Protecting Children Online?

Tijana Milosevic, Sonia Livingstone

Published by The MIT Press

Milosevic, Tijana and Sonia Livingstone.
Protecting Children Online? Cyberbullying Policies of Social Media Companies.
Project MUSE. muse.jhu.edu/book/57843.

For additional information about this book
https://muse.jhu.edu/book/57843

For content related to this chapter
https://muse.jhu.edu/related_content?type=book&id=2082393
It was a sunny Sunday morning in November 2014. I was standing in front of a building in San Francisco, eyeing one of the many floors on which the start-up company I was hoping to interview should have been located. I had tweeted its representative earlier—having found no personal email addresses on the app site—and he was more than kind to answer (something that few did when contacted without a reference). He gave me his address, to which I emailed the interview questions. After several back and forth attempts and little success in setting up a mutually convenient time for the interview, the representative sent a brief message: “I should have some time on Sunday,” he wrote, leaving me wondering when on Sunday, exactly, because that day would be the last one of my stay in the city. So, I took my chances by showing up at the doorstep of the building, semi-uninvited, thrilled to recognize the name of the company on the intercom. I rang the bell and began to explain myself after hearing a male voice at the other end. He let me in, into what I remember as a light, spacious office that belied the covert reputation the company had acquired in the media.

This story reflects substantially more availability than I experienced in my early attempts to interview people from many other companies; it seemed nearly impossible to reach the human beings behind generic contact info on websites, never mind finding a company phone number. Of course, social media representatives are too busy to talk to researchers—imagine if they had to answer dozens or hundreds of queries (or more?) from researchers daily—and I acknowledge how impractical that would be. When I was allowed in, I considered myself lucky, even when I learned very little.

The person whose kind willingness to help began to open the industry’s doors for me was Sonia Livingstone, a professor at the London School of Economics and Political Science. As a well-known academic and founder of the EU Kids Online, a network of more than 150 researchers in 33 countries in Europe who research children and digital media, Livingstone reached
out to some of the company representatives and e-safety experts, asking if she could put me in touch regarding the interviews.

Until then, I had not fully appreciated the extent of nontransparent behavior, perhaps best exemplified in the widespread practice of non-disclosure agreements (NDAs) that many companies would ask visitors to sign upon entering their premises, and an overall lack of perception of any obligation to explain their operational policies to the public. Several e-safety advisers to companies told me that they, too, had to sign NDAs; even they were not always privy to companies’ operational policies. If the advisers were expected to provide critiques on companies’ policies, they may have found it difficult to fulfill their role under such circumstances. It further appeared that this situation was accepted as the (perhaps lamentable) “way things are” in the community of e-safety experts (with few exceptions: see Carr, 2013b), NGOs, policy advisers, and regulators.

In the context of the differences between legalistic versus coordinated self-regulation discussed in chapter 5 (Newman & Bach, 2004), some policy advisers in the EU thought that just getting the US companies to the table as part of the self-regulatory effort had been a significant sign of accomplishment—given that such actions were not something that the US companies had necessarily been accustomed to. Soliciting the good will and cooperation from the industry, as one e-safety expert observed, was important for the EU’s economic competitiveness. But asking US companies to provide more transparency might have been too high of a bar, especially given that they were regarded primarily as private businesses—and less as public utilities or venues where the public sphere unfolds.

As I explained in earlier chapters, each company tailors its intervention and prevention tools to its specific technological affordances, but there are no written rules or minimum standards to which every company must adhere. These minimum standards and sometimes very elaborate policy and enforcement mechanisms tend to emerge nonetheless as part of the self-regulatory and self-organizational efforts that I analyze in this chapter.

Broader questions I pose about this process include: What can be known about the companies’ efforts to address cyberbullying and about the companies’ policy enforcement mechanisms? What do companies consider as an effective policy and enforcement mechanism, and why? What assumptions about the nature and occurrence of bullying on their platforms do companies’ documents and representatives make? How do they explain the rationales about the particular tools of enforcement they use and react to perceived ineffectiveness?
What Does It Mean to Moderate Content?

In the summer of 2016, YouTube came under fire when some of its popular channels accused the company of having “vague” TOS and “a selective approach to moderation” (Kiberd, 2016). (Companies use in-house or outsourced employees as moderators, whose role is to look into content reported for TOS or Community Guidelines/Standards/Rules violations and decide which action to take if the content is determined to violate the company policy.) The allegations made by the channels were elaborated in a video titled “The YouTube Rant (I’m getting banned off YouTube),” which allegedly inspired many others of its kind. The video—which had been uploaded by what a news article posted on the Vice Media platform Motherboard described as an “already infamous” YouTube “‘commentary’ channel” called LeafyIsHere—had at that time already gathered more than 5 million views (LeafyIsHere, 2016). According to the article, it had previously been common to see smaller channels shut down for “roasting” and “occasional parody,” whereas bigger channels “consistently . . . got away with it” (Kiberd, 2016). In an apparent change of policy, however, YouTube was now said to have become stricter in its moderation practices and perhaps more consistent, given the video’s allegation that now even some bigger channels were receiving warnings and restrictions on their platform activity (so-called strikes) for violating its Community Guidelines (Kiberd, 2016).

The interplay of any platform’s business model with its policies on taking down abusive content may be difficult to discern, and it is a particularly interesting question that seems rarely discussed in public. For instance, does the shutting down of widely popular accounts or removal of their content negatively affect any platform’s commercial interests, and if so, to what extent and in which ways? I raise questions such as these throughout this chapter.

Problems regarding the ways in which moderation is put into practice with respect to abuse, harassment, and bullying are by no means peculiar to YouTube; most companies struggle with these behaviors in one way or another.

Definitions, Behaviors, and Levels of Transparency

Most social media platforms include some sort of anti-bullying, anti-abuse, or anti-harassment provisions in their TOS or other corporate documents, whether they’re labeled Community Standards, Community Guidelines,
Rules, or Principles. (Some companies do not make any references to “community,” however, a factor that notably emerged in my analysis.)

Such documents rarely provide a more specific or detailed level of explanation for what these companies consider to be “bullying” on their platforms. Some companies provide examples of related behaviors that they consider to be “bullying,” such as “abuse” or “harassment,” but these instances are less common. A company may sometimes elaborate on bullying in its Safety or Help Center, but, as I discuss later in the chapter, not all companies have Safety Centers.

Some corporate documents that did stipulate examples include the following:

Twitter: “Some of the factors that we may consider when evaluating abusive behavior include: if a primary purpose of the reported account is to harass or send abusive messages to others; if the reported behavior is one-sided or includes threats; if the reported account is inciting others to harass another account; and if the reported account is sending harassing messages to an account from multiple accounts” (Twitter, 2016d).

Facebook: “We don’t tolerate bullying or harassment. We allow you to speak freely on matters and people of public interest, but remove content that appears to purposefully target private individuals with the intention of degrading and shaming them” (Facebook, 2016f).

Facebook had been updating its Community Standards to provide more specific explanations of what the company considered to be “bullying and harassment,” The document explains further:

This content includes, but it is not limited to: pages that identify or shame private individuals, images altered to degrade private individuals, photos or videos of physical bullying posted to shame the victim, sharing personal information to blackmail or harass people, and repeatedly targeting other people with unwanted friend requests or messages (Facebook, 2016a).

YouTube’s specified examples of “harassment and cyberbullying” include:

Abusive videos, comments, messages . . . making hurtful and negative comments/videos about another person . . . deliberately posting content in order to humiliate someone . . . revealing someone’s personal information . . . maliciously recording someone without their consent . . . incitement to harass other users or creators . . . [and] unwanted sexualization, which encompasses sexual harassment or sexual bullying in any form. (YouTube, 2016)

Ask.fm specified that content containing “rude words or . . . intended to embarrass anyone” was not allowed, but this was not clustered under
“bullying” in the company’s policy, and “mean” content, or content intended to “harass, scare or upset,” was not allowed either (Ask.fm, 2016a).

These four companies had the most specific examples to be found as I conducted my research, but they still fell short of providing guidelines as to how the companies put them into practice. What would count as “abusive”? Or, how are “purposeful targeting” and “deliberate humiliation” assessed in practice? What counts as “hurtful” or “shaming”?

Company documents did not explain whether bullying captured the concept of power imbalance among children, whether the action had to be repeated, what would constitute repetition, or how the company went about determining whether it had been repeated. “Bullying” and “harassment” were often used interchangeably. But, as I discussed earlier in this book, these nonetheless distinct terms may also carry different legal consequences in various geographic locations.

Twitter preferred to use a broader term, “abuse,” rather than “bullying,” a decision that the company explained as part of its effort to protect freedom of speech by not becoming involved with the content on the platform. Laying out very specific provisions for “bullying” would demand content mediation (such as removing tweets in bullying cases), an action that may infringe on freedom of speech and that the company preferred not to undertake: “Regarding the accounts whose sole purpose is to be abusive, we wouldn’t necessarily want to take down tweets, we just want to remove the entire account,” explained Patricia Cartes, the head of Global Trust and Safety Outreach, Public Policy at Twitter, in our personal communication. The company has other strategies, which I detail later in the chapter.

It was typically the representatives of more established companies who provided an explanation for the language found in their policies.² They specified that the decision not to include definitions of the word “bullying” reflected the similar difficulties of researchers: if the academic community could not arrive at a single, agreed-upon definition of what constituted bullying or digital bullying, it would not be helpful for companies to be bound to one or guided by one.

