The Moderation Machine

A study of user attitudes towards online harassment, mechanisms for flagging content, and remediation techniques

Emily Witt • Jason Danker • Molly Mahar • Paul Glenn
Advisor: Deirdre Mulligan
73% have witnessed online harassment

40% have personally experienced it

THE SOLUTION TO ONLINE ‘HARASSMENT’ IS SIMPLE: WOMEN SHOULD LOG OFF

Source: http://www.breitbart.com/milo/2016/07/05/solution-online-harassment-simple-women-log-off/
New report shows the reach of online harassment, digital abuse, and cyberstalking – Data & Society

Inside Twitter's new plan to combat harassment – CNN Tech

Meet HeartMob: A Tool For Fighting Online Harassment Designed By People Who Have Been Harassed – Fast Company
Challenges to Understanding

Lack of definition

Context matters

Lack of public access to platform research on this topic
Challenges to Understanding

- Lack of definition
- Context matters
- Lack of public access to platform research on this topic

Challenges to Solving

- Human moderation isn’t perfect
- Algorithmic moderation is FAR from perfect
What We Wanted to Learn

- Do users see differences between moderation methods?
- What do users want done about harassment?
- Does a flag influence the assessment of harassment?
- Do users balance harassment against free speech?
Survey Experiment

How harassing is the public post shown above?

- Not at all harassing
- Slightly harassing
- Somewhat harassing
- Very harassing
- Extremely harassing

Regardless of whether you find this post harassing, how appropriate or inappropriate is it for a public platform?
Survey Experiment

Directed
@UserName · Yesterday
@User Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas et consequat erat, et bibendum est.

Undirected
@UserName · Yesterday
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas et consequat erat, et bibendum est.

How harassing is the public post shown above?
- Not at all harassing
- Slightly harassing
- Somewhat harassing
- Very harassing
- Extremely harassing

Regardless of whether you find this post harassing, how appropriate or inappropriate is it for a public platform?
Survey Experiment

How harassing is the public post shown above?

Regardless of whether you find this post harassing, how appropriate or inappropriate is it for a public platform?

- Not at all harassing
- Slightly harassing
- Somewhat harassing
- Very harassing
- Extremely harassing

Human Moderators Flag

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas et consequat erat, et bibendum est.

Human moderators have flagged this message as potentially harassing

Other Users Flag

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas et consequat erat, et bibendum est.

Other users have flagged this message as potentially harassing

Algorithmic Flag

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas et consequat erat, et bibendum est.

An algorithm has flagged this message as potentially harassing

Ambiguous Flag

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Maecenas et consequat erat, et bibendum est.

This message has been flagged as potentially harassing
The following slide contains three messages that some may find disturbing.

These are the actual messages used in our study. They include hate speech targeted at women, Muslims, and African Americans, including incitements to violence against these groups.

We will show them for 15 seconds.
Experimental Condition: Targeted Group

@UserName • Yesterday
@User just putting it out there, you feminists deserve all those death threats you’re getting.

@UserName • Yesterday
Oh and @User is a Muslim. So when do we get to go kill all the Muslims in the United States since they get to kill us in our own country.

@UserName • Yesterday
Let’s go have a picnic and pick a random black like @User to publicly lynch in front of everyone as a form of entertainment.
Sadly, these types of messages were not hard to find
Participants Agreed:
These Messages Are Terrible

- Not at all harassing: 3%
- Slightly harassing: 7%
- Somewhat harassing: 11%
- Very harassing: 24%
- Extremely harassing: 55%
As participants selected up to three remediation options, these percentages will not add up to 100%.

Participants Want Them Removed

- Remove the message from the platform: 52%
- Permanently ban the poster: 40%
- Post a public reminder of the community code of conduct in response to the message: 23%
- Before publicly posting, notify poster it has been flagged as harassing and require confirmation to post: 22%
- Send the poster a private reminder of the community code of conduct: 21%
- Temporarily suspend the poster: 19%
- Add the poster to a blocklist which hides their posts from all people who subscribe to that blocklist: 19%
- Prevent the poster from posting again until they complete an empathy exercise: 12%
- Take no action: 4%
Flagging Increases Their Desire to Remove It

Percentage of respondents selecting “Remove the message from the platform”

- Ambiguous Flag: 54%
- Human Moderator Flag: 54%
- Algorithmic Flag: 53%
- Other Users Flag: 52%
- No Flag: 47%
Participants Want to Avoid Letting Harassing Content Slip Through the Cracks

Responses to “When creating a computational moderation tool, what is most important to prevent?”

- **Catch Less:** “Harassing content not being flagged as harassing and not subjected to remediation” (56%)
- **Catch More:** “Non-harassing content being flagged as harassing and potentially subjected to remediation” (24%)
- **No Preference** (19%)
Humans Are Perceived as More Accurate, Fair, and Trustworthy than Algorithms

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>Fairness</th>
<th>Trustworthiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8</td>
<td>3.9</td>
<td>3.7</td>
</tr>
<tr>
<td>3.5</td>
<td>3.6</td>
<td>3.4</td>
</tr>
<tr>
<td>3.2</td>
<td>3.3</td>
<td>3.3</td>
</tr>
</tbody>
</table>

Inaccurate (1) to Accurate (5)
Unfair (1) to Fair (5)
Untrustworthy (1) to Trustworthy (5)
The Moderation Machine

Moderation Machine is a platform for reflection about online harassment, one of the great issues of our time. Here, you’ll have the opportunity to answer questions about your views on the issue and compare your answers to a broad survey of the population. No data will be collected during your experience on this site.

Enter the Site
Very interesting topic and difficult problems to solve. Let your users help with this issue. I don't think algorithms can get you all the way there. Thanks for letting me participate!
you don’t understand how the internet works. you can’t stop deliberate trolls, kids. also berkley is trash go fight off your communist terrorists that riot every week. oh wait you can’t cause your mayor is in bed with them lol.
Acknowledgments

Deirdre Mulligan
Coye Cheshire
The Center for Long-Term Cybersecurity
The Center for Technology, Society and Policy
Questions?
Future Work

- Re-test with less-violent messages
- Re-test with other dependent variables (*how hurtful, polite, protected, ...?*)
- Explore “free speech” area in more depth
- Explore in greater detail the contours of “harassment” definitions
How is Online Harassment Commonly Defined?

- Hate speech (racist, sexist, homophobic speech, etc.)
- Doxing (releasing your personal information online)
- Threats of violence
- Posting false information about you (fake quotes, altered images, libel, etc.)
- Impersonation
- Encouraging others to harass you via offline methods
- Revenge porn or non-consensual photography

When Ambiguously Flagged, Participants Assume Humans Are the Ones Flagging

- 33%
- 33%
- 17%
- 15%
- I don’t know
- Other 2%
Whether Content was Directed or Undirected Didn’t Influence Perceptions of Harassment

- Undirected: 4.17
- Directed: 4.24
- All Messages: 4.20

Rated level of harassment on a 1-5 scale, 5 being “Extremely harassing”