What is a flag for?
Social media reporting tools and the vocabulary of complaint

Kate Crawford
Microsoft Research;
New York University; MIT

Tarleton Gillespie
Cornell University

published in *New Media & Society*, 2014. See http://nms.sagepub.com/

Abstract

The flag is now a common mechanism for reporting offensive content to an online platform, and is used widely across most popular social media sites. It serves both as a solution to the problem of curating massive collections of user-generated content and as a rhetorical justification for platform owners when they decide to remove content. Flags are becoming a ubiquitous mechanism of governance—yet their meaning is anything but straightforward. In practice, the interactions between users, flags, algorithms, content moderators, and platforms are complex and highly strategic. Significantly, flags are asked to bear a great deal of weight, arbitrating both the relationship between users and platforms, and the negotiation around contentious public issues. In this essay, we unpack the working of the flag, consider alternatives that give greater emphasis to public deliberation, and consider the implications for online public discourse of this now commonplace yet rarely studied sociotechnical mechanism.

Keywords

Community, Facebook, flagging, norms, platforms, Twitter, YouTube

An image of two men kissing is posted to Facebook: a still from the British TV show *EastEnders*, the characters are embracing, eyes closed, a kiss on the lips. The author of the Facebook post said he chose the image because he thought it was a fairly uncontroversial one: "The photos I had considered using before I chose that one are much more racy. Oh the irony!"[1] Soon afterward, several Facebook users flagged the image for being graphic sexual content, and it was removed. So began a public controversy in which Facebook was accused of hypocrisy and homophobia, with critics noting that gay kisses were being flagged and removed while straight kisses went unremarked. Ultimately, Facebook was compelled to reinstate the image and apologize (Zimmer, 2011).

"Flagging"—a mechanism for reporting offensive content to a social media platform—is found on nearly all sites that host user-generated content, including Facebook, Twitter, Vine, Flickr, YouTube, Instagram, and Foursquare, as well as in the comments sections on most blogs and news sites. Flagging mechanisms allow users to express their concerns within the predetermined rubric of a platform's "community guidelines." But a flag is not merely a technical feature: It is a complex interplay between users and platforms, humans and algorithms, and the social norms and regulatory structures of social media.

Flags play an important and specific role in the "techno-cultural construct" (Van Dijck, 2013) of social media platforms (p. 29). They are, of course, part of a commonly available suite of tools that facilitate, compartmentalize, and quantify user feedback (Gerlitz and Helmond, 2013). This includes Facebook's "like"

button and “share” mechanisms designed to convey user approval or to republish it on another platform, as well as more intricate ranking tools such as Reddit’s “upvoting” and “downvoting” buttons, that aggregate user reactions to boost the visibility of some content over others. Each of these mechanisms allows users to participate in how the platform content is organized, ranked, valued, and presented to others. And it is a participation that is profoundly guided by the efforts and design of platforms: “Sociality coded by technology renders people’s activities formal, manageable, and manipulable, enabling platforms to engineer the sociality in people’s everyday routines” (p. 12)—a sociality they claim to merely facilitate. Our interest in flags draws attention to one particular element of that coded participation, whereby users participate—or appear to—in the governance of social media platforms and the imposition and reification of community norms.

Flags may have become a ubiquitous mechanism of governance, but they are by no means a direct or uncomplicated representation of community sentiment. While a flag may appear to be single data point—something is reported or it is not—this simplicity belies a tangle of system designs, multiple actors and intentions, assertions and emotions. Engagement with complaints, and the user communities from which they come, is mediated and modulated by the procedures instantiated in the flag interface and the algorithms behind it (Beer, 2009; Malaby, 2006). And they are not stable expressions: Their effects are often uncertain and their meaning unclear. They may harken to other democratic and governance processes, but they do not operate within a transparent or representative system. Echoing Bowker and Star’s (1999) observations of classification and standards, flags are ubiquitous, have a material force in the world, and are indeterminate and steeped in practical politics, yet they have maintained a kind of invisibility from scholarly inquiry.

We argue that the flag represents a little understood yet significant marker of interaction between users, platforms, humans, and algorithms, as well as broader political and regulatory forces. Multiple forces shape its use: corporate strategies, programming cultures, public policy, user tactics and counter-tactics, morals, habits, rhetorics, and interfaces (Burgess and Green, 2009; Lievrouw, 2014; Van Dijck, 2013). By analyzing the ambiguous meanings and effects of flagging content, we aim to bring more clarity to the range of functions that flags play in social media, and the way sites reconcile their inability to directly know or quantify community values with their practical and rhetorical value. Our argument includes both an ideological critique in which we reveal the complex meanings beneath this simple tool, and a pragmatic critique, in which we consider alternatives—not as solutions per se, but to demonstrate that there are other possibilities, each with their own political choices and implications.

We see flags serving two general purposes for social media site operators. First, flagging provides a practical mechanism for addressing the daunting task of regulating such vast and changing collections of content. More than 100 hours of video are being uploaded every minute, as YouTube claims,[2] sites of that scale may have little choice but to place the weight and, to some degree, the responsibility of content decisions entirely on flagging by users. The obligation to police is sometimes a legal one, but social media platforms generally go well beyond what is legally required, forging their own space of responsibility (Grimmelmann, 2010; Tushnet, 2008). For many sites, user input is understood to be a genuinely important and necessary part of maintaining user-friendly spaces and learning from their community. Flags, then, act as a mechanism to elicit and distribute user labor—users as a volunteer corps of regulators.

Second, flagging offers a powerful rhetorical legitimation for sites when they decide either to remove or to retain contentious content, as they can claim to be curating on behalf of their user community and its expressed wishes. Traditional media organizations and regulatory bodies have long faced the challenge of assessing and applying community values to content, be it film, television, and literature. Flags in social media sites can appear, at first blush, to offer a more direct expression of what communities find unacceptable. In short, sites can point to user flagging to legitimize their own content management decisions as a measurable expression of “community standards,” invoking a long Western tradition of media regulation and communication policy that has historically informed both private and public content regulation (Gillespie, 2013; Napoli, 2001). This is doubly meaningful, given that most governments claim enormous difficulty regulating the activities of large social media sites, and these services excel at exceeding traditional state boundaries, policy frameworks, and technical means for pre-release content-rating systems. As yet, no national or international body seems

able to selectively monitor and regulate such volumes of content in real time (Crawford and Lumby, 2013). Thus, flags are a practical and symbolic linchpin in maintaining a system of self-regulation (Price and Verhulst, 2005; Verhulst, 2010)—and avoiding government oversight.

