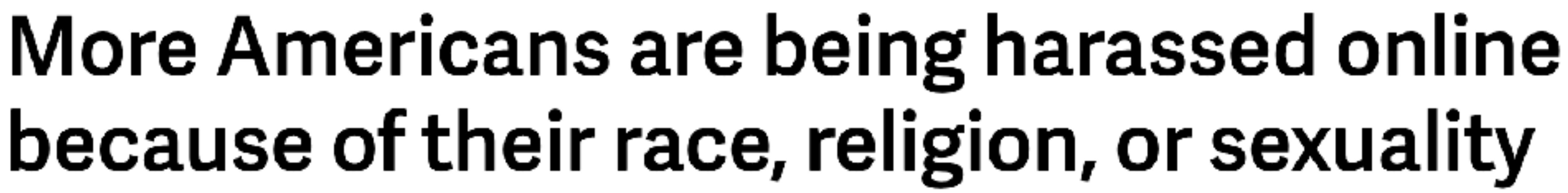
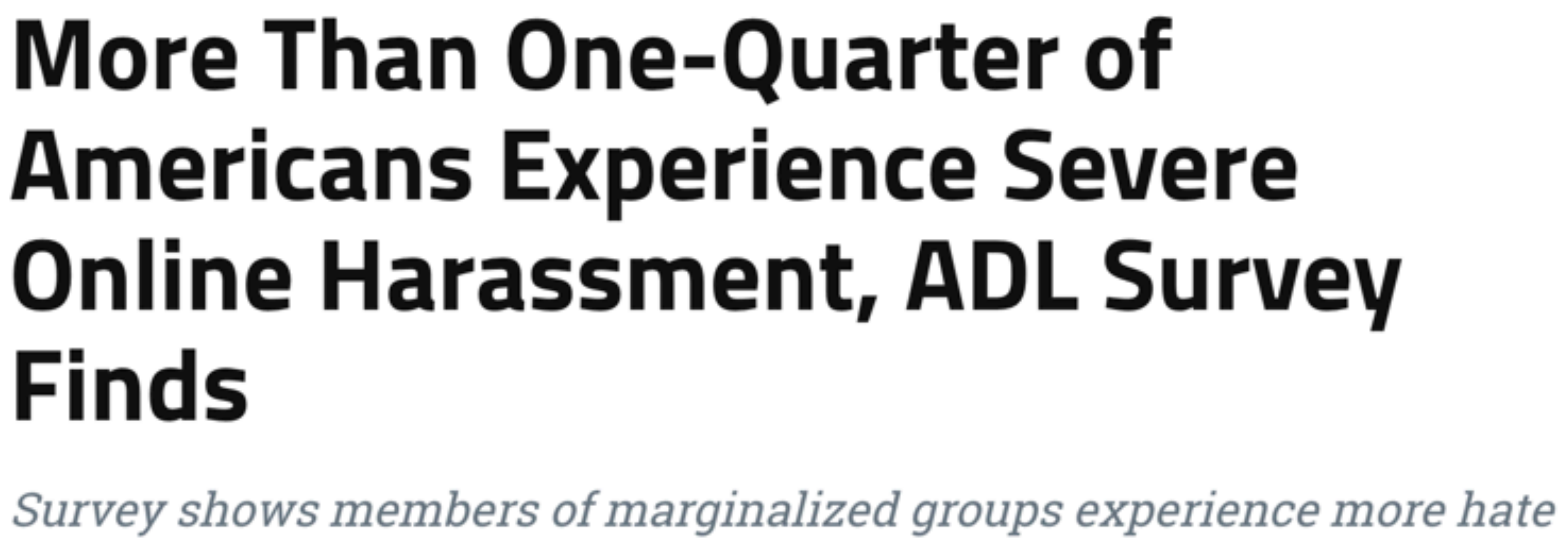
SoK: Hate, Harassment, and the Changing Landscape of Online Abuse

Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, **Deepak Kumar**, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, Gianluca Stringhini

Content warning: Potentially triggering language and difficult subject material ahead.

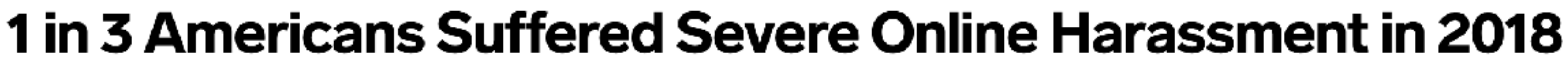


More Americans are being harassed online because of their race, religion, or sexuality

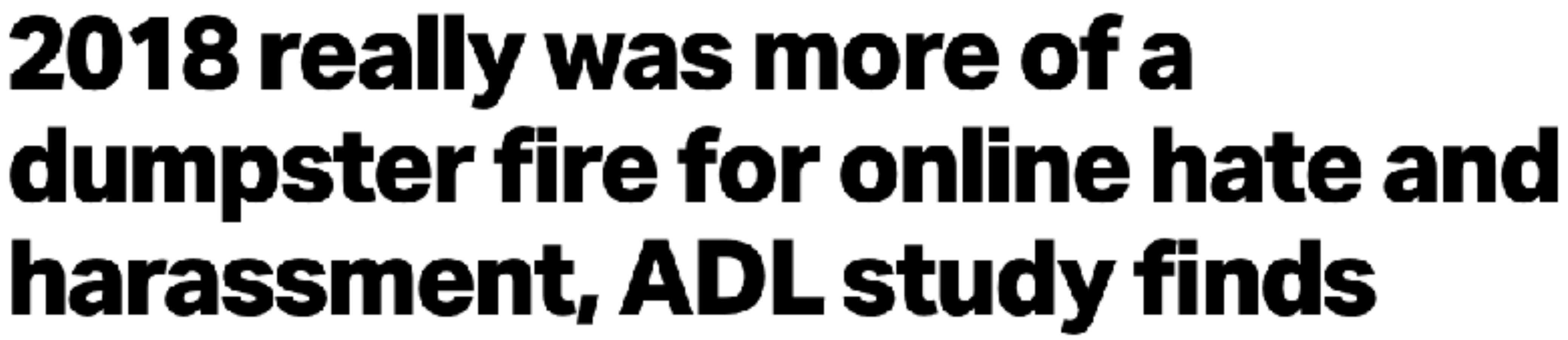


More Than One-Quarter of Americans Experience Severe Online Harassment, ADL Survey Finds

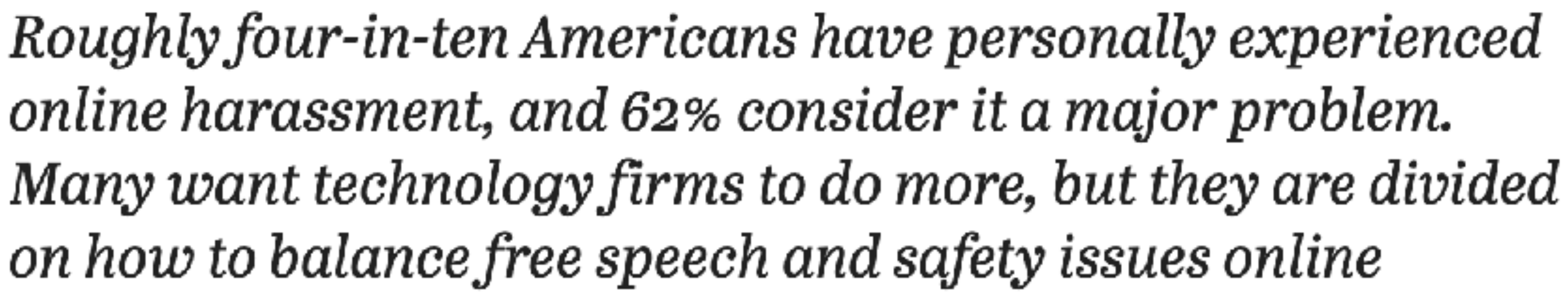
Survey shows members of marginalized groups experience more hate



