# The Moderation Machine

Jason Danker, Paul Glenn, Molly Mahar, Emily Witt
Advisor: Deirdre Mulligan
UC Berkeley School of Information
MIMS Final Project 2017

## Acknowledgments

We would like to acknowledge and thank the Center for Long-Term Cybersecurity and the Center for Technology, Society & Policy for funding this work. We are also deeply appreciative of our advisor Deirdre Mulligan for lending her expertise and guiding us on this project, as well as Coye Cheshire for providing input on our research design and analysis and Steve Weber for his support. We would also like to thank Beth McBride and Taylor Honda for their assistance with usability testing.

## Introduction

This study focuses on user attitudes towards online harassment, methods of flagging content, and remediation mechanisms. In order to understand what factors might contribute to how people identify and experience harassment and judge the appropriateness of content, we explored how sociodemographic and personal characteristics, along with visible message "flagging" indicating potential harassment, influenced individual judgments of harassment and preferences for remediation techniques.

Social media platforms utilize a combination of automated technical measures and human content moderators to address issues like abusive language and online harassment. These efforts can be hindered by the individual and fluid nature of what constitutes harassment to different users, in different contexts, and on different platforms. Though harassment has a legal definition in the law enforcement context, people seem to understand it differently, and differentially, in online communities. In 2016, Jigsaw, a subsidiary of Alphabet that works on online harassment and other global security challenges, convened a workshop with researchers, representatives from online platforms, and advocates to identify areas for research and action, with the potential to curtail online harassment. One question they identified, and to which we hope to contribute, was: *How can we improve definitions of online harassment and responses to it?*[1]

Motivated by this definitional question, we did not define harassment for our participants. Instead, we hoped to surface information about how they define it for themselves by asking them to assess a piece of content and determine to what extent they considered it to be harassing. Our intention was to collect this data on messages that included a wide range of dimensions, including level of harassment, characteristic being targeted (i.e., race, religion, gender, sexual orientation, physical ability), degree and type of violence (none, threat of violence, hope for violence, humor about violence, as well as sexual violence and type of violence historically perpetrated on certain groups), and directedness (whether the message is directed toward a particular person). In the interest of maintaining internal validity in our experimental conditions, we narrowed this range to include a set of messages with similar types of threats of violence targeting three groups: feminists, Muslims, and African Americans. These messages had been publicly posted on Twitter, and each one was modified slightly to include both a directed and undirected version, meaning that one version included a mention to a specific user (@User), while the other did not.

Flagging is one way that platforms identify potentially harassing and abusive content. Scholars Kate Crawford and Tarleton Gillespie have investigated the uses of the "rarely studied sociotechnical mechanism" of the flag in the moderation of platforms. They point out that "the interactions between users, flags, algorithms, content moderators, and platforms are complex and highly strategic".[2] We wondered how aware typical internet users were about flagging and reporting mechanisms. Crawford and Gillespie note that "flags are asked to bear a great deal of weight, arbitrating both the relationship

---

[1] High Impact Questions and Opportunities for Online Harassment Research and Action. Cambridge, MA. https://civic.mit.edu/sites/civic.mit.edu/files/OnlineHarassmentWorkshopReport-08.2016.pdf

[2] Crawford, Kate, and Tarleton Gillespie. "What is a flag for? Social media reporting tools and the vocabulary of complaint." New Media & Society 18.3 (2016): 410-428.

between users and platforms, and the negotiation around contentious public issues."[3] Furthermore, Crawford and Gillespie highlight the invisibility of the flagging mechanism when applied to specific pieces of content. Once a piece of content is removed, other users are typically not informed that something was removed, nor whether it was removed based on the presence, type, or source of flagging. We found this absence compelling. *If the flagging process were visible, would that change user judgments about whether material is harassing and how they'd want it addressed by platforms?* To investigate this issue, we appended a visual flag indicator to the messages as one of our experimental conditions and varied the source of the visual flag among an algorithm, a human moderator, other users, and an ambiguous source.

Though the bulk of the work of content moderation is currently done by human moderators,[4] more and more platforms are turning to automated, algorithmic solutions and support mechanisms in their attempts to address the problem of harassment at scale. This shift raises issues around the morality of delegating decisions of social norms to automated systems.[5] As this shift in authority progresses from trained human moderators to computational processes determining what is and is not acceptable, we pose the question: *How do users perceive the fairness, trustworthiness, and accuracy of these types of authority?*

Finally, we looked at user preferences for remediation techniques. While platforms term the practice of removing objectionable content "content moderation," we seek to differentiate between the repair of harms (remediation) and the minimization of negative extremes (moderation). To do this, we proposed additional methods of dealing with potentially harassing content that could encourage pro-social behavior rather than simply minimizing antisocial behavior. We sourced ideas for prevention and remediation options from those currently in use on a variety of platforms, and generated a few using a values-based approach. As such, our remediation techniques were not limited to those aimed solely at content moderation—punitive measures such as the removal of content or user suspension—but instead incorporated a wider range of values. To generate these techniques, we drew from Bowler, et. al.'s paper on using storytelling to elicit user-generated design principles for responding to cyberbullying,[6] Stuart Geiger's conceptualization of values in collective blocklist responses to Twitter harassment,[7] and Shilton et al.'s study of values dimensions.[8] We wondered: *Are internet users receptive to these pro-social remediation mechanisms?* To determine that, we examined the relationship between harassment assessment, flagging, and the preferred choice of remediation techniques.

---

[3] Crawford, Kate, and Tarleton Gillespie. "What is a flag for? Social media reporting tools and the vocabulary of complaint." New Media & Society 18.3 (2016): 410-428.
[4] Chen, Adrian. "The laborers who keep dick pics and beheadings out of your Facebook feed." Wired, October 23 (2014).
[5] Bruno, Latour. "Where Are the Missing Masses? The Sociology of a Few Mundane Artefacts'." (1992): 225-258.
[6] Bowler, Leanne, Cory Knobel, and Eleanor Mattern. "From cyberbullying to well–being: A narrative–based participatory approach to values–oriented design for social media." Journal of the Association for Information Science and Technology 66.6 (2015): 1274-1293.
[7] Geiger, R. Stuart. "Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space." Information, Communication & Society 19, no. 6 (June 2, 2016): 787–803. doi:10.1080/1369118X.2016.1153700.
[8] Shilton, Katie, Jes A. Koepfler, and Kenneth R. Fleischmann. "How to see values in social computing: methods for studying values dimensions." Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014.

These three factors—perceptions of harassment, flagging, and remediation methods—create a complex network of interactions on social media platforms today. We undertook this study with the hope of producing information that will help designers and platforms better understand consumer attitudes, concerns, and perspectives, so that platforms can be more responsive and reflective of their communities. We make the findings more accessible to the general public by surfacing them on an interactive website that allows visitors to explore and reflect on their own understanding of these issues in comparison to the study findings.

We did not find that different flagging conditions led to any statistically significant differences in harassment level or appropriateness. However, we did find that messages which included the threat or hope for violence toward a certain group—such as feminists, African Americans, and Muslims—regardless of whether directed to a specific individual and without including any indication of prior interaction between a poster and recipient, were found to be at least "Very harassing" on average. Additionally, the most selected remediation option across all conditions was "Remove the message from the platform," which indicates that, while we were easily able to find the tested messages on public social media pages, the vast majority of people think this type of content should not be so available. We theorize that this may have occurred because of the inclusion of violence within our messages, and we recommend further work to explore whether messages lacking violence would be rated differently.

Our research also indicated, although not at a statistically significant level that users are more comfortable with human flagging of harassing content than algorithmic processes. When asked to assess the accuracy, fairness, and trustworthiness of algorithmic, moderator, and user-driven processes, moderator and user scored higher across all metrics. This preference for human moderation carried over into user designs of algorithmic tools with the "Flagged by other humans" feature being selected most often. In addition to indicating a preference for human driven moderator, when a visual flag was present with no source was specified, respondents assumed a human, either moderator or user, had flagged the message.

Finally, we discuss ways in which platforms could incorporate these findings into the creation of their policies and the design of their remediation techniques, and suggest avenues for further research.

## Background

Existing studies have sought to define harassment[9] and quantify the number of Americans who have experienced it,[10] to understand the effects of harassment on recipients,[11] to understand what factors contribute to users engaging in abusive behavior and how the affordances of social networks might

[9] 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.
[10] Pew Center Study: Duggan, Maeve. (Oct, 2014). Online Harassment. Pew Research Center. Retrieved March 16, 2016, from: http://www.pewinternet.org/2014/10/22/online-harassment/.
[11] Citron, Danielle Keats. Hate crimes in cyberspace. Harvard University Press, 2014.

allow users to engage in abusive behaviors who otherwise might not,[12] and to understand the effects of different types of moderation.[13] Matias, et. al., provide an outline of the different types of harassment one might receive online, which highlights the fluid nature of harassment. Hate speech, doxing (release of one's private information), threats of violence, posting false information, impersonation, encouraging others to harass someone else, and revenge porn can all constitute harassment.[14]

**Values-Based Approaches to Understanding Harassment**

There are few studies that seek to surface how social media users' attitudes toward remediation may be influenced by their values. One such study was done by Bowler, Knobel, and Mattern, in which teenagers were asked to create cyberbullying narratives and design remediation proposals to address the type or degree of cyberbullying present in the narrative.[15] A number of values-based design principles were then extracted from the solutions the teens suggested. These included designing for consequence, control and suppression, fear, empathy, reflection, empowerment, and attention. Our study builds on this work and asks adult internet users to make assessments of, and decisions about, examples of potentially harassing content found online. These decisions include a wide range of moderation and remediation options, some of which are based on the principles put forth by Bowler et al., such as inclusion of a remediation option designed to promote reflection by allowing the poster to pause and consider their actions before inflicting harm: "Before publicly posting, notify the poster that it has been flagged as potentially harassing and require confirmation to post."[16]

Our work is also grounded in the social justice orientation proposed by Dombrowski, Harmon, and Fox, who outline "strategies [to] target the goals of social justice along six dimensions: transformation, recognition, reciprocity, enablement, distribution, and accountability."[17] While these dimensions have some overlap with those put forth by Bowler et al., their social justice frame provides additional context. Designing for recognition and reciprocity is embodied by remediation options such as "Post a public reminder of the community code of conduct in response to message," which identifies an unjust action and responds to its perpetrator so as to "engender different forms of participation that could lead to more equitable engagements" in the future.[18] Related to reciprocity, remediation options similar to "Prevent poster from posting again until they have completed an empathy exercise" are designed for enablement in that they encourage the poster to reflect and provide an opportunity "to develop their own capacity."[19] Finally, remediations can be designed for accountability, which aims to hold people responsible for their actions. This dimension is embodied by options such as "Temporarily suspend the poster" or "Permanently ban the poster."

---

[12] Cheng, Justin, et al. "Anyone Can Become a Troll." American Scientist 105.3 (2017): 152.
[13] Grimmelmann, James, The Virtues of Moderation (April 1, 2015). 17 Yale J.L. & Tech. 42 (2015); U of Maryland Legal Studies Research Paper No. 2015-8. Available at SSRN: https://ssrn.com/abstract=2588493
[14] Matias, et al., *supra*.
[15] Bowler, Leanne, Cory Knobel, and Eleanor Mattern. "From cyberbullying to well−being: A narrative−based participatory approach to values−oriented design for social media." Journal of the Association for Information Science and Technology 66.6 (2015): 1274-1293.
[16] See the Values-Based Remediation Options section of the appendix for the full mapping of remediation options to the design principles from Bowler et al., Dombroski et al., and Shilton et al.
[17] Dombrowski, Lynn, Ellie Harmon, and Sarah Fox. "Social Justice-Oriented Interaction Design." Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16 (2016): 662. Web.
[18] *Ibid.*
[19] *Ibid.*

Although the dimensions of distribution and transformation do not directly map to the remediation options we considered, they provide context for understanding the role remediation plays in governing online interactions and the need for remediation options to be flexible. Designing for distribution aims for "the equitable distribution of the benefits and burdens of social systems."[20] As harassment is not evenly distributed, moderation and remediation are means through which the inequity in the benefits and burdens of social media can be addressed. Additionally, designing for transformation highlights the evolving nature of social justice, and therefore the need for remediation options to adjust over time based on social norms of acceptable behavior and to address new issues that arise.

Our study is also grounded in principles of values sensitive design. Our research draws from Shilton, Koepfler, and Fleischmann, who formalize a method for "study[ing] values in social computing."[21] Specifically, they outline six dimensions on the sources and attributes of values. By mapping remediation options to their dimensions of salience, intention, and enactment, we hope to provide context for how these actions could achieve their intended effects. For the actions aligned with salience, the aim is to appeal to users' underlying values and highlight how their messages may be out of alignment with those values. For intention, the aim is to get individuals to reflect on their messages such that they don't accidentally send messages others may find harassing. For enactment, the aim is to prevent harassing messages from being sent, or seen, even if a user sends such a message. We believe that better understanding how remediation options map to value dimensions could help online platforms design remediation options to address specific issues and be more responsive to users' needs regarding harassing content.

**Moral Issues of Delegating Moderation**

Presently, the bulk of the work of content moderation is done by human moderators,[22] but platforms are more and more turning to automated, algorithmic mechanisms to support or take over the work of addressing the problem of harassment at scale. For example, Jigsaw recently released Perspective, an algorithm that attempts to counter abuse by assigning a piece of text a "toxicity" score, which could conceivably be used by a platform to remove messages over a certain threshold.[23] This algorithm was trained through the use of data tagged by CrowdFlower workers, and focuses entirely on the content of the message itself. Shortly after its release, shortcomings in this approach were outlined by various members of the tech community, including ethnographer and machine learning researcher Caroline Sinders[24] and scholar Hossein Hosseini.[25] Sinders points out that the algorithm is trained on a dataset

---

[20] Dombrowski, Lynn, Ellie Harmon, and Sarah Fox. "Social Justice-Oriented Interaction Design." Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16 (2016): 662. Web.