Flagging mechanisms not only individualize expressions of concern, they transform them into data points. Just as users are understood as profiles, and their navigation of the site as traces, their concerns are also apprehended as data: a flag, a date stamp, a classification. This logic of "datafication" (Mayer-Schönberger and Cukier, 2013) poses its own dilemmas, codifying complaints as a legible regime of feeling, and using it to tap into otherwise undetectable patterns of user sentiment (Andrejevic, 2013). These data subsume and come to stand in for the users and their objections. Some of the moderation is then handed off to algorithms, or to human-algorithm hybrids, designed to sort flags into manageable expressions of the user base, prioritized by abstract judgments of content egregiousness and the inferred flag motivation, all designed to simulate human judgments of importance and urgency. Human operators tasked with responding to these flags are similarly bound by the categories and sorting procedures, and find they must respond in algorithmically rule-bound ways: approve, deny, or escalate (Gillespie, 2012).

While flagging and reporting mechanisms do represent a willingness to listen to users, they also obscure the negotiations around them. They leave little room for the articulation of concern, and that articulation is bound up within an individual message to the platform, resulting in an opaque decision by the platform. Disagreements about what is offensive or acceptable are inevitable when a diverse audience encounters shared cultural objects. From the providers' point of view, these disagreements may be a nuisance, tacks under the tires of an otherwise smoothly running vehicle. But they are also vital public negotiations. Controversial examples can become opportunities for substantive public debate: Why is a gay kiss more inappropriate than a straight one? Where is the line drawn between an angry political statement and a call to violence? What are the aesthetic and political judgments brought to bear on an image of a naked body—when is it "artistic" versus "offensive?" The debate over particular values and visibilities, these disagreements are also public negotiations about the contours of public discourse itself. They are a debate about *how* we should debate: what we should be allowed to see, what we will and won't accept as part of the cultural terrain, and the conditions under which content should be excised from public view. Flags matter because social media platforms are not just providers of content or spaces of online activity, but keepers of the public discourse (Balkin, 2007; Baym and boyd, 2012; boyd, 2010; Silverstone, 2007). As such, they have an obligation not just to the individual who posts and the individual who views, but to the broader public life constituted by their interaction.

The limited vocabulary of flags

Flags are a thin form of communication, remarkable more for what they cannot express than what they can. Although their vocabularies have been growing in recent years across a number of sites, they are fundamentally constrained in how they facilitate communication between users and platform operators.

Although not all platforms call this function "flagging" (it is also referred to as reporting), the term and its iconography has come into common usage in the industry and among users. The term "flag" has four significant historical precursors: military, nautical, nationalistic, and bureaucratic. The first use of flags can be traced back to the battle standards of the First Persian Empire and Roman legions, used to differentiate armies during combat, as both co-ordination mechanisms and field signs (Leepson, 2005). By the early 17th century, maritime flags came into common use, first to signal a ship's origin and then as a form of communication at sea. A "red flag" was for sending a warning: marking danger ahead or something treacherous in the waters.[3] National flags only became common at the end of the 18th century, with the rise of nationalism, as markers of state legitimacy and authority. Finally, in 20th century office contexts, one might flag a detail in a document that requires following up or review, a practice that was materialized into an actual flag of sorts with the invention of the Post-It. The purpose being served by social media flags is a curious blend of these histories: They act both as a small, mundane markers of

something to be attended to and adjusted, and an evocative warning sign of danger ahead. And curiously, the flagging mechanism as a whole and that little, pervasive icon has come to represent an assurance of the authority of the platform as a whole.

The specific implementation of social media flagging mechanisms differs between platforms, as does the system of human and algorithmic governance at work behind the scenes. Some sites, such as YouTube and Twitter, rely exclusively on these reports from users as the trigger for reviewing and potentially removing offensive content. Other sites use flags in tandem with proactive review of content, like Facebook. Some sites position a flag alongside each and every bit of content, some allow flagging of a particular user or channel, some provide site-wide feedback mechanisms, and some do all of the above. The specific kinds of inappropriate content being policed—nudity and sexual content, bullying and harassment, hate speech, violence, drug use, self-harm, and terrorism—differ by platform as well. However, the visible mechanisms for identifying and reporting offensive content are very similar across many of these services.

In some sites, a flag constitutes a single bit of expression. For example, the short video service Vine offers users the ability to "report a post." Click, and the offending video is immediately reported. There is no mechanism offered to articulate the reason why the video is being reported, what rule it violates, or how egregious the user found it. There is no way to "unreport" that video once it is selected. It is the simplest expression of complaint: a technical rendition of "I object."

Across the wide range of social media services, some offer slightly more expressive vocabularies by which complaints about content may be articulated, and these have been growing slowly over the years. These vocabularies can be used to trace out the distinct *logics* of flagging. Some, like Pinterest, offer a limited menu of options by which a user can articulate their complaint in a few data points. Tumblr allows users to report each other; once the "report" button is selected, the user is offered the options "Spam," "Harassment," and "No Thanks." This, in effect, is an act of flagging offensive *people* as opposed to just offensive *content* (Figure 1).

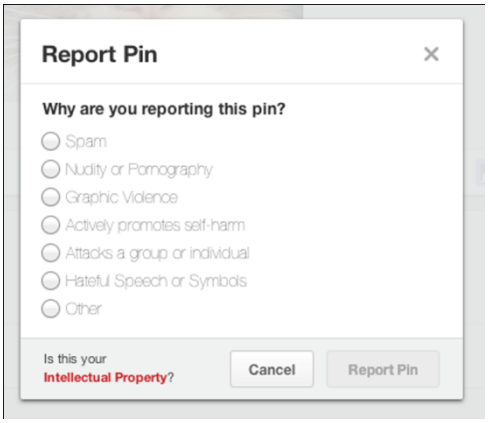


Figure 1. Pinterest flagging pop-up window. Source: Pinterest.

Others, like WordPress, offer a slightly broader menu of concerns: mature content, abusive content, self-harm/suicide, or copyright infringement. If, for example, "mature content" is selected, a text box then appears, where the user must outline in their own words why the flagged content should be considered "mature." Many news and magazine sites that offer comment threads on their articles include a flag, with slight variations in the details of how the system works. Flagging on *The New York Times* includes a simple sub-menu of choices, focusing more on the kind of violations experienced in comment threads: such as vulgarity, going off topic, and being inflammatory. *The Guardian* includes a text box for elaboration and makes typing something in that box mandatory before the complaint can be registered (Figure 2).

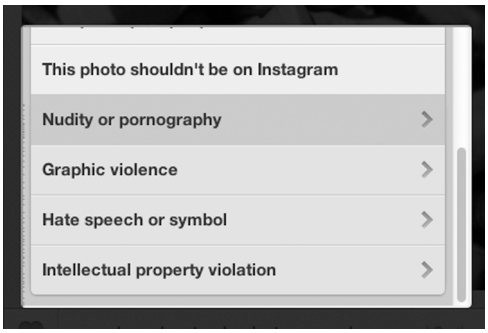


Figure 2. Instagram flagging pop-up window. Source: Instagram.

