Envisioning Online Hate and Harassment as a Security Problem

Deepak Kumar

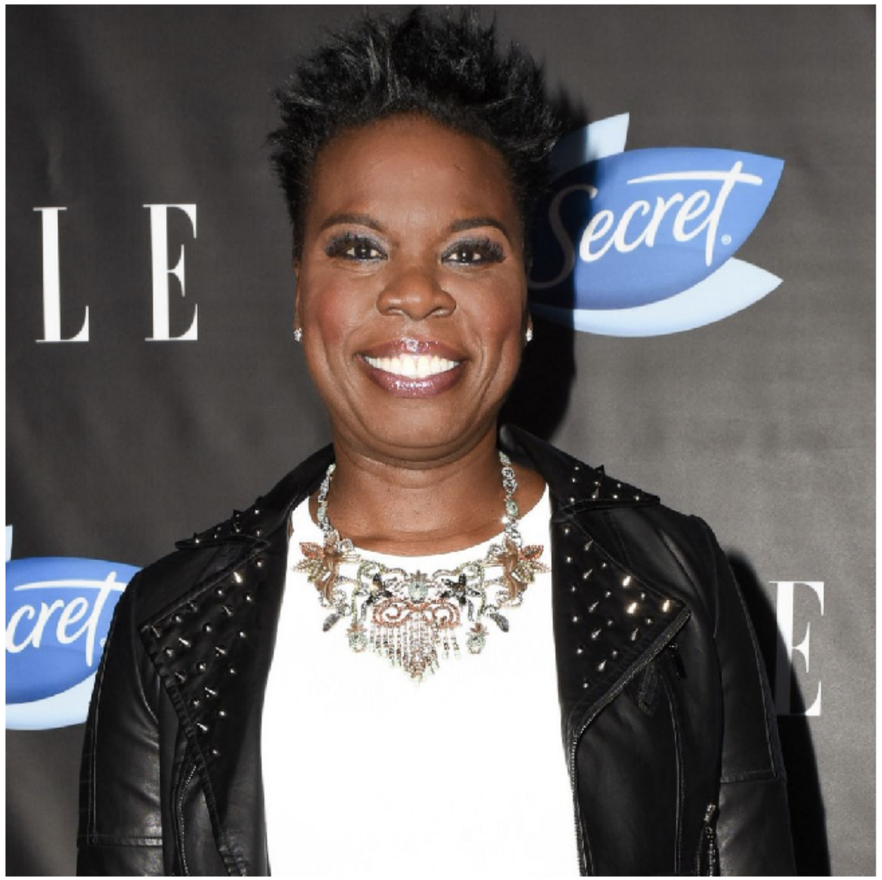
Content warning: Potentially triggering language and difficult subject material ahead.

What does online 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

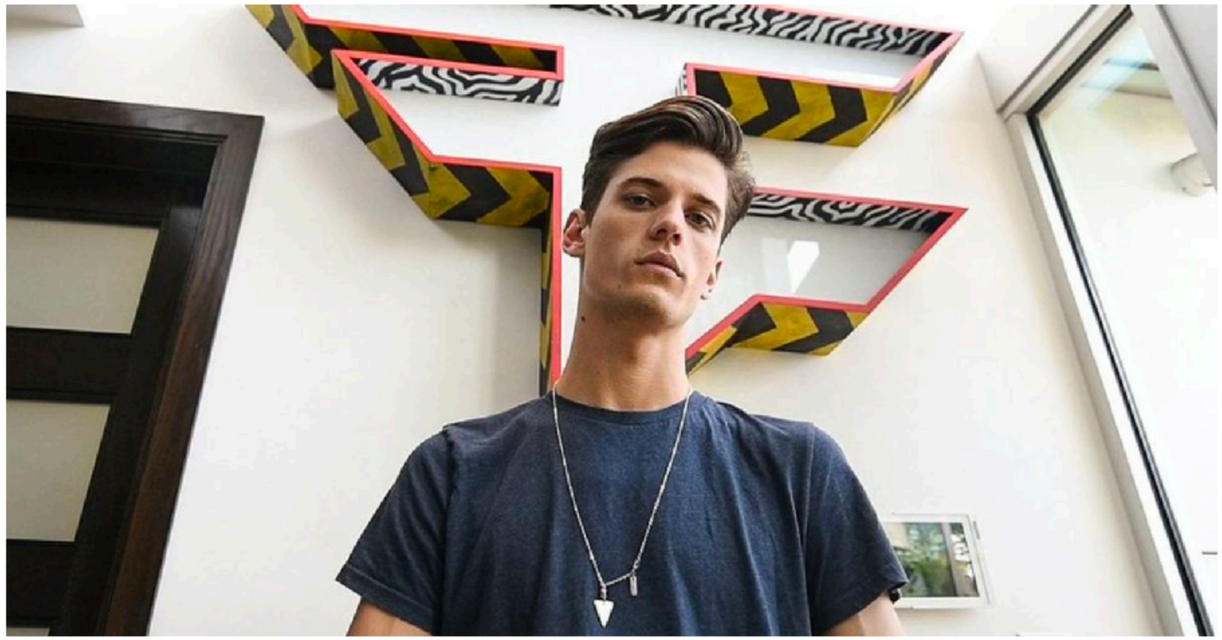
Coordinated campaigns of **toxic comments** on social media that attempt to silence voices.

