International Journal of Communication 10(2016), 5164–5185 1932–8036/20160005

Social Media Companies' Cyberbullying Policies

TIJANA MILOSEVIC[1]
University of Oslo, Norway

This article examines social media companies' responsibility in addressing cyberbullying among children. Through an analysis of companies' bullying policies and mechanisms that they develop to address bullying, I examine the available evidence of the effectiveness of the current self-regulatory system. Relying on the privatization-of-the-digital-public-sphere framework, this article signals concerns regarding transparency and accountability and explains the process through which these policies develop and can influence the perceptions of regulators about what constitutes a safe platform. The article is based on a qualitative analysis of 14 social media companies' policies and interviews with social media company representatives, representatives of nongovernmental organizations, and e-safety experts from the United States and the European Union.

Keywords: cyberbullying, social media, online platforms, intermediaries, digital public sphere, digital bullying, freedom of speech, privacy, e-safety, youth and media, children

When 14-year-old Hannah Smith died by suicide, she had allegedly been cyberbullied on Ask.fm (Smith-Spark, 2013). Anonymous questions are a hallmark of the social networking site, available in 150 countries with 150 million users, around half of whom were under 18 at the time (Ask.fm, 2016; Henley, 2013). Ask.fm suffered public rebuke (UK Government and Parliament, n.d.) and the UK prime minister asked its advertisers to boycott the site. Yet, a year after the suicide, the coroner's report concluded that the girl had been sending harassing messages to herself and no evidence of cyberbullying was found (Davies, 2014).

Although the case of Hannah Smith is an anomaly because cyberbullying did not seem to take place, it nonetheless joins a long list of actual cyberbullying incidents on social media platforms that drew

Tijana Milosevic: tijana.milosevic@media.uio.no
Date submitted: 2016–01–10

[1] This material is from *Cyberbullying Policies of Social Media Companies*, forthcoming from MIT Press, Spring, 2018. I would like to thank my doctoral dissertation committee for their invaluable guidance and especially Dr. Laura DeNardis, Dr. Kathryn Montgomery and Dr. Patricia Aufderheide for their continuous support; Dr. Sonia Livingstone for her kind help in securing the interviews; Dr. Elisabeth Staksrud for her thoughtful feedback on this article and support of my research; and anonymous reviewers for their constructive and helpful feedback. The research was supported by American University's Doctoral Dissertation Research Award.

public attention because of their connection to self-harm (Bazelon, 2013). Such cases can put pressure on companies' businesses and influence the development of policies and mechanisms to address cyberbullying.

Cyberbullying policies are enforced through self-regulatory mechanisms that social media companies have in place to address incidents on their platforms. These mechanisms can include reporting tools, blocking and filtering software, geofencing,[2] human or automated moderation systems such as supervised machine learning, as well as antibullying educational materials. Companies tend to provide tools for their users to report a user or content that they find abusive. After looking into the case, the company can decide whether the reported content violates its policy and hence whether it wants to block the user who posted it, remove the abusive content, or take some other action (O'Neill, 2014a, 2014b). Some companies also develop educational materials in cooperation with e-safety nongovernmental organizations (NGOs) that teach children about positive online relationships in an effort to prevent bullying.

Although social media companies' official policies tend to be written on their websites, these policies do not always explain how the mechanisms against bullying work. Social media platforms are online intermediaries that enable user-generated content and allow for interactivity among users and direct engagement with the content (DeNardis & Hackl, 2015). In the United States, under Section 230 of the Communications Decency Act, online intermediaries are exempt from liability for cyberbullying incidents that take place on their platforms on the grounds of being intermediaries only, which means that they do not get involved with content. However, social media platforms' policies against cyberbullying and the mechanisms of their enforcement include extensive involvement with content, which can put their intermediary status into question. In a number of countries, specific laws have provisions that ask the companies to collaborate with law enforcement to reveal the identity of perpetrators (Dyer, 2014) or to take specific content down upon the request of government representatives, such as a child commissioner (Australian Government, Office of the Children's eSafety Commissioner, n.d.). However, no laws stipulate which mechanisms every social media company must develop to address bullying.

An Underresearched Area

A limited amount of academic research addresses this aspect of online intermediation. Previous studies have either focused on only one platform or examined a broader range or harassment concerning adults (Citron, 2014; Matias et al., 2015), raising issues about effectiveness and how a lack of transparency of specific mechanisms such as flagging leaves users with few options and can limit companies' responsibility (Crawford & Gillespie, 2016). Others have proposed theoretical solutions for the reported ineffectiveness of some aspects of these mechanisms (van der Zwaan, Dignum, Jonker, & van der Hof, 2014) or examined the effectiveness of reporting in the context of sexual harassment (Van Royen, Poels, & Vandebosch, 2016). The few studies that refer specifically to cyberbullying among children and adolescents did not set out to provide a systematic analysis of how the effectiveness of mechanisms is

[2] Geofencing leverages the global positioning system to ban certain geographic locations from accessing a social media platform.

conceptualized and measured across a range of the most popular social media companies among youth (Bazelon, 2013). Bazelon's study raised concerns about the effectiveness of Facebook's mechanisms and the company's ability to handle reported cases in a timely manner. In a competitive environment where e-safety is part of the corporate business model, no study has analyzed long-term trends in policy development that continue to exist, even when individual companies that adopted them may no longer be on the market—a topic that this research seeks to address.

Government-initiated independent evaluations of the effectiveness of these mechanisms have been sparse, and are by now dated (Donoso, 2011; Staksrud & Lobe, 2010), and raise significant concerns about the effectiveness of the mechanisms provided, which is another reason for undertaking this study. This study is limited by a paucity of data and the lack of previous research on this topic as well as by the fact that only one company in the sample did not originate in the United States. This article does not set out to provide an analysis of companies' actions in the face of meeting legal requirements in different geographic locations or how different views on privacy and freedom of speech in the United States and the European Union may be reflected in companies' actions (see, e.g., Schwartz & Solove 2014; Whitman, 2004).

Scope of Study

This study seeks to explain the process of online intermediation in intervening with and preventing cyberbullying cases concerning children and teens. What can be known about the relative effectiveness of companies' mechanisms to address cyberbullying? This research is based on a qualitative analysis of 14 social media companies' terms of service and all other publicly available corporate documents that reference "abuse," "harassment," or "(cyber)bullying." I also conducted 27 interviews with social media company representatives, the representatives of e-safety nongovernmental organizations that work with companies on developing their cyberbullying policies in the United States and the European Union, independent e-safety consultants, and academics and government representatives working on e-safety.

