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*

Pew Center Study: Duggan, Maeve. (Oct, 2014). Online Harassment. Pew Research Center. http://www.pewinternet.org/2014/10/22/online-harassment/

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

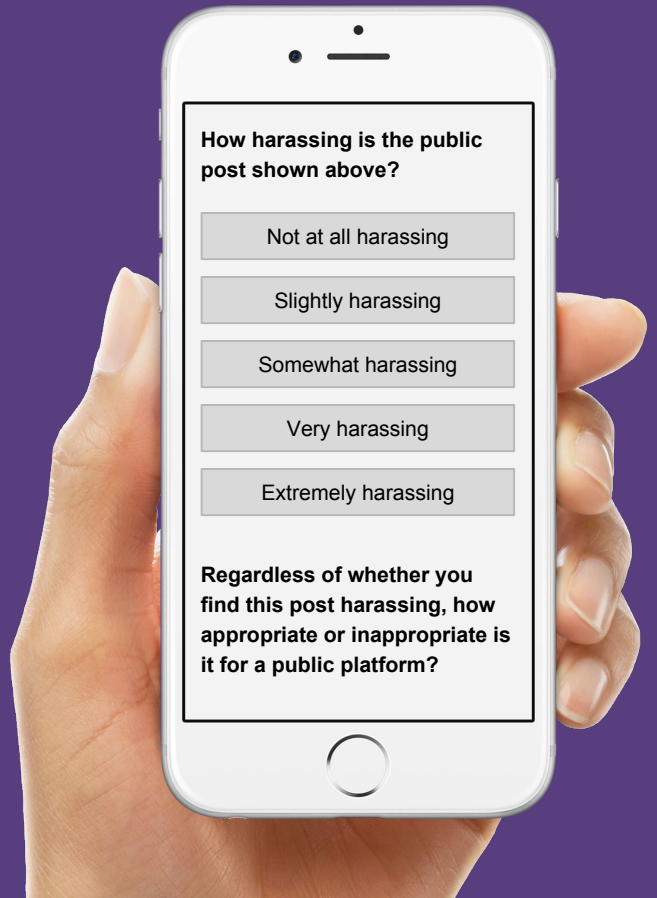
Do users see differences between moderation methods?

Does a flag influence the assessment of harassment?

What do users want done about harassment?

Do users balance harassment against free speech?

Survey Experiment



Survey Experiment

Directed

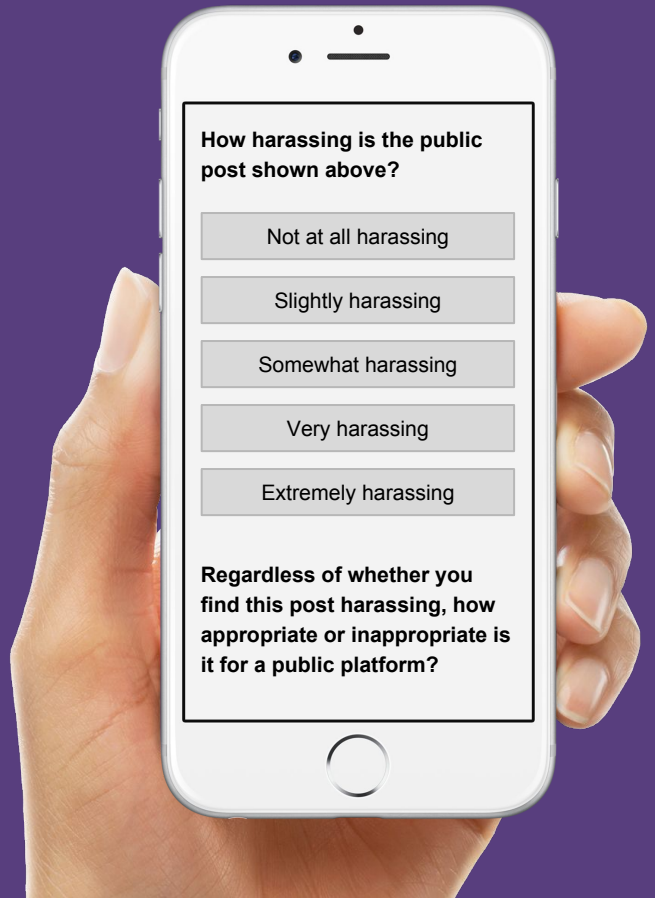
@UserName • Yesterday

@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.