Falsely reporting targets to authorities or platforms to take action against their person or accounts.

Twitch Streamer Nate Hill Swatted While Streaming Fortnite

A swatting incident is a terrifying event for all involved, which is why fans were concerned when streamer Nate Hill had to cut his stream suddenly.

BY MICHAEL LEE
PUBLISHED FEB 24, 2021



Online Hate and Harassment is Ubiquitous

41% of people in US

40% of people globally

Source: PEW Research Center Online Harassment 2021, Microsoft Digital Civility Index

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.

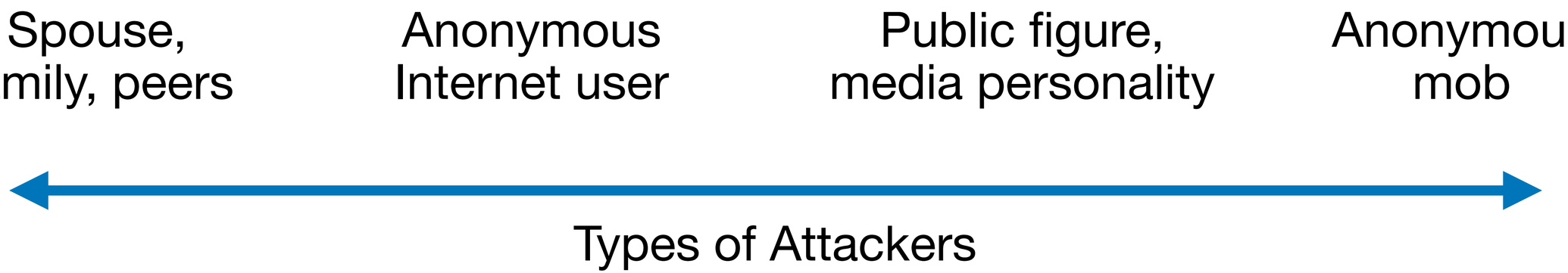
Literature Review

- Examined the last five years of research and journalism on online hate and harassment
  - IEEE S&P, USENIX Security, CCS, CHI, CSCW, ICWSM, Web, 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

Threat Model: Targets and Attackers

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

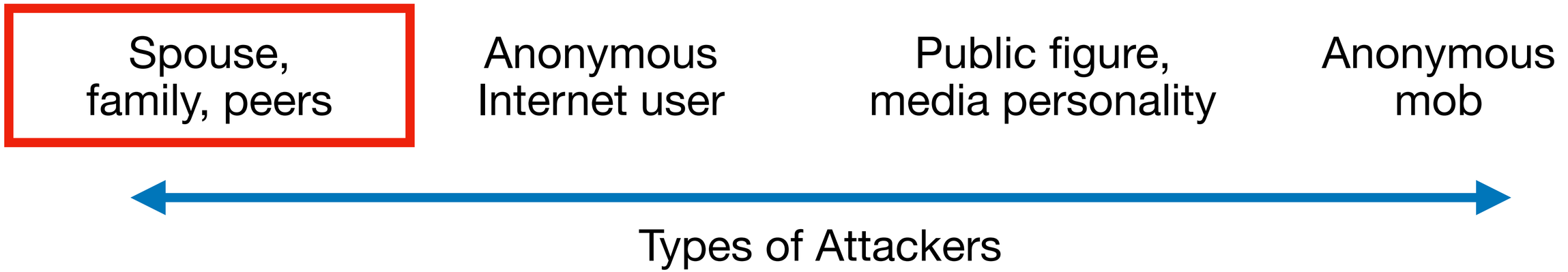
An attacker's main goal is to emotionally harm or coercively control the target.