None of the interviewed companies reported to publicly disclose the guidelines that their respective (or outsourced) moderators used to determine whether a case constituted bullying (and hence whether to act on it or not, i.e. take such content down or leave it on the platform due to possible violation of the company’s Terms).

As a rationale for not disclosing these guidelines, the representatives of some (especially more established) companies explained that determining whether a case constituted bullying was done case by case, and the decision
was almost entirely context-dependent, which is why it was difficult to draw such generalizations. Furthermore, the guidelines for moderators were characterized as complex, extensive, and involving long training, which is why they could not be easily regurgitated into simple explanations for the public. I wondered if companies would consider a post as “bullying” even if it were one post only without evidence of repeated bullying communication. One representative of an established company explained that they would take into account repetition and power imbalance. But if a single post was reported, they still might consider it to be bullying and take it down; without knowing if such activity is part of offline bullying, they would prefer to err on the side of caution.

For YouTube, bullying was used interchangeably with “harassment,” and a single post could qualify:

We can take it down even if it’s just one comment, or one video, for instance; if it crosses that line that we consider as harassment—and this is a case-by-case situation—we don’t have a clear “that word is harassment and that word isn’t harassment” [policy]. It doesn’t work that way. It’s always case by case – who is attacking who and in which manner. (YouTube representative, personal communication with the author, November 12, 2014)

In a 2014 interview, an Ask.fm representative said they believed one mean comment would not be enough for the company’s moderators to determine that a case constituted bullying; it would have to be repeated commenting. But two years later, in my communication with Justin Patchin, a well-known cyberbullying expert who acted as an e-safety adviser for the company, I learned that in practice it was nonetheless context-dependent, and that one comment could be enough (e.g., if it was accessible to others or if someone had created a public profile about someone else and made a single post). He observed that companies in general tend to be interested not in definitions but in behaviors. Because moderation was outsourced to one or more companies, he said, specific guidelines (including those for filtering) might have been left at their discretion and there may not have been a formal, company definition of bullying. As I stated in the beginning of this chapter, very few companies provide examples of specific behaviors that they consider as “bullying.”

Furthermore, under such circumstances where context was the determining factor, it became increasingly important to use human moderators rather than an automatic detection system (despite its levels of sophistication). Several companies that preferred not go on record provided another rationale for not disclosing their moderation guidelines: they would not
want to make it easy for users who wanted to abuse the policy to circumvent the rules.

Through my interviews I learned a great deal about what companies consider to be bullying, much of it not detailed on the companies’ websites. This was especially the case with smaller start-ups.

The Evolution of Self-Organizational Efforts

The more established a company was, or if it found itself at the center of a high-profile incident, the more likely it was to provide a greater degree of elaboration of its policy and enforcement tools.

In 2014, among the documents of relevance to bullying (as well as harassment or abuse), some of the surveyed companies merely had Terms of Service. Nonetheless, by 2016, they had gone on to develop more documents in the form of Principles, Community Standards/Guidelines/Rules, or Safety Centers. These further elaborated the companies’ anti-bullying policies and enforcement mechanisms, or referred users to NGOs, or provided more information about bullying.

Safety Centers sometimes exhibited videos and educational texts about bullying that the companies developed with the assistance of NGOs. The companies who had such features characterized them as an important part of their self-organizational effort and as an example of an evolving company policy.

Facebook, YouTube, and Twitter seemed to lead the way: Facebook first introduced “Community Standards,” YouTube had “Community Guidelines,” and Twitter had “Twitter Rules.” Other companies followed suit, often with remarkably similar wording. Facebook had also developed a “Bullying Prevention Hub,” a section of the company’s website containing information about how to prevent bullying.

From the standpoint of self-regulation, it is important to observe that the development of these online documents, Safety Centers, and enforcement mechanisms was not a self-regulatory requirement for all the companies surveyed here, as a number of them had never taken part in any formal self-regulatory initiative related to bullying. Some of the established companies may have submitted these texts and enforcement mechanisms as best-practice evidence within self-regulatory initiatives in Europe, such as ones I discussed in chapter 5 (Safer Social Networking Principles, CEO Coalition, and ICT Coalition), and in the US (Internet Safety Task Force). Consider, for instance, that Principle 6 of the ICT Coalition specifically advised the companies to provide educational provisions and “links to
other sources of relevant, independent and authoritative advice for parents and carers [sic], teachers and for children” (ICT Coalition, n.d., p. 4), which Safety Centers illustrate. But other companies that did not participate in such initiatives could develop and adopt these provisions via informal industry-wide collaboration or simply by observing the model of the more established companies.

The Spillover of Expertise

Representatives of these younger companies attended e-safety conferences, such as the Family Online Safety Institute’s (FOSI) annual conference, which brought together industry and NGO representatives as well as educators, academics, government representatives, and other interested parties to discuss relevant issues in the e-safety field. These younger companies could thus create contacts with the older companies and even hire people who had previously worked for them in a process that was sometimes described as “the spillover of expertise.”

Establishing contacts with e-safety NGOs who may later advise them can also happen at these conference venues, and exemplifies how self-organization evolves. Companies could hire people who had previously worked for state attorneys general offices. Some very new platforms that offered similar types of affordances (e.g., anonymity) reported that some of their competitors were not so willing to share best practices among themselves.

A possible reason why some older companies, such as Voxer, or other companies such as WhatsApp, did not have Safety Centers at this time was that they might not have had the need to develop them, and might not have been publicly perceived as having the properties of social media. Given that their platforms were primarily enablers of private chats between two or several people, they might not have seen their platforms as conducive to bullying either. (See table 6.1.)

Technological Affordances and Varieties of Bullying

Varieties of technological affordances among these platforms can account for quite different types of bullying. Many informants observed that bullying on Facebook and Twitter tends to be subtler than on anonymous platforms, such as Ask.fm. On Facebook, children can communicate in less straightforward ways than by writing mean comments or swear words on each other’s profiles; they can tag the target in a post that contains an ironic photo, for instance. Or, they can easily exclude someone by creating
### Table 6.1
Presence of Community Guidelines/Principles/Rules/Standards and Safety Centers at the Time of Research

<table>
<thead>
<tr>
<th>Company</th>
<th>Founded</th>
<th>Community Guidelines/Principles/Rules</th>
<th>Safety Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>2004</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>YouTube</td>
<td>2005</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Twitter</td>
<td>2006</td>
<td>Yes</td>
<td>Yes (within Help Center)</td>
</tr>
<tr>
<td>Tumblr</td>
<td>2007</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Voxer (primarily messenger)</td>
<td>2007</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>WhatsApp (primarily messenger)</td>
<td>2009</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Ask.fm</td>
<td>2010</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Instagram</td>
<td>2010</td>
<td>Yes</td>
<td>Yes (as part of Help Center)</td>
</tr>
<tr>
<td>Google+</td>
<td>2011</td>
<td>Yes in 2014 but not apparent in 2016</td>
<td>Yes</td>
</tr>
<tr>
<td>Snapchat</td>
<td>2011</td>
<td>Yes</td>
<td>Yes (significantly smaller scale in 2014 than in 2016)</td>
</tr>
<tr>
<td>Whisper</td>
<td>2012</td>
<td>Yes (in 2014, there were Community Guidelines but not as a separate document from TOS). In 2016, there was a separate document (Whisper, 2016b)</td>
<td>No</td>
</tr>
<tr>
<td>Yik Yak</td>
<td>2013</td>
<td>No in 2014 but Yes in 2016. They were just called “Guidelines” rather than “Community Guidelines” (Yik Yak, 2016b)</td>
<td>No in 2014 but Yes in 2016</td>
</tr>
<tr>
<td>Secret</td>
<td>2014</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
closed groups or by ignoring that person in conversations. In all instances it might be especially difficult for a platform to discern that bullying is in fact taking place, and thus hard to take action to remove the content or block the offending user. Similar subtleties can be seen on Twitter in a practice called “subtweeting,” which refers to tweeting about someone without using their handle and mentioning them, often in a mocking or derisive way.

On Ask.fm, however, swear words and openly mean questions were more common. As I discussed in chapter 2, anonymity can disinhibit users. It can be easier to bully someone online than in person, and easier still on an online platform that allows anonymity or the use of a pseudonym. This is why the affordances of anonymous apps such as Whisper, Secret, or Yik Yak are seen as conducive to what some informants characterized as more “blatant” bullying.

**Delegating the Reporting to the Community**

One of the functional purposes of these corporate documents was to convey the idea that e-safety is an effort of the entire community of users on a given platform. This ability of the platform to get the community to “regulate/ moderate/police itself” was understood by some informants as an “advanced or evolved” self-regulatory mechanism and cyberbullying policy. Such self-directed language also tries to convey the idea that the company does not want to interfere in free speech on the platform unless it “decides” or “feels” that it is necessary (e.g., regarding a violation of Community Guidelines), and that, when it comes to bullying, it is primarily the responsibility of the community to regulate itself rather than have the company regulate the community. Here are examples from YouTube and Facebook:

Respect the YouTube Community:

*We’re not asking for the kind of respect reserved for nuns, the elderly, and brain surgeons. We mean don’t abuse the site. Every cool, new community feature on YouTube involves a certain level of trust. We trust you to be responsible, and millions of users respect that trust. Please be one of them.* (YouTube, n.d.a).

