Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms

Jessica A. Pater
Georgia Institute of Technology
Atlanta, GA
pater@gatech.edu

Moon K. Kim
Georgia Institute of Technology
Atlanta, GA
moonkim@gatech.edu

Elizabeth D. Mynatt
Georgia Institute of Technology
Atlanta, GA
mynatt@gatech.edu

Casey Fiesler
University of Colorado, Boulder
Boulder, CO
casey.fiesler@colorado.edu

ABSTRACT

Harassment in online spaces is increasingly part of public debate and concern. Pervasive problems like cyberbullying, hate speech, and the glorification of self-harm have highlighted the breadth and depth of harassment taking place online. In this study we conduct a content analysis of the governing policies for fifteen social media platforms as they relate to harassment (of oneself and/or of community members) and other associated behaviors. We find that there is a striking inconsistency in how platform-specific policies depict harassment. Additionally, how these policies prescribe responses to harassment vary from mild censuring to the involvement of law enforcement. Finally, based on our analysis and findings, we discuss the potential for harnessing the power of the online communities to create norms around problematic behaviors.

Keywords

harassment; governance; online communities; policy; social media; terms of service; community norms

1. INTRODUCTION

Increasingly, online harassment is an issue of public debate and concern, particularly in the context of pervasive problems like cyberbullying [1, 4], hate speech and trolling [6, 17], and the glorification of self-harm [3, 22]. To address these types of issues, the computing research community has looked towards technological and design solutions for combatting these known problems [1, 4] while lawmakers tackle the same issues by creating new laws to govern behavior [2, 18]. However, these decisions are often made in an environment removed from where the actual harassment takes place. Operators of online social platforms have a unique advantage where they can create policies to address harassment in the same environment where it takes place.

But *are* these sites developing policy-based measures to address harassment, and if so, *what* do they look like?

Harassment itself is complex to define. Traditionally thought of as repeated annoyance or attacks, harassment is often operationalized as external, targeting another individual [23, 24]. However, harassment can also be focused inward, with an individual harassing oneself through self-harm and self-injurious behaviors [10]. Furthermore, harassment online has unique properties due to the asynchronous, and often anonymous, nature of online interactions. Pinning down certain behaviors as harassment within policy is also complicated due to values around freedom of expression.

Previous studies have examined the most common manifestations of online harassment and their associated behaviors inflicted on others, such as cyberbullying [4], hate speech [6], and sexual harassment [21], and those inflicted on oneself, like self-harm, suicide [15], and eating disorders [22]. This body of work examines these activities within specific contexts, and seeks to learn more about how platform-specific policies (such as Terms of Service, Community Guidelines, and other forms of formalized site governance) regulate these activities [17].

Though policy can be as powerful a design agent as technology [8], it is often only considered as an afterthought in the context of user research and design [13]. To understand the landscape of online harassment and intervene in appropriate ways—whether with rules (legal or platform-specific policies), technologically, or through community-enforced social norms—we first need a clear picture of what social platforms are attempting in terms of written regulations.

To understand this current state, we devised a study to examine how social platforms and their communities characterize harassment. In this work we address several key questions:

1. Do social media policies address the issue of harassment?
2. How do social media policies define and/or characterize harassment? How consistent and comprehensible are these policies? Are there similarities across platforms?

3. What actions in response to harassment do platforms outline in their policies?

To address these questions, we conducted a content analysis of policy documents from fifteen popular social media platforms, examining harassment-specific content. It should be noted that policies *as written* do not always reflect policies as enacted. However, analyzing stated policies is a required first step in understanding how social platforms work towards governing harassment. Through our analysis of policies as written, we show that harassment is not specifically defined, but it is commonly listed alongside other various prohibited behaviors. Additionally, platform responses to harassment and prohibited behaviors are varied and inconsistent.

We discuss the policy and design challenges that this vagueness and inconsistency might present to both online platforms and their users, and what research directions these observations suggest for the social computing research community. Going forward, a better understanding of the characterization of harassment in the policies that govern online social platforms provides the foundation to examine the differences between what is articulated through policies and what happens in practice.

