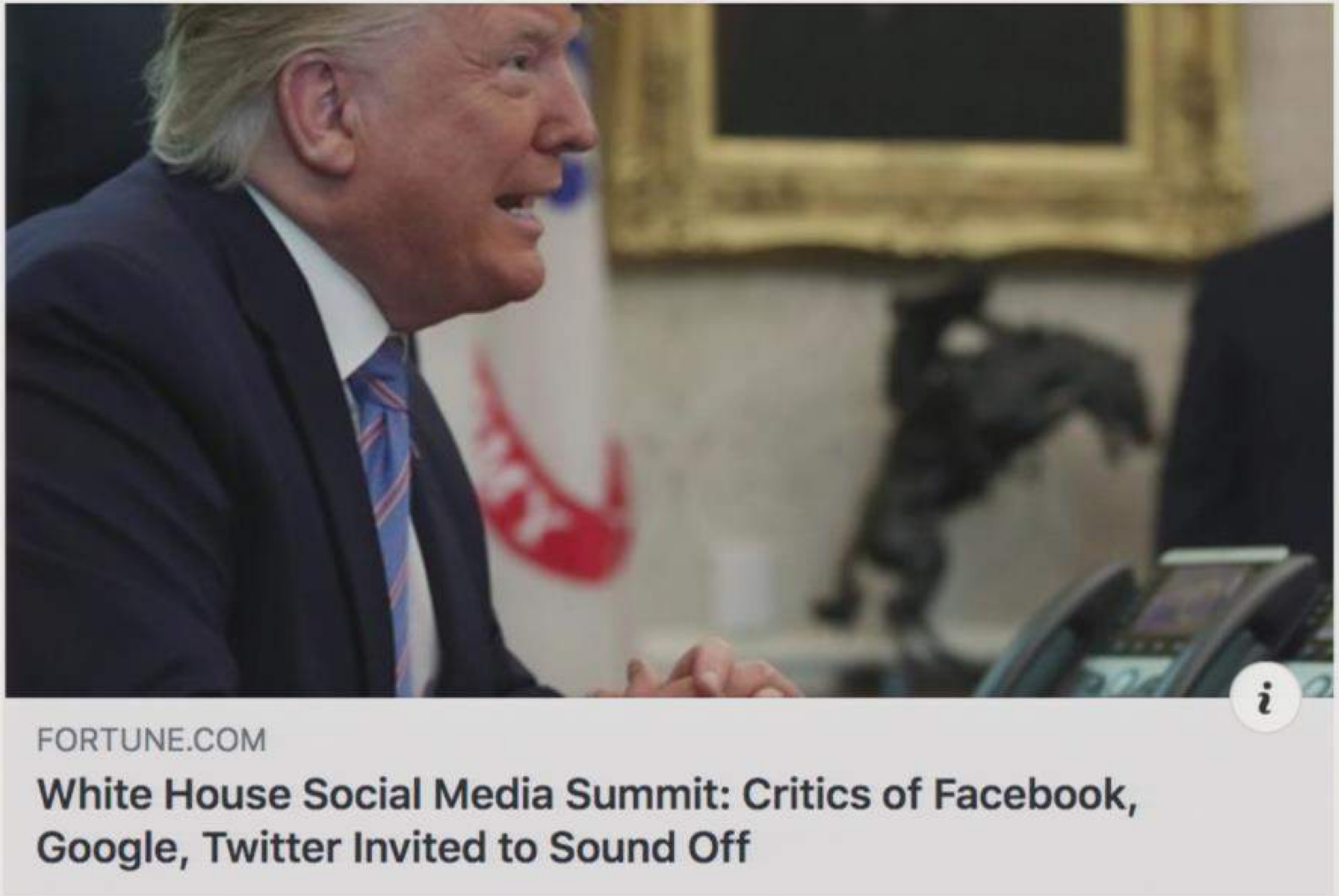
Site policies

# Section 230 of the Communications Decency Act of 1996



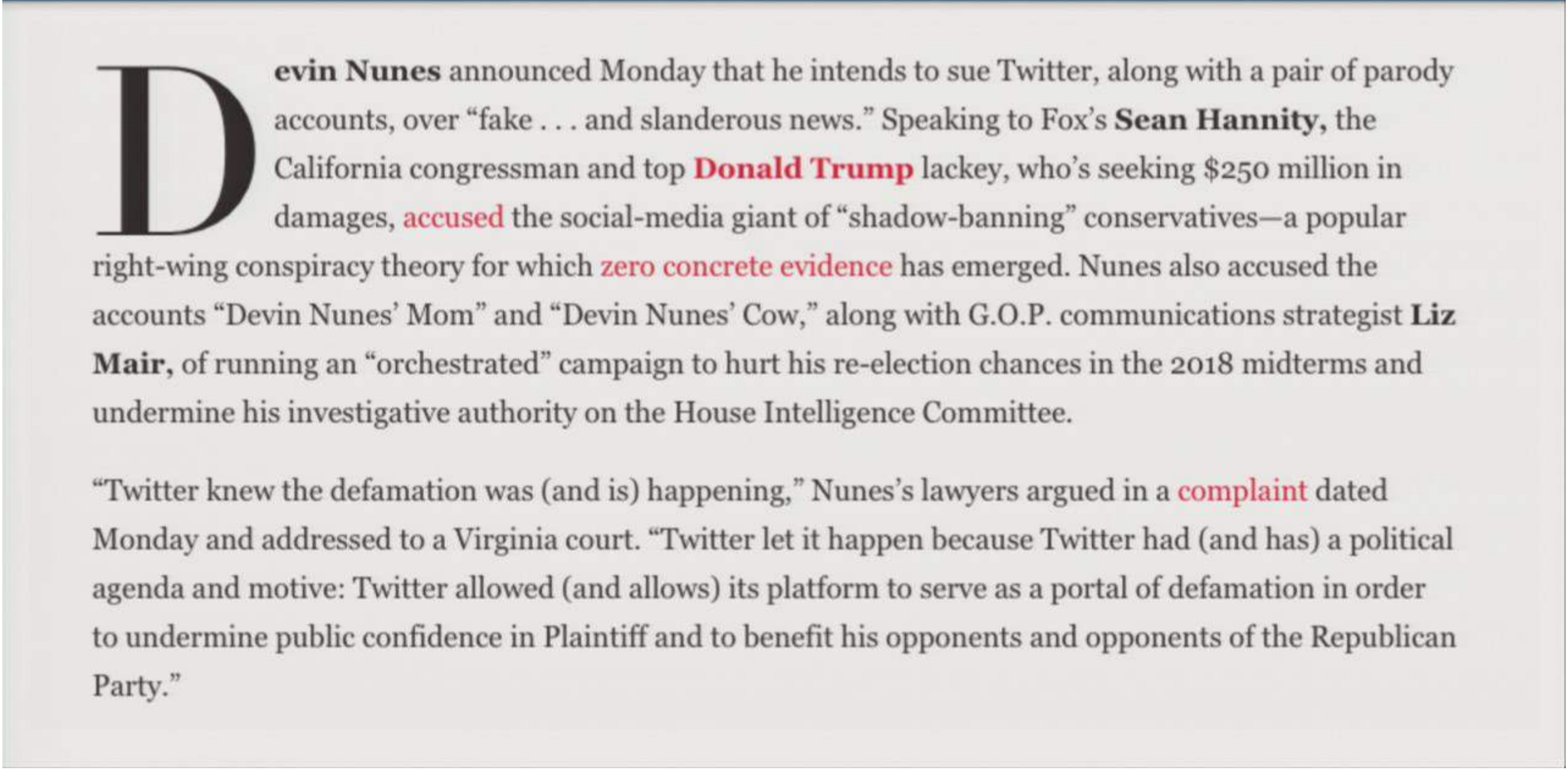
[1] The term "information content provider" means any person or entity that is responsible, in whole or in part, for the creation or development of information provided through the Internet or any other interactive computer service. *See Title 47 U.S.C. § 230(f)(3).* The word *responsible* ordinarily has a normative connotation. *See* The Oxford English Dictionary 742 (2nd ed. 1998) (stating one definition of *responsible* as "Morally accountable for one's actions."). As one authority puts it: "[W]hen we say, 'Every man is *responsible* for his own actions,' we do not think definitely of any authority, law, or tribunal before which he must answer, but rather of the general law of right, the moral constitution of the universe...." James C. Fernald, Funk & Wagnalls Standard Handbook of Synonyms, Antonyms, and Prepositions 366 (1947). Synonyms for *responsibility* in this context are *blame, fault, guilt,* and *culpability*. *See* Oxford American Writer's Thesaurus 747 (2nd ed. 2008). Accordingly, to be "responsible" for the development of offensive content, such as defamation, one must be more than a neutral conduit for that content. One is not "responsible" for the development of offensive content if one's conduct was neutral with respect to the offensiveness of the content (as would be the case with the typical Internet bulletin board). We would not ordinarily say that one who builds a highway is "responsible" for the use of that highway by a fleeing bank robber, even though the culprit's escape was facilitated by the availability of the highway. Twitter is "responsible" for the development of offensive content on its platform because it in some way specifically encourages development of what is offensive about the content. *FTC v. Accusearch, Inc.*, 570 F.3d 1187, 1198-1199 (10th Cir. 2009) (citing *Fair Housing of Council of San Fernando Valley v. Roommates.com, LLC*, 521 F.3d 1157, 1168 (9th Cir. 2008) ("a website helps to develop unlawful content …if it contributes materially to the alleged illegality of the conduct.").