Literature Review

Although *cyberbullying* is a term that social media companies tend to use interchangeably with *online harassment*, it is nonetheless a specific form of peer aggression that is distinct from online harassment in general (Ševčíková & Šmahel, 2009) and from other specific examples such as trolling or revenge porn. Cyberbullying is not easily defined and operationalized (Tokunaga, 2010; Vandebosch & Van Cleemput, 2009). Similar to off-line bullying, research considers cyberbullying as "willful and repeated harm inflicted" toward another person, which involves a power imbalance (Hinduja & Patchin, 2009, p. 5). No academic consensus exists on how many times abuse needs to happen in order to classify as "repeated." Although there are disagreements about the prevalence of cyberbullying (Kowalski, Giumetti, Schroeder, & Lattanner, 2014), some recent studies in different countries find an increase in the frequency of cyberbullying across various age groups (EU Kids Online, 2014). Hurtful information includes, but is not limited to, offensive remarks, deception, spreading gossip via digital media, and exclusion from an online community (Vandebosch & Van Cleemput, 2009).

Masking one's identity in the online world (anonymity) as a contributing factor to cyberbullying, which can desensitize the perpetrator, is frequently emphasized in the literature (Hinduja & Patchin, 2009; Vandebosch & Van Cleemput, 2009). However, some research suggests that children know who the so-called perpetrators are, and cyberbullying can happen within close relationships between one-time friends (Kowalski et al., 2014; Mishna, 2012;). Some studies show a strong overlap between cyberbullying perpetration and victimization as well as between cyberbullying and off-line bullying (Görzig & Machackova, 2015). The role of bystander behavior and engagement in helping victims is an increasingly important area of cyberbullying research (Bastiaensens et al., 2014).

Theoretical Framework: Privatization of the Digital Public Sphere

Of particular relevance for this work is the privatization-of-the-digital-public-sphere framework, which examines ways in which the increasing role of online intermediaries affects civil liberties and democratic culture that this public sphere affords (DeNardis, 2014; Gillespie, 2010, 2015; Hestres, 2013) The social media companies of relevance for this study are corporations that delineate the boundaries of which content is allowed on their platforms or which user behavior is permitted through end-user agreements such as terms of service and community guidelines/standards. "This private ordering, rather than (or in addition to), laws, norms, or governments determines the conditions of freedom of expression in the public sphere" (DeNardis, 2014, p. 158).

For instance, Facebook decided to allow users to post breast-feeding images, which had previously been banned because of nudity concerns (Facebook, n.d.). However, the company does not allow images of childbirth with explicitly displayed parts of a woman's body (Hill, 2014). Facebook recently caused significant controversy over banning an iconic Vietnam War photo because it contained an image of a nude child, which was against the company policy, only to reinstate the photo after public pressure (Scott & Isaac, 2016, para. 1). It is not clear which factors influence the companies' decision making—to what extent it is the interests from pressure groups and the media, purported editorial values of the company, or economic interests combined with the perceived will of the majority of users. Similarly, by deciding which content constitutes bullying or is offensive to someone, the companies regulate free speech on their platforms. Protecting one person's privacy can mean curbing someone else's free speech (DeNardis, 2014; Solove, 2007).

Yet social media companies' business models have well-documented implications for end-users' privacy (Cohen, 2012; DeNardis, 2014; Solove, 2007; Vaidhyanathan, 2008, 2011). While social media companies offer their services for free, they embed online advertising strategies, which allows them to target users based on their characteristics in an effective manner and sell data to third parties. What is not always clear, however, is the interplay of companies' economic incentives and decisions regarding censorship and content removal in connection to bullying.

A number of works analyze how private companies participate in the discursive production of sharing and human connection and address the transparency of social media companies' business models. Most of the companies analyzed here use the word *platform* to describe their ethos and activity. Tarleton Gillespie analyzed the discursive deployment of the word *platform* to demonstrate how these companies

use it strategically to frame their services in ways that allow them to "both pursue current and future profits, to strike a regulatory sweet spot between legislative protections that benefit them and obligations that do not, and to lay out a cultural imaginary within which their service makes sense" (Gillespie, 2010, p. 348; cf. Wyatt, 2004). This study analyzes the implications behind the companies' provision of cyberbullying policies, which are germane to the tension between cultivating a community of young users and ensuring advertising revenue as well as for the issue of whether to intervene with content or remain neutral.

José van Dijck (2013) explains how coded structures engineered by these platforms essentially alter the nature of user connections and interactions. The crucial by-product of a sharing culture and peer production on these platforms are valuable behavioral data that companies monetize through advertising. Platform owners tend to call for more transparency and openness from their users, which ensures "maximum sharing and frictionless online traffic" (van Dijck, 2013, p. 21) while failing to apply the same transparency standards to their own practices and business models.

Self-Regulation

Industry self-regulation can be defined as the creation and enforcement of rules by a set of actors, especially the industry, with little or no state intervention (Lievens, 2010, 2016). E-safety nongovernmental organizations play an active role in this multistakeholder self-regulatory system in several important ways (Staksrud, 2013; Taraszow, 2013). For instance, they can pilot the mechanisms for the companies, help users bring cases to companies' attention, design antibullying educational materials for the companies' websites, or run help lines where companies can refer children who might have a bullying problem. This is why interviews with e-safety NGO representatives have been conducted as part of this research.

Among cited rationales for adopting self-regulation over traditional command-and-control regulation is its ability to adapt to fast-paced technological developments (Latzer, Just, & Saurwein, 2013). Drawbacks include insufficiently effective enforcement, limited sanctions for transgressors, less transparency and accountability (McLaughlin, 2013), and favoring private interests over public interests (Lievens, 2010; Tambini, Leonardi, & Marsden, 2008). Industry self-regulation is sometimes described as the regulation "by raised eyebrow" (McLaughlin, 2013, p. 81), meaning that when policy makers are dissatisfied with how the industry self-regulates, they indicate the possibility of legislation in an ambiguous manner, which acts as an incentive for the industry to improve.