Threat Model: Targets and Attackers

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

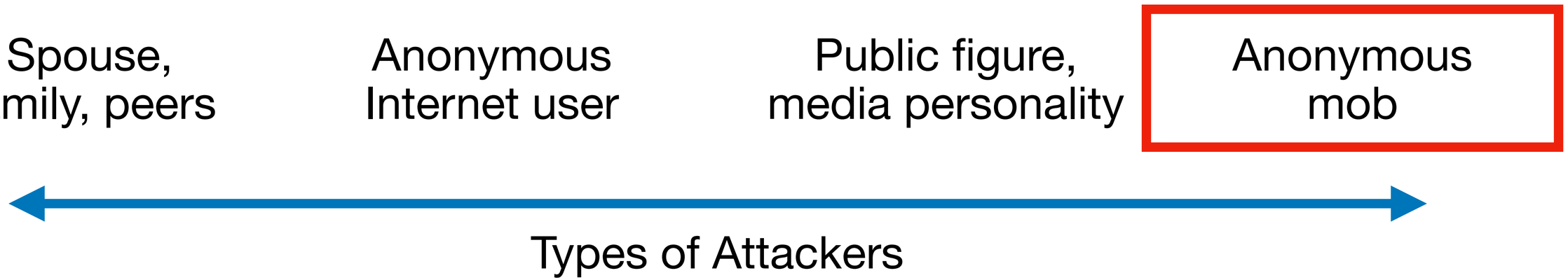
An attacker's main goal is to emotionally harm or coercively control the target.



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.



Differentiating Attacks

We 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 Online Hate and Harassment

Attack Type	Security Principle
Toxic Content	Availability
Content Leakage	Confidentiality
Overloading	Availability
False Reporting	Integrity
Impersonation	Integrity
Surveillance	Confidentiality
Lockout and Control	Integrity, Availability

Seven Classes of Online Hate and Harassment

Attack Type	Security Principle	Classic Abuse
Toxic Content	Availability	Spam
Content Leakage	Confidentiality	Data Breaches
Overloading	Availability	DoS, DDoS
False Reporting	Integrity	Mark not-spam
Impersonation	Integrity	Phishing
Surveillance	Confidentiality	RAT, Tracking
Lockout and Control	Integrity, Availability	Ransomware

Seven Classes of Online Hate and Harassment

Attack Type	Security Principle	Classic Abuse
Toxic Content	Availability	Spam
Content Leakage	Confidentiality	Data Breaches
Overloading	Availability	DoS, DDoS
False Reporting	Integrity	Mark not-spam
Impersonation	Integrity	Phishing
Surveillance	Confidentiality	RAT, Tracking
Lockout and Control	Integrity, Availability	Ransomware

Seven Classes of Online Hate and Harassment

Attack Type	Security Principle	Classic Abuse
Toxic Content	Availability	Spam
Content Leakage	Confidentiality	Data Breaches
Overloading	Availability	DoS, DDoS
False Reporting	Integrity	Mark not-spam
Impersonation	Integrity	Phishing
Surveillance	Confidentiality	RAT, Tracking
Lockout and Control	Integrity, Availability	Ransomware