Facebook gives people around the world the power to publish their own stories, see the world through the eyes of many other people, and connect and share wherever they go. The conversation that happens on Facebook and opinions expressed here mirror the diversity of people using Facebook. To balance the needs and interests of a global population, *Facebook protects expression that meets the community standards* outlined on this page (Facebook, 2016a, emphasis added.).
Such messages can convey that bullying is not a norm but an outlier on the platform, and that the community comprises participants who do not bully. Belonging to the community implies self-moderating by refraining from bullying:

Have Fun: The language below is meant to support this rule. It allows us to continue providing and improving our Services, and it helps ensure that a few mean users don’t ruin the fun for everyone else. Your part in that is simple. Just use common sense—keep sending awesome Snaps to your friends and please don’t send Snaps that they don’t want to receive. (version of a guideline formerly on Snapchat; emphasis added).

**Newer Companies: Liability, Community, and Freedom of Speech?**

When companies did not have Community Guidelines and Safety Centers—whether it was because they were in an early stage of e-safety development or just didn’t see the need to develop such policies—they were more likely to provide anti-harassment provisions without making specific references to “bullying.”

These provisions were typically stipulated in TOS and housed under the sections of company websites labeled “Legal.” Indeed, their purpose seemed geared to outlining the legal responsibility of the company and protecting it from possible liability in case harassment took place—quite a different task than establishing a sense of community. In such cases companies may have even openly emphasized their discretionary right to take action on the content they considered to be in violation of TOS.

Apps allowing anonymity faced particular pressure from media reports about the alleged severity of cyberbullying on their platforms. That is why, especially if they were newly established, they addressed the negative media attention by emphasizing that they did not hesitate to take down questionable content.

Guaranteeing freedom of speech, then, was not a rationale they used to avoid taking action in certain circumstances while gearing the company’s efforts toward creating a community that can manage itself. The guidelines that these companies followed when they did take down content were not publicized either. Takedowns were resolved on a case-by-case basis, which again points to the question of the power that private companies have over the digital public sphere. But the longer the company maintained its presence on the market, and especially if its name appeared in the media in relation to bullying, the more likely it was to abandon this discretionary discourse.
Community Autonomy and Transparency

Companies that rely on the language of community responsibility appear to convey the idea that users have a voice in delineating what is allowed on their platforms. While such wording may imply user empowerment, it also appears to downplay the extent to which the companies really are the final arbiters on what takes place on their platforms.

Facebook, in particular, the oldest company and the one in the sample with the largest number of users, presented itself as a platform with participatory governance. It had a Site Governance Page (Facebook Site Governance, n.d.) where users could leave feedback to any announced changes to the company’s TOS, which Facebook termed “Statement of Rights and Responsibilities” (Facebook, 2015f).

In this regard, Facebook behaved like a government entity that created a public debate prior to introducing a new law. Users were portrayed as having a significant amount of leverage over values and norms the company embraced—as exemplified in the debate on nudity. While Facebook had initially prohibited breastfeeding images, it changed its policy in 2014 to allow photos of mothers breastfeeding on the platform. The company posted then its decision with the following explanation:

We agree that breastfeeding is natural and beautiful, and we’re glad to know that it’s important for mothers to share their experiences with others on Facebook. The vast majority of these photos are compliant with our policies (Facebook, 2015b).

This suggests (perhaps much like the Vietnam War photo explanation) that the company imbues its own values as to what amounts to “natural” or “beautiful.” According to a Tech Times article, Facebook’s decision arrived only after the pressure from feminists and breastfeeding advocates, clustered primarily around #FreeTheNipple movement (Arce, 2014; Esco, Richards, & Azuelos, 2015). Whether the company acceded to the pressure from advocacy groups, or the decision had little to do with it, is not at all clear and transparent. What matters when explaining the logic behind cyberbullying policies, however, is the idea that such discourse promotes the notion of user autonomy, participation, and shared governance.

According to “Facebook Principles,” a document that the TOS and Community Guidelines are based on, transparency is the cornerstone of the company’s ethos:

We are building Facebook to make the world more open and transparent, which we believe will create greater understanding and connection. Facebook promotes openness and transparency by giving individuals greater power to share and connect and
certain principles guide Facebook in pursuing these goals. Achieving these principles should be constrained only by limitations of law, technology, and evolving social norms (Facebook, 2016b).

This statement suggests that Facebook understands itself or wishes to be understood not merely as a corporate entity, or a brand whose purpose is to provide a satisfying product to its users, but as a platform with a mission to enhance human connection as embodied in “sharing culture” and a focus on the inherent value of “sharing,” as José van Dijck aptly explains:

Facebook’s business model is most certainly a contentious balancing act between stimulating users’ activity and exploiting it; its success ultimately depends on customers’ willingness to contribute data and allow maximum data mining. . . . Values of connectedness and community are equated with connective values, smoothly aligning business models with user interests. (van Dijck, 2013, p. 64)

Based on what social media companies publicly reveal about their moderation practices, it is difficult to tell how content moderation (e.g., taking posts down because of cyberbullying) may reflect on their business models and revenue. Might it affect user satisfaction or even the process of data collection? And how does investment in safety relate to company revenue? These aspects of the discussion are carefully elided in company discourse that focuses on user benefits, autonomy, and community empowerment. In the following chapters I return to some of the relevant points.

Enforcement: From Formal Document to Practical Operation

When it comes to enforcing policy, much like with defining terms such as harassment and cyberbullying, what is formally written in the company documents may only be a fraction of what actually takes place at the operational level. Most companies provide users with the ability to report or flag abusive content, or at least to contact customer service. The companies will only take action on reported content if the moderation process establishes that the content constitutes bullying, harassment, or abuse, thereby violating TOS / Community Guidelines (see also Crawford & Gillespie, 2016). Most companies provide their users with tools to block abusive users or posts, which can, under some circumstances, lead to an account suspension.

Some companies at one point provided a form of filtering for the words and phrases that were not allowed on the platform (e.g., Ask.fm, Secret, Whisper, Yik Yak), and some reported doing so in different languages in markets where they had significant numbers of users (although what “significant” meant in exact numbers was not necessarily specified).4
The companies whose discursive cornerstone was freedom of speech were reluctant to employ filtering. Companies did not disclose whether they had moderators for every language represented in their user base, how many such moderators they employed (or outsourced moderation to), or what criteria they used to decide that they needed to introduce a moderator for a specific language.

The more experienced social media companies tended to provide more elaborate explanations of their tools for enforcing the policy. Some of them also invested heavily into research on optimizing enforcement tools, which seemed to be especially the case with Facebook and Twitter, although no company revealed any details about the scale of financial investment. Despite more extensive explanation, the information necessary for determining how the effectiveness of enforcement mechanisms was measured nonetheless remained largely undisclosed.

Lessons Learned from Facebook

Given that the e-safety experts I interviewed frequently characterized Facebook as the company leading the self-regulatory effort when it came to cyberbullying, I analyze Facebook’s policies in more detail here, and use the company as an illustration how industry-wide policies in this area have evolved.

Facebook’s relationship with regulators in the US and the EU seemed to be going rather smoothly while I was writing this book. But, as I briefly outlined in chapter 5, and as the following case of “panic button” illustrates, this may not have always been the case. Early in their development, companies have the potential to trigger fears in regulators and the public, especially if their user base outgrows the e-safety awareness of their leadership and consequently their companies’ e-safety capacities.

Once they adopt a visible e-safety strategy (e.g., Safety Centers, NGO partnerships, or Community Guidelines), however, and start developing the language of advanced policies, they can successfully attenuate these fears and build a collaborative relationship with relevant stakeholders.

In May 2017, as this book was being completed, The Guardian leaked some of Facebook’s previously undisclosed operational policies, including instructions the company gave to its moderators on how to handle some types of abuse. The content of these files caused significant criticism for the company and even calls on behalf of some policy makers in the UK for the company to become more transparent in how it handles its content moderation (Grierson, 2017). Citing these documents, The Guardian
Untangling the Companies’ Motives and Actions

reported the company even allowed “the ‘sharing of footage of physical bullying’ of children under seven” as long as such sharing was done without a caption (Hopkins, 2017, para 7; see also Hopkins & Wong, 2017). These documents revealed that Facebook’s operational policy for moderators defined bullying broadly as “an attack on private persons with the intent to upset or silence them,” and one is a private person if they are not a public figure (Hopkins, 2017). It will be most interesting to see how the company will respond to the public reactions in terms of policy changes and whether this incident will affect the company’s relationship with regulators internationally.

“The Panic Button”

The introduction of the reporting button (“flag”) on this platform did not take place without a public debate that illustrates how e-safety design can be important for any platform’s business model, and how little is publicly known about this dynamic. The ways in which a platform’s business interests may be related to the choice and design of the preferred e-safety tools rarely finds its way into public discussions around e-safety.

The Child Exploitation and Online Protection Center (CEOP), a command within the UK’s National Crime Agency, is a body that works primarily on protecting children from sex offenders and child pornography, but also engages in other e-safety issues. In 2010, CEOP wanted to ensure that major social networks introduced a standardized “safety button,” dubbed as “the panic button,” which would be located on the profile page of every user who was under 19 (Facebook refuses to add safety buttons, 2010). By clicking on that button, the report would go directly to CEOP. Furthermore, the button allowed 10 different reporting options. For instance, if a child did not want to report to the police they could receive help via phone from counselors working at an NGO charity called Childline.