1 in 3 Americans Suffered Severe Online Harassment in 2018

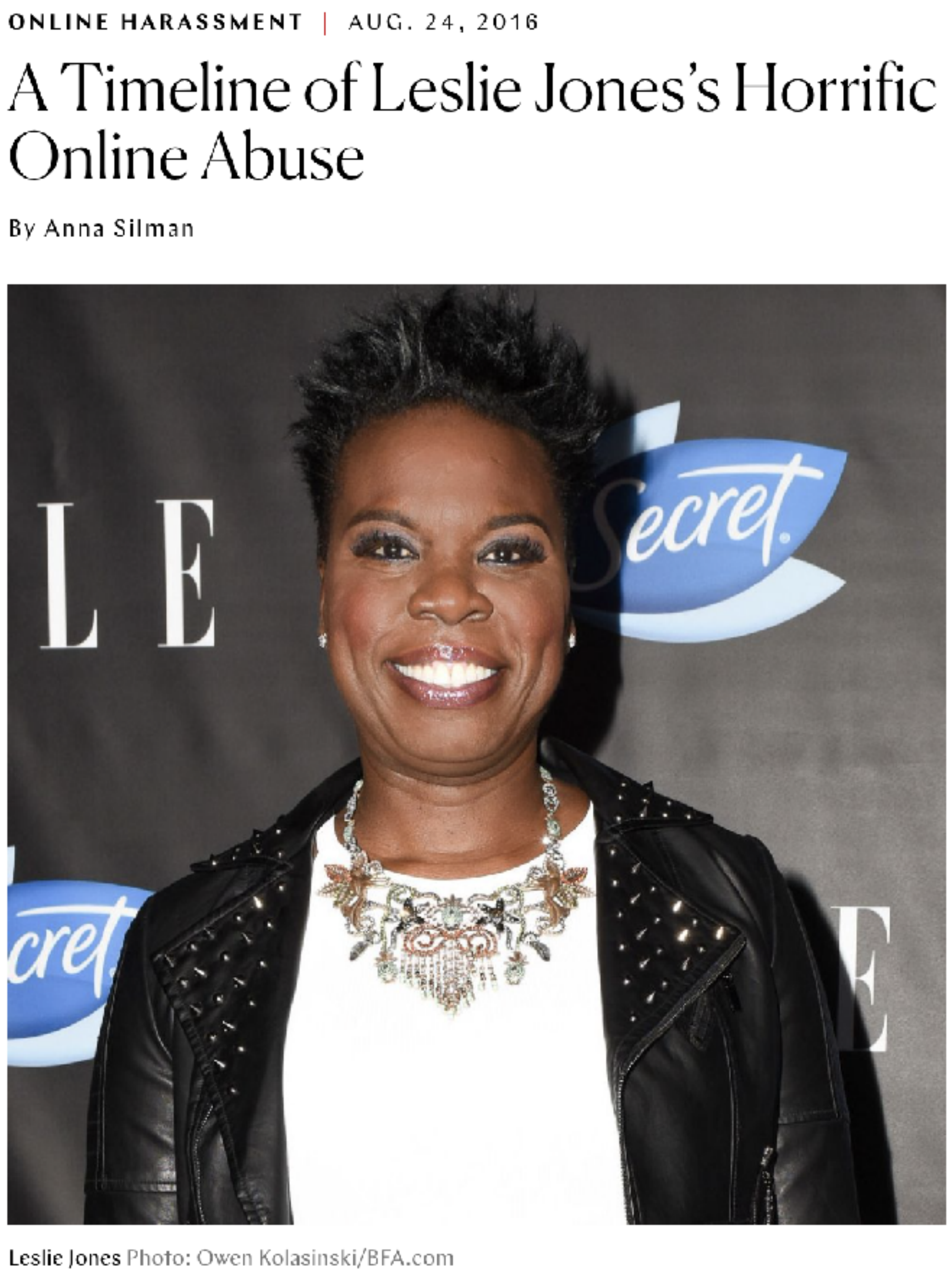


2018 really was more of a dumpster fire for online hate and harassment, ADL study finds



Roughly four-in-ten Americans have personally experienced online harassment, and 62% consider it a major problem. Many want technology firms to do more, but they are divided on how to balance free speech and safety issues online

What does hate and harassment look like?



ONLINE HARASSMENT | AUG. 24, 2016

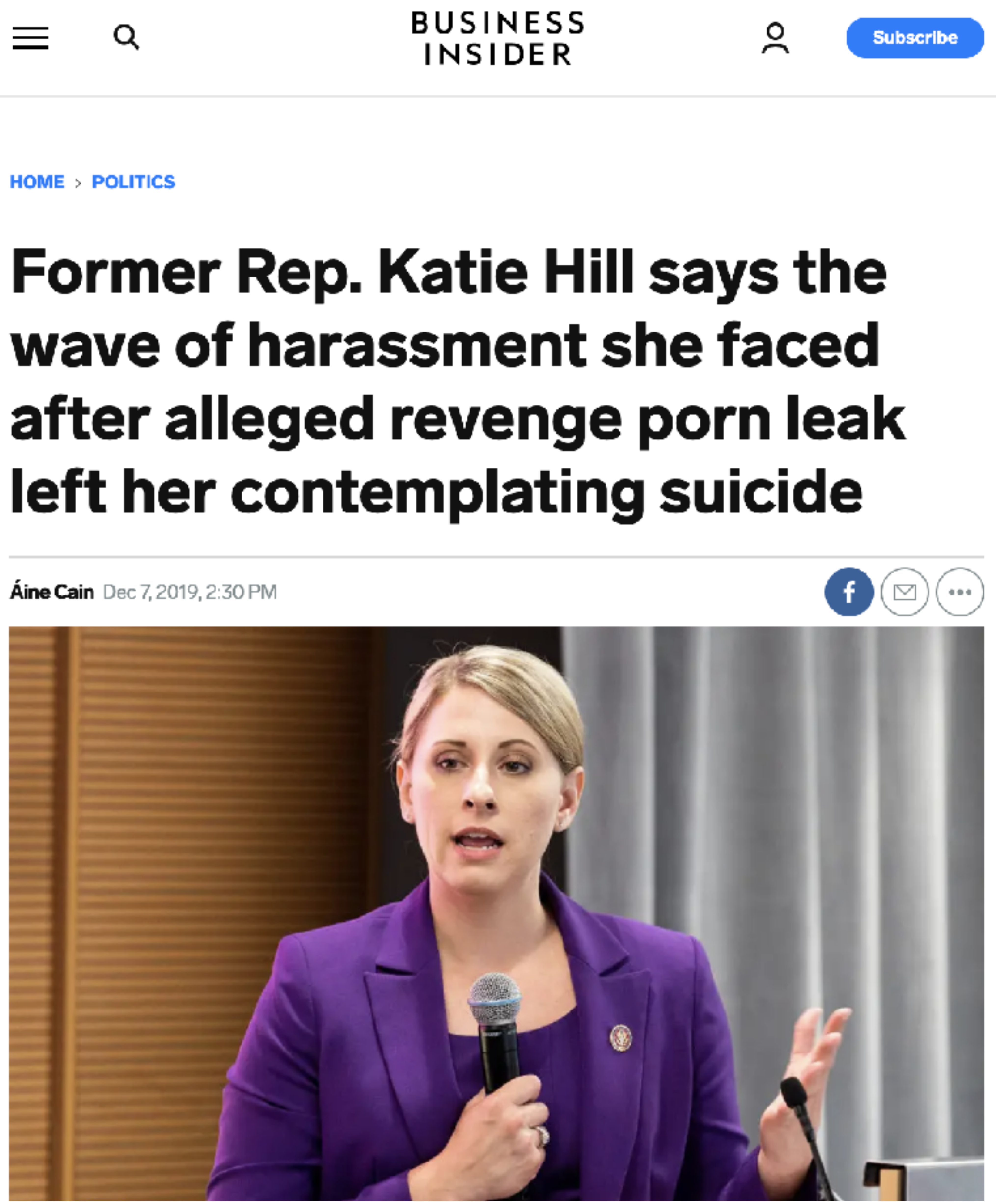
A Timeline of Leslie Jones's Horrific Online Abuse

By Anna Silman

Leslie Jones Photo: Owen Kolasinski/BFA.com

Coordinated campaigns of **toxic comments** on social media that attempt to silence voices.

Non-consensual leaking of **intimate images and other personal information** by former partners or anonymous attackers.