Thus far, globally, independent evaluations of mechanisms that social media companies provide to address bullying among children in particular have been scarce.[3] Children's status as minors guarantees them heightened standards of protection, which applies to harm from bullying; however, they also have the right to provision and participation, and striking a balance between these can be a challenge (Livingstone, 2016). Independent evaluations were conducted as part of two self-regulatory frameworks in

[3] The term *children*, for the purposes of this study, refers to anyone under the age of 18, as specified in the United Nations Convention on the Rights of the Child (Livingstone, 2009, 2016).

the EU and one self-regulatory initiative in the United States. In the EU, within the Safer Social Networking Principles framework, researchers analyzed self-declaration statements and tested a number of services offered by these companies (Staksrud & Lobe, 2010). A follow-up evaluation was conducted a year later (Donoso, 2011). These evaluations found multiple issues with companies' ability to address users' complaints effectively. The most recent independent evaluation in the EU was conducted as part of the ICT Coalition, which outlined the need for improved reporting in reference to mobile platforms. However, in this case, the research did not carry out actual tests of the tools provided by the companies (O'Neill, 2014b). This was also the case with the only independent evaluation relevant to bullying conducted in the United States as part of the Internet Safety Technical Task Force (Berkman Center for Internet & Society at Harvard University, 2008). With this background in mind, the study poses the following research questions:

RQ1: What can be known about the relative effectiveness of social media companies' cyberbullying policies and mechanisms against bullying?

RQ2: How do companies conceptualize the effectiveness of antibullying policies and mechanisms on their platforms, and what evidence do they provide of this effectiveness?

RQ3: How might cyberbullying policies be related to the process of "monetize[ing] engineered streams of information" (van Dijck, 2013, p. 12), and what role could the lack of transparency play in this process?

RQ4: How might companies present information about their policies in a way that minimizes the perception of any conflict of interest in companies' services to various constituencies, from advertisers to users and policy makers (Gillespie, 2010)?

Method

The companies analyzed in this study were selected based on the following criteria: the high number of users they had at the time of this study's execution (see the appendix); their reputation as being popular among teen and preteen users (e.g., Elgersma, 2016; Lenhart, 2015); and their appearance in the media in relationship to bullying, especially in reference to self-harm or suicide of children and teens (e.g., Wallace, 2015). Furthermore, the platforms selected in the sample have varied technological affordances, allowing for the study to capture a range of responses. Social media companies included in this sample are Facebook, Instagram (Facebook-owned), Twitter, Ask.fm, YouTube (Google-owned), Yik Yak, Secret app, Google+, Tumblr (Yahoo! owned), Snapchat, Whisper, and messenger apps Voxer, WhatsApp (Facebook-owned), and KIK. Company representatives interviewed were not members of public relations departments; rather, they were either CEOs or had titles such as head of global safety, director of public policy, and product manager, safety.

Apps can gain (or lose) a large user base as well as media attention in the course of a couple of years or even months. This is why the analysis presented here does not purport to be comprehensive.

Rather, it synthesizes key features of and similarities and differences in the policies and enforcement mechanisms, which form the basis of the industry's private regulation, and which will inform the policies of companies that are yet to appear on the market.

The community of e-safety NGOs that work with social media companies is a relatively small, tight-knit community where experts tend to know one another. For instance, at the time of this research, Twitter listed 30 organizations as expert resources in handling online abuse, and the list included both government agencies and NGOs. Facebook labeled 20 international NGOs and experts as those able to provide advice on bullying. These lists are updated periodically. Ask.fm listed 7 NGOs as its trusted partners. Some companies did not cooperate with NGOs at all. After interviewing five NGOs and e-safety consultants, the boundaries of the community of NGOs that cooperated with social media companies on cyberbullying became clear. Furthermore, social media companies rely on multiple NGOs for the same purpose—for instance, to test new cyberbullying prevention features. Hence, by interviewing multiple NGOs that fulfill the same purpose, I was able to arrive at a level of saturation.

The interviews with social media companies' representatives were difficult to obtain, and I would not have secured most of the interviews without connections. They took place on the companies' and NGOs' premises, during international conferences, on Skype, or via e-mail. Some companies ask all visitors to sign nondisclosure agreements (NDAs) to enter the premises. The interviewees cleared me from the NDAs under the condition that the interview quotes would be approved prior to inclusion in the study. Some of the quotes were changed or omitted as a result of this process. NDAs are a standard practice in the industry (Carr, 2013).

Results

Definitions of Bullying, Harassment, and Threats

The corporate documents of all the companies in the sample stipulate that harassment, abuse, or cyberbullying are not allowed on their platforms. However, not all the companies provide a rationale for why a specific term was chosen, and they do not disclose the criteria for determining whether a case constitutes cyberbullying. Overall, older companies in the sample that are perceived by e-safety consultants and NGO representatives as having significant resources to invest in e-safety (e.g., Facebook, Twitter, YouTube) and are often characterized in the interviews as "the more established companies," and those that have undergone significant e-safety audits after high-profile cyberbullying incidents (e.g., Ask.fm) provide more information in their policies about what they consider to be cyberbullying and how their enforcement mechanisms work.

To act on cyberbullying content, the companies need to first establish that such content is cyberbullying and violates the company policy. Research-based cyberbullying definitions typically stipulate that behavior needs to be "repeated." In light of various technological affordances, what counts as repeated on a platform? Is only one mean post enough for a behavior to qualify as cyberbullying? Only a fraction of companies provide answers to these questions. Facebook's and Facebook-owned Instagram's representatives, for instance, were aware of research-based definitions that include the notion of

repetition. They explained that the company's moderators could, however, take content down as cyberbullying even if it were only one mean post, because it may be part of a larger incident taking place off-line. For YouTube, cyberbullying, which the company considered interchangeably with harassment, could also constitute one comment only, and according to a YouTube representative: "It's always case by case—who is attacking who and in which manner." Under the circumstances where context determines what constitutes a policy violation, it becomes increasingly important for these companies to have human rather than automatic moderation, which a number of informants characterized as a substantial cost.

Other, less established companies in the sample were less likely to provide examples of what they considered to be cyberbullying in their corporate documents, and the interviewed representatives did not refer to research-based definitions, including what "repeated" or "power imbalance" meant for them. They frequently conflated threats with harassment and did not distinguish cyberbullying as a specific form of harassment.

Cyberbullying and Technological Affordances

A number of company and NGO representatives emphasized that some platforms lend themselves to subtle forms of bullying, while other sites witness "more blatant" bullying. The nature of cyberbullying on these platforms has important implications for how antibullying mechanisms were designed and how their effectiveness is conceptualized.