YouTube has perhaps gone the farthest in expanding its flagging vocabulary, offering a menu of choices, each of which leads to a sub-menu, requiring them to specify the nature of the concern. In 2013, YouTube added a way to indicate the time code of the video where the offending material appears, and a text box for the user to "please provide additional details about" the offending content—in 500 characters or less. This classification and data provided by the user help channel the reports into the categories YouTube prefers, a categorization used in important ways in their review process. For instance, some categories (e.g. "sexual content: content involving minors") are viewed immediately, as platforms have strict legal obligations to report child pornography; videos flagged as "sexual content: graphic sexual nudity" might be prioritized over those flagged as "harmful dangerous acts: pharmaceutical or drug abuse" (Figure 3).

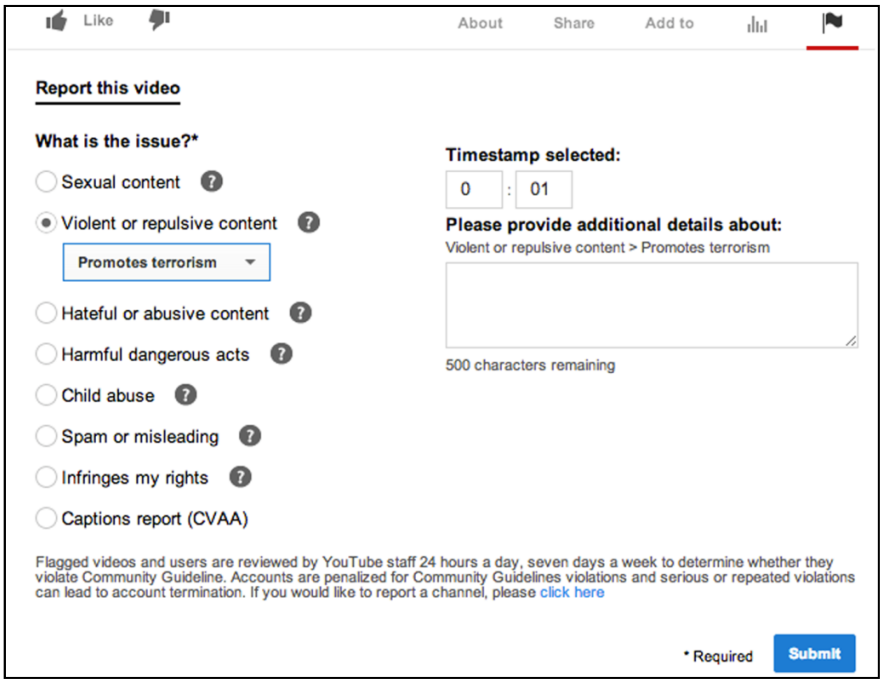


Figure 3. YouTube flagging menu. Source: YouTube.

However, YouTube offers little indication of what happens with the report, how or even if a decision was made to remove content. This is not uncommon. On some sites, a short message appears indicating that the report was received, or thanking the user for their feedback, in the well-worn language of customer service. But

the process by which a flag is received, sorted, attended to, and resolved remains completely opaque to users. The user is left to presume that their report is being added to a queue, perhaps to be adjudicated by a human, or an algorithm, or a combination of both. When a video is removed, there is no specific explanation as to why; often no indication remains that the content even existed.

In terms of process transparency, Facebook has gone the farthest to date. Facebook's flagging apparatus is intricate, in part because the service handles so many different kinds of user content, and in part because it has historically taken a more interventionist approach to moderation. Currently, once a report is made, a "support dashboard" allows the user to monitor the flags they have registered. The dashboard displays a list of what was complained about, when it was logged, and its status in the review process; users can even cancel a complaint before it has been reviewed. In our informal tests, complaints were responded to within 24 hours, with brief descriptions of why an action was or was not taken (Figure 4).

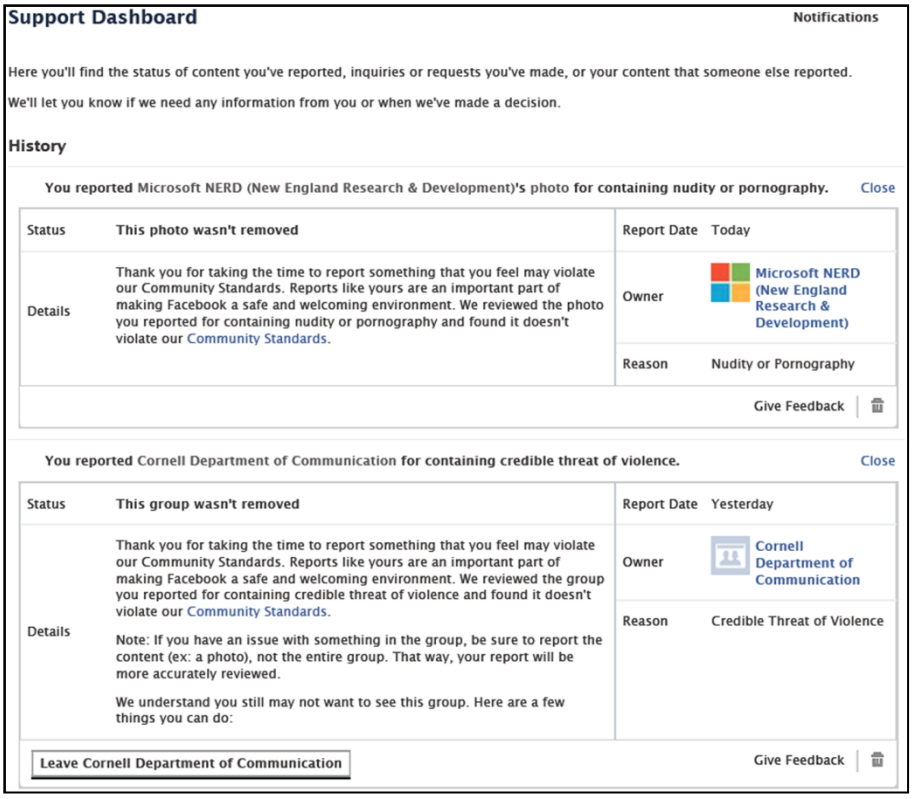


Figure 4. Facebook support dashboard. Source: Facebook.

So while flagging may appear on first glance to be the same on most sites, we can observe a spectrum of procedural micro-practices, from the most narrow feedback channel to more articulated forms (see Appendix 1).[4] Rather than thinking of this spectrum as representing sites that are more or less responsive to users, more or less willing to listen, these design choices are strategic and contextual: differences in types of content matter, as do the sites' age and size. Sites also differ in their philosophical and tactical approach to content moderation, and in the kind of public scrutiny they have received for past efforts. For example, Facebook—arguably because of its size and the fact that its content rules are a more conservative than many other content platforms—has received an inordinate amount of criticism for being not just strict or hypocritical, but unresponsive. Its "support dashboard" that makes visible (if not transparent) its review process is a highly tactical response to that charge.