[21] Shilton, Katie, Jes A. Koepfler, and Kenneth R. Fleischmann. "How to see values in social computing: methods for studying values dimensions." Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, 2014.

[22] Chen, Adrian. "The laborers who keep dick pics and beheadings out of your Facebook feed." Wired, October 23 (2014).

[23] http://www.perspectiveapi.com/

[24] Sinders, Caroline. "Toxicity and Tone Are Not The Same Thing: analyzing the new Google API on toxicity, PerspectiveAPI." Medium, 23 Feb 2017, medium.com/@carolinesinders/toxicity-and-tone-are-not-the-same-thing-analyzing-the-new-google-api-on-toxicity-perspectiveapi-14abe4e728b3#.ei8iavabd. Accessed 23 Feb 2017.

[25] Hosseini, Hossein, et al. "Deceiving Google's Perspective API Built for Detecting Toxic Comments." arXiv preprint arXiv:1702.08138 (2017).

of narrow conversational range, as it includes only comments from three news sites and Wikimedia talk pages, which all have specific community standards. It does not include content from social media platforms where conversations may be vastly different in standards, structure, tone, and range of topic.

Additionally, while Jigsaw calls their predictions "toxicity," the output appears to be more akin to tone than toxicity, a concept which is far more complex.[26] The first release of the Jigsaw API gives a toxicity score of 78% to "I hate bananas," while it assigns a toxicity score of only 37% to "We must secure our existence and a future for white children."[27] In his work, Hosseini demonstrated the ease with which adversaries of the system can subtly modify their content to receive a much lower toxicity score, thereby undermining the efficacy and usability of the tool.

This shift to algorithmic moderation raises interesting and important questions about the role technology should take, and the extent to which communities, online and offline, should delegate the work of enforcing its mores and morals to machines. The shift, in essence, further engorges Bruno Latour's "missing mass of morality," previously largely identified as controlling our *actions*, to the area of *words*.[28] While Latour has said that the delegation of moral authority allows us to be "more moral than our predecessors," did he envision the delegation to extend to deciding what is acceptable in public discourse?[29]

Each source of moderation—algorithms, human moderators, and users—has benefits and drawbacks that further complicate the choice of how to delegate moderation activities. Algorithms readily scale to large amounts of data and content, but "inherit the prejudices of prior decision makers" that exist within the data that trained them.[30] Algorithmic recognition of abusive or harassing speech is also extremely difficult: keyword spotting can be foiled by obfuscation, insults change frequently, abuse can cross sentence boundaries, and sarcasm requires contextual knowledge.[31] Human moderators possess the nuanced judgment we expect, but scaling up raises concerns about the mental health risks facing professional content moderators—those who look at thousands of violent, obscene, and harassing messages and images every hour.[32] Tasking the community of users to police itself provides benefits in scalability and community norm-setting, but "there is evidence that strategic, coordinated flagging has

---

[26] Sinders, Caroline. "Toxicity and Tone Are Not The Same Thing: analyzing the new Google API on toxicity, PerspectiveAPI." Medium, 23 Feb 2017, medium.com/@carolinesinders/toxicity-and-tone-are-not-the-same-thing-analyzing-the-new-google-api-on-toxicity-perspectiveapi-14abe4e728b3#.ei8iavabd. Accessed 23 Feb 2017.

[27] According to the Anti-Defamation League, this is the most popular white supremacist slogan in the world. It is commonly referred to as the "14 words." Retrieved 5 May 2017 from: https://www.adl.org/education/references/hate-symbols/14-words

[28] Bruno, Latour. "Where Are the Missing Masses? The Sociology of a Few Mundane Artefacts'." (1992): 225-258.

[29] *Ibid.*

[30] Barocas, Solon and Selbst, Andrew D., Big Data's Disparate Impact (2016). 104 California Law Review 671 (2016). Available at SSRN: https://ssrn.com/abstract=2477899

[31] Nobata, Chikashi, et al. "Abusive language detection in online user content." Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.

[32] Chen, Adrian. "The laborers who keep dick pics and beheadings out of your Facebook feed." Wired, October 23 (2014).

occurred, widely and systematically."[33] That is, organized groups of users can game the system and subvert it into another form of silencing or harassment.

The need for content moderation at all may represent a failure of the legal and normative regulatory constraints described by Lawrence Lessig.[34] Of course the very ability to remove content from an online platform requires an architectural component of regulation, one that is supported legally via Section 230 of the Communications Decency Act. A movement from human moderation to algorithmic moderation diverts regulatory power from norms to architecture. Roger Brownsword also identifies the effects of moving from a normative to non-normative regulatory regime and questions whether such a move may rob us of our ability to make normative decisions in other contexts.[35] His work raises important questions applicable to algorithmic moderation, especially how an algorithmic moderator might respond to evolutions of human morality and community standards within social networks.

With these issues in mind, we included an experimental condition in which a visual flag indicated that one of four flagging sources had identified the piece of content as potentially harassing. We sought to determine whether the indicator's presence influenced participants' assessment of harassment as compared to the absence of a flag or a flag from a human or humans. We included several questions in our survey to understand the comfort level our participants had with having an algorithm moderating content on a social media platform. Finally, we asked them what factors they would want such an algorithm to take into account.

**Legal Issues of Moderation**

Under Section 230 of the Communications Decency Act, platforms in the United States can make these moderation and delegation choices behind the scenes, without any requirements to publicize any methodology or details of the process—regardless of whether the speech would be Constitutionally protected.[36] As Crawford and Gillespie note, "The opacity of the [flagging] process means that the site is not obligated to honor the flags it does receive, and that any decision to remove content can be legitimized as being in response to complaints from the community."[37] The humans involved in this process, who "are exposed to heinous examples of abuse, violence, and material that may sicken others," are typically prohibited under nondisclosure agreements from speaking about their employment duties.[38] This lack of transparency at multiple levels raises questions about the role of free speech—both of and on platforms. As private corporations, platforms can legally make these moderation choices, yet "complex, politically charged decisions to keep or remove content"[39] are made

---

[33] Crawford, Kate and Gillespie, Tarleton L., What is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint (August 5, 2014). New Media & Society, 2014. doi: 10.1177/1461444814543163 . Available at SSRN: https://ssrn.com/abstract=2476464

[34] Lawrence, Lessig. "Code: Version 2.0." New York (2006). www.socialtext.net/codev2/what_things_regulate

[35] Brownsword, Roger. "Lost in translation: Legality, regulatory margins, and technological management." Berkeley Technology Law Journal 26.3 (2011): 1321-1365.

[36] 47 U.S. Code § 230

[37] Crawford, Kate and Gillespie, Tarleton L., What is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint (August 5, 2014). New Media & Society, 2014. doi: 10.1177/1461444814543163 . Available at SSRN: https://ssrn.com/abstract=2476464

[38] Chen, Adrian. "The laborers who keep dick pics and beheadings out of your Facebook feed." Wired, October 23 (2014).

[39] Roberts, Sarah T. "Social Media's Silent Filter." The Atlantic, March 8 (2017).

without public input or reflection. Indeed, “A flag-and-delete policy obscures or eradicates any evidence that the conflict ever existed.”[40] In this way, “the logic for the existence of the material on a site becomes opaque and the content becomes normalized.”[41] As Crawford and Gillespie conclude, “without any public record or space for contestation, [flags] leave us little chance to defend the right for those things to be said, or to rally against something so egregious that it warrants more than a quiet deletion.”[42]

**Encouraging Reflection**

Two instruments that inspired the original concept and design for our values-intervention reflection site were The New York Times’ comment moderation test[43] and MIT’s Moral Machine.[44] In its test, the Times asks users to make decisions about whether to approve or reject comments based on the NYT’s moderation policy. For our tool, we want to elicit values by exploring a wider range of decision options than simply taking down or leaving up a comment, and consider how those decisions reflect values that are, could, or should be explicitly designed for when developing a tool that allows users to post messages. For our reflection site, we intended to show users how their decisions compared to those of others who had used the tool, which could help them better understand the extent to which their values are shared or unique. For this, we drew from the concept of the Moral Machine, which has similar functionality. However, concerns about data security and confidentiality raised by UB Berkeley’s Committee for the Protection of Human Subjects drove us to bifurcate this process. In the end, we collected data through a Qualtrics survey and feed that data into visualizations in our web experience that users interact with after assessing content themselves. We do not collect any research data from users of the web experience.

## Research Questions

We began with five primary research questions:

RQ1: To what degree do sociodemographic and personal characteristics influence an individual’s assessment of the severity of potentially harassing or abusive messages?

RQ2(a): To what degree does the presence of a visual element indicating a message has been flagged as potentially harassing or abusive, by either a human or an algorithm, influence an individual’s assessment of severity?

RQ2(b): How does an individual’s perceptions about the fairness, accuracy, and trustworthiness of human moderators, other users, and algorithms influence their assessment of harassment when flags are present, as well as their choice of remediation options?

---

[40] Crawford and Gillespie, *supra*.
[41] Roberts, Sarah T. "Commercial Content Moderation: Digital Laborers' Dirty Work." (2016).
[42] Crawford and Gillespie, *supra*.
[43] Etim, Bassey. “Approve or Reject: Can You Moderate Five New York Times Comments?” The New York Times, 23 Feb 2017. https://www.nytimes.com/interactive/2016/09/20/insider/approve-or-reject-moderation-quiz.html
[44] Scalable Cooperation. "Moral Machine." Moral Machine. MIT Media Lab, n.d. Web. 06 Mar. 2017. http://http://moralmachine.mit.edu/

RQ3: Do individuals make different choices regarding appropriate remediation action(s) in response to potentially abusive or harassing messages based on sociodemographic and personal characteristics and/or the presence, absence, or type of flagging context surrounding a message?

RQ4: Are people receptive to a broader range of potential moderation and remediation mechanisms than currently exist on public platforms for responding to potentially abusive or harassing messages?

## Methods

### Content Selection

The messages we chose for the survey were found on public social media platforms, either currently available or from screenshots that other internet users had saved and referenced online. Originally our selection of messages had a large number of what we'd consider "gray area" messages—that is, those that could potentially be harassing but could just as easily go the other way. We coded each of these messages for targeted characteristic, directedness, and presence of threat of or hope for violence. As we revised our survey and constrained the number of variables, we realized the need to hold some of these coded elements constant across our messages. Eventually, we settled on three messages containing some element of violence to keep them at similar harassment levels (as we perceived them) in order to ensure internal validity in our experiment. In a future survey we would make different choices in this regard due to the lack of spread among participants' assessments of harassment.

Participants were also asked to rate additional messages which were used within the reflection site visualizations but were not included in the experiment or reported with our results. The independent variables were not held constant for these additional messages. These messages targeted homosexuals, Mexicans, Jews, transgender people, and Nazis.

### Instrument

Data for the study were collected using a multi-part survey instrument we developed[45] in Qualtrics, a platform for designing and hosting online research. Each participant was shown a piece of content and then asked to respond to questions about harassment level (5-point scale), appropriateness (7-point scale), and frequency of seeing similar posts (5-point scale). The participant was then asked to choose the top three remediation options they would prefer be applied to the post by the platform. To test our hypotheses about the effects of directedness and presence/source of flag, we presented each participant with one piece of content that was either directed or undirected, and either did not have a flag at all (control), or had a flag indicator conveying the content had been flagged as potentially harassing (by an unspecified source, by an algorithm, by a human moderator, or by other users). Participants in the ambiguous source condition were also asked what they perceived the source of the flag to be. The list of content and flags is provided in the Flag Conditions section of the appendix.

After the content assessment section, participants were asked to rate the effectiveness of all of the remediation options in preventing future posts similar to the one they saw, and to rate the severity of all of the remediation options. Next they answered Likert-type questions about the fairness, accuracy, and trustworthiness of algorithms, human moderators, and other users in identifying harassing content,

[45] CPHS Approval #2017-02-9647, PI Deirdre Mulligan

before selecting the top three features they would include in an algorithmic tool of this nature. For all participants, we collected basic demographic data, basic social media usage, whether or not they felt they had personally experienced online harassment, and whether they thought anything they'd done online could be perceived as having been abusive or harassing.

### Survey Design and Pilots

We conducted a pilot of the survey with a convenience sample of 41 people, mostly Berkeley students. We also conducted talk-aloud walkthroughs with four undergraduate students to get their reactions to the survey flow, questions, and timing. Based on those talk-alouds, we updated the survey by removing or clarifying some questions, and adding others that we felt we were still missing.

We also conducted face validity sessions with other students, to validate what assumptions people were making and how they interpreted the sources of flag indicators. For the ambiguous source flag, in each case, the participants interpreted the flag as being placed by a human, either a moderator or user of the platform. Based on this, we considered removing the ambiguous source condition, but determined that how users interpret the source of that flag is in itself an interesting question, which may change over time. These interviews also resulted in the consideration of some small changes to the design of the flagged messages. We considered including a visual indicator of the source of the flag to aid in quick comprehension by participants, such as including a computer inside the flag icon for the algorithmic flag, a person icon inside the flag for the human moderator flag, and multiple people icons inside the flag icon for the other users flag. We ultimately decided against this out of concern that it would add unnecessary additional variables to our experiment. We did however move the source of the flag to the beginning of the sentence based on insights from these interviews, as in "A human moderator has flagged this content as potentially harassing."