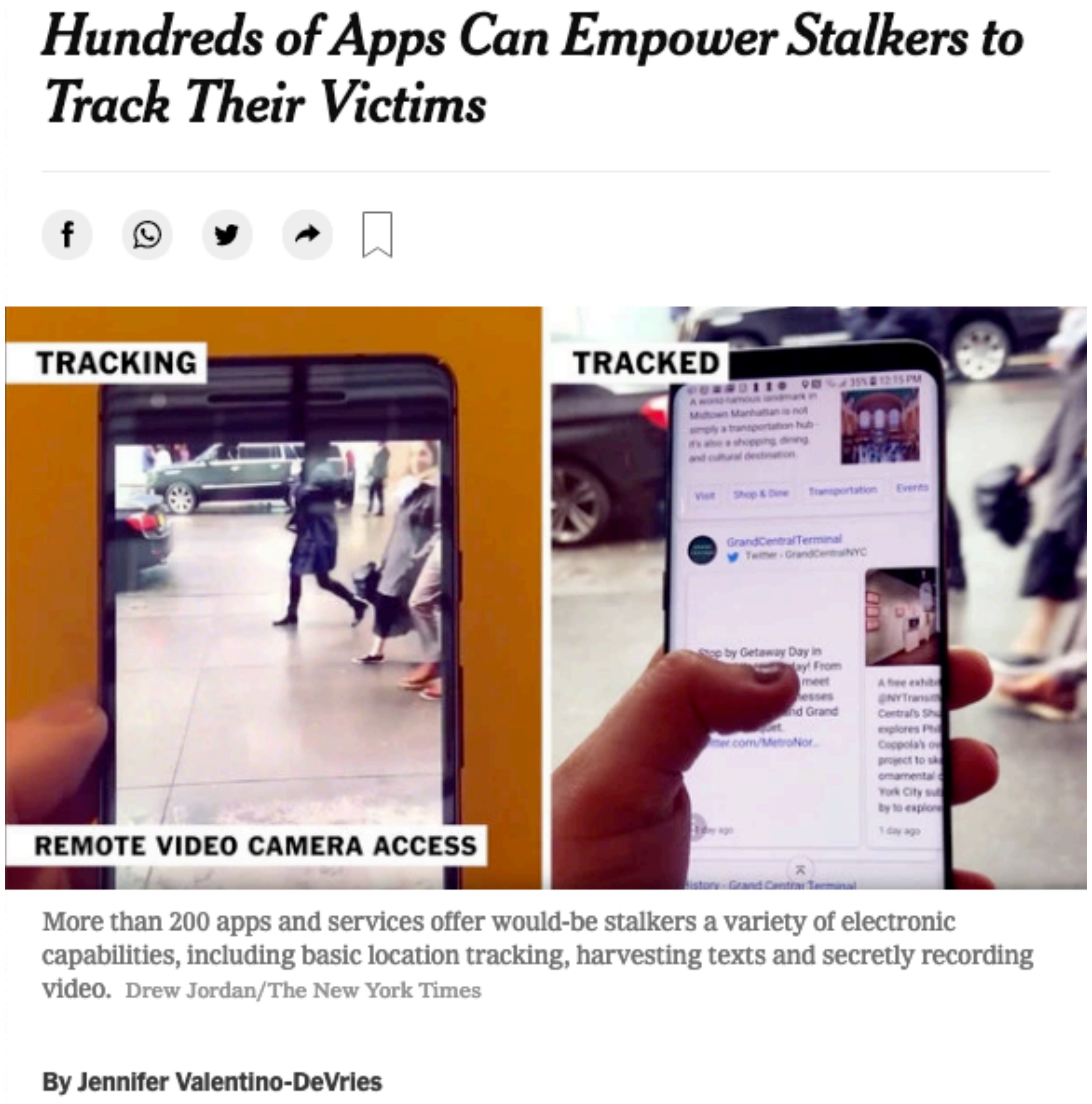


BUSINESS INSIDER

HOME › POLITICS

Former Rep. Katie Hill says the wave of harassment she faced after alleged revenge porn leak left her contemplating suicide

Áine Cain Dec 7, 2019, 2:30 PM



Hundreds of Apps Can Empower Stalkers to Track Their Victims

More than 200 apps and services offer would-be stalkers a variety of electronic capabilities, including basic location tracking, harvesting texts and secretly recording video. Drew Jordan/The New York Times

By Jennifer Valentino-DeVries

Spyware and tracking can aid in **surveilling** intimate partners through their devices and accounts.

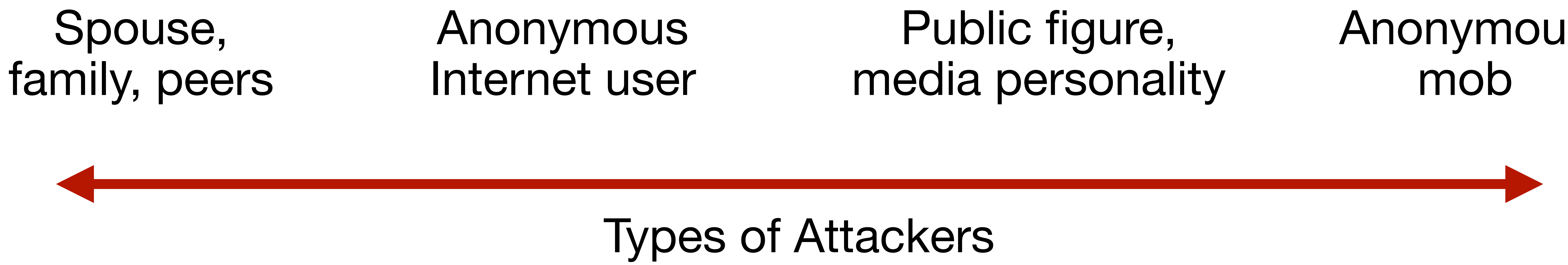
Intent is to **inflict emotional harm,** includes coercive control or instilling a fear of sexual or physical violence.

We should address online hate and harassment as a *security* problem.

Threat Model: Targets and Attackers

Targets of harassment can be individuals or at-risk groups (e.g., LGBTQ+ people)

An attacker’s main goal is to emotionally harm or coercively control the target.



Spouse, family, peers

Anonymous Internet user

Public figure, media personality

Anonymous mob

Types of Attackers

Literature Review

- Examined the last five years of research and journalism on online hate and harassment
  - IEEE S&P, USENIX Security, CCS, CHI, CSCW, ICWSM, WWW, SOUPS, and IMC
    - Used related papers as a “seed set”, manually searched through related works, and expanded search to include findings from social sciences
  - Also included major news events (e.g., Gamergate) and related attacks and news coverage
  - Reviewed over **150 news articles and research papers** in online hate and harassment

Differentiating Attacks

Research team synthesized criteria that differentiate attacks, falling into **three broad categories – Audience, Medium, Capabilities**

Category	Criteria
Audience	Intended to be seen by the target?
Audience	Intended to be seen by an audience?
Medium	Does attack use media, such as text or images?
Capabilities	Require deception of the audience?
Capabilities	Deception of a third-party authority?
Capabilities	Amplification?
Capabilities	Privileged access to information?

Differentiating Attacks – Audience

Category	Criteria	Examples
Audience	**Intended to be seen by the target?**	**Bullying, Trolling**
Audience	**Intended to be seen by an audience?**	**Doxxing**
Medium	Does attack use media, such as text or images?	Hate Speech
Capabilities	Require deception of the audience?	Impersonated profiles, Deepfakes
Capabilities	Deception of a third-party authority?	SWATing
Capabilities	Amplification?	Raiding, Dogpiling
Capabilities	Privileged access to information?	IPS, GPS monitoring

Differentiating Attacks – Medium

Category	Criteria	Examples
Audience	Intended to be seen by the target?	Bullying, Trolling
Audience	Intended to be seen by an audience?	Doxxing
Medium	**Does attack use media, such as text or images?**	**Hate Speech**
Capabilities	Require deception of the audience?	Impersonated profiles, Deepfakes
Capabilities	Deception of a third-party authority?	SWATing
Capabilities	Amplification?	Raiding, Dogpiling
Capabilities	Privileged access to information?	IPS, GPS monitoring