YouTube has made a very public statement that it does not proactively look for inappropriate content, that it only reviews content that has been flagged; they have also endured a year-long legal battle with Viacom over copyright infringement, in which they have had to claim to be very responsive to copyright takedown notices. Historically, it makes a great deal of sense that their flagging mechanism would be deeply articulated. With its announcement of a curated list of "super flaggers" in 2014, individuals and organizations who are allowed to flag up to 20 videos at a time, YouTube publicly signaled that the value of a flag changes depending on who is doing the flagging. The UK's Metropolitan Police's Counter Terrorism Internet Referral Unit now uses its "super flagger" status to list for removal what it perceives as extremist content. In other words, some flags are worth more than others.

Flickr has, since its inception, taken a different approach to regulating inappropriate content: Most content is allowed, but users must rate their photos as Safe, Moderate, or Restricted. Flickr opts to patrol the rating rather than patrolling content. When a user flags an image in Flickr, the complaint is phrased as "I don't think this photo is flagged at the appropriate level." This particular version of flagging maps the complaint into a framework that understands violations not as "beyond the pale" or "against the rules" but as "incorrectly classified"—leaving the user no way to articulate that egregious material should be off the site, not just "restricted."

However, this seeming expansion of vocabulary and process belies a fundamental narrowness to the flagging mechanism. First, as is clear with Flickr's flagging tool, the vocabulary is prefigured by how the site understands inappropriate content, and how it prefers to deal with its removal or restriction. Categories are not only powerful in the way they leave out things that do not fit; they also embody the structural logics of a system of classification (Bowker and Star, 1999). YouTube's submenus organize according to genres of bad content; Flickr's according to degrees of raciness; the New York Times' according to values of what constitutes proper debate in news spaces; for Vine, the only question is to report or not to report.

But more importantly, flags speak only in a narrow *vocabulary of complaint*. A flag, at its most basic, indicates an objection. User opinions about the content are reduced to a set of imprecise proxies: flags, likes or dislikes, and views. Regardless of the proliferating submenus of vocabulary, there remains little room for expressing the degree of concern, or situating the complaint, or taking issue with the rules. There is not, for example, a flag to indicate that something is troubling, but nonetheless worth preserving. The vocabulary of complaint does not extend to protecting forms of speech that may be threatening, but are deemed necessary from a civic perspective. Neither do complaints account for the many complex reasons why people might choose to flag content, but for reasons other than simply being offended. Flags do not allow a community to discuss that concern, nor is there any trace left for future debates.

The strategic value of thin and ambiguous flags

Despite the thin vocabulary of flags and the opaque adjudication process into which they figure, these mechanisms are asked to do a considerable amount of important work—including policing content, placating users, and suggesting to external bodies that they represent a functioning system of self-regulation. Given this burden, it would be easy to see the narrow vocabulary as either a design flaw, an institutional oversight, or an insurmountable obstacle. Given the enormous number of complaints they receive from users, perhaps social media platforms can only keep up with the leanest forms of feedback. But another possibility is that the thinness of the flagging mechanism may be of strategic value to the sites that employ them.

The flag is merely the first step in a process of content regulation, a process that most sites hope either disappears completely beneath the smooth operation of the site, or when it must appear, to present itself as a rational and fair policing mechanism. But the regulation of contentious user content is an invariably messy process, fraught with the vagaries of human interpretation and shaped by competing institutional pressures. In fact, it benefits social media platforms to retain the ability to make judgments on content removal, based on ad hoc and often self-interested assessments of the case at hand. The opacity of the process means that the

site is not obligated to honor the flags it does receive, and that any decision to remove content can be legitimized as being in response to complaints from the community. Given that flags remain open to interpretation and can be readily gamed, they can also be explained away when the site prefers to ignore them.

Furthermore, flags respond to several legal frameworks that, especially in the United States, draw clear distinctions between user actions and platform responsibilities. Section 230 of the Communications Decency Act immunizes most online platforms from state tort liability (including claims for defamation and invasions of privacy), if they do not participate in the creation of the offending content or if they choose to remove content "in good faith to restrict access to or availability of material that the provider or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected." This parallels the safe harbors provided by the Digital Millennium Copyright Act that immunize online service providers from any monetary liability for copyright infringement as long as those providers play no direct role in the posting of the content and lack the so-called red flag knowledge of infringing content. Since knowledge renders the site open to liability, there is little incentive for sites to review content before users flag it. This has had heavy influence on the design of many US sites where user-generated content is posted (Lee, 2009), reinforcing a "responsive" model of content management that relies on user-initiated actions. These two legal frameworks together reinforce designs and practices that start with the user acting and the platform responding, but require no transparent correspondence between the two.

For most sites, flagged content is first examined by a reviewer employed by the site or by a third-party contractor (Gillespie, 2012). Many decisions end there, with a flagged piece of content either removed or vindicated. But content that falls into a gray area, where the first reviewer is unsure of how to decide, is directed up to the content policy team. There are many reasons why a content policy team might want to retain something that has generated mass complaints. In such moments, it helps a great deal that flags can "mean" a number of things.

For example, YouTube's team have decided to retain certain videos despite having been heavily flagged, leaning on a number of justifications in the process. The widely discussed video of Neda, the Iranian woman caught up in street protests and shot by military police in 2009, was deemed graphic enough to be in violation of their rules on violence, but newsworthy enough to keep anyway.[5] A teen pregnancy public service video produced by the Leicester, UK National Health Service that included simulated footage of a teenage girl giving birth in a schoolyard was first removed after being flagged by users for being too graphic; when YouTube was alerted to its origin, it was reinstated.[6] In such cases, flags are treated as useful data to consider, but can be overruled by situated judgments about the aims and obligations of the site as a public resource.

Further troubling the value of flags is another nearly illegible population: non-flaggers. The number of flags a piece of content receives generally represents a tiny fraction of the total number of views. The degree to which flagged content should be understood as having concerned the community depends, in principle, on some understanding of what proportion of users were disturbed. Views might be a useful stand-in metric here, but for understanding the users' response to a piece of content, they are even more narrow than the flag as an expressive gesture. When, for example, a user does *not* flag a video they just watched, this absence could represent full-throated approval, or tacit support, or ambivalence. Some of these users may have in fact been offended, but did not bother to flag, or did not know they are expected to, or did not know it was "for" them. Some might not believe it would make a difference. Others may have been offended, but also believed politically that the content should remain, or that the site shouldn't be removing content at all. Invariably, the population of non-flaggers is a murky mix of some or all of these. But all the site has, at most, is an aggregate number of views, perhaps paired with some broad data about use patterns. This makes views, as a content moderation concern, an unreadable metric. Neither views nor flags can be read as a clear expression of the user community as a whole.