As the UK Telegraph reported, at the time Facebook was confronted with “mounting pressure” from parents and the UK government to improve its safety strategy (“Facebook refuses to add safety buttons,” 2010). Facebook agreed on the point that allowing users to report cases to CEOP directly should be instituted on the platform. Yet they did not agree on the design of the button and the wording that should be used to report offensive content. According to the Telegraph, the company representatives said that in their experience users disliked big buttons and that such graphics “intimidate and confuse people,” thus lowering the likelihood of reporting. Instead, Facebook preferred to include its own “report abuse” link and allow users the option to report to CEOP directly as well as to company moderators.
Bebo, a social networking site that was popular with young people at the time but whose user base had been rapidly declining, had adopted the CEOP report button (Barnett, 2009). CEOP criticized Facebook for refusing to do the same. How increasingly stringent safety measures may have influenced Bebo’s popularity is a debate I will revisit in chapter 8, adding Ask.fm and Formspring to the list of examples. Microsoft’s MSN chat and a number of websites also introduced the button.

At the time, the *Telegraph* reported, CEOP was receiving as many as 10,000 clicks per month, which resulted in 5,000 criminal investigations (the article did not specify how many were in reference to bullying vs. other e-safety issues). It also reported that the number of complaints to the police in reference to Facebook had “almost quadrupled” at that time in comparison to the previous year, which was why CEOP deemed it was especially important to have Facebook adopt the procedure (“Facebook refuses to add safety buttons,” 2010).

Facebook delivered its rationale to the public for not adopting the button’s design, saying that its own testing showed such action would decrease the likelihood of reporting; the *Times* newspaper, on the other hand, reported that according to the CEOP director, Facebook’s reason for not adopting the button may have been the possible impact of such an action on the company’s advertising revenue (Monaghan, 2011, cf. Fresco, 2009). Whether there would be any tangible impact on advertising is not a topic that frequently finds its way into public discussions, and the company to my knowledge does not disclose such information.

**Other Options for Addressing Abuse**

In addition to reporting abusive posts to Facebook, if some content bothered users but did not violate Community Standards, Facebook recommended that users block or unfriend people, hide them from their Newsfeed (updates users received from their friends), or send them a message and try to resolve issues on their own. Most companies provided a variation of at least some of these options (typically significantly more modest, unless the company was established), which had been adapted to the technological affordances of their particular platforms.

A detailed diagram explaining how the reporting worked was available on Facebook (Facebook Reporting Guide, n.d.), which stated Facebook had “hundreds of team members” working in 24 languages located in Menlo Park, and Austin in the US, and in Dublin, Ireland, and Hyderabad, India (Facebook Reporting Guide, n.d.). Moderators were then divided across four teams: the Safety Team, the Hate and Harassment Team, the Abusive
Content Team, and the Access Team. The majority of reports were handled within 72 hours, according to the diagram.

**Difficulties Determining the Details**

Similarly to other companies surveyed here, at the time of my research Facebook did not disclose information about the exact number of staff and moderators working on each team across its four offices; such information was classified, as was all documentation regarding the number of reports that were processed as “bullying” by the company’s moderators in a given period of time and evidence that all or the majority of reports were indeed handled in 72 hours.

Companies with a longer history typically emphasize in their discourse that their users’ safety is paramount to them, assuring that a robust e-safety effort is in place:

At Facebook, nothing is more important than the safety and security of the people who use our service. With a community of over 901 million people, Facebook maintains a robust reporting infrastructure made up of dedicated teams all over the world and innovative technology systems. (Facebook Reporting Guide, n.d.)

Nevertheless, specific criteria of effectiveness, evidence of effectiveness, or the details behind how the effectiveness of this effort is measured remain difficult to determine in the case of established and new companies alike.

**Support Dashboard**

Some of the e-safety experts I interviewed characterized Facebook’s “Support Dashboard” as an important move made by the company to address complaints that it was taking too long to respond to reports about inappropriate content (e.g., Donoso, 2011; Bazelon, 2013b), and to introduce more transparency about its handling of these reported cases.

Having reported the content in question, the user would receive a link to what in 2014 was called Support Dashboard and in 2016 the Support Inbox. The dashboard/inbox explained whether the reported post was being reviewed, and, once it had been, notified the user about the moderators’ decision (Facebook, 2016e). In 2014, the time span in which Facebook made its decisions was not specified; in 2016, it stipulated in a section titled “What to expect now” that the reported content would be assessed against Community Standards the same day.

Some e-safety experts, who believed the dashboard to be an important move on Facebook’s part toward fostering transparency, said that from their purely anecdotal experience, its introduction appeared to coincide with
fewer complaints about Facebook’s flagging tools. Statistics as to whether or how the introduction of this tool has improved the effectiveness of the company’s reporting system, however—either provided by the company or from any other source of evaluation or publicly available document—were not readily available at the conclusion of my research. Neither was such information provided by NGOs that I had the opportunity to interview and that were members of Facebook’s Advisory Board at the time of my research.  

**Moderation 2.0: Toward Advanced Policies**

As social media platforms began to introduce flagging or reporting options for users, part of the responsibility for moderation was transferred to users—and hence the discourse on safety as a joint effort evolved, as summed up in this quote from an NGO representative:

> Because in this world of web version 2.0 it is in fact the users who are moderators—their reports first alert service providers of content that breaks the rules and when people start losing confidence in that (reporting tools), then there is a problem. (Anonymous, personal communication with the author, June 30, 2014)

The difficulties of moderating vast amounts of reported content while avoiding the loss of user confidence—for instance, in cases where the effectiveness of the moderation was not to their liking—led companies to move toward more “advanced policies” that attempted to empower users by further deferring moderation to them. Such a move might also allow some companies to hire or subcontract fewer moderators. Consider an explanation given by one company representative:

> But we have a lot of users and the ratio we have with the number of moderators is very good because we make sure the community polices itself. So as these communities get bigger you’ll see them “moderating” each other quicker. (Anonymous, personal communication with the author, November 12, 2014)

**Social Reporting and Community Empowerment**

Social reporting is an example of a policy that relies on the logic of community moderation. It seeks to empower users by providing them with tools designed to help them resolve conflicts among themselves, thereby delegating part of the responsibility for conflict resolution onto users. After extensive research, Facebook introduced the social reporting tool in an effort to better address the needs of its users; the company had noticed that moderators kept receiving a large number of reports they could not act upon because they could not establish that a case constituted bullying and had thus violated company policy.
For instance, the company may not necessarily be able to act if it receives a report of a photo of two girls smiling at each other, with no mean comments underneath. Even after looking into the context behind the post, the information provided might not be sufficient for moderators to determine that a case constitutes bullying. By taking such a photo down, the perceived peril for the company is that it would be curbing users’ freedom of expression.

Facebook introduced “social reporting” in 2011 (Facebook, 2016d). It allows users to send a message to the person whose content they think is bullying or abusive in an effort to try to resolve the issue without reporting it to the company’s moderators. In other words, 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 abuse or bullying, but that may not qualify as such according to corporate policy.

Social reporting also allows users to reach out to a third party (e.g., parent/caregiver, teacher, friend) in order to seek help when one feels bullied. As part of social reporting, Facebook also provided “premade” (prewritten) messages (see examples in the next section) that a user could send when reaching out to another user who may have bullied them, or to a third party.

Social reporting was perceived as a more advanced enforcement tool than the simple content takedown because it could help address conflicts that could persist or originate offline. Facebook provided pages called “Details on Social Reporting” and “What is Social Reporting,” which outlined the entire process (Facebook Safety, 2011; Facebook, 2016d) and provided step-by-step explanations on how to report (Facebook Help Center, 2016b).

Infrastructure for Community Autonomy

Facebook didn’t provide a separate “social reporting” button. Rather, when users wanted to report (e.g. by clicking on “report” option), they were led through a series of prompts (“flows”) which asked them to specify why they were reporting. At the time of this book’s writing, not all of these flows would lead to the option of reporting the post to the company’s moderators (i.e. “regular reporting”). Depending on which options the user decided to choose in these flows to explain to the company why they were reporting, at the end of the flow they would be given one or more of the following options: report to the company (i.e., “regular reporting,” by which the content is supposed to go to the company’s moderation system for inspection); engage in social reporting (i.e., reach out to the person who posted the
abusive content, in an attempt to resolve the issue, or to a trusted person, in a plea for help or advice); or they would be given both options: to report to the company and engage in social reporting (then they could choose which one they prefer). At the end of the flow, users were also typically advised they could block, unfollow, or unfriend the person whose content they dislike.

For example, when I clicked on a post to report it, a dialog box opened with the following prompt (referred to by the company as “a flow”): “Help us understand what’s happening.” I could then choose: (a) “It’s annoying or not interesting,” (b) “I think it shouldn’t be on Facebook,” (c) “It’s about me and I don’t like it” (provided as an option in 2014 and 2015 but not in 2016), and (d) “It’s spam.”

Consider the next level of choices: If I clicked on “It [the content] shouldn’t be on Facebook,” the following dialog box opened: “What’s wrong with this post?” and I could then choose: (a) “It’s rude, vulgar or uses bad language,” (b) “It’s hurtful, threatening or suicidal,” (c) “It’s private information like my phone number or address,” (d) “It goes against my beliefs,” and (e) “Something else.”