Attempts at moderation inevitably lead to false positives.



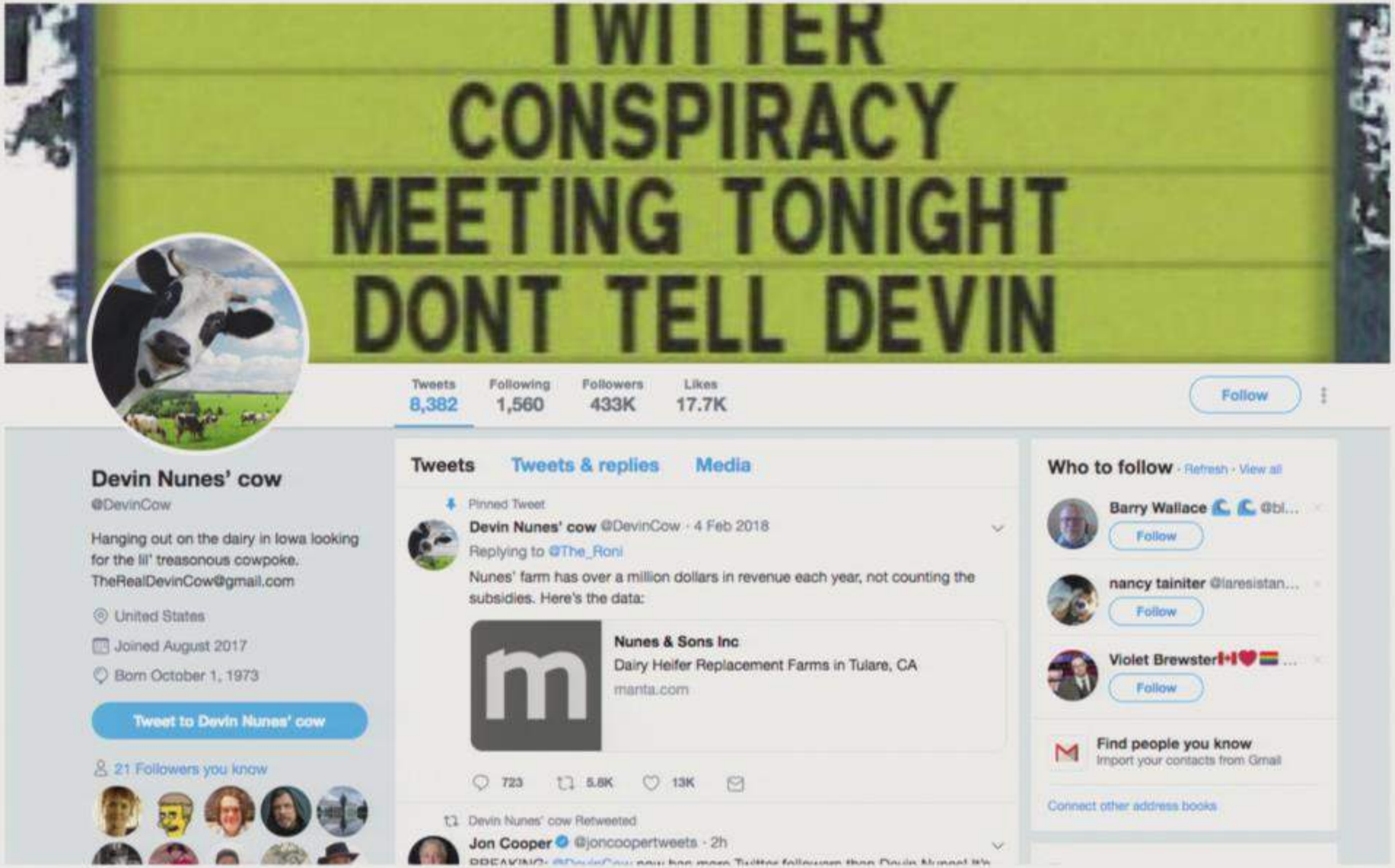
FORTUNE.COM

White House Social Media Summit: Critics of Facebook, Google, Twitter Invited to Sound Off