2. RELATED WORK

The problem of online harassment has seen increasing attention within HCI and social computing research. Much of this prior work has focused on cyberbullying, especially towards technological and design solutions to curtail behavior [1, 4]. However, alongside design strategies, platform policies and community norms share an important role in encouraging certain behaviors and discouraging others. For example, in researching Reddit in the context of Gamergate, Massanari found that the site's policies and approach to community governance contributed to the presence of harassment [17]. Others have noted that law, platform policy, and norms can all play a role in regulating behavior, and conflicts between these multiple sources of governance can be problematic [9]. Social media platform designers and operators are in a special position of authority. Through their design and policy choices, they can play a critical role in governing user behavior. Routinely, they decide what content or behavior is allowable or not, and by doing so, they are effectively in the "gatekeeping position" of freedom of expression online [5].

Jackson et al. therefore argue that policy can be as important as design and practice within CSCW systems, and should be considered alongside rather than as an afterthought [13]. Accordingly, within the computing community there is a growing corpus of research focused on understanding online policy, especially with respect to Terms of Service (TOS) [16], end-user license agreements [11], copyright policies [8, 9], and privacy policies [14]. Additionally, research has

shown that there are often considerable disconnects between a community's understanding and its platform's actual policies [8, 11]. With the work described in this paper, we bring this policy-centric analysis to the problem of online harassment, as an extension to the prior work in both areas.

3. METHODS
3.1 Selecting Social Media Platforms
We adapted methods from Fiesler et al. [8] to analyze the TOS and supplementary policy documents of fifteen widely used social media platforms. We referenced a popular online trend aggregator to provide a list of widely used social media platforms for our analysis.[1] Table 1 highlights these platforms, along with proxies for popularity (unique monthly visitors) and cultural homogeneity (% user base that is in the United States).

Platform	Year Founded	Unique Monthly Visitors (US)	User base (% in U.S.)
Facebook	2004	122,875,81	21.8%
Twitter	2006	81,026,590	21.6%
LinkedIn	2002	67,633,842	26.8%
Pinterest	2010	61,327,968	45.6%
Instagram	2010	55,608,250	24.4%
Tumblr	2007	40,086,036	33.9%
Flickr	2004	23,893,368	28.3%
Meetup	2002	11,943,825	48.1%
Vine	2013	10,202,168	43%
Classmates	1995	4,758,561	89.6%
VK	2006	3,119,051	0.8%
Tagged	2004	1,445,300	17.5%
AskFM	2010	1,356,355	4.3%
MeetMe	2005	903,532	50.3%
Google +	2011	*Alexa doesn't differentiate from Google.com statistics*	

Table 1. Social media platforms in our dataset, sorted by popularity.

3.2 Policy Documents
In October of 2015, we collected policy documents from the fifteen platforms, a total of 56 documents (see Table 2). We identified two primary types of policies, formal and informal. Formal policy documents are legal and binding, and they include TOS, privacy policies, and acceptable use policies. Informal policy documents are non-legal, and they include normative artifacts such as community guidelines, safety guides, and other informative documentation. Most sites had more than three documents (80%), with an average of 3.7 per platform; Facebook had the most with eight documents.

We also assessed the readability of each document using a measure common in prior studies of online policies, the Flesch-Kincaid Grade Level (FKGL) score [8, 14]. In line with this prior work, the level of difficulty is high, an average

---

[1] http://www.ebizmba.com/articles/social-networking-websites

of 11.2 (equivalent to a high school junior) across all 56 documents. Formal policy documents in our dataset scored an average FKGL score of 12.2 while informal policy documents scored an average FKGL score of 9.8. These results suggest that informal policies rather than legal documents present information in a more accessible way.

3.3 Qualitative Coding of Policy Documents

Drawing from previous methods of policy document analysis [8], we used an iterative, inductive approach to qualitative coding. Two researchers independently coded a randomized sample of documents. They coded for general themes in the data, then met to discuss and synthesize the themes. We then refined and expanded the finalized codebook. High-level thematic codes included "moderation," "prohibited activity," "response to prohibited activity," "community standards," and "specific guidance." A total of 289 codes were applied to the dataset.