If I chose “(a)” I would then be able to send a message to the person who posted the content, unfollow them, or submit the post to Facebook for review. If I chose “(b)” I would see a different dialog box, which said, “Help us understand the problem: how is it harmful?” I could then choose one of the following options: “It’s mean,” or “It offends gender, race, sexual orientation or ability,” or “It’s threatening or violent,” or “I think they might hurt themselves.” If I chose “It is mean,” (the option perhaps very likely to correspond to bullying content), I would not be offered the choice to report the content to Facebook. I could only message the person, unfriend her, or reach out to a friend to resolve the issue. Hence, not all options in dialog boxes that the user might choose would necessarily result in the option to report the content to Facebook. This is why, perhaps, some of the content that could constitute bullying might not have the opportunity to be reported to the company.

The exact wording of these flows was said to vary based on the user’s age (e.g., teens see different texts in these flows than adults do) and it was said to be tweaked regularly (sometimes every few weeks) based on extensive research, in an attempt to optimize the flows in a way that would match the language users employ to describe bullying behavior. In August 2016, a flow that a 14-year-old user might see, which is very different from the flow I described above, looked like the illustration in figure 6.1)
Untangling the Companies’ Motives and Actions

As seen in figure 6.1., the word “bullying” may not come up in the flows. This reflects the finding that teens tend not to identify with either the word “bullying” or with “bully–victim” roles (see chapter 2).

Interplay of Effectiveness and Efficiency
Facebook’s organization of the reporting flows is designed to minimize the number of reports that users would file regarding content that does not violate the company’s Community Guidelines, and about which Facebook could potentially do nothing about. It is an attempt to increase reporting efficiency.

The increase in efficiency is considered to be effective because users tend to understand the context of the conflict and are therefore perceived as best positioned to solve the cases for which the company cannot establish to have violated the company policy (Milosevic, 2015a). Community autonomy, then, implies that users participate in keeping the platform safe. Safety is portrayed as a coordinated effort between the users and the platform. The process of creating the community implies taking responsibility for one’s actions and helping others; the entire e-safety onus does not fall entirely on the company:

We believe safety requires a coordinated effort from everyone—whether by reporting inappropriate behavior or making sure your account and passwords are secure. Let’s work together to create an environment where we can all share comfortably. (Facebook, 2015e, emphasis added)

“Sharing comfortably” ensures “frictionless sharing” and seamless functioning of the company’s business model—not just for Facebook, but also for social media companies in general (van Dijck, 2013)—and can be seen as another connotation behind such wording.

Scale of Effort
The wording used in social reporting flows is by no means arbitrary. Facebook partnered with psychologists and neuroscientists at the Yale Center for Emotional Intelligence, as well as with the Greater Good Science Center at the University of California, Berkeley, to develop the language for social reporting as well as for Facebook’s cyberbullying prevention initiative, the “Bullying Prevention Hub” (Yale Center for Emotional Intelligence, 2013). In the view of some e-safety experts, such extensive effort was affordable primarily to large companies such as Facebook.
Figure 6.1
Reporting Flow
c) **Help Us Understand What's Happening**

This is hateful towards

- A race or ethnicity
- A religious group
- A gender or orientation
- People with disability or disease
- An individual

[Back] [Continue]

---

d) **What You Can Do**

Here are some things you can do to handle this.

- **Submit to Facebook for Review**
  Report this post if it goes against our [Community Standards](#).

- **Block Tijana**
  You won't be able to see or contact each other.

- **Unfollow Tijana**
  You won't see her posts in News Feed, but you'll remain friends.

- **Unfriend Tijana**
  You'll no longer be friends on Facebook.

[Back] [Cancel]
Compassion Research Day
In 2013 a team headed by an engineering director at Facebook organized an annual event called Compassion Research Day as a way to present, on a regular basis, the results of Facebook’s research and collaboration with Yale, Berkeley and other academic institutions; Facebook also organized an in-house “compassion team” in order to help sociologists, psychologists, and neuroscientists, among other researchers, “conduct experiments” and “implement (their) findings on the network,” all “funded in part via stipends from Facebook” (Tsui, 2014).

There seems to be an understanding among companies that users will not return to their sites if the platforms are infested with bullying or harassment. In line with this thinking, Bonnie Tsui summed up her report about Facebook’s efforts to facilitate the ongoing Compassion Research Days: “If people enjoy their experience more, Facebook will do better as a business” (Tsui, 2014). But as the controversy over the “panic button” perhaps illustrates, a legitimate question (for any company) to ask might be: How does a company’s business model benefit from its preference for using certain moderation tools rather than those methods characterized as ineffective, or as threatening to freedom of speech?

A researcher who consulted with teams at Facebook on social reporting, explained that one of the difficulties in providing statistics on effectiveness was that this research did not have a “typical before and after.”

We say, “hey let’s tweak this little thing,” [referring to wording in social reporting flows]. We don’t always tweak only that one thing and have the opportunity to evaluate just the change that that one little thing was able to impart upon people’s completion rates. (Anonymous, personal communication with the author, December 3, 2014)

For example, a question in one of the platform’s flows would ask the person who reported disliking the photo, “How does this photo make you feel?” The prefilled messages would read: “embarrassed,” “sad,” and so forth, with each option naming a particular emotion. The Facebook team found that a subtle change in the wording from “embarrassed” to “it’s embarrassing,” resulted in a lot more people completing the flow.

The data collected over 30 days of research revealed that of all the teens entering this flow, 76% were reporting photos and 24% were reporting posts (textual content). (The sample was collected in the US and consisted of “all 13–16 year olds who entered [the] resolution tool within a thirty-day period” see Compassion Research Day, 2013.) Within this sample, 15% selected the option “it’s bullying” while 66% selected the option “it’s
annoying” (suggesting that perhaps teens might not find the word “bullying” to be helpful or to accurately reflect what they were trying to report).\textsuperscript{9}

Out of the 25\% of teens who used the flows, 90\% messaged the person who posted the content they had a problem with, and 10\% messaged a trusted adult or a friend. As many as 75\% used prewritten messages as opposed to tailoring them themselves—which suggests that these were useful to teens.

Once the content creators were contacted, 75\% of them were reported to respond back. Statistics relating to the type of response provided—whether it was favorable to the person who felt bullied or not—were not readily available. However, 37\% of them deleted the problematic content upon request.

**Community Autonomy Efforts by Twitter, YouTube, Google+, and Tumblr**

Twitter, another more established company, has also developed a reporting system that was explained in detail on the company’s website (Twitter, 2016a, 2016c). The company created a Trust and Safety Council, a partnership with advocacy groups, e-safety experts, and researchers working to prevent abuse (Twitter, 2016b), similar to Facebook’s Safety Advisory Board (see chapter 7).

Using the word “targeted abuse” in its policy rather than “bullying” allowed the company to intervene in those cases that met the conditions for targeted abuse rather than develop guidelines for specifically cyberbullying tweets. The company’s focus on freedom of speech permeates its abuse policy discourse:

At Twitter we look at abuse from a holistic perspective. Other tech companies follow a different approach and break down their policies very granularly. You may see hate speech, cyberbullying, etc. called out on their policies. When we outlined our Twitter Rules, which are our content boundaries, we emphasize behavior; we look at the intent of accounts. It’s important to bear in mind that [they were] built on the principle of freedom of speech. We want the tweets to flow. (Patricia Cartes, head of Global Trust and Safety Outreach, Public Policy at Twitter, personal communication with the author, June 17, 2014, emphasis added)

As of the summer of 2016, the closest options to bullying in the reporting queues (comparable to Facebook’s flows) were “abusive or harmful” content, which then prompted users to choose among a number of sub-options, such as “targeted harassment” and “disrespectful or offensive content.” Just like Facebook, Twitter provided users who wanted to report a tweet with means to explain why they were reporting it.
This approach, whereby Twitter prefers to warn users rather than to take down specific cyberbullying tweets, is also conceptualized as effective, placing trust in the community’s ability to autonomously regulate itself when provided with the right tools by the platform. But if the case is not one of a temporary fall-out between people, and when an account’s sole purpose is judged to be abusive, the platform prefers to shut down such an account.

When we contact the person and give them a warning or even permanently suspend them we see in the very vast majority of cases—people will remove the tweets that they deem abusive. We see that people react to warnings and educational messages: in the great majority of cases there is no bad intent. You have two friends who fall out and one starts tweeting in a way that is not constructive. That type of user reacts very well to our educational messaging and our warnings and rectifies their behavior. (Patricia Cartes, Twitter, June 17, 2014, personal communication with the author)

Similarly, Google-owned YouTube provides for reporting abusive videos, and it relies on the logic of user autonomy and the user’s ability to self-regulate:

You may not like everything you see. Some of the content here may offend you—if you find that it violates our Terms of Use, then click the button that says “Flag” under the video you’re watching to submit it to YouTube staff. If it doesn’t, then consider just clicking on something else—why waste time watching videos you don’t like? (YouTube, n.d.a)

A “how to” video explains how to flag content, emphasizing that YouTube depends on its community to make sure that the Guidelines are respected. The rationale behind this shared responsibility for preventing abuse relies on the volume of user posts: with so much content being uploaded to YouTube, the company could not possibly be expected to monitor all the content on its own.

Users are warned again that “mildly annoying” content should be ignored rather than reported. Yet, specific guidelines on what YouTube will consider as “mildly annoying” were not publicly explained.