Survey Experiment

Human Moderators Flag

@UserName • Yesterday

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

@UserName • Yesterday

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

@UserName • Yesterday

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

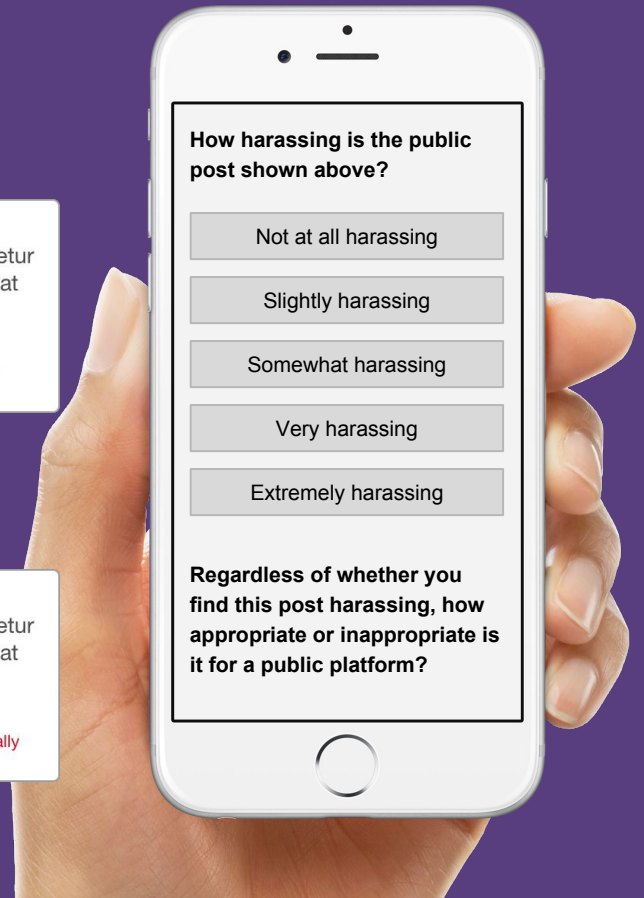
An algorithm has flagged this message as potentially harassing

Ambiguous Flag

@UserName • Yesterday

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.