Differentiating Attacks – Capabilities

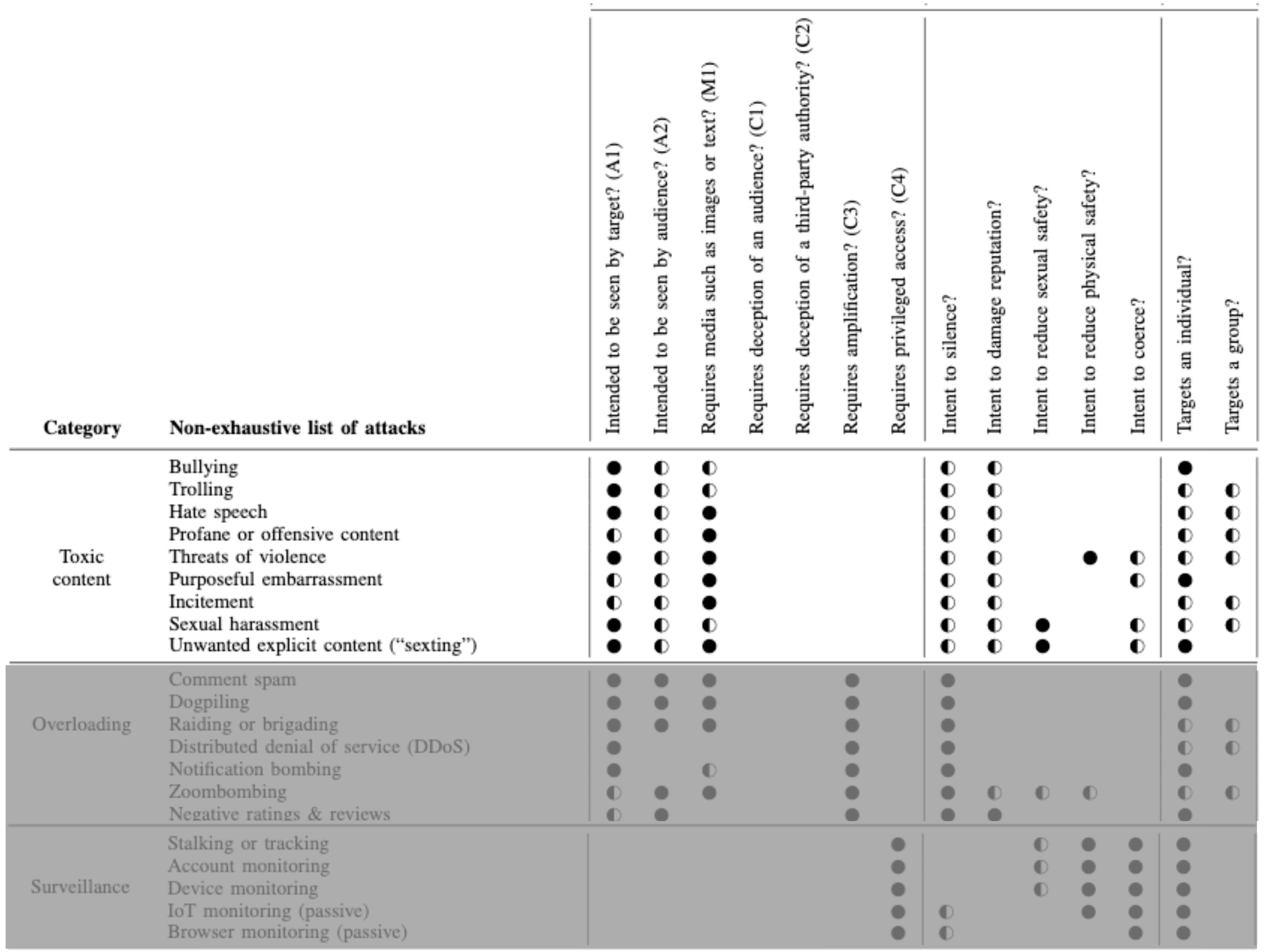
Category	Criteria	Examples
Audience	Intended to be seen by the target?	Bullying, Trolling
Audience	Intended to be seen by an audience?	Doxxing
Medium	Does attack use media, such as text or images?	Hate Speech
Capabilities	**Require deception of the audience?**	**Impersonated profiles, Deepfakes**
Capabilities	**Deception of a third-party authority?**	**SWATing**
Capabilities	**Amplification?**	**Raiding, Dogpiling**
Capabilities	**Privileged access to information?**	**IPS, GPS monitoring**

Seven Classes of Hate and Harassment

Category
Toxic Content
Content Leakage
Overloading
Fake Reporting
Impersonation
Surveillance
Lockout and Control

Seven Classes of Hate and Harassment

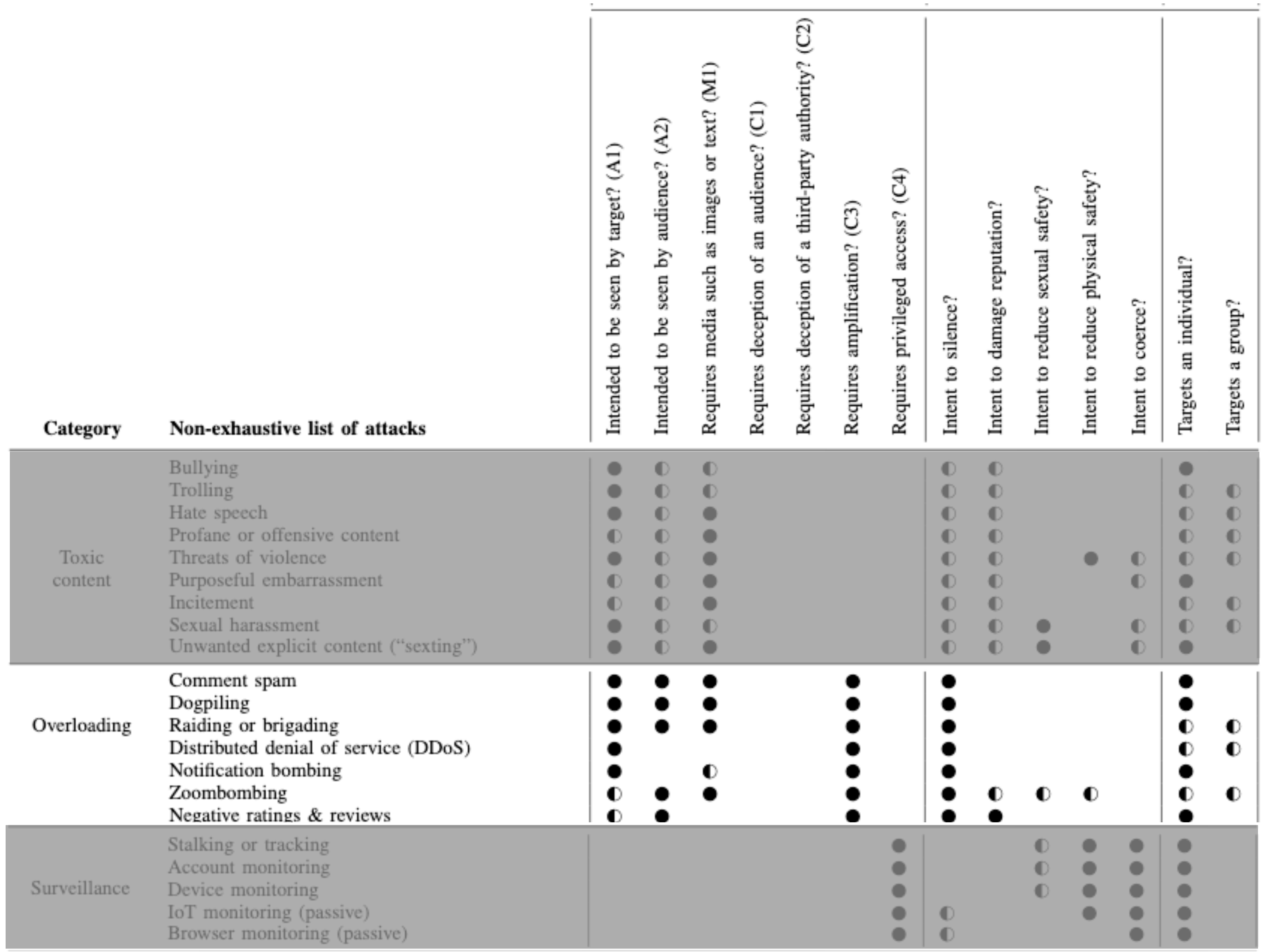
Category
Toxic Content
Content Leakage
Overloading
Fake Reporting
Impersonation
Surveillance
Lockout and Control