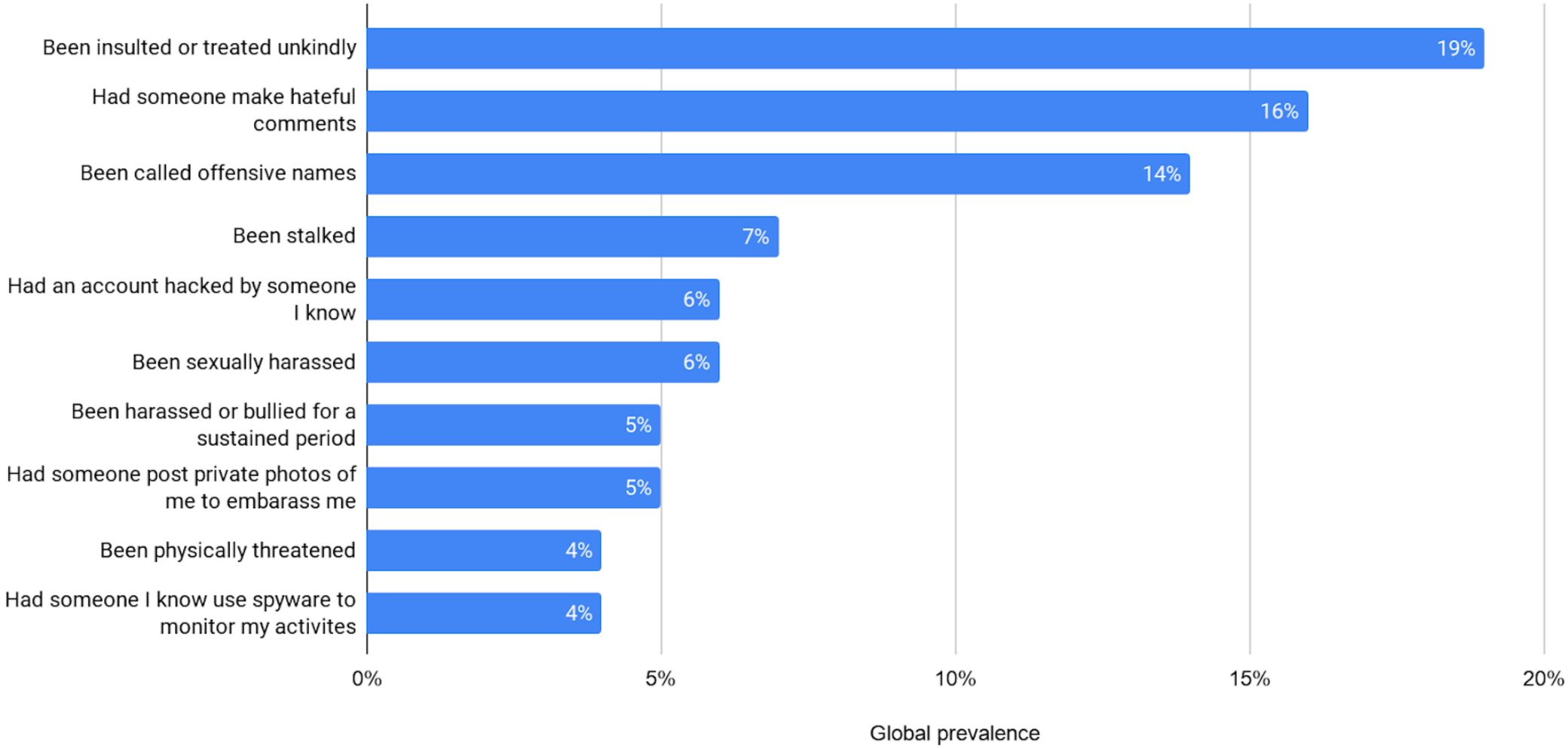
There is no single solution to address the diverse set of hate and harassment attacks.

But it gets more complicated than that.

Survey Instrument

- Surveyed ~1000 participants from 22 countries each 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)