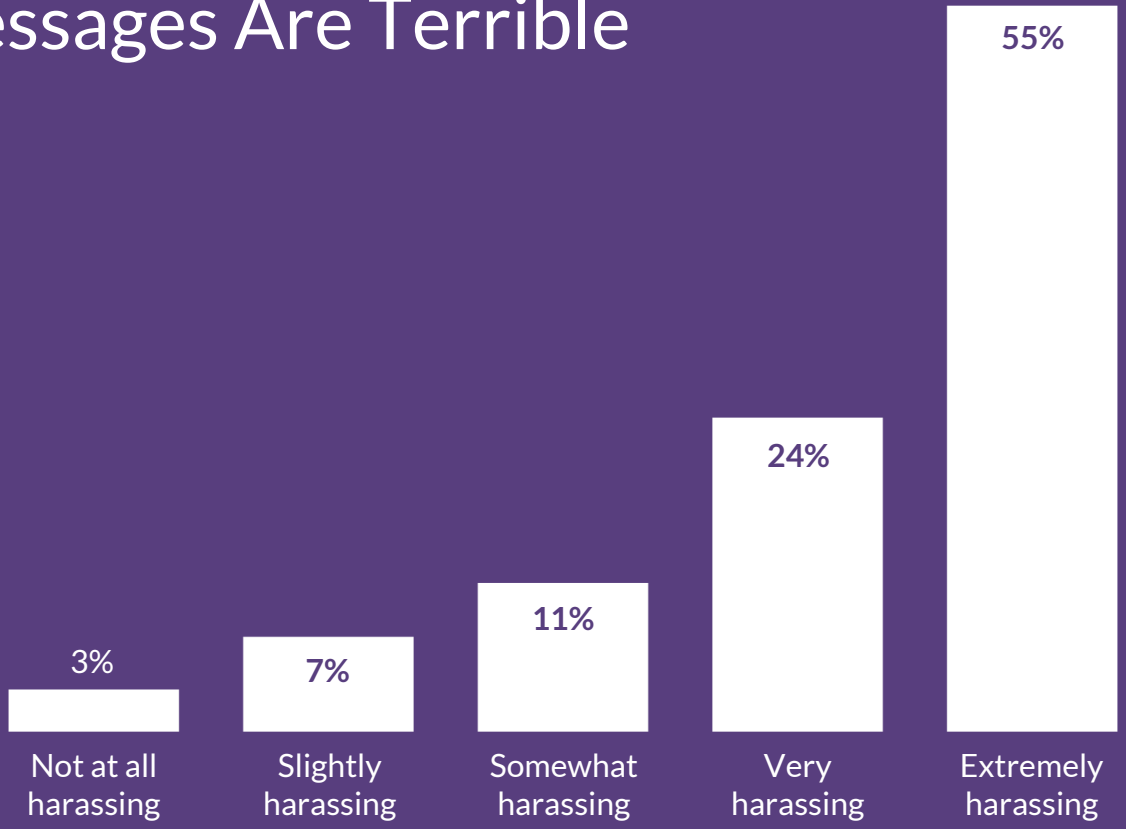
These are the actual messages used in our study. They include hate speech targeted at women, Muslims, and African Americans, including incitements to violence against these groups.

We will show them for 15 seconds.