Category	Non-exhaustive list of attacks	Intended to be seen by target? (A1)	Intended to be seen by audience? (A2)	Requires media such as images or text? (M1)	Requires deception of an audience? (C1)	Requires deception of a third-party authority? (C2)	Requires amplification? (C3)	Requires privileged access? (C4)	Intent to silence?	Intent to damage reputation?	Intent to reduce sexual safety?	Intent to reduce physical safety?	Intent to coerce?	Targets an individual?	Targets a group?
Toxic content	Bullying	●	◐	◐					◐	◐				●	
	Trolling	●	◐	◐					◐	◐				◐	◐
	Hate speech	●	◐	●					◐	◐				◐	◐
	Profane or offensive content	◐	◐	●					◐	◐				◐	◐
	Threats of violence	●	◐	●					◐	◐		●	◐	◐	◐
	Purposeful embarrassment	◐	◐	●					◐	◐			◐	●	
	Incitement	◐	◐	●					◐	◐				◐	◐
	Sexual harassment	●	◐	◐					◐	◐	●		◐	◐	◐
	Unwanted explicit content (“sexting”)	●	◐	●					◐	◐	●		◐	●	
Overloading	Comment spam	●	●	●			●		●					●	
	Dogpiling	●	●	●			●		●					●	
	Raiding or brigading	●	●	●			●		●					◐	◐
	Distributed denial of service (DDoS)	●					●		●					◐	◐
	Notification bombing	●		◐			●		●					●	
	Zoombombing	◐	●	●			●		●	◐	◐	◐		◐	◐
	Negative ratings & reviews	◐	●				●		●	●				●	
Surveillance	Stalking or tracking							●			◐	●	●	●	
	Account monitoring							●			◐	●	●	●	
	Device monitoring							●			◐	●	●	●	
	IoT monitoring (passive)							●	◐			●	●	●	
	Browser monitoring (passive)							●	◐				●	●	

Seven Classes of Hate and Harassment

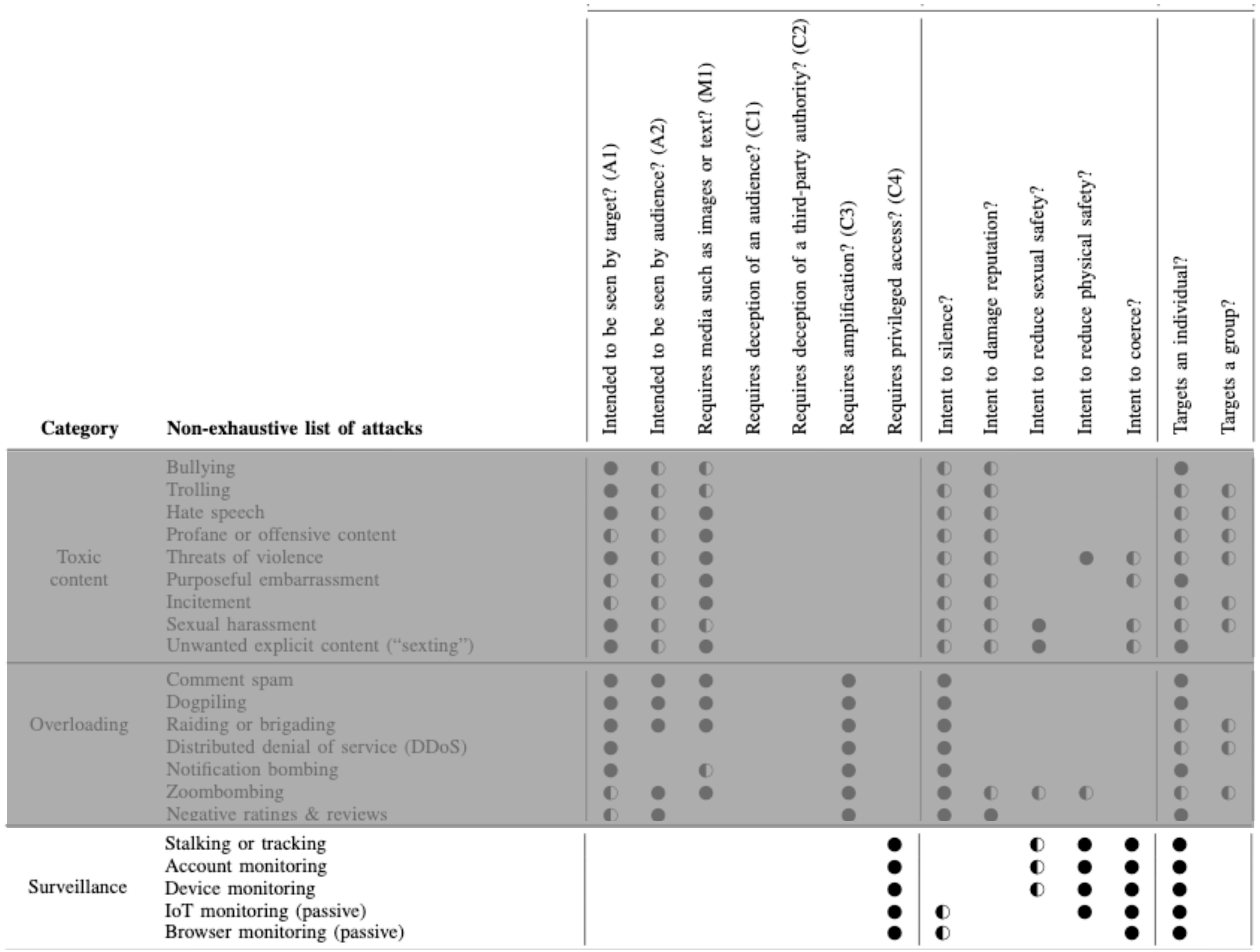
Category
Toxic Content
Content Leakage
Overloading
Fake Reporting
Impersonation
Surveillance
Lockout and Control



Category	Non-exhaustive list of attacks	Intended to be seen by target? (A1)	Intended to be seen by audience? (A2)	Requires media such as images or text? (M1)	Requires deception of an audience? (C1)	Requires deception of a third-party authority? (C2)	Requires amplification? (C3)	Requires privileged access? (C4)	Intent to silence?	Intent to damage reputation?	Intent to reduce sexual safety?	Intent to reduce physical safety?	Intent to coerce?	Targets an individual?	Targets a group?
Toxic content	Bullying	●	◐	◐					◐	◐				●	
	Trolling	●	◐	◐					◐	◐				◐	◐
	Hate speech	●	◐	●					◐	◐				◐	◐
	Profane or offensive content	◐	◐	●					◐	◐				◐	◐
	Threats of violence	●	◐	●					◐	◐		●	◐	◐	◐
	Purposeful embarrassment	◐	◐	●					◐	◐			◐	●	
	Incitement	◐	◐	●					◐	◐				◐	◐
	Sexual harassment	●	◐	◐					◐	◐	●		◐	◐	◐
	Unwanted explicit content ("sexting")	●	◐	●					◐	◐	●		◐	●	
Overloading	Comment spam	●	●	●			●		●					●	
	Dogpiling	●	●	●			●		●					●	
	Raiding or brigading	●	●	●			●		●					◐	◐
	Distributed denial of service (DDoS)	●					●		●					◐	◐
	Notification bombing	●		◐			●		●					●	
	Zoombombing	◐	●	●			●		●	◐	◐	◐		◐	◐
	Negative ratings & reviews	◐	●				●		●	●				●	
Surveillance	Stalking or tracking							●			◐	●	●	●	
	Account monitoring							●			◐	●	●	●	
	Device monitoring							●			◐	●	●	●	
	IoT monitoring (passive)							●	◐			●	●	●	
	Browser monitoring (passive)							●	◐				●	●	

Seven Classes of Hate and Harassment

Category
Toxic Content
Content Leakage
Overloading
Fake Reporting
Impersonation
Surveillance
Lockout and Control