Google+ provided for similar procedures. User Content and Conduct Policy contained links explaining how to report content. Much like on the other social networks I’ve examined here, when clicking on the post to report it, you were asked to choose one option to explain why you were reporting the post, and led to subsequent options thereafter. Finally, you could choose to ask the owner of the post to remove it; remove the user from your circles (block the user); or report the post for review to Google. While Google+ did not label the procedure of asking the other person to
remove the post as “social reporting,” the idea behind this process appears to have been the same as with Facebook’s social reporting tool.

Google+ also emphasized, similar to other companies, that while users may not like some content, or they may think it is mean or negative, Google+ would only remove it if it violated its “Content and Conduct Policy” (Google, 2016b, cf. Google, n.d.). A link to a webpage with more information on how to protect oneself from online bullying was provided (Google, 2016b). The webpage contained standard recommendations on what to do if one was involved in a bullying case, including channels for reporting, and an admonition to not retaliate.

At the time of this research, Tumblr was an older company that exhibited surprisingly little of the standard evolutionary trajectory in terms of cyberbullying policies and little of the discursive turn toward community autonomy I discuss in this chapter.10 Because of its age, one might have expected it to have had a more elaborate Safety Center (Tumblr, n.d.a); in 2014 nothing resembling a report button was apparent next to posts on the platform itself. The Guidelines only specified that one should not harass or try to circumvent the “ignore feature”; they did not seem to explain what “ignore feature” meant and what it allowed users to do, or give examples of circumventions that it had referred to (the term was later changed to “block” [Tumblr 2015]). The Guidelines also specified that users could report any violation thereof to Tumblr, but no further explanation was given, outside of a link that was supposed to specify what the reporting options were, but the link only provided an email address to which users could report abuse.

By 2016, after updates to the platform, the Community Guidelines specified that “bullying” was not allowed, but much like on other platforms, the company provided no explanation for what “bullying” was specifically considered to be (Tumblr, 2016). “Targeted abuse” and “harassment” were defined as “sending you unwanted messages or reblogging your posts in an abusive way” (Tumblr, 2016). In the reporting flow, harassment was specified in a playfully vague manner (as “being a jerk”) and as “circumventing the ignore feature in order to send someone mean or hateful messages” (Tumblr, n.d.c). When I clicked “report harm to minors” there appeared a follow-up notice that further defined such harm as “sexually suggestive or violent depictions of minors, or bullying even by another minor” (Tumblr, n.d.d, emphasis added)—which is an interesting distinction given that bullying is a term more often used to refer to conflict that happens precisely between minors. Elsewhere on the website, the company provided links to counseling services with helplines through which users could seek...
professional help (Tumblr, n.d.a.). The company never responded to inter-
view requests.

Interventions on the Anonymous Platform Ask.fm

After Tinder owner InteractiveCorp (IAC) bought Ask.fm in 2014, and into early 2015, the platform’s abuse policy specified that enforcement included blocking, reporting, and filtering. The words that were to be filtered, however, or any guidelines as to what determined whether the content would be filtered, were not apparent. Swear words typed in the search box, in an attempt to define them as filterable, yielded zero results. Suggestive terms, such as “hang yourself,” also yielded zero results.11

Responding to public pressure, the company had been forced to adopt the discourse of a more established company early on, when the new owners pledged they would cut down on bullying. In 2016, the Ask.fm Community Guidelines specified: “Respect is the foundation of our community” (Ask.fm, 2016a). The company had already been partnering with well-known experts in the field of e-safety, with NGOs, and with other organizations (Ask.fm, 2016c, 2016f).

The focus that Ask.fm places on the shared responsibility for e-safety makes the platform particularly interesting. The company uses terms broader than “community” to reinforce that a complex phenomenon such as bullying is a wide-ranging social problem:

Rather than searching in vain for a safety silver bullet, ASK.fm leadership is committed to doing the hard work of digging deep into the potential causes behind the complex issues that online platforms like ours face with respect to safety. Where does bullying start? How is it defined? Where are we successful as a society at combatting this issue in the offline world, and how can we best apply a holistic lens to addressing this issue to our digital platform? These aren’t just technical or product design questions, they are societal ones. Only by working together across communities, educators, government, law enforcement and industry can we find the most effective answers. (Ask.fm, 2016b, emphasis added throughout)

It is notable that information about how effectiveness of enforcement was defined and measured was not explained on this platform either. Although it provided detailed screenshots and instructions about reporting tools (Ask.fm, 2016d), along with links to e-safety expert organizations and individuals (e.g., Ask.fm, 2016c), it was difficult to tell how this “digging deep” took place, at least based on publicly available information.
Secret, Whisper, and Other Anonymous Apps

Other platforms that allowed anonymity, especially those early in their existence, provided even less information about policy enforcement and what the company did to handle reports of abuse. Despite media reports about the company’s moderation process, the anonymous app Whisper nonetheless disclosed very little information about how it regulated cyberbullying cases. In 2014 and 2016, the only reference to enforcement was that a violation of TOS and Community Guidelines could result in a termination of one’s account or blocking of future access to the service.

Secret, an app that closed down in 2015, similarly did not provide an explanation in its corporate documents about how it enforced its policy and how it made sure that cyberbullying did not take place. Its Community Guidelines mentioned the option to report bullying content but did not explain how the reporting system worked or how the company handled such posts. Although it was not published or explained in detail to the public, an elaborate system for handling such posts did exist, as I discovered in an interview with Secret. A legitimate question to ask, therefore, is to what extent do other young companies employ elaborate enforcement mechanisms that they are under no obligation to discuss publicly, and whether additional disclosure or full transparency should be a requirement.

Secret’s Proactive Moderation Procedures

As part of its bullying prevention initiative, Secret employed “advance screening” or “sentiment analysis” which was “automatic.” Since the platform was strictly anonymous, 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 pre-made database of first and last names and curse words. If the system detected a name or a bad word, it prompted the user with questions such as: “Are you sure you want to post this,” or “say something kind?” Or it warned: “This is not a place where you should post negative things.” Such prompts were meant to act as deterrents to posts that could violate Community Guidelines. The system also screened the post for its “severity level,” and if it was determined to be “high” or “significant,” the post was not published but instead withheld and sent to a human moderator for further review. The guidelines and standards for determining the severity level were not available to the public.

Once a post was published, a button could be used to report inappropriate content, a process called “community flagging.” If a post was reported
by numerous accounts, it was taken down automatically—a procedure that the companies focusing on freedom of speech would find invasive. According to Secret, every time a case was reported, it was reviewed by a human moderator. The moderators had a checklist to determine if a case constituted bullying. The details of the checklist were not available to the public but every case was said to be heavily dependent on context.

Typically, if a post contained a personal name, once it was reported, it would be taken down. If a user’s post was taken down, that account was temporarily suspended. Suspension meant that the user was not able to post and comment on things but could be present on the platform and could “like” other posts. If a person’s posts were reported two to three times, the person was permanently blocked from the platform.

These decisions were described to me as contextual, decided on a case-by-case basis, as there were no clear-cut rules. The blocked person could appeal the decision and file a complaint. Because each user was registered on the platform with either a phone number or a Facebook account, or both, the user could only open another account with a new phone number or a new Facebook account. According to Secret, such a situation would rarely occur, as it constituted too high of a barrier.

**Yik Yak’s Trajectory**

Similarly, Yik Yak did not publicize the details of its policy enforcement procedures at the time I spoke with company representatives in late 2014. But my interview revealed an elaborate system that was otherwise not explained publicly, and the representative asked that I not publish it in this book because it had become outdated.

In 2016, however, some of the mechanisms that existed at the time of my interview were explained by company policy documents; the platform moved, to an extent, away from anonymity, adopting the discourse of a more established company with a Safety Center and partnerships with e-safety expert organizations. Much like the other companies, Yik Yak’s definition of bullying remained undisclosed to the public except for the following command: “Do not bully or specifically target others. This includes but is not limited to, defaming, abusing, harassing, stalking and threatening others” (Yik Yak, 2016b).

An automatized tool that could also be considered a form of advanced policy—as it allows users to solve conflicts among themselves by providing them with the infrastructure adapted to particular technological affordances of the given platform—was the ability to “upvote” and “downvote,” similar to the option to “like” and idea behind “dislike”). Five downvotes on Yik Yak
“Downvoting” and “upvoting” is illustrated in figure 6.2, whereby downward arrows indicate “downvotes” and upward arrows “upvotes.”

In February 2015, *TechCrunch* ran a story alleging that Yik Yak was “systematically” downvoting “mentions of competitors” on its platform (Constine, 2015). Yik Yak responded that this happened with a lot of other sites and games and not just competitor apps because “when we see repeated posts that say ‘Go get this (app)’ or ‘go download this’ we consider it spam” (DeAmicis, 2015). Whatever the case may be (see ThirdParent, 2015), it illustrates the potential interplay between e-safety policies and companies’ business models, which is less understood in the absence of transparency.

**Digital Messengers: Affordances Informing Enforcement**

Apps that were primarily messengers or that did not perceive themselves as social media also provided little explanation as to how their bullying
policies were enforced. Furthermore, they did not necessarily have flag/report buttons the way that previously analyzed companies did.