The algorithmic flag was interpreted to have been placed by a third party app installed by the platform where the algorithm was based on the presence of certain words. One user felt that the word "algorithm" was not accessible, but we determined that it is common enough in the public conversation to be included. No other option, such as "computational processes" or "systems" quite captured the intended meaning. The human moderator flag was interpreted to have been placed by a moderator employed by the platform and trained to enforce the rules of that platform. The other user flag was interpreted to have been placed by other, untrained users who were not in positions of authority and who were not necessarily placing the flag based on the rules of the platform.

### Question Vocabulary

We chose the specific wording of questions as the result of our research team's discussions of how to best capture the concepts we wished to measure, and validated them using qualitative comments from our pilot survey, as well as in-person user testing.

One question, "How abusive is this social media post?" was removed from the survey from the pilot when it became clear—as we had suspected it might—that users were unclear on the meaning of the word in this context. One user left the comment, "For questions using the word abusive, does this mean abusive towards a person or abusive towards social media," which solidified the decision to remove.

Another major vocabulary problem surrounded the word severe in the question “In general, how severe do you think the following remediation options are?” With this question we were trying to ascertain whether participants’ opinions of the various remediation options would create a scale by which platforms could escalate the handling of a piece of content based on its level of harassment. However, capturing this sense proved difficult. Other terms we considered were: “serious,” “intense,” “tough,” “strong,” “consequential,” and “to what extent does this punish?” We opted for “severe” after users in an in-person qualitative user testing exercise confirmed its applicability. After using “severe” in our final instrument, we received a respondent comment asking “I want to say that the previous questions were confusing to me. You asked how "severe" I thought remediation options were. By ‘severe,’ did you mean ‘effective’ (from the public's point of view) or ‘harsh’ (from the posters)?” Based on this response, we now think that “harsh” would perhaps have been a good alternative.

### Survey Participants, Sampling, and Demographics

We sought English-speaking, adult internet users for our study. We opted to purchase our sample from Qualtrics for timing and convenience. Our total sample size was 1089 participants, assigned to each of the experimental conditions to reach N > 30 for each condition. All respondents completed the survey within two days of each other. The demographic makeup of the survey respondents differs substantially from the U.S. population in terms of gender, but only slightly in terms of race. Given this, we explored weighting the data to correct for the gender disparity, but found that, while the weighting led to a slight difference in the overall means, it did not affect statistical tests between the experimental conditions. As such, we have not weighted the data included here. See Table 0 for a breakdown of participant demographics.

## Table 0: Respondent Demographics

| **Gender Identity** | **Response Count** | **Percent of Total** |
|---|---|---|
| Female | 663 | 61.7% |
| Male | 408 | 38.0% |
| Non-Conforming | 3 | 0.3% |
| *Total* | *1074* | |
| | | |
| **Racial Identity** | | |
| White | 869 | 78.9% |
| Black or African-American | 124 | 11.3% |
| Hispanic, Latino, or Spanish origin | 61 | 5.5% |
| Asian | 49 | 4.5% |
| American Indian or Alaska Native | 16 | 1.5% |
| Other | 13 | 1.2% |
| Native Hawaiian or Other Pacific | 3 | 0.3% |
| Middle Eastern or North African | 1 | 0.1% |
| *Total* | *1101* | |
| | | |
| **Sexual Identity** | | |
| Heterosexual or straight | 965 | 91.3% |
| Bisexual | 39 | 3.7% |
| Homosexual, Gay, or Lesbian | 33 | 3.1% |
| Something else | 11 | 1.0% |
| I don't know | 9 | 0.9% |
| *Total* | *1057* | |
| | | |
| **Age** | | |
| 18 - 24 | 80 | 7.3% |
| 25-34 | 232 | 21.2% |
| 35-44 | 219 | 20.0% |
| 55-64 | 198 | 18.1% |
| 65+ | 145 | 13.2% |
| 45-54 | 222 | 20.3% |
| *Total* | *1096* | |

*Note: Due to the nature of the content in the research instrument, all demographic questions were optional. As such, the total counts will not necessarily match across demographics. Additionally, respondents were able to select multiple options when self-identifying their race. Due to the possibility of multiple selections, the total here exceeds the total sample size.*

## Variables

### Dependent Variables

**Level of Harassment**

After seeing a randomly selected message, participants were asked to rate the perceived level of harassment of the message. At its most basic, this survey was intended to set a baseline for the general public's perception of what constitutes online harassment, thus we needed to determine whether or not a piece of content was perceived as harassing to each participant. This dependent variable was measured on a five-point Likert scale: (1) Not at all harassing, (2) Slightly harassing, (3) Somewhat harassing, (4) Very harassing, and (5) Extremely harassing.

**Appropriateness**

After rating the level of harassment, participants were asked to rate the level of appropriateness or inappropriateness of the message for a public social media platform. We chose "appropriateness" in case participants wanted to highlight differences between messages that they might or might not view as harassment, but might feel differently about whether they belonged on social media. Because this was a binary scale, this dependent variable was measured on a seven-point Likert scale: (1) Absolutely appropriate, (2) Appropriate, (3) Slightly appropriate, (4) Neutral, (5) Slightly inappropriate, (6) Inappropriate, and (7) Absolutely inappropriate.

**Frequency**

Having assessed the message, participants were then asked how often they come across similar content. While we did not have any specific hypotheses on how frequency of exposure would relate to the assessed levels of harassment or appropriateness, we wanted to capture this information for exploratory purposes. As we are most interested in perceived frequency of exposure, this dependent variable was measured on a five-point Likert scale: (1) Never, (2) Rarely, (3) Occasionally, (4) A moderate amount, and (5) A great deal.

**Preferred Method(s) of Remediation**

Participants were also asked to select the remediation options that they would choose to apply to the given social media message they were shown. This consisted of a list of nine options, from which participants were invited to select up to three options. We offered the possibility for up to three options, recognizing that in some cases there might be more than one option that seems appropriate and effective. Some of these options were actions that are widely adopted by various social media platforms, and some were ideas derived from our values-based research.

**Remediation Effectiveness**

For each of the nine remediation options, plus a free text "Other" option wherein participants could offer us their own suggestion, we asked them to rate how effective they thought each method was to discourage that type of content on social media platforms. We did not ask them specifically about the message they had seen earlier. This dependent variable was measured on a five-point Likert scale: (1) Not at all effective, (2) Slightly effective, (3) Somewhat effective, (4) Very effective, and (5) Extremely effective.

**Remediation Severity**
For each of the nine remediation options, plus the same "Other" option if it had been filled out previously, we asked participants to rate how severe they thought each remediation method was. We did not ask them specifically about the message they had seen earlier. This dependent variable was measured on a five-point Likert scale from (1) Not at all severe to (5) Extremely severe.

**Fairness**
Participants were asked to rate each of the three main flagging agents (an algorithm, a human moderator, and other users) according to how fair they perceived each agent to be when flagging content for potential remediation. This variable was measured on a five-point Likert scale: (1) Unfair, (2) Somewhat unfair, (3) Neither unfair nor fair, (4) Somewhat fair, and (5) Fair.

**Accuracy**
Participants were asked to rate each of the three main flagging agents according to how accurate they perceived each agent to be when flagging content for potential remediation. This variable was measured on a five-point Likert scale: (1) Inaccurate, (2) Somewhat inaccurate, (3) Neither inaccurate nor accurate, (4) Somewhat accurate, and (5) Accurate.

**Trustworthiness**
Participants were then asked to rate each of the three main flagging agents according to how trustworthy they perceived each agent to be when flagging content for potential remediation. This variable was measured on a five-point Likert scale: (1) Untrustworthy, (2) Somewhat untrustworthy, (3) Neither untrustworthy nor trustworthy, (4) Somewhat trustworthy, and (5) Trustworthy.

### Independent Variables

**Directedness**
Messages were either directed toward a specific user (indicated by blue highlighted "@User" within the message text), or undirected.

**Targeted Characteristic**
Messages in the experiment either targeted feminists, African Americans, or Muslims. We chose feminists as a stand-in for females, who are widely reported to experience the highest levels of harassment online. We further chose African Americans and Muslims, as they are populations that are anecdotally receiving high levels of harassment in our current political and social climate.

**Flag Indicator**
Messages were identified as being flagged by human moderators, by other users, by an algorithm, or by an ambiguous source. Text describing the flagging agent (an algorithm, human moderator, other users, or an ambiguous/non-specified source) was added in red text to the footer of the message, along with a red flag icon. Messages in a control group for this variable did not display the visual flag and text.

**Demographics**

We also collected demographics on our participants and expected these demographics to have some effect on the results of the dependent variables. We collected: race, sexual orientation, age, gender, social media usage, most frequently used social media platform, and prior harassment experience.

## Hypotheses

Because harassment typically has a personal element to it, we hypothesized that messages that were directed at a particular user would be rated as more harassing than those that were not aimed at anyone in particular.

Because flagging indicates agreement between different parties on what constitutes questionable content, we hypothesized that messages carrying flags would be rated as more harassing than those without flags. We also thought that messages flagged by “a human moderator” and by “other users” would be rated as more harassing than the algorithmic or generic flagging conditions.

Assuming that participants would want the punishment to match the crime, as it were, we hypothesized that the harassment level of the post and the rated severity of the preferred remediation options would correlate positively.

Assuming that a member of an in group is more likely to recognize and want to act on harassment aimed at their particular group or demographic, we hypothesized that participants who identified as belonging to the targeted group in the potentially harassing message would rate that message more highly than someone who was not a member of the group.

Additionally, we hypothesized that participants who rated a particular moderation agent (human moderator, other users, or an algorithm) as most trustworthy, fair, or accurate, and also rated a message containing the corresponding flag, would rate that message as more harassing.

We had additional hypotheses dealing with a participant’s prior experiences of harassment and their ratings of a message’s harassment level or their sense of effectiveness or severity of a particular remediation option. These hypotheses are available in the detailed list in the Enumerated Research Hypotheses section of the appendix.

## Results

We did not consistently find any statistically significant differences in the means for harassment level and appropriateness when testing our directedness and flagging hypotheses. We also did not find consistent statistically significant differences when comparing the responses from respondents identifying as female and male. While we saw statistical significance when testing a few of the experimental conditions, this appears to be an artifact of the number of conditions rather than a true finding. See Table 1 for the key statistical results for harassment level.

## Table 1: Key Statistical Results, Harassment Level

| Condition | Flag | Target | t-stat | p-value |
|---|---|---|---|---|
| Directed / Undirected | - | - | 1.092 | 0.275 |
| | - | African Americans | -0.727 | 0.468 |
| | - | Feminists | 0.350 | 0.727 |
| | - | Muslims | 2.295 | 0.022 |
| | No Flag | - | 1.053 | 0.294 |
| | Generic | - | -0.020 | 0.984 |
| | Algorithm | - | -0.461 | 0.645 |
| | Moderators | - | 1.872 | 0.063 |
| | Users | - | 0.079 | 0.937 |

| Condition | Directedness | Target | t-stat | p-value |
|---|---|---|---|---|
| No Flag / Generic | - | - | 1.412 | 0.159 |
| | Directed | - | 1.554 | 0.122 |
| | Undirected | - | 0.445 | 0.657 |
| | - | African Americans | 0.620 | 0.536 |
| | - | Feminists | 0.410 | 0.682 |
| | - | Muslims | 1.622 | 0.107 |
| No Flag / Algorithm | - | - | 0.824 | 0.410 |
| | Directed | - | 1.389 | 0.166 |
| | Undirected | - | -0.179 | 0.858 |
| | - | African Americans | 1.513 | 0.133 |
| | - | Feminists | 0.138 | 0.890 |
| | - | Muslims | -0.082 | 0.935 |
| No Flag / Moderators | - | - | 0.362 | 0.718 |
| | Directed | - | -0.159 | 0.874 |
| | Undirected | - | 0.578 | 0.564 |
| | - | African Americans | 1.506 | 0.134 |
| | - | Feminists | -0.205 | 0.838 |
| | - | Muslims | -0.617 | 0.538 |
| No Flag / Users | - | - | 1.067 | 0.287 |
| | Directed | - | 1.221 | 0.223 |
| | Undirected | - | 0.265 | 0.792 |
| | - | African Americans | 0.945 | 0.347 |
| | - | Feminists | -0.066 | 0.948 |
| | - | Muslims | 0.967 | 0.335 |

| Independent Variable | Directedness | Flag | t-stat | p-value |
|---|---|---|---|---|
| Male / Female | - | - | -1.673 | 0.103 |
| | Directed | - | -0.880 | 0.391 |
| | Undirected | - | -1.568 | 0.136 |
| | - | No Flag | 0.033 | 0.975 |
| | - | Generic | -0.659 | 0.543 |
| | - | Algorithm | -10.402 | 0.000 |
| | | Moderators | -1.698 | 0.117 |
| | | Users | -0.714 | 0.506 |

| Statistical Significance: | $p < 0.1$ | $p < 0.05$ |
|---|---|---|

### Harassment Level

Harassment level was assessed on a 5-point scale from (1) “Not at all harassing” to (5) “Extremely harassing.” The mean harassment level across all conditions was 4.20, and the response distribution is in Chart 1 below. While not statistically significant, respondents that received the “No Flag” condition assessed the post as being more harassing than those that received any condition with a flag. As we had hypothesized flagged content would be assessed as more harassing, better understanding this difference warrants further research.

### Appropriateness

Appropriateness was assess on a 7-point scale from (1) “Absolutely appropriate” to (7) “Absolutely inappropriate.” The mean appropriateness across all conditions was 6.14, and the response distribution is in Chart 2 below. Similarly to harassment level, the “No Flag” condition was assessed as being less appropriate than the other flag conditions.