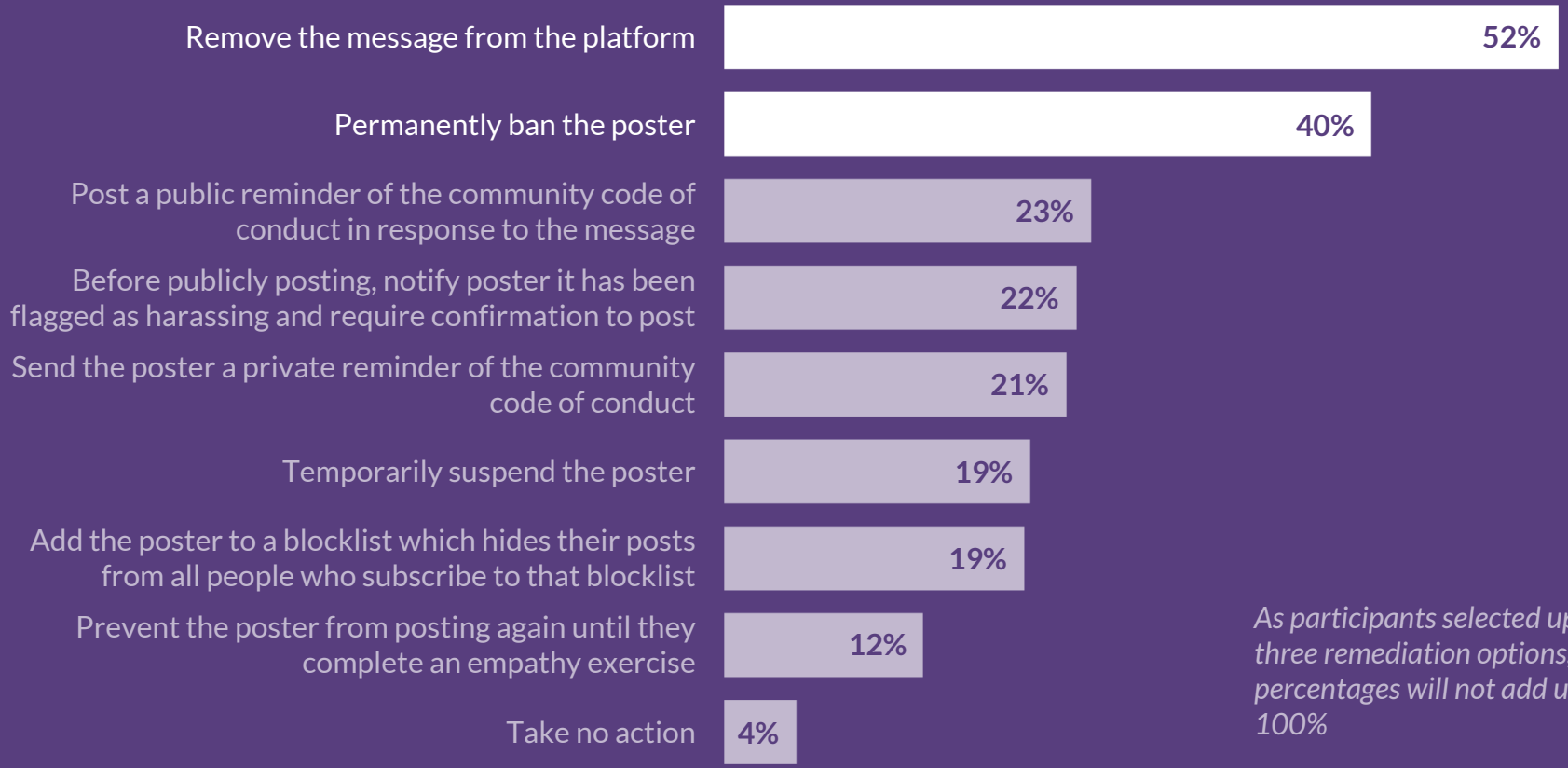
Experimental Condition: Targeted Group

Sadly, these types of messages were not hard to find

Participants Agreed: These Messages Are Terrible