**Devin Nunes** announced Monday that he intends to sue Twitter, along with a pair of parody accounts, over "fake . . . and slanderous news." Speaking to Fox's **Sean Hannity,** the California congressman and top **Donald Trump** lackey, who's seeking $250 million in damages, accused the social-media giant of "shadow-banning" conservatives—a popular right-wing conspiracy theory for which zero concrete evidence has emerged. Nunes also accused the accounts "Devin Nunes' Mom" and "Devin Nunes' Cow," along with G.O.P. communications strategist **Liz Mair,** of running an "orchestrated" campaign to hurt his re-election chances in the 2018 midterms and undermine his investigative authority on the House Intelligence Committee.

"Twitter knew the defamation was (and is) happening," Nunes's lawyers argued in a complaint dated Monday and addressed to a Virginia court. "Twitter let it happen because Twitter had (and has) a political agenda and motive: Twitter allowed (and allows) its platform to serve as a portal of defamation in order to undermine public confidence in Plaintiff and to benefit his opponents and opponents of the Republican Party."



**Devin Nunes' cow**
@DevinCow

Hanging out on the dairy in Iowa looking for the lil' treasonous cowpoke. TheRealDevinCow@gmail.com

United States

Joined August 2017

Born October 1, 1973

Tweet to Devin Nunes' cow

21 Followers you know

Pinned Tweet

**Devin Nunes' cow** @DevinCow · 4 Feb 2018

Replying to @The_Roni

Nunes' farm has over a million dollars in revenue each year, not counting the subsidies. Here's the data:

**Nunes & Sons Inc**
Dairy Heifer Replacement Farms in Tulare, CA
manta.com

723 5.8K 13K

Devin Nunes' cow Retweeted

**Jon Cooper** @joncoopertweets · 2h

BREAKING: @DevinCow now has more Twitter followers than Devin Nunes! It's

**Barry Wallace** 🌊 🌊 @bl...
Follow

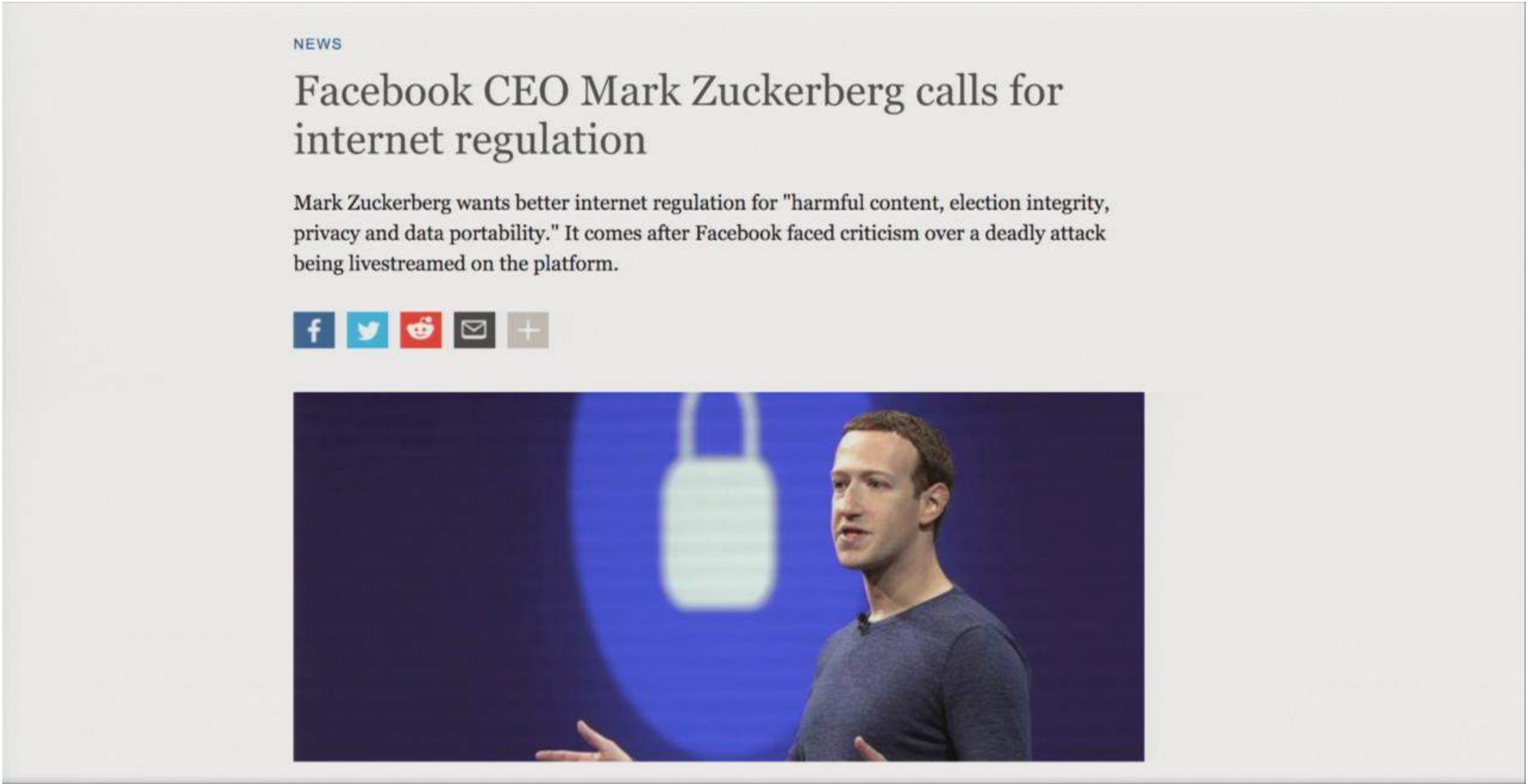
**nancy tainiter** @laresistan...
Follow