Chart 1: Harassment Level Response Distribution

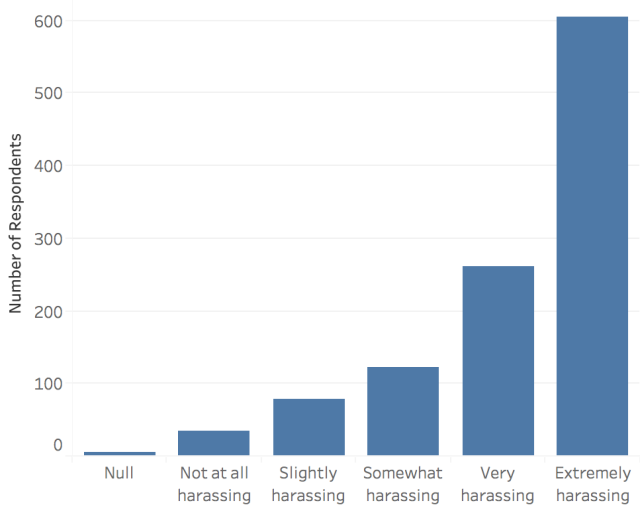
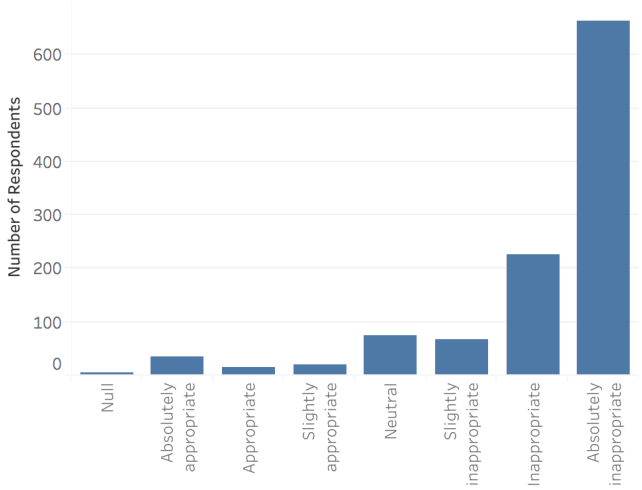



Chart 2: Appropriateness Response Distribution




### Frequency

While not a dependent variable, we captured the perceived frequency with which respondents come across content similar to that in the instrument and have included this in our tables as a check to ensure the conditions were not influenced by this frequency.

*Perceived Frequency: Respondent Percentage*

| Never | Rarely | Occasionally | A moderate amount | A great deal |
|---|---|---|---|---|
| 27% | 31% | 26% | 10% | 6% |

*See Tables 2, 3, and 4 on the next page for breakdowns of the specific conditions.*

## Table 2: Directedness

***Descriptive Results***

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***Directed*** | Mean | 4.24 | 6.15 | 2.36 |
| | N | 552 | 551 | 553 |
| | Std. Deviation | 1.070 | 1.475 | 1.197 |
| ***Undirected*** | Mean | 4.17 | 6.12 | 2.42 |
| | N | 549 | 551 | 552 |
| | Std. Deviation | 1.105 | 1.422 | 1.134 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1101 | 1102 | 1105 |
| | Std. Deviation | 1.087 | 1.448 | 1.166 |

## Table 3: Flags

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***Algorithm*** | Mean | 4.19 | 6.06 | 2.47 |
| | N | 221 | 220 | 221 |
| | Std. Deviation | 1.105 | 1.516 | 1.263 |
| ***Generic*** | Mean | 4.13 | 6.11 | 2.35 |
| | N | 219 | 219 | 220 |
| | Std. Deviation | 1.128 | 1.400 | 1.134 |
| ***Moderators*** | Mean | 4.24 | 6.19 | 2.41 |
| | N | 226 | 226 | 226 |
| | Std. Deviation | 1.036 | 1.439 | 1.163 |
| ***No Flag*** | Mean | 4.28 | 6.22 | 2.30 |
| | N | 215 | 216 | 217 |
| | Std. Deviation | 1.035 | 1.372 | 1.134 |
| ***Users*** | Mean | 4.17 | 6.11 | 2.41 |
| | N | 220 | 221 | 221 |
| | Std. Deviation | 1.132 | 1.514 | 1.131 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1101 | 1102 | 1105 |
| | Std. Deviation | 1.087 | 1.448 | 1.166 |

## Table 4: Target

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***African Americans*** | Mean | 4.49 | 6.34 | 2.04 |
| | N | 367 | 369 | 370 |
| | Std. Deviation | 0.969 | 1.392 | 1.124 |
| ***Feminists*** | Mean | 3.99 | 5.96 | 2.65 |
| | N | 372 | 371 | 372 |
| | Std. Deviation | 1.182 | 1.549 | 1.112 |
| ***Muslims*** | Mean | 4.13 | 6.11 | 2.48 |
| | N | 362 | 362 | 363 |
| | Std. Deviation | 1.039 | 1.373 | 1.178 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1101 | 1102 | 1105 |
| | Std. Deviation | 1.087 | 1.448 | 1.166 |

## Discussion of Findings

### Remediation Preferences

Of all the remediation techniques, respondents selected “Remove the message from the platform” the most at 575 times, followed by “Permanently ban the poster” at 445 times. While we did not find statistically significant differences in the harassment level and appropriateness assessment between the “No Flag” condition and the “Flagged” conditions (generic, algorithmic, moderator, user), we did observe that respondents in the “No Flag” condition selected “Remove the message from the platform” at a greater rate than the other conditions. Although this observation is also not statistically significant, it warrants further investigation, as it indicates providing users with additional information may increase their comfort level when applying remediation. This  See Chart 3 for a breakdown of the selected remediation options.

Respondents assessed efficacy on a five point scale from (1) “Not at all effective” to (5) “Extremely effective.” The top three respondent  ratings for the efficacy of the remediation options in discouraging people from posting the type of content in the research instrument were: “Permanently ban the poster” (3.8), “Remove the message from the platform” (3.6), and “Add the poster to a block-list which hides their posts from all people who subscribe to that block-list” (3.3). See Chart 4 for a breakdown of all the remediation options.

Respondents assessed severity on a 5 point scale from (1) “Not at all severe” to (5) “Extremely Severe.” The top three respondent ratings were: “Permanently ban the poster” (3.6), “Add the poster to a block-list which hides their posts from all people who subscribe to that block-list” (2.9), and “Remove the message from the platform” (2.9). See Chart 5 for a breakdown of all the remediation options.

These findings do not support our hypothesis that users may desire a broader range of remediation options beyond the commonly used punitive measures of banning posters and removing content. Instead, many users requested in qualitative feedback harsher punitive options be considered, such as a user ban that carries across different platforms or that platforms impose monetary penalties. Several users responded that the options we provided were “not nearly severe enough” with one user writing “most of your solutions would be under that heading (not nearly severe enough)”. A commonly requested additional remediation option was for platforms to provide an easy method for users to notify law enforcement of the harassment offense.

We did, however, have several participants respond qualitatively that the poster of the content they saw seemed to be in need of counseling or mental health support. In future research, it would be interesting to explore whether users would be more supportive of remediation options framed as providing such support to a poster in place of the empathy exercise remediation option, which was rarely preferred and commonly viewed as not very effective with an average rating of 2.86.

Chart 3: Remediation Option Selection Rate

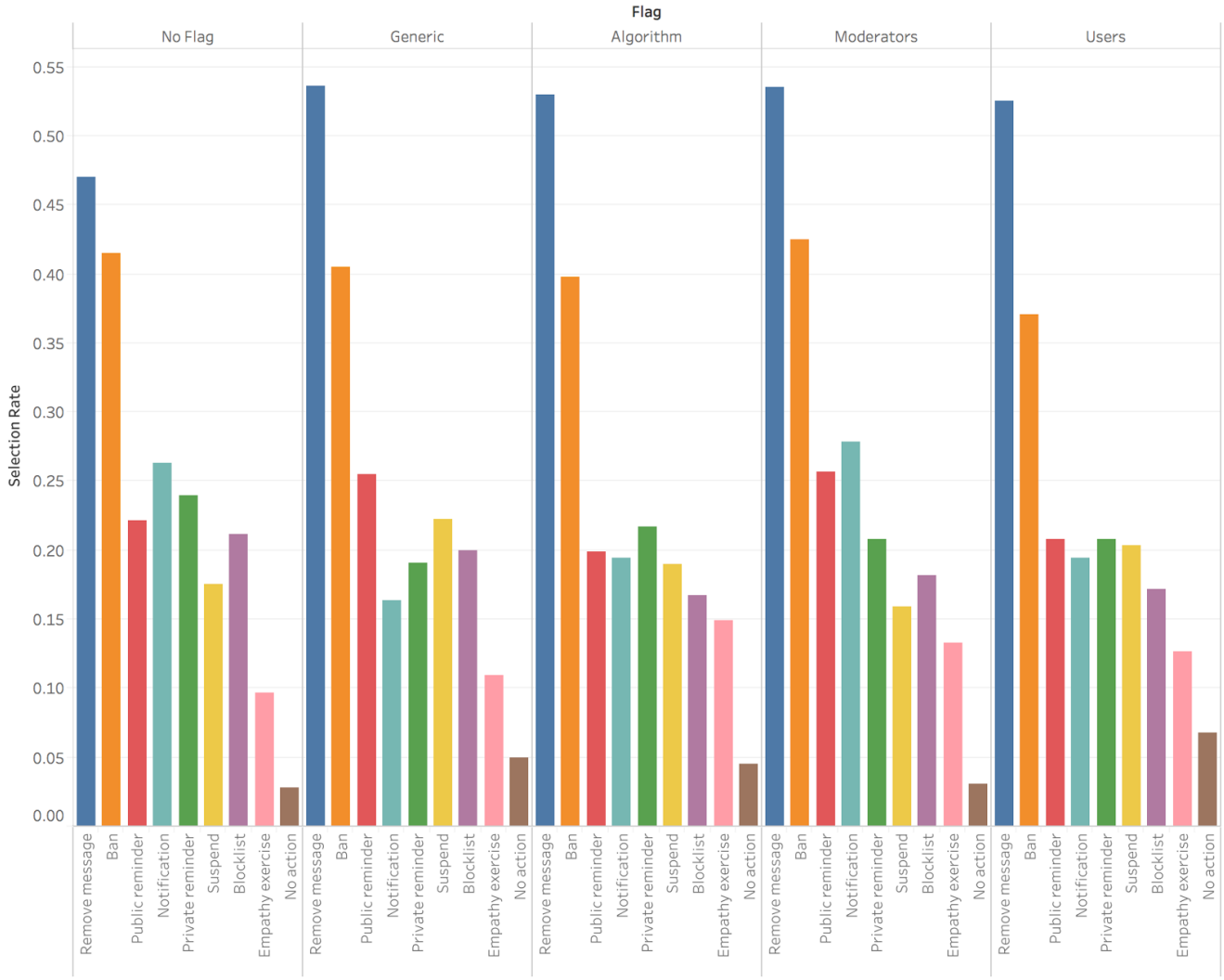



Chart 4: Average Remediation Efficacy Rating

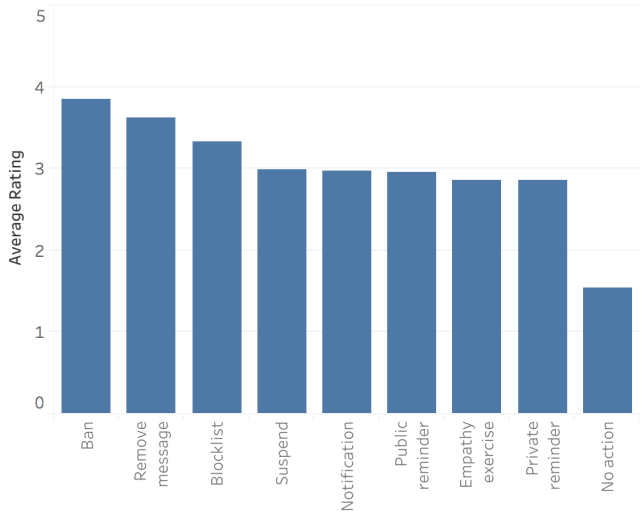
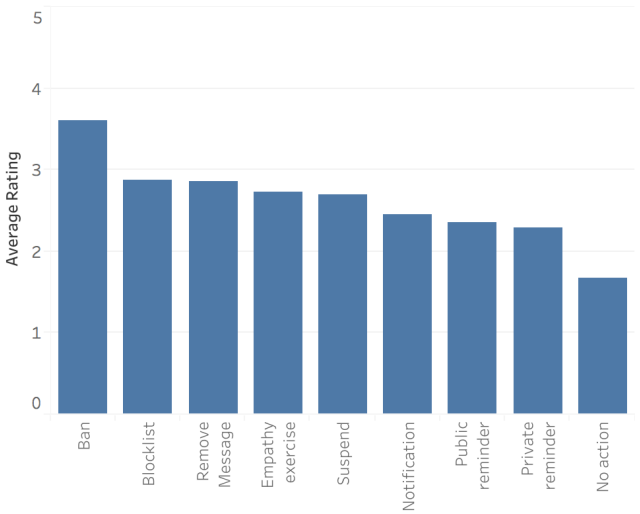