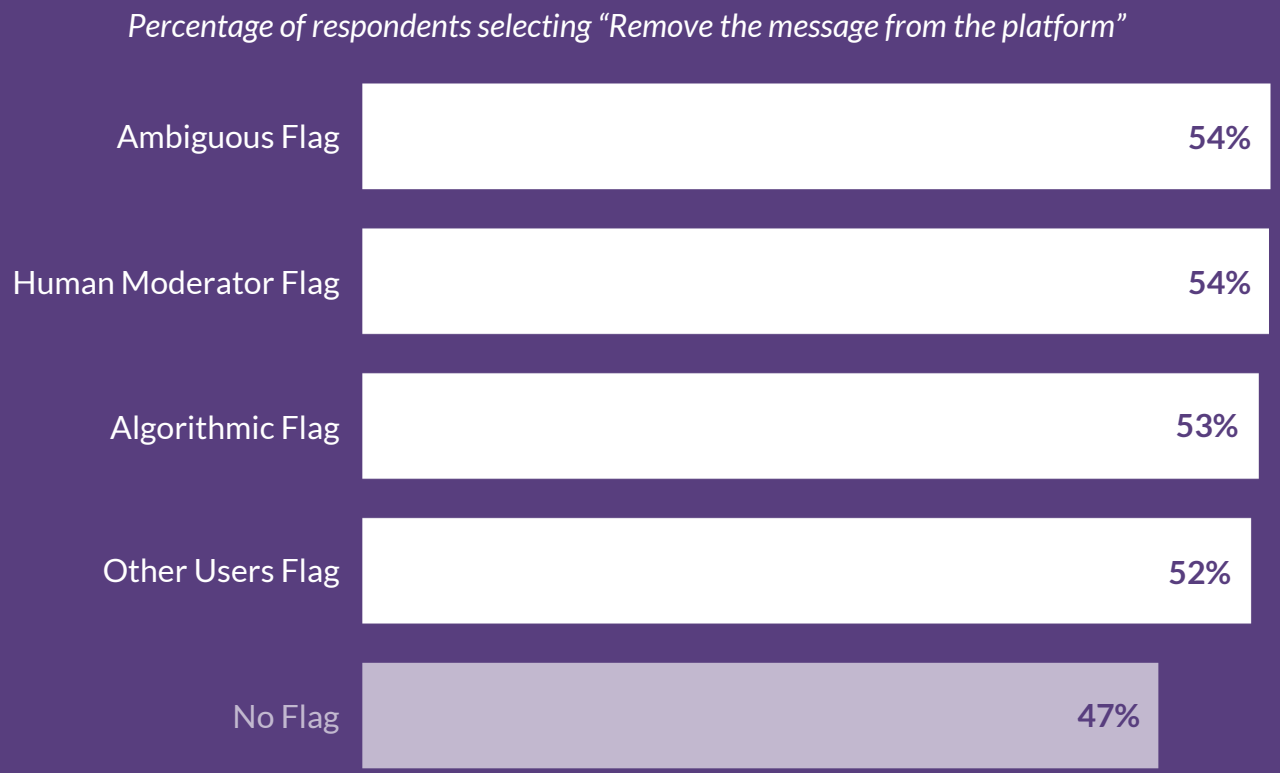


Participants Want Them Removed



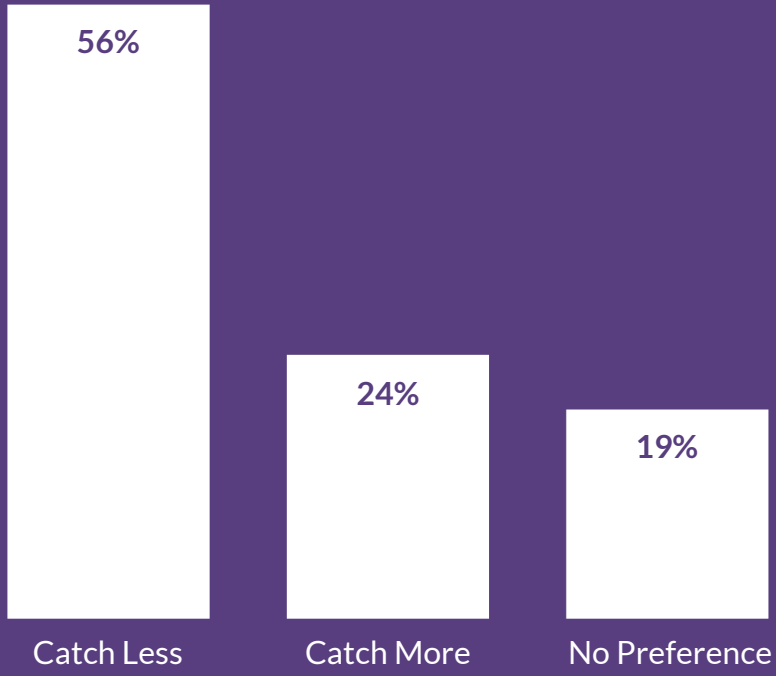
As participants selected up to three remediation options, these percentages will not add up to 100%