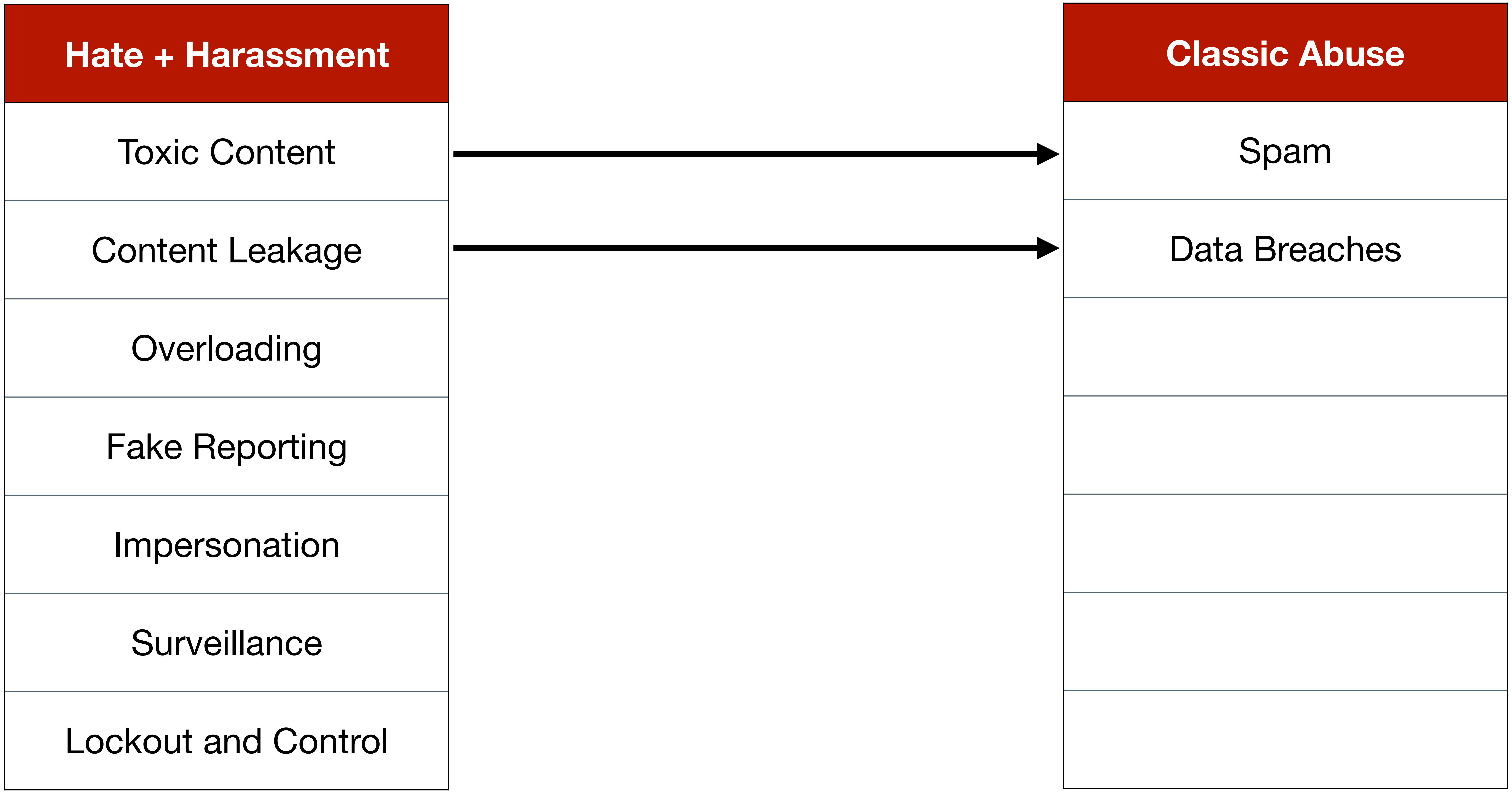


Category	Non-exhaustive list of attacks	Intended to be seen by target? (A1)	Intended to be seen by audience? (A2)	Requires media such as images or text? (M1)	Requires deception of an audience? (C1)	Requires deception of a third-party authority? (C2)	Requires amplification? (C3)	Requires privileged access? (C4)	Intent to silence?	Intent to damage reputation?	Intent to reduce sexual safety?	Intent to reduce physical safety?	Intent to coerce?	Targets an individual?	Targets a group?
Toxic content	Bullying	●	◐	◐					◐	◐				●	
	Trolling	●	◐	◐					◐	◐				◐	◐
	Hate speech	●	◐	●					◐	◐				◐	◐
	Profane or offensive content	◐	◐	●					◐	◐				◐	◐
	Threats of violence	●	◐	●					◐	◐		●	◐	◐	◐
	Purposeful embarrassment	◐	◐	●					◐	◐			◐	●	
	Incitement	◐	◐	●					◐	◐				◐	◐
	Sexual harassment	●	◐	◐					◐	◐	●		◐	◐	◐
	Unwanted explicit content (“sexting”)	●	◐	●					◐	◐	●		◐	●	
Overloading	Comment spam	●	●	●			●		●					●	
	Dogpiling	●	●	●			●		●					●	
	Raiding or brigading	●	●	●			●		●					◐	◐
	Distributed denial of service (DDoS)	●					●		●					◐	◐
	Notification bombing	●		◐			●		●					●	
	Zoombombing	◐	●	●			●		●	◐	◐	◐		◐	◐
	Negative ratings & reviews	◐	●				●		●	●				●	
Surveillance	Stalking or tracking							●			◐	●	●	●	
	Account monitoring							●			◐	●	●	●	
	Device monitoring							●			◐	●	●	●	
	IoT monitoring (passive)							●	◐			●	●	●	
	Browser monitoring (passive)							●	◐				●	●	

Parallels to Security Attacks

Hate + Harassment		Classic Abuse
Toxic Content	→	Spam
Content Leakage		
Overloading		
Fake Reporting		
Impersonation		
Surveillance		
Lockout and Control		

Parallels to Security Attacks



Hate + Harassment	Classic Abuse
Toxic Content	Spam
Content Leakage	Data Breaches
Overloading	
Fake Reporting	
Impersonation	
Surveillance	
Lockout and Control	

Parallels to Security Attacks

Hate + Harassment	Classic Abuse
Toxic Content	Spam
Content Leakage	Data Breaches
Overloading	DoS, DDoS
Fake Reporting	Mark not-spam
Impersonation	Phishing
Surveillance	RAT, Tracking
Lockout and Control	Ransomware

Hate and harassment attacks span many different kinds of attackers, targets, and methods.

Hate and harassment impacts diverse users in different ways.

Survey Instrument

- Surveyed ~1000 participants from 22 countries around the world for three years and asked about hate and harassment experiences
  - Survey was translated for countries that do not primarily speak English
  - Some countries do not appear for all three years to maximize unique countries
- Asked participants “Have you ever personally experienced any of the following online?”
  - Asked about hate and harassment experiences documented in prior work
  - Collected demographic data (e.g., gender, LGBTQ+ status, age, social media usage)

Measuring hate and harassment outcomes

- Modeled experiencing any form of hate and harassment as a binomial distribution
  - Input variables are categorical demographic data

Demographic	Treatment	Reference	Odds
LGBTQ+	LGBTQ+	non-LGBTQ+	1.9x
Social Media Usage	Daily	Never	2.5x
	Weekly	Never	2.3x
Age	18 – 24	65 and up	4.0x
	25 – 34	65 and up	3.4x
Year	**2017**	**2016**	**1.2x**
	2018	**2016**	**1.3x**

Measuring hate and harassment outcomes

- Modeled experiencing any form of hate and harassment as a binomial distribution
  - Input variables are categorical demographic data
- Odds of experiencing online hate and harassment has increased over time

Demographic	Treatment	Reference	Odds
LGBTQ+	LGBTQ+	non-LGBTQ+	1.9x
Social Media Usage	Daily	Never	2.5x
	Weekly	Never	2.3x
Age	18 – 24	65 and up	4.0x
	25 – 34	65 and up	3.4x
Year	**2017**	**2016**	**1.2x**
	2018	**2016**	**1.3x**

Measuring hate and harassment outcomes

- Modeled experiencing any form of hate and harassment as a binomial distribution
  - Input variables are categorical demographic data
- Odds of experiencing online hate and harassment has increased over time
- Participants from *minority* groups experience more online hate and harassment