4. POLICY ANALYSIS

4.1 Defining Harassment

Surprisingly, we found that not a single platform-specific policy in our dataset explicitly defines what constitutes harassment. However, while they do not explicitly define harassment, Instagram and Twitter were the only platforms that describe specific behaviors or activities that are taken into account when determining if something is harassment. These behaviors include "repeated unwanted contact" (Instagram) or "reported behaviors [that are] one-sided or include threats" (Twitter).

More commonly, we saw the word "harassment" (including derivations of "harass" and "harassing") embedded in a string of other prohibited activities and behaviors. Table 3 highlights these activities and behaviors that co-occur with harassment in these documents. The most common include:

	Formal Documents				Informal Documents				
	TOS	Privacy Policy	Additional Terms	Acceptable Use Policy	Safety Guide	Community Guideline	Parental Guide	Law Enforce. Guide	Teen/Youth Guide
Facebook	(X)	X	X			X	(X)	X	X
Twitter	X	X	(X)				(X)	X	
LinkedIn	X	X		X		(X)			
Pinterest	X	X		(X)					
Instagram	(X)	X				(X)	(X)	X	
Tumblr	X	X			X	(X)			
Flickr	(X)	X			(X)	(X)	X		
Meetup	(X)	X				X			
Vine	X	X				X			
Classmates	(X*)								
VK	X	X							
Tagged	(X*)				X			X	
Ask.fm	(X)	X					(X)	X	(X)
MeetMe	(X*)								
Google+	X	X	(X)			X			

* indicates that the Privacy Policy is found embedded in the TOS
( ) indicates the policy directly mentions "harassment"

Table 2. Policy documents present within dataset

abuse, bullying, harm, hate, stalking, and threats. The term "harassment" and its derivations only appear in 37.5% of all documents in our dataset (see Table 2). All but two platforms (Vine and VK) either directly mention the term "harassment" or "harass" in their formal and informal policy documents.

Despite the lack of guidance for what it actually means, "harassment" is explicitly mentioned within 21 of the 56 policy documents within our dataset (see Table 2). In comparing the different types of policy documents, we found that six platforms mention harassment only in formal policy documents, two platforms mention harassment only in

	Abuse	Attack	Bullying	Defame	Eating Dis.	Harm	Hate	Impersonate	Intimidate	Libelous	Racist	Self-harm	Self-injury	Self-mutilation	Stalking	Threat	Torture	Vulgar
Facebook		X	X		X		X		X			X	X			X		
Twitter	X	X																
LinkedIn	X					X				X								
Pinterest		X	X		X	X	X					X				X		
Instagram	X		X	X	X		X	X	X			X			X	X		
Tumblr					X							X	X	X	X			
Flickr	X			X		X	X	X	X	X	X				X	X	X	X
Meetup			X			X	X		X	X	X	X			X	X	X	X
Vine								X				X				X		
Classmates	X		X	X			X		X			X			X	X		
VK			X									X	X			X		
Tagged	X		X			X	X			X	X				X	X		X
Ask.fm			X								X	X				X		
MeetMe	X					X										X		
Google+			X				X	X								X		

Table 3. Terms associated with harassment within policy documents

informal policy documents, and five platforms mention them in both formal and informal policy documents. Several platform-specific policies used similar language in describing these unwanted behaviors. In one case (Flickr and MeetMe), it was more than just similar language but actually a direct duplication.

Few platforms delineate directionality of harassing behaviors and activities between those focused on others and those focused on oneself. For example, Ask.fm's TOS states in the rules of conduct that users must not post or send anything that "contains any threat of any kind, including threats of physical violence to yourself or others." Other platforms, like Facebook and Instagram, have informal policy documents that discuss harm to oneself at a more granular level, calling out eating disorders, self-injury, and self-harm. Our analysis shows that over half (eight) of the platforms consider harm to oneself to be a behavior that is important to regulate.