Flagging Increases Their Desire to Remove It



Participants Want to Avoid Letting Harassing Content Slip Through the Cracks

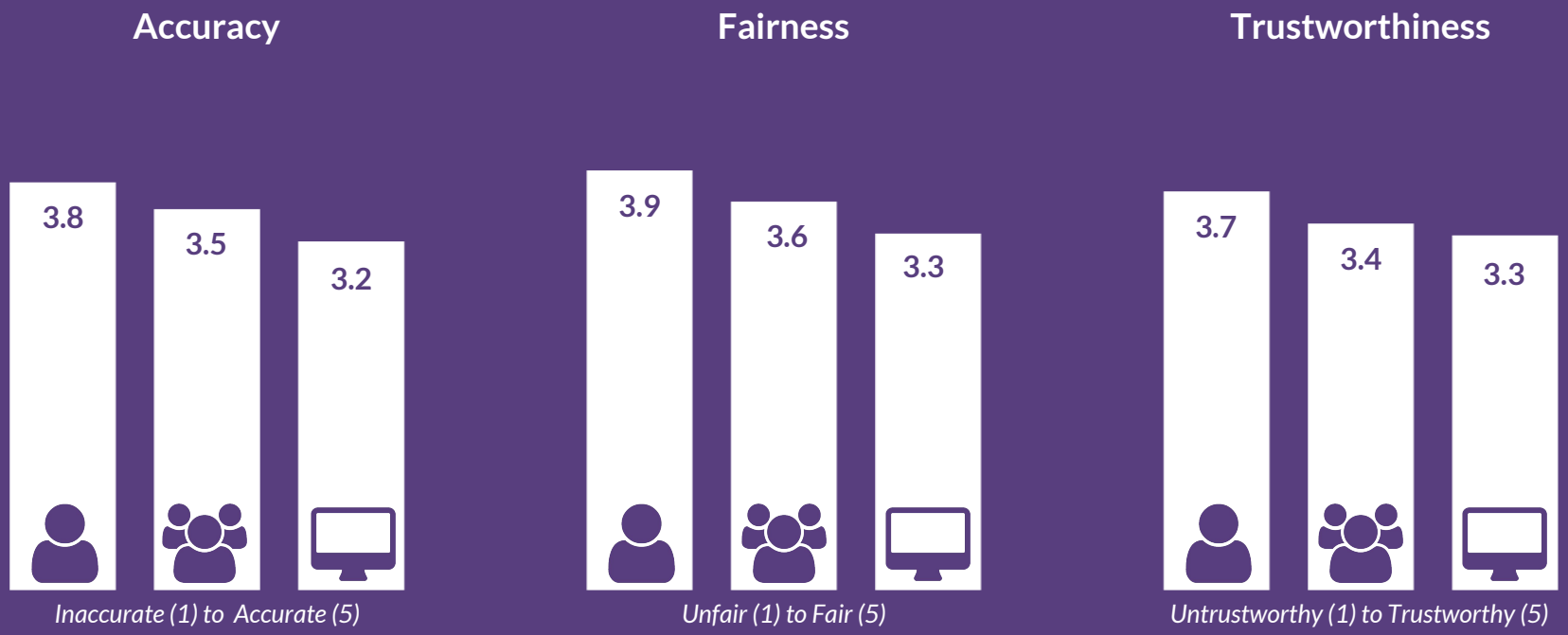
Responses to "When creating a computational moderation tool, what is most important to prevent?"