For instance, Facebook's moderators might receive a photo of two children smiling at each other that is reported to them as cyberbullying. Since the photo itself does not have any insignia of cyberbullying, the company's moderators cannot take the photo down unless they can establish that the content indeed violates the company's policy based on the guidelines that the company uses. On Facebook, children do not necessarily need to write mean comments or swear words on each other's profiles in order to bully each other. Photos that are ironic can be shared while tagging the bullied person, for instance. Similarly, on Twitter, a practice called subtweeting refers to tweeting about someone without using the person's handle and mentioning him or her—a form of irony that allows for subtle bullying that is difficult for the platform to establish as bullying.

On Ask.fm, on the other hand, bullying could take place in a more overt manner: Swear words and openly mean questions are more common. Likewise, the affordances of other anonymous apps such as Whisper, Secret, and Yik Yak, where users' posted content is not attached to user names or profiles, were reported to be more conducive to blatant bullying. Hence, enforcement in such cases may be more straightforward and bullying cases more evident. Both NGO informants and companies that do not provide for anonymous sharing often characterized anonymous platforms as particularly conducive to cyberbullying. Yet, since no platform in the sample discloses the statistics on the incidence of bullying cases detected or processed, it is difficult to verify such observations.

On platforms where cyberbullying is more blatant (e.g., Ask.fm, Secret, Whisper, Yik Yak), the companies may employ filtering for prohibited words (which were nonetheless not disclosed). They do so in different languages in the markets where they have significant numbers of users. On the other hand,

the companies that particularly emphasize free speech on their platforms (e.g., Facebook, Twitter, and YouTube) report being especially reluctant to employ such filtering.

Proactive Content Moderation

Supervised machine learning is a form of proactive content moderation—automatic crawling (monitoring) of the network that allows for the detection of cyberbullying cases as they happen (Dinakar, Jones, Havasi, Lieberman, & Picard, 2012; Xu, Jun, Zhu, & Bellmore, 2012). Thus, content could be flagged as potentially cyberbullying even before a user reports it to the company.

Companies that had a rationale for *not* using supervised machine learning (e.g., Facebook, Twitter, YouTube) seemed to have a consensus for why they preferred not to use it: Cyberbullying is context-dependent and varies from case to case, which is why supervised machine learning may result in a lot of false positives and much of the valid content could be taken down, potentially infringing upon users' freedom of speech. For example, the word *bitch* is often used by youth to mean *friend* or *mate*. Algorithms used as part of supervised machine learning may signal the word *bitch* as problematic, whereas this is often not the case. On the other hand, cyberbullying on these platforms can be subtle and involve irony, and hence the algorithm may not be able to detect it. Some companies reported privacy concerns as a reason for their reluctance to use it. Applying supervised machine learning to the content that had not been reported or shared with settings set to "public" could amount to surveillance.

On the other hand, the less established companies could have extensive automatic content screening that was not publicly disclosed and that they perceived as effective. For example, in its policy, anonymous app Whisper provides very little information about how it regulates cyberbullying cases. The only enforcement specified is that a violation of terms of service/community guidelines could result in a suspension or blocking of one's account. Similarly, Secret did not[4] provide any explanation in its corporate documents for how it enforced its policy and how it ensured that cyberbullying did not take place. Community guidelines mentioned the option to report bullying content, but there was no explanation about how the reporting system worked or how such posts were handled by the company. However, an elaborate system for handling such posts did exist, as revealed in the interview with the company.

As part of its bullying prevention initiative, Secret employed "advance screening" or "sentiment analysis," which was "automatic." Since no personal names were allowed on the platform, when someone wanted to post a name or a bad word, the system automatically checked the post against a premade database of first and last names and swear words/bad words. If the system detected a name or a bad word, it prompted the user with the following messages: "Are you sure you want to post this"? or "Say something kind"—which should act as a deterrent to posts that violated community guidelines. In addition to sending this automatic deterrent to the user, the system screened the post for "severity level," and if it determined that it was "high" or "significant," the post was not published but withheld and sent to a human moderator for further review. The guidelines and standards for determining the severity level had not been disclosed. Such proactive moderation was similar to Whisper's (Chen, 2014).

[4] The company shut down in 2015.

Once a post was published, a "report" button could be used to flag inappropriate content. If a post was reported by numerous accounts, it was taken down automatically—a procedure that the companies focusing on freedom of speech values would find invasive. Similarly, at the time of this research, Yik Yak did not publicize the details behind enforcement of its policy. However, an elaborate system that was, to an extent, similar to Secret's did exist.

In addition to company moderation, both Secret and Yik Yak relied on their user communities to moderate themselves, and this mechanism that deferred moderation from the company to users was perceived as especially effective by the companies. For example, users can either "upvote" (similar to Facebook's "like") or "downvote" a post. A specific number of downvotes (in Yik Yak's case, five) makes a post automatically come down. The policies aimed at empowering users were seen by e-safety experts as effective as well.

Transparency in Moderator Guidelines

At the time of this research, none of the companies in the sample publicized the definitions that its moderators used to assess whether a reported case constituted cyberbullying, abuse, or harassment. The more established companies provided examples of what they consider to be cyberbullying or harassment, but they nonetheless did not reveal the guidelines that they provided for their moderators. And while NGO representatives tended to hold the view that newer companies present a more substantial problem because their user base can grow more quickly than their e-safety expertise, it was not necessarily the case that the more established companies were more transparent.

Whisper was a rare company that publicly stated it employed 120 human moderators and also named the companies it outsourced its moderation to (Ortutay, 2014). No company revealed how many moderators could handle cases in different languages spoken in various markets where the platform was available. Given the subtleties of cyberbullying, understanding linguistic and cultural nuances is recognized by a number of companies as being especially important in being able to effectively moderate the content.

Evolution of Self-Regulatory Effort: Advanced Mechanisms

In addition to the terms of service, the older the company, the more likely it is to have other corporate documents such as "principles" or "community standards/guidelines/rules," which elaborate cyberbullying-related provisions and their enforcement. Some companies also created "safety centers," which can refer users to NGOs for assistance with cyberbullying or can provide information about cyberbullying for children, parents, or educators. These safety centers can also contain videos and educational texts about bullying that the companies develop with NGOs; these materials are described as empowering young users and providing them with tools to help themselves (and solicit assistance of parents, caregivers, educators, and NGOs in doing so). NGO representatives characterize these safety centers as an effective policy and an important evolution in self-regulation. The companies and NGOs did not publish data on the extent to which children and teens were using these materials and whether they found them useful in helping them resolve bullying incidents.