**Violet Brewster**🇨🇦❤️🏳️‍🌈 ...
Follow

**Find people you know**
Import your contacts from Gmail

Connect other address books

Platforms do not want to be the arbiters of free speech.



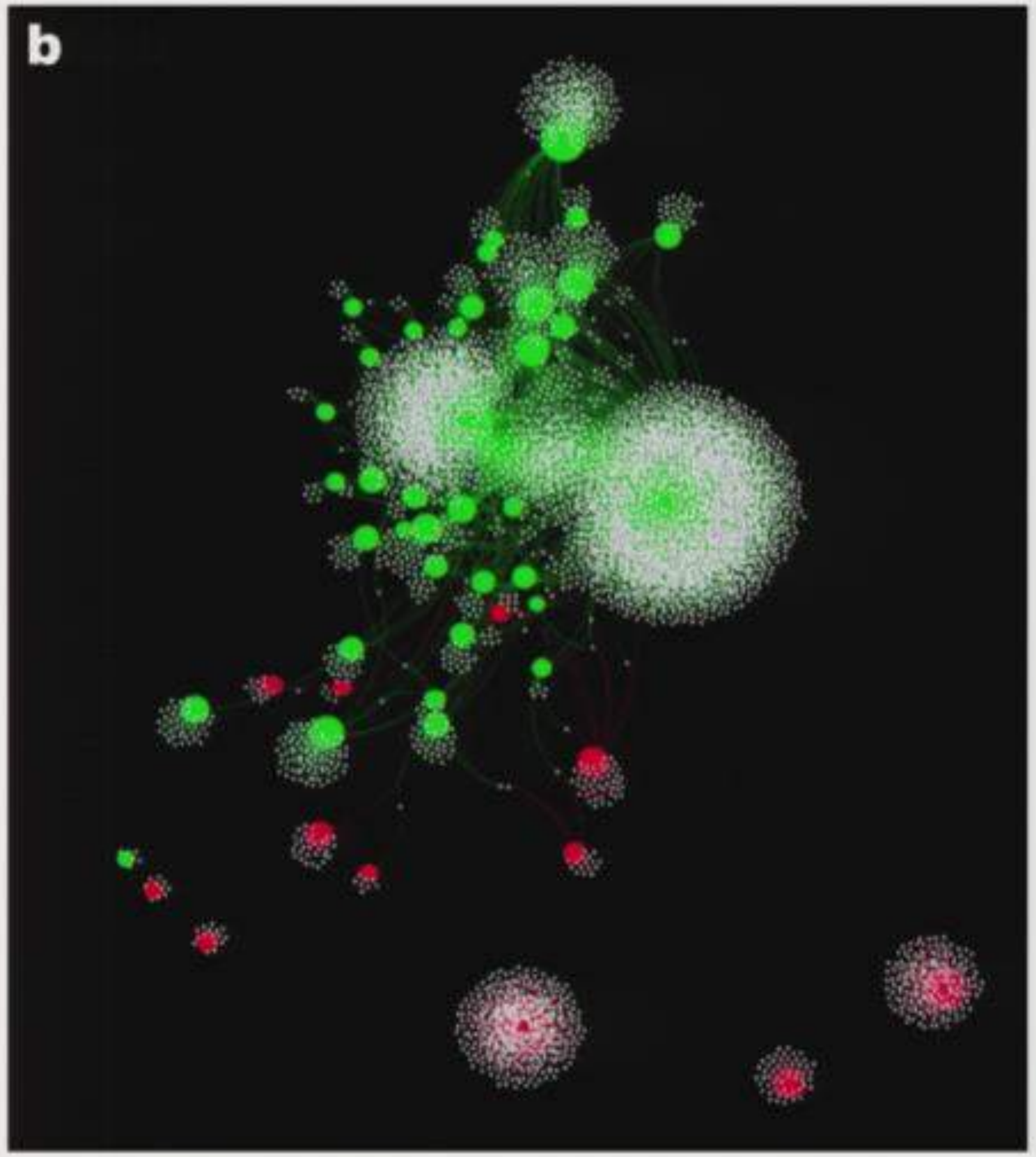
NEWS

# Facebook CEO Mark Zuckerberg calls for internet regulation

Mark Zuckerberg wants better internet regulation for "harmful content, election integrity, privacy and data portability." It comes after Facebook faced criticism over a deadly attack being livestreamed on the platform.

Scale is a huge problem.

# Platforms are federated.



If I stop a hate group on Facebook, do they move to MeWe?

So, typical retributive approaches to online moderation aren’t working well.

# Restorative justice

a system of criminal justice which focuses on the rehabilitation of offenders through reconciliation with victims and the community at large

embedded in feminist theory, African-American and Native studies, and other counter-narratives.

# Most famous example: South African Truth and Reconciliation Commission

TRUTH AND RECONCILIATION COMMISSION
HEALING OUR PAST
FIRST HEARING - EAST LONDON 15 - 18 APRIL 1996

# Goals of restorative justice.

- put key decisions into the hands of those most affected by crime
- make justice more healing and more transformative
- reduce the likelihood of future offenses

Howard Zehr

# Guiding questions of restorative justice

1. Who has been harmed?
2. What are their needs?
3. Whose obligations are these?
4. Who has a stake in this situation?
5. What are the causes?
6. What is the appropriate process to put things right?

# Signposts of restorative justice

- Focus on the harms of wrongdoing rather than the rule that was broken.
- Show equal concern and commitment to those victimized and those who have offended, involving both in the process.
- Work towards the restoration of those harmed, empowering them and responding to their needs.
- Support those who have offended, while encouraging them to understand, accept, and carry out their obligations.
- Recognize that while obligations may be difficult for those who offended, those obligations should not be intended as harms and they must be achievable.
- Provide opportunities for dialogue, direct or indirect, between those who harmed and those harmed, as desired by both parties.
- Find a meaningful way to involved community members and to respond to the community bases of crime.

# Critiques of restorative justice

Places too much burden on the victim.