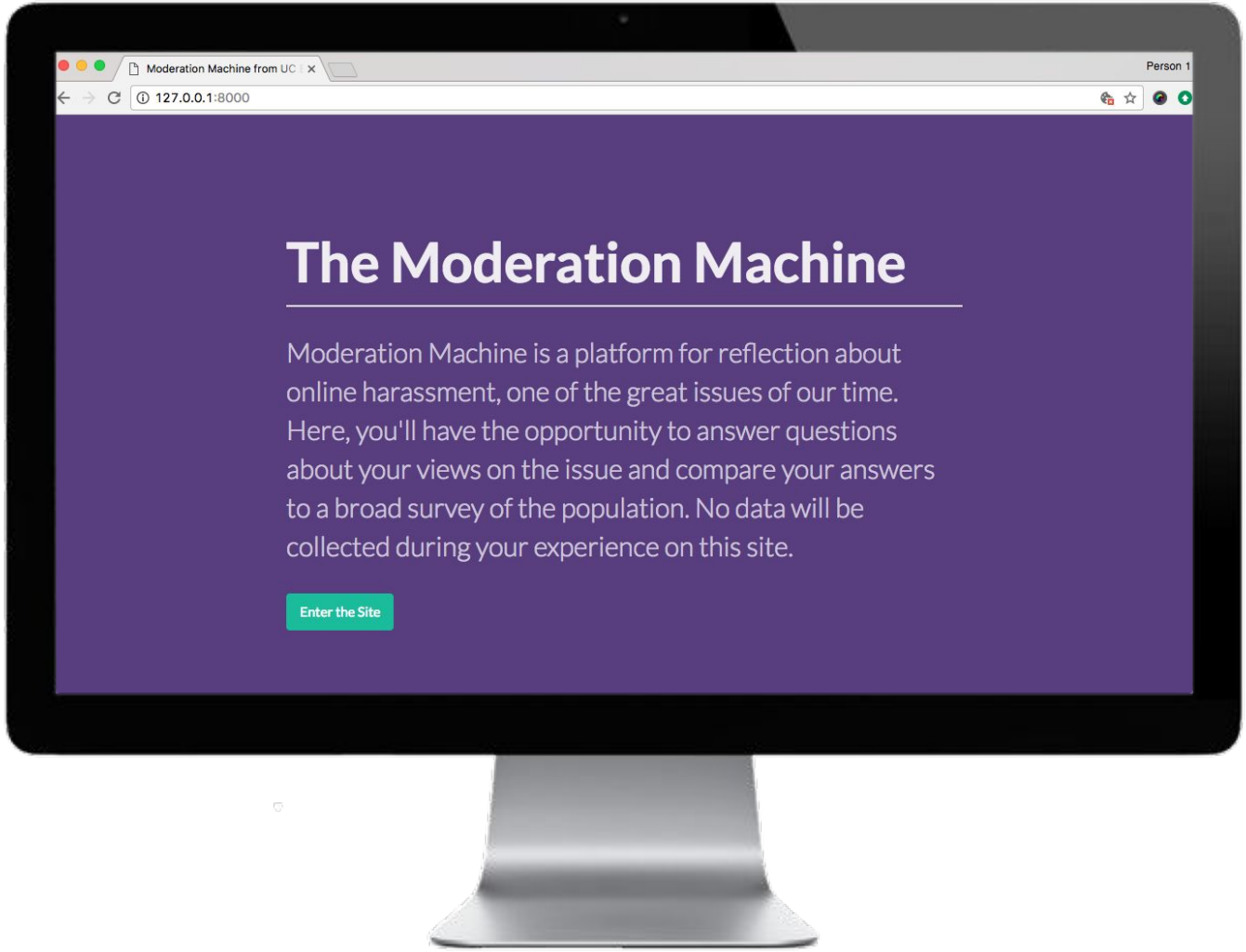


Catch Less: "Harassing content not being flagged as harassing and not subjected to remediation"

Catch More: "Non-harassing content being flagged as harassing and potentially subjected to remediation"

Humans Are Perceived as More Accurate, Fair, and Trustworthy than Algorithms




Moderation Machine from UC
Person 1
127.0.0.1:8000
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 cant stop deliberate trolls, kids. also berkley is trash go fight off your communist terrorists that riot every week. oh wait you cant 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