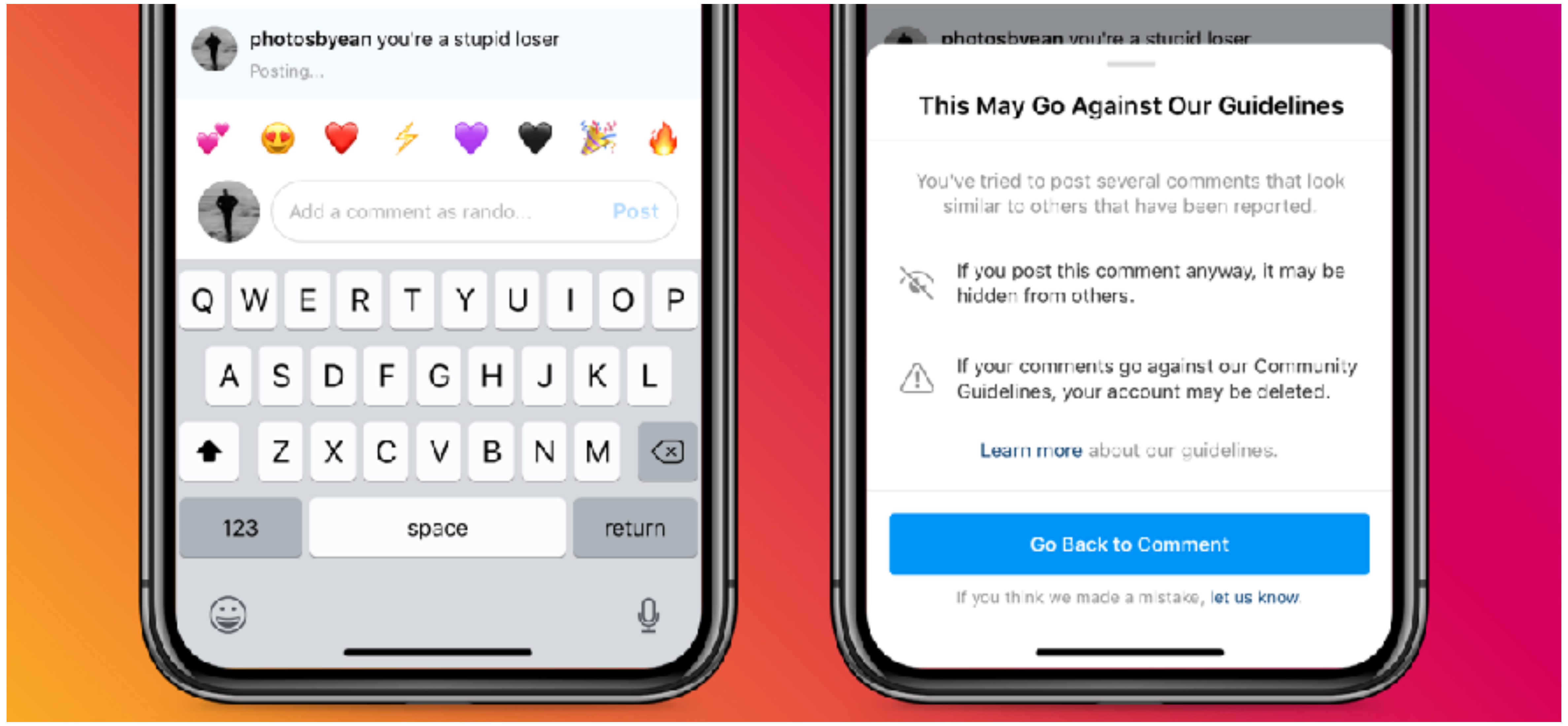
Demographic	Treatment	Reference	Odds
LGBTQ+	**LGBTQ+**	**non-LGBTQ+**	**1.9x**
Social Media Usage	Daily	Never	2.5x
	Weekly	Never	2.3x
Age	**18 – 24**	**65 and up**	**4.0x**
	25 – 34	65 and up	3.4x
Year	2017	2016	1.2x
	2018	2016	1.3x

Designing hate and harassment defenses **must** take into account diverse online experiences.

What can we do about it?

Towards Solutions and Interventions

- **Nudges, indicators, warnings**
- Human moderation, review, and delisting
- Automated detection
- Conscious design
- Policies, education, awareness



photosbyean you're a stupid loser

Posting...

Add a comment as rando...

This May Go Against Our Guidelines

You've tried to post several comments that look similar to others that have been reported.

If you post this comment anyway, it may be hidden from others.

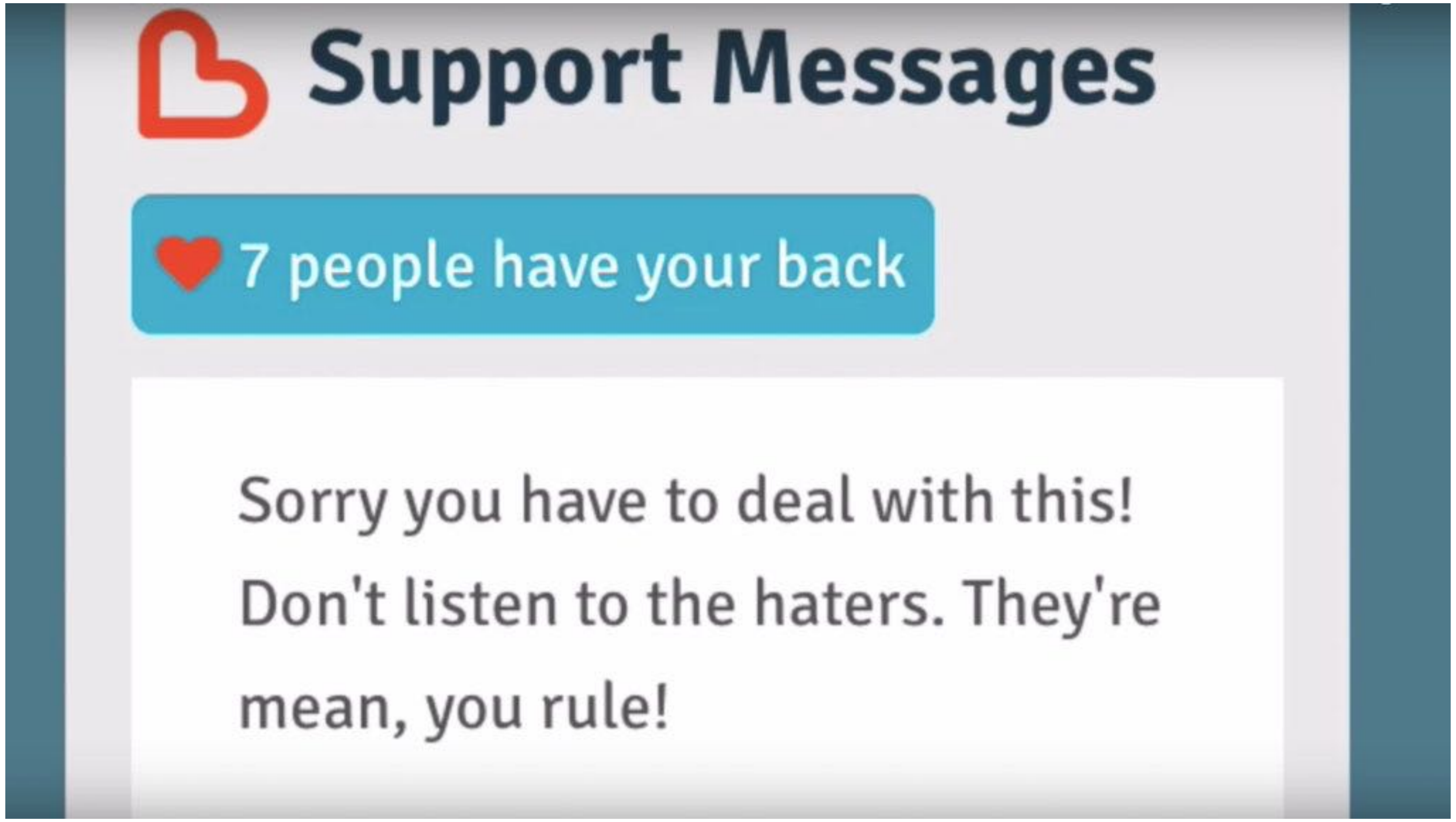
If your comments go against our Community Guidelines, your account may be deleted.

Learn more about our guidelines.

If you think we made a mistake, let us know.

Towards Solutions and Interventions

- Nudges, indicators, warnings
- **Human moderation, review, and delisting**
- Automated detection
- Conscious design
- Policies, education, awareness



Support Messages

7 people have your back

Sorry you have to deal with this! Don't listen to the haters. They're mean, you rule!