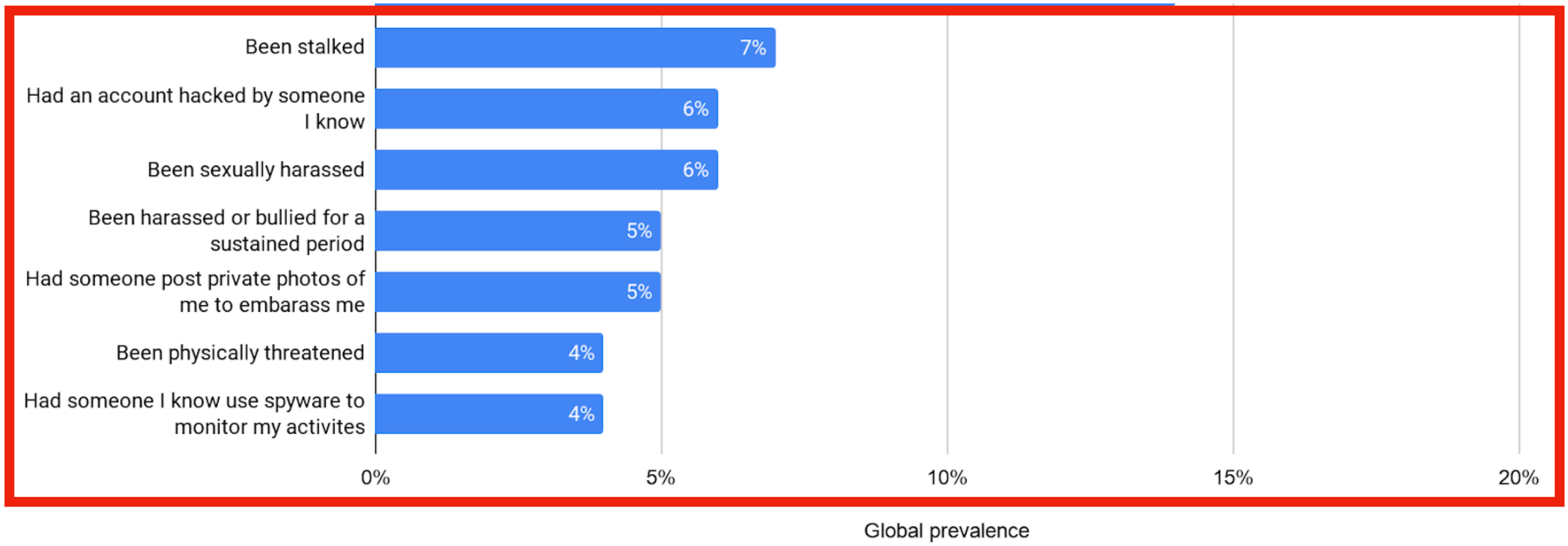
Breakdown of Harassment Experiences



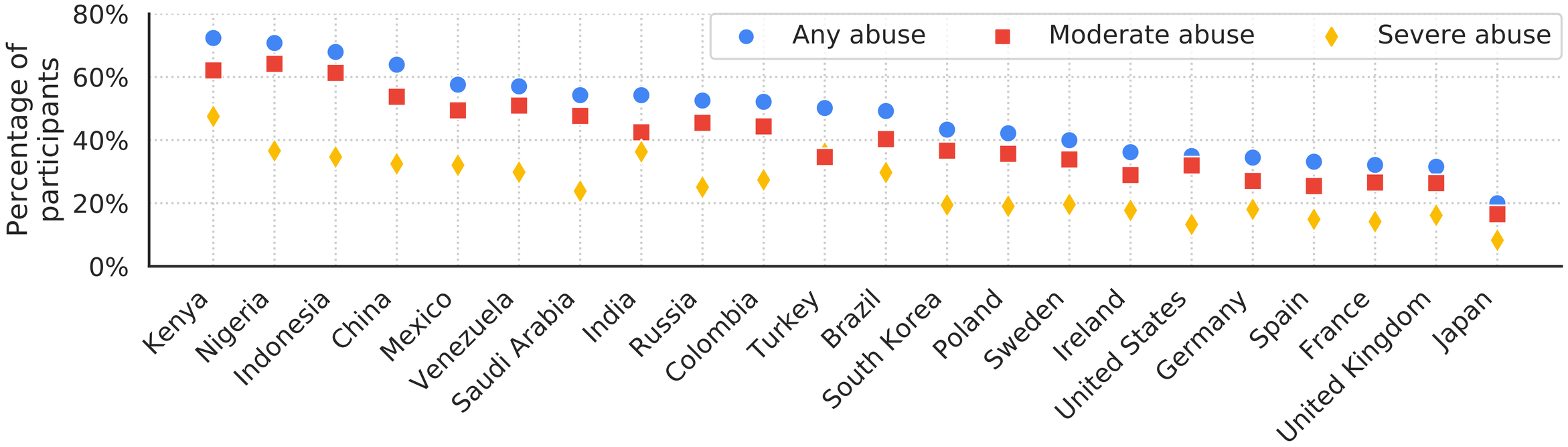
Breakdown of Harassment Experiences

Toxic content is one of the largest threats Internet users face.

Breakdown of Harassment Experiences



Prevalence of Online Hate and Harassment



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
- 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

Measuring hate and harassment outcomes

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

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

How Google's Jigsaw Is Trying to Detoxify the Internet

Can Facebook Use AI to Fight Online Abuse?

The task of detecting abusive posts and comments on social media is not entirely technological

Instagram to use artificial intelligence to detect bullying in photos

The move highlights efforts from tech companies to use automation to moderate their platforms.

07-17-20

Twitter automatically flags more than half of all tweets that violate its rules

07-17-20

Twitter automatically flags more than half of all tweets that violate its rules

Twitter still failing women over online violence and abuse

Users may disagree about what constitutes toxic content online, leading to “gray areas” in automated classification

How do users from diverse backgrounds interpret toxic content online?