Chart 5: Average Remediation Severity Rating



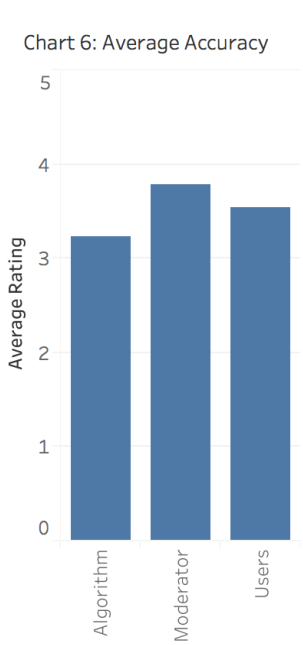
## Moderation Source

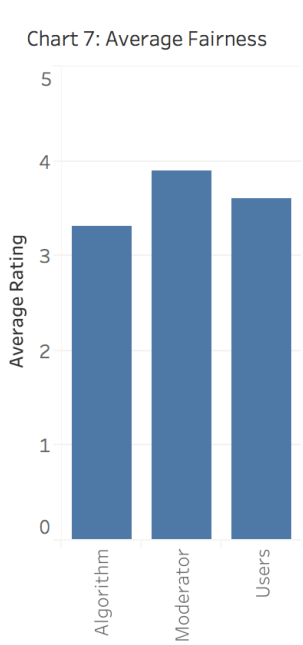
**Content with no flag was assessed as more harassing**

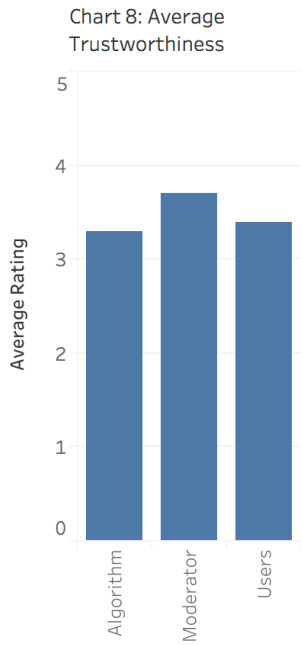
One interesting finding is that, while we hypothesized flagging would increase respondents' harassment level and appropriateness assessments, the data, although not statistically significant, consistently indicated the opposite: content with no flag was assessed as more harassing, and less appropriate, than the same content with any flag condition. This could be an opportunity for future qualitative research to understand the mechanism through which this is occurring: *Are users less concerned about content when they know it has already been flagged? Do users assume that the content and poster is already being remediated in some way when it has been flagged? Or are users suspicious or concerned that platforms are being overly censorial with their use of flagging and/or content moderations?*

**Users are more comfortable with human flagging of harassing content than algorithmic**

Across accuracy, fairness, and trustworthiness, human moderators scored the highest, followed by other users, and then an algorithm; this held true across all of the visual flag conditions. While no explanation was provided to participants as to how any of the three sources had made the determination to flag the message, recent scholarship has suggested that disclosure of how an algorithm reached its decision can increase human trust in that decision.[46] [47] This could be an area for future research, to explore whether this finding about transparency increasing trust applies in the context of content moderations and in head-to-head comparison. Additionally, longitudinal research to understand how these perceptions may change over time as people become more or less comfortable with delegation to algorithms.



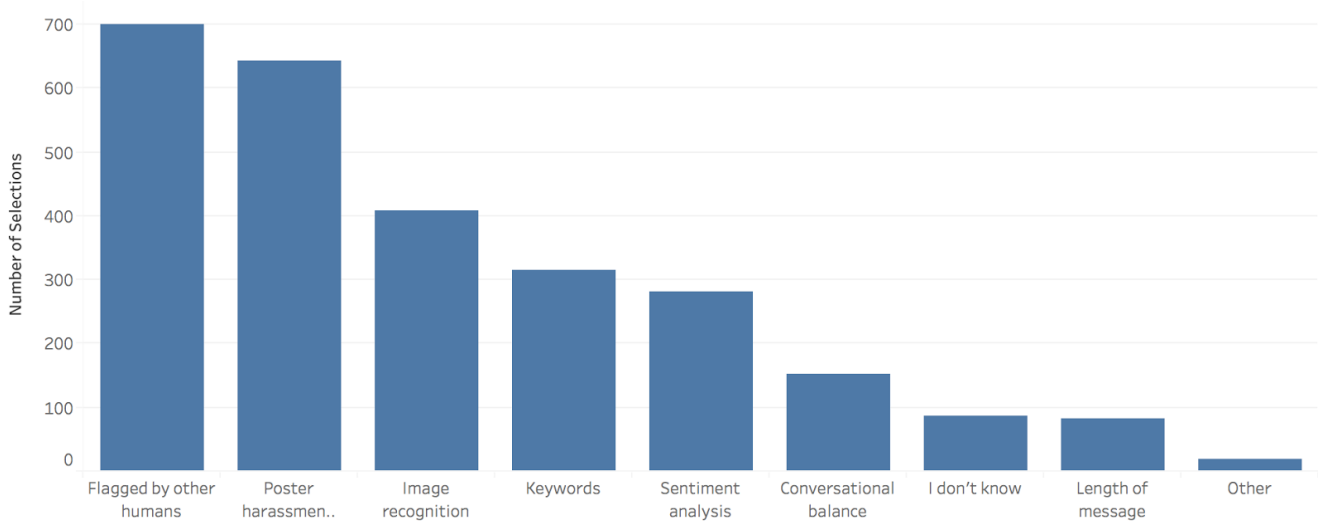




[46] Kizilcec, René F. "How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016.
[47] Wang, Ning, David V. Pynadath, and Susan G. Hill. "Trust calibration within a human-robot team: Comparing automatically generated explanations." *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016.

**Preference for human moderation carried over into user designs of algorithmic tool**

When identifying the criteria that would be most useful in detecting harassing posts online, “Flagged by other humans” was selected the most at 701 times, which aligns with our results across perceptions of accuracy, fairness, and trustworthiness. This was followed by “Poster harassment history” at 643 times, and “Image recognition - does the message contain any images that are commonly recognized as harassing?” at 408 times. We found it interesting that a large number of respondents selected image recognition, perhaps indicating that they have come across offensive or harassing images quite often, though we didn’t explicitly ask participants about this. Surprisingly, “Keywords” and “Sentiment Analysis” were chosen fourth and fifth highest, respectively, despite being commonly assumed to be the mechanism by which the algorithmic flag was operating in both our face validity interviews and in the qualitative responses in our survey. This may indicate respondents are unimpressed by the usefulness of the algorithmic tools they currently interact with or hear about. If this is the case, further exploration could provide deeper context for our finding that users perceive algorithms to be less trustworthy, fair, and accurate than humans.

Chart 9: Algorithmic Features Response Distribution




**Users are more concerned about harassing content not being flagged than non-harassing content being flagged**

When considering what outcome was most important to prevent when flagging content as potentially harassing, respondents selected “Harassing content not being flagged as harassing and not subjected to remediation” as being most important 623 times, while “Non-harassing content being flagged as harassing and potentially subjected to remediation” was selected 269 times, and “No preference” was selected 215 times. This finding was surprising to us as anecdotal evidence led us to expect a much larger proportion of respondents would cite free speech as a reason for keeping content up online, even if it were potentially harassing.

**When no source was indicated, users assumed the source to be human**

Participants in the ambiguous flag experimental condition were asked what they assumed to be the source of the flag. “A human moderator” and “Another user or users” were each selected 72 times, and “An algorithm” was selected 38 times. This result was consistent with the pre-survey face validity

interviews we conducted. Similarly to fairness, trustworthiness, and accuracy, this could be an area for longitudinal research. We posit that user assumptions may change over time as platforms expand their use of algorithmic tools for moderation, unless the use of algorithmic tools is not made visible to users. As survey respondents placed greater trust in human moderation, we advise platforms to be forthright about their use of algorithms in these decisions, lest they be faced with consumer backlash and loss of confidence.

Chart 10: False Positives and Negatives

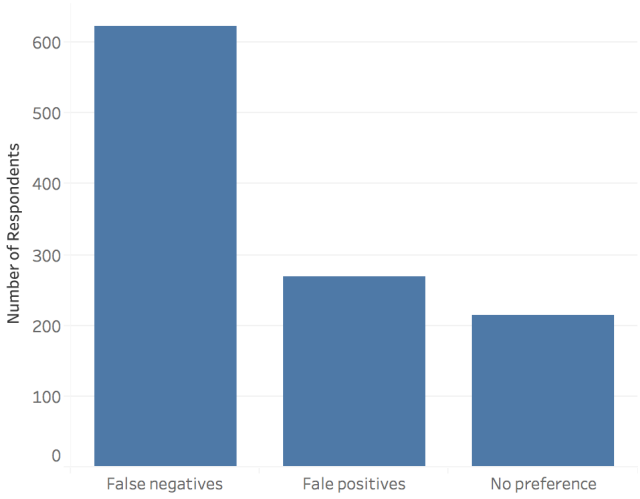
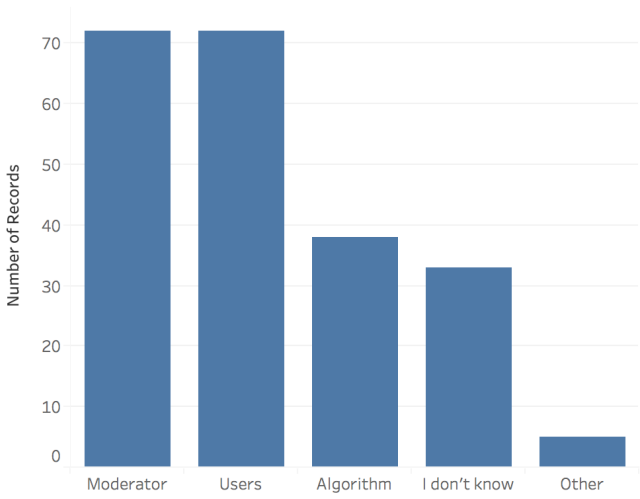



Chart 11: Assumed Flag Source




### Harassment Experience

When asked if they had personally experienced online harassment, respondents who responded "Definitely no" and "Definitely yes" had higher average harassment level assessments that those that responded otherwise. When comparing these means to those who responded "Might or might not have," the difference is statistically significant ($p < 0.01$). This finding was unexpected and warrants further research. See tables 5 and 6 for details.

**Table 5: Personal Experience of Online Harassment**

*Do you think you've ever personally experienced online harassment?*

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| **Definitely not** | Mean | 4.35 | 6.40 | 1.85 |
| | N | 352 | 351 | 353 |
| | Std. Deviation | 1.051 | 1.254 | 1.000 |
| **Probably not** | Mean | 4.14 | 6.18 | 2.35 |
| | N | 261 | 261 | 262 |
| | Std. Deviation | 1.037 | 1.355 | 0.943 |
| **Might or might not have** | Mean | 3.98 | 5.94 | 2.44 |
| | N | 182 | 182 | 182 |
| | Std. Deviation | 1.198 | 1.476 | 0.994 |
| **Probably yes** | Mean | 4.12 | 6.00 | 2.79 |
| | N | 167 | 168 | 168 |
| | Std. Deviation | 1.124 | 1.540 | 1.090 |
| **Definitely yes** | Mean | 4.40 | 5.83 | 3.32 |
| | N | 131 | 132 | 132 |
| | Std. Deviation | 0.990 | 1.810 | 1.464 |
| **Total** | Mean | 4.21 | 6.14 | 2.39 |
| | N | 1093 | 1094 | 1097 |
| | Std. Deviation | 1.086 | 1.449 | 1.165 |

**Table 6: Personal Experience of Online Harassment**

*Statistical Results: Harassment Level*

| Condition | t-stat | p-value |
|---|---|---|
| Definitely no / Might or might not have | 3.510 | 0.001 |
| Probably no / Might or might not have | 1.494 | 0.136 |
| Probably yes / Might or might not have | 1.140 | 0.255 |
| Definitely yes / Might or might not have | 3.379 | 0.001 |

Statistical Significance: $p < 0.01$

## Free Speech

**Participants who cited free speech rated content as not very harassing**

On average, respondents that referenced “free speech” or “freedom of speech” in any of the open text fields in the instrument rated the content as less harassing and more appropriate than those that did not. These respondents had an average harassment level rating of 3.68 (versus 4.23) and average appropriateness rating of 5.50 (versus 6.16). For example, one heterosexual male participant, who rated the anti-feminist message as “not at all harassing” and “absolutely appropriate” wrote “free speech is a constitutional right. as long as the poster does not make a direct physical threat ‘that is actionable' no restriction is legal. political correction/censoring is first found in Mao's 'little red book' and it is being applied today and there is a growing backlash coming in the USA.” While our instrument did not explicitly capture this variable, these responses indicate additional research is needed to understand the relationship between harassing content and the desire for freedom of speech.

**Table 7: Free Speech Reference**

***Descriptive Results***

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***Free Speech Reference*** | Mean | 3.68 | 5.50 | 2.39 |
| | N | 44 | 44 | 44 |
| | Std. Deviation | 1.343 | 1.745 | 1.280 |
| ***No Free Speech Reference*** | Mean | 4.23 | 6.16 | 2.39 |
| | N | 1057 | 1058 | 1061 |
| | Std. Deviation | 1.071 | 1.429 | 1.161 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1101 | 1102 | 1105 |
| | Std. Deviation | 1.087 | 1.448 | 1.166 |

### Additional Qualitative Responses

**Respondents displayed a range of qualitative responses from “something must be done” to “this battle can’t be won” to “there is nothing wrong”**

Many respondents took the opportunity to express their frustration with the continued problem of harassing messages online.

> *“Harassment online/bullying/hate messages all are totally unacceptable behavior and should be treated as such. Not acceptable. period. forget freedom of speech. not applicable, in my opinion with regard to hate speech, etc.”*

> *“They've got to find a way to stop the people from harassing.”*

However, other respondents expressed a belief that this problem is a fundamentally human problem that can not and will not be solved through platform moderation or design.