Freedom Under Responsibility to Young Users

The more established companies' community guidelines/principles/rules and safety centers emphasize the idea that ensuring bullying-free spaces is a shared responsibility between the company and its users. This ability of the platform to get the community to "regulate," "moderate," or "police itself" (sometimes with the help of NGOs via help lines or NGOs' assistance in the creation of safety centers) was seen by a number of both corporate and NGO interviewees as an advanced or evolved self-regulatory mechanism.

Social reporting is the best example of a policy that relies on the logic of community moderation. Facebook pioneered this tool when it noticed that its moderators were receiving a large amount of reported content that they could not establish constituted cyberbullying. It can involve reaching out to "a bully" on behalf of "the victim" in an effort to resolve the issue without reporting it to the company's moderators. Hence, when social reporting takes place, it happens without any notice being provided to the company's moderators. It is primarily intended as a remedy for the content that users mind or think constitutes cyberbullying but that may not qualify as such according to the company policy.

Social reporting also allows users to reach out to a third party (e.g., parent, teacher, friend) for help when they feel bullied. As part of social reporting, Facebook provides "premade" messages that users can send when reaching out to another user who may have bullied them. The company as well as a number of NGOs perceive helping users resolve conflicts among themselves as a more advanced enforcement tool than the moderators' takedowns, because it has the potential to get to the heart of conflicts, which may originate off-line. Facebook's social reporting was based on an extensive research effort in partnership with academic institutions (Facebook Talks Live, 2013; Tsui, 2014).

Perceived and Measured Effectiveness

A number of e-safety consultants and NGO representatives reported that they did not perceive the established companies to be problematic in terms of having effective policies in place. Rather, they frequently perceived start-ups as a problem because these companies could gain large numbers of users over a short period of time without having adequate e-safety mechanisms in place.

Fact-based evidence of effectiveness, however, is rare. For instance, evidence on how many reported cases companies *actually* handle within 48 or 72 hours is typically not publicly available (nor are any estimates provided by companies independently audited), making it difficult to know if the companies are slow to respond. The companies do not provide any other data on measurement of young users' satisfaction with their cyberbullying resolution tools. Facebook provided some evidence of effectiveness for its social reporting policy based on U.S. teens and the company's privately contracted research (Facebook Talks Live, 2013). The platforms did not disclose information on how they measure the effectiveness of each of the antibullying mechanisms on their platforms.

Within this self-regulatory effort, no governmental body in the United States or in the European Union mandates that companies provide data on effectiveness, nor is there a continuous effort to independently evaluate the effectiveness of antibullying mechanisms.

Discussion

From the theoretical standpoint of privatization of the digital public sphere (DeNardis, 2014) and discursive positioning of the platforms (Crawford & Gillespie, 2016; Gillespie, 2010, 2015; van Dijck, 2013), this study indicates how companies' cyberbullying policies in the context of children can help companies pursue profits, "lay out a cultural imaginary within which their service makes sense," and "strike a regulatory sweet spot between legislative protections that benefit them and obligations that do not" (Gillespie, 2010, p. 348).

While company representatives tend to agree that there has to be a concerted effort on their behalf to address cyberbullying and that cyberbullying content is not tolerable, by emphasizing that they may not be best positioned to intervene in children's conflicts and that cyberbullying is a behavioral phenomenon that will take place no matter what the companies do, they downplay their own responsibility for the extent to which technologies may facilitate such behavior or at least provide a venue for it (and monetize it). The companies perceive the tools that enable community self-moderation as advanced and effective, because they see users as better positioned to understand the context behind their own conflicts. The texts that describe such policies represent the user as an empowered actor and emphasize the shared nature of responsibility for e-safety between users and platforms. From this perspective, standard reporting to companies' moderators, which can result in content removal, is not as advanced a policy because it does not necessarily address the heart of the conflict or solve the bullying incident, which can continue off-line. Standard policies such as flagging may allow companies to deflect responsibility through a lack of transparency in how such flagging is handled or lead users to "shoehorn" their "complex feelings into pull-down menu in order to be heard" (Crawford & Gillespie, 2016, p. 424). However, the move toward advanced policies, which are meant to address those concerns, shifts part of the responsibility for handling cyberbullying away from companies by providing children and caregivers with tools to address it themselves or by referring them to NGOs.

Yet it is questionable to what extent users are empowered and these mechanisms effective given that companies do not publicly provide a set of standards by which they evaluate this effectiveness. Only some of the more established companies employ extensive research-based efforts, and only one company offered statistics on the effectiveness of its advanced policy, and based only on a sample of U.S. teens although the company operates globally. Consider also that children's and teen's status as minors puts into question the extent to which such shifting of responsibility for risky behavior (even when the tools are described as specifically designed for the empowerment and education of children) can be applied to these young users (see Staksrud, 2013). From the companies' perspective, advanced policies could be seen as empowering children through provision and participation rights guaranteed to children by the United Nations Convention on the Rights of the Child (see, e.g., Livingstone, 2016). Nonetheless, this only makes a stronger case for giving children a voice in whether these advanced mechanisms are working for them,

and the shared nature of responsibility that such policies imply places the burden of evidence of their effectiveness on the companies.

Further, the focus on e-safety as a joint effort elides any mention of the companies' interests in pursuing profits. One reason why companies characterize community self-regulation efforts as "effective" is "efficiency." The tools that allow users to moderate themselves signify that companies can delegate part of the responsibility onto users and thus receive fewer reports, many of which companies may not be able to act upon, because their moderators would not have enough evidence to establish that a case constitutes cyberbullying. Having in mind the vast amount of content that is shared on these platforms[5] and the subsequent cost of hiring human moderators, community self-moderation can also be beneficial for the companies' business models.

One such elided question from the standpoint of profits is how those policies that are characterized by companies as "(in)effective" affect "frictionless sharing" inherent to companies' business models (van Dijck, 2013, p. 58). Might some other mechanisms be effective from children's perspective, yet are not pursued by companies because they go against their business models? These questions can hardly be answered without an independent assessment of the effectiveness of the current policies and enforcement mechanisms from children's perspective, which currently does not exist. An independent assessment, for instance, would not be commissioned by, paid for, or executed by the companies themselves.