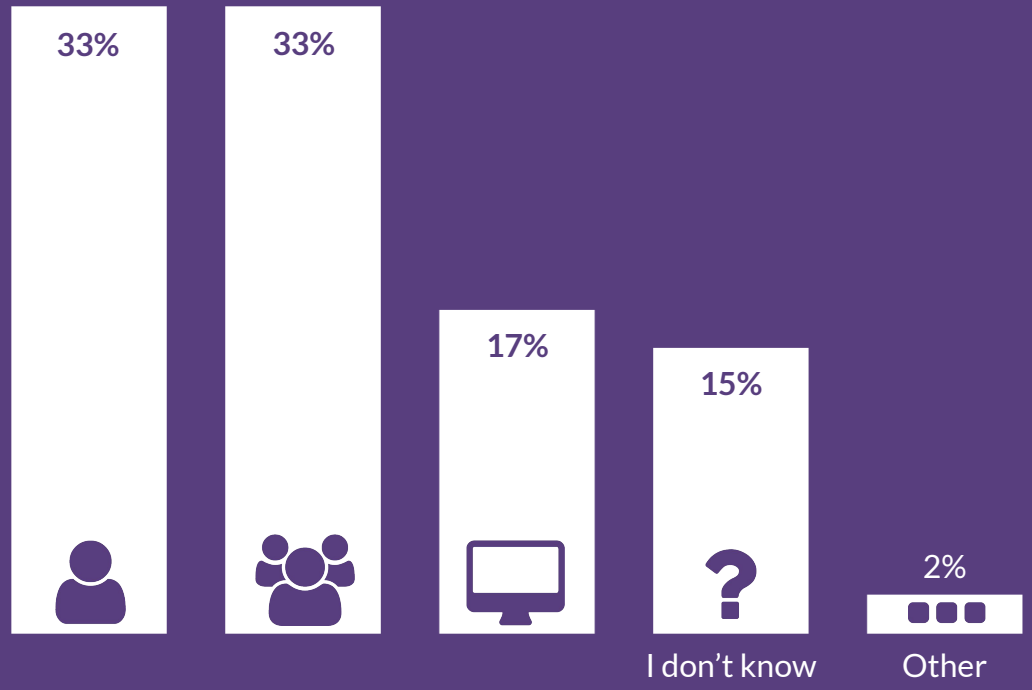
- Re-test with less-violent messages
- Re-test with other dependent variables (*how hurtful*, *polite*, *protected*, ….*?*)
- Explore “free speech” area in more depth
- Explore in greater detail the contours of “harassment” definitions

How is Online Harassment Commonly Defined?

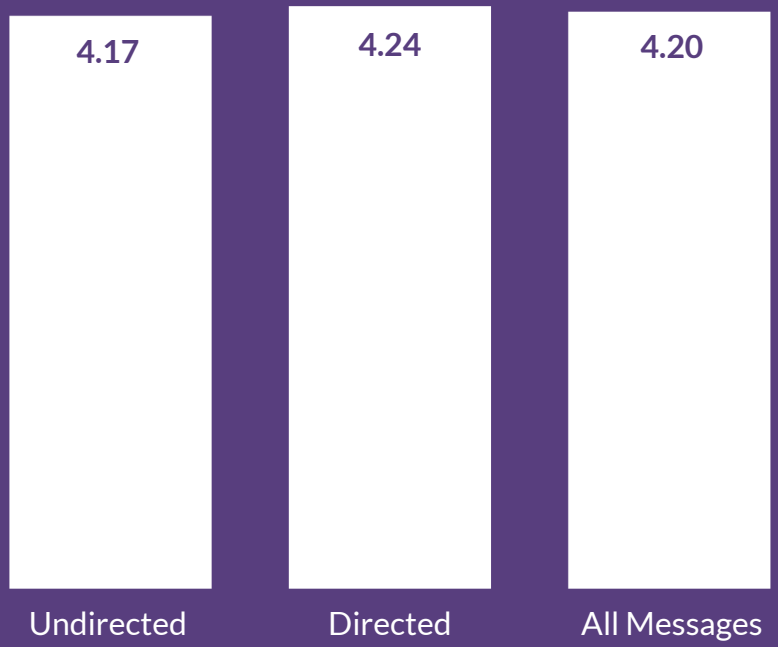
- Hate speech (racist, sexist, homophobic speech, etc.)
- Doxing (releasing your personal information online)
- Threats of violence
- Posting false information about you (fake quotes, altered images, libel, etc.)
- Impersonation
- Encouraging others to harass you via offline methods
- Revenge porn or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

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). Reporting, Reviewing, and Responding to Harassment on Twitter. Women, Action, and the Media. May 13, 2015. http://womenactionmedia.org/twitter-report.

When Ambiguously Flagged, Participants Assume Humans Are the Ones Flagging



Whether Content was Directed or Undirected Didn’t Influence Perceptions of Harassment



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