4.2 Responses to Prohibited Activities

There are a variety of different *responses* to these prohibited behaviors outlined in platform-specific policies. We identified nine unique responses that range in scope and severity (see Table 4). For example, some platforms have policies that enable operators to send warnings to community members, delete content or accounts, or work with third party entities like law enforcement, federal governments, or organizations (i.e., the National Center for Missing and Exploited Children). On the other hand, we found little to no distinction between how platform operators respond to harassment and other prohibited activities like hacking and spamming. Policies did not distinguish nuanced responses specifically to *harassment*, instead ascribing general responses to all prohibited activities.

	Restrict account	Send user warning	Remove content	Disable account	Delete account	Work w/Law enforce.	Removed from search	Interface with 3rd party(s)
Facebook		X	X	X		X		X
Twitter					X	X	X	X
LinkedIn	X	X		X	X	X		X
Pinterest			X	X	X	X		X
Instagram			X	X	X	X		X
Tumblr		X	X		X	X	X	X
Flickr	X	X	X	X	X	X		X
Meetup			X	X	X	X		X
Vine		X	X	X	X	X		
Classmates		X	X	X	X	X		X
VK			X	X	X	X		
Tagged			X		X	X		X
Ask.fm			X	X		X		
MeetMe					X	X		X
Google+			X	X	X	X		X

Table 4. Platform responses to harassment and prohibited behaviors, as described in policy documents

As described in these policy documents, the most common response to prohibited activity is for platform operators to cooperate and work with law enforcement officials. Every platform-specific policy in our dataset stated the operator's willingness to work with law enforcement, often under the guise of complying with state and federal laws. Most commonly, these activities included accessing and sharing content and personal information from within the platform. Another potential response is to suspend or delete an individual's account if it is deemed that the TOS has been broken or that a user has violated community rules and standards. All platforms reserved this right.

For platform-specific policies that provide *some* criteria for determining "harassment" (Instagram and Twitter), we found that responses focused on empowering their users to find help. For example, with respect to how Instagram interfaces with third party groups, informal policy documents provide links to major bullying hotlines and help centers like Connect Safely, Stop Bullying, and the National Crime Prevention Center on Cyberbullying, as well as eating disorder hotlines. Tumblr also provides a document that lists free and confidential counseling on several harassment issues, including rape, bullying, suicide, and self-harm. Not only do they provide U.S.-based resources, they also include information for nine other countries.

4.3 Encouraging Positive Norms

Several platform-specific policies described the types of activities listed in Table 3 in informal policy documents like community guidelines and safety tips for parents and teens. In these documents, policies around harassment tend to be more normative than prescriptive. For example, Instagram's community guidelines begin, "We want to foster a positive, diverse community," before enumerating unacceptable content and activities like those found in Table 3.

Other informal policies discuss the prohibited behaviors in a more relatable manner without calling out specifics. For example, Flickr's community guidelines state, "Play nice … Don't be creepy. You know the guy. Don't be that guy." These informal codes of conduct are much less about providing rules and definitions and more about instilling in the user a deeper understanding of what type of behavior is acceptable and condoned by the community.

Within formal policy documents, some platform designers and operators take extra steps to help users understand their policies by providing exemplars. For example, Pinterest provides for credible exceptions to their prohibited activities. In their acceptable use policy, they provide examples of instances of prohibited activity that *are* acceptable. They tease out the importance of taking into consideration the context and the intent of the content that is being shared (see

Image 1). Pinterest was also a site praised for including plain language explanations for policies in prior TOS work [8].

This image of a hate group is used in a documentary way, so it's okay

GRAMMAR NAZI

In this context, a modified swastika is satirical, not hateful

Image 1. Hate speech policy described through exemplars on Pinterest

5. DISCUSSION

Harassment online is a large and growing issue. Currently platform operators are employing various technological measures, including algorithms that detect certain types of language and text [3] or using context of the user's activities to detect harassing behaviors like cyberbullying [4]. We believe it is imperative to understand platform-specific characterizations of harassment before examining effectiveness of policy or the ethically appropriate ways of regulating harassing behaviors online. Once we have this foundation, we can begin to assess the embedded social norms surrounding harassment within online platforms and the roles that written policies play in mitigating and reducing harassment within these communities.