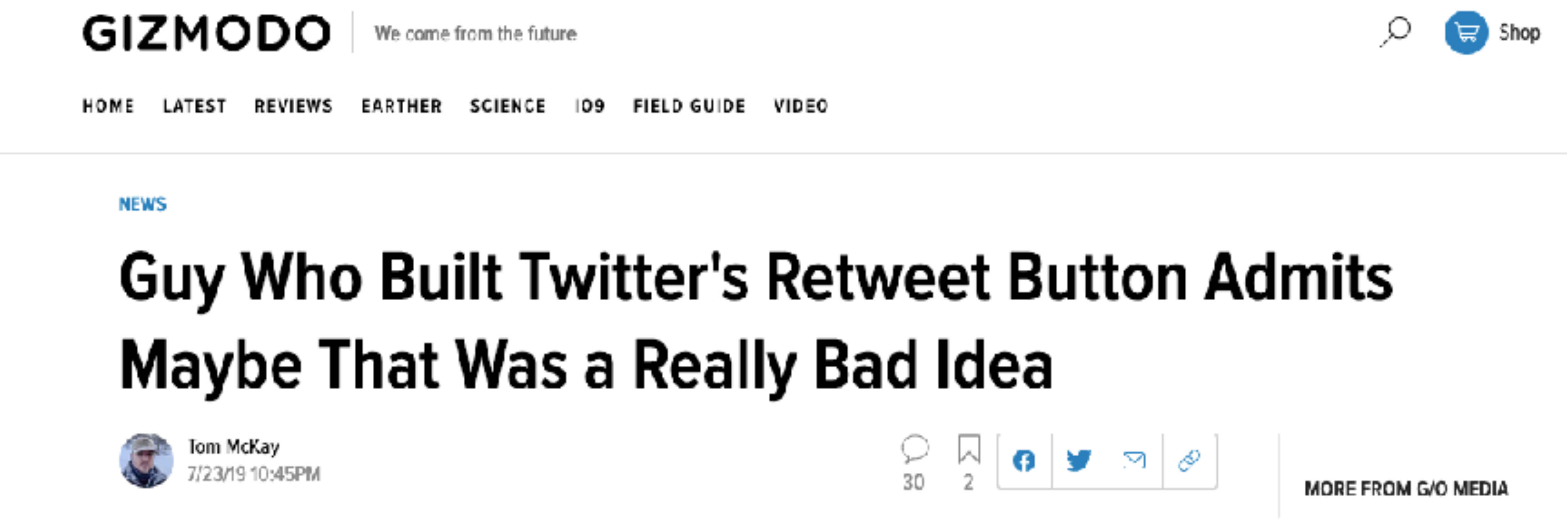
Towards Solutions and Interventions

- Nudges, indicators, warnings
- Human moderation, review, and delisting
- **Automated detection**
- Conscious design
- Policies, education, awareness

Perspective

Towards Solutions and Interventions

- Nudges, indicators, warnings
- Human moderation, review, and delisting
- Automated detection
- **Conscious design**
- Policies, education, awareness



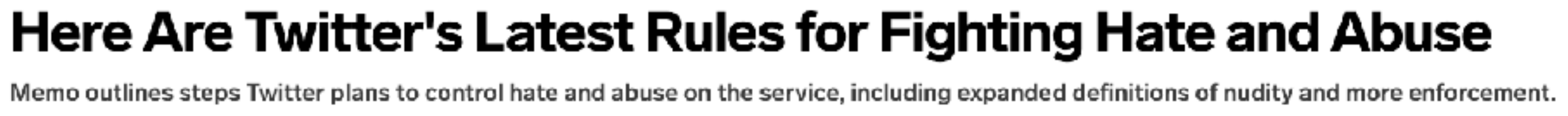
NEWS

Guy Who Built Twitter's Retweet Button Admits Maybe That Was a Really Bad Idea

Tom McKay
7/23/19 10:45PM

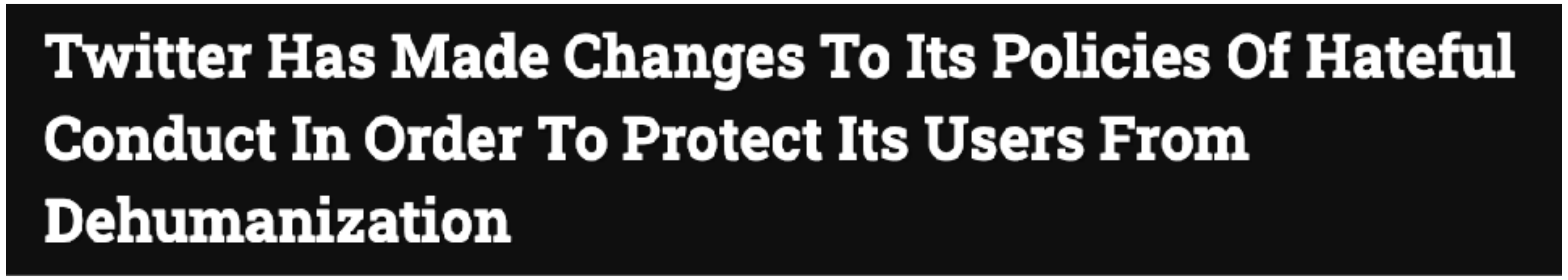
Towards Solutions and Interventions

- Nudges, indicators, warnings
- Human moderation, review, and delisting
- Automated detection
- Conscious design
- **Policies, education, awareness**

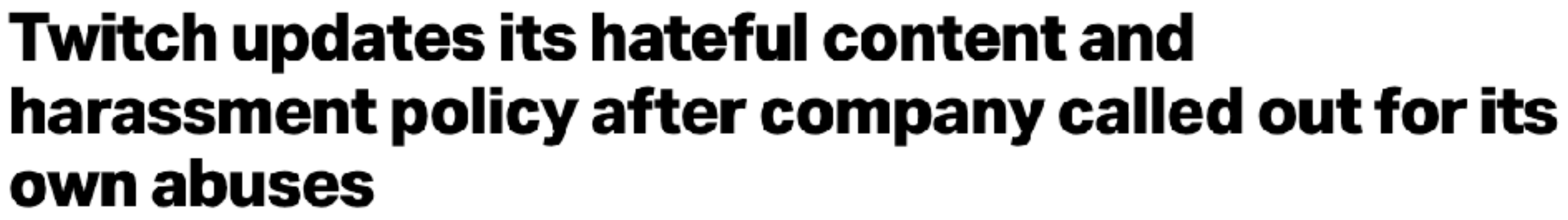


Here Are Twitter's Latest Rules for Fighting Hate and Abuse

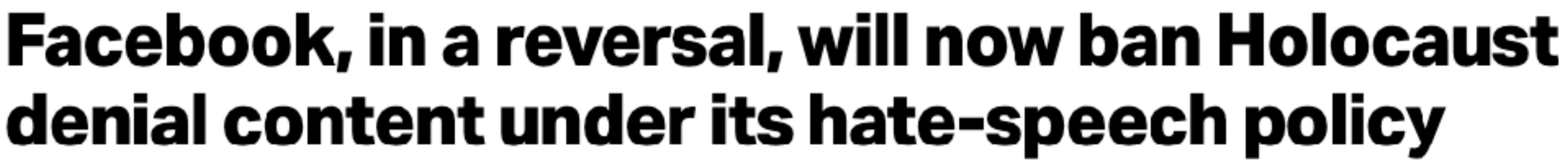
Memo outlines steps Twitter plans to control hate and abuse on the service, including expanded definitions of nudity and more enforcement.



Twitter Has Made Changes To Its Policies Of Hateful Conduct In Order To Protect Its Users From Dehumanization



Twitch updates its hateful content and harassment policy after company called out for its own abuses



Facebook, in a reversal, will now ban Holocaust denial content under its hate-speech policy

Tensions, Challenges, Outstanding Questions

- How do we empower targets of abuse instead of burdening them with choice?
- How do we balance user privacy with accountability?
- How do we define success in abuse research?

Tensions, Challenges, Outstanding Questions

- How do we empower targets of abuse instead of burdening them with choice?
- How do we balance user privacy with accountability?
- How do we define success in abuse research?

Deepak Kumar

kumarde@cs.stanford.edu

@_kumarde