Survey Participants (US only)

17K
participants

108K
comments

Survey Participants (US only)

17K
participants

27%
minorities

13%
LGBTQ+

108K
comments

51%
religious

50%
parents

Sourcing and Sampling Toxic Content

- Aggregated a corpus of 550K comments from Twitter, Reddit, 4chan
- Each comment was labeled through Google's Perspective API, which provides a score from 0 – 1 on toxicity
- Wanted to prioritize comments with *mixed* agreement across raters to understand why and where perspectives were different
    - Conducted a pilot study (N = 200) rated 800 comments, 80 from each stride, to identify comments with least agreement

Stride	% Agreement	% Final Dataset
0.0 – 0.1	90%	5%
0.1 – 0.2	81.8%	5%
0.2 – 0.3	80%	5%
0.3 – 0.4	76.4%	10%
0.4 – 0.5	71.4%	10%
0.5 – 0.6	65.2%	15%
0.6 – 0.7	68.3%	15%
0.7 – 0.8	65.2%	20%
0.8 – 0.9	76.4%	10%
0.9 – 1.0	80%	5%

Overall Perceived Toxicity

- 53% of comments have a median rating of “Not at all toxic” across all five raterswhile 1% of comments have a median rating of “Extremely toxic”
    - *“They’re like the polar opposite of limp-writes smug douchebag homo’s [sic] and liberals who buy a Prius and think they’re better than everyone else.”*

Overall Perceived Toxicity

- 53% of comments have a median rating of “Not at all toxic” across all five raterswhile 1% of comments have a median rating of “Extremely toxic”
  - *“They’re like the polar opposite of limp-writes smug douchebag homo’s [sic] and liberals who buy a Prius and think they’re better than everyone else.”*
- 39.1% of comments were rated either “Slightly toxic” or “Moderately toxic”
  - *“Kids with hoodies are going to be our future criminals”*
  - *“Women can’t be responsible for hiring people. It is foolish to entrust hormonal women with the most important part of the company.”*

Participants Regularly Disagree on Comment Toxicity

“So you don’t want money.... Just free college, loan forgive-ness, and (and I’m not sure how this is relevant) healthcare for veterans? I presume you believe only blacks were slaves? Also, your last sentence implies you believe all blacks were slaves..