In the case of Voxer, in 2014, I could not locate a report button that, when strategically placed next to a message for instance, would have allowed users to report abusive content. Nevertheless, users could report abusive messages to Voxer by contacting customer service. The company had two primary ways in which it addressed bullying and abuse: by allowing users to block other users and by enabling “privacy mode,” also illustrated on the company’s website (Voxer n.d.a). Any user could turn on the “privacy mode,” which meant that only those people who had that person's email could find the person on the platform. The user could sign up either with an email address or with a phone number. According to the company representative, a vast majority of users signed up with their emails. They could therefore change the email address in case of harassment and enable the privacy mode. Blocking prevented abusive users from sending any further messages to a user who had blocked them. If the abusive user went through the trouble of creating another account to continue the harassment, the victim could create an account with another email unknown to the abuser, and enable the privacy mode. The company considered such measures to be effective, given that it did not perceive itself as social media and hence not as a platform that was particularly conducive to harassment or bullying.

An option to report abusive messages to WhatsApp was not apparent at the time of this research; users could, however, block each other, and an account could be temporarily banned from the service. This could happen, among other reasons, if a user was blocked by “too many people in a short period of time” (WhatsApp, 2016a). A permanent ban could result from “being annoying to other users.” At one point, its TOS even playfully stated that “an annoying person is anyone who is (capriciously or not) determined to be annoying by authorized WhatsApp employees, agents, subagents, superagents or superheros” (WhatsApp, 2016b). WhatsApp was a rare company that between 2014 and mid-2016 seemed to make no move toward establishing the e-safety discourse discussed above or explaining e-safety mechanisms in its corporate documents. Its TOS mentioned harassment, but did not reference bullying.

Kik's policy in 2014 only seemed to provide options to block an abusive user or report someone’s message “as spam.” Any reporting provisions specific to cyberbullying were not apparent. However, in 2016, there was an option to report a user based on “being abusive” (see Kik Interactive, 2016a, 2016b).
In an August 2016 email to me, Kik reported that it was “still in the early stages of formalizing policies” and requested that I contact the company in October 2016. When I initiated the follow-up communication, Kik declined to provide an interview at that time as well. It is particularly interesting to observe how a company whose e-safety issues were the repeated subject of mainstream media coverage since 2013 would not feel more pressured to have developed a policy several years later.

Snapchat in late 2014 did not provide a detailed explanation on how its enforcement mechanisms worked, but it did allow a user to report abusive content or the user who posted it, and to submit a written explanation about the incident to the company directly. Over time, the company developed a Safety Center, partnered with e-safety NGOs that created e-safety guides for the platform and started to employ the discourse of a more established company with an emphasis on community (Snapchat, n.d.a, n.d.b; ConnectSafely, n.d.). Bullying was mentioned in its 2016 TOS: “you may not upload, post, send or store content that: bullies, harasses or intimidates” (Snapchat, 2016); however, the TOS did not specify the behaviors that defined these terms, and neither did the Community Guidelines, which under the section “harassment, bullying or spamming” said: “This one should be simple: Do not bother or make other people feel bad on purpose. If someone blocks you, it’s not okay to contact them from another account” (Snapchat, n.d.a).

There was no explanation as to how the company decided a user was bothering someone and making them feel bad, apart from circumventing the blocks. A cyberbullying guide detailing how children can stay safe on Snapchat, developed by ConnectSafely.org, a well-known NGO in the US e-safety community, said that the platform was “not the most likely ‘place’ for cyberbullying to occur because a lot of what is shared in the app isn’t public, doesn’t stick around long, and is usually shared only among friends, or—in the case of Stories — fans or people who care about what you’re sharing” (ConnectSafely, n.d., para 2). While these features may limit public humiliation, bullying can also happen among one-time friends, as discussed earlier in the book (and as some company representatives pointed out). On the “Snapchat Support” website, one could “Report a Safety Concern” (Snapchat, n.d.c) using options to report various types of content on Snapchat such as Stories, Snaps, or Chats. Similar options were provided inside the app itself.
Degrees of Content Moderation

One significant factor that cannot be ignored in the consideration of e-safety policies is that the amount of content uploaded on these platforms is vast. Perhaps YouTube is an excellent illustration, with 300 hours of uploaded video per minute (Dormehl, 2015). At one point, 20,000 questions were asked on Ask.fm platform per minute (Ask.fm Safety Guide for Schools and Educators, 2017). Although no company revealed how many bullying-related reports it received in a particular timeframe (given the volume of sharing on each platform), one can assume that moderating the services can be extremely difficult and costly, especially if humans do the moderating.

Many companies I surveyed for this book tended to emphasize that the reported content was, to a large extent, reviewed by humans and not by automated means. Typically, though, they could not guarantee that a human reviewed every piece of reported content—not even the companies with substantial resources could make that claim. Companies emphasized that because bullying cases are so context-specific, they need to be reviewed by humans. Most companies do not disclose the numbers of in-house and outsourced moderators they employ, who they outsource their moderation to, and the amount of financial resources they invest in moderation and e-safety in general.

No company provided a breakdown of its moderator base according to language expertise. When users speak a variety of languages, it can become increasingly difficult to moderate bullying cases, which require moderators to understand the subtleties of context. For example, with a global user base, more than 80% of Facebook’s daily active users were from outside the US and Canada (Facebook, 2015a). Having such expertise for every language where a company operates can become quite costly.

Some companies reported that even if they were to publish how many moderators they employed, there would always be those to whom these numbers would appear insufficient when compared against millions of users. Therefore, revealing these numbers was not deemed to be conducive to a constructive public debate (Milosevic, 2015a).

Outsourcing Moderation

Although the companies I researched tended to use some degree of in-house moderation, the vastness of this effort often required them to outsource their moderation services, and this was overwhelmingly done in
developing countries, resulting in a number of media reports about moderation, which accused the companies of poor outsourcing practices.

In 2012, the UK newspaper the *Telegraph* wrote that Facebook outsourced its moderation to companies abroad where moderators were paid $1 per hour, and that there were no security measures in place to prevent moderators from uploading and sharing this content further or from accessing user data (Barnett & Hollingshead, 2012). According to the *Telegraph* reporters, Facebook had previously disclosed: “No user information beyond the content in question and the source of the report is shared.”

The *Telegraph* claimed, however, that the names of those tagged in the reviewed posts (as well as the users who uploaded the posts) had been clearly visible; as per the article, Facebook explained that such user data was necessary to provide context for the moderators’ decisions. Citing an interview with a moderator conducted by Gawker, and previous Facebook policy statements as well, the *Telegraph* reported that when moderators in places like Africa or Asia review the content, they can ignore it, delete it, or escalate it to employees in California (Barnett & Hollingshead, 2012). The article stated that such moderation was widespread across the Silicon Valley.

Twitter told the *Telegraph* that it had in-house moderation but would not answer the question on outsourcing for the article. An anonymous Facebook moderator also alleged in a blog post that outsourced moderators were paid $1 per hour unless they were American, in which case they were paid $30 (TheInternetOffendsMe, 2013; see also Solon, 2017). This moderator further suggested that Facebook received about 250,000 reportable issues (those that actually violated Community Standards) per hour and that the company relied on automation in the moderation process, so a number of reports were not handled by human moderators.

Facebook reported having very strict rules regarding its outsourced moderation. Its contractors were subject to rigorous quality controls and the company had implemented several layers of safeguards to protect the data of those using their service. Outsourced moderators addressed only simple, clear issues or false reports and forwarded any potentially contentious ones to in-house moderators. No exact numbers of either in-house or outsourced moderators were provided (Milosevic, 2015a). YouTube did not disclose such details and neither did Ask.fm. I asked some of the companies if I could witness the moderation process, but I was not allowed to, and my understanding based on this research is that the companies rarely allow non-employees to observe the moderation process (see Bazelon, 2013b; the
author had the opportunity to witness Facebook’s moderation process and describe it in her book).

It would seem that the less established companies that allowed for anonymity would be even less likely to discuss their moderation practices. Yet one such company, Whisper, was a rare example of a company willing to speak publicly about the exact numbers of moderators it had employed and its moderation process in general. Whisper told Mercury News in 2014 that it employed 120 human moderators to sift through posts in real time (Ortutay, 2014). Articles on the website Gigaom and in the Guardian alleged that Whisper was using TaskUs, an international outsourcing company with a team in the Philippines (DeAmicis, 2014; Lewis & Rushe, 2014 respectively), and the San Francisco-based Metaverse Mod Squad to outsource its cyberbullying moderation (DeAmicis, 2014). As the Gigaom reporter disclosed, “Moderators delete the bad stuff, shuffle cyberbullies into ‘posts-must-be-approved-before-publishing’ category and stamp suicide Whispers with a ‘watermark,’ the Number for the National Suicide hotline” (DeAmicis, 2014).

According to a Wired article from October 2014, Whisper allowed the reporter to view the moderation process and was one of the few companies that practiced “active moderation, an especially labor-intensive process in which every single post is screened in real time” (Chen, 2014). The article did not specify if all real-time screening was conducted by humans. If it was automated, then Whisper was not the only company providing such moderation; Secret did something similar, as I discussed above (at the time of that interview, all of Secret’s moderation was outsourced to companies in Guatemala and the Philippines). According to Secret, the moderators received a few days of training until they learned the workflow and were never shown images, just text; according to the company representative, that diminished the peril of their transgression (e.g., taking someone’s personal photo and sharing it on another platform).

Another start-up revealed that it had employed six in-house moderators for its then couple of million users, the exact number of which the company preferred not to reveal. In addition to employing in-house moderators, the process was also outsourced to a company in the Philippines whose name could not be revealed.