> *“You can't win this battle unfortunately since internet is anonymous, all you can do is to block the source and hope that the culprit moves on to another platform of his choice.” (straight asian male who rated picnic message as somewhat harassing and neutral on the appropriateness scale)*

> *“I've been cyberbullied myself. The people who do it are idiot losers and "consider the source" applies. You're on a wrong track and a wild goose chase here.” (straight F of unknown ethnicity who rated the picnic message as somewhat harassing but absolutely appropriate)*

Some respondents were openly hostile to us and our research purposes, which were stated in the informed consent section at the beginning of the survey:

> *“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.”*

> *“In general, your whole premise is screwed up. Harassment is a crime, and it's something directed at a particular person who has standing to seek relief from the criminal justice system. This is merely speech that is offensive to some. You must be liberals; that would be the only explanation for such nonsense and lack of logic. People with more on the ball don't waste attention and energy with online "communities," "forums," etc.” (straight Female of unknown ethnicity who rated the picnic message as somewhat harassing but absolutely appropriate)”*

Taken together, the qualitative comments we received highlight the sensitivity and controversy that surround the issue of moderating online harassment. We posit that it is unlikely platforms will find a “silver bullet” to please all users, however improvements in the design of moderation processes and transparency could help to ameliorate some concerns, while tapping into the wishes of the large majority that agrees these messages are harassing.

### Limitations

The content we used in the research instrument, with its inclusion of the threat of violence against a person or group, may have been too harassing and inappropriate for the experimental conditions to register a difference from the average. The negative skew observed in the response distributions for both harassment level and appropriateness lends credence to this possibility.

Although we view our decision not to include a definition of harassment for our respondents as primarily a feature, we also acknowledge that the responses we received to our survey questions about harassment experience may have been influenced or anchored by the messages we showed respondents as being potentially harassing. The results for our harassment experience questions do not align with the Pew Study on the same topic.[48] 27% or our respondents reported that they definitely or probably have personally experienced harassment, while 40% of Pew respondents said they’ve experienced it. An additional 12% of our respondents said they “might or might not have” experienced harassment, which could account for the difference between our sample and the Pew finding in this area.

## Reflection Website

### Technical Implementation and Iteration

The development of a web application that allows the general public to take part in our research and reflect on their own attitudes has always been a core part of our goal for this project. Our web application was initially architected to support a full survey-taking experience, including collecting and storing results securely. Over time, as we narrowed focus and opted to use the Qualtrics platform for gathering results (due to privacy, confidentiality and security concerns from CPHS), the focus of the web application significantly narrowed.

It is because of the initial focus that the application was developed in Python using the Django web framework. Django is an open-source web framework designed for building large-scale applications. It

[48] Pew Center Study: Duggan, Maeve. (Oct, 2014). Online Harassment. Pew Research Center. Retrieved March 16, 2016, from: http://www.pewinternet.org/2014/10/22/online-harassment/.

is notable in that much of the functionality required for a major web development is built-in, including form building and database management, two key features necessary for conducting a survey.

After migrating our survey to Qualtrics, we redesigned the application to mimic the experience of taking the full Qualtrics survey, but without storing individual users' answers in long-term storage for confidentiality reasons. For this, we utilized Django's built-in support for session variables, flushing the session both on a visit to the home page, and arrival at the results page, as well as with standard time-outs. The challenges in this design were two-fold: effectively loading all of the survey questions and answers to a database so they could easily and quickly be accessed by the application; and effectively designing the forms, by extending Django's standard layouts, to allow for an easy user experience.

We again re-focused the experience design to allow users to rate as much or as little content as they desired, or merely see visualizations of the overall survey results. The decision to do so came in light of our findings that most of the content was found to be very harassing. In addition to limiting the opportunities for reflection if it's likely that the user will simply agree with all respondents, we thought it would not be a useful exercise for our users to repeatedly subject themselves to this content. This new solution still requires light use of Django forms and session variables, but a more lightweight solution, like Flask or Pyramid, would have been appropriate were the initial Django architecture not already created. Still, some of the Django infrastructures remained useful, in particular its affordances for managing forms via a database (even though we were not using the same data model to store results).

Upon receiving the results of the survey, where most messages we put forward were rated as "very harassing," we questioned the value of allowing users to continue to expose themselves to content near-universally viewed as bad. We then migrated the experience site design again to an interactive storytelling model that gave users a tour of our most interesting findings. We incorporated our original goal of allowing users to reflect on a choice—this time not just how harassing the message is, but also some of the questions we found more interesting, such as which factors they would like to see an algorithmic moderation tool take into account and what the source of a visual flag indicator was—to build an interactive experience where users see charts of other respondents' data combined with their own choice and a paragraph of text outlining the fundamental choice behind the question. We outline the full user flow below.

### User Flow

Within our reflection site, visitors tour the research findings in a manner that presents a coherent narrative of the survey results and the issues and choices that must be confronted by social media users and moderators. By comparing the user's responses with those of the survey respondents, we promote reflection while providing additional context on both the issue and the user's position.

First, a user rates a piece of content on a scale from "Not at all harassing" to "Extremely harassing," as survey participants were asked to do. This question frames the discussion to come. After submitting their rating, a chart appears, showing the ratings of survey participants along with a bullet indicating their own rating. This allows them to consider their own opinion within the context of the larger population's responses.

Next, the user is shown two pieces of content. One is innocuous, but carries a visual flag indicator that it has been marked harassing by an algorithm, while the other is harassing (as determined by our survey) yet carries no such flag. Users are asked which of these two scenarios they would most like to avoid in moderation: false positives or false negatives. Upon submitting, they again see a visualization of how their choice compares to the wider survey result. We provide a discussion about the values of freedom of speech and the avoidance of censorship outweighing the desire to prevent all harassing content among our survey participants, when it comes to algorithms.

We then address the issues of fairness, accuracy and trustworthiness across human and algorithmic moderators by showing the user a piece of content three times, each with a different visual flagging indicator that the source was humans, an algorithm, or ambiguous. We ask the user to choose among the three which they find most trustworthy, most accurate and most fair. What follows is again a comparison to the results of our survey participants and a discussion of the trade-offs between having humans moderate content versus algorithms, especially as it pertains to mental and physical health risks and scalability.

The narrative returns to the question of designing an algorithm to moderate content, asking the user which factors they would like such an algorithm to consider. We provide a comparison of results and discussion of some of the trade-offs implied by including or not including various factors.

Next, the user is presented with a choice of two remediation options which had similar scores in the survey for severity, but differ in their prevalence, and asked the user to choose which action they would prefer to take for a given piece of content. While this is not a question we asked in the survey, it draws on our results. We present a visualization of the remediation preferences of survey respondents and discuss how other options, beyond those typically found on platforms (like content removal), could be effective and were popular among respondents.

The user is then presented with the first piece of content they rated, now with the “ambiguous” visual flag indicator and asked to rate it again. What follows is a discussion about how the flags, despite not having much of an impact on harassment ratings among survey respondents, did affect their choice of remediation option, making them more likely to want the content taken down. We discuss how there may be a “safety in numbers” effect, making people more likely to agree that something should be taken down if they know that others have found it harassing.

Lastly, we ask the user which social media they spend the most time on, allow them to compare their response with our survey respondents, and discuss how context and community matters in establishing social norms on the internet.

On the final page, we present the user with resources to learn more about online harassment and the various issues discussed including remediation options, delegation to algorithms, and free speech and censorship.

## Conclusion

### Design and Policy Implications for Platforms

While the lack of statistically significant results does not inspire a strong call to action for platforms, there are several key takeaways from our responses relevant to social media platform operators. First, we have established that there is very little popular support for the presence of violent content targeting women and minorities on social media regardless of whether that content targets specific users or not. To the extent that platforms may hesitate to remove content of this nature, fearing a backlash from users, we again highlight that even users using the phrase “free speech” in their free-text comments rated the harassment level of content at 3.68, which falls between “somewhat harassing” and “very harassing,” and the appropriateness level at 5.50, which falls between “slightly inappropriate” and “inappropriate.” This indicates that, while there may be a vocal minority that would be upset by this content’s removal, the impact to the platform is likely to be minimal.

Additionally, while there was no significant difference between responses to our main experimental conditions based on the visual flag indicator presented to the respondent, we observed that users who saw a flag were more likely to choose taking down a message as a preferred remediation option. This is in the same vein as research by Crawford and Gillespie,[49] as well as Kizilcec,[50] emphasizing the positive effects of transparency in these processes. We suggest that platforms may be well-served to introduce such visual flag indicators in order to increase trust in their moderation processes—especially while a specific message is undergoing review, and particularly if platforms are using algorithms. This effort could be buttressed by the release of annual moderation transparency reports by platforms detailing the moderation process and statistics on the number and nature of pieces of content removed. We believe a combination of the flagging indicator and transparency reports would help promote reflection, deliberation, and discussion in more public channels.

Finally, in light of many of our qualitative results—especially those from the angriest of our respondents—we think platforms would be well-served to make clearer the distinction between government censorship prohibited under the First Amendment and content moderation governed by the platforms. Platforms could consider promoting learning and discourse on the principles, purpose, and application of the First Amendment—potentially via several of the values-based approaches to remediation that we included—through their reporting, flagging and takedown mechanisms. Alternatively, one could envision an addition to the remediation process that explains to the posters of moderated content that their First Amendment rights have not been violated; they are, as cartoonist Randall Munroe says, merely being shown the door.[51]

### Future Work

While we were unable to reject our null hypotheses, we conducted additional analysis on the data and identified a few areas for future research:

---

[49] Crawford, Kate, and Tarleton Gillespie. "What is a flag for? Social media reporting tools and the vocabulary of complaint." New Media & Society 18.3 (2016): 410-428.

[50] Kizilcec, René F. "How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface." Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 2016.

[51] Munroe, Randall. “Free Speech.” *xkcd.* Retrieved 5 May 2017. https://xkcd.com/1357/

We hypothesize that the presence of violence in the tested messages may have resulted in our messages being so universally viewed as harassing that we were unable to successfully discern effects of flagging conditions that would have had a significant effect if opinions about the level of harassment were more diverse. To this end, we are undertaking a secondary, much smaller study using a non-violent message: “@User you’re why women shouldn’t speak in public,” which we expect will results in a wider range of harassment ratings. In the future we would like to both increase the ranges of messages we test as well as revise our instrument to increase its sensitivity in collecting information that illuminates the ways in which people think about this type of content, beyond whether it is harassing or not. For instance, we may see more variability in our flagging conditions if included a wider range of questions about how people experience and understand this content, such “how painful do you think this content would be to the recipient?”, “how marginalizing do you think this content would be to the recipient?” or, along this lines of the Perspective API, “How toxic do you think this content would be to the conversation on this platform?”

Future work in this space should also look to explicitly capture people’s attitudes toward free speech as it relates to online harassment. While we were able to gain limited insight into this by analyzing the free text fields in our survey, explicitly capturing this information will allow for generalizable insights.

While we intentionally left harassment undefined in our research, more work is needed to understand how people define the contours of online harassment. We at one point envisioned a survey where respondents would rank 10-20 pieces of content in order from most to least harassing, but designing a desirable interface for such a ranking proved difficult, as did selecting pieces of content likely to fall on a clear spectrum. Such a tool, successfully designed, would have the potential allow researchers to model which features of a message caused it to be ranked as more or less harassing and further the understanding of what makes a particular message harassing.

# Appendix

## Content

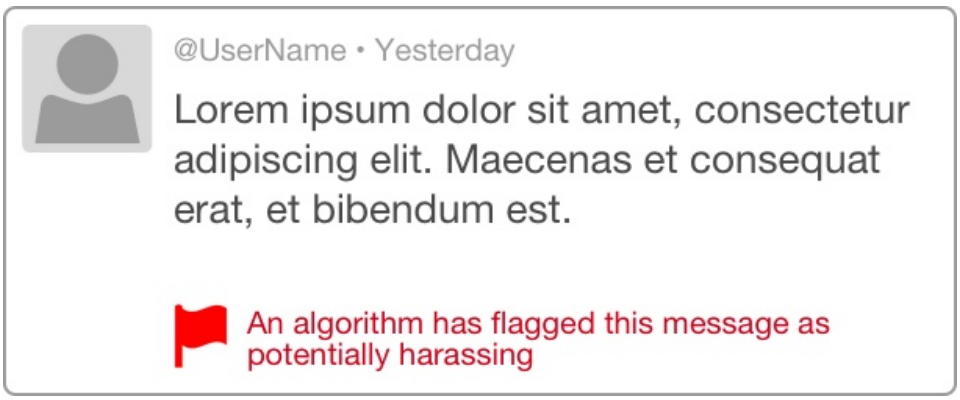
| ***Messages used in Experiment Phase 1*** |
|---|
| Oh and @User is a Muslim. So when do we get to go kill all the Muslims in the United States since they get to kill us in our own country |
| Oh and he's a Muslim. So when do we get to go kill all the Muslims in the United States since they get to kill us in our own country |
| This is also where picnics originated. Pick a random black and lynch them publicly in front of everyone as a form of entertainment. |
| This is also where picnics originated. Pick a random black like @User and lynch them publicly in front of everyone as a form of entertainment. |
| @user Just putting it out there, you feminists deserve all those death threats you are getting. |
| Just putting it out there, feminists deserve all those death threats they're getting. |

| ***Message used in Experiment Phase 2*** |
|---|
| @user you’re why women shouldn’t speak in public. |

| ***Messages used non-experimentally, for Reflection Tool visualizations*** |
|---|
| @User Because you're trans, your family will never love you. You are living a lie and you know it. I hope someone ends your miserable existence. |
| If you're trans, your family will never love you. You are living a lie and you know it. I hope someone ends your miserable existence. |
| @User If I saw your gay ass in real life I would beat the fuck outta you. |
| If I saw any gay asses in real life I would beat the fuck outta them. |
| @User has a $1.5m home and wants to call themself average. These NYC Jews need a good gassing |
| They have a $1.5m home and want to call themselves average. These NYC Jews need a good gassing |
| @User With the U.S.A. being overrun by dirty Mexicans like you we have to take things into our own hands and shoot and kill them on sight. |
| With the U.S.A. being overrun by dirty Mexicans we have to take things into our own hands and shoot and kill them on sight. |
| @User said "White Supremacy is a religion of peace." This is incredible. I have no words. I need to punch the fucking nazi right now. |
| "White Supremacy is a religion of peace." This is incredible. I have no words. I need to punch a fucking nazi right now. |