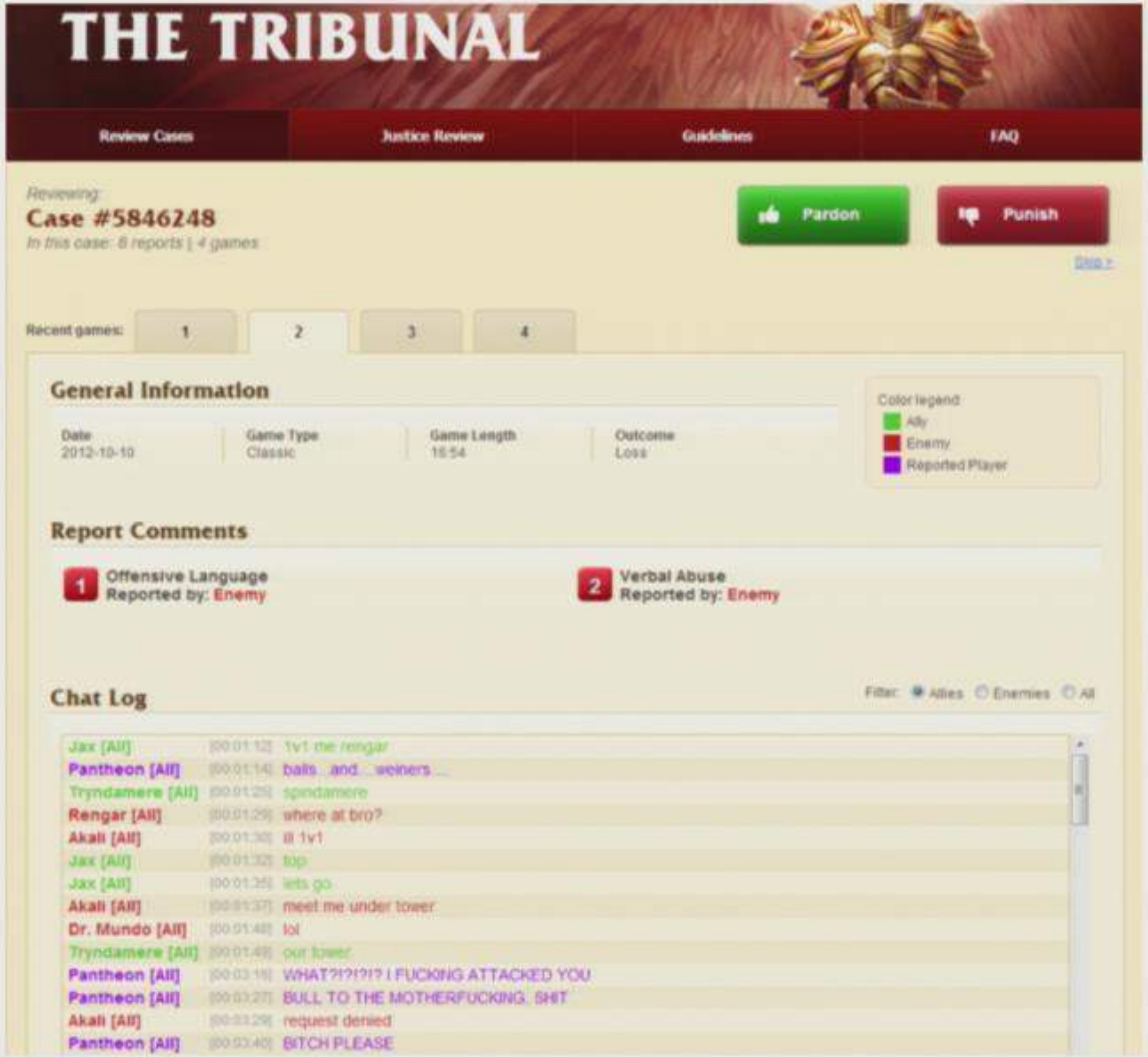
May not fit every crime.

Outcomes are not operationalized well.

Doesn’t scale.

Can we use restorative justice principles to address online adversarial interactions?

# League of Legends Tribunal System



*Reviewing:*

**Case #5846248**

*In this case: 8 reports | 4 games*

Recent games: 1 | 2 | 3 | 4

**General Information**

| Date | Game Type | Game Length | Outcome |
|---|---|---|---|
| 2012-10-10 | Classic | 16:54 | Loss |