**Proactive Moderation, Machine Learning, and AI**

The type of moderation that occurs if users must first report content, and then wait for moderators to examine it and determine whether it constitutes
bullying, can be described as “reactive moderation.” On the other hand, “proactive moderation” implies the screening of content—even when it is not reported—as a means of cyberbullying prevention.

In September 2016, Wired magazine introduced Google’s project Jigsaw, which had been developing a set of open-source tools called Conversation AI (artificial intelligence) to end harassment through algorithmic learning. Wired wrote that if Jigsaw were to find a way through “free speech paradox . . . it will have pulled off an unlikely coup: applying artificial intelligence to solve the very human problem of being nicer on the internet” (Greenberg, 2016). Informants representing the more established companies that invoked free speech protections would not necessarily see these proactive approaches as effective tools when childhood bullying is concerned (I discuss the particulars in the following paragraphs). Nonetheless, it will be interesting to observe how Jigsaw’s strategy with respect to AI pans out, now that Google has seemingly embraced this approach and made it open source. Many interested parties wonder whether Google, a multinational for-profit company with a hand in digital public sphere, will gain financially from these apparently altruistic Jigsaw tools, as Google eventually hopes to do. Julian Assange, the founder of WikiLeaks, raised similar concerns about the fact that Jigsaw is run by Jared Cohen, a former State Department employee, implying that Cohen is extending government interests via Google (Greenberg, 2016). The downsides of supervised machine learning that the informants lay out below are said to hold for Google’s Conversation AI as well (see Greenberg, 2016).

Supervised machine learning in the context of bullying is a form of proactive content moderation: automatic crawling (monitoring) of the network that would allow for detection of cyberbullying cases as they happen (Dinakar, Jones, Havasi, Lieberman, & Picard, 2012; Xu, Jun, Zhu, & Bellmore, 2012). Hence, the content could be flagged for moderators’ attention as “potentially cyberbullying” even if a user does not report it to the company. This approach can involve the development of an algorithm, a form of automated system to search the content on a given platform for bullying indicators (Dinakar et al., 2012).

Companies that prefer not to use such proactive moderation, may adopt a common rationale to defend their choice. Bullying is context-dependent and varies from case to case, which is why supervised machine learning may result in many false positives (cases misidentified as “bullying”), thus infringing on users’ free speech. Consider, for instance, that young people sometimes use the word “bitch” to mean “friend or mate” which can result in the following problem:
When we were looking at the amount of tweets we were seeing where the word “bitch” was used in an abusive way, the percentage was tiny. It was surprisingly small. . . . It’s just hard to determine the factors [behind bullying], because, as I said before you can have subtweeting, you can have abuse through images not just the key words. . . . So I don’t think we would be satisfied at this time that there’s an automated way to deal with cyberbullying. . . . I think if that technology was extremely accurate so if you could flag an extremely tiny percentage of false positives then perhaps we would consider it. (Patricia Cartes, Twitter, personal communication with the author, June 17, 2014)

But in August 2016 the Verge ran a story about Twitter’s use of an algorithm to filter out abuse from responses to a question-and-answer session with the then US president Barack Obama, “and potentially” from another session with Caitlyn Jenner, a former Olympic gold medalist who underwent a sex change. Some tweets were said to have been moderated manually as well (Robertson, 2016). Twitter’s erstwhile CEO, Dick Costolo, denied such accusations in a tweet as “sensationalist nonsense” (Robertson, 2016). In November 2016, in a follow-up conversation with the author, Twitter representative Patricia Cartes reported that Twitter did analyze patterns of behavior and at times would decrease visibility of certain tweets (however, primarily in connection to violent and extremist behavior). She reported Twitter had used “propriety spam tools to see if we (Twitter) could prevent abuse before it even happens.” This tool, however, was not based on “specific words” but rather on “patterns of behavior” and it was distinct from “supervised machine learning” or “algorithmic learning.” She said Twitter had never done any “manual manipulation of results.”

The company was also accused of banning conservative voices on Twitter after a rightwing blogger, who the Guardian described as a “notorious troll,” was permanently banned from the platform on grounds of abusive tweets (Hunt, 2016). The blogger responded by claiming that the platform was engaging in editorializing its content (Yiannopoulos, 2016).

Bullying does not have to involve swear words; it can be subtle and involve ironic comments or hidden mockery—all cited as another reason why designing an accurate supervised machine learning system would be difficult. I conducted an interview with an academic who researched extensively the use of machine learning for preventing bullying, and who noted that the sophistication of this particular technique was not yet at the level where it could catch such subtle forms of bullying. However, as the informant explained, waiting for a perfect tool is not the right approach on behalf of the companies: “I think what should happen is you go fight with the army you have, not the army you wish [to have].”
Companies report that privacy concerns are another reason why they prefer not to apply these proactive approaches; they are wary of giving users the idea that they monitor content on the platform, especially the content that is not shared by users publicly but only among their friends. This point was reinforced in interviews with NGO representatives and e-safety consultants as well.

Efficiency is also cited as a problem when trying to apply proactive tools: given the large volume of content shared on these platforms, screening it in advance, especially using human moderators, would be almost impossible, as YouTube and Twitter acknowledged.

We don’t do any proactive monitoring on the platform. You know—300 hundred hours of uploaded videos per minute—it’s nearly impossible to do this sort of large crawling that is automatic. (YouTube representative, personal communication with the author, November 12, 2014)

When it comes to digital messengers, because of the private nature of such communication, employing proactive content moderation was regarded by some companies as not only undesirable, but also as unethical, in the words of one company representative:

This is private communication. We’re not Facebook; we’re not Twitter. Because these are not public forums . . . not only is there not a responsibility for us to monitor but we think it would be unethical for us to monitor what’s going on. These are private conversations. (Anonymous, personal communication with the author, November 6, 2014)

Toward Transparency and Evaluation of Effectiveness

A number of important details that emerge behind the scenes in online intermediation by private companies can have significant implications for resolving the cyberbullying incidents that occur on their platforms and affect as well their efforts to create a more dignified paradigm of social relations among youth.

The more established companies realize 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 assure its users that the platform is a safe one, but also in the ability to cite these efforts in front of the public and the regulators as evidence that the company is doing its best to address cyberbullying.

Company documents sometimes single out transparency as an important aspect of the policy and its enforcement. But even the more established
companies provide little or no palpable evidence for the effectiveness of their enforcement tools. Even the results of in-house evaluation efforts tend to be sporadic and not readily available to users.

The promised transparency also belies what the companies do not reveal: the guidelines that their moderators use when deciding whether a case constitutes bullying (or harassment and abuse), for instance, or how many reports are received and processed per unit in time, or what percentage of reports are handled by humans (a point especially relevant for addressing subtle bullying), or any details about the amount the companies invest in e-safety measures. Many companies do not even stipulate what they mean by “bullying,” such as the specific behaviors that their understanding or definitions of bullying, harassment, or abuse entail, which makes it difficult to understand how they approach and practice enforcement.

Particularly important is the move toward “advanced policies,” those that allow the community of users to moderate itself. The language adopted for such policies tends to represent the user as an empowered actor and emphasize the shared nature of responsibility in addressing cyberbullying between users and platforms. The discourse of shared responsibility also signals that part of enforcement responsibility is delegated to users, caregivers, educators, law enforcement (if necessary), and NGOs (see chapter 7). These policies are described as effective because the companies themselves are not always able to determine that a reported case constitutes bullying and hence whether to act on it; a takedown may not solve the offline problem; and users are said to be best positioned to handle such conflicts because they understand the context behind the incident.

A policy that is defined as effective in this way, however, also tends to be efficient for the companies not only because it can cut down on the amount of non-actionable reports, but also because of the responsibilities and amount of work delegated to users. In the absence of an independent evaluation, or even in-house evaluation in the majority 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, especially from users’ perspective.

What such discourse elides, despite the promise of transparency, is how the various enforcement mechanisms could affect the companies’ business models and how such considerations may influence the decisions as to the preferred enforcement mechanism or the one described as “effective.” For example, consider the case of the companies that invoke freedom of speech: Might active content mediation (takedowns, filtering, and proactive content moderation such as in the application of supervised machine
learning) negatively affect their business models (user satisfaction but also data collection)? And how might these considerations play a role in deciding which policy will be adopted and characterized as effective?

Such use of advanced policies also seems to secure less independent scrutiny for the companies who adopt it. The more established a company is, the greater the likelihood that it will have adopted the logic of advanced policies (by using a greater degree of what companies describe as empowerment, and by providing tools for delegation). Some of the newer companies that garnered negative media attention in the context of e-safety may liaise with the more established companies in order to acquire this expertise, or they may partner with e-safety NGOs or hire professionals who are aware of the value of e-safety for the company’s business model.

Because the user base of the new companies can grow very quickly, and faster than the necessary e-safety policies they attempt to put in place, these companies tend to be perceived by experts and sometimes regulators as in need of a greater scrutiny. While this may be the case, the more established companies that tend to have a greater skill at ensuring the perception of e-safety are not independently evaluated either.

This is not to say that the policies in place at more established companies are necessarily ineffective; there is evidence based on some of these companies’ in-house research that specific measures, such as social reporting, could be helping some young users solve their conflicts. Nonetheless, until the industry considers the potential for independently established standards of effectiveness and evaluation—and those standards may indeed be different for each company if they aim to take into account of diverse technological affordances—it will remain difficult to even discuss this concept.