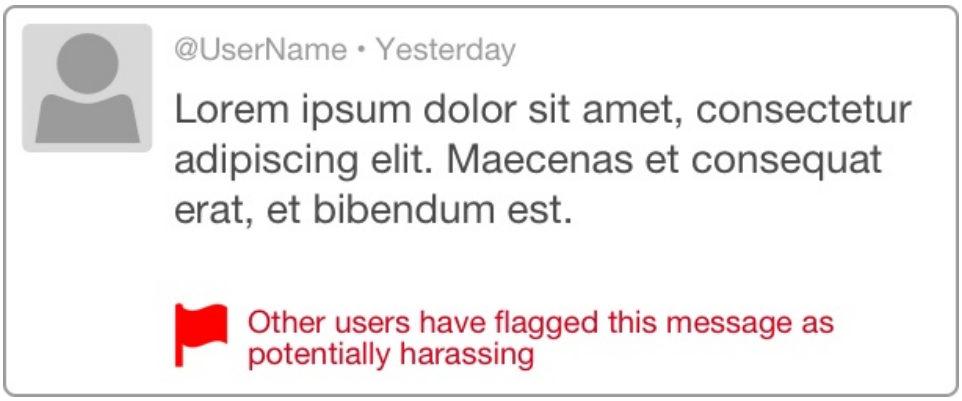
## Flag Conditions

**Condition 1: Content Flagged by Algorithm**




**Condition 2: Content Flagged by Human Moderators**


**Condition 3: Content Flagged by Other Users**



**Condition 4: Content Flagged, No Source**

### Enumerated Research Hypotheses

1. **Harassment level of the post && directedness**
    a. Null: There is no difference in rated level of harassment between messages that are directed at a specific user and messages that are not directed.
    b. Alt: Participants rate messages directed a specific user as more harassing than those that are not directed at a specific user.
2. **Harassment level of the post && presence/type of flag**
    a. Null: The presence and type of visual flagging makes no difference in a message's rated level of harassment.
    b. Alt 1: The presence of visual flags will increase a message's rated level of harassment over messages lacking these flags.
    c. Alt 2: Messages flagged by humans (moderators or other users) will have higher ratings of harassment than those flagged by algorithms.
3. **Harassment level of post && preferred remediation option && remediation severity**
    a. Null: There will be no relationship between rated level of harassment and the rated level of severity of the preferred remediation option(s).
    b. Alt: The severity rating of the preferred remediation option(s) will increase with increases in the message's rated level of harassment.
4. **Harassment level of post && targeted characteristic && demographics**
    a. Null: The rated harassment level of the message will not affected by a person's membership in the group that is targeted in the message.
    b. Alt: Messages will be rated as more harassing when they are rated by participants who are members of the message's targeted group.
5. **Harassment level of post && harassment experience (victim)**
    a. Null: Rated level of harassment will not be influenced by whether or not a participant has experienced harassment previously.
    b. Alt: Those who have previously experienced harassment will rate messages as more harassing than those who have not previously experienced harassment.
6. **Harassment level of post && harassment experience (perpetrator)**
    a. Null: Rated level of harassment will not be influenced by whether or not a participant has committed harassment previously.
    b. Alt: Those who have previously committed harassment will rate messages as less harassing than those who have not previously committed harassment.
7. **Harassment level of post && preferred social media site**
    a. Null: A participant's preferred social media site will not have an effect on the message's rated level of harassment.
    b. Alt 1: Participants who prefer Twitter will rate messages as less harassing.
    c. Alt 2: Participants who prefer Pinterest will rate messages as more harassing.
8. **Harassment level of post && presence/type of flag && assessment of fairness of flagging methods**
    a. Null: A user's conception of the fairness of some flagging process will have no relationship to the rated harassment level of the post.
    b. Alt: A user who rates a particular flagging process as fair or somewhat fair will rate messages containing that flag as more harassing.

9. **Harassment level of post && presence/type of flag && assessment of accuracy of flagging methods**
    a. Null: A user's conception of the accuracy of some flagging process will have no relationship to the rated harassment level of the post.
    b. Alt: A user who rates a particular flagging process as accurate or somewhat accurate will rate messages containing that flag as more harassing.
10. **Harassment level of post && presence/type of flag && assessment of trustworthiness of flagging methods**
    a. Null: A user's conception of the trustworthiness of some flagging process will have no relationship to the rated harassment level of the post.
    b. Alt: A user who rates a particular flagging process as trustworthy or somewhat trustworthy will rate messages containing that flag as more harassing.
11. **Remediation severity && harassment experience (victim)**
    a. Null: There will be no relationship between a participant's prior experience as the victim of harassment and the rating of remediation severity.
    b. Alt: Participants who have previously experienced harassment will rate all remediation options as less severe than those who have not experienced harassment.
12. **Remediation severity && harassment experience (perp)**
    a. Null: There will be no relationship between a participant's prior experience as the perpetrator of harassment and the rating of remediation severity.
    b. Alt: Participants who have previously perpetrated harassment will rate all remediation options as more severe than those who have not committed harassment.
13. **Remediation effectiveness && harassment experience (victim)**
    a. Null: There will be no relationship between a participant's prior experience as the victim of harassment and the rating of remediation effectiveness.
    b. Alt: Participants who have previously experienced harassment will rate all remediation options as less effective than those who have not experienced harassment.
14. **Remediation effectiveness && harassment experience (perpetrator)**
    a. Null: There will be no relationship between a participant's prior experience as the perpetrator of harassment and the rating of remediation effectiveness.
    b. Alt: Participants who have previously perpetrated harassment will rate all remediation options as less effective than those who have not committed harassment.

**Table 8: Racial Identity**

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***White*** | Mean | 4.19 | 6.15 | 2.36 |
| | N | 869 | 868 | 871 |
| | Std. Deviation | 1.104 | 1.446 | 1.166 |
| ***Hispanic, Latino, or Spanish origin*** | Mean | 4.25 | 6.19 | 2.38 |
| | N | 61 | 63 | 63 |
| | Std. Deviation | 0.994 | 1.268 | 1.263 |
| ***Black or African-American*** | Mean | 4.34 | 6.17 | 2.41 |
| | N | 124 | 125 | 125 |
| | Std. Deviation | 1.011 | 1.496 | 1.185 |
| ***Asian*** | Mean | 3.94 | 5.98 | 2.67 |
| | N | 49 | 49 | 49 |
| | Std. Deviation | 1.180 | 1.108 | 1.088 |
| ***American Indian or Alaska Native*** | Mean | 4.19 | 5.31 | 2.44 |
| | N | 16 | 16 | 16 |
| | Std. Deviation | 1.167 | 1.991 | 1.263 |
| ***Middle Eastern or North African*** | Mean | 5.00 | 7.00 | 2.00 |
| | N | 1 | 1 | 1 |
| | Std. Deviation | - | - | - |
| ***Native Hawaiian or Other Pacific Islander*** | Mean | 4.33 | 6.33 | 2.33 |
| | N | 3 | 3 | 3 |
| | Std. Deviation | 1.155 | 1.155 | 0.577 |
| ***Other*** | Mean | 4.54 | 6.85 | 2.54 |
| | N | 13 | 13 | 13 |
| | Std. Deviation | 0.877 | 0.376 | 0.877 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1101 | 1102 | 1105 |
| | Std. Deviation | 1.087 | 1.448 | 1.166 |

**Table 9: Sexual Identity**

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***Heterosexual or straight*** | Mean | 4.21 | 6.17 | 2.37 |
| | N | 965 | 966 | 968 |
| | Std. Deviation | 1.071 | 1.401 | 1.170 |
| ***Homosexual, Gay, or Lesbian*** | Mean | 4.48 | 6.18 | 2.52 |
| | N | 33 | 33 | 33 |
| | Std. Deviation | 0.939 | 1.467 | 1.149 |
| ***Bisexual*** | Mean | 4.44 | 6.43 | 2.68 |
| | N | 39 | 40 | 40 |
| | Std. Deviation | 1.021 | 1.338 | 1.095 |
| ***I don't know*** | Mean | 4.56 | 6.44 | 2.11 |
| | N | 9 | 9 | 9 |
| | Std. Deviation | 1.014 | 0.726 | 0.928 |
| ***Something else*** | Mean | 4.00 | 5.64 | 2.18 |
| | N | 11 | 11 | 11 |
| | Std. Deviation | 1.342 | 1.567 | 1.250 |
| ***Total*** | Mean | 4.22 | 6.18 | 2.38 |
| | N | 1057 | 1059 | 1061 |
| | Std. Deviation | 1.068 | 1.398 | 1.166 |

## Table 10: Gender Identity

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***Male*** | Mean | 4.03 | 5.81 | 2.38 |
| | N | 408 | 408 | 409 |
| | Std. Deviation | 1.189 | 1.695 | 1.196 |
| ***Female*** | Mean | 4.32 | 6.37 | 2.38 |
| | N | 663 | 664 | 666 |
| | Std. Deviation | 0.988 | 1.171 | 1.138 |
| ***Non-Conforming*** | Mean | 4.33 | 6.33 | 1.67 |
| | N | 3 | 3 | 3 |
| | Std. Deviation | 0.577 | 1.155 | 1.155 |
| ***Total*** | Mean | 4.21 | 6.16 | 2.38 |
| | N | 1074 | 1075 | 1078 |
| | Std. Deviation | 1.076 | 1.418 | 1.160 |

## Table 11: Age

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***18 - 24*** | Mean | 4.09 | 6.13 | 2.64 |
| | N | 80 | 80 | 80 |
| | Std. Deviation | 1.203 | 1.641 | 1.117 |
| ***25-34*** | Mean | 4.31 | 6.01 | 2.87 |
| | N | 232 | 232 | 232 |
| | Std. Deviation | 1.009 | 1.519 | 1.255 |
| ***35-44*** | Mean | 4.00 | 5.86 | 2.49 |
| | N | 219 | 218 | 219 |
| | Std. Deviation | 1.232 | 1.739 | 1.182 |
| ***55-64*** | Mean | 4.26 | 6.17 | 2.11 |
| | N | 198 | 200 | 200 |
| | Std. Deviation | 1.085 | 1.417 | 1.024 |
| ***65+*** | Mean | 4.36 | 6.48 | 1.94 |
| | N | 145 | 144 | 145 |
| | Std. Deviation | 0.955 | 1.024 | 0.944 |
| ***45-54*** | Mean | 4.18 | 6.31 | 2.24 |
| | N | 222 | 223 | 224 |
| | Std. Deviation | 1.037 | 1.130 | 1.113 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1096 | 1097 | 1100 |
| | Std. Deviation | 1.089 | 1.442 | 1.165 |

## Table 12: Actions Potentially Described as Harassing

*Do you think anyone would describe something you've done online as harassing?*

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***Definitely not*** | Mean | 4.38 | 6.44 | 2.08 |
| | N | 600 | 600 | 603 |
| | Std. Deviation | 0.963 | 1.142 | 1.066 |
| ***Probably not*** | Mean | 4.00 | 6.05 | 2.45 |
| | N | 218 | 219 | 219 |
| | Std. Deviation | 1.128 | 1.478 | 0.968 |
| ***Might or might not have*** | Mean | 3.85 | 5.82 | 2.52 |
| | N | 147 | 147 | 147 |
| | Std. Deviation | 1.235 | 1.408 | 0.982 |
| ***Probably yes*** | Mean | 4.00 | 5.33 | 3.37 |
| | N | 81 | 81 | 81 |
| | Std. Deviation | 1.194 | 2.006 | 1.289 |
| ***Definitely yes*** | Mean | 4.31 | 5.19 | 3.75 |
| | N | 52 | 52 | 52 |
| | Std. Deviation | 1.261 | 2.301 | 1.412 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1098 | 1099 | 1102 |
| | Std. Deviation | 1.088 | 1.450 | 1.166 |

## Table 13: Initiated a Complaint to an Online Platform

*Have you ever initiated a complaint to a platform about harassing content?*

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***Definitely not*** | Mean | 4.24 | 6.28 | 1.96 |
| | N | 522 | 520 | 523 |
| | Std. Deviation | 1.071 | 1.366 | 0.987 |
| ***Probably not*** | Mean | 4.15 | 6.16 | 2.52 |
| | N | 145 | 145 | 145 |
| | Std. Deviation | 1.139 | 1.321 | 0.987 |
| ***Might or might not have*** | Mean | 3.79 | 5.88 | 2.58 |
| | N | 145 | 145 | 145 |
| | Std. Deviation | 1.201 | 1.353 | 1.065 |
| ***Probably yes*** | Mean | 4.13 | 5.90 | 2.74 |
| | N | 128 | 129 | 129 |
| | Std. Deviation | 1.065 | 1.565 | 1.134 |
| ***Definitely yes*** | Mean | 4.56 | 6.08 | 3.18 |
| | N | 158 | 160 | 160 |
| | Std. Deviation | 0.863 | 1.741 | 1.364 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1098 | 1099 | 1102 |
| | Std. Deviation | 1.088 | 1.450 | 1.164 |