Flags can be gamed

Despite being characterized as a mechanism of the "community," the flag is a fundamentally individualized mechanism of complaint, and is heard as such. But that does not mean it is always used that way. Flags are also deployed by users in more social and tactical ways, as part of an ongoing relationship between users. Flags get pulled as a playful prank between friends, as part of a skirmish between professional competitors, as retribution for a social offense that happened elsewhere, or as part of a campaign of bullying or harassment—and it is often impossible to tell the difference. Flagging is also used to generate interest and publicity, as in the cases where YouTube put a racy music video behind an age barrier and promoters decried the "censorship" with mock outrage.[7] The fact that flagging can be a tactic not only undercuts its value as a "genuine" expression of offense, it fundamentally undercuts its legibility as a sign of the community's moral temperature.

Furthermore, flagging content as offensive is not just an individual tactic, it can be a collective one. White (2012) notes that, in the personals ads of Craigslist, flagging often serves as a policing mechanism not just for content that violates the site's rules, but for forms of sexual expression that do not conform to traditional gender norms. And there is evidence that strategic, co-ordinated flagging has occurred, widely and systematically. In other words, flagging systems are gamed.

Organized flagging is generally managed surreptitiously, and it depends on the content moderation process being imperfect. While only a few sites prohibit it, most see it as an inappropriate use of the site's reporting tools. Yet, there are several examples of coordinated attacks. In 2012, accusations swirled around a conservative group called "Truth4Time," believed to be coordinating its prominent membership to flag pro-gay groups on Facebook.[8] One of the group's administrators claimed that this accusation was untrue, and that the group had formed in response to pro-gay activists flagging their anti-gay posts.[9] Either way, it seems that surreptitious, organized flagging occurred. Here, the thinness of the flag is of strategic value not for the platform but for aggrieved groups.

Strategic flagging is most prominent when visibility is perceived to be a proxy for legitimacy.[10] As Fiore-Silfvast (2012) describes, a group of bloggers angered by the presence of pro-Muslim content on YouTube began an effort called "Operation Smackdown." Launched in 2007 and active as recently as 2011, the group coordinated their supporters to flag specific YouTube videos under the category of "promotes terrorism" (a submenu under "violent repulsive content"—see Appendix 1). They offered step-by-step instructions on how to flag content, set up playlists on YouTube of the videos they wanted to target, and added a Twitter feed announcing a video to be targeted that day.[11] Participating bloggers would celebrate the number of targeted videos that YouTube removed, and would lambast YouTube and Google for allowing others to remain.

Fiore-Silfvast calls the effort "user-generated warfare," as it had the hallmarks of amateur info-war tactics: "'countering the Cyber-Jihad one video at a time' as the group proudly proclaimed" (Fiore-Silfvast, 2012: 1973). The flags generated by this group are of a very different sort: not expressing individual and spontaneous concern but as a social and coordinated proclamation of collective, political indignation—all through the tiny fulcrum that is the flag, which is asked to carry even more semantic weight.

A more social mechanism for a more social problem

We've seen that flags can have many meanings and functions, and users' engagement with flags can be both individual and social. But even when representing an individual expression, a flag must be read as a signal within much larger, complex social systems. When a flag is used to successfully remove content, it often does so without leaving a trace: Generally, users won't know that the content was removed, or why. This raises a significant question: should the action of flags be made more social and more visible over time? and should the history of how a piece of content has been flagged be made public, or even open to debate, so that communities could discuss whether it is offensive or harmful, or worthy of being kept regardless?

Public debates about the appropriateness of content do occur in online spaces designed to support them, such as Wikipedia, Reddit, and in some multiplayer game worlds. Each has its own distinct characteristics. For example, League of Legends, which claims around 12 million players per day,[12] established "The Tribunal" in 2011, where certain types of flagged in-game behavior can be evaluated collectively by the player community, and determinations made about disciplinary action. However, only frequent game players are eligible to participate in the Tribunal, and this voluntary participation comes with a point structure to provide compensation: Participants get points for being "correct"—which is understood as voting with the majority. Users who consistently vote against the majority will gradually get less access to the reporting process, or their voting access may be removed altogether.[13] The process is not open or visible to all, but it functions as a delimited form of community moderation.

In Wikipedia discussion pages, on the other hand, the quality of content is openly debated and the decisions to keep or remove content on that basis are visible and preserved over time. These backstage discussions reveal the ongoing contests over knowledge that shape the entries that users read, and can be accessed by Wikipedia editors as well as casual readers. The discussion page format lifts the veil from the apparently bloodless recitation of facts that constitutes the main view of a Wikipedia page, to show reversals, fights, or unresolved questions (Reagle, 2004, 2010). It also functions as a space where automated content agents are significant in making updates, and most of the traces of both human and non-human activity are legible to all (Geiger and Ribes, 2010; Niederer and Van Dijck, 2010). Rather than being a space for a form of deliberative democracy, the debates of Wikipedia editors can be understood as more agonistic in character—not leading toward a perfect consensus, but as part of what Mouffe (Castle 1998) describes as an ongoing conflict within "an arena where differences may be confronted." Wikipedia's hope is that this ongoing conflict can produce, as a result or perhaps as a residue, an encyclopedia entry of reasonable quality.

On the other hand, YouTube's current system of content determinations is politically closer to a monarchic structure, in the traditional sense of the "rule of one." A plea can be made, but the decision to remove or suspend is solely up to the platform. When there are debates about the appropriateness of YouTube's decisions, it has to be done outside of the system: taken up by interest groups, on blogs, and in the media. But YouTube's recent addition of the text box to allow users to qualitatively describe the reasons behind a flag suggests different models are possible. Allowing users to detail their concerns offers an expressive value for the user, as well as possibly helping to determine whether content should be removed. Regardless of how YouTube ultimately chooses to heed (or ignore) these written complaints, the option to describe rather than merely flag offending content provides deeper, lasting context for the user and platform. However, it still suffers from being isolated and invisible to other users.

Content platforms could adopt a more Wikipedia-like model, one that was closer to the philosophic ideals of agonistic pluralism. This might involve a space for "backstage" discussion, one that preserved the history of debates about a particular video, image, or post. This would provide a space for engaged debate where objections and counter-objections could be made, rather than the current process, which is inscrutable to other users. This could also allow some complaints to be worked out without site interference, such as when a complaint is withdrawn or an argument is conceded. Users could see where complaints had previously occurred and been rejected, with explanations from the site's policy team when possible, saving them time and uncertainty in the long run. Finally, even if content is removed, debate about it could remain and continue.

However, there are legitimate concerns about open, history-preserving editorial processes. As the number of active contributors in Wikipedia continues to decline, and the gender of the administrators continues to skew male, several researchers have pointed out how quality control mechanisms can be restrictive, and biased against changes introduced by newer editors (Geiger and Ford, 2011; Halfaker et al., 2013). As Van Dijck and Neiborg (2009) point out, for the majority of users, their activity is not driven by a desire for openness and shared knowledge, but by their personal curiosities and hierarchical desires to be a top-posting user (p. 862). Still, a more legible system in the vein of Wikipedia's could support services like YouTube in providing quicker, more transparent responses to users. The intensity and speed

of comments could be used as an additional signal to mark something as problematic, and could also be used to shape the appropriate action. An open backstage area could also be used to legitimize and strengthen a site's decision to remove content. Significantly, it would offer a space for people to articulate their concerns, which works against both algorithmic and human gaming of the system to have content removed.