Regarding self-regulation and relationships with regulators, the more established companies in the sample understand the value of e-safety for the success of their business models, and tend to invest significant research-based efforts into policy enforcement. This value of e-safety is reflected in the ability of the company to leverage the policy to assure its users that the platform is a safe one, and thus minimize the consequences of high-profile cyberbullying incidents. Companies are then also able to cite these efforts to regulators as evidence that they are doing their best to address cyberbullying, thus averting regulation and minimizing conflict resulting from the lack of transparency. If self-regulation in respect to cyberbullying and children is to be conceptualized as "regulation by raised eyebrow" (McLaughlin, 2013), then the raised eyebrow, even in the form of independent evolutions, seems to be, to a large extent, missing, given multiple indications of ineffectiveness (Bazelon, 2013; Donoso, 2011; Staksrud & Lobe, 2010). Companies should be more specific about what they mean by cyberbullying, abuse, and harassment, because this is the necessary step toward understanding how they handle these cases. In the absence of an independent evaluation, or even in-house evaluation in most of the cases, the public, however, cannot know against which standards of effectiveness these policies and enforcement mechanisms are evaluated and, in the light of such standards, how effective they are. Until there are independently established standards of effectiveness of antibullying policies, which would need to take into account the companies' diverse technological affordances, it will remain difficult to even discuss the concept of effectiveness.

[5] On YouTube, shared content is 300 hours of uploaded video per minute; on Ask.fm, it is 20,000 questions per minute.

Although anonymous platforms have a particularly negative reputation and tend to be seen by nonanonymous companies and some NGOs as especially conducive to bullying, given that no company releases statistics on the number of incidents it processes in a specific time frame, there is little evidence to support this proposition. Anonymity can be an enabling technological affordance for freedom of expression of political minorities, especially in oppressive societies (DeNardis, 2014), and its dismissal should be considered with care.

These policies have rarely been conceptualized as an important aspect of an increasingly privatized digital public sphere, where the companies outgrow their intermediary function. A call for transparency in this respect is *not* an end goal in itself, but rather a necessary stepping-stone toward understanding which solutions are effective from young users' perspective.

Limitations and Future Research

Leveraging different methods may help overcome the inherent limitations of interviewing company representatives who, even when they are not representatives of PR departments, remain restricted in what they are able to say. Ethnography and participant observation, for instance, may turn this limitation into an aspect of analysis. The same applies to the work of NGOs whose relationship with companies is complex. Nonetheless, given the widespread use of NDAs in the industry, securing permission for such research may prove difficult. Researchers could consider creating fictitious profiles of child users on various platforms and attempt to test enforcement mechanisms (Staksrud & Lobe, 2010).[6] Finally, surveys, focus groups, in-depth interviews, and participant observation with children with the goal of understanding how they use these enforcement mechanisms and think about effectiveness could also constitute a way forward.

References

Ask.fm. (2016). *About us.* Retrieved from http://about.ask.fm/about/

Australian Government, Office of the Children's eSafety Commissioner. (n.d.). *The Enhancing Online Safety for Children Act 2015.* Retrieved from https://www.esafety.gov.au/about-the-office/legislation

Bastiaensens, S., Vandebosch, H., Poels, K., Van Cleemput, K., DeSmet, A., & De Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior*, *31*, 259–271.

Bazelon, E. (2013). *Sticks and stones: Defeating the culture of bullying and rediscovering the power of character and empathy*. New York, NY: Random House.

[6] For a discussion on ethics regarding fictitious profiles, see Staksrud (2015).

Berkman Center for Internet & Society at Harvard University. (2008, December 31). *Enhancing child safety and online technologies: Final report of the Internet Safety Technical Task Force.* Cambridge, MA: Author. Retrieved from http://cyber.law.harvard.edu/sites/cyber.law.harvard.edu/files/ISTTF_Final_Report.pdf

Carr, J. (2013, January 16). Non-disclosure agreements—enforced silence? *Desiderata*. Retrieved from https://johnc1912.wordpress.com/2013/01/16/non-disclosure-agreements-enforced-silence/

Chen, A. (2014, October 23). The laborers who keep dick pics and beheadings out of your Facebook feed. *Wired*. Retrieved from http://www.wired.com/2014/10/content-moderation/

Citron, D. (2014). *Hate crimes in cyberspace*. Cambridge, MA: Harvard University Press.

Cohen, J. E. (2012). *Configuring the networked self: Law, code, and the play of everyday practice*. New Haven, CT: Yale University Press.

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, *18*(3), 410–428. doi:10.1177/1461444814543163

Davies, C. (2014, May 6). Hannah Smith wrote "vile" posts to herself, say police. *The Guardian*. Retrieved from http://www.theguardian.com/uk-news/2014/may/06/hannah-smith-suicide-teenager-cyber-bullying-inquests

DeNardis, L. (2014). *The global war for Internet governance*. New Haven, CT: Yale University Press.

DeNardis, L., & Hackl, A. M. (2015). Internet governance by social media platforms. *Telecommunications Policy*, *39*(9), 761–770.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., & Picard, R. (2012). Common sense reasoning for detection, prevention and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligence Systems*, *2*(3), 1–30.

Donoso, V. (2011). *Results of the assessment of the implementation of the Safer Social Networking Principles for the EU. Individual reports of testing of 14 social networking sites.* Retrieved from https://lirias.kuleuven.be/bitstream/123456789/458077/1/Individual+Reports+SNS+Phase+A.pdf

Dyer, E. (2014, October 20). Cyberbullying bill draws fire from diverse mix of critics. *CBC News*. Retrieved from http://www.cbc.ca/news/politics/cyberbullying-bill-draws-fire-from-diverse-mix-of-critics-1.2803637

Elgersma, C. (2016, February 26). Snapchat, KIK and 6 more iffy messaging apps teens love. *Common Sense Media*. Retrieved from https://www.commonsensemedia.org/blog/snapchat-kik-and-6-more-iffy-messaging-apps-teens-love

EU Kids Online. (2014). *EU Kids Online: Findings, methods, recommendations* (Report). London, UK: London School of Economics. Retrieved from http://eprints.lse.ac.uk/60512/

Facebook. (n.d.). *Does Facebook allow photos of mothers breastfeeding?* Retrieved from https://www.facebook.com/help/340974655932193

Facebook Talks Live. (2013, December 5). Compassion Research Day. Presentation: Bullying: New lessons from teenagers. *Livestream*. Retrieved from http://new.livestream.com/facebooktalkslive/events/2564173

Gillespie, T. (2010). The politics of "platforms." *New Media & Society*, *12*(3), 347–364.