Through our analysis of the policies as written, we found that there was no standard definition of harassment across the fifteen social media platforms. The documentation of activities that could be perceived as harassment was inconsistent across platforms and vague within platforms (see Table 3). This lack of specificity and consistency for what constitutes harassment makes future work on understanding harassment policies at a deeper, more granular level challenging. This ambiguity is exacerbated by issues associated with the directionality of harassment—it is unclear if harassment towards others should be treated in a similar fashion to those who focus their harm on themselves. Harassment towards oneself (e.g. eating disorders, self-mutilation) is in itself a public health issue [22] and not just an offensive behavior which may require a more nuanced and ethically appropriate response.

Is harassment just a container for a set of offensive behaviors, or is it something more complex that is being overlooked? According to official policies, platform operators respond to harassment in the same way they respond to any prohibited activity. This position means that someone who is sending threating messages could potentially be treated in the same manner as someone who is spamming a group's wall with unwanted content. In offline communities, these offenses are regarded differently, with different sets of repercussions.

Online responses to harassment often happen away from the eyes of the community. Social acknowledgement and understanding that can come from "witnessing" responses to harassment are stymied, depriving the community from opportunities to learn and voice concern. Because we do not know if the community is aware of the platform operator's responses to harassment, we also do not know if they feel that the actions taken uphold their vision of community norms. This gap highlights the inherent tension within the platforms' governance structures to create open and free spaces for expression while still upholding policies that regulate behaviors.

The manner and severity in which a platform chooses to handle harassment varies across platforms. Our data shows that these range in severity from simple content removal to informing law enforcement. However, the platform operators do not specify which prohibited activities are associated with which responses. Some platforms utilize responses that range in severity (Flickr) while some take a more uniform stance on how they respond to harassment (Twitter) as shown in Table 4. With a lack of consistency in the definition as well as the types and severity of the responses used by the different platforms, community members who engage with multiple platforms have a higher burden of understanding the rules and repercussions of breaking those rules.

Platform operators play a pivotal role when it comes to translating platform policies. Beyond the issues of inconsistency of the characterization of harassment, our data shows that policies are written in complex terms and require higher comprehension levels than the average user—two known barriers to the internalization of policy [8, 11]. Further research is needed to understand the potential differences in how communities receive and engage platform policies. For example, in determining how best to craft policy solutions, a next step might be to assess whether users understand existing materials and if their concept of harassment matches that of the platform(s) they use.

Prior work has shown that when a site's policies are unclear, the rules and norms are often community-constructed and enforced [8] and potentially have more staying power than externally imposed rules [20]. Some online social platforms have successfully integrated these principles by either designing *to support* existing social norms within a community or designing *to encourage* more positive norms [7, 19]. In these spaces, community norms, policies, and technologies are more closely aligned [7]. This observation highlights the importance of integrating policy and design into a cohesive development strategy [12]. To achieve this alignment, platform operators could work with communities through participatory design [1] or purposeful design for social norms [7, 19]. In considering best practices for policies, creating harassment policies *with* the community could empower the creation of positive social norms *for* the community, creating a more favorable approach to platform governance.

6. CONCLUSION

Harassment is a serious issue within online social media platforms. However, platform-specific policies that address harassment currently lack definition and consistency. As a first step in better characterizing the ecosystem of governance around harassment, and in thinking about best practices for harassment policy, we suggest that rules derived from users may have more salience and a better chance of adoption in the community. We should empower community members to contribute to informal harassment policies, harnessing the power of communities to create norms around problematic behaviors. As researchers continue to work towards technological and design solutions to curtail harassment, we hope this work on characterizing the current state of platform policies provides a foundation and greater context for future research.

7. ACKNOWLEDGEMENTS

8. REFERENCES

[1] Ashktorab, Z. and Vitak, J. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems*, 3895–3905.