Above all, a shift to an open backstage model would recognize that social media sites are not only about individual content consumption, but also intensely relational, social spaces where debates over content and propriety play a role in shaping the norms of a site. This contest of ideas could remain even if the content must disappear from view. Decision-making about whether content is seen or not is always a political contest, whether open or not, a choice between counter-posed perspectives where beliefs and emotions may be strongly felt.

Whatever moderation mechanisms these commercial providers craft, they will at some point arbitrate between equally viable and incommensurate positions on morality, political identity, and the norms of the "community" they help constitute. Systems based on a more agonistic ethos might better accommodate intense conflict around content removal, and offer a space of "multiple constituencies honoring different moral sources" (Connolly, 1999: 51). This is particularly significant given that flags by themselves can in fact obscure inherent tensions in user communities: Some may flag religious videos as spreading hate speech,[14] while religious groups may flag political content that they perceive as blasphemous.[15] Some beliefs are fundamentally incompatible. While a social media site must in many cases make a single decision to keep or remove content, visible traces of how and why a decision was made could help avoid the appearance that one perspective has simply won, in a contest of what are in fact inherently conflicting worldviews. A flag-and-delete policy obscures or eradicates any evidence that the conflict ever existed.

Epilogue

In July 2013, after a popular campaign, Caroline Criado-Perez convinced The Bank of England that more women should appear on banknotes. Jane Austen would be represented on the £10, the only woman other than the Queen on national currency.[16] It might seem an uncontroversial decision. But after the news was made public, Criado-Perez received a deluge of violent rape threats on Twitter. Criado-Perez complained to Twitter about the attacks, and noted that their complaint mechanism was inadequate for online abuse of such intensity.[17] After a petition requesting a full review of Twitter's policies garnered more than 120,000 signatures, and various public figures, including members of the UK Parliament expressed concern, Twitter changed its position. It offered a flag to report individual tweets, and more staff to respond to them. Twitter's flag, like others, was borne out of cultural and political strife.

Giving Twitter users the ability to express their concerns about a particular tweet may be seen as a victory. It does represent a significant recognition that action should be taken against online abuse, and it offers a more granular approach to making complaints. But is a flag enough? The Criado-Perez case offers a hint of the difficulties of governing the flows and spikes of obscene and abusive behavior online. Even after anonymous accounts were suspended, the perpetrators sprang up again with new usernames, bragging about their ability to continue their attacks with impunity.[18] And just a few weeks after the Criado-Perez case, Twitter's new flagging mechanism was being used by White supremacists to shut down accounts of feminists who were using the #solidarityisforwhitewomen hashtag.[19] Although online platforms can install them and invoke them as solutions, flags may be structurally insufficient to serve the platforms' obligations to public discourse, failing to contend with the deeper tensions at work.

We have raised several questions about the suitability of flags in the many situations where they are deployed, and whether their actions could be made more legible. Flags proceduralize and perform collective governance while simultaneously obscuring it: Flagged content is not apparent to other users, nor are the reasons for removal or retention made public, and social media sites are intentionally silent on

when and how a flag "counts'. The combination of proprietary algorithms assessing relevance, opaque processes of human adjudication, and the lack of any visible public discussion leaves critical decisions about difficult content in the hands of a few unknown figures at social media companies. Flags may allow us to disapprove of what others say. But without any public record or space for contestation, they leave us little chance to defend the right for those things to be said, or to rally against something so egregious that it warrants more than a quiet deletion.

Finally, to the extent that the flag provides the dominant vocabulary of complaint, users tend to adjust their expression to adhere to that vocabulary. Whether a user shoehorns their complex feelings into the provided categories in a pull-down menu in order to be heard, or a group decides to coordinate their "complaints" to game the system for political ends, users are learning to render themselves and their values legible within the vocabulary of flags. And as long as these large-scale platforms, which host so much of our contested, global public discourse, continue to rely on flags and understand their obligation to respond only in terms of flagging, users wishing to object will find they have little alternative.

Acknowledgements

We are very grateful to our friends and colleagues who read an early draft of this article and offered insightful advice. Thanks to Mike Ananny, Danah Boyd, Brittany Fiore-Silfvast, Lee Humphreys, Jessica Lingel, Kate Miltner, Jonathan Sterne, T.L. Taylor, the graduate students of the "new media and society working group" at Cornell University, and the members of the Social Media Collective at Microsoft Research.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. See http://dangerousminds.net/comments/setting_the_facts_straight_on_the_facebook_fiasco for a full description.
2. See http://www.youtube.com/yt/press/statistics.html for the latest. As of September 2013, YouTube was citing the "100 hours of video are uploaded to YouTube every minute" stat.
3. See the entries for "flag" and "red flag" in the Oxford English Dictionary. Available at: http://www.oed.com/view/Entry/70892 and http://www.oed.com/view/Entry/160299
4. We have tested some well-known social media sites as examples, but bear in mind that these procedures are regularly updated, and may have changed since the writing of the essay.
5. Jemima Kiss, "YouTube looking at standalone 'SafeTube' site for families." *The Guardian*, 29 May 2009. Available at: http://www.theguardian.com/media/pda/2009/may/29/youtube-google.