Identities, Experiences, and Toxicity

1.2x
18-24 vs. 35-44

1.1x
minorities vs. non-minority

1.6x
LGBTQ+ vs. non-LGBTQ+

1.3x
parent vs. non-parent

Identities, Experiences, and Toxicity

1.2x
18-24 vs. 35-44

1.1x
minorities vs. non-minority

0.8x
witnessed toxic content

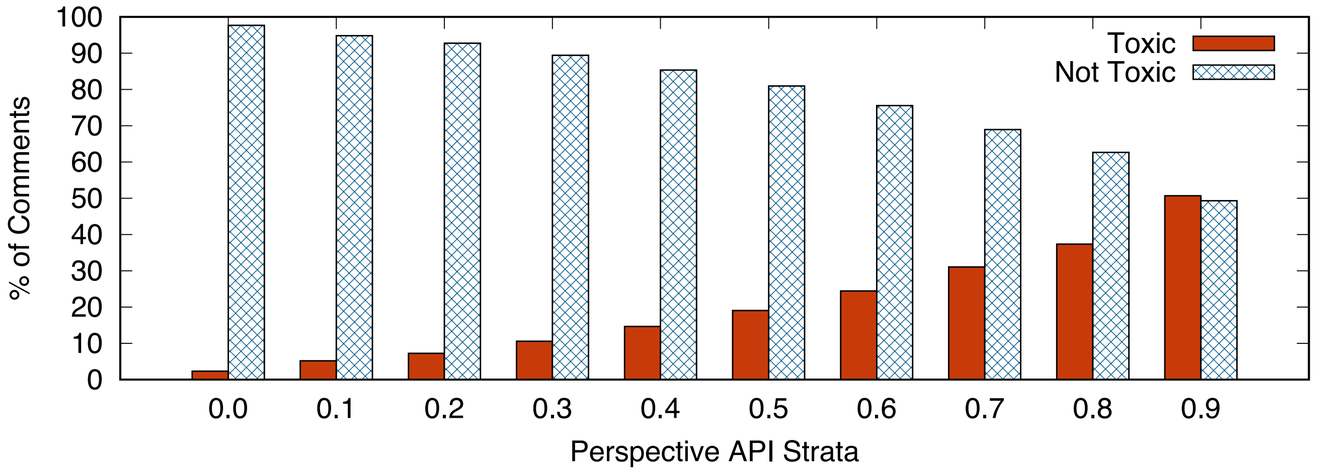
1.6x
LGBTQ+ vs. non-LGBTQ+

1.3x
parent vs. non-parent

1.5x
target of toxic content

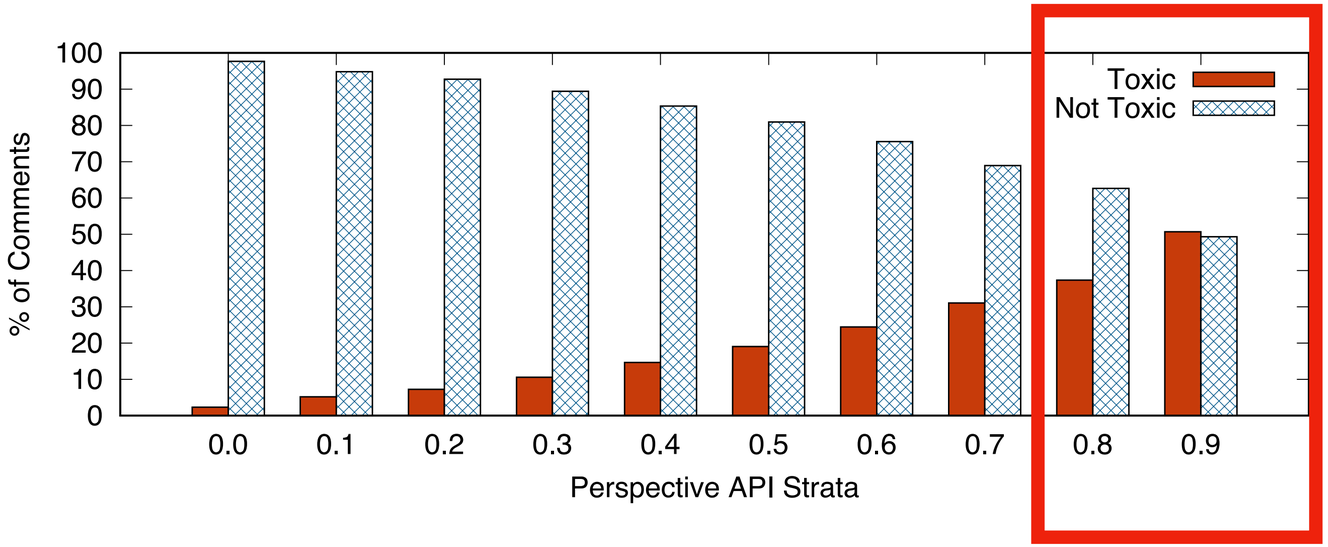
Benchmarking Toxicity Classifiers

- Benchmarked Google Jigsaw's Perspective API, which is a state of the art classifier with our collected dataset



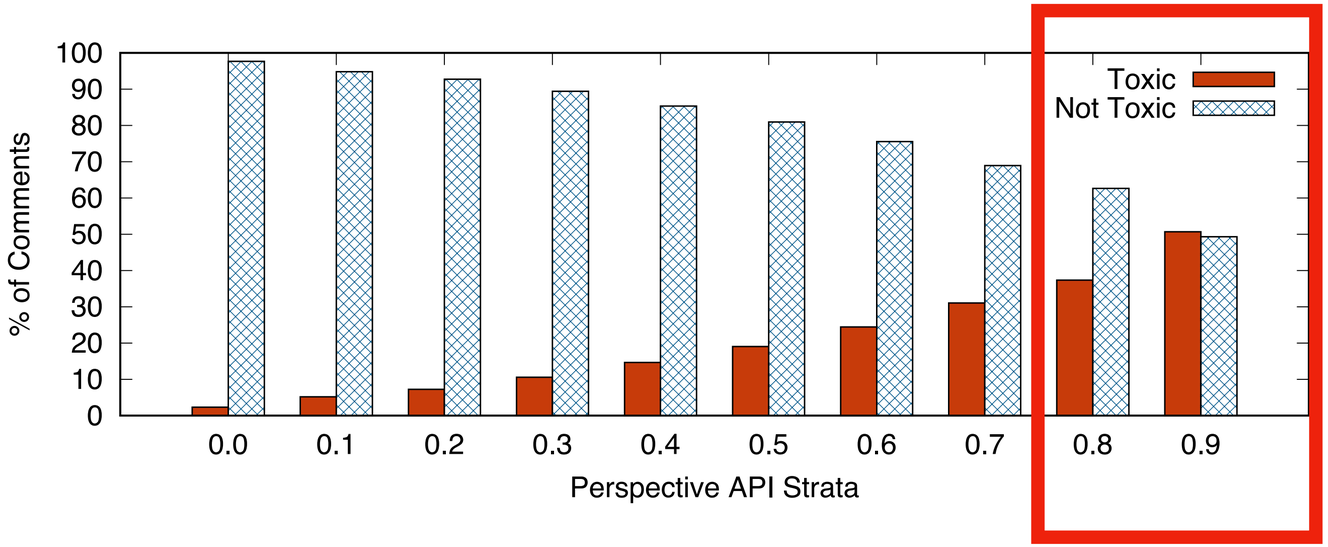
Benchmarking Toxicity Classifiers

- Benchmarked Google Jigsaw's Perspective API, which is a state of the art classifier with our collected dataset
- At highest strides, accuracy of classifier was at best 51%
- Existing classifiers **fail to capture** divergent toxicity perspectives



Benchmarking Toxicity Classifiers

- Benchmarked Google Jigsaw's Perspective API, which is a state of the art classifier with our collected dataset
- At highest strides, accuracy of classifier was at best 51%
- Existing classifiers **fail to capture** divergent toxicity perspectives



Can we do better?