## Table 14: Most Used Social Media Platform

*Which social media platform do you use most frequently?*

| | | Harassment Level | Appropriateness | Frequency |
|---|---|---|---|---|
| ***Facebook*** | Mean | 4.24 | 6.18 | 2.40 |
| | N | 805 | 807 | 809 |
| | Std. Deviation | 1.047 | 1.385 | 1.151 |
| ***Reddit*** | Mean | 4.00 | 5.89 | 2.56 |
| | N | 9 | 9 | 9 |
| | Std. Deviation | 1.118 | 1.364 | 1.424 |
| ***Twitter*** | Mean | 3.91 | 5.77 | 2.75 |
| | N | 75 | 75 | 75 |
| | Std. Deviation | 1.337 | 1.721 | 1.220 |
| ***Pinterest*** | Mean | 4.18 | 6.61 | 2.09 |
| | N | 33 | 33 | 33 |
| | Std. Deviation | 1.185 | 0.788 | 0.947 |
| ***Instagram*** | Mean | 4.21 | 5.76 | 2.81 |
| | N | 62 | 62 | 62 |
| | Std. Deviation | 1.058 | 1.905 | 1.353 |
| ***Other, please specify:*** | Mean | 4.35 | 6.57 | 2.39 |
| | N | 23 | 23 | 23 |
| | Std. Deviation | 0.982 | 0.728 | 1.158 |
| ***I don't use social media*** | Mean | 4.07 | 6.09 | 1.78 |
| | N | 92 | 91 | 92 |
| | Std. Deviation | 1.203 | 1.582 | 0.875 |
| ***Total*** | Mean | 4.20 | 6.14 | 2.39 |
| | N | 1099 | 1100 | 1103 |
| | Std. Deviation | 1.088 | 1.444 | 1.164 |

### Survey Instrument

*Notes:*

- *One message from Text Content will be inserted in **[Content Placeholder]** below.*
- *Note: Control and experimental groups will receive the same survey questions (with the exception of one additional questions for one experimental group). All other questions, content, and flow will remain the same between the control and experimental conditions. Only the display of the information in the **[Content Placeholder]** will vary between the groups. Please see the 'Control and Experimental Conditions' section for further details on the conditions.*
- *The exact remediation actions (questions 5,6,7), computational moderation features (question 8), and types of online harassment (question 17) may vary slightly in the final instrument. If necessary, these changes will be confirmed with the IRB board prior to running the survey.*
- *Each respondent will assess one piece of content.*

**Content Assessment**

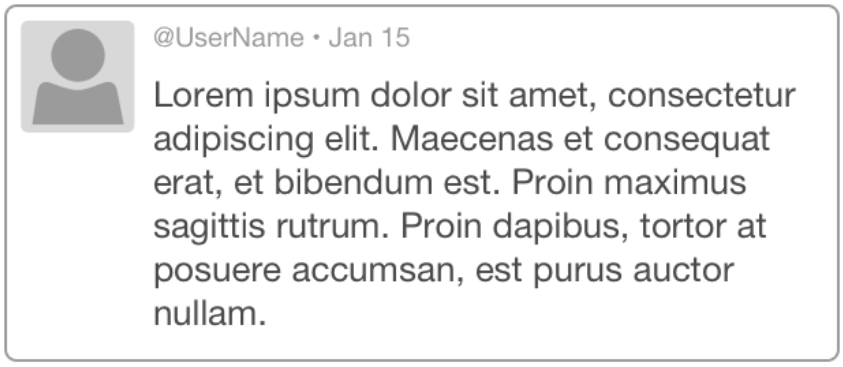
*User Instructions:*

Please consider the social media post below when responding to the following questions.

***[Section start]***

***[Content]***

*Content will be presented below in either the control or one of the experimental condition formats. The control format is used below.*




1. How harrassing is this social media post? *[Single select]*
    - Not at all harassing
    - Slightly harassing
    - Somewhat harassing
    - Very harassing
    - Extremely harassing

2. How appropriate is this social media post? *[Single select]*
    - Absolutely inappropriate

- Inappropriate
- Slightly inappropriate
- Neutral
- Slightly appropriate
- Appropriate
- Absolutely appropriate

3. How often do you come across this type of social media post? *[Single select]*
   - Never
   - Rarely
   - Occasionally
   - A moderate amount
   - A great deal

4. Which of the following remediation actions would you prefer be applied to this social media post? *[Select up to three]*
   - Take no action
   - Remove the message from the platform
   - Add the user to a block-list which hides their posts from all subscribers to that block-list
   - Prevent the user from posting again until they complete an empathy exercise
   - Send the user a private reminder of community code of conduct
   - Post a public reminder of the community code of conduct in response to message
   - Notify the user that the post has been deemed potentially abusive or harassing by an automated system and require confirmation prior to posting
   - Temporarily suspend the user
   - Permanently ban the user
   - Something else? Write In:

***[Section end]***

**Remediation Assessment**

User Instructions:
We would now like you to consider the severity of a few possible remediation options which could be used when responding to social media posts online.

5. Regardless of your answer to the previous question, how effective do you think the following remediation options are in discouraging people from posting this type of content? *[Grid; single select per row; up to 10 options will be shown per respondent]*

| | Not at all severe | Slightly severe | Somewhat severe | Moderately severe | Extremely severe |
|---|---|---|---|---|---|
| Take no action | | | | | |
| Remove the message from the platform | | | | | |
| Add the poster to a block-list which hides their posts from all people who subscribe to that block-list | | | | | |
| Prevent the poster from posting again until they complete an empathy exercise | | | | | |
| Send the poster a private reminder of the community code of conduct | | | | | |
| Post a public reminder of the community code of conduct in response to message | | | | | |
| Before publicly posting, notify the poster that it has been flagged as potentially harassing and require confirmation to post | | | | | |
| Temporarily suspend the poster | | | | | |
| Permanently ban the poster | | | | | |

6. Platforms have a variety of methods for discouraging harassing behavior. In general, how severe do you think the following remediation options are? *[Grid; single select per row; up to 10 options will be shown per respondent]*

| | Not at all severe | Slightly severe | Somewhat severe | Moderately severe | Extremely severe |
|---|---|---|---|---|---|
| Take no action | | | | | |
| Remove the message from the platform | | | | | |
| Add the poster to a block-list which hides their posts from all people who subscribe to that block-list | | | | | |
| Prevent the poster from posting again until they complete an empathy exercise | | | | | |
| Send the poster a private reminder of the community code of conduct | | | | | |
| Post a public reminder of the community code of conduct in response to message | | | | | |
| Before publicly posting, notify the poster that it has been flagged as potentially harassing and require confirmation to post | | | | | |
| Temporarily suspend the poster | | | | | |
| Permanently ban the poster | | | | | |

***[Section end]***

Moderation Source

7. *This question will be asked only of participants in the No-Source Flag condition:*
   The public post above is flagged and states: "This message has been flagged as potentially harassing." Considering this statement, what did you assume to be the source of the flagging? *[select all that apply]*
   - An algorithm
   - A human moderator
   - Another user or users
   - Other, please specify:
   - I don't know

Please answer the following questions regarding different methods for flagging content as potentially harassing.

8. How fair do you think each of the following methods is?

| | Unfair | Somewhat Unfair | Neither Unfair nor Fair | Somewhat Fair | Fair |
|---|---|---|---|---|---|
| An algorithm | | | | | |
| A human moderator | | | | | |
| Another user or users | | | | | |

9. How trustworthy do you think each of the following methods is?

| | Unfair | Somewhat Unfair | Neither Unfair nor Fair | Somewhat Fair | Fair |
|---|---|---|---|---|---|
| An algorithm | | | | | |
| A human moderator | | | | | |
| Another user or users | | | | | |

10. How accurate do you think each of the following methods is?

| | Unfair | Somewhat Unfair | Neither Unfair nor Fair | Somewhat Fair | Fair |
|---|---|---|---|---|---|
| An algorithm | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| A human moderator | | | | | |
| Another user or users | | | | | |

*11.* To help us better understand the rationale for your answers to the above questions, please explain your answers below. (*Optional)*

**Computational Moderation Features**

*User Instructions:*
For the following questions, imagine you've been tasked with building a computational tool to automatically detect abusive or harassing content online.

12. Select up to three criteria that you think will be most useful in detecting this type of content. *[Multi-select up to three; randomized order]*
    - Presence or lack of certain words that you define
    - Sentiment analysis - does the comment seem negative or positive overall?
    - Length of message
    - Conversational balance - have both parties contributed to the conversation, or is one person doing all of the communication?
    - Image recognition - does the message contain any images that are recognized as commonly harassing or abusive
    - Message has been flagged by other humans
    - Message poster has previously been flagged for harassing content
    - Other: [what other elements would you include?]
    - I don't know

13. As you design this tool, what do you think is more important to prevent?*[single select]*
    - Non-harassing and non-abusive messages being flagged as "harassing" or "abusive" and potentially subjected to remediation
    - Harassing or abusive content not being flagged and not subjected to remediation
    - No preference
    - I don't know

14. If you'd like to explain your response to the above question, please do so here: *[Free text; optional]*

**Demographic Information**

15. How old are you? *[Single select; optional]*
    - 18 - 24
    - 25 - 34
    - 35 - 44
    - 45 - 54
    - 55 - 64
    - 65+

16. Which racial or ethnic categories describe you? *[Select all that apply; optional]*
    - White
    - Hispanic, Latino, or Spanish origin
    - Black or African-American
    - Asian
    - American Indian or Alaska Native
    - Middle Eastern or North African
    - Native Hawaiian or Other Pacific Islander
    - Other

17. Which gender category best describes you? *[Single select; optional]*
    - Male
    - Female
    - Non-Conforming

18. Do you consider yourself to be: *[Single Select; optional]*
    - Heterosexual or straight
    - Homosexual, Gay, or Lesbian
    - Bisexual
    - Something else
    - I don't know

**Harassment Experience**

19. Have you ever witnessed any of the following types of online harassment? *[Select all that apply; optional]*
    - Harassing or abusive messages sent in public channels
    - Harassing or abusive messages sent through private channels
    - Doxxing (public release of someone's personal information)
    - Hate speech
    - Threat of violence
    - False information posted about someone
    - Revenge porn
    - Messages encouraging others to harass someone
    - Impersonation
    - Other

20. Do you think you've ever personally experienced online harassment or abuse? *[Single select; optional]*
    - Yes
    - No
    - I don't know

21. Do you think anyone would describe something you've done on the internet as abusive or harassing? *[Single select; optional]*
    - Yes
    - No
    - I don't know

**Social Media Use**

*User Instructions:*
Please answer a few questions about yourself.
All the remaining questions are optional. You may skip any that make you uncomfortable or that you would prefer not to answer. However, the more information we are able to gather from people, the more useful our findings will be, so we appreciate you answering as many questions as possible.

22. On average, how often do you visit the following social media sites? *[Grid; single select per row, optional]*

| | Multiple times a day | Daily | Weekly | Monthly | Rarely/Never |
|---|---|---|---|---|---|
| Facebook | | | | | |
| Reddit | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Twitter | | | | | |
| Pinterest | | | | | |
| Instagram | | | | | |
| Snapchat | | | | | |
| Other, please specify [optional] | | | | | |
| Other, please specify [optional] | | | | | |

23. Which social media platform do you use <u>most</u> frequently? *[Single select; optional]*
    - Facebook
    - Reddit
    - Twitter
    - Pinterest
    - Instagram
    - Snapchat
    - Other, please specify *[Open end; optional]*

**<u>Request for Comments</u>**

24. Do you have any thoughts you'd like to share with us regarding online harassment or moderation?
    Optional

## Remediation Techniques Used in Survey

| Potential Corrective Action | Design for... (Bowler et al.) | Value Attribute Shift (Shilton et al.) | Social Justice Dimensions (Dombroski et al.) | Potential Value Conflicts |
|---|---|---|---|---|
| Take no action | - | - | - | Proliferation for harassing messages |
| Remove the message from the platform | Consequence<br>Control and Suppression | Performed to Potential | Accountability | Suppression of free speech<br>Potential for false positives and negatives<br>Does not promote behavior change |
| Before publicly posting, notify the poster that it has been flagged as potentially harassing and require confirmation to post | Reflection<br>Empathy | Accidental to Purposive<br>Peripheral to Central | Enablement | Suppression of speech |
| Add the poster to a block-list which hides their posts from all people who subscribe to that block-list | Empowerment<br>Control and Suppression | Performed to Potential | Accountability | Does not promote behavior change<br>"Concerns that criteria to place users on list may be inconsistently applied; Concerns that appearing on list implies some value judgement about a person" (Geiger) |
| Prevent poster from posting again until they have completed an empathy exercise | Consequence<br>Empathy | Peripheral to Central | Enablement | Suppression of speech |
| Send the poster a private reminder of community code of conduct | Reflection | Peripheral to Central | Enablement | Reminder could easily be ignored |
| Post a public reminder of the community code of conduct in response to message | Reflection<br>Empathy | Peripheral to Central | Recognition<br>Reciprocity | Potential for deleterious effects from "naming-and-shaming" |
| Temporarily suspend the poster | Consequence | Peripheral to Central | Accountability | Suppression of speech |
| Permanently ban the poster | Consequence | Purposive to Impossible | Accountability | Suppression of speech |

## Remediation Techniques Considered but not Used in Survey

| Potential Corrective Action | Design for... (Bowler et al) | Value Attribute Shift | | Potential Value Conflicts |
|---|---|---|---|---|
| Notify friends of the recipient of the harassing message | Attention | Performed to Potential | Recognition | ? |
| Post a public message which fact-checks the post to validate the recipient of the harassing message | Empowerment | Performed to Potential | Recognition | ? |
| Activate a bot which targets offensive comments to the poster for 12 hours | Consequence<br>Empathy | Peripheral to Central | Reciprocity<br>Accountability | Tool could become another avenue for harassment |
| Publicly flag the poster's profile indicating the user has posted harassing content | Fear | Accidental to Purposive | Reciprocity<br>Accountability | The button could be abused and become another avenue for harassment |

## Reflection Site Screenshots