6. Mark Sweeney, "NHS viral video on teen pregnancy banned by YouTube" *The Guardian* 15 May 2009. Available at: http://www.theguardian.com/media/2009/may/15/nhs-teenage-pregnancy-viral-video-youtube-banned.
7. See, for example: David Itzkoff, "M.I.A. Video for 'Born Free' Is Pulled From YouTube." *New York Times*, 27 April 2010. Available at: http://artsbeat.blogs.nytimes.com/2010/04/27/m-i-a-video-for-born-free-is-pulled-from-youtube/?_r=0. Perez Hilton, "Flagged! Rihanna's New Video TOO Hot For YouTube!" *PerezHilton.com*, 1 February 2011. Available at: http://perezhilton.com/2011-02-01-rihannas-new-music-video-is-deemed-in-appropriate-on-youtube-for-users-under-18-years-old. "New David Bowie video reinstated after YouTube ban." *NME.com*, 8 May 2013. Available at: http://www.nme.com/news/david-bowie/70180. Ken Lombardi, "Justin Timberlake's 'Tunnel Vision' video back on YouTube after brief ban." *CBSnews.com*, 5 July 2013. Available at: http://www.cbsnews.com/8301-207_162-57592492/justin-timberlakes-tunnel-vision-video-back-on-youtube-after-brief-ban/
8. The group was brought to public attention by blogger and activist Alvin McEwen, then profiled on the Huffington Post by Michelangelo Signorile. Alvin McEwen, "Secret religious right Facebook group plotting cyber attack on gay community" *Holy Bullies and Headless Monsters*, 23 April 2012. Available at: http://holybulliesandheadless-monsters.blogspot.com/2012/04/secret-religious-right-facebook-group.html. Michelangelo Signorile, "Truth4Time, Secret Religious Right Facebook Group, Included NOM Co-Founder, Fox News Pundit And More." *Huffington Post*, 24 April 2012. Available at: http://www.huffingtonpost.com/2012/04/24/truth4time-secret-religious-right-facebook-group-_n_1449027.html.
9. Michael Brown, "Gay Activists Expose Secret Rightwing Cabal (Or Not)" TownHall.com, 25 April 2012. Available at: http://townhall.com/columnists/michaelbrown/2012/04/25/gay_activists_expose_secret_rightwing_cabal_or_not.
10. Similar strategic interventions happen more commonly on sites like Reddit and Digg, where users can push good content up or downvote it into obscurity. Both sites prohibit organized and automatic downvoting, and both have faced difficulties where groups organized to systematically vote some kinds of political content off the site. See Peterson (2013).
11. The Smackdowncorps.com website has been taken down, but is available in the Internet Archive. Here are their instructions for flagging: Available at: http://web.archive.org/web/20110717001423/http://smackdowncorps.org/fieldmanual/howto.html. Other blogs and YouTube playlists remain active; see for example: Available at: http://smackdownoftheday.blogspot.com/; http://www.youtube.com/playlist?list=PLF0B6F7B771615DE9; http://my-petjawa.mu.nu/archives/cat_jihadtube.php.
12. Christopher MacManus, "League of Legends the world's 'most played video game'." *CNet/CBS Interactive*, 1 October 2012. Available at: http://news.cnet.com/8301-17938_105-57531578-1/league-of-legends-the-worlds-most-played-video-game/.
13. For details, see http://beta.na.leagueoflegends.com/legal/tribunal/ and http://leagueo-flegends.wikia.com/wiki/The_Tribunal.
14. Jennifer LeClaire, "Are Facebook, Apple and Google Censoring Christian Speech?" *Charisma News,* 25 May 2012. Available at: http://www.charismanews.com/us/33478-are-facebook-apple-and-google-censoring-christian-speech.
15. Tom Zeller Jr., "A Slippery Slope of Censorship at YouTube." *New York Times*, 6 October 2006. Available at: http://www.nytimes.com/2006/10/09/technology/09link.html.
16. Katie Allen and Heather Stewart, "Jane Austen to appear on £10 pound note." *The Guardian*, 24 July 2013. Available at: http://www.theguardian.com/business/2013/jul/24/jane-austen-appear-10-note
17. Jamie Doward, "Twitter under fire after bank note campaigner is target of rape threats." *The Guardian*, 27 July 2013. Available at: http://www.theguardian.com/uk-news/2013/jul/27/twitter-trolls-threats-bank-notes-austen.

18. Alexandra Topping and Ben Quinn, 'Stella Creasy receives Twitter photo of masked, knife-wielding man', *The Guardian*, 6 August, 2013. Available at http://www.theguardian.com/technology/2013/aug/06/stella-creasy-twitter-photo-masked-man-knife.
19. For example, see @ItsJulie1964 requesting her followers to report other Twitter users for spam as retribution. Her account has since been suspended, but reactions to her are still apparent: https://twitter.com/search?q=itsJulie1964&src=typd.

References

Andrejevic M (2013) *Infoglut: How Too Much Information is Changing the Way We Think and Know*. London: Routledge.

Balkin J (2007) Media access: a question of design. *George Washington Law Review* 76(4): 933.

Baym N and boyd d (2012) Socially mediated publicness: an introduction. *Journal of Broadcasting & Electronic Media* 56(3): 320–329.

Beer D (2009) Power through the algorithm? Participatory web cultures and the technological unconscious. *New Media & Society* 11(6): 985–1002.

Bowker G and Star SL (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: The MIT Press.

boyd d (2010) Social network sites as networked publics: affordances, dynamics, and implications. In: Papacharissi Z (ed.) *A Networked Self: Identity, Community and Culture on Social Network Sites*. New York: Routledge, pp. 39–58.

Burgess J and Green J (2009) *YouTube: Online Video and Participatory Culture*. Cambridge: Polity Press.

Castle D (1998) Hearts, minds and radical democracy (interview with Ernesto Laclau and Chantal Mouffe). *Red Pepper*, 1 June. Available at: http://www.redpepper.org.uk/article563.html (accessed 17 June 2014).

Crawford K and Lumby C (2013) Networks of governance: users, platforms, and the challenges of networked media regulation. *International Journal of Technology Policy and Law* 2(1): 270–282.

Fiore-Silfvast B (2012) User-generated warfare: a case of converging wartime information networks and coproductive regulation on YouTube. *International Journal of Communication* 6: 24.

Geiger RS and Ford H (2011) Participation in Wikipedia's article deletion processes. In: *WikiSym '11 / Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, Mountain View, CA, 3–5 October, pp. 201–202. New York: ACM.

Geiger RS and Ribes D (2010) The work of sustaining order in Wikipedia: the banning of a vandal. In: *CSCW '10 / Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, Savannah, GA, 6–10 February, pp. 117–126. New York: ACM.

Gerlitz C and Helmond A (2013) The like economy: social buttons and the data-intensive web. *New Media & Society* 15(8): 1348–1365.

Gillespie T (2012) The dirty job of keeping Facebook clean. In: Culture Digitally. Available at: http://culturedigitally.org/2012/02/the-dirty-job-of-keeping-facebook-clean/ (accessed 17 June 2014).

Gillespie T (2013) Tumblr, NSFW porn blogging, and the challenge of checkpoints. In: Culture Digitally. Available at: http://culturedigitally.org/2013/07/tumblr-nsfw-porn-blogging-and-the-challenge-of-checkpoints/ (accessed 17 June 2014).

Grimmelmann J (2010) The internet is a semicommons. *Fordham Law Review* 78(6): 2799–2842.

Halfaker A, Geiger RS, Morgan JT, et al. (2013) The rise and decline of an open collaboration system: how Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist* 57(5): 664–688.

Lee E (2009) Decoding the DMCA Safe Harbors. *Columbia Journal of Law & the Arts* 32: 233–273.

Leepson M (2005) *Flag: An American Biography*. New York: St Martin's Press.