Personalized abuse protections can help account for diverse perspectives in toxic content classification

Fine Tuning Toxicity Classifiers

0.35

avg. precision

0.37

avg. accuracy

Fine Tuning Toxicity Classifiers

0.35

avg. precision

personalized tuning

0.37

avg. accuracy

Fine Tuning Toxicity Classifiers

0.35
avg. precision

0.37
avg. accuracy

personalized tuning

0.6
avg. precision

0.68
avg. accuracy

What kinds of comments do people not want to see?

Sounds like you’re a no one who’s gonna die bitter and alone and forgotten

What kinds of comments do people not want to see?

Sounds like you're a no one who's gonna die bitter and alone and forgotten

Store them in an unventilated room with hoses that run between the room and your car's exhaust pipe. That'll solve your problem.

Taking into account divergent perspectives **can improve** existing automated tools for toxicity detection.

But there’s a long way to go.

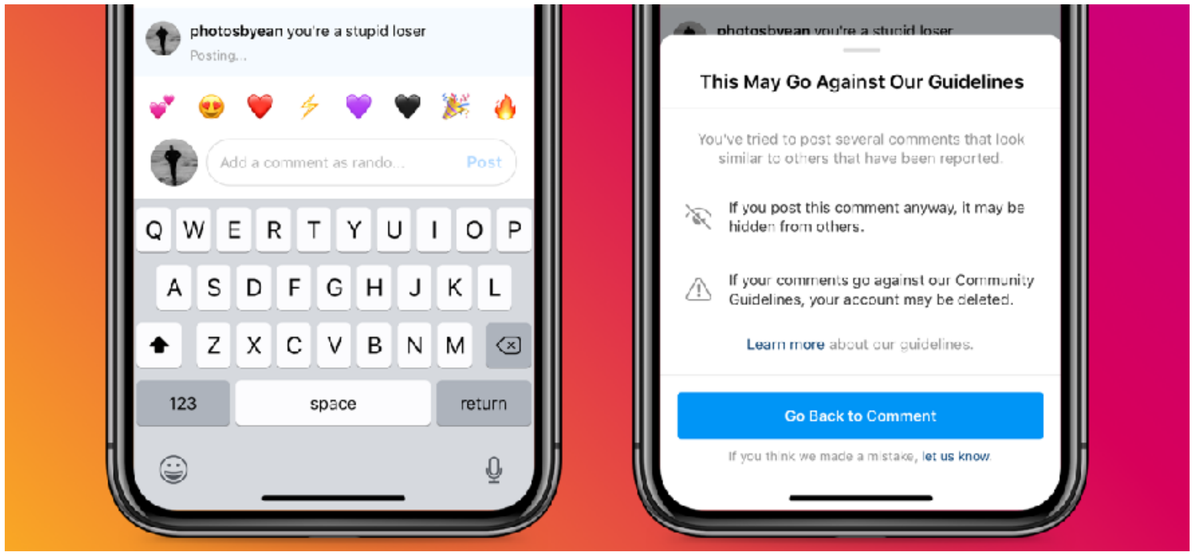
Tensions and Challenges

- How do we empower targets of abuse instead of burdening them with choice?
- How do you balance moderation with filter bubbles and free speech?
- How do we enable both privacy and accountability?

TikTok Admits It Suppressed Videos by Disabled, Queer, and Fat Creators

Towards Solutions and Interventions

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



Perspective

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

Key Takeaways

- Online abuse is *changing*, the security community can and should work towards tackling the problem
- Online hate and harassment is growing over time and especially dangerous to some Internet users
- Many techniques and defenses are already well studied in the security community, can draw on these for future research

Key Takeaways

- Online abuse is *changing*, the security community can and should work towards tackling the problem
- Online hate and harassment is growing over time and especially dangerous to some Internet users
- Many techniques and defenses are already well studied in the security community, can draw on these for future research

Deepak Kumar

kumarde@cs.stanford.edu

@_kumarde