Gillespie, T. (2015). Platforms intervene. *Social Media + Society*, *1,* 1–2. doi:10.1177/2056305115580479

Görzig, A., & Machackova, H. (2015). Cyberbullying from a socio-ecological perspective: A contemporary synthesis of findings from EU Kids Online (Media@LSE Working Paper Series). London, UK: London School of Economics and Political Science. Retrieved from http://www.lse.ac.uk/media@lse/research/mediaWorkingPapers/pdf/WP36-FINAL.pdf

Henley, J. (2013, August 6). Ask.fm: Is there a way to make it safe? *The Guardian*. Retrieved from http://www.theguardian.com/society/2013/aug/06/askfm-way-to-make-it-safe

Hestres. L. E. (2013). App neutrality: Apple's App Store and freedom of expression online. *International Journal of Communication*, *7*(15), 1265–1280.

Hill, M. (2014, October 22). By removing photos of childbirth, Facebook is censoring powerful female images. *The Guardian*. Retrieved from http://www.theguardian.com/commentisfree/2014/oct/22/facebook-removing-childbirth-female-images

Hinduja, S., & Patchin, J. W. (2009). *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Thousand Oaks, CA: SAGE Publications.

Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. *Psychological Bulletin*, *140*(4), 1073–1137.

Latzer, M., Just, N., & Saurwein, F. (2013). Self and co-regulation: Evidence, legitimacy and governance choice. In M. E. Price, S. G. Verhulst, & L. Morgan (Eds.), *Routledge handbook of media law* (pp. 373–397). New York, NY: Routledge.

Lenhart, A. (2015, April 9). *Teens, social media and technology overview 2015.* Washington, DC: Pew Research Center. Retrieved from http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015/

Lievens, E. (2010). *Protecting children in the digital era: The use of alternative regulatory instruments*. Leiden, Netherlands: Martinus Nijhoff.

Lievens, E. (2016). Is self-regulation failing children and young people? Assessing the use of alternative regulatory instruments in the area of social networks. In S. Simpson, H. Van den Bulck, & M. Puppis (Eds.), *European media policy for the twenty-first century: Assessing the past, setting agendas for the future* (pp. 77–94). New York, NY: Routledge.

Livingstone, S. (2009). *Children and the Internet*. London, UK: Polity.

Livingstone, S. (2016). Reframing media effects in terms of children's rights in the digital age. *Journal of Children and Media*, *10*(1), 4–12.

Lynley, M. (2015, April, 29). After passing 10 million monthly active users, Whisper hires its first president. *TechCrunch*. Retrieved from https://techcrunch.com/2015/04/29/whisper-hires-its-first-president/#.hzgvrt:CTAX

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, & the Media*. Retrieved from http://womenactionmedia.org/twitter-report

McLaughlin, S. (2013). Regulation and legislation. In B. O'Neill, E. Staksrud, & S. McLaughlin (Eds.), *Towards a better Internet for children? Policy pillars, players and paradoxes* (pp. 77–91). Gothenburg, Sweden: Nordicom.

Mishna, F. (2012). *Bullying: A guide to research, intervention and prevention*. Oxford, UK: Oxford University Press.

O'Neill, B. (2014a, April). *The ICT Coalition: First report on the implementation of the ICT Principles.* Dublin, Ireland: Dublin Institute of Technology. Retrieved from http://www.ictcoalition.eu/gallery/75/ICT_REPORT.pdf

O'Neill, B. (2014b). *Policy influences and country clusters: A comparative analysis of Internet safety implementation.* London, UK: London School of Economics. Retrieved from http://www.lse.ac.uk/media@lse/research/EUKidsOnline/EU%20Kids%20III/Reports/D6.3-Policy-Influences-May-2014-Final.pdf

Ortutay, B. (2014, March 24). Anonymous apps like Secret and Whisper find a niche in Silicon Valley. *San Jose Mercury News*. Retrieved from http://www.mercurynews.com/business/ci_25409802/anonymous-apps-like-secret-and-whisper-find-niche

Phaneuf, W. (2012, July 26). Source: Voxer approaches 70M users, encouraged by global smartphone adoption. *PandoDaily.* Retrieved from https://pando.com/2012/07/26/voicetext-app-voxer-expands-encouraged-by-global-smartphone-adoption/

Schwartz, P. M., & Solove, D. J. (2014). Reconciling personal information in the United States and in the European Union. Retrieved from http://paulschwartz.net/wp-content/uploads/2014/08/Schwartz%20and%20Solove%20Reconciling%20PII%20(FINAL%20CLR%202014).pdf

Scott, M., & Isaac, M. (2016, September 9). Facebook restores iconic war photo it censored for nudity. *The New York Times.* Retrieved from http://www.nytimes.com/2016/09/10/technology/facebook-vietnam-war-photo-nudity.html?_r=0

Ševčíková, A., & Šmahel, D. (2009). Online harassment and cyberbullying in the Czech Republic. *Journal of Psychology, 217*(4), 227–229.

Smith, D. (2015, March 30). This is the next major messaging app. *Business Insider*. Retrieved from http://www.businessinsider.com/yik-yak-the-next-major-messaging-app-2015-3?r=US&IR=T&IR=T

Smith-Spark, L. (2013, August 9). Hannah Smith suicide fuels calls for action on Ask.fm cyberbullying. *CNN*. Retrieved from http://edition.cnn.com/2013/08/07/world/europe/uk-social-media-bullying/

Solove, D. J. (2007). *The future of reputation: Gossip, rumor, and privacy on the Internet*. New Haven, CT: Yale University Press.

Staksrud, E. (2013). *Children in the online world: Risk, regulation, rights*. Surrey, UK: Ashgate.

Staksrud, E. (2015). Counting children. On research methodology, ethics and policy development. In H. Ingierd & H. Fossheim (Eds.), *Internet research ethics* (pp. 98–121). Oslo, Norway: Cappelen Damm Akademisk.

Staksrud, E., & Lobe, B. (2010, January). *Evaluation of the implementation of the Safer Social Networking Principles for the EU Part 1: General report*. Luxembourg: European Commission. Retrieved from http://ec.europa.eu/information_society/activities/social_networking/docs/final_report/first_part.pdf

Tambini, D., Leonardi, D., & Marsden, C. (2008). The privatization of censorship: Self-regulation and freedom of expression. In D. Tambini, D. Leonardi, & C. Marsden (Eds.), *Codifying cyberspace: Communications self-regulation in the age of Internet convergence* (pp. 269–289). Abingdon, UK: Routledge.