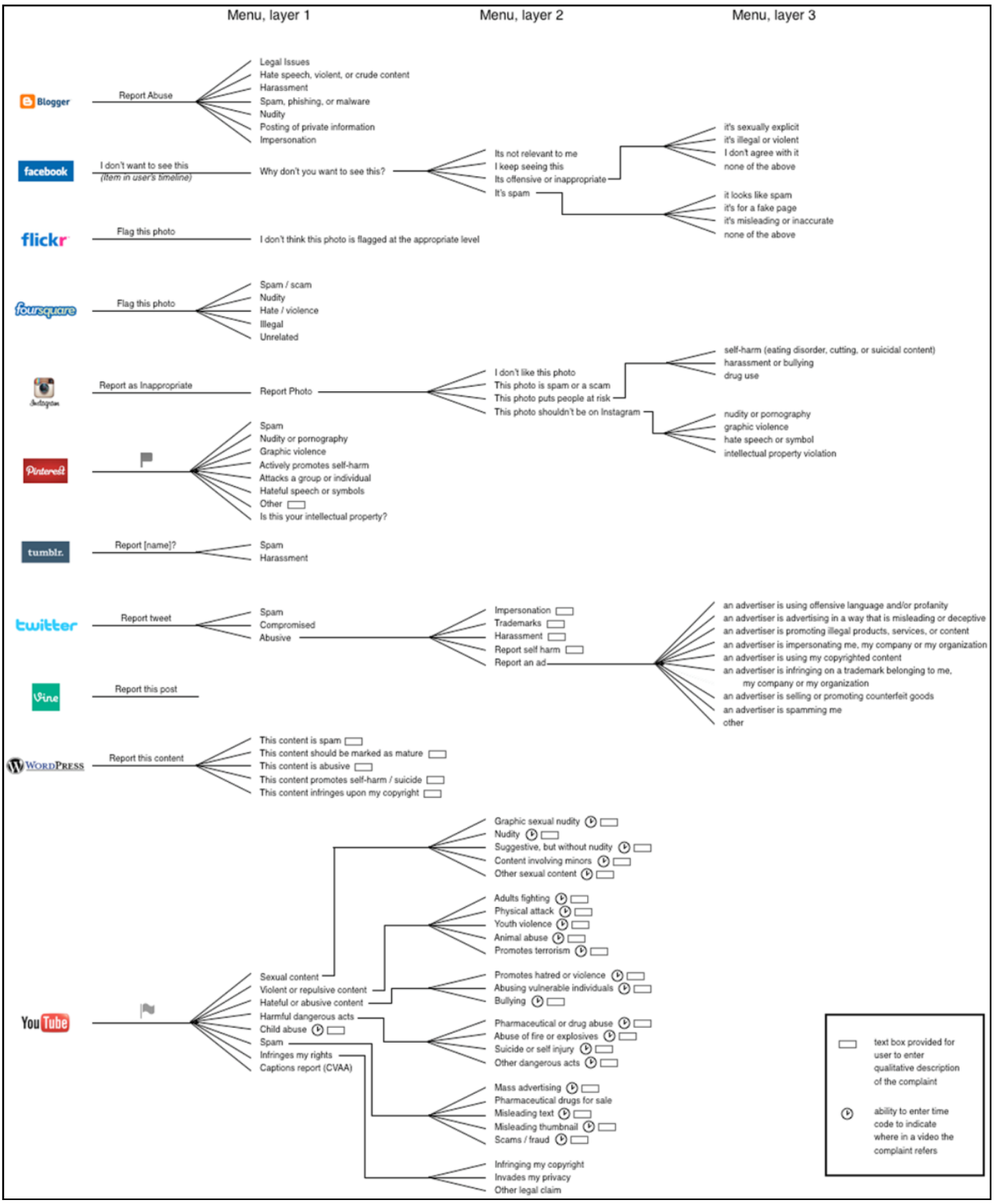
Lievrouw L (2014) Materiality and media in communication and technology studies: an unfinished project. In: Gillespie T, Boczkowski P and Foot K (eds) *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: The MIT Press, pp. 21–51.
Malaby T (2006) Introduction: contingency and control online. *First Monday* (special issue 7). Available at: http://firstmonday.org/ojs/index.php/fm/article/view/1606/1521
Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. New York: Eamon Dolan/Houghton Mifflin Harcourt.
Napoli P (2001) *Foundation of Communications Policy: Principles and Process in the Regulation of Electronic Media*. Cresskill, NJ: Hampton Press.
Niederer S and Van Dijck J (2010) Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society* 12(8): 1368–1387.
Peterson C (2013) *User generated censorship: manipulating the maps of social media*. Masters Thesis, Massachusetts Institute of Technology, Cambridge, MA.
Price M and Verhulst S (2005) *Self-Regulation and the Internet*. The Hague: Kluwer Law International.
Reagle JM (2004) A case of mutual aid: Wikipedia, politeness, and perspective taking. In: *Proceedings of the First Wikimania Conference*, Frankfurt, DE, 5–7 August 2005. Available at: http://reagle.org/joseph/2004/agree/wikip-agree.html (accessed 17 June 2014).
Reagle JM (2010) "Be Nice": Wikipedia norms for supportive communication. *New Review of Hypermedia and Multimedia* 16(1–2): 161–180.
Silverstone R (2007) *Media and Morality: On the Rise of the Mediapolis*. Cambridge: Polity Press.
Tushnet R (2008) Power without responsibility: intermediaries and the First Amendment. *George Washington Law Review* 76(4): 986–1016.
Van Dijck J (2013) *The Culture of Connectivity: A Critical History of Social Media*. Oxford: Oxford University Press.
Van Dijck J and Nieborg D (2009) Wikinomics and its discontents: a critical analysis of Web 2.0 business manifestos. *New Media & Society* 11(5): 855–874.
Verhulst S (2010) The regulation of digital content. In: Lievrouw L and Livingstone S (eds) *Handbook of New Media: Social Shaping and Social Consequences of ICTs (Updated Student Edition)*. London: SAGE, pp. 329–350.
White M (2012) *Buy it Now: Lessons from eBay*. Durham, NC: Duke University Press.
Zimmer M (2011) Facebook's censorship problem. *Huffington Post*, 22 April. Available at http://www.huffingtonpost.com/michael-zimmer/facebooks-censorship-prob_b_852001.html

Authors

Kate Crawford is a Principal Researcher at Microsoft Research, a Visiting Professor at the MIT Center for Civic Media, and a Senior Fellow at New York University's (NYU) Information Law Institute. She researches the spaces where people, algorithms, and data interact. Her work has been widely published in venues such as *Information, Communication & Society*, *Feminist Media Studies,* and the *International Journal of Communication* (2014). In 2013, she received a Rockefeller Foundation Bellagio fellowship for her work on data and ethics.

Tarleton Gillespie is an Associate Professor in the Department of Communication at Cornell University and is currently a visiting researcher at Microsoft Research New England. He is the author of *Wired Shut: Copyright and the Shape of Digital Culture* (MIT Press, 2007) and the co-editor (with Pablo Boczkowski and Kirsten Foot) of *Media Technologies: Essays on Communication, Materiality, and Society* (MIT Press, 2014). He is also co-founder (with Hector Postigo) of the National Science Foundation (NSF)-sponsored scholarly collective Culture Digitally (culturedigitally.org).

Appendix 1: The structure of some platform complaint mechanisms.



Please note that Facebook has slightly different vocabulary and structure depending on whether the user is flagging a news item, a page, a group, or a photo. Here, we have mapped the details from flagging a single item, which is a comparable exemplar.