[2] Brown, K. et al. 2006. Cyber-Bullying: Developing Policy to Direct Responses that are Equitable and Effective in Addressing this Special Form of Bullying. *Journal of Educational Administration*. 57, 1–36.

[3] Chancellor, S. et al. 2016. #thygapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. *Proc. of the ACM Conference on Computer Supported Cooperative Work and Social Computing*.

[4] Dadvar, M. and Jong, F. De. 2012. Cyberbullying detection: a step toward a safer Internet yard. *Proc. of the 21st International Conference Companion on World Wide Web*, 121–125.

[5] DeNardis, L. and Hackl, A.M. 2015. Internet Governance by Social Media Platforms. *Telecommunications Policy*. 39, 9, 761–770.

[6] Djuric, N. et al. 2014. Hate Speech Detection with Comment Embeddings. *Proc. of the 24th International Conference on World Wide Web Companion*, 29–30.

[7] Fiesler, C. et al. 2016. An Archive of Their Own : A Case Study of Feminist HCI and Values in Design. *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems*, 2574–2585.

[8] Fiesler, C. et al. 2016. Reality and Perception of Copyright Terms for Online Content Creation. *Proc. of the 19th ACM Conference on Computer Supported and Cooperative Work & Social Computing*.

[9] Fiesler, C. et al. 2015. Understanding Copyright Law in Online Creative Communities. *Proc. of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 116–129.

[10] Gilbert, P. and Irons, C. 2005. Focused therapies and compassionate mind training for shame and self-attacking. *Compassion: Conceptualisations, research and use in psychotherapy*. 263–325.

[11] Good, N.S. et al. 2007. Noticing notice: a large-scale experiment on the timing of software license agreements. *Proc. of the SIGCHI conference on Human factors in computing systems*, 607.

[12] Jackson, S. et al. 2014. The policy knot: re-integrating policy, practice and desgin in CSCW studies of social computing. *Proc. of the 17th ACM Conference on Computer Supported Cooperateive Work & Social Computing*, 588–602.

[13] Jackson, S. et al. 2014. The Policy Knot: Re-integrating Policy, Practice and Design in CSCW Studies of Social Computing. *Proc. of the Conf. on Computer Supported Cooperative Work & Social Computing*, 588–602.

[14] Jensen, C. and Potts, C. 2014. Privacy Policies as Decision-Making Tools: An Evaluation of Online Privacy Notices. *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 471–478.

[15] Lewis, S.P. et al. 2011. The scope of nonsuicidal self-injury on YouTube. *Pediatrics*. 127, 3 (2011), e552–e557.

[16] Luger, E. et al. 2013. Consent for All: Revealing the Hidden Complexity of Terms and Conditions. *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, 2687–2696.

[17] Massanari, A. 2015. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*.

[18] Meredith, J.P. 2011. Combating Cyberbullying Emphasizing Education over Criminalization.pdf. *Federal Comm. Law Journal*. 63, 1 (2011), 311–340.

[19] Morgan, J.T. et al. 2013. Tea & Sympathy : Crafting Positive New User Experiences on Wikipedia. *Proc. of the 2013 conference on Computer supported cooperative Work and Social Computing*, 839–848.

[20] Ostrom, E. 2015. *Governing the Commons*. Cambridge Univeristy Press.

[21] Panko, R.R. and Beh, H.G. 2002. Monitoring for pornography and sexual harassment. *Comm. of the ACM*. 45, 1 (2002), 84–87.

[22] Pater, J. et al. 2016. "Hunger Hurts but Starving Works:" Characterzing the Presentation of Eating Disorders Online. *Proceedings of the 2016 Conference on Computer Supported Cooperative Work and Social Computing*.

[23] Wolak, J. et al. 2007. Does Online Harassment Constitute Bullying? An Exploration of Online Harassment by Known Peers and Online-Only Contacts. *Journal of Adolescent Health*. 41, 6 SUPPL. (2007).

[24] Yin, D. et al. 2009. Detection of Harassment on Web 2.0. *In Proc. of the Content Analysis in the WEB 2.0 Workshop at WWW*.