Taraszow, T. (2013). The influence of NGOs on safer Internet policy making. In B. O'Neill, E. Staksrud, & S. McLaughlin (Eds.), *Towards a better Internet for children? Policy pillars, players and paradoxes* (pp. 173–192). Gothenburg, Sweden: Nordicom.

Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in Human Behavior 26*(3), 277–287.

Tsui, B. (2014, March 18). Friends with benefits: Inside Facebook's Compassion Research Day. *Pacific Standard*. Retrieved from http://www.psmag.com/nature-and-technology/friends-benefits-facebook-sociologists-social-media-74781

UK Government and Parliament. (n.d.). *Petition: Government to take a safeguarding children position against sites like Ask.fm.* Retrieved from http://epetitions.direct.gov.uk/petitions/48886

Vaidhyanathan, S. (2008, February 15). Naked in the "Nonopticon": Surveillance and marketing combine to strip away our privacy. *Chronicle of Higher Education*. Retrieved from http://www.chronicle.com/article/naked-in-the-nonopticon/6197

Vaidhyanathan, S. (2011). *The Googlization of everything: (And why we should worry)*. Oakland, CA: University of California Press.

van der Zwaan, J. M., Dignum, V., Jonker, C. M., & van der Hof, S. (2014). On technology against cyberbullying. In J. van den Hoven, N. Doorn, T. Swierstra, B. J. Koops, & H. Romijn (Eds.), *Responsible innovation 1: Innovative solutions for global issues* (369–392). Dordrecht, Netherlands: Springer.

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media.* Oxford, UK: Oxford University Press.

Van Royen, K., Poels, K., & Vandebosch, H. (2016). Help, I am losing control! Examining the reporting of sexual harassment by adolescents to social networking sites. *Cyberpsychology, Behavior and Social Networking*, *19*(1), 16–22.

Vandebosch, H., & Van Cleemput, K. (2009). Cyberbullying among youngsters: Profiles of bullies and victims. *New Media & Society, 11*(8), 1349–1371.

Wallace, K. (2015, January 9). Parents, beware of bullying on sites you've never seen. *CNN.* Retrieved from http://edition.cnn.com/2013/10/10/living/parents-new-apps-bullying/

Whitman, J. Q. (2004). *The two Western cultures of privacy: Dignity versus liberty* (Faculty Scholarship Series Paper 649). Retrieved from http://digitalcommons.law.yale.edu/fss_papers/649

Woollaston, V. (2015, May 1). The Secret's out: Anonymous app shuts down following legal battles and claims that it encouraged bullying. *Daily Mail.* Retrieved from http://www.dailymail.co.uk/sciencetech/article-3064077/The-secret-s-Anonymous-app-shuts-following-legal-battles-claims-encouraged-bullying.html

Wyatt, S. (2004). Danger! Metaphors at work in economics, geophysiology, and the Internet. *Science, Technology, & Human Values, 29*(2), 242–261.

Xu, J. M., Jun, K. S., Zhu, X., & Bellmore, A. (2012, June). *Learning from bullying traces in social media*. Paper presented at the conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Montreal, Canada.

Appendix

E-safety NGOs and Organizations Interviewed

- A member of UK Council for Child Internet Safety, who did not speak on behalf of the organization but in the capacity of an e-safety expert
- A representative of the Child Exploitation and Online Protection Center in the United Kingdom
- South West Grid for Learning
- Childnet
- Family Online Safety Institute
- European Schoolnet
- Member of the Attorney General of Maryland's office
- No Bully
- iKeepSafe
- ConnectSafely.org
- Greater Good Science Center at Berkeley

Not all e-safety experts interviewed were affiliated with organizations, and their views were reported anonymously. The interviews were conducted as part of doctoral dissertation research.

Core Interview Questions for Companies

- Please describe the cyberbullying policies that you currently have.
- How have these policies changed over time?
- In your corporate documents, you use the term "abuse"/"cyberbullying"/"harassment." Why have you chosen to use this particular language?
- How does this language inform the actions you take regarding cyberbullying/harassment/abuse on your platform (in the context of children and teens—those under 18)?
- Which criteria do your moderators use to determine if a case reported on your platform constitutes cyberbullying (how do you define it)?
- Do you use any proactive monitoring of your network—for example, supervised machine learning? Why/why not?
- Have you ever considered introducing "notice-and-takedown" for cyberbullying content? Why/why not?
- How would you characterize the effectiveness—how do you define and measure the effectiveness—of your enforcement tools?
- Do you provide any statistics/official company reports on the number of cyberbullying reports that you receive per unit in time, and how quickly you handle them?
- Do you provide some publicly available evidence of the effectiveness of the enforcement tools?
- Do you provide any information on the amount of investment into efforts to prevent cyberbullying?

- How might the selected enforcement mechanisms impact your business?

Core Interview Questions for NGOs

- Do you work with any of the social media companies (companies listed) on their cyberbullying intervention/prevention efforts, and, if so, could you please describe your role?
- How do you see the role of NGOs in the creation of companies' cyberbullying policies?
- Do you have a formal partnership with the company? Please describe.
- Do you receive any financial (or other) benefits from the company?
- What would you consider to be an effective policy (and enforcement mechanism), and why?
- Might you be able to refer me to evidence of effectiveness?
- How do you perceive the effectiveness of self-regulatory efforts, and does your NGO take a stand on this issue? Please elaborate.

Numbers of Users at the Time of This Research

- Facebook: 1.39 billion monthly active users
- YouTube: More than 1 billion users
- WhatsApp: 700 million monthly active users
- Google+: 300 million monthly active users (however, the popularity of this company has been widely disputed since)
- Tumblr: 420 million users or 224 million blogs
- Instagram: More than 300 million users
- Twitter: 284 million monthly active users
- Kik: 200 million monthly active users
- Ask.fm: 150 million monthly active users
- Snapchat: 100 million monthly active users

The following companies did not release official numbers, which is why I cite media reports where the numbers appeared:

- Whisper: 10 million monthly active users (Lynley, 2015)
- Yik Yak: 3.6 million monthly active users (Smith, 2015)
- Voxer: 70 million users (Phaneuf, 2012)
- Secret: Specific numbers were not reported in the media, but the app generated significant controversy around bullying relevant for this study (e.g., Woollaston, 2015).