Color legend

- 🟩 Ally
- 🟥 Enemy
- 🟪 Reported Player

**Report Comments**

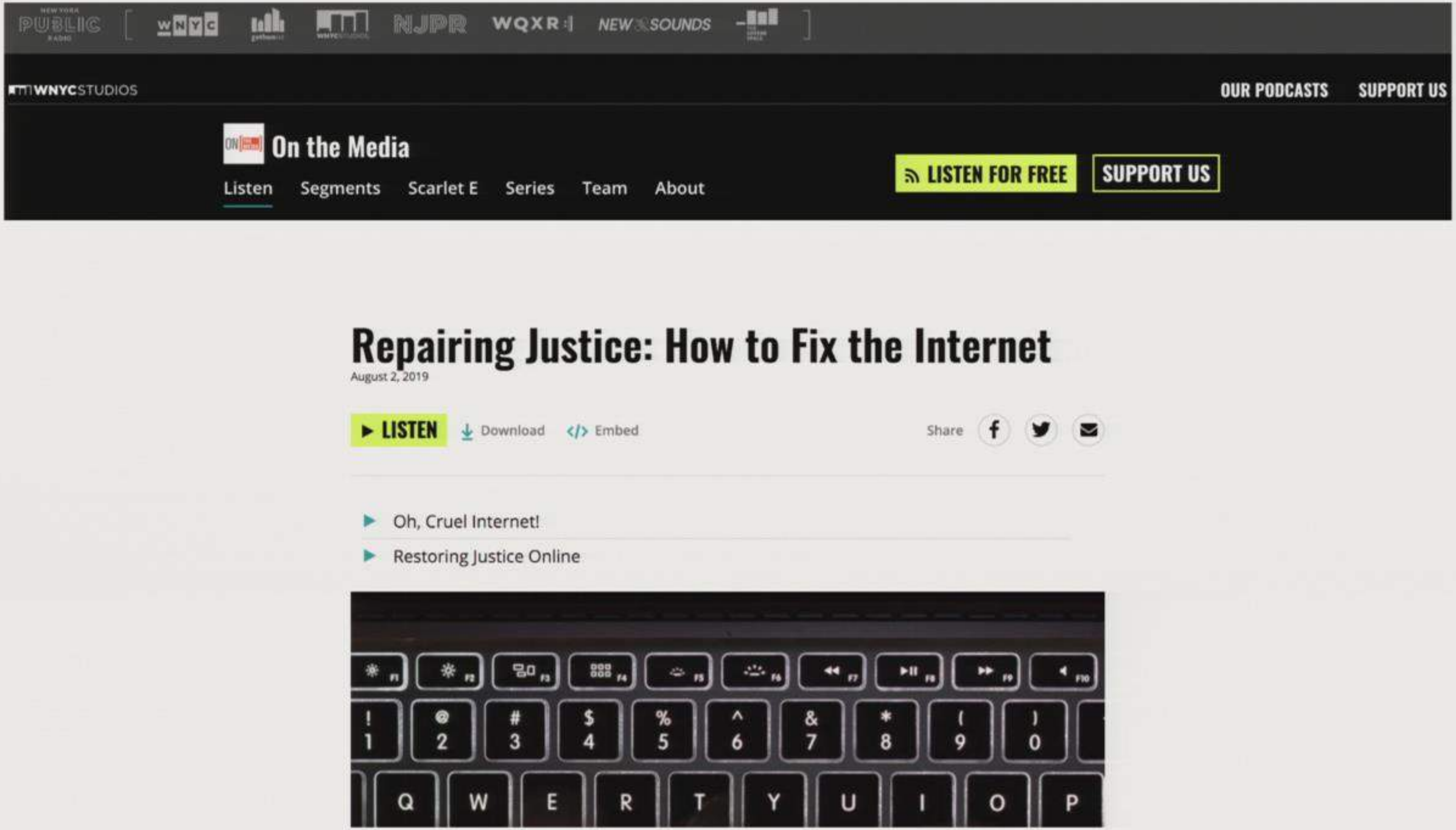
1 Offensive Language
Reported by: Enemy

2 Verbal Abuse
Reported by: Enemy

**Chat Log**

Filter: ● Allies ○ Enemies ○ All

```
Jax [All]          [00:01:12]  1v1 me rengar
Pantheon [All]     [00:01:14]  balls...and...weiners...
Tryndamere [All]   [00:01:25]  spindamere
Rengar [All]       [00:01:29]  where at bro?
Akali [All]        [00:01:30]  ill 1v1
Jax [All]          [00:01:32]  top
Jax [All]          [00:01:35]  lets go
Akali [All]        [00:01:37]  meet me under tower
Dr. Mundo [All]    [00:01:48]  lol
Tryndamere [All]   [00:01:49]  our tower
Pantheon [All]     [00:03:18]  WHAT?!?!?!? I FUCKING ATTACKED YOU
Pantheon [All]     [00:03:27]  BULL TO THE MOTHERFUCKING. SHIT
Akali [All]        [00:03:29]  request denied
Pantheon [All]     [00:03:40]  BITCH PLEASE
```



# Repairing Justice: How to Fix the Internet

August 2, 2019

- ▶ Oh, Cruel Internet!
- ▶ Restoring Justice Online

# Next steps

Representative survey to measure dimensions of justice people expect in online interactions.

Experiments testing different features that could promote justice.

Field testing of justice mechanisms.

Cliff Lampe

cacl@umich.edu
@clifflampe

Can we use restorative justice principles to address online